UNIVERSITY OF SOUTHAMPTON

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING

Electronics and Computer Science

Measuring the Social Influence of Online Communications at the Individual and Collective Level: A Causal Framework

by

Dimitra (Mimie) Liotsiou

Thesis for the degree of Doctor of Philosophy

October 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF PHYSICAL SCIENCES AND ENGINEERING
ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

MEASURING THE SOCIAL INFLUENCE OF ONLINE COMMUNICATIONS AT THE INDIVIDUAL AND COLLECTIVE LEVEL: A CAUSAL FRAMEWORK

by Dimitra (Mimie) Liotsiou

A central problem in the analysis of observational data is inferring and measuring causal relationships - what are the underlying causes of the observed outcomes? With the recent proliferation of Big Data from Web-mediated social communications, it has become important to measure the social influence of online communications, i.e. to determine to what extent online social communications cause certain messages, ideas, behaviours to be widely adopted (to ‘go viral’), and to what extent other causes also play a role. This thesis proposes a critique and a causal conceptual and methodological framework for analysing, measuring and qualifying the social influence of online text-based communications in a given setting, while accounting for the effects of other relevant causes, at the individual and the collective level, based on ‘found’ observational digital data. At the individual level, this thesis demonstrates theoretically and through an analytical discussion how the proposed causal framework can successfully address the key limitations of the popular contagion-based paradigm for online social influence, enabling researchers to disentangle, measure and qualify the social influence of online communications, versus the effects of other (social and non-social) causes. At the collective level, by applying the proposed causal framework, this thesis empirically shows that the assumption of the contagion-based paradigm that the influence of online communications can be measured in isolation, without regard for other causes, does not hold, as it is empirically found that other causes can introduce non-negligible confounding bias to estimates of the social influence of online communications, and that these confounding causes themselves can be stronger causes of the outcomes of interest than online social communications, more robust to bias, with their effects following a much steadier pattern over time. Overall, the proposed causal framework enables researchers to empirically test claims and assumptions about which causes should be accounted for when measuring the social influence of online communications on outcomes of interest, and to pick apart and compare the social influence of online communications versus the influence of other causes, over time and across contexts, at the individual and at the collective level.

Contents

List of Figures

List of Tables

Declaration Of Authorship

I, DIMITRA LIOTSIOU, declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

MEASURING THE SOCIAL INFLUENCE OF ONLINE COMMUNICATIONS AT THE INDIVIDUAL AND COLLECTIVE LEVEL: A CAUSAL FRAMEWORK

I confirm that:

1. This work was done wholly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: Liotsiou et al. (2016).

Signed: . . . .

Date: . . . .

Acknowledgements

For their outstanding and invaluable guidance and support, throughout the progress of this PhD thesis, I would like to express my gratitude and sincere thanks to my supervisors, Prof Luc Moreau, and Prof Susan Halford. Thank you for always being there to provide constructive advice and fruitful recommendations, always with keen attention to detail, towards the methodical development of this thesis.

Moreover, I would also like particularly to express my gratitude and thanks to Prof Luc Moreau, for trusting me and giving me the opportunity to work on this challenging research topic.

I am grateful to the department of Electronics and Computer Science (ECS) at the University of Southampton, and to Roke Manor Research, for providing the funds for this PhD studentship.

Furthermore, I thank Pete Lockhart and Mike Hook, at Roke Manor Research, for always offering useful comments and feedback in our meetings over the course of this research.

Many thanks also to my fellow researchers and members of staff in the Web and Internet Science (WAIS) group, for creating such a pleasant and friendly environment in the lab.

Last, but not least, I would like to offer special thanks to my parents, my grandmother, my uncle and aunt, and to my friends, for their continued support and encouragement, each in their own way, that has meant so much to me, throughout my PhD research.

Thank you all very much!

Nomenclature

ACE	Average Causal Effect
ACF	Abstract Causal Framework
ARD	Absolute Relative Difference
ASE	Average Selection Effect
CCF	Collective-level Causal Framework
DAG	Directed Acyclic Graph
F	In Chapter 5, a variable representing focal item traits, and in Chapters 6, 7, a variable representing the presence of a term in the most frequent terms of the document draft at the end of the current interval
I	In Chapters 6, 7, a variable representing the presence of a term in the most frequent terms of the document draft at the start of the current interval
ICF	Individual-level Causal Framework
LSTM	Long Short-Term Memory
P	Participation variable, for the participation in email communications, in Chapters 6, 7
S	Sentiment variable, for the sentiment expressed in email communications, in Chapters 6, 7
SEM	Structural Equation Model
U	A variable representing factors external to the two individuals under study, in Chapter 5, and external to the collective setting under study, in Chapter 6
W	A variable representing the shared personal traits between two individuals under study, in Chapter 5, and the traits internal to the collective setting under study, in Chapter 6
W_k	A variable representing the personal traits of a person k that are not shared by the other person under study, in Chapter 5
$Y_{j,t-1}$	A variable representing person j's online message (Y) about a focal item (e.g. an idea, opinion, behaviour), at time $t-1$, in Chapter 5
$Y'_{i,t}$	A variable representing person i's adoption outcome, Y', of a focal item (e.g. an idea, opinion, behaviour), at time t, in Chapter 5
Y_{t-1}	A variable representing a group's online communications, about a focal item at time $t-1$ at the collective level, in Chapter 6
Y'_t	A variable representing a group's collective outcome of adoption, Y', of a focal

item (e.g. an idea, opinion, behaviour), at time t, in Chapter 6

Chapter 1

Introduction

This chapter begins by motivating the importance of the research problem of measuring the social influence of Web-mediated text-based communications on outcomes of interest, based on 'found' observational digital trace data. It then presents the research questions that are the focus of this thesis, and subsequently presents the thesis statement. It next outlines the key contributions of this thesis. Finally, it describes the structure of the rest of the thesis.

1.1 Motivation

Social influence has long been an important research topic in the social sciences, in a wide range of disciplines, including sociology, psychology, social psychology, social and cognitive neuroscience, political science, communications, behavioural science, economics, finance, management, and marketing (e.g. Berger, 2013; Dahl, 1957; Falk and Scholz, 2018; Katz and Lazarsfeld, 1955; Kelman, 1961; Lakoff, 2004; Lee and Leets, 2002; Liu et al., 2017; Mason et al., 2007; Parsons, 1963; Robins et al., 2001; Rogers, 2003; Sperber, 1996; Thaler and Sunstein, 2009).

In this thesis, to say that 'A influenced B' is to say that A was one cause of the occurrence of B, whether intentionally or unintentionally, and provided that no force or coercion was used in an attempt to ensure the occurrence of B. Hence, the 'social influence' of A on B can be defined as the role A had in causing the occurrence of B. In the case of 'social influence', A represents a social factor, i.e. a cause from the social world, such as a person's or a group's action(s), behaviour(s), statement(s), message(s), or other communication(s), and B represents another person's or group's action(s), decision(s), opinion(s), belief(s), message(s), or other outcome(s). This definition follows the way in which the term ‘social influence’ has been defined and used in the social sciences.[1]

With the emergence of online communications such as email, and of online social network platforms like Facebook and Twitter over the last decade, Big Data from online social interactions has been produced at an unprecedented volume and detail, offering scientists new kinds of ‘found’ observational data through which to examine social processes. The focus of this thesis is the social influence of online text-based communications (or, for short in this thesis, ‘online social influence’), i.e. the effects of these social communications that happen online, mediated by a Web technology, e.g. communications on a social media platform, or email communications, rather than interactions that happen offline and then digital records of them become archived and posted on the Web.

This proliferation of online communications has led to online social influence becoming an increasingly prominent topic of study in the field of computer science, as well as to the birth of the interdisciplinary field of computational social science (Lazer et al., 2009). In this new field of computational social science, a need has been articulated for methods to be developed that can systematically combine the social and the computational sciences (Counts et al., 2014; Mason et al., 2014; Wallach, 2016).

Understanding social influence is pivotal since it has been claimed that social influence drives the spread of behaviours and attitudes as diverse as smoking cessation, obesity, happiness, loneliness, getting divorced, and political participation, along social ties, in a process analogous to the contagious spread of viruses (Alshamsi et al., 2015; Aral and Walker, 2012; Cacioppo et al., 2009; Centola and Macy, 2007; Christakis and Fowler, 2007, 2008, 2013; Domingos and Richardson, 2001; Easley and Kleinberg, 2010, Chapter 19; Fowler and Christakis, 2008; González-Bailón et al., 2011; Katz and Lazarsfeld, 1955; Kempe et al., 2005; McDermott et al., 2013; Nickerson, 2008), to the extent that ensuring a select few trend-setting individuals (the so-called *influentials*) adopt a behaviour would suffice to lead a large population to follow their example and also adopt this behaviour. If social influence does operate in this contagious manner, then harnessing its power would have an immense impact on areas like marketing, public policy, politics, and public health: it would suffice for one to target a desirable idea, opinion, innovation, or behaviour (e.g. stopping smoking) only at a select few ‘influentials’ and ensure that they advertise it (e.g. ensuring a few ‘influentials’ advertise that they have stopped smoking) in order for a large population to also adopt that desirable idea, opinion, innovation, or behaviour. An important and recent example that highlights the importance of understanding how social influence online operates, and whether it does in fact operate in this contagious manner or not, is the recent concerning evidence regarding Russia’s attempts to influence public opinion in the United States (and in the

[1] Even though there is some degree of variation in how the term has been used in different contexts, this definition captures the elements that these social science definitions generally have in common. More details on the definitions and literature on social influence are given in Chapters 2.3 and 2.6.2, while Chapter 4.1 provides more details on the definition used in this thesis.

United Kingdom) by promoting divisive political and social messages on social media. In this case, it was found that such Russia-sponsored content was seen by millions of social media users in the US in the period before the 2016 presidential elections, as well as after (Levin, 2017; Schrage, 2017; Solon and Siddiqui, 2017). Such incidents have again brought to the fore the issue of the influence of online communications on important outcomes such as public opinion formation, voter turnout and election results.

The above contagion-based paradigm for conceptualising and analysing social influence, which assumes that social influence drives the spread of diverse behaviours and attitudes along social ties in the manner that a contagious virus spreads, has been extensively employed in the context of Web-mediated communications, in theoretical and observational studies of online social influence. Some of the goals of these studies were to 'build models of influence' (Goyal et al., 2010), to identify 'influentials' (Bakshy et al., 2011; Cha et al., 2010) or 'influential spreaders' (Kitsak et al., 2010), to detect online communities (Barbieri et al., 2013), and to characterize (Leskovec et al., 2006, 2007; Liben-Nowell and Kleinberg, 2008), predict (Goyal et al., 2010; Ghosh and Lerman, 2010b), or maximise (Domingos and Richardson, 2001; Kempe et al., 2003) the online spread of information, recommendations, ideas, behaviours, or other actions, or indeed to characterize the 'diffusion of protests' (in González-Bailón et al., 2011), using email data or data from online social network platforms like Twitter, Digg and Flickr.

Contagion-based studies of online interactions (via social media, emails, blogs, or other modes of online communication) such as the above should be applauded for being based on detailed and extensive analyses, using large online datasets, and for contributing valuable insights, and indeed some of the very first insights, on a variety of social phenomena and activities taking place on these new and increasingly pervasive channels of online communication. Still, in this contagion-based paradigm, problematic claims have also been made, equating observed actions with social influence from a specific online source (an online message or set of messages), based on actions that can be readily measured given data from a particular online communications setting, in an ad-hoc manner. This has been done without justification of why such actions can be claimed to represent evidence of social influence from that particular online source, i.e. without data-based evidence (or any other kinds of evidence) that the observed actions are indeed due to social influence from that source and not due to some other causes. There has been a tendency to conflate data and methods with the concept of social influence, where data and methods have been used which are inadequate to make claims on the social influence of online communications. This problematic relationship between the data and methods used on the one hand, and the claims made on the other, leads to a methodological and epistemological question of what data and methods does one need, in order to make different kinds of claims about the social influence of online communications? Indeed, this practice of retro-fitting the concept of social influence to what can be readily measured, without justification, and without testing its assumed validity, has also been pointed out as problematic in the literature (e.g. Freelon, 2014; Marres, 2017; Tufekci, 2014).

Such claims range from equating social influence with the number of, or other quantitative properties of, a user's social network ties to other users, to interpreting as social influence various behaviours and actions that may occur along such social network ties. In the former case, it is far from self-evident why the properties of a person's social ties alone suffice to claim the behaviour or actions of others are due to this person's influence, without having also considered the qualities of the ties, of the people involved, of the context, and of the focal behaviour or action itself (this problem has also been pointed out in, for example, Mason et al., 2007; Freelon, 2014; Shalizi and Thomas, 2011; Tufekci, 2014). In the latter case, some more concrete common examples of how online social influence has been defined in terms of social platform-specific actions are the following: if user j's social connection i mentions the same entity as them (e.g. a URL or a hashtag), within a narrow time window, or if i re-shares or up-votes j's post, or chooses to follow j, or mentions j's username (Bakshy et al., 2011; Barbieri et al., 2013; Cha et al., 2010; Ghosh and Lerman, 2010b; González-Bailón et al., 2011), then i's action is assumed to be due to social influence from j.

One may say that such measures of online activity represent the levels of attention or interest that a given piece of content has received (Ackland, 2013; Freelon, 2014; Watts, 2007, 2011), where this content was posted online by someone, in relation to some topic, in a given context. Indeed, the above studies that apply the contagion paradigm to online communications offer valuable contributions, as they have brought to the fore the social contagion hypothesis, they have applied it in a variety of ways to data from online communications, and they have produced valuable insights in terms of modelling, characterising, or performing statistical (associational) predictions for online attention, social network patterns, or patterns and types of online action and response. However, beyond indicating some degree of attention (paid to some content posted by a given person online, in relation to some topic, and in a specific context), it is far from straightforward to infer the *meaning* or the *causes* behind the measures of observed actions used in the contagion-based paradigm, and indeed Anagnostopoulos et al., 2008 and Bakshy et al., 2011 note that this approach yields an overestimate of the social influence of the online communications under study. In addition, it has been acknowledged that the ideal way to make causal claims in empirical settings is to use controlled experiments, but this can often be difficult or infeasible in practice (Anagnostopoulos et al., 2008; Shalizi and Thomas, 2011; Sharma et al., 2015; Spirtes, 2010).

The difficulty in estimating the social influence of a particular factor of interest based on non experimental, observational data is that the social influence of that factor is only one of many possible causes behind an observed outcome. Rather than social influence from a given factor, there may be other unobserved common causes influencing (causing) the observed outcome and its assumed cause. Observationally determining that a cause of a given outcome is a particular causal factor, rather than any other possible causal factor, or a mix of many of these causal factors, is known to be a very difficult problem (Anagnostopoulos et al., 2008; Aral et al., 2009; Bakshy et al., 2011; Shalizi and Thomas, 2011; Sharma and Cosley, 2016; Sharma et al., 2015).

1.2 Research questions

The research questions that are the focus of this thesis are the following:

How can the social influence of Web-mediated human communications be measured, using 'found' observational (non-experimental) digital trace data? What kinds of methods and what kinds of data are needed to measure it?

'Found' and 'observational' digital traces means data that is non-experimental, i.e. that is not the product of randomised controlled experiments, and that is not the product of surveys, interviews or other similar qualitative methods. In this thesis, 'found' and 'observational' data will be referred to as 'observational' data from now on, as is customary in the computer science and the statistics literature.

As mentioned, the focus of this thesis is social influence from online communications, i.e. the effects of these social communications that happen online, mediated by a Web technology, e.g. communications on a social media platform, or email communications (rather than interactions that happen offline and then digital records of them become archived and posted on the Web).

The thesis statement and contributions, presented below, outline how this thesis aims to answer the above research questions.

1.3 Thesis statement

Given that the popular contagion-based paradigm for measuring online social influence using observational data has been criticised for making untested claims about social influence based on inadequate data and methods, the thesis statement is formulated below:

This thesis challenges the contagion-based paradigm for understanding and measuring the social influence of Web-mediated communications, and proposes an alternative conceptual and methodological framework. The proposed causal conceptual and methodological framework can more accurately measure and qualify the social influence of Web-mediated communications, while accounting for other relevant causes (social or non-social). It can *do this at the individual and at the collective level, based on observational digital trace data.*

That is, this thesis starts with a critique of the core limitations of the contagion paradigm, based on which it then proposes a causal conceptual and methodological framework that can measure the social influence of online communications more accurately than the contagion-based paradigm, as it addresses the core limitations of the contagion paradigm. By applying the proposed framework, this thesis finds that one should measure and account for, in addition to the social influence of online communications, the effects of three other types of causes (personal traits, focal item traits, and external circumstances), in individual-level analyses, and the effects of causal factors such as previous outcomes, in collective-level analyses. Therefore, the thesis provides evidence that other causes should not be ignored in the analysis of social influence online, contrary to the untested assumption of the contagion-based paradigm; rather, since these causes introduce confounding bias to estimates of online social influence, they should be measured and adjusted for appropriately, in order to remove confounding bias from estimates of online social influence, as much as possible. The framework proposed and applied in this thesis enables one to determine what kinds of claims can and cannot be made based on the available data (i.e. which causal factors are and are not measured in the data), and to assess what additional data would be needed in order to strengthen and/or qualify those claims.

In more detail, the framework proposed in this thesis is based on causality theory and is informed by the social sciences, constituting a methodological contribution of the type that is much needed in the emergent interdisciplinary area of computational social science. It is demonstrated theoretically and empirically how this framework can successfully address the core limitations of the contagion-based paradigm, and enable researchers to systematically disentangle, measure and qualify the effects of social influence from online interactions, versus those of other causes, at the individual and at the collective level.

1.4 Key contributions

The above thesis statement is substantiated through the the following contributions.

In brief, this thesis begins by contributing an analytical and empirical critique of the contagion-based paradigm. It then proceeds to formulate a causal conceptual and methodological framework for understanding and measuring the social influence of online communications. Next, the thesis applies this framework, first at the individual and then at the collective level. In both cases, it is found that the assumption of the contagion paradigm that the social influence of online communications can be studied without taking other causes into account does not hold. Rather, it is found that other causes can introduce large confounding bias to estimates of the social influence of online communications, and should therefore be measured and appropriately adjusted for, using causal concepts and methods.

In more detail, the contributions of this thesis are enumerated below.

1. The critique (Chapter 3) first contributes a new and more comprehensive analysis and classification of key conceptual and methodological limitations of the contagion-based paradigm for the social influence online text-based communications. The following four classes of limitations are identified: use of vague language; the assumption that the only cause of an observed outcome is online communications; the assumption that ambiguous outcomes mean adoption, endorsement, or agreement; the practice of taking assumptions as self-evident, rather than testing them empirically or otherwise (e.g. theoretically, based on domain expertise or previous findings). Next, the empirical critique demonstrates how these limitations manifest in practice, attempting to assess whether this paradigm can offer insights about the social influence of online communications on observed outcomes. Using this paradigm's most common measures of the social influence of online communications, it is found that, while valuable descriptive insights are produced, these do not allow for any inferences about whether online communications, or other influencing factors, were behind the observed outcomes.

2. Building on the critique, an abstract causal conceptual and methodological framework is proposed (the Abstract Causal Framework, ACF, in Chapter 4) for defining, conceptualising and measuring the social influence of online communications, which can address the key limitations of the contagion-based paradigm, in a manner that is flexible and not specific to any one platform or channel of online communication. It is 'abstract' (i.e. general) in the sense that it is not specific to individual-level or collective-level analyses only; rather, it is applicable to both. The ACF is comprised of the following principles: having a clear definition of social influence, in line with how it has been understood in the social sciences; distinguishing outcomes from causes (not conflating observed outcomes with a specific cause); accounting for other possible causes behind the observed outcome of interest; applying causal methods for the empirical analysis and measurement of the social influence of online communications.

3. The ACF is instantiated for individual-level analyses, resulting in the Individual-level Causal Framework (ICF), in Chapter 5, and this proposed ICF contributes the following: it covers the space of types causes other than online communications that may be behind observed outcomes, and offers a flexible classification scheme for them; using causal methods, it finds that the social influence of online communications is confounded with the effects of each of these types of causes (namely, personal traits, focal item traits, and external circumstances), hence each of them

should be measured and accounted for; it considers and classifies key qualitative aspects of how different combinations of causes can lead to different qualities in the outcome. The usefulness and versatility of these contributions are demonstrated in an analysis using previous studies of online datasets from different real-world settings.

4. Next, the ACF is instantiated for collective-level analyses (in Chapter 6), resulting in the Collective-level Causal Framework (CCF). The CCF proposed here presents a set of principles for how the influence of online communications can be conceptualised and measured at the collective level specifically (something which has received little attention in the literature). It tailors the general principles of the ACF to collective-level analysis, by addressing issues that are specific to collective-level analyses, including: mapping any variables captured at the individual-level to collective-level variables; determining the appropriate unit of analysis (which is no longer individual people, as it was for the ICF); and offering a flexible classification scheme for possible confounding causes (causes internal to the collective setting, causes external to the collective setting, traits of the focal item). The CCF constitutes a contribution applicable to a wide range of collective settings, i.e. any setting where there is a record of collective outcomes and of online communications, such as projects in organisations and in other professional settings of collaboration, as well as projects in the newer and fast growing areas of crowd-sourcing and citizen science. The CCF is generic and flexible, so that that an investigator may apply and tailor it to any such setting of collectively-produced outcomes. The flexibility and real-world applicability of the CCF is also demonstrated, showing how the CCF can be applied empirically, to a real-world setting of collectively-produced outcomes, using public data, from concepts, to causal modelling, to variable extraction from the data and to the implementation of the causal estimation formulae.

5. Finally, in Chapter 7, the CCF is applied empirically to a real-world setting of collaboration, using public data. This collective framework and its empirical application contribute the empirical finding that the assumption of the contagion-based paradigm that, in analysing the social influence of online communications on outcomes, other causes can safely be ignored, does not necessarily hold. That is, in the particular setting studied, this assumption is found to not hold: there is large confounding bias in estimates of the social influence of online communications, due to other causes; the magnitude of the influence of online communications is small compared to the effects of other causes, less robust to confounding bias, and less stable over time; ignoring confounding causes leads to very different patterns and conclusions; and a model accounting for other causes better fits the empirical data than the implied model of the contagion-based paradigm. In addition, the influence of online communications is found to vary over time and across contexts,

> dimensions which have received little attention in the contagion-based paradigm. Therefore, this chapter contributes the finding that, contrary to the contagion-based paradigm's assumptions, causes other than online communications cannot be safely ignored, as such causes may introduce confounding to the estimate of the social influence of online communications on outcomes, and this confounding may be large. Moreover, these other causes may even be stronger causes than online communications, and their effects might even be more robust to confounding and more steadily evolving over time (as happens in the setting studied here) than the effects of online communications. In addition, the empirical application of the CCF in this chapter also serves to demonstrate how the collective-level framework can be employed and adapted to settings of Web-mediated or Web-assisted collective action, and how it enables one to measure and compare the social influence of online communications, versus the effects of other causes, on the outcome of interest, over time and across contexts, based on observational digital trace data.

It is noted here that empirically analysing and measuring the social influence of online communications on real-world outcomes of interest is often difficult, as the online communications data and/or the outcome data is often held privately and not available to researchers, often due to (at least in part) commercial or privacy concerns (Barbieri et al., 2013; Marres, 2017, p. 17). For instance, Facebook and Twitter limit the data they make available to external researchers, and the outcomes observable in such data are typically ambiguous. Organisations typically do not publish their employees' internal online communications, such as emails or Slack messages, or project outcomes (particularly not over time). Similarly, individuals do not typically make publicly available detailed records of all their online communications or of their personal outcomes such as beliefs, opinions, attitudes or actions. For the empirical analyses in this thesis (in Chapters 3.2, 6, 7), real-world observational digital traces from the public online archives of a World Wide Web Consortium (W3C) Working Group are used, as a case study, specifically the Provenance Working Group.[2] W3C Working Groups are settings of collective decision-making and collaboration, where domain experts in a topic area participate in order to produce a set of Web standardisation documents, in this case on the topic of Provenance on the Web. Furthermore, W3C Working Groups are settings of international collaboration, and lead to the production of Web standards that can shape Web practices and Web usage worldwide. (More details on this dataset are given in Chapters 3.2.1 and 6.2.) This public dataset was chosen because it offers the advantage that the outcomes of interest are captured in the data, rather than assuming unmeasured outcomes can be inferred by using the ambiguous outcomes in the data as proxies for adoption or endorsement outcomes, as is commonly the case in the contagion-based paradigm. An additional advantage of this dataset is that outcomes and online communications are captured over time, and also across contexts (different sub-groups of

[2] https://www.w3.org/2011/prov/wiki/Main_Page

people working on separate standardisation documents), enabling one to investigate how the social influence from online communications on the outcomes may vary over time, and across different contexts, dimensions which are often not considered in contagion paradigm studies of the social influence of online communications. Hence, the public and relatively rich dataset of this Working Group serves as a case study that allows one to test the contagion paradigm's assumptions about the social influence of online communications, by measuring the influence of online communications, as well the influence of other factors, on the nature and content of outcomes produced collectively (the outcomes in this case being global-reaching Web standards).

1.5 Structure of the thesis

The rest of this thesis is structured as follows.

Chapter 2 discusses the literature landscape and background for social influence on the Web, including relevant methodological paradigms for conceptualising and for measuring the social influence of Web-mediated interactions on outcomes of interest.

The contributions of this thesis are presented in Chapters 3-7.

Chapter 3 presents an analytical critique of the key limitations of the contagion-based paradigm for analysing social influence in online interactions, followed by an empirical critique based on a worked example to help empirically assess the merits and limitations of applying such a paradigm.

Chapter 4 introduces a causal conceptual and methodological framework (called here the Abstract Causal Framework, ACF) for conceptualising and measuring the social influence of text-based online communications, based on observational digital trace data, which can address the limitations of the contagion-based paradigm identified in the critique.

Next, Chapter 5 presents an instantiation of the ACF for the analysis of the social influence of online communications, and of other causes, on the outcome of interest, at the individual level. This instantiation is called here the Individual-Level Causal Framework (ICF). Most of the work in this chapter was published as a poster paper in *Dimitra Liotsiou, Luc Moreau, and Susan Halford. Social influence: From contagion to a richer causal understanding. In* International Conference on Social Informatics, *volume 10047, pages 116-132. Springer, 2016.* (Liotsiou et al., 2016), which was honoured with the Best Poster Award for the accompanying poster.[3]

Chapter 6 presents the Collective-level Causal Framework (CCF), i.e. the collective-level instantiation of the abstract causal framework (ACF) of Chapter 4, for the analysis of

[3] See https://web.archive.org/web/20170208203814/http://usa2016.socinfo.eu/.

the social influence of online communications, and of other causes, on the outcome of interest, at the collective level. It also discusses the nature and the data of the real-world collective setting to which this CCF is next applied, the W3C Provenance Working Group, and describes the details of the empirical application of this CCF to this dataset.

Chapter 7 then presents the analysis and results from applying the CCF to this real-world collaborative setting, including measurements and patterns of the influence of online communications on the outcome of interest, versus the influence of other causes, over time and across contexts, together with the discussion and evaluation of these findings.

Finally, in Chapter 8, the key findings and contributions of this thesis are summarised, followed by an outline of possible future directions and a presentation of concluding remarks.

Chapter 2

Background

This chapter reviews the literature on social influence, particularly in the context of Web-mediated interactions. It discusses how social influence has been conceptualised and measured in the literature, and the practical challenges in measuring it, using observational data.

First, the context of Web-mediated interactions is presented, and how social influence has been studied in that context, in Section 2.1. Section 2.2 considers how online social influence (that is, the social influence of Web-mediated communications) has been studied at the collective level, an area that has received relatively little attention compared to individual-level analyses of online social influence. Next, Section 2.3 investigates how social influence can be defined. Then, in Section 2.4, a discussion follows of how social influence on the Web has been conceptualised and measured under the popular contagion-based paradigm, with Section 2.5 discussing known limitations of this paradigm. Finally, the causal methodological approach to conceptualising and measuring social influence is presented in Section 2.6, discussing how this paradigm can be used to address the limitations of the contagion-based paradigm. The chapter concludes with a brief summary, in Section 2.7.

2.1 The study of social influence in Web-mediated interactions

Over the last decade, the Web has become more and more embedded in people's lives (at least in parts of the world where Internet access is common), and today it is common for people to exchange information, interact socially and collaborate over the Web, in some form or other. Unlike offline forms of interaction, the Web continuously produces digital traces of the interactions it mediates, in real time and over time, at a previously unprecedented level of detail and scale, and such traces are widely archived (e.g. by social media platforms, Internet service providers, organisations, government and administrative authorities). Social influence on the Web has been studied widely, focusing on online communications via email, or more recently, on social networks like Twitter, since, over the last decade or so, such modes of online communications have emerged as important channels through which millions of interactions happen every day, producing large and detailed digital traces that scientists can study.

By examining these traces, scientists aim to understand the types and evolution of actions, discourses, relationships and processes that occur in these dynamic and often complex online settings. Given the prevalence of the Web, it has become important to understand how such complex systems of online interactions behave and evolve, as these encompass millions of vital activities of our everyday lives, from professional collaboration, to informal, but often large-scale, online interactions. Identifying and understanding the processes of social influence in such social systems is important in the effort to understand how people interact, how they form opinions and make decisions, and how their opinions and decisions are affected by those of others and by events and information they observe. Understanding social influence is also important when trying to examine why certain pieces of information, rumours, trends, ideas, opinions, behaviours, innovations, or products spread widely (e.g. become culturally or socially significant and influential over a specific time period), whereas others fail to be adopted by a large mass of people. Examining forms of social influence has long been central in fields as diverse as communication (Watts, 2007), business and marketing (Berger, 2013; Domingos and Richardson, 2001), sociology (Granovetter, 1973), social psychology (Mason et al., 2007), political science (Dahl, 1957; Margetts et al., 2015), and medicine (Christakis and Fowler, 2007).

As discussed in Chapter 1, understanding social influence online is important given that it has been claimed that social influence drives the spread of many different behaviours and attitudes, ranging from smoking cessation, to obesity, happiness, loneliness, getting divorced, and even to political participation. It has been claimed that social influence drives the spread of these behaviours along social ties in a process analogous to the contagious spread of viruses (e.g. Alshamsi et al., 2015; Christakis and Fowler, 2007, 2013; Katz and Lazarsfeld, 1955; Kempe et al., 2005; McDermott et al., 2013; Nickerson, 2008), to the extent that ensuring a select few trend-setting individuals (termed *influentials*) adopt a behaviour would suffice to lead a large population to follow their example and also adopt this behaviour. Therefore, as discussed, if social influence does operate in this manner, then harnessing its power would have an immense impact on areas like marketing, public policy, politics, and public health. In that case, it would suffice for one to target a desirable behaviour only on a select few 'influentials' in order for a large population to also adopt that behaviour.

Given such strong and broad claims, it is important to examine and test whether social influence does indeed work in a contagious manner (as has been called for in the literature, e.g. in Freelon, 2014; Tufekci, 2014), and whether the assumptions of this contagion-based paradigm for social influence are accurate and hold empirically.

If the contagion-based paradigm is found to be limited and to not hold empirically, it is important then to understand under what circumstances might certain behaviours and actions be successfully adopted by a large population, and what is the role of social influence, and of any other relevant factors, when studying the occurrence of an outcome of interest. For example, as mentioned, in the context of political participation, there has recently been concern regarding Russia's attempts to influence public opinion in the United States (and in the United Kingdom) by promoting divisive political and social messages on social media, where it was found that such Russia-sponsored content was seen by millions of social media users in the US in the period before the 2016 presidential elections, as well as after (Levin, 2017; Schrage, 2017; Solon and Siddiqui, 2017). Such incidents have again brought attention onto the impact of social media on important outcomes such as public opinion, voter turnout and election results, and to answer such questions one would need to measure the extent of influence from such divisive online content (and, as this thesis stresses, to also disentangle that from the influence of other relevant factors, to try and determine how much impact such social media posts ended up having, versus how much impact other relevant causes had).

In the field of computer science, and in the emergent computational social science field, social influence is typically conceptualised in terms of information propagation which is assumed to occur contagiously along social network ties. So it is assumed that, as viruses spread contagiously from one person to another, so do pieces of information, actions, behaviours, opinions, and attitudes spread like a virus from one person to another. In this contagion-based paradigm, it is often assumed that people's actions, choices and behaviours are caused by another person, the so-called *influential*, who had previously endorsed the same action, choice, or behaviour. So, studies of social influence that use this contagion-based paradigm for social influence often assume that the cause of a person's observed action is another 'influential' person. In this vein, common topics of study include the 'theory of influentials' (Watts, 2007), and the 'influence maximisation problem' (Kempe et al., 2003; Domingos and Richardson, 2001), where the latter is about finding who these influentials are who can make the adoption or endorsement of a focal item such as a message, product or behaviour spread the furthest, based on assumptions about people's propensity to adopt a focal item that has been advertised by someone in their social network.

Empirically, under this paradigm, a person's online social influence has commonly been measured in terms of their social network ties on a given social media platform (followers or friends, depending on the platform), and/or in terms of the attention or response levels attained by a person's posts (response messages, likes, comments, retweets, reshares, votes, and so on, depending on the online medium) (e.g. Bakshy et al., 2011; Cha et al., 2010; Cheng et al., 2014; Ghosh and Lerman, 2010b; González-Bailón et al., 2011;

Watts, 2007). Common properties of social ties or responses measured are their number, and the properties of the network they form, particularly the person's position in the network formed by these social (friend or follower) ties or response ties, particularly various network centrality measures. That is, one's social influence is measured in terms of the audience that was *actually* reached (responses), or in terms of the audience that could *potentially* be reached (followers or friends).

This is the established framing and conceptualisation of the notion of social influence in Web-mediated interactions. But is this conceptualisation an accurate representation of social influence on the Web? It has been acknowledged in the literature (e.g. Aral et al., 2009; Bakshy et al., 2011; Shalizi and Thomas, 2011) that the assumptions on which this paradigm is based, particularly the assumption that the cause of one person's action is an 'influential' in their online social network, do not always hold, and in fact often it is very difficult to establish with any reliability that these assumptions hold. In addition, other criticism and limitations include the ambiguity of online actions, such as following someone on social media, resharing a social media post, or responding to an online message or post, used as outcomes (observed actions) of endorsement or adoption, in efforts to measure the social influence of online communications on the adoption or endorsement of information, ideas, behaviours, or people (Ackland, 2013; Avnit, 2009; Cha et al., 2010; Tufekci, 2014). For instance, Ackland (2013) notes how 'influence in social media environments in largely conceptualised as attention', and the appropriateness of influence outcomes on Twitter is discussed, saying 'In Twitter, how do we know when someone has been influenced, that is, what is the appropriate outcome? It appears that the best measure of influence in Twitter is retweeting, since this is a clear indication that someone has made a conscious decision to pass information on'. As this thesis will discuss in more detail, in this chapter and the next, it is worth considering whether and to what extent such evidence of attention are adequate to make claims on social influence.

This thesis investigates the above question of whether this conceptualisation is an accurate representation of social influence on the Web. As shall be described in the following chapters, this thesis finds that this conceptualisation is rather limited, and offers at best a partial view of online social influence, and at worst a severely biased view of online influence. By failing to consider and appropriately adjust for other causes of the observed actions and behaviours, and assuming that other causes can be ignored, the so-called *confounding bias* from other causes is left unaccounted for, leading to potentially severely biased measurements of online social influence, and to a lack of insights on how strong or weak a cause online social influence is compared to other causes.

Given that this is the established paradigm for social influence, and having raised the question of whether it is adequate in conceptualising social influence online and in measuring it, the next section takes a step back, to consider established and commonly accepted definitions of social influence, in order to help clarify what a conceptualisation and measurement methodology of social influence should capture. Then, the following two sections revisit the contagion-based paradigm, describing its methods and assumptions in more detail, as well as its limitations.

2.2 Social influence in collective online settings

Studies of online social influence have largely taken an individual-level perspective of online interactions (e.g. on online social networks, over email), modelling people as nodes and their social ties or interactions as edges in a network, and analysing node-to-node social influence (Tufekci, 2014), and individual actions such as an individual 'liking' or resharing another's post. Collective-level studies, studying actions at the collective rather than the individual level (e.g. majority outcomes such as the results of elections; outcomes that are results of group effort such as the outputs produced in organised collaboration, professional or otherwise), are rare in comparison. This focus on individual-level analysis of node-to-node interaction may be reinforced by factors such as the proliferation of social media platforms where social interactions are already represented in terms of networks of individual users, or due to the interest in applying findings on social influence to online marketing and advertising of products (Barbieri et al., 2013; Berger et al., 2008; Domingos and Richardson, 2001; Goyal et al., 2010; Keller and Berry, 2003; Kempe et al., 2003; Leskovec et al., 2007), as indeed some of the earliest studies on social influence took an individual-level approach to social influence and studied the spread of innovations and products (Rogers, 2003 (first published in 1962); Strang and Soule, 1998; Katz and Lazarsfeld, 1955). Indeed, these marketing applications have been quite central in the development and usage of models for social influence, to the extent that, for example, Goyal et al. (2010, p. 1-2) go as far as to claim that 'any models proposed for influence should be compatible with the assumptions made in applications such a[s] viral marketing', where it is later explained that the assumptions made in viral marketing are of contagious spread of product adoptions due to social influence from one's immediate social network ties, as per threshold models and per the influence maximisation problem.

However, even if not studied as extensively, there are also several collective-level analyses of the influence of online (and offline) social interactions on outcomes of interest. For instance, some studies designed and deployed so-called *sociometric badges* in organisations to capture (offline) face-to-face communication and study the relationship between collective behaviour and performance outcomes like productivity or job satisfaction (Alshamsi et al., 2015; Olguín and Pentland, 2010). More generally, there have been several attempts to study the influence of online social communications of political candidates on Twitter on whether they will be elected (Bright et al., 2017; Metaxas et al., 2011). Yet other studies have focused on statistical prediction based on social media data of outcomes even further removed from the social media platform and the users' whose social posts are analysed, to predict (i.e. find associations with) outcomes such as box-office movie success (Mishne et al., 2006), and the stock market (Bollen et al., 2011; Gilbert and Karahalios, 2010).

In the context of Web-mediated collaboration, Jensen et al. (2000) studied the influence of different kinds of online communications on the level of cooperation in online environments. Indeed, the Web today also mediates many kinds of group-level interactions and collaboration efforts, with various types of online platforms hosting different, and often new, types of collective projects. Such online collective settings that have received considerable researcher attention include crowdourcing platforms, and citizen science platforms.

Crowdsourcing is a sourcing model whereby individuals or organisations use contributions from Web users to obtain needed services or ideas (Estellés-Arolas and González-Ladrón-de Guevara, 2012; Taeihagh, 2017). Wikipedia is perhaps the best known crowdsourcing platform, where the goal is for Web users to voluntarily populate Wikipedia with encyclopaedic articles, and where contributors to each article can also communicate with each other online (on the so-called Talk pages[1]). There are also crowdsourcing marketplaces (such as Amazon Mechanical Turk[2]) that connect crowdsourcing workers with tasks commissioned by companies or other institutions, however in these cases crowdsourcing workers work independently and usually cannot directly communicate with each other on such marketplaces.

In addition, citizen science, also known as crowdsourced science, is one type of crowdsourcing where members of the public help in scientific research by completing specific tasks (Bonney et al., 2009; Silvertown, 2009). Examples of citizen science projects include various astronomy projects (e.g. Galaxy Zoo[3])), projects to map the brain (Eyewire project[4]), and projects to raise awareness on light pollution (the 'GLOBE at Night' project,[5] Kyba et al., 2013), and the Zooniverse platform which hosts various scientific citizen science projects from various research disciplines.[6] Citizen science platforms have already attracted the interest of researchers, with existing studies on, for example, how the social interaction functionality provided on crowdsourcing platforms may be associated with the quality of the collective outcomes produced (Tinati et al., 2015, 2016). Besides such studies, however, the main form of influence that has been studied on such platforms relates to incentivisation strategies, i.e. how can the platform designer provide appropriate incentives to the workers, financial or otherwise, in order to entice them to produce work of good quality, under certain resource constraints (e.g. Allahbakhsh et al., 2013; Gao et al., 2015; Kamar and Horvitz, 2012).

[1] https://en.wikipedia.org/wiki/Help:Talk_pages
[2] https://www.mturk.com/mturk/welcome
[3] https://www.galaxyzoo.org/
[4] https://eyewire.org/explore
[5] https://www.globeatnight.org/
[6] https://www.zooniverse.org/

Overall, the area of collective-level online social influence has received relatively little attention (compared to individual-level analyses). This thesis will study this topic, in Chapters 6 and 7, using causal analysis methods.

The remainder of this chapter focuses on the much more popular and more developed area of individual-level analysis of social influence, which has been extremely active, and has made very big claims on the concept and processes of online social influence.

2.3 Defining social influence

An important question one needs to address before studying online social influence is, what do the terms 'influence' and 'social influence' mean? Social influence online has come to typically be thought of as information propagation potential through contagion, but the accuracy and adequacy of this conceptualisation has been criticised, as mentioned above and as shall be discussed further in the following two sections. The term 'social influence' is rather broad, as it describes an intangible and complex social process. As discussed, this term has often been used simplistically (equated with number of followers or number of responses on social media, or equated with particular network topology measures from network theory, e.g. in Ghosh and Lerman, 2010a; Ghosh and Lerman, 2014; Kitsak et al., 2010; González-Bailón et al., 2011). It has typically been defined and used in context-specific terms, with reference to the actions allowed by the online social platform under study. For example, social influence has been defined in terms of the action of up-voting an article on Digg (Ghosh and Lerman, 2010b); in a different study, a separate type of a person's influence has been defined for each of three types of action on Twitter: following, retweeting, or mentioning that person (Cha et al., 2010). Therefore, care is needed in defining and using this term.

Influence is defined in the Oxford Dictionaries as 'The capacity to have an effect on the character, development, or behaviour of someone or something, or the effect itself [...]'. [7] In the Merriam-Webster Dictionary it is defined as 'the act or power of producing an effect without apparent exertion of force or direct exercise of command' and 'the power or capacity of causing an effect in indirect or intangible ways'.[8] The causal nature of influence is already apparent in the above definitions, and indeed the word 'causing' is used in the Merriam-Webster definition. (This causal nature will be discussed further in Section 2.6.)

In the social sciences and humanities literature, similar definitions can be found in Leenders (1997) (as cited in Robins et al., 2001), where social influence is defined as: 'Social influence occurs when an individual adapts his or her behavior, attitudes or beliefs to the behavior, attitudes or beliefs of others in the social system'. In Cha et al.

[7] http://www.oxforddictionaries.com/definition/english/influence

[8] http://www.merriam-webster.com/dictionary/influence

(2010) (from computer science), influence is defined as per Katz and Lazarsfeld (1955), an important work from the field of communications, as ‘capacity of causing an effect’, very similarly to the above dictionary definitions. From the field of philosophy, Morriss (1987) defines influence as a form of causation, occurring in a possibly covert, unclear, or unintentional way, that does not involve force or coercion (unlike, for example, the related concept of power, which may involve force or coercion, per the definition in Dahl, 1957, a heavily cited political science paper). Similarly in Rogers (2003), another very influential work from the communications literature (first published in 1962), the term ‘social influence’ is used in the sense of a person causing another person to change their behaviour, through the use of appropriate incentives. From the area of communication studies again, in Parsons (1963), influence is defined, first broadly, as the generalised mechanism by which attitudes or opinions are determined. It is then defined more specifically, for the purposes of that particular study, with an element of intentionality, as ‘a way of having an effect on the attitudes and opinions of others through intentional (though not necessarily rational) action – the effect may or may not be to change the opinion or to prevent a possible change’.

The causal nature of the term ‘influence’ is also well accepted in the causal methods literature (e.g. Morgan and Winship, 2014; Pearl, 2009a; Shalizi and Thomas, 2011; Thomas, 2013). Notably, Pearl (2009a), in his discussion of associational versus causal concepts, lists ‘influence’ in the causal concepts. Specifically in relation to the contagion-based paradigm for social influence, Thomas, 2013 in his critique of the social contagion theory in the context of the widely publicised and scientifically controversial work of Christakis and Fowler (specifically responding to their discussion paper Christakis and Fowler, 2013, as well as referring to their monograph Christakis and Fowler, 2009, both of which discuss their numerous previous studies) stresses that ‘The word “influence” should be used exclusively in the causal manner in which it has been historically known’ and states that ‘The word “influence” is a well-established proxy for a directional causal effect’. This causal nature and usage of the term ‘influence’ is also recognised and adopted in this thesis.

The causal nature of influence will be examined in more detail in Section 2.6. The causal nature of influence is also recognised in empirical and theoretical studies of online interactions, in fields including computer science (e.g. Aral et al., 2009; Eckles and Bakshy, 2017), communications (e.g. Watts, 2007), and social psychology (e.g. Mason et al., 2007), and including in studies that do not actually use causal methodology in their empirical analyses and acknowledge the limitations of their (associational, non-causal) claims on influence due to this (e.g. Bakshy et al., 2011, from the field of computer science; Bright et al., 2017, from the political sciences).

Similarly to the above definitions, which are platform- and context-independent, and as discussed in Section 1.1, in this thesis, to say that ‘A influenced B’ is to say that A was one cause of the occurrence of B, whether intentionally or unintentionally, and requiring that no force or coercion was used in an attempt to ensure the occurrence of B. Therefore, the 'social influence' of A on B in this thesis is defined as the role A had in causing the occurrence of B, where A represents a social factor, i.e. a cause from the social world, such as a person's or a group's action(s), behaviour(s), statement(s), message(s), or other communication(s), and B represents another person's or group's action(s), decision(s), opinion(s), belief(s), message(s), or other outcome(s). Chapter 4 will describe in more detail how the term 'social influence' will be used in the analyses presented in this thesis.

It is noted here that the term 'online social influence' is used in this thesis merely as shorthand, to denote 'the social influence of online text-based communications', as mentioned in Chapter 1.1. Online text-based communications include communications such as social media posts, re-posts, comments, messages sent on online messenger applications, emails, or textual content posted on message boards or blogs. The intention is not to claim that the only kind of online social influence that exists is the social influence of online text-based communications. Rather, it is recognised that there may be other kinds of online social influence, besides the social influence of online text-based social communications, that one may be exposed to online. For example, other kinds of online social influence include the social influence of online news websites, or the social influence of online images or videos. Even though the principles and frameworks proposed in this thesis may apply to other kinds of online social influence as well, this thesis does not seek to explore this, as its main focus is the social influence of online text-based communications. Therefore, where the term 'online social influence' is used in this thesis, it is used only for brevity's sake, to denote 'the social influence of online text-based communications', and not to claim that the latter is the only kind of online social influence that may exist.

2.4 The contagion-based paradigm for social influence

This section discusses the contagion-based paradigm of social influence on the Web, including contagion-based formal models for social influence, whereby social influence is assumed to spread the adoption or endorsement of information, actions, attitudes, opinions, ideas, behaviours, products or innovations along social network ties like a contagious virus, and the related methods used for conceptualising and empirically measuring online social influence. The next section discusses the literature on the limitations of this contagion-based paradigm, both in terms of conceptualisation and in terms of empirical measurement of online social influence.

This contagion-based paradigm has been quite popular, both for online and offline studies of social influence. For example, Lyons (2011, p. 2) notes that the series of studies by Christakis and Fowler (Christakis and Fowler, 2007, 2008; Fowler and Christakis, 2008;

Cacioppo et al., 2009) on the purported contagious spread of obesity, happiness, smoking cessation, loneliness and many other states and behaviours through social influence along social ties has received considerable acclaim in the popular press and the academic community, e.g. having been covered on the front page of the New York Times, having garnered accolades from a Nobel Prize winner and from a National Academy of Sciences member, and having received several millions in research funding from national institutes following this work, while a popular book they published on their work (Christakis and Fowler, 2009) has been translated into twenty languages.

Similarly, the contagion paradigm has also been popular in the context of online social influence. The widespread use of this contagion paradigm for conceptualising and measuring social influence in online communications has been noted, along with the merits and limitations of this paradigm, in several empirical and theoretical studies (e.g. Ackland, 2013; Alshamsi et al., 2015; Bakshy et al., 2011; Shalizi and Thomas, 2011; Tufekci, 2014).[9]

As mentioned, in the contagion paradigm literature, settings of online (and offline) social interaction are represented in the form of networks (or graphs), where the nodes (or vertices) typically represent people, i.e. Web users in the online communications setting under study, and edges may represent social ties, or interactions. That is, as discussed in Chen et al. (2013), these networks may represent social affiliation networks, or interaction networks, depending on what data is available to the investigators, on the methodology uses and on the research question studied.

Edges in social affiliation networks may represent social affiliation among the people studied, commonly followership relationships (an edge from node A to B means that A follows B), or friendship relationships (e.g. in platforms with mutual friendships, such as Facebook, friendship edges will always be bidirectional), depending on the kinds of social ties the given social network platform allows (e.g. friendships of Facebook, followership on Twitter). In interaction networks, two nodes (people) are connected by an edge if one reacted to another's online post (e.g. social media post, email, blog post) in some way, depending on what is allowed on each platform, such as re-shares (in a reshare network, an edge from A to B would mean B re-shares A's post), or 'likes' or comments. An edge from A to B may also represent that B posted about the same information as A (e.g posted the same URL or hashtag), denoting the assumed flow of information from one node to the other, across people in an assumed underlying social network. Edges can have an associated weight, to denote how many times B responded or reacted to A's posts over the time period studied.

[9] Indeed, several works in the contagion paradigm have received much attention in academia in terms of citations. Theoretical models such as those proposed in Domingos and Richardson (2001), Easley and Kleinberg (2010, Chapter 19) and Kempe et al. (2003) have each received thousands of citations, and several empirical studies of social influence online which apply such concepts and models of contagion have received thousands (e.g. Cha et al., 2010; Kitsak et al., 2010; Kwak et al., 2010) or hundreds of citations (e.g. Ghosh and Lerman, 2010b; González-Bailón et al., 2011; Goyal et al., 2010; Kempe et al., 2005). (Citation numbers were retrieved from Google Scholar, on the 4th of April 2018).

2.4.1 The theory of influentials

The theory of influentials, or 'the influentials hypothesis' (Watts and Dodds, 2007), states that a small proportion of the population, which is 'informed, respected and well-connected' (Cha et al., 2010; Katz and Lazarsfeld, 1955), disproportionately influences the behaviours and opinions of a large section of the population. For instance, Keller and Berry (2003) claim that 'one in ten Americans tells the other nine how to vote, where to eat, and what to buy'. These extremely influential individuals have been called *opinion leaders, influentials, innovators* in the diffusion of innovations theory, and *hubs, connectors*, or *mavens*, and this theory has become well known in the world of marketing (Cha et al., 2010; Watts, 2007).

As presented in Watts and Dodds (2007), this theory has its roots in the 1950s, when a new theory of public opinion formation was presented, called the 'two-step flow' theory. This theory claimed that information and social influence does not flow directly, in a single step, from media to individuals; rather, it first flows through a small minority of 'opinion leaders' and then flows to these people's followers. In the following decades, this theory, and the idea of 'opinion leaders', later also called 'influentials', rose to prominence in the literatures of the diffusion of innovations, communication theory, and marketing, becoming a contagion-based paradigm by the late 1970s, and leading to thousands more research studies by the end of the twentieth century. It is this paradigm for social influence that has been the typical one encountered in computer science studies of social influence in settings of online communication.

The theory of influentials has been widely used, but it has also been criticised, with more modern theories of information flow and peer influence challenging this traditional view of influence. In some empirical studies (e.g. Cha et al., 2010; Ghosh and Lerman, 2010b), it has been claimed that some evidence has been found to support the influentials hypothesis – it has been found that response ranks follow a *power law*: a small minority individuals get disproportionate levels of response and engagement, which are orders of magnitudes larger than for the majority of individuals (and influence is defined in terms of response levels in those studies). Nevertheless, predicting who will be the people whose posts will receive large volumes of response has proven elusive in other empirical studies, such as Bakshy et al. (2011).

2.4.2 Models of contagion, influence maximisation, and network centrality

A big section of the literature on influence-as-contagion is concerned with theoretical network models of contagion. In these models, it is generally assumed either that an action of re-posting of some content must be due to the person who originally posted that content, or due to the content itself. For example, resharing a user's post containing some information on social media is taken to mean being influenced by that user to reshare that information; forwarding an email is assumed to be due to that email's content influencing the person to forward it. It is also generally assumed that the action of publishing a post about some topic, idea, or attitude is the same as the adoption or endorsement of that topic, idea, or attitude, possibly due to some person of interest who previously posted, or posted about, that topic, idea, or attitude (e.g. using a hashtag related to a protest in a Twitter post is interpreted as having joined that protest). It is generally assumed that (the adoption of) opinions, behaviours and attitudes spread in the same way as (the awareness of) information spreads.

Many of these models are threshold-based models, which have been popular for many years (Granovetter, 1978; Watts, 2002). In these, it is assumed that a person will do something (e.g. buy a product; join a group in an online community; join a protest) if a certain number (Granovetter, 1978) or fraction (Watts, 2002), representing the threshold, of the population or of their immediate social connections (neighbours) have also been influenced (Bonchi, 2011; Easley and Kleinberg, 2010, Chapter 19; Kempe et al., 2003, and the much-cited work from mathematical sociology Granovetter, 1978). These threshold models of contagious spread assume that 'individuals make decisions based on the choices of their neighbours' (Easley and Kleinberg, 2010, p. 565). This assumption is very common in contagion-based network models of influence (Mason et al., 2007). For instance, another very popular formal model, from social psychology, that also assumes that individuals are only influenced by their neighbours is the dynamic social impact model by Nowak et al. (1990). Similarly, Friedkin's widely used structural theory of social influence (Friedkin, 2006) also relies on this assumption. Such threshold models in effect assume each node makes a cost-benefit calculation when deciding whether to adopt something, which depends on their neighbours' adoptions (Watts and Dodds, 2007).

As an example of what a threshold means in threshold models, it is stated in Easley and Kleinberg (2010, p. 584) that 'A threshold of k means, 'I will show up for the protest if I am sure that at least k people in total (including myself) will show up'. The rationale behind the assumption that this is the decision process by which people adopt behaviours is stated in (Easley and Kleinberg, 2010, p. 564): 'We saw in fact that there are two distinct kinds of reasons why imitating the behavior of others can be beneficial: informational effects, based on the fact that the choices made by others can provide indirect information about what they know; and direct-benefit effects, in which there are direct payoffs from copying the decisions of others for example, payoffs that arise from using compatible technologies instead of incompatible ones'. In the payoffs used in the models presented subsequently in Easley and Kleinberg (2010, p. 566), it is assumed that for any pair of neighbours, if they do not adopt the same behaviour then each node gets a 0 payoff, while if they adopt the same behaviour they each get a positive payoff.

This assumption, that a person will adopt something if enough of their immediate social ties (neighbours) have, mirrors how a virus spreads from one person to those he/she may contact. In these models, the whole 'contagious' process starts from a single person (node) who is 'infected' or 'actiavated' first, the so-called *seed node*. So one well-known problem stemming from this kind of model is how to choose a seed node such that the maximum number of people are 'infected' (i.e. buy a product, adopt an idea, join a movement) – this is the *influence maximisation problem* (Kempe et al., 2003; Lappas et al., 2010; Kleinberg, 2007), and depends on the threshold (susceptibility to infection) of each person (node) in the network. The problem was originally posed by Domingos and Richardson, 2001). Such threshold models simulate the contagious process for several time steps, where at each time step, some nodes are infected or active, and it is caclulated which other nodes will be infected based on whether their neighbours are infected and their thresholds. Some models also involve an element of probability: each node is assigned a threshold probabilistically, e.g. uniformly drawn from [0, 1] (Kempe et al., 2003), or, in other cases, even if the adoption threshold is reached, there is still a chance that the person may not be infected (Kleinberg, 2007; Kleinberg, 2008). For example, two fundamental models that work in this manner, that are commonly used to model how information, actions, behaviours, attitudes, and other focal items spread along social network ties due to social influence, are the *Linear Threshold Model* and the *Independent Cascade Model* , where the latter is based on models of interacting particle systems (Kempe et al., 2003; Bonchi, 2011).

Similarly, models that were originally proposed for modelling the spread of actual contagious diseases have also been commonly used to model the spread of information, actions, behaviours, attitudes, and other focal items, particularly the *SIS* (Susceptible-Infected-Susceptible) and *SIR* (Susceptible-Infected-Recovered) models (Anderson et al., 1992; Castellano et al., 2009; Castellano and Pastor-Satorras, 2012; Kitsak et al., 2010; Klemm et al., 2012; Shin et al., 2016). These models are named after the possible states that the nodes of the network (or vertices, representing people) can be in, where states represent health status with respect to the pathogen under study. These kinds of models that compartmentalise members of a population according to their state with respect to the disease are known as *compartmental models* in epidemiology (one of the earliest works in this paradigm is Kermack and McKendrick, 1927). In the SIR model, initially the node that is chosen as the seed is in the 'Infectious' state, and the others are in the 'Susceptible' state. Each node in the 'Infectious' state infects each of its neighbours in the 'Susceptible' state with a given probability (infection rate) and then enters the 'Recovered' state. This is repeated until no node is in the 'Infectious' state. Shin et al., 2016 use this model, and they quantify the influence of the seed node by measuring the number of vertices infected at any time during the above process. Some types of infections, like infections from the common cold and influenza, do not give an infected person immunity to future infections, so the 'Recovered' state does not apply. SIS models capture this: for them, there is no 'Recovered' state, and after a node has been 'Infected' it becomes 'Susceptible' again at the next time step. Compared to threshold models, SIR and SIS models have been considered 'pure' contagion models, as every contact between an infected and a susceptible person is treated independently of any other. Also, mathematically, the threshold function of threshold models and the infection functions in SIR and SIS models have different properties (Watts and Dodds, 2007).

Further, the contagion-based literature for the spread of information, actions, attitudes, and behaviours through social influence, which uses threshold models often distinguishes between two contrasting types of contagion, *simple* contagion and *complex* contagion (developed in Centola and Macy, 2007; Granovetter, 1973; Watts, 2002, and further used in e.g. Centola, 2013; Romero et al., 2011; Lerman, 2016). These model different types of contagions, and have been found to apply well to different types of contagious diffusion. The difference between them is that, in simple contagion, having one 'infected' neighbour (e.g. a neighbour who has purchased a product, or adopted an idea, or shared a news article) is sufficient to get infected, whereas in complex contagion, one neighbour is not sufficient; rather, exposure to more than one infected neighbours is necessary to get infected. Complex contagion has been used to model the spread of behaviour on online social networks (González-Bailón et al., 2011), whereas it has been argued that simple contagion is sufficient for the spread of information but not behaviour (Centola and Macy, 2007), as adopting a behaviour (e.g. joining a socio-political movement) may entail more risks than simply quoting some information (e.g. a news item).

As discussed, social influence has been studied extensively using network models (as with the models of contagion above). It has often been quantified using measures of the network link structure, i.e. the network topology. Examples of topology-based measures include a definition of influence as 'the number of paths, of any length, that exist between two nodes' (Ghosh and Lerman, 2010a). In particular, so-called *centrality* measures, which describe the position of a person in relation to the social network under study, are very commonly used as measures of how influential actors in social networks are (Bakshy et al., 2011; Dubois and Gaffney, 2014; Ghosh and Lerman, 2014; Kitsak et al., 2010). Centrality measures have very widely been used to theoretically model flows along networks representing physical systems as well as information systems (Ghosh and Lerman, 2014).

Since online social influence in the literature is commonly conceptualised and measured in terms of how many online followers or friends one has, and/or in terms of how much interactions and responses one receives, such quantitative measures of one's friendship and/or interaction network have been used to measure one's social influence online. Several different measures of centrality exist, ranging from capturing local to more global properties of the underlying network (or graph). The simplest and most local one is called *degree* centrality, which is the number of edges adjacent to a given node (or vertex). This is also called the node's degree. In bidirectional networks, *in-degree* centrality is the number of incoming edges to this node, and *out-degree* is the number of outgoing edges from this node to other nodes. In the follower network, a node's in-degree is the number of their followers, and in a friendship network the number of their friends. In an interaction network, in-degree is the number of incoming edges representing the given interaction, e.g. likes, reshares, responses. (As such edges are commonly weighted, with weights numerical representing counts of occurrences of each interaction, a node's in-degree centrality is the sum of the weights of all these incoming edges.) In-degree centrality has very widely been used as a measure of a person's online social influence. Beyond in-degree, more complex measures of centrality exist, that account for the properties of the wider network's link structure.

Centrality measures are used very commonly in the field of bibiliometrics and citation analysis. The number of citations an academic paper receives is its in-degree centrality in the citation network, and has been widely used as a measure of importance (Newman, 2010; Newman, 2014). Another global measure of centrality is *closeness* centrality (Bonacich, 1972), which measures how close a node is to all other nodes in a network, hence considering nodes beyond one's immediate neighbours as well. The closeness centrality of a node is the inverse of its average distance of to all other nodes in the graph. The distance between two nodes is defined as the length of the shortest path between them, where the path length is the number of edges along this path. Furthermore, *betweenness* centrality indicates how many shortest paths pass through a node, which is often thought to be a measure of importance of nodes within a network (Freeman, 1977).

An important measure of centrality in the context of measuring the 'flow of information and influence' is *eigenvector* centrality, which accounts for the degree of indirect neighbours of a vertex too, as well as the immediate network. The basic idea is that a node should be considered important (high eigenvector centrality) according to not only how many other nodes point to it but also according to how important those other nodes are (and so on for each of the latter nodes, recursively). So, nodes ranked as important in the sense of having high eigenvector centrality are those that have either many immediate neighbours, or have neighbours that themselves have high eigenvector centrality (or both). Measures similar to eigenvector centrality are Katz centrality, the famous PageRank algorithm that Google uses to rank search results based on hyperlinks (citations) that reference them (Page et al., 1999), and Kleinberg's much-cited work on the HITS algorithm for ranking webpages using the hyperlinks that reference them (Kleinberg, 1999). In the highly cited work of Kleinberg, 1999 (more than 8,600 citations on Google Scholar), it is claimed that webpages that have a lot of hyperlinks referencing them must be authorities in their field, and hence should be ranked highly in search results about a relevant topic. The rationale given is that 'Hyperlinks encode a considerable amount of latent human judgement, and we claim that this type of judgement is precisely what is needed to formulate a notion of authority'. This logic has since been extended to the concept of social influence. In the popular paradigm of the online influence literature, instead of hyperlinks the links studied are friendship and/or interaction links, and instead of interpreting the number and properties of those links as indicators of one's authority, they are interpreted as indicators of social influence.

2.4.3 The million follower fallacy

In popular culture, in everyday parlance and in the news, the term social influence is often used overly simplistically. For instance, having a large number of subscribers on online social media is often assumed to mean having great influence (Stringhini et al., 2012). However, having many subscribers merely indicates one has the potential for their content to reach this large audience; whether this audience actually sees and engages with that content is not at all guaranteed (let alone whether it adopts the ideas, opinions, or behaviours described in that content). Being part of someone's audience does not necessarily guarantee that you have an interest in what they have to say. For example, on Twitter, a user may have fake followers (artificial accounts run by automated programs known as Twitter-bots) or paid-for followers, as there is a big underground market for buying followers to artificially increase one's apparent prestige and popularity (Ghosh et al., 2012; Sridharan et al., 2012; Stringhini et al., 2012). In addition, sometimes when a Twitter user reveives a new follower, the former proceeds to reciprocate the gesture and follow back, out of politeness, not out of interest in the content that this new follower posts (Avnit, 2009). That is, some followers in one's audience might have only followed just in reciprocation, out of politeness, not interest. Moreover, even if one is a legitimate follower, they may still never see certain tweets, for example because they rarely log into Twitter, or because they missed that specific post among the other posts in their timeline.

Therefore, one's follower count does not guarantee a real, interested and responsive audience of that size, or imply any influence. So a user's audience size is not a sufficient or particularly robust measure of their influence. And these practices of buying fake followers to increase one's perceived popularity and influence status feed back into a vicious circle, making follower counts even more unreliable and meaningless in terms of influence.

Indeed, in the scientific literature, this simplistic linking of influence to number of followers has been empirically shown to be flawed (Cha et al., 2010), since influence is a much more complex concept than merely having the potential to reach a large audience. An indication of actual engagement of the audience, in response to the content in question, is also necessary (but generally not sufficient) in order to be able to begin to make more robust claims regarding influence. In Cha et al. (2010) it was found that, on Twitter, having many subscribers (a large audience) is not strongly correlated with actual audience engagement and responsiveness (in the form of retweets and mentions), where engagement is considered by the authors a better indicator of a user's influence than follower counts. Stringhini et al. (2012) also use the term 'influence' in a similar way on

Twitter, defining it as a measure of engagement with a user account, and including in it the number of followers an account has, but also the number of retweets and mentions it has received.

In the private sector, companies dedicated to measuring user influence on social media, like Klout[10] and PeerIndex[11] also state that they do not only use follower count but also other measures that indicate audience responsiveness and engagement on online social networks, in order to come up with a more holistic 'influence score'. (As these are private, for-profit companies, details of how their influence measures are calculated are not publicly available.)

2.4.4 Social influence as response

In addition to interpreting the number of followers and friends (and structural properties of the follower or friend network) as indicators of social influence, and given the problems of that approach, interpreting response to a post, or attention to a topic, as social influence is another assumption that is widely used in theoretical and observational studies of online communications (Anagnostopoulos et al., 2008; Bakshy et al., 2011; Barbieri et al., 2013; González-Bailón et al., 2011; Goyal et al., 2010; Kwak et al., 2010; Leskovec et al., 2006).

So, an online message or post is considered influential if it receives a lot of responses. By extension, a Web user is considered influential if their online posts receive a lot of responses, or if (on social media) their username is mentioned a lot (e.g. with Twitter's @mention functionality). Similarly, a topic or keyword (e.g. represented by a hashtag on Twitter) or a URL is considered more influential the more people mention it in their social media posts. In more detail, if user j's social connection i mentions the same entity as them (e.g. a URL or a hashtag), within a narrow time window, or if i re-shares or up-votes j's post, or chooses to follow j, or mentions j's username (Cha et al., 2010; Ghosh and Lerman, 2010b), then i's action is assumed to be due to social influence from j. One may say that such measures of online activity represent the levels of attention or interest that a given piece of content has generated (Ackland, 2013; Watts, 2007).

For instance, Dubois and Gaffney (2014) compare six metrics commonly used to measure the social influence of communications on social media, four of which are based on the topology of how a user is embedded in the Twitter social network (i.e. follower network), and the other two relate to Twitter actions or responses involving username mentions and hashtag usage: one being based on how much a user's Twitter handle is mentioned by others, another being based on how many topic-related keywords a user includes in their tweets (which is assumed to be an indication of the user's 'knowledge', or 'expertise').

[10] `https://klout.com/corp/score`
[11] `https://www.brandwatch.com/p/peerindex-and-brandwatch`

Moreover, in Bakshy et al. (2011), a study about 'quantifying influence on Twitter' based on the usage of URLs, influence is defined as follows: 'if person B is following person A, and person A posted the URL before B and was the only of B's friends to post the URL, we say person A influenced person B to post the URL'.[12] Barbieri et al. (2013) go as far as to take an approach whereby, if a group of users is observed acting on an item i within a short period of time, and if this occurs on several different items, then the authors assume that 'then we can infer that these users are connected in some social network, that they communicate and can influence each other'.

2.5 Known limitations of the contagion-based paradigm

Having discussed how the social influence of Web communications has been studied, at the individual and the collective level, how the term 'social influence' can be defined, and how the contagion-based paradigm has conceptualised and measured the social influence of online communciations, this section now turns to some important known limitations of the contagion paradigm.

In several theoretical and empirical studies, it has been found that information awareness and actions of response do not 'spread' due to social influence as per the contagion-based paradigm. For example, in empirical studies, Bakshy et al. (2011) studied the propagation of URL usage on Twitter, and found that it is not sufficient for the originator of a message to be an 'influential' (having many Twitter followers, i.e. high in-degree in the follower network) for a URL's usage to spread widely, and no reliable predictions could be made regarding which URLs will spread widely on online social networks. Cha et al. (2010) came to a similar conclusion, finding that having many followers on Twitter (hence being considered an 'influential' by traditional standards) is neither necessary nor sufficient in order for one's posts to receive many responses in the form of retweets and mentions (with influence being defined by the authors in terms of response volume).

Other known limitations of the contagion-based paradigm include limitations that arise when considering causes of observed actions other than social influence, the issue of the so-called *post-hoc ergo propter-hoc* fallacy, and conceptual and methodological criticisms. These will be discussed next.

2.5.1 Causes other than the social influence of online communications

Even if one accepts the conceptualisation of social influence in terms of contagion, and hence the definition of one being influential as a person's online posts receiving high

[12] As we shall see, the authors themselves recognise this definition has a significant disadvantage: the two postings of the URL may be coincidental, and it may a consequence of homophily (personal similarity between the two people) rather than influence. Therefore, they state that their 'estimates of influence should be viewed as an upper bound'.

levels of response or engagement, it is still difficult to tell whether these responses are due to social influence from that person's online post(s). This is because there are other factors that could be the cause of the response, or even factors that could be the common cause behind both the response and the original prompt (e.g. another social media post). Therefore, one must be careful not to conflate responses that are caused by these other factors with manifestations of the social influence of the particular user, or the particular post, being responded to. Other factors are not part of the models of the contagion-based paradigm for the spread of information, actions, and behaviours due the social influence of online communications. This is an important limitation of the contagion paradigm, as it has been found that failing to account for such factors (factors other than online social influence from a given person, group, or set of online communications) means that the contagion paradigm does not accurately model empirical settings of online interactions.

For example, in Lerman (2016) it is noted how the standard contagion paradigm failed in accurately predicting spreading patterns in several empirical studies, as spreading dynamics observed empirically were much smaller than what is predicted by the popular Independent Cascade Model of the contagion paradigm (Bakshy et al., 2011; Goel et al., 2012; Ver Steeg et al., 2011). It is then discussed how cognitive factors that may limit a user's attention span when looking at their social media feed, including factors such as the position of a post in one's social feed (how prominently displayed or hidden it might be in a user's social media feed), that are not part of contagion models, can help explain why the standard contagion paradigm often fails at accurately predicting spreading patterns.

Homophily is another such factor that has been studied in several works on online social influence. Homophily, or assortative mixing, or social selection, are terms used to describe the observation that people tend to associate with others whom they perceive as similar to themselves in some way (McPherson et al., 2001; Newman, 2010). For example, this similarity may be in terms of demographic characteristics, like age, race, gender, social class, in terms of personality traits, like extroversion or shyness, or in terms of interests and beliefs. When the focus of investigation is social tie formation and similarity, homophily and influence are often taken to capture causally reverse social processes, where becoming friends with someone because of pre-existing similarity is considered homophily, whereas becoming similar to someone (e.g. adopting their behaviours or ideas) because of pre-existing friendship is considered social influence (Aral et al., 2009; Crossley and Ibrahim, 2012). This means that, when studying online social networks, homophily leads to ties being more likely among similar people, and so neighbours' (friends') outcomes could be correlated because inherent personal similarities rather than because one influenced the other (Aral et al., 2009). Therefore, in studies of social influence, it is important to be aware of this factor, as it has been found that failing to account for homophily may lead to mistaking cases of homophily (pre-existing similarity among social connections) for influence, hence overestimating influence, possibly gravely so (Anagnostopoulos et al., 2008; Aral et al., 2009; Bakshy et al., 2011; Shalizi and Thomas, 2011). For instance, in the empirical results of Aral et al. (2009), it was found that if one ignores homophily and rather interprets all outcomes (product adoption) using the standard contagion paradigm, one would overestimate social influence by 300-700%.

An additional factor that is often not included in contagion-based models of social influence, and particularly in the 'influentials hypothesis' (discussed in Section 2.4.1), which has been studied elsewhere in the literature, is the role of susceptible, or easily influenced, individuals (Aral and Walker, 2012; Watts and Dodds, 2007; Watts, 2007). In Watts and Dodds, 2007 and Watts, 2007, through the use of simulations and threshold models, it is found that whether a trend spreads or not depends on whether there are enough people that are generally open, or susceptible, to it, and this is regardless of who started it (whether the person who started it was an 'influential' or not). That is, whether a message spreads does not depend on whether an 'influential' started it, but rather on whether the circumstances in the wider environment are such that benefit the adoption of this sort of message. They argue that this is especially dependent on the link structure of the social network. In Aral and Walker (2012), the role of 'influentials' is tested empirically against the role of susceptible individuals in a large-scale randomized experiment on Facebook, in the context of the spread of adoption of a commercial Facebook application. It is found that an individual's importance in the spreading of application downloads is determined not only by his/her attributes and personal network (i.e. whether they are an 'influential')' but is jointly determined by the presence of influential and susceptible individuals and the local properties of their social network.

2.5.2 Conceptual and methodological criticisms

In addition to the specific issues raised above, there exist in the literature additional criticisms at the level of logical reasoning, concepts, and methodology used in the contagion-based paradigm for social influence online.

A related problem to the lack of consideration of other causes in contagion-based studies of online social influence is the *post hoc ergo propter hoc* ('after this, therefore because of this') logical fallacy. It asserts that if event B happens after event A, it must be that event B happened because of event A (Watts, 2007). The problem with this fallacy is noted in Watts (2007), in the context of using anecdotal evidence to support the influentials hypothesis, particularly in the media, and with reference to Malcolm Gladwell's best-selling popular science book 'The Tipping Point' (Gladwell, 2002). Watts (2007) notes how simplistic narrative descriptions of the form 'this happened, then that happened' are often used when discussing anecdotal evidence, in an effort to give a snappy and simple explanation of an observed phenomenon, which means overlooking the possibility of multiple, and possibly indirect, causes, in favour of a simple, linear narrative of causation. As shall be described in Chapter 3 of this thesis, this fallacy is not only common in discussions of anecdotes informally and in the media. It is also common in scientific studies using contagion-based paradigm for social influence, which typically assume that observed outcomes occurred due to the social influence of specific online communications, without taking other possible causes into account.

Tufekci (2014), Freelon (2014), and Marres (2017) each present important limitations and problems related to the contagion-based paradigm for studying social influence in online interaction settings (social media in particular), and are worth discussing individually.

In Tufekci (2014), a range of crucial limitations in analyses of social media Big Data are discussed. One of the most important criticisms that is relevant to social media studies of social influence under the contagion paradigm is interpretational. It concerns how the meanings of outcomes (e.g. retweeting) are assumed to be endorsement, agreement, and 'influence', when retweets can in fact have a range of meanings, from positive to neutral to negative, expressing emotions as negative as anger, denunciation, or disgust. As case studies two cases are discussed (on page 6 of Tufekci, 2014), in one case describing an instance of thousands of backlash retweets and mentions against a user's insensitive tweet, and in another a case where a political figure's tweets were widely retweeted with a negative or mocking intent. This discussion concludes with the statement that '"influence" and "popularity" may not be the best term to use [when measuring online outcomes such as the above]. Some portion of retweets and follows are, in fact, negative or mocking, and do not represent "influence" in the way it is ordinarily understood. The scale of such behaviour remains an important, unanswered question' (Tufekci, 2014, p. 6).

An additional and very important methodological criticism of the contagion paradigm in Tufekci (2014) is two-fold: it is noted that importing network methods from other fields to study human behaviour often happens without evaluating their appropriateness, and that solely focusing on 'node-to-node' interactions on networks and ignoring 'field effects' (events that affect whole groups of people, e.g. through shared experience or broadcast media) is problematic as it is the latter that may often account for observed phenomena. It is discussed how the answer to the question of whether information spreads in the way germs do is implicitly assumed to be affirmative, however this is problematic. That is because not only does information not diffuse among humans solely on a single social media platform (humans receive information from a wide rage of sources), but crucially the spread of epidemics and contagious diseases is well understood, has been extensively empirically verified, with its mechanism well defined: small microbes travel in physical space to infect people that are physically close by entering their body. However the spread of information, behaviours, actions, beliefs and attitudes through social influence does not operate under such a mechanism. Further, the physical process of disease spread has well-understood properties, and the underlying probabilities of infection can often be calculated with precision. One key issue here is that the conditions and assumptions under which information, behaviours, actions, beliefs or attitudes spread among online social network 'neighbours' (users that are friends with or follow each other on the social platform) may be analogous to physical proximity and disease spread, but these assumptions and conditions are 'rarely subjected to critical examination'. An important point raised is that 'whether there is a straightforward relation between information exposure and the rate of "influence", as there often is for exposure to a disease agent and the rate of infection, is something that should be empirically investigated, not assumed' Tufekci, 2014, p. 8).

Taking a further step back, (Tufekci, 2014, p. 8) also highlights the representation of social media interactions as networks as a choice that 'requires a whole host of implicit and important assumptions that should be considered explicitly rather than be assumed away'. (For example, modelling everything as a node-to-node network leads to ignoring field effects, i.e. other factors that affect groups of people simultaneously, not spreading in a node-to-node manner, and may themselves be causal factors behind observed behaviours.) It is also stressed that importation of network methods from other fields 'needs to rely on more than some putative universal, context-independent property of networked interaction simply by virtue of the existence of a network' (Tufekci, 2014, p. 8).

The issue of interpreting the meaning of online behaviours is also raised in Freelon (2014), a paper that is also cited in Tufekci (2014). This paper studies some of the top cited papers in communications and in social computing, published in recognised venues in each field. It examines how hyperlinks on the Web, and followers and retweets on Twitter are commonly interpreted in the communications and social computing literatures, and highlights how interpretative strategies of the meaning of digital human interaction traces are 'scattered and ad hoc', with many studies not justifying or explaining the validity of their interpretations, and instead interpretations are often 'stated as plainly as any self-evident fact'. Particularly the case of retweets is most relevant in this thesis, as those are commonly studied actions in online communications settings, and are often assumed to indicate 'influence'. (Freelon, 2014, p. 67) states that '"Retweets are not endorsements" is a disclaimer commonly seen in Twitter biographies', but most of the social computing papers studied offer no justification for their interpretations. A noted exception is Bakshy et al. (2011), which discusses in detail the kinds and limitations of inferences they make about social influence from retweet data. Indeed Freelon (2014, p. 68,69) notes how influence is an 'expansive concept' and an 'abstract, elusive concept', highlighting the importance of thoughtful consideration of what digital traces of social media behaviour, like retweets, might mean in terms of this concept. Overall, the key fining in Freelon (2014, p. 69) is that 'substantial proportions' of the articles studied, from both disciplines, 'failed to justify the social implications they imputed to trace data', which occurrence more extensively in social computing articles. It is then advocated that ‘while claims that traces represent influence, trust, credibility, etc. may sound intuitive, they need convincing support’ (Freelon, 2014, p. 69).

Similarly to Tufekci’s point (in Tufekci, 2014) on the problematic implicit assumption underlying contagion-based methods, that all network interaction has ‘some putative universal, context-independent property’, this kind of assumption is also criticised in Marres (2017). In this book, Marres (2017, p. 16, 1-22) specifically criticises the universalist undertones of certain claims made about what can be inferred from digital traces of social interactions, claims which have appeared both in reputable media sources like the New York Times, and in scientific publications. Marres discuses such claims, starting with a New York Times article (Markioff, 2011), promising the realisation of the aims of ‘universal social science’, to uncover for the first time laws of human behaviours and enable predictions of social and economic instability just as natural scientists can predict natural phenomena. Marres notes that this ‘grand narrative about a “new science” that will uncover the eternally valid laws of society [...] is called sociological positivism and [...] tends to be regarded among serious sociologists as too naive and/or hubristic to entertain’. She then discusses similar claims made in scientific publications, for example in Lazer et al. (2009), where a vision is outlined for the new field of computational social science, and it is claimed that this field would ‘compile [digital traces] into comprehensive pictures of both individual and group behaviour’, as well as in other articles and books, such as the claim in Pentland (2014) that digital data would ‘give us the chance to view society in all of its complexity, through the millions of networks of person-to-person exchanges’. As noted also in Tufekci (2014), social media platforms represent only one of the many sources that expose people to information, attitudes, beliefs, and behaviours, so such claims that data from online social platforms will give investigators a comprehensive picture of how information, actions, behaviours, attitudes or opinions are adopted through social influence, or of how any other similar social process operate, are problematic. Marres argues (Marres, 2017, p. 22) that such statements ‘re-activate a universalist vision of social science, one which much sociological training and literature precisely teaches us to qualify’.

In addition, Marres (2017, p. 20) also raises the issue of the empiricist idea of data-driven research, fashionable within and outside academia in the wake of Big Data and the digital data deluge, and how several scholars have criticised it. It is noted how analyses advertised as ‘data-driven’ actually are highly dependent on theory, as they rely on specific models, and often depend on a limited set of data formats, and come with a set of conceptual assumptions. As presented above in the discussion of Tufekci (2014), this is very much the case with the contagion-based paradigm, which makes many strong assumptions about what factors do and do not matter in influencing people (only immediate friends matter), how influence works (like a disease), without questioning these assumptions, and without appropriately justifying or empirically proving that these assumptions are applicable in online social interaction settings.

Moreover, Marres (2017, p. 18) talks about the argument made in the sociological literature that 'the digital data deluge makes possible a shift from theory-driven causal "explanation" to a more empirical style of "description" as the dominant mode of sociological analysis, as digital data analysis enables fine-grained description of social life'. Marres does not go into the extent to which this may have happened. As this thesis shall present in Chapter 3, this tendency for the focus to be on descriptive analyses, rather than on causal analyses, is exactly what has been happening in the contagion-based paradigm; the problem, however, is that causal claims about social influence (who is influencing whom, who is an 'influential') are made, based on inappropriate methods, that is, descriptive, not causal, methods.

Finally, similarly to the point in Tufekci (2014) about how using a network model cannot be justified simply by virtue of the fact that data from online social network platforms is already in a network form, Marres (2017, p. 112), in the context of the claim that '"the digital" drives methodological innovation in social research', which she finds problematic, she urges researchers 'to avoid adopting methods and techniques *simply because they are easy to use*'. This tendency is also noted in Alshamsi et al. (2015, p. 3), where the term 'network measurability bias' is employed to describe 'the tendency to focus on processes that are easily observable within digital social networks (such as "likes" and "re-tweets"), while neglecting key latent processes such as the ideological, cultural, and economic incentives of actors'. In the critique of Chapter 4, this thesis carries this point further, and discusses also how the common tendency of using as outcomes the actions that are readily recorded and easily measured in common settings of online communications (email, social media) is not always appropriate for making meaningful claims on social influence.

2.6 A causal paradigm for online social influence

Given the limitations of the contagion-based paradigm for online social influence, and having discussed how the term 'social influence' can be defined, this section turns to an alternative, causal, paradigm, for conceptualising and measuring online social influence. As this thesis will propose in subsequent chapters, this causal paradigm is an important component of the effort to address the limitations of the contagion-based paradigm, as it is useful for conceptualising social influence in a manner aligned with how the term has been traditionally understood in the social sciences, and for measuring and disentangling the social influence of online communications from the influence of other causes on the outcome of interest.

This section begins by introducing causal methods, and then proceeds to discuss the causal nature of the concept of influence. It next presents graphical causal models, a key tool in causal modelling and inference. Then, methods are discussed for estimating numerical values for causal effects, and finally evaluation methods are presented for assessing the fit of a causal model to a given empirical dataset.

2.6.1 Introduction

It is an often-repeated cautionary phrase in statistics that 'correlation does not imply causation', and that 'association does not imply causation'. The field of causality theory, which saw rapid developments in the last thirty years, allows one to go beyond correlations and associations and to reason about causation in a rigorous, formal way, using tools like graphical causal models, which in turn are based on directed graphs and probability theory (Pearl, 2009b). This thesis (particularly in Chapters 4 - 6) will be using graphical causal models to reason about the social influence of online communications versus the effects of other possible causes of observed outcomes (expanding upon the work presented in Shalizi and Thomas, 2011). The relevant theory is presented here, based on Pearl (2009a,b), Pearl (2010), Pearl et al. (2016), Morgan and Winship (2014), and Shalizi (2013).

One common and powerful way for demonstrating why traditional statistics must be enriched with new ingredients in order to deal with cause and effect relationships is through Simpson's paradox. This is a well known statistical paradox, discovered more than a century ago, that has puzzled statisticians for many decades (Pearl (2009a, p. 118-9). This paradox is said to occur in datasets where a statistical association that holds for an entire population is reversed for each subpopulation. For instance, consider a dataset based on which taking a drug is associated with better recovery rates than not taking it, in a patient population, but if filtering the patients based on gender, female drug-takers are found to have worse recovery rates than female non-takers, and male drug-takers are found to have worse recovery rates than male non-takers (Pearl et al., 2016, Example 1.2.1). Then the question arises of which finding should be used - should one go with the population-level result, or with the subpopulation-level result? In the example, should a doctor administer the drug to a female patient, a male patient, a patient of unknown gender? Causal methods, particularly graphical causal models, allow one to deal with this impasse, and decide whether the population or subpopulation level result should be chosen, based on the causal assumptions and relations encoded in the graphical causal model. (Specifically, they offer formal causal rules for reasoning about which variables should and should not be used for filtering a population into subpopulations.) That is, the data alone cannot guide this decision; rather, one needs to first understand the causal mechanism that led to, or *generated*, the observed data. Statistical methods cannot determine the causal mechanism from the data alone, so no statistical methods can help with the above decision. (Simpson's paradox and the impossibility of using statistical methods and the data alone to resolve it is discussed and illustrated with examples in Pearl et al. (2016, p. 1-6), and it is also explained in great depth in Pearl (2009b, Chapter 6).)

2.6.2 Influence is a causal concept: causal versus statistical notions

The causal literature makes an important distinction between notions of association versus of causation. Traditional statistical methods allow one to make associational claims, but in order to make causal claims one needs to use the additional language, semantics and mathematical apparatus of causal methods (Pearl, 2009a; Shalizi, 2013). As an illustration of the insufficiency of standard statistics and probability for reasoning about causal relations, Pearl (2009a, Section 2.3) notes that the syntax of probability calculus does not let one express the fact that ‘symptoms do not cause diseases’, let alone draw mathematical conclusions from such facts. As stated in Pearl (2010), ‘Traditional statistics is strong in devising ways of describing data and inferring distributional parameters from samples. Causal inference requires two additional ingredients: a science-friendly language for articulating causal knowledge and a mathematical machinery for processing that knowledge, combining it with data and drawing new causal conclusions about a phenomenon’.

Pearl (2009a, Section 2.1) explains that the basic distinction between statistical and causal methods relates to coping with change. The aim of statistical analysis is to calculate parameters of distributions, based on samples of those distributions, based on which to then ‘infer associations among variables, estimate beliefs or probabilities of past and future events, as well as update those probabilities in light of new evidence or new measurements’. If experimental conditions remain the same, standard statistical analysis is adequate for those tasks. However, causal analysis goes one step further, as it also aims to infer ‘the dynamics of beliefs or probabilities *under changing conditions*, for example, changes induced by treatments or external interventions’. A distribution function and the laws of probability cannot tell us how that distribution would change if external conditions changed. Rather, one needs causal assumptions to obtain this information, which ‘identify relationships that remain invariant when external conditions change’.

Therefore, in order to help clearly formulate a distinction between associational (statistical) versus causal concepts, Pearl (2009a, Section 2.2) uses as a demarcation line the question of whether a concept can be defined solely in terms of a joint distribution of observed variables. An associational concept is any relationship that can be defined thus, whereas a causal concept is any relationship that cannot be defined from the distribution alone. It is worth quoting here the examples of associational and of causal concepts that Pearl lists next. ‘Examples of associational concepts are: correlation, regression, dependence, conditional independence, likelihood, collapsibility, propensity score, risk ratio, odds ratio, marginalization, conditionalization, “controlling for”, and so on. Examples of causal concepts are: randomization, influence, effect, confounding, "holding constant", disturbance, spurious correlation, faithfulness/stability, instrumental variables, intervention, explanation, attribution, and so on. The former can, while the latter cannot be defined in terms of distribution functions'. It is then highlighted how this demarcation line helps investigators trace the assumptions needed for substantiating various types of scientific claims, and how every claim that invokes causal concepts 'must rely on some premises that invoke such concepts; it cannot be inferred from, or even defined in terms of statistical associations alone'.

As can be seen above, 'influence' is listed as a causal concept, as well as 'effect', 'explanation', and 'attribution'. And indeed, throughout Pearl (2009a), the verb 'influence' is often used interchangeably with the verb 'cause', and the term 'causal effect' is used interchangeably with the term 'causal influence'. The terms 'causation' and 'influence' are also used interchangeably in Shalizi (2013, e.g. on p. 483). In Pearl (2009a, p. 107) it is stated that 'effect' is defined as a general capacity to transmit changes among variables (based on simulating hypothetical interventions in the causal model, when actual interventions through experiments are not possible and one is working with observational data). As discussed in Chapter 2.3 of this thesis, Thomas (2013) also notes that influence has historically been understood as a causal concept, and several studies of influence on social media recognise the causal nature of influence, in fields such as computer science (e.g. Aral et al., 2009; Bakshy et al., 2011; Eckles and Bakshy, 2017), communications (e.g. Watts, 2007), and social psychology (e.g. Mason et al., 2007). All this is in line with the dictionary definitions and the social scientific definitions of influence and social influence discussed in Chapter 2.3. These definitions inform how the terms 'influence' and 'social influence' will be defined and used in this thesis, as shall be presented in Chapter 4.1.

It is noted that contagion-based models make claims on 'influence' by explicitly assuming that observed actions (actions such as re-sharing of a post on social media, that are assumed to mean adoption, e.g. 'activation' or 'infection' of a node in threshold models) can be 'attributed' to 'influence' from one's social network neighbours, or from the person whose post is being acted upon (e.g. re-shared) (e.g. Cha et al., 2010; Kleinberg, 1999; Kempe et al., 2003; González-Bailón et al., 2011; Kitsak et al., 2010; Ghosh and Lerman, 2010b; Kwak et al., 2010). However, as mentioned in Section 2.5, and as shall be further discussed in Chapter 3, whether those assumptions are appropriate and realistic is typically not examined in the contagion model (and there is scientific evidence to suggest those assumptions do not hold, e.g. as noted in Lerman, 2016). Despite using, and making claims on, causal concepts, contagion-based studies by and large do not use any causal methodology. It is also problematic that they do not use causal methodology to reason about their assumptions, or to empirically test the validity of their assumptions (as can be done in the causal paradigm, with the methods described in Section 2.6.5).

Based on this reasoning, as well as on the definitions from Chapter 2.3, Chapter 4 will describe how the term 'influence' and 'social influence' will be used in the analyses presented in this thesis.

2.6.3 Graphical causal models

This section describes graphical causal models, also known as causal networks, and also discusses the other main tools (structural equation models, the potential outcomes framework, propensity score methods and matching methods) available currently for causal analysis. It begins by introducing graphical causal models, then connects those to structural equation models (a different but equivalent way of representing causal relationships). It returns to graphical causal models to discuss how they can be used to address confounding bias in causal analyses, a topic that is very important in disentangling the social influence of online communications from the effects of other (social and non-social) causes of observed outcomes. It then outlines an alternative framework for causal analysis, the potential outcomes framework, which has been proven to be logically equivalent to the structural equations and graphical approaches for causal analysis (Pearl, 2009b; Pearl et al., 2016).

2.6.3.1 Introduction

As mentioned, one important tool used in causality theory is graphical causal models (also called causal networks, causal diagrams, or causal graphs). A graphical causal model can be represented as a ***directed acyclic graph (DAG)*** G, comprised of a set set of nodes, N, and a set of directed edges, or arrows, E – that is, $G = \{N, E\}$. Nodes represent variables, and edges denote causal relationships. The fact that this directed graph is acyclic means that edges (arrows) cannot form cycles, that is, starting at any node, it is not possible to follow a sequence of arrows emanating from it, along the direction in which the arrows are pointing, and end up at the same node. This means that an outcome of a variable cannot also be a cause of that variable.

A pair of nodes A and B have an arrow from A to B if A is a *possible* cause of B. An absence of an arrow between A and B represents the assumption that neither can be a direct cause of the other. That is, an arrow merely indicates the *possibility* of a causal relation, the strength of which will be determined from the data (the strength may be found based on the data to be zero), while ‘a missing arrow represents a claim of zero influence’ (Pearl (2009a, p. 105). That is, the causal relationships depicted in causal diagrams represent probabilities, not absolute certainties. For example, as we shall see more formally, the numerical estimate of the influence (causal effect) of one variable A on an outcome variable B, when both A and B are binary variables, represents the difference in the probability of B's occurrence when A is made to occur versus when A is made to not occur.

When there is an arrow from A to B, A is called a ***parent*** or ***ancestor*** of B, and B a ***child*** or ***descendant*** of A. If a node has no arrow pointing to it, i.e. no parents, it is called ***exogenous***, otherwise it is called ***endogenous***. A ***path*** is a sequence of consecutive edges that do not all necessarily point in the same direction.

Which nodes are connected to which depends on the modeller's causal assumptions, which should be well-justified and grounded in domain expertise (Pearl, 2009b). A graphical causal model represents the assumed data generating process (the assumed process according to which nature generates the observed data): the causal *mechanisms* that are assumed to determine the value of each variable. The rules for manipulating graphical causal models then show what causal inferences can be made from these causal assumptions.

For any three nodes A, B, C linked in a causal DAG, the names commonly used for them or for the possible paths between them are given in Definition 2.6.1.

Definition 2.6.1. Types of paths The names commonly used for paths involving nodes A, B, C are the following:

Chain: A path $A \rightarrow B \rightarrow C$, or $A \leftarrow B \leftarrow C$;

Fork: In a path $A \leftarrow B \rightarrow C$, where node B is the common parent of both A and C., node B is called a ***fork*** since two arrows emanate from it;

Collider: In a path $A \rightarrow B \leftarrow C$, where node B is the common child of nodes A and C, node B is called a ***collider***, as two arrowheads 'collide' on it.

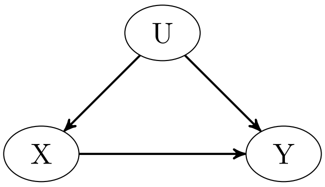


Figure 2.1: Example causal diagram

Every causal graph encodes the joint probability distribution of the variables (nodes) in it. Given that the ***probability distribution*** for a variable X is the set of probabilities assigned to each possible value of X, the ***joint probability distribution*** of a set of variables is the probability distribution of the set of variables that are depicted as nodes in the causal graph. That is, the joint probability distribution represents the probability of every possible event as defined by the values of all the variables involved (Pearl, 2001, p. 3). Causal graphs are governed by the ***Causal Markov Condition***, whereby endogenous variables only depend on their parents (Pearl, 2009b). For example, using $P(w)$ as shorthand for $P(W = w)$, which denotes the probability that a variable W takes value w, for any variable (node) W, the joint probability distribution representing Figure 2.1

is: $P(y, x, u) = P(u)P(x|u)P(y|x, u)$. That is because Y's parents are U and X; U is the parent of X; and U has no parents. This causal diagram denotes that U is a common cause of X and Y, and X is also a cause of Y.

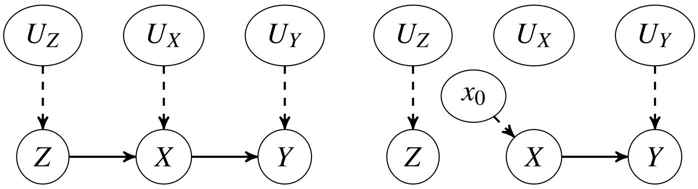
Causal effects represent ***interventions***, or ***surgeries***, or ***manipulations***, onto the arrows of the causal DAG, which changes the joint distribution the DAG represents, resulting in a new, altered DAG which represents the new ***post-intervention joint distribution***. For instance, in a DAG containing the path $A \rightarrow B$, to estimate the causal effect of a variable A on the variable B, this causal effect is written as $P(B = b|do(A = a))$ in Pearl's ***do-notation***. This quantity denotes the probability (or frequency) that event $(B = b)$ would occur, if, hypothetically, B were set to the particular value b through experimental manipulation or intervention. This would mean performing a surgery on the causal DAG, which would involve: deleting all arrows into A, setting A's value to a, and leaving the rest of the DAG unchanged. This post-intervention distribution of $\mathbf{B}$ is not in general the same as the ordinary conditional distribution $P(B = b|A = A)$, as the latter represents taking the original, pre-intervention, population and ***selecting*** from it only the sub-population where $A = a$ (i.e. ***filtering*** the pre-intervention population based on A and only keeping cases where $A = a$).[13]

2.6.3.2 Structural equation models

The causal relationships encoded in causal DAGs can also be expressed using Structural Equation Modelling (SEM). Structural equation models are particularly common for effect analysis in econometrics, behavioural science and in some other areas of social science, where the bulk of SEM methodology is for linear analysis. However, SEM can be extended to nonparametric form (models where the functional form of the equations is unknown).

In structural equation models, it is common to include in the causal model (both in the causal diagram and in the SEM) any unobserved exogenous variables (or *background variables*) that influence the values of the variables of interest, where the modeller decides to keep otherwise unexplained. Conventionally an unobserved exogenous variable that determines the value of variable X is named U_X, and there is a dashed arrow (to denote that U_X is unobserved) from U_X to X. These variables are sometimes called ***disturbances*** or ***errors***.

[13] The concept of a graphical causal model, or causal network, is related to the more widely known concept in computer science, particularly in artificial intelligence, of a Bayesian network. Bayesian networks constitute an important tool for reasoning in the presence of uncertainty, i.e. for reasoning based on probabilistic information (Darwiche, 2010). A Bayesian network represents the joint probability distribution of the variables encoded in the network as nodes, but the directions of the arrows do not necessarily encode the directions of the causal relationships. As stated in Pearl (2001, p. 4), 'A causal network is a Bayesian network with the added property that the parents of each node are its direct causes'.



(a) Model M (b) Model M_0

FIGURE 2.2: Example causal diagrams for SEM

For instance, for the causal diagram in Figure 2.2(a), the associated (nonparametric) SEM (structural model M) is a system of three functions, one for each observed variable, shown in Equation 2.1:

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= f_X(z, u_X) \\ y &= f_Y(x, u_y) \end{aligned} \tag{2.1}$$

where U_Z, U_X and U_Y are assumed to be jointly independent but otherwise arbitrarily distributed. Two variables X and Y are said to be ***independent*** if $P(X = x|Y = y) = P(X = x)$ for all values x and y of X and of Y respectively. That is, knowledge that Y has occurred gives us no additional information about the probability of X occurring. Otherwise, X and Y are said to be dependent. This is a symmetric relation - if X is independent of Y then Y is independent of X (Pearl et al., 2016, p. 10). This definition extends to the notion of independence of a finite number of variables, so the variables in set of variables S are said to be ***jointly independent*** if, for all subsets R of set S, all the variables in set R are independent. For example, for three variables A, B, C to be jointly independent, all of the following must hold for all values a, b, c of A, B, C respectively: $P(A = a|B = b, C = c) = P(A = a)$, $P(B = b|A = a, C = c) = P(B = b)$, $P(C = c|A = a, B = b) = P(C = c)$, $P(A = a|B = b) = P(A = a)$, $P(A = a|C = c) = P(A = a)$.

Each observed variable in the causal diagram has one structural equation in the SEM system associated with it, denoting how the value of the variable (on the left hand side) is a function of its parents (on the right hand side), i.e. a function representing the causal process or mechanism that determines the value of this variable. The absence of a variable from the function arguments in the right hand side of the equation encodes the assumption that the data generating process ignores that variable in the causal mechanism that determines the value of that variable. A system of such functions are called ***structural*** if they are assumed to be autonomous, i.e. if each function is invariant to possible changes in the form of other functions (Pearl, 2009a, p. 107).

In all causal diagrams in this thesis except the one in Figure 2.2, to simplify the analysis, background factors are not explicitly included in the causal diagrams (following the conventions in Shalizi and Thomas (2011), where causal DAGs are used to reason about social influence versus homophily).[14]

In SEM, interventions are represented as follows (per Pearl, 2009a, p. 107-108). In the example causal model M of Figure 2.2(a) and SEM model M (system 2.1), to emulate an intervention that sets X to the value x_0 in a causal model M with a pre-intervention joint distribution $P(x, y, z)$, one must replace the equation for X from the SEM system (i.e. the equation $x = f_X(z, u_X)$ in system 2.1) with the equation $x = x_0$, and leave all other equations unchanged, thus obtaining the SEM system 2.2. This SEM system denotes a new structural model, M_0, which represents the post-intervention joint distribution $P(y, z|do(x_0))$, and is depicted in the causal diagram in Figure 2.2(b).

$$\begin{aligned} z &= f_Z(u_Z) \\ x &= x_0 \\ y &= f_Y(x, u_y) \end{aligned} \tag{2.2}$$

2.6.3.3 Dealing with confounding

Graphical causal models are particularly useful for identifying latent (unobserved) variables that introduce *confounding bias* to the estimate of the causal effect of a variable X on another variable Y, and for then *adjusting for* those variables to obtain the unbiased causal effect of X on Y. This is illustrated using the simple example causal model in Figure 2.1, whose structure appears in our model of the influence (the causal effect) of online communications versus the effects of other factors, on observed outcomes, as we shall see. Figure 2.1 represents a situation where the observed variable X is a cause of the observed variable Y, but variable U is a latent (unobserved) cause of both X and of Y. Let us next consider two toy examples of what this Figure may represent.

To borrow and adapt a toy example commonly employed by Pearl (e.g. in Pearl, 2001, 2009b; Pearl et al., 2016), in Figure 2.1, let Y represent whether the pavement is wet ($Y = 1$) or dry ($Y = 0$), let X represent whether the sprinkler is on ($X = 1$) or off ($X = 0$), and let U represent whether the season is rainy ($U = 1$) or dry ($U = 0$). If the season is a dry season ($U = 0$), like summer, that may cause the sprinkler to be on ($X = 1$) which in turn may cause the pavement to be wet ($Y = 1$). However, if the season is not dry but rainy ($U = 1$), e.g. autumn, then this will likely cause the sprinkler to be off ($X = 0$), but the pavement may again be wet ($Y = 1$) due to the seasonal rain. This sprinkler-season-pavement example helps illustrate three kinds of inferences one can derive from causal

[14] To avoid confusion on how structural equation methods are interpreted versus how traditional regression methods are interpreted, Pearl (2009a, p. 104) notes that background (or exogenous) factors in structural equations are fundamentally different from residual terms in regression equations. This is because the latter are artifacts of analysis, which by definition are uncorrelated with the regressors (the independent variables), while the former are part of physical reality, i.e. of the causal mechanisms that determine the variations observed in the data, and are treated as any other variable.

diagrams, by ‘propagating information in any direction’ (Pearl, 2001, p. 2): ***prediction, abduction***, and ***explaining away***, in the following way (adapted from Pearl (2001, p. 2)). If the sprinkler is on, then the pavement is probably wet (prediction). On the other hand, if one sees that the pavement is wet, this makes it likely that the sprinkler is on, or that it is a rainy season (abduction); but if one then observes that the sprinkler is on, then that reduces the likelihood that it is a rainy season (explaining away).

As another toy example closer to the topic of this thesis, in Figure 2.1, let Y represent whether Bob decides to watch (e.g. via online streaming) a crime series ($Y = 1$) or not ($Y = 0$), with X representing whether Bob’s friend Alice posted on social media that she just watched and enjoyed that same crime series ($X = 1$) or whether she did not ($X = 0$), and U representing the presence ($U = 1$) or absence ($U = 0$) of an interest in crime series shared by Bob and Alice. Let’s assume that an investigator has data on X and Y but does not have on U, i.e. Bob and Alice’s interest or not in crime series in general is a latent, or unobserved, factor. In this model, the possible existence of a shared interest between Alice and Bob in crime series (U) affects whether Bob will watch this particular series (Y) not only directly (via the arrow $U \rightarrow Y$) but also indirectly, via affecting whether Alice will also watch (and post about watching) this particular series, X which in turn causally affects Y (the path $U \rightarrow X \rightarrow Y$). So, under this model, if we want to estimate the effect of Alice’s posting online that she watches a given series (X) on whether Bob will also watch this series (Y) we need to observe and account for whether they both share a pre-existing interest in series of this kind (U), in order to disentangle the effect ($X \rightarrow Y$) of Alice’s online post (X), which is the effect of interest, from the effects of any pre-existing interest both Alice and Bob might have had ($U \rightarrow X \rightarrow Y$ and $U \rightarrow Y$). We shall discuss this more formally below. Of course, in a realistic scenario there will likely be many more causal factors affecting what series Alice and Bob watch and post about (X and Y), which for the sake of simplicity and clarity are not shown in this toy example. A more complete picture of causal factors that may affect both online communications (X) and individual-level outcomes of interest (Y) will be presented in Chapter 5 of this thesis.

For Figure 2.1, the goal is to estimate the causal effect of X on Y, which is written as $P(Y = y|do(X = x))$ in Pearl’s *do-notation*, denoting, as above, the distribution of Y which would be generated, hypothetically, if X were set to the particular value x through experimental *manipulation* or *intervention*. In the causal graph this would mean deleting all arrows into X, setting X’s value to x, and leaving the rest unchanged. Again, this post-intervention distribution of Y is not in general the same as the ordinary conditional distribution $P(Y = y|X = x)$, as the latter represents taking the original, pre-intervention, population and *selecting* from it only the sub-population where $X = x$. The mechanisms that set X to that value may have also influenced Y through other channels, so the latter distribution would not typically really tell us what would happen if we externally manipulated X.

Figure 2.1 illustrates this point. If one considers the dependence of Y on X, in the form of the conditional $P(Y = y|X = x)$, one sees that there are two channels of information flow between X and Y: one is the directed causal path from X to Y, represented by $P(Y = y|do(X = x))$. However, there is also another, indirect path, between X and Y through their unobserved common cause U, where observing X gives information about its parent U, and U gives information about its child Y. If one just observes X and Y, one cannot distinguish the causal effect from the indirect inference - the causal effect is *confounded* with the indirect dependence between X and Y created by their common cause U. That is, this path through U creates *spurious associations* between X and Y, and is therefore called a 'spurious path' or a 'backdoor path' (Pearl et al., 2016, Section 3.3). More generally, the causal effect of X on Y is confounded whenever $P(Y = y|do(X = x)) \neq P(Y = y|X = x)$. If there is a way to write $P(Y = y|do(X = x))$ in terms of distributions of observables, then one says that the confounding can be removed by an ***identification***, or ***deconfounding, strategy***, which renders the causal effect ***identifiable***. Rendering the causal effect identifiable constitutes solving the ***identification problem*** with respect to that causal effect.

Formally, to test whether there is confounding, one must first test whether some variables ***block*** (stop the flow of information or dependency along) all paths from X to Y, using the ***d-separation criterion***, presented in Definition 2.6.2 as per Pearl (2009b).

Definition 2.6.2. D-separation or blocking. A set of nodes Z *block* or *d-separate* a path p if and only if (i) p contains a *chain* $i \rightarrow m \rightarrow j$ or a *fork* $i \leftarrow m \rightarrow j$ such that the middle node m is in set Z, or (ii) p contains a *collider* $i \rightarrow m \leftarrow j$ such that neither the middle node m, nor any of its descendants, are in set Z. Then, a set Z d-separates X from Y if and only if Z blocks *every* path from X to Y.

The graphical criterion of d-separation is related to the probabilistic concept of conditional independence, defined below in Definition 2.6.3. This relationship is very useful as it allows one to test graphical properties of the graphical causal model using standard conditional probabilities. For instance, this is particularly useful in evaluating whether a proposed graphical causal model violates any of the dependencies and independencies in a given dataset (which shall be applied in this thesis, in Chapter 7.2).

Definition 2.6.3. Conditional independence. When any variable Z d-separates a variable X from a variable Y, then X is ***conditionally independent*** of Y given Z, written as $X \perp\!\!\!\perp Y|Z$. That is, once one knows that Z has occurred, learning that X has occurred does not change the probability of occurrence of Y.

This notion of d-separation or blocking is then used in defining the *backdoor criterion*, in Definition 2.6.4 per Pearl (2009b). This is a graphical criterion allowing one to identify those variables that should be adjusted for in order to remove the confounding bias from the causal estimate of interest.

Definition 2.6.4. backdoor criterion and deconfounding set. A set of variables Z satisfies the backdoor criterion (as per Pearl (2009b)) relative to X and Y if: (i) no node in Z is a descendant of X, and (ii) Z blocks every path between X and Y that contains an arrow into X. Then the set Z is called a ***sufficient, admissible*** or ***deconfounding set***. The ***minimal deconfounding set*** is the smallest such set Z that satisfies the backdoor criterion.

Finding the deconfounding set permits the confounding bias to be removed, thus rendering the causal effect X on Y identifiable from non-experimental data, using the ***backdoor adjustment formula*** presented in Definition 2.6.5 (Pearl, 2009a,b):

Definition 2.6.5. backdoor adjustment formula

$$P(Y = y|do(X = x)) = \sum\nolimits_{z} P(Y = y|X = x, Z = z)P(Z = z) \tag{2.3}$$

Since the right-hand side of Equation 2.3 contains only standard conditional probabilities, which are estimable from observational, non-experimental data, the causal effect of X on Y can be estimated from such data without bias.

In the example of Figure 2.1, one sees that variable U satisfies the backdoor criterion, and hence, to obtain the causal effect of X on Y, one should simultaneously measure X, Y and U for every member of the randomly-selected sample under study, and then obtain the causal effect by using the backdoor adjustment formula (Equation 2.3) for $Z = \{U\}$.

In summary, to remove confounding and obtain the unbiased causal effect of X on Y, the ***deconfounding strategy*** is described in Definition 2.6.6.

Definition 2.6.6. Deconfounding strategy. The deconfounding strategy for obtaining the unbiased causal effect of X on Y is composed of the following three steps:

1. Select a large random sample from the population of interest;
2. For every individual in the sample, measure X, Y, and all variables in Z; and
3. Adjust for Z by partitioning the sample into groups that are homogeneous relative to Z (i.e. one group per value of Z), assess the effect of X on Y in each homogeneous group, and then average the results, as per Equation 2.3.

2.6.3.4 Potential outcomes and counterfactuals

The potential outcomes framework, of Neyman (1923) and Rubin (1974), is useful for asking ***counterfactual*** questions, that is, 'what if' questions where the 'if' portion is untrue and unrealised; hypothetical questions about what would the value of the outcome Y be if a causal factor X, had had the value x', given that in reality X had value x, and Y had value y.

It has been demonstrated that definitions of counterfactuals, and the symbolic machinery for analysing them, emerge as natural by-products of SEM (Pearl, 2009a, Section 3.4, Definition 4). As done with the SEM system 2.1 in the example above, the phrase 'had X been x'' is interpreted as an instruction to replace the equation for X with the equation $X = x'$ in the SEM system, and this replacement allows the constant x' to differ from the actual value of X (namely $f_X(z, u_X)$) without rendering the system of equations inconsistent. Therefore, one obtains a formal interpretation of counterfactuals in multistage models, i.e. models where the dependent variable in one equation may be an independent variable in another.

It has also been demonstrated how Structural Causal Models (composed of SEM and causal diagrams), that can be used to compute counterfactuals, unify and subsume the potential outcomes framework and other approaches to causal analysis (see Pearl, 2009a, especially Section 3.4), and provide a logically equivalent methodology that yields the same answers as the potential outcomes apparatus to any given causal question (Pearl et al., 2016). So, even if counterfactual notation and the potential outcomes framework was used in the causal analyses of this thesis, the same results would be obtained: expressions of the form $P(Y = y|do(X = x'))$ would be written instead in the form $P(Y_{x'} = y)$, and the backdoor criterion (2.3) that all the analyses here are based on (in Chapters 5 to 7), which uses the causal structure of the causal diagram to derive the admissible deconfounding set Z, is mirrored in the ***conditional ignorability condition*** of the potential outcomes framework which results in the same admissible set Z (Pearl et al., 2016, p. 102-3; Pearl, 2009a, p. 129).

In terms of the potential outcomes framework, Pearl has criticized the opacity of its language and the difficulties this introduces in causal reasoning, as well as the lack of guidance in covariate selection in this framework, in contrast to what graphical causal models and the backdoor criterion offer (Pearl, 2009b, p. 350; Pearl, 2009a). It is argued that this opaqueness of counterfactual independencies in the potential outcomes framework explains why many researches in that tradition wrongly assume that adjusting for as many pre-treatment covariates as possible will help decrease confounding bias, when it has been found that this practice can in fact *increase* bias when the covariates do not satisfy the backdoor criterion (this will be discussed further in Section 2.6.4.2).

2.6.4 Estimation of causal effects

In the empirical analysis in this thesis (Chapters 6 and 7), where the effects of online social influence versus the effects of other causal factors will be estimated, nonparametric estimation methods will be used, and the probabilities in the backdoor formula will be calculated directly from our frequency data. (This will be described in more detail in Chapter 6.4.3).

This section explains the advantages and rationale of this approach. It then briefly covers some of the most commonly used estimation methods from the literature: propensity score methods, matching methods, and linear causal models, and their respective merits and limitations.

2.6.4.1 Nonparametric estimation

In this thesis, in Chapters 6 and 7, nonparametric estimation is used for the causal (and associational) effects. The causal (and associational) quantities that have been used this far, most importantly in the backdoor formula (Equation 2.3), will be calculated directly from the data, using frequencies for each probability. When using causal methods, the probabilities in the formulae may come from frequency data (per the frequentist school of statistics) or from subjective assessment (per the Bayesian school of statistics), i.e. causal methods are orthogonal to the frequentist versus Bayesian debate, as explained in Pearl (2009a, p. 119).

An advantage of using nonparametric estimation is that there is no need to make additional assumptions about the form of the function f in the structural equation models. Making such assumptions is often considered a big simplification, and parametric assumptions may not always apply well to the given setting under study. For instance, linear models are particularly common in causal analyses, but these require the strong assumption that relationships between variables are linear, and that all error terms follow a Gaussian distribution, which may not always be appropriate (Pearl, 2009a; Tsapeli and Musolesi, 2015). A potential disadvantage of nonparametric estimation is intractability, as when there are many confounders to adjust for calculating estimates may become intractable (the state space, i.e. the number of possible value combinations among the confounders, increases exponentially with the number of confounders). However, this does not apply in the empirical causal analyses in this thesis, as there are few confounders to adjust for (since the analysis is done on the collective level, and individual-level confounders are aggregated). So, there is no need here for parametric estimation, as there are not too many confounders to adjust for.

In Chapters and 6 and 7 of this thesis, one important area of focus will be the analysis of the presence or absence of confounding, so it is of interest to compare causal versus associational effects, recalling that if these two are equal, there is no counfounding bias, and otherwise there is confounding bias present. That is, as discussed in Section 2.6.3.3, the causal effect of X on Y is confounded whenever $P(Y = y|do(X = x)) \neq P(Y = y|X = x)$. The conditional probability with the *do* opetator on the left hand side is called the causal effect in the literature (e.g. Pearl, 2009a). In this thesis, the conditional probability on the right hand side will be called the *selection effect*, as conditional probabilities on frequency data represent merely filtering the data, i.e. selecting out of the data the subpopulation for which X has the value x (Pearl et al., 2016; Shalizi, 2013, Chapter 25.1).

As discussed, causal analysis is about dealing with *change*, i.e. with how would the value of outcome Y change if the value of one of its causal factors, X, were to change from x to x'. So, it is common to ask how the outcome would change when the causal factior (or 'treatment') changes from being absent to being present. For exampe, how would patient recovery rates change if they were given a drug, versus if they were not given a drug. Therefore, we are interested in *comparing* $E(Y = y|do(X = x))$ to $E(Y = y|do(X = x'))$. This comparison can be expressed mathematically using the difference between the two causal quantities, or using other expressions, e.g. involving ratios of the causal quantities (Morgan and Winship, 2014; Pearl, 2009a).

The *average difference* is one such common comparative measure: $E(Y|do(X = x)) - E(Y|do(X = x))$. This quantity is also known in the literature as the *Average Causal Effect (ACE)* or the *Average Treatment Effect (ATE)* (from medical contexts).

For example, in cases of a binary outcome Y and a binary cause of interest X, the non-causal quantity $P(Y = 1|X = 1) - P(Y = 1|X = 0)$ has been called in Pearl (2009a, p. 114-115) 'the risk difference', while the difference $P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$ has been called 'causal effect difference', or the 'average causal effect (ACE)', or the 'average treatment effect (ATE)', or sometimes the 'causal effect (of X on Y)'. [15]

As noted in Morgan and Winship (2014, p. 47), it can be shown that, when the outcome Y is binary, the difference of conditional expectations equals the difference of conditional probabilities. Given that $E(Y|X = x) = \sum_y y \times P(Y = y|X = x)$, the difference $E(Y|X = x) - E(Y|X = x')$ expands to:

$$\begin{aligned}
&E(Y|X = x) - E(Y|X = x') = \\
&= \textstyle\sum_{y=\{0,1\}} yP(Y = y|X = x) - \sum_{y\{0,1\}} yP(Y = y|X = x') = \\
&= 1 \times P(Y = 1|X = x) + 0 \times P(Y = 0|X = x) \\
&\qquad -(1 \times P(Y = 1|X = x') + 0 \times P(Y = 0|X = x')) = \\
&= P(Y = 1|X = x) - P(Y = 1|X = x')
\end{aligned} \tag{2.4}$$

[15] For variables X and Y that can take multiple values, the quantity $P(Y = y|do(X = x)$ is what one normally wishes to estimate, where x and y are values that X and Y can take. This is the the general causal effect, usually called the 'causal effect'. As causal effects are about change, though, one also wishes see the causal effect of increasing Y's exposure to X by one unit, i.e. to compare $P(Y = y|X = x)$ to $P(Y = y|X = x+1)$ (Shalizi (2013, p. 574)). For binary variables X and Y, the goal is commonly to estimate the difference, $P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$, (or the difference, $E(Y = 1|do(X = 1)) - E(Y = 1|do(X = 0))$, which as we shall see is equal to the Probability-based expression). This is usually called the 'causal effect difference', or the 'average causal effect (ACE)', or 'the average treatment effect (ATE)' (Morgan and Winship, 2014; Pearl et al., 2016, p. 55-56; Pearl, 2009a; Shalizi, 2013). However, the latter quantity is sometimes also referred to 'the causal effect', e.g. in Pearl (2009a, p. 114-115), Shalizi (2013, p. 574).

So, the derivation in Equation 2.4 shows that, for binary Y, $E(Y = 1|X = x) - E(Y = 1|X = x') = P(Y = 1|X = x) - P(Y = 1|X = x')$. The same can be shown for the respective causal difference, i.e. that $E(Y = 1|do(X = x)) - E(Y = 1|do(X = x')) = P(Y = 1|do(X = x)) - P(Y = 1|do(X = x'))$ (the only difference would be the addition of adjustment for, i.e. marginalisation over, the deconfounding set Z, which does not affect the derivation).

Therefore, in this thesis, the simplified, probability-based, differences will be used for the ACE (Equation 2.5, for binary Y), and for the respective Average Selection Effect (ASE, Equation 2.6, for binary Y). The ASE is also known in the literature as the *Average Risk Difference* (Pearl, 2009a, p. 115).

$$ACE = P(Y = 1|do(X = x)) - P(Y = 1|do(X = x')) \tag{2.5}$$

$$ASE = P(Y = 1|X = x) - P(Y = 1|X = x') \tag{2.6}$$

More specifically, as the causal variables (X) studied here are also binary, as is commonly the case in causal analysis (comparing the results in the presence versus absence of a treatment), the formulae used are given in Equation 2.7 for the ACE, and Equation 2.8 for the ASE.

$$ACE = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0)) \tag{2.7}$$

$$ASE = P(Y = 1|X = 1) - P(Y = 1|X = 0) \tag{2.8}$$

As stated in Pearl (2009a, p. 115), whenever Z is an admissible deconfounding set, the backdoor criterion (Equation 2.3) also applies for the effect differences ACE, ASE, in which case the ASE can be expanded via the backdoor criterion, by marginalising over all strata (values) of Z (Equation 2.9). So Equation 2.9, which is the expansion of Equation 2.8, will equal the ACE (Equation 2.8) whenever Z is an admissible deconfounding set.

$$\begin{aligned}ASE &= P(Y = 1|X = 1) - P(Y = 1|X = 0) = \\ &= \textstyle\sum_z [P(Y = 1|X = 1, Z = z) - P(Y = 1|X = 0, Z = z)]P(z = z)]\end{aligned} \tag{2.9}$$

It is this property that is used in order to analyse the existence or absence of confounding in the causal effects of online communications on outcomes, by comparing the ACE and the ASE, in Chapters 6 and 7.

In this thesis, in the empirical analysis of Chapters 6 and 7, the terms ‘causal effect’ and ‘selection effect’ will be used for short, to refer to the binary ACE and ASE. That is, to refer to the differences of the respective causal conditional probabilities, with Y and X being binary variables: Equation 2.7 for the ACE, and Equation 2.8 for the ASE.

So, overall, the estimation approach used here is nonparametric, frequentist, with binary causal and outcome variables, using the above difference-based equations for the average causal and selection (associational) effects.

2.6.4.2 Other estimation methods

This section briefly discusses some other estimation methods, specifically some of the most frequently used estimation methods from the literature: propensity score methods, matching methods, and linear causal models, and outlines their respective merits and limitations.

Propensity score methods. The propensity scores framework (Rosenbaum and Rubin, 1983) is one way of estimating causal effects. There has been some controversy around propensity scores, as it is a very popular method that has often been used under the wrong assumption that one can choose any set of covariates S, and the propensity score method itself would remove the confounding bias. However, this is not the case: the propensity score method is not an identification method, and it only yields unbiased estimates if the covariate set S is admissible (a valid deconfounding set, satisfying the backdoor criterion). This is explained in detail in Pearl (2009b, Chapter 11.3.5), Pearl, 2009a, and also in Shalizi and Thomas (2011) and Shalizi (2013). Another related problematic assumption in the propensity scores tradition is that it is safe for one to adjust for as many covariates as possible (regardless of whether they satisfy the backdoor criterion); however, it has been shown that adjusting for variables that do not satisfy the backdoor criterion can in fact increase confounding bias (discussed in Pearl, 2009a). Propensity scores are also commonly used as part of matching methodology, which is described next.

Matching methods. Matching is another method that can be used for the estimation of causal effects of X on Y. In the special case where X is binary, for every value of the covariates S, s_i, let us take any unit (participant) i where X=1 (treatment unit). Then, suppose we can find another unit $i*$ with the same value s_i and with $X = 0$ (control unit). The latter unit is called the *matched* control unit, of the former treatment unit. Then, $Y_i - Y_{i*} = E(Y|X = 1, S = s_i) - E(Y|X = 0, S = s_i)$, i.e. the comparison between the outcome of the treated unit and this matched control unit is an unbiased estimate of $E(Y|X = 1, S = s_i) - E(Y|X = 0, S = s_i)$, provided that S is an admissible (i.e. deconfounding) set. If a match $i*$ can be found for every unit i, then, by the law of large numbers, the average $Y_i - Y_{i*}$ over all units is approximately the ACE. This is a nonparametric method, and if S is large, the curse of dimensionality may make the estimate intractable. In those cases, one could use the propensity score of S, $L(S)$, as an attempt to lighten the curse of dimensionality in estimating the causal effect (although it remains in the estimation of the propensity score, if the latter is estimated nonparametrically). Contrary to what has often been assumed in the literature, even though matching methods may mimic the logic of randomisation, they are not identification methods: if S is not admissible (per the backdoor criterion), the causal effect will not be free of confounding bias (Arceneaux et al., 2010; Shalizi, 2013, Section 27.1.3). In addition, as mentioned, propensity scores are very commonly used for matching (e.g. in the online social influence literature, in Anagnostopoulos et al., 2008; Aral et al., 2009). However, there has been some debate on whether one should match on propensity scores (King and Nielsen, 2016; Pearl, 2009b, Section 11.3.5). Either way, as mentioned, neither matching nor propensity scores can remove confounding bias; they both rest on the assumption that S satisfies the backdoor criterion.

Other parametric methods: linear methods. When the deconfounding set Z is large, one is faced with the curse of dimensionality, where causal effects may be intractable. In that case, using parametric, rather than nonparametric, estimation may be necessary. This is not the case for the analyses in this thesis, but as linear methods are particularly common in the causal literature, they are briefly discussed here. Linear estimation requires the strong assumption that relationships between variables are linear, and that all error terms have Gaussian distribution, which may not always be appropriate for any given dataset (Pearl, 2009a). However, making this assumption simplifies the estimation process greatly (Pearl et al., 2016, Section 3.8), as e.g. standard statistical methods such as linear regression can be used to estimate causal effects. Similarly to nonparametric estimation, the backdoor criterion (Definition 2.6.4) is again needed in order to determine the admissible set Z, in order to regress Y on X and Z. The coefficient of X will then give the total effect of X on Y (Pearl et al., 2016, Section 3.8.3). Linear regression on its own is a descriptive statistical method, and cannot tell us what the admissible set is, unless we invoke the graphical backdoor criterion; it is not a method for removing confounding (Pearl et al., 2016, Section 3.8.1; Ogburn, 2017, Section 6.2).

2.6.5 Evaluating the fit of a graphical causal model to data

As stated in Pearl et al. (2016, p. 48-50), causal models have testable implications in the data sets they generate. That is, the structure of a proposed graphical causal model has implications that can be tested in the empirical data, using standard statistical tests of independence and conditional independence, by applying the d-separation criterion.

As discussed in Section 2.6.3.3, the d-separation criterion (Definition 2.6.2) states that 'If Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z (written $X \perp\!\!\!\perp Y|Z$)'. So the notation $X \perp\!\!\!\perp Y|Z$ stands for $P(Y|X, Z) = P(Y|Z)$. In consequence, one can test the blocked paths of the graphical causal models using standard statistical conditional independence tests in the empirical data.

There are other ways to test the fitness of a causal model to data, which require testing a statistical hypothesis over the entire model, having fully specified the functional form of the model and estimated its parameters, in order to then be able to evaluate the likelihood that the data was generated by the hypothesized model and not by sheer chance. As discussed in Pearl et al. (2016, p. 49), this can be done (approximately) by assuming a linear and Gaussian model (all functions linear and all error terms normally distributed) because then the joint distribution (also Gaussian) can be expressed in terms of the model's parameters, and then the likelihood that the data was generated by the fully parametrized model can be evaluated. However, these kinds of global tests have the disadvantage that they require a fully specified and parametrized model, with all parameters estimated. So if any parameter cannot be estimated, then the joint distribution cannot be estimated, hence no part of the model can be tested. This can occur when some of the error terms are correlated, or equivalently when some of the variables are unobserved. Another drawback of this standard testing procedure is that it tests models globally, so if it is found that the model does not fit the data well, one cannot determine why that is, i.e. which edges should be removed or added to improve the fit. In addition, when one tests the model globally, the number of variables involved may be large, and if there is measurement noise of each variable then the test will not be reliable.

In contrast, the d-separation based tests have the advantages that, first, d-separation is nonparametric, so it is only based on the structure of the causal DAG, and does not require the specification of specific functions that connect variables. Secondly, it tests the models locally, rather than globally, allowing one to identify specific areas where the proposed model is flawed, and repair those, instead of having to start from scratch with a whole new model.

In this thesis, in the empirical implementation of causal modelling and estimation of the influence of online social communications on collective outcomes, these d-separation based tests will be used to assess the fit of the proposed model to the empirical data, in Section 7.2.

2.7 Summary

This chapter has reviewed the literature on understanding, analysing and measuring the social influence of online communications, discussing how it has been conceptualised in the literature and the practical challenges in measuring it.

Having first introduced how social influence has been studied in the context of Web-mediated interactions, and how this has been done not only at the individual level but also, less frequently, at the collective level, this chapter proceeded to present definitions of the term ‘social influence’. These definitions include definitions from the social sciences, and from the causal methods literature, which inform the definition of social influence that will be used for the analyses in this thesis (discussed in this chapter and presented in more detail in Chapter 4.1). Next, this chapter presented the popular contagion-based paradigm for understanding the social influence of online communications, and its various and important limitations as they have been discussed in existing literature. As one promising way to address the limitations of the contagion paradigm, this chapter next presented an alternative, causal, paradigm, for conceptualising and measuring online social influence. As this thesis will propose and discuss in the following chapters, this causal paradigm is an important component in the effort to conceptualise social influence in a manner that is better aligned with how the term ‘social influence’ has traditionally been understood in the social sciences, and in the effort to measure and disentangle the social influence of online communications from the influence of other causes on the outcome of interest.

Building on this chapter, the next chapter (Chapter 3) will present an analytical and empirical critique of the contagion-based paradigm, which covers the limitations presented in this chapter and expands upon them, proposing a more comprehensive classification of key limitations of the contagion paradigm. The following chapters (Chapters 4 - 7) will propose a causal conceptual and methodological framework which, drawing upon the definitions and the causal paradigm’s methods described in this chapter, can address the limitations of the contagion paradigm presented in the critique of Chapter 3.

Chapter 3

Critique of the contagion-based paradigm

As an initial contribution, this chapter presents a critique of the contagion-based paradigm for understanding social influence online. This critique is two-fold: it is composed of an analytical critique, and of an empirical critique which is based on the empirical analysis of a real-world dataset. Based on the limitations presented in this critique, Chapters 4 to 7 will present a causal conceptual and methodological framework that can successfully address these limitations.

For the discussions in this chapter, it will be useful to remember the definitions of 'influence' and 'social influence' presented earlier, in Chapters 1.1, 2.3, and 2.6.2. As described there, this thesis defines these two terms in the following manner, which is aligned with how these terms have been historically understood in the social sciences and in everyday parlance. In this thesis, to say that 'A influenced B' is to say that A was one cause of the occurrence of B (whether intentionally or unintentionally, and requiring that no force or coercion was used in an attempt to ensure the occurrence of B). Hence, the 'social influence' of A on B in this thesis is defined as the role A had in causing the occurrence of B, where A represents a social factor, i.e. a cause from the social world, such as a person's or group's action(s), behaviour(s), statement(s), message(s), or other communication(s), and B represents another person's or group's action(s), decision(s), opinion(s), belief(s), message(s), or other outcome(s). (Chapter 4.1 will discuss these definitions in more detail).

Before proceeding to present the two-fold critique itself, it is worth first clarifying its scope and aims.

Firstly, it is stressed here that, as discussed in Section 1.1, contagion-based studies of online interactions (on social media, emails, blogs, or other modes of online communication) should be commended for being based on detailed and extensive analyses, using large online datasets, and for contributing valuable insights (indeed often some of the very first insights), on a range of social phenomena and activities taking place in these new and pervasive settings of online communication. These studies have often offered important results in terms of associational findings about patterns, characteristics and properties of online social networks and/or of online responses, which relied on the assumption that the adoption of focal items (e.g. ideas, opinions) spreads in the manner of contagious diseases along social network ties.

The critique in this chapter discusses the key limitations of the contagion assumption on which these studies rely. The goal of this critique is not to claim that the contagion-based paradigm should never be used. Rather, it stresses that, before proceeding to use the contagion-based paradigm, an investigator should first test whether the contagion assumption holds in the setting under study, rather than automatically assuming that it must hold (as has been also pointed out elsewhere, e.g. in Freelon, 2014; Lerman, 2016; Tufekci, 2014).

That is, as discussed in Chapters 2.4 and 2.5, the contagion paradigm assumes that, when one wants to measure the influence of some online communications posted by a person's or group's social network ties (relating to a focal item, and generally assumed to be interpretable as endorsements or adoptions of the focal item) on whether this person or group adopts the focal item, the only cause that needs to be considered and measured is this set of online communications from the person's network (and, often, the quantitative properties of social ties and network), and no other types of causes (such as the personal traits of those involved, the traits of the focal item, and external circumstances) need be measured.

So, in order to use the contagion paradigm to measure the influence of online communications, this contagion assumption should be treated as a hypothesis to be tested. One should first establish that the contagion assumption holds in the given setting, i.e. that it is safe to ignore other causal factors, and ignoring them does not distort the estimate of the influence of online communications, i.e. does not introduce confounding bias due to the non-zero effects of common causes of the online communications and of the outcome (as discussed in Chapter 2.6, particularly in Section 2.6.3.3). So, one must establish that indeed other confounding causal factors do not distort the measurement of online social influence, i.e. that all other possibly confounding causal factors actually introduce zero or negligible confounding bias to the estimate of the influence online communications on the outcome, in the context under study.[1]

In addition, it is also noted here that the focus of this critique chapter is specifically contagion-based studies, rather than studies that use multivariate statistical methods (e.g. multivariate regression) in order to model and estimate the influence of a factor

[1] And indeed this thesis does test this assumption of contagion, by drawing upon established findings from the literature in Chapter 5, and empirically in Chapter 7, using causal methods in both cases, as these are particularly well suited in examining such causal assumptions.

of interest (e.g. online communications) while taking into account and adjusting for the effects of other possible causes of the outcome by including them as variables in the multivariate model. Indeed, as the contagion-based paradigm does not take other possible causes into consideration, studies in this paradigm generally do not use such multivariate statistical methods that include other causes – rather, they use different methodology. As discussed in Sections 2.1, 2.4, and 2.5 of the Background chapter, contagion-based studies instead tend to either model a setting as a social network and essentially 'simulate' how a focal item would contagiously spread under a threshold model of contagious infection (then measuring properties of this social network's topology, of the infection threshold distribution and of how widespread the focal item eventually is among network nodes), and/or they focus on properties of response patterns, possibly measuring how these are associated with the social network topology.

So, if one were to use standard multivariate statistical methods which adjust for other possible causes, this would go some way towards addressing the above limitation of the contagion-based paradigm in the context of studying online social influence. Still, using causal methods can improve upon the use of these multivariate statistical methods even further. As described in Section 2.6, it has been shown that the (not unusual) practice of adjusting for as many other variables as possible can actually result in increased confounding bias in the estimate of the effect of the variable of interest (Pearl, 2009a, p. 117, 130). To address this, causal methods offer an identification or deconfounding strategy (per Definition 2.6.6, using graphical causal models and the backdoor criterion of Definition 2.6.4, as discussed in Section 2.6.3.3) in order to determine which variables should and should not be adjusted for so as to remove confounding bias.

And even in cases where domain expertise is used to identify relevant possible causes of the outcome to adjust for in the multivariate statistical formula (rather than indiscriminately including for adjustment as many variables as possible), explicitly representing those causal assumptions in a graphical causal model yields a powerful and formal visual representation which, together with the above graphical and computational identification strategy, can formally and systematically show which of these other causal variables should indeed be adjusted for and which (if any) should not be adjusted for as that would increase confounding bias. So, using graphical causal methodology would even in this case offer added value, as it would not only result in a transparent visual representation of the full picture of causal assumptions (in a causal graph), but it would also add an extra layer of a formal and systematic 'check' of which variables should indeed be adjusted for to minimise confounding bias, drawing upon the formal rules of graphical causal methods.

So, once these two causal steps are performed, of 1) drawing the graphical causal model, and 2) using the model and the backdoor criterion to perform the identification step in order to determine which variables should be adjusted for, then standard multivariate statistical methods can be used in step 3), for the estimation of the causal effect (i.e. the influence) of online communications on the outcome. That is, as discussed in Section 2.6.4, estimation methods do not solve the identification problem; rather, the identification step (based on the causal graph) must first be performed before proceeding to the use of standard statistical methods for estimation. Having performed the first two steps, for the third step of estimation there are various multivariate statistical estimation methods that one can use, whether parametric (like multivariate linear regression) or non parametric, depending on the problem and the data at hand, as discussed in Section 2.6.4.

Having clarified the scope of this chapter's critique, the next section will present the first branch of this critique, namely the analytical critique.

3.1 Analytical critique of the contagion-based paradigm

The various known limitations of the contagion paradigm that were presented in Chapter 2.5 have generally not been discussed specifically in the context of the problem of analysing the social influence of Web-mediated communications, but rather in the broader context of analysing Big Data, either digital Big Data (in Marres, 2017), or more specifically Big Data from the Web (e.g. hyperlinks and Twitter in Freelon, 2014), or from online social media platforms (Tufekci, 2014). On the other hand, studies that are concerned with the social influence of online communications specifically have tended to note only one part of the limitations of the contagion-based paradigm (e.g. Bakshy et al., 2011; Shalizi and Thomas, 2011; Watts, 2007) and have tended to not consider the rest of the problems. As a result, the problems that are relevant particularly to the social influence of online communications have generally either been presented interspersed among other broader problems that are less specific to the social influence of online communications, or presented only partially.

Given that this thesis is concerned specifically with the issue of analysing the social influence of online communications, and building on the valuable contributions of existing critiques discussing the limitations of contagion-based practices of Big Data analysis and social influence analysis, this section proposes an analytical critique of the contagion-based paradigm for online social influence. This analytical critique contributes a new, and more comprehensive, classification of the key conceptual and methodological limitations of the contagion-based paradigm for the social influence of online communications. It attempts to systematically capture and analyse what this thesis argues are the key limitations of the contagion-based paradigm, for the problem of analysing the social influence of online communications specifically. Section 3.2 will next demonstrate how these limitations manifest if one attempts to empirically apply the concepts and methods of the contagion-based paradigm to analyse the social influence of online communications in a real-world setting. Chapters 4 to 7 will then present a causal approach that can successfully address these limitations.

In the analytical critique proposed in this section, the key problems of the contagion-based paradigm are classified into the following four classes:

Language. In the contagion-based paradigm, several empirical studies say 'we define influence as [an easily measured platform-specific action]'. However, this kind of definition is not aligned with how influence has historically been understood in the social sciences or in everyday parlance. Further, in this paradigm, it is common to talk in vague, rather than concrete, terms, about the 'spread' (rather than the *adoption*, which is often implied) of 'ideas' (as a blanket term, even for things such as products or news items that are not ideas).

Assumed cause of outcomes. The contagion-based paradigm assumes that the only possible cause of observed outcomes is social influence from someone in one's immediate or extended social network, in the setting of online interaction under study (e.g. on the given social media platform). In empirical studies using this assumption, one common problem that this thinking leads to is the *post hoc ergo propter hoc* fallacy (meaning 'after this, therefore because of this').

Assumed meaning of outcomes. Observed outcomes (actions or behaviours, such as posting content online or responding to online content) in the contagion paradigm are assumed to signify adoption of or agreement with the topic being discussed in that online content, when actually these outcomes are ambiguous and can be interpreted in any number of ways.

Untested assumptions. The assumptions made in this paradigm, often cloaked in vague or ill-suited language (as per the first point above) are often stated as self-evident, as facts of nature, when in reality they are assumptions, the validity of which should be tested empirically or otherwise (e.g. theoretically, drawing upon domain expertise or findings). This kind of logic leads to the *begging the question*, or *circular reasoning*, fallacy being widespread in this paradigm, whereby that which is to be empirically proven (particularly whether someone is an 'influential') is assumed as a given. One important consequence of this problematic reasoning is the issue of what an 'influential' is, as, in the usage of this term in the literature, the defining characteristic of being an 'influential' (which is succeeding in getting others to adopt what they propose) has been conflated with one of the traits (specifically, being well-connected socially) that are often associated with, but not defining of, such people.

Each of the above limitations are discussed in turn, in the following sections.

3.1.1 Language

In the contagion-based paradigm, several empirical studies contain statements of the form ‘we define influence as [a readily measurable platform-specific action]’.

For instance, Ghosh and Lerman (2010b, p. 1) state that, on the Digg social platform, ‘We empirically define influence as the number of in-network votes a users post generates’. Barbieri et al. (2013, p. 1) say that ‘our method [...] defines the level of influence of each user in each community’. Their method is based on the assumption that ‘if we see a group of users acting on item i in a short time frame, and we observe this occurring on various different items, then we can infer that these users are connected in some social network, that they communicate and can influence each other’, and:

> ‘our proposal assumes that item adoptions are governed by an underlying stochastic diffusion process over the unobserved social network, and that such diffusion model is based on community-level influence. By fitting the model parameters to the user activity log D, we learn the community membership and the influence level of each user in each community’. (Barbieri et al., 2013, p. 1)

As another example, in Cappelletti and Sastry (2012), where an algorithm is developed for ranking the top ‘influential’ Twitter users, the authors argue that dictionary definitions of influence are ‘vague’ and that instead one must define influence in terms of platform-specific actions:

> ‘For this work, which is focused on the Twitter environment, the term influence was defined as how much excitation a user causes in the information network, constrained by the topic of interest. In this case, the term excitation [...] was defined as how much interest a user received. In the Twitter platform, one user may reveal interest to others by reading and further interacting with their tweets or username, utilizing the available mechanisms: reply, mention, and retweet. Thus, finding influential users is the equivalent of finding the users that are more related to the tweet or username that is being propagated in the Twitter timeline’. (Cappelletti and Sastry, 2012, p. 5)

Cappelletti and Sastry (2012, p. 1) further explain that their ranking method ‘is based on the notion that the highly influential users in Twitter are those whose usernames are being currently amplified by the Twitter network, via mentions, replies or retweets by other users’.

However, it is argued in this thesis that such definitions and uses of the term 'influence' or 'social influence' are not aligned with how these terms have historically been understood in the social sciences literature or in everyday parlance. The terms 'influence' and 'social influence' have been understood in a manner independent of the kind of online communication channel one may be studying, both in everyday parlance (e.g. per dictionary definitions), and in the academic literature, as discussed in Chapter 2.3. Rather, what the above studies are doing is *interpreting* certain observable and measurable actions in their dataset as *manifestations of* influence from a particular social media user. That is, they are *assuming* that certain actions can be interpreted as manifestations of influence from a particular social media user. Therefore, it is important to not attempt to define or use the 'influence' in an ad hoc manner (as is also pointed out in Freelon, 2014) that clashes with how this term has been historically understood, and to make clear and explicit what the assumptions are (about what is interpreted as influence) that one's work is based on.

That is, it is argued here that one cannot define as influence an arbitrary action that one observes, as influence has been historically understood in a manner that does not equate it to an observed response to something on social media, such as an upvote or a retweet. Rather, as discussed in Chapters 1.1 and 2.3, influence (from a specific source, e.g. a social media user, or a post on social media) is the *(causal) effect*, or the role, or the impact, that this specific source had in the occurrence of such observed actions. Therefore, such observed actions might (if appropriate) be considered *outcomes* or *evidence* or manifestations of social influence from some source (provided that there is evidence to support that this source did indeed play a role in causing the occurrence of the outcome), but they are not the social influence itself. The above language leads to confusion, and to conflation of observed outcomes (e.g. retweets) with one very specific source of causation, or of influence, that may (or may not) have caused them (the source being a social media user or post). That is, there is a conflation between the causal concept of social influence from a given source and measurable platform-specific actions which represent outcomes that may have occured *due to* that online source (or due to other influencing factors). Therefore, it is stressed here that it is important, when studying the social influence of certain online actions, events, or people on specific outcomes of interest, to be clear about what actions or events in the dataset represent the outcomes, and what actions, events or people may represent causes (sources of social influence) of such outcomes.

In addition, in this popular paradigm, it is common to encounter language that relies on vague, rather than concrete, terms, about the 'spread' of 'ideas' (e.g. Cha et al., 2010; Kempe et al., 2003; Kitsak et al., 2010). That is, the word 'spread' is used, rather than the word '*adoption*', and the word 'ideas' is often used as a blanket term, even for things such as products or news items, that are not ideas. But, generally, when saying that 'an idea has spread', or that 'an idea is widespread', what that usually means concretely is that this idea has become held, or espoused, or adopted widely, by many people. Alternatively, one might argue that saying that ‘an idea is widespread’ may in some contexts mean merely that many people have become aware of this idea, even if they may have not adopted it. But adoption or endorsement is a very different outcome than mere awareness. Being aware of an idea, but not having adopted it, may mean that one is still skeptical about that idea, or is not particularly interested in it. For example, purchasing a product is very different from merely being aware of it; espousing an idea or an opinion is very different from merely being aware of it. If what is meant by ‘something being widespread’ is ‘something being adopted widely’, as is likely, given that contagion models, and their empirical applications, generally talk about adoption outcomes (e.g. Barbieri et al., 2013; Cha et al., 2010; Easley and Kleinberg, 2010; Kempe et al., 2003; González-Bailón et al., 2011), then this should be stated clearly. Hence, this thesis argues that it would be best to specify what is a meaningful outcome in a given context (e.g. is it adoption? Is it awareness?) and rather than talking about an idea, or some information, or some other focal item, ‘spreading’, to talk about *the outcome relative to the focal item* that is being observed to occur widely - is the adoption of this idea ‘spreading’, i.e. being observed in many people? Is it awareness of this information that is ‘spreading’, i.e. have many people become aware of this information?

The usage of the more vague language about things ‘spreading’ is problematic also because if instead one were to talk about the ‘adoption of ideas’ then this claim might invite more questions than a claim about the ‘spread of ideas’. Hence the ‘spread of ideas’ may seem like a more neutrally occurring and anodyne phenomenon than the ‘adoption of ideas’. That is because ‘adoption’ implies an adopter, a person on the receiving end, and this implies some degree of agency on the part of that person. However usage of the word ‘spread’ does not suffer from this, as the person on the receiving end (to whom the idea has spread) is not part of the picture. Indeed, ‘spread***er***s’ are part of the vocabulary of these studies (e.g. González-Bailón et al., 2011; Kitsak et al., 2010), as they are also studied in medicine and epidemiology, but ‘spread***ee***s’ are not. People on the receiving end are only assigned a numerical threshold to determine whether or not they will be ‘infected’ or ‘activated’ by the ‘idea’, which is a function of the number (or proportion) of their immediate social ties that have already been ‘infected’ by the ‘idea’. There is generally no clear explanation as to what this threshold value represents in the real world, it is not discussed whether it is supposed to reflect the respective person’s agency, and if so based on what criteria and what procedure might a person’s agency be encoded into a single numerical threshold value, and generally the agency of the people on the receiving end is not addressed in those contagious models.

This language is part of the contagious disease analogy: people do not have a say in whether a disease spreads to them, e.g. in whether they catch the flu. However, people do not ‘catch ideology’ like they catch the flu.[2] So it seems that this kind of

[2] To borrow a phrase from Shalizi’s blog post Shalizi, 2010 about the paper Shalizi and Thomas (2011)

borrowed language (and analogy) from the spread of diseases may have been taken too far (as also noted in Tufekci, 2014), and is no longer applicable when it comes to the spread of ideologies. Indeed it is not clear that this contagion paradigm serves a purpose when studying the adoption of other focal items like ideas, beliefs, opinions, or even products, but the contagion paradigm has long been applied to make inferences about how online (and offline) communications influence such outcomes. For example, González-Bailón et al. (2011) use the contagion paradigm and Twitter data to make claims on how social media communications influence political protest recruitment and participation. However, this outcome (whether one will be recruited to or join a protest), is deeply tied in with one's political ideologies, which are not included in the contagion model. In addition, Cha et al. (2010) notes that so-called *influentials*, special people who supposedly can make others adopt anything they propose, are also known as 'opinion leaders' – that is, it is assumed that opinions are also focal items that spread like a contagious disease. Further, product purchases (adoptions) have long been one of the key application areas of the contagion paradigm (two of the first papers proposing this were Domingos and Richardson (2001) and Kempe et al. (2003), which have been frequently cited by other works in the contagion paradigm), even though one might question how purchasing a product might be analogous to catching the flu, as if the purchaser had no agency in whether to spend their money on a given product. (This issue will be discussed in further detail in Chapter 5.) This kind of language and analogy is also related (and may have led to) the problematic 'influentials hypothesis' (Watts, 2007; Watts, 2011, Chapter 4) whereby it is hypothesised that a few special people get others to adopt whatever they tell them (e.g. Keller and Berry, 2003), and the 'others' seem to have no say, no agency (at least no consideration is given to the latter). Continuing to employ this kind of vague language can also reinforce the problematic 'influentials hypothesis' mindset and the contagious disease paradigm.

There are also several instances of even vaguer language than the 'spread' of 'ideas' in the contagion paradigm, talking about 'the spread of influence', often alongside 'the spread of information' (e.g. in Bonchi, 2011; Goyal et al., 2010; Kempe et al., 2003; Lerman and Ghosh, 2010; Leskovec et al., 2007). While 'information' has some specificity, as it is an item of interest, the belief in which, or awareness of which, might become widely observed, due to some causes, 'influence' is not a focal item that can 'spread', i.e. that can be adopted, believed, endorsed, agreed with, known, etc. Rather, influence (from some source or sources) might be *why* a certain focal item, such as information, news, an idea, opinion, idea, belief, behaviour, action, product, has become widely known or espoused/adopted/believed/purchased. Influence is the causal effect, or role, some factor may have in bringing about the adoption (or similar), or the awareness, of some 'thing' (item) - it is not the thing itself.

Overall, the use of ill-suited definitions and of vague language ('spread of ideas', 'spread of influence') is an important limitation of the contagion-based paradigm, as it reinforces the use of problematic concepts and models for social influence, and is also connected to the three limitations in the following sections. Indeed, the use of this language from the contagion paradigm, imported from epidemiological models for the spread of actual diseases, has become so widespread and ingrained in research about online social influence, and so tied to the concept of social influence, that, for example, Goyal et al. (2010, p. 1-2) go as far as to suggest that the very concept of influence should be retrofitted to popular models of contagion, by claiming that 'any models proposed for influence should be compatible with the assumptions made in applications such a[s] viral marketing', where these assumptions are of contagious spread of information, actions, behaviours and beliefs due to social influence, as per threshold models of contagion.

3.1.2 Assumed cause of outcomes

The contagion-based paradigm for online social influence assumes that the only possible cause of observed outcomes (e.g. retweets or mentions, as noted in Tufekci, 2014) is social influence from someone in one's immediate or extended social network, on the given social media platform. As discussed, interpreting response as social influence is an assumption that is widely used in theoretical and empirical studies of online social networks such as Twitter, Digg and Flickr (e.g. Anagnostopoulos et al., 2008; Bakshy et al., 2011; Barbieri et al., 2013; Cappelletti and Sastry, 2012; Cha et al., 2010; Dubois and Gaffney, 2014; Ghosh and Lerman, 2010b; González-Bailón et al., 2011; Goyal et al., 2010; Kwak et al., 2010). Here, given a user *i* and a user *j*, if user *i* follows *j* and mentions the same entity as *j* (e.g. a URL or a hashtag) within a narrow time window, or if *i* re-shares or up-votes *j*'s post, or chooses to follow *j*, or mentions *j*'s username (Cha et al., 2010; Dubois and Gaffney, 2014; Ghosh and Lerman, 2010b), then *i*'s action is assumed to be due to social influence from user *j*. Indeed, one may say that such types of online activity or response represent the levels of attention or interest that a given piece of content has generated (Ackland, 2013; Watts, 2007). However, beyond indicating some degree of attention, it is far from straightforward to infer that *j*'s response is indeed attributable to social influence from user *i*, and is not attributable to any other cause.

Still, in the contagion-based literature, a person's social influence has come to be thought of as synonymous with the volume of responses their posts or messages have received (e.g. Barbieri et al., 2013; Bakshy et al., 2011; Cappelletti and Sastry, 2012; Cha et al., 2010; Ghosh and Lerman, 2010b).

Social influence on the Web has come to be synonymous with contagious information propagation online, e.g. on social media, over emails, in blogs (e.g. the claim in Goyal et al., 2010, mentioned above, that the concept of influence should be modelled in terms of the contagious threshold models used in the domain of viral marketing). That is, such outcomes are automatically interpreted as evidence, or symptoms of only one cause: social influence, from another user on the given social network platform. As also noted in the previous section, observing such outcomes is essentially equated with this specific cause, i.e. there is a conflation of such outcomes with one specific cause (social influence from a given user), which may or may not have been behind this outcome. This claim is done by assumption, without having proven that indeed this is the cause of the observed outcome, and that the outcome is not attributable to any other cause(s), such as the interests of the responder, or external circumstances (e.g. a current news item or a trend). Indeed, as discussed in Chapter 2.5.2, an important point raised in Tufekci (2014) is that 'whether there is a straightforward relation between information exposure and the rate of 'influence', as there often is for exposure to a disease agent and the rate of infection, is something that should be empirically investigated, not assumed'.

Let us now consider some examples of this practice. In Cha et al. (2010), which studies Twitter, it is assumed that 'the number of retweets containing ones name, indicates *the ability of that user* to *generate* content with pass-along value' (emphasis added), so when observing a retweet of someone's tweet, this outcome is automatically attributed solely to some *ability* of behalf of the user. Similarly, when considering mentions of a user's Twitter name as an outcome, that outcome is also assumed to indicate '*the ability of that user* to engage others in a conversation', so again the cause of this outcome is assumed to be the user referenced in the mention. In addition, in Kwak et al. (2010), in order to identify the top 'influentials' on Twitter, the authors rank users by number of followers, and by number of retweets, i.e. they assume that outcomes such as pressing the 'Follow' button or retweeting are solely due to the user followed or retweeted. In Ghosh and Lerman (2010b), studying the Digg social platform, it is stated that 'We empirically define influence as the number of in-network votes a users post generates', where 'in-network' means from the user's social network ties, and then all such votes are aggregated for all posts made by each user considered, and this number is assumed to indicate that user's influence.

In addition, in Goyal et al. (2010), without reference to a particular social network platform, influence is defined as contagious propagation, whereby a person can only be influenced by their (immediate, in this case) social network ties, as per the threshold models of contagion described in Chapter 2.4.2. That is, in such a model, if a person is observed to take a given action, it is assumed that this outcome must be due to one or more people in their immediate social network who had already taken that action. Barbieri et al. (2013) make an even larger logical jump, by assuming that whenever a group is observed to act on the same items on social media, that then these outcomes (actions) must be because those people are socially connected and influence each other along those presumed but unobserved ties systematically. That is, not only is it assumed that these people influenced each other to act on those specific items, but a much larger assumption is made, that, based on these observations, these people must influence each other in general, and there must exist social ties among them. In their own words: 'if we see a group of users acting on item i in a short time frame, and we observe this occurring on various different items, then we can infer that these users are connected in some social network, that they communicate and can influence each other'. This is assumed even though there is no evidence that these people know each other, or that they talk to each other, and there is no evidence of social ties. And it is assumed that the observed actions (as well as any future actions) are due to social influence from each other, and cannot be due to any other causal factors, such as these people happening to have common interests even if they do not know each other, or shared external circumstances (e.g. the items acted upon may be related to a series of current news events, like a current election cycle, or unfolding news coverage of a natural disaster).

Similarly, in Leskovec et al. (2006), a study based not only on online communications data (emails of product recommendations and discounts) but also on data of actual purchases, it is assumed that, if a person has received an email recommending a given product, and that person later purchases that product, the only cause is that email recommendation (without accounting for, for example, any pre-existing interests or preferences that person might have had in products of that type, which might have been part of the cause). In such a case, according to Leskovec et al. (2006), the purchase is attributed wholly onto the email recommendation, which is then considered a successful recommendation, even though other causal factors such as the purchaser's interests have not been measured or accounted for, in order to disentangle the influence of the recommendation from the influence of other factors.

As a last example, Bonchi (2011) assumes that one can safely ignore causal factors other than social influence when attempting to analyse social influence, i.e. that it is possible to analyse and measure social influence accurately without considering and accounting for alternative causes: 'We do not even discuss further how to distinguish between social influence, homophily and other factors, although we agree that it is an interesting research problem. Instead, we prefer to take an algorithmic and data mining perspective, focussing on available data and on developing learning frameworks for social influence analysis'. That is, it is assumed here that causes other than social influence are not part what they call 'an algorithmic and data mining perspective', and that one can analysing social influence adequately without considering any other causal factors.

As noted in Chapter 2.4.2), the justification for the assumption that the only influencing factor of observed outcomes is people in one's social network, stated in (Easley and Kleinberg, 2010, p. 565) as 'individuals make decisions based on the choices of their neighbours', is presented in (Easley and Kleinberg, 2010, p. 564), as follows: ' We saw in fact that there are two distinct kinds of reasons why imitating the behavior of others can be beneficial: informational effects, based on the fact that the choices made by others can provide indirect information about what they know; and direct-benefit effects, in which there are direct payoffs from copying the decisions of others for example, payoffs that arise from using compatible technologies instead of incompatible ones'. In the payoffs used in the models presented subsequently in Easley and Kleinberg, 2010, p. 566, it is assumed that for any pair of neighbours, if they do not adopt the same behaviour then each node gets a 0 payoff if, while if they adopt the same behaviour they each get a positive payoff.

However, there are problems with each of the two reasons presented above as justification. The first reason assumes that every decision that individuals make is about information (i.e. that the object of the decision, or the focal item for short, is information), or operates like information. However, many decisions that people make, and indeed that social influence has been claimed to be their cause, are about other focal items, like beliefs, ideas, opinions, emotions, which also tie in with one's subjective experience, personality, identity, feelings, psychology, and other factors. In such cases, information from others may be useless to an individual who has different personal tastes, beliefs, personality, or who is under different external circumstances (e.g. financial, geographical), and the individual would have no reason to imitate others. Or, indeed, an individual might want to do the opposite from what they observe others do in a given context, rather than imitating them, if their goal is to do something new and original, or to stand out (e.g. Mason et al., 2007). Then, assuming that such cases operate like information maximisation, as suggested, would be too simplistic.

The second reason is about focal items that have so-called *network effects* or *network externalities*, whereby one gains some benefit (or *utility*) not only from adopting the given item and using it (e.g. buying and using a product), but also one gains additional benefit the more others adopt and use the same item. The classic examples for network effects concern telephony as the focal item, or an online social network platform as the focal item, where the more the people who start using the focal item (who own a telephone or join a social network platform), the more the average user benefits from having adopted and from using the focal item. More generally, technological products are often used as examples of focal items with externalities (including in Kleinberg, 1999, p. 563-566), where users may benefit more from adopting a techonology compatible with the technologies their friends, colleagues use, rather than adopting an incompatible one . However, network effects do not universally apply to every focal item. They are only relevant for a specific sub-class of focal items, when there are many products, ideas, beliefs, opinions, and other items that do not exhibit network effects, therefore this second reason for the contagion assumption does not apply. Indeed, as above, for some cases where an individual may value uniqueness and non-conformism, this might outweigh any benefits from network effects, or there may even be negative network effects, where the more people adopt something, the less the given individual may want to adopt it (Mason et al., 2007).

Even though, in Easley and Kleinberg (2010, p. 565), some passing allusion is made to other factors, like the extrnal and social environment, and the perceived risk of a behaviour, that may affect the adoption and diffusion of innovations (in a discussion about Rogers, 2003), ultimately those considerations are not part of contagion-based models for influence. For example, in Easley and Kleinberg (2010, p. 579) it is assumed that the case of a‘risky’ behaviour that may not be immediately adopted will be adequately reflected in that one’s immediate social network neighbours will have likely mostly not adopted it. However, any personal views that may differ from those of one’s neighbours and that may shape an individual’s perception of what is risky or what is not, any wider societal attitudes towards a focal behaviour, beyond the immediate neighbours, and any inherent traits of the focal behaviour itself are not addressed.

The individual who is making the decision about whether to adopt a given focal item, and who is on the receiving end of social influence from his/her neighbours (immediate social ties), as this model assumes, is only assigned a numerical threshold to determine whether or not he/she will be ‘infected’ or ‘activated’ by the focal item, which is a computed as a function of the number (or proportion) of his/her neighbours that have already been ‘infected’. As an example of what a threshold means in threshold models, it is stated in Easley and Kleinberg (2010) that ‘A threshold of k means, ‘I will show up for the protest if I am sure that at least k people in total (including myself) will show up’. Again, there is no regard for that individual’s beliefs (e.g. do they view going to a protest as worthwhile, are their beliefs aligned well with the aims of the protest; regardless of what their social ties believe), the features of the protest (e.g. its aims), external circumstances (e.g. how safe it may be to protest in the given political climate). There is no discussion as to what the value of this threshold represents in the real world, it is not discussed whether it is supposed to somehow reflect the individual’s views, beliefs and other personal traits, or their external circumstances.

It is well known in the social sciences literature, and also often recognised in the computer science literature, that a person is not influenced solely by their immediate social ties, but by a range of factors (e.g. Allahbakhsh et al., 2013; Aral et al., 2009; Bakshy et al., 2011; Kelman, 1961; Mason et al., 2007; Sperber, 1996; Watts, 2007). And indeed, contrary to the common asumption in empirical social media studies using contagious models, one is not only influenced by their social ties *on that particular social platform*; a social media platform only represents a thin slice of one’s sources of influence (e.g. Tufekci, 2014). For instance, it is worth quoting here from Sperber (1996, p. 106), where the author, a social and cognitive scientist, presents a vivid criticism of the assumption that people are influenced solely by their social ties, using as an example how one might form an opinion about Bill Clinton:

> ‘However, it is unlikely that you formed your own views simply by copying, or by averaging other people’s views. Rather, you used your own background knowledge and preferences to put into perspective information you were given about Clinton, and to arrive by a mixture of affective reactions and inferences at your present view. The fact that your views are similar to many other people’s may be explained not at all by a copying process, and only partly

> by an influence process; it may crucially involve the convergence of your affective and cognitive processes with those of many people towards some psychologically attractive type of views in the vast range of possible views on Clinton'. (Sperber, 1996, p. 106)

This assumption that one is only influenced by their social media ties, or by the online communications on a given platform, in contagion-based studies of online social influence, is an also an instance of the *post hoc ergo propter hoc* fallacy ('after this, therefore because of this'), which asserts that if an event B happened after an event A, it must be that event B happened because of event A (Watts, 2007). As discussed in Chapter 2.5, the problem with this fallacy is noted in Watts (2007), in the context of using anecdotal evidence to support the influentials hypothesis, particularly in Gladwell (2002), a best-selling popular science book. However, this fallacy is very common not only in anecdotal stories used in the news or in popular science books, but also in scientific studies of online influence as contagion.

One illustration of the extent to which this assumption can lead to misestimation of the magnitude of online social influence in shaping observed outcomes on social media is presented in Myers et al. (2012). This work simulated a process which picked Twitter users purely at random and marked them as having posted a given piece of content. After this process had marked 10% of the users, about 30% of posts appeared, falsely, to be a result of social influence: the random process picked a user that had, simply by chance, a Twitter connection who had already posted the same content. Of course, it is not argued in this thesis that people's decisions to post things on social media is governed purely by chance. Rather, there must be a reason why they decide to post about something, there must be a cause behind it, e.g. that they personally have always found that topic interesting, or maybe someone or some event outside the social media platform sparked their interest in it. Still, this study is useful as an illustration of how causes external to the given social network, causes other than influence from one's Twitter followees, can be misattributed to influence from one's Twitter neighbours under the contagion-based paradigm, to the extent that a large misestimation of the extent of such influence may be made. Similarly, Aral et al. (2009) empirically found a very large extent of misestimation of social influence online, due to ignoring the effects of personal similarity among people's interests, hence confounding the social influence of the online communications of others with the influence of personal similarity. (This assumption of the contagion paradigm is also tested in this thesis, in Chapter 5 for individual decisions, and in Chapters 6 and 7 for collective-level decisions.)

For example, on Twitter, when one observes that a person retweets (reshares) a given tweet, what may the cause be? It may be any of the following causes, or a mix of them:

- The tweet's author. This is what is usually assumed in empirical studies of social influence (e.g. Cha et al., 2010; Dubois and Gaffney, 2014), without testing to rule out other causes, such as the below.

- The tweet itself. For example, did the tweet present a very persuasive argument? Or a very emotive argument? This may be the cause of adoption, regardless of who posted the tweet, or of how interested the adopter is in that specific topic. For instance, in the offline epxeriments presented in Kelman (1961), the source of influence is a message from a stranger, not from a friend or social tie, and adoption or not of the focal view is based on the persuasiveness of argument itself. So, social influence does not have to come only from one's immediate social ties (neighbours), or even distant social ties, as is assumed in the threshold models of contagion.

- The specific topic of the tweet. It has been found in the social psychology, management, and marketing literature (Berger, 2013; Kilduff et al., 2010) that topics that invoke emotional arousal, specifically *activating* emotions such as excitement or anger, have been found to increase the chances that the viewer will then reshare the given message. For instance, cute videos or memes of pets that invoke excitement and joy.

- The retweeter's interests. It may be that regardless of the person who posted the tweet, the phrasing of the tweet, or the exact topic of the tweet, the retweeter has an interest in this area, and that is why they retweeted the message. E.g. the retweeter may in general like to retweet funny pet videos, or the retweeter may be maintaining an account that retweets news or political tweets in general.

- External circumstances. For instance, it might be that the tweet is about a very current and important news item, such as a political event during an election cycle, or a flood or earthquake, and people reshare all relevant tweets as they come, in an effort to help ensure people are informed with the latest developments, even though they do not in general have an interest in e.g. earthquakes, and not because they particularly liked the particular phrasing of the given tweet.

A similar rationale applies to other social media platforms, and to other modes of text-based online communication, and also to other kinds of outcomes, such as comments, 'likes', and mentions. These additional causes (i.e. additional to the social influence from specific online communications) are discussed further in Chapter 5.

Therefore, given that the assumptions of the contagion based paradigm are problematic, if one still wants to establish whether an observed outcome is due to social influence from specific online communications, one should also consider other causes, and establish the extent to which the influence of the latter can be disentangled from the influence of the former, based on the available data, using appropriate causal methods (as described in Chapter 2.6). Without this process, claims that observed outcomes such as retweets are manifestations of influence from the use who posted the tweets (e.g. as in Cha et al., 2010; Dubois and Gaffney, 2014), or claims that actions upon an item i are due to invisible social influence among the users who took those actions (e.g. as in Goyal et al., 2010) are unsubstantiated.

This problematic post hoc ergo propter hoc practice of claiming that observed outcomes must be due to one's social network ties is not unique to the study of social influence *online*, but is also common in studies of offline settings. In a criticism of the use of the term 'influence' in the work of Christakis and Fowler (Christakis and Fowler, 2009) Thomas (2013) says (with 'CF' referring to Christakis and Fowler):

> 'Here I fear that the (literally) hypothetical use of the word 'influence' has manifested its own undue effect. For example, in their monograph [2], CF immediately jump to the term 'three degrees of influence' when describing the extent of network autocorrelation, whether we may also call it clustering, association, correlation, or peer effects. Whereas I understand that a choice of wording is 'evocative', this basis is needlessly confusing. The word 'influence' is a well-established proxy for a directional causal effect in every scientific and sociological context in which it appears (including every citation in the CF discussion paper containing the word 'influence' in the title.' (Thomas, 2013)

3.1.3 Assumed meaning of outcomes

As discussed in Section 3.1.1, the contagion-based paradigm talks about 'the spread of ideas', but what 'spread' usually means is'adoption'. And when saying that someone is an 'influential', that means that they are exceptionally effective in getting others to adopt what they propose (as per Cha et al., 2010 in reference to Katz and Lazarsfeld, 1955).

Therefore, in order to study the social influence of online social communications on outcomes of interest, or the social influence particular people exert for outcomes of interest, the outcomes must indicate adoption (or non-adoption). And in order to claim that someone has been influenced (by some person or event) based on an observed action of theirs, that action must non-ambiguously indicate adoption or agreement (with that person or event). However, the online outcomes commonly studied (e.g. retweets and reshares, mentions of a username or of a hashtag, URL or other keyword, comments, 'likes') do not necessarily always indicate adoption or agreement. Still, such ambiguous outcomes are systematically assumed under the contagion-based paradigm to mean agreement with a social media user, or adoption of a given online message.

For example, in González-Bailón et al. (2011), the goal is to study recruitment into a political movement on Twitter. There, if a user uses a protest-related hashtag, it is assumed that this means that the user has been recruited into the political movement and protest, i.e. that they have joined it. However, using this hashtag merely means that a person is talking about something relevant to the protest (if it is not a spam tweet), and the tweet could be expressing neutral sentiment, or even complaints about or disagreement with the protest. The content of the tweet is not examined to indicate if there are any traces of endorsement (and even then, this would not suffice to indicate the person has physically joined the protest, or is fully accepting all of the ideology of the political movement wholesale). Instead, an appropriate outcome would need to indicate adoption of the political movement's core beliefs into this person's belief system, and it would help to qualify the extent and intensity of that adoption, and whether it stays strong or fades over time. One possible way to establish might be with an appropriately designed repeated survey. But using a hashtag on Twitter is not a meaningful outcome, and it is insufficient to deduce adoption (or any other disposition, beyond some degree of attention, as noted in Ackland, 2013; Freelon, 2014; Watts, 2007), as, per Tufekci (2014, p. 6),'the same act can have multiple, even contradictory meanings'.

Even retweeting, which has been very commonly been used in the contagion paradigm as an indicator of influence, does not necessarily mean agreement and adoption, i.e. successful influence from some unspecified cause(s). It can mean disagreement, mocking, passively reporting an event, any number of things, as discussed, with real life examples, in Tufekci (2014) and in Freelon (2014). At best, it means that the retweeter considered this tweet worthy of resharing, *for some unknown reason.* Similarly, even 'likes' and 'favourites' do not necessarily mean agreement with the message posted: rather, a user may take that action just out of politeness, to reciprocate for likes received in the past from that user, to indicate they have seen a post to the person who posted it, ironically or to poke fun at the post, or for yet other reasons. For instance, Meier et al. (2014) conduct a survey of user motivations for using the 'Favourite' button on Twitter, and find that motivations are very heterogeneous, and not always consistent within and between users. They come up with a taxonomy of 25 motivations, including re-finding a tweet, liking the content of a tweet, or wishing for a more private conversation (than what a retweet would allow), among others.

Therefore, when using the contagion paradigm and such observed outcomes from online social media, beyond indicating some degree of attention paid to that particular post (or user, in the case of a mention or follow action), it is far from straightforward to infer their *meaning* (let alone their source of causation), and indeed the empirical studies by Anagnostopoulos et al. (2008) and Bakshy et al. (2011) recognize that this contagion-based approach yields an overestimate of social influence.

This practice of assuming that an action that references someone or something means agreement with that person or thing has pre-dated social media platforms. As discussed in Chapter 2.4.2, with the proliferation of the Web, the need arose for search engines to automatically find and rank the most relevant webpages for a user's query. One successful idea, based on information science, library science, and specifically citation analysis (e.g. Newman, 2014), was to use the hyperlinks that pointed to a given webpage to help determine its relevance to a given query topic. In Kleinberg's work on the HITS algorithm for ranking webpages using the hyperlinks that reference them (Kleinberg, 1999), it is claimed that webpages that have a lot of hyperlinks referencing them must be authorities in their field, and hence should be ranked highly in search results about a relevant topic. The rationale given is that 'Hyperlinks encode a considerable amount of latent human judgment, and we claim that this type of judgment is precisely what is needed to formulate a notion of authority'. This logic has since been extended to the concept of social influence. In the contagion-based paradigm of the online influence literature, instead hyperlinks the links studied are friendship and/or interaction links (e.g. reshares, comments, 'likes', mentions, of a user or a keyword), and instead of interpreting the number and properties of those links as indicators of authority, they are interpreted as indicators of social influence. However, as discussed above, relevance with something does not mean influence from that thing. Just as academic papers may cite other papers not because they consider them authorities, but because they disagree with them, e.g. in order to criticise them, or for any number of other reasons, webpages may also cite other webpages which they do not consider authorities, and people communicating online may reference other people or entities for any number of reasons, not necessary because they agree with them or endorse them.

Therefore, the observed actions that are easily measured and quantified when using social media data do not generally warrant interpretation as agreement or adoption, and hence cannot be used as meaningful outcomes to substantiate claims on social influence (as also noted in e.g. Alshamsi et al., 2015; Tufekci, 2014; Olteanu et al., 2016). For instance, Watts (2011, p. 105) recognizes this, and comments on one of the studies he worked on:

> 'influence - the number of retweets - was the wrong measure. We measured retweets because thats what we could measure, and that was definitely better than nothing. But presumably what you really care about is how many people click through to a story, or donate money to a charitable cause, or buy your product. [...] In the end, we simply don't know who is influential or what influencers, however defined, can accomplish. Until it is possible to measure influence with respect to some outcome that we actually care about, and until someone runs the real-world experiments that can measure the influence of different individuals, every result - including ours - ought to be taken with a grain of salt'. (Watts, 2011, p. 105)

While social media actions such as the above may be useful for making claims on the spread of information (or, more precisely, of the awareness of information), the spread of information is not the same as the spread of (the adoption of) focal items other than information, such as actions, attitudes, behaviours, beliefs, or opinions, due to social influence. When someone shares, re-shares, 'likes', comments, or responds to some information, such indications of attention, or interest may be adequate, or at least useful, in analysing how information spreads (in terms of information awareness), even though they are not appropriate outcomes for influence, as it is not guaranteed that they denote endorsement or adoption. This is also recognised in Centola and Macy (2007, p. 27), where they discuss their own results as well as those from other studies that 'show that it can be very dangerous to generalize from the spread of information and disease to whatever is to be diffused. Network topologies that make it easy for everyone to know about something do not necessarily make it likely that people will change their behavior. '

Overall, in order for one to substantiate claims on influence, outcomes are needed that indicate whether influence (from some source) has occurred, i.e. whether adoption or agreement has occurred (due to that source). The social media actions that are readily measurable and have been used as outcomes in the contagion-based paradigm do not satisfy this requirement. Rather, one must take care to use meaningful outcomes, that do unambiguously indicate adoption or endorsement, as appropriate, for the focal item (e.g. idea, behaviour, product) studied.

3.1.4 Untested assumptions

The assumptions made in the contagion-based paradigm (discussed above in Sections 3.1.2 and 3.1.3), often cloaked in vague or ill-suited language (as per Section 3.1.1) are often stated as self-evident, as facts of nature, when in reality they are assumptions, the validity of which should be tested empirically or otherwise (e.g. theoretically, drawing upon domain expertise or findings), as noted in e.g. Tufekci (2014), and discussed in Chapter 2.5.2, in order to substantiate them as claims.

That is, one cannot define or assume that a user being well-connected (e.g. high centrality value) in social network, or user's posts getting a lot of response, or a topic (e.g. hashtag, URL, other keyword) getting a lot of response and participation is the same as that user or topic being 'influential'. That is because the latter means that an online message (or by extension its creator) is successful in obtaining agreement and adoption from others, and those adoptions are due to it (or by extension its creator) more than due to any other causes, i.e. this online message is the strongest cause of adoption outcomes. In order to substantiate such claims that someone or something is an 'influential' in general, or is the source of influence in a more specific context (with respect to a specific focal item), one must prove, rather than assume, that this person or thing is the strongest cause of the observed outcomes, which requires measuring, or at least considering in some way, other relevant causes (per Section 3.1.2), and also requires that the observed outcomes unambiguously denote adoption of, or endorsement of, or agreement with the focal item (per Section 3.1.3).

Indeed (as mentioned previously), in the context of analysing the spread of information online, Lerman (2016) discusses studies that found the contagion-based paradigm does not hold empirically, as information was found to spread empirically less broadly than what would follow from contagion-based models.

For example, in the context of analysing online social influence, the assumption that being 'well-connected' on a given social media platform makes that user an 'influential' is made in Sun et al. (2009), and in Kempe et al. (2003), Kitsak et al. (2010), under the untested assumption that the adoption of a given item due to online social influence operates as a contagious virus along social ties based on threshold models of contagion, so the better connected one is, the more people will be 'infected' by his or her online messages. The assumption that social media posts (created by a given user, and/or related to a topic denoted by a hashtag or keyword or URL) getting a lot of response or engagement is the same as that user or topic being 'influential' is made in several studies (e.g. Cha et al., 2010; Dubois and Gaffney, 2014; Ghosh and Lerman, 2010b; Ghosh and Lerman, 2014; González-Bailón et al., 2011). Barbieri et al. (2013) make a mixture of both assumptions, assuming that observing posts about the same topic, by multiple users, for multiple topics, must imply that these users are socially connected, communicate, and influence each other.

The case of 'influentials' (or 'opinion leaders'), how they are defined, and how they are studied, is an important topic to discuss here. Cha et al., 2010 begin by citing a dictionary definition: 'The Merriam-Webster dictionary defines influence as "the power or capacity of causing an effect in indirect or intangible ways", which makes clear the causal nature of the concept. They then present the traditional literature on social influence, from which the 'theory of influentials' stems (in particular referring to Katz and Lazarsfeld, 1955; Rogers, 2003; and also referring to the popular science book Gladwell, 2002) state how the term 'influentials' refers to people who are 'exceptionally persuasive in spreading ideas to others'. So, an 'influential' is someone who is exceptionally capable in causing others to adopt what he/she is proposing (e.g. an idea, belief, opinion, or a product). Cha et al. (2010) note how the traditional view (i.e. the theory of influentials) assumes that these 'influentials' are 'a minority of members in society', who 'drive trends on behalf of the majority of ordinary people', and it is assumed that they 'possess qualities that make them exceptionally persuasive'. That is, in this paradigm, attention is only paid to the inherent qualities of these people, as causes for why they manage to persuade many others, and there is no consideration of whether the external circumstances in society may also be beneficial in helping people adopt the proposed item (as is countered e.g. in Watts, 2007). Cha et al. (2010) further note that such 'influentials' are 'loosely described as being informed, respected, and well connected'. Similarly, in Dubois and Gaffney (2014), in discussing Katz and Lazarsfeld (1955), state that in that work, 'four core facets of influence are suggested: *having a following, seen* *as an expert, knowledgeable/ have expertise*, and [being] in a position within their local community to exert social pressure and social support/ *social embeddedness*'.

It is argued here that, while the above 'four core facets' might indeed be the *causes* why one is persuasive, this needs to be tested and proven, rather than assumed. It is not enough for these four factors to be generally associated or correlated with one's persuasiveness (if indeed such an association exists), as such an association may be due to causal pathways other than direct causation (i.e. the 'correlation does not imply causation' slogan). For instance, both persuasiveness and these four traits might be due to an unmeasured common cause, such as a person having been born in an exceptionally wealthy family, or being exceptionally well educated.

More than this, based on this conceptualisation of an 'influential', one may not conclude that 'well connectedness' alone (one of the four traits associated with persuasiveness/ influence) can be equated with being an influential. This is only one of the four traits that are associated with being an influential; it is not the defining trait of an influential (which is persuasiveness, or capability to non-coercively cause others to adopt what he/she is proposing). However, this is what has been commonly done in the contagion-based literature (in the contagion tradition, upon which 'influence maximisation' is also based, where it is assumed the only cause and mechanism of adoption is imitation of one's social ties, e.g. Domingos and Richardson, 2001; Easley and Kleinberg, 2010; Kempe et al., 2003; Kitsak et al., 2010; Sun et al., 2009). Indeed, being well-connected is easily measurable from social media network data, because of the networked nature of those platforms and resulting data, whereas other qualities are not (e.g. the other traits: expertise, knowledgeability, not to mention traits outside this individual: traits of the individuals in his/here 'audience', traits of the focal item itself, external circumstances in society and the current culture.) As noted in Marres (2017), it is common to see methodology used just because it is convenient, but as pointed out in Tufekci (2014) and Watts (2011), such methodology is not sufficient to substantiate claims on social influence.

Therefore, if one wants to claim that being 'well connected' means being 'influential', one needs to investigate whether the former is indeed is a strong causal factor of the latter, rather than conflating the two. To investigate this, one needs observed outcomes that unambiguously denote agreement or adoption. And to establish if well-connectedness is indeed the only cause behind such outcomes, or at least the strongest cause, or a relatively strong causal factor, one must also account for and measure other relevant causal factors (e.g. perceived quality of arguments and perceived knowledgability of that individual, as well as factors outside this 'influencer' candidate: external factors, personal traits of audience members persuaded and not persuaded, traits of the focal item, etc.). Therefore, establishing that 'well connected' people are 'influentials', or establishing that people do indeed imitate their immediate social ties regardless of other circumstances (as the contagion paradigm assumes untestedly), or establishing that any person (or group, or event) is the source of influence in a given setting for a given focal item, is a much more complex problem, requiring more robust data (unambiguous outcomes, and measurements of other causal factors) and methods (not assumed contagion, but causally modelling and accounting for other causes). And other causes will generally lie outside a given social media platform too, i.e. studying only one platform means only considering one thin slice of each user's reality (Marres, 2017; Tufekci, 2014), as discussed in Chapter 2.5.2

Begging the Question Fallacy This kind of logic, of assuming as a given that which should be proven, is an example of the *begging the question*, or *circular reasoning* logical fallacy. This is an informal fallacy, whereby one fails to prove anything other than what is already assumed. This is frequently done in the contagion-based paradigm, where studies typically *assume* one or both of the below:

1. In an empirical study of online social influence, an observed outcome (an action or response such as a reshare, comment, mention) means adoption, and the source of causation must be the person or item referenced: an observed response to a post X means adoption of X and that influence is from (the cause is) the person who posted X; referring to an item (denoting some topic, and represented e.g. as a hashtag, URL, other keyword) T means adoption, endorsement or agreement with T and that influence is from (the cause of this outcome is) the person (or group) who posted T earlier (possibly within a time window) from the immediate (or not) social network. This is as per the contagion paradigm: focal items (behaviours, views, products, etc.) spread like diseases and infect people (adoptions) along social network ties.

2. In a theoretical study of online social influence, people will imitate (imitation, i.e. adoption, is assumed as the only possible action, or outcome) what most of their immediate neighbours do, as per threshold models of contagion. So social ties are all that matters, and social network topology measures (e.g. centralities) are indicator's of one's social influence capacity.

These assumptions are made, but it is *not proven* that these assumptions would result in accurate measurements of the social influence of online communications. Rather, what is done is an untested acceptance of these assumptions, and a description of what these assumptions imply in a given social network, theoretical or empirical.

For example, many studies use and apply just the second assumption: Barbieri et al. (2013) apply it to the extreme, assuming when observing many users mentioning the same topic, for several topics, that this warrants the inferences that these people must communicate, there must be social ties between them, and that they must influence each other along those inferred ties. Kitsak et al. (2010) use the known social networks in various online and offline settings, use a threshold model of contagion to simulate adoptions (without testing whether adoptions did actually happen in this way in each network), and assuming each simulated adoption is due to the seed 'spreader' who started it, test different types of centralities and topology measures of 'influence' to see which one best correlates with the most 'influential' 'spreaders' of the simulation. Similarly, Domingos and Richardson (2001) and Kempe et al. (2003) also assume that the adoption, acceptance or uptake of information, actions, behaviours, attitudes or opinions spreads due to social influence as per threshold models of contagion, regardless of other causes (second assumption above), and the Kempe et al. (2003) present efficient approximation methods for finding 'influentials' under this assumption, with no consideration of whether this assumption does hold in practice.

Other studies that draw on empirical data of social network interactions employ the first assumption, and assume that actions like retweets and mentions mean adoption, and that they are due to the person whose tweet is retweeted (or whose name is mentioned) and not due to any other causes (e.g. Cha et al., 2010; Dubois and Gaffney, 2014; Ghosh and Lerman, 2010b). Of course, as discussed, outcomes like retweets and mentions do not necessarily imply endorsement or adoption (e.g. Freelon, 2014). But even if an investigator can somehow establish that, in a given context, a social media outcome does warrant interpretation as endorsement, agreement, adoption (e.g. perhaps with outcomes that may have more positive connotations than retweets, like 'likes' or upvotes of a post), again one does not know if that outcome is due to the user who posted the online communications under study.

Finally, other studies examine the correlations between the implications of the two assumptions. That is, they correlate the theoretical claims of the second assumption, which are based on threshold models of contagion, with empirical interaction data from social media interpreted under the first assumption above. For instance, they use different types of social network topology measures (e.g. different centrality metrics) to rank users (second assumption) and then see if the top-ranked users by each measure do indeed receive the most responses in their empirical data (first assumption), often for different types of responses (e.g. mentions and retweets on Twitter, upvotes on Digg), for instance in Cha et al. (2010), Dubois and Gaffney (2014), and Ghosh and Lerman (2014).

Again, these assumptions are made, but it is *not proven* that these assumptions and their implications amount to social influence. Rather, what is done is a description of these assumptions, of their implications, and of their correlations, in a given social network, either theoretically or with empirical interaction data.

Overall, as has already been called for in the literature (Bright et al., 2017; Freelon, 2014; Tufekci, 2014; Watts, 2011), in analysing human behaviour in online social settings, and in the analysis of social influence in particular, assumptions of the kind discussed above should be empirically tested, not presented as if they were self-evident facts.

3.1.5 Summary

In summary, the above four classes of problems of the contagion-based paradigm highlight the key limitations of this paradigm for making claims on social influence online. It has shown the problems of applying the analogy and language of the contagious spread of diseases to the problem of measuring the social influence of online communications, highlighted the unfoundedness of assumptions on the causes and on the meaning of easily measured outcomes, and discussed how these assumptions have been taken as self-evident facts, rather than being empirically tested.

Overall, the critique in this section has demonstrated how the methods and concepts (contagion-based models) and the data (ambiguous outcomes, consideration of only one out of many causes) used in the contagion-based paradigm are insufficient to make claims on the social influence of online communications.

Section 3.2 will next demonstrate how these limitations manifest if one attempts to empirically apply the concepts and methods of the contagion-based paradigm to a real-world setting employing of Web-mediated communications, this time outside social media platforms. Following that, Chapters 4 to 7 will present a causal approach that can successfully address these limitations and allow investigators make more robust claims on the social influence of online communications.

3.2 Empirical critique through a real-world worked example

Given the key limitations of the contagion-based concepts and methods for social influence on the Web, described in the analytical critique, this section helps empirically demonstrate how much the established methods can and cannot tell us about online social influence in practice, through an empirical critique based on a worked example, using a real-world dataset of Web mediated interactions.

As discussed at the beginning of this chapter (and in Chapters 1.1, 2.6.2), to claim that *A* influenced *B* is to claim that *A* was one of the causes of *B*. Hence, claims about the social influence of online communications are causal claims. Therefore, this section aims to explore the question of whether the contagion paradigm allows one to make such claims, i.e. whether it allows one to go beyond associational, descriptive claims about patterns of online response, and how those might be associated with various characteristics of an interaction setting, towards causal claims about the role of the social influence of online communications on outcomes of interest.

Thus, this section attempts to empirically assess whether the contagion-based paradigm can offer any insights on the social influence of online communications. To that end, this paradigm's most common measures of online social influence (number and network properties of responses) are employed, using primarily the kinds of online data this paradigm usually considers (online communications data; in this thesis, email communications from the public W3C Provenance Working Group archives). This empirical worked example serves as an illustration (in addition to the existing social media studies discussed) of how far one can go in making claims on social influence, based on online communications data and on the (descriptive, non causal) methods of the contagion based paradigm alone. It is also useful as an illustration of the contrast of using this paradigm (with online response as the assumed outcome of adoption), versus using a causal paradigm on the same dataset (with actual, contextually meaningful outcomes of adoption) which will be presented in Chapters 6 and 7.

It is noted that no claims are made here that the findings from this dataset generalise to other W3C Working Groups, or to other collaborative efforts. Rather, the goal of these analyses is to provide an additional example of what kinds of claims can be made on the social influence of online communications under the contagion-based paradigm (and, in Chapters 6 and 7, how causal methods can be used to measure online social influence and to assess the presence of confounding in estimates of online social influence, from this setting of group interaction).

The online communications channel in the empirical setting of the W3C Provenance Working Group studied here is email. Empirical contagion-based studies have recently largely focused on settings where the online communications channel is social media, but there are also several, often slightly older, empirical contagion-based studies that used data from blogs or from emails (Lerman and Ghosh, 2010). For example Leskovec et al. (2007) aimed to study the influence of email product recommendations and discounts on subsequent product purchases and the properties of the chains of forwarded recommendations; Liben-Nowell and Kleinberg (2008) studied the properties of the forwarding chains of two email chain letter petitions in various mailing lists; Wu et al. (2004) studied the mail boxes of 40 employees of a US organisation, analysing how attachments and URLs spread via email forwarding chains. The goal for the proposals in this thesis is to be online channel-agnostic, i.e. independent of whether the online communications channel is email, or a specific social media platform, or some other mode of online communication. So, since the contagion paradigm focuses on the actions of sending (or posting) content and responding to content, and on the structure of the online social network and/or of the communications network (on social media, over email, or over other modes of online communication such as blogs), this section also focuses on the actions of sending and responding to content, and on the structure of the online interpersonal network (in email communications). That is, this section uses the same types of actions, methods, and metrics as contagion-based empirical studies commonly use.

This section begins by presenting the real-world online interaction dataset that will be analysed here, then it outlines the implementation and design of the analysis. Next the findings are presented and discussed, and the section concludes with a summary.

3.2.1 The dataset

As a case study, or a worked example, this section uses observational digital trace data from the public online archives of a World Wide Web Consortium (W3C) Working Group, specifically the Provenance Working Group. This is a setting of collective decision-making and collaboration, where domain experts in a topic area participate in order to produce a set of standardisation documents, in this case on the topic of Provenance on the Web.[3] This group produced four *Recommendation* documents, and eight *Notes* documents.

The advantages of using this dataset include the fact that it allows one to extend the analysis of online social influence, from social media or individual-level email recommendations and individual-level outcomes only, to other types of online interaction data, which are also ubiquitous (email communications in settings of collaboration), where individual as well as collective-level outcomes (decisions, actions, events) can be studied. There have been calls for researchers working on the analysis of digital data of social interactions to extend their focus beyond popular social media platforms, and particularly Twitter, which have recently been over-represented in studies of online human behaviour and which carry their own limitations in terms of the kinds of data they generate. For instance, a variety of such limitations are the focus of Tufekci (2014). Alshamsi et al. (2015) discuss how online outcomes, such as 'support' for political movements, has often failed to translate to lasting offline outcomes in the real world. In addition, Blank and Blank and Lutz (2017) discuss the limitations of online social media platforms, in terms of representativeness, which can limit the validity of one's inferences about real-world outcomes in the general population when using social media data. Indeed, the Economic and Social Research Council (ESRC) in the UK recently issued a call for research using 'New and Emerging Forms of Data', as part of their Big Data Network incentive. This call invited researchers to make use of data beyond social media data, such as data that is routinely collected by business, voluntary bodies, and other organisations, to inform evidence-based policies and make such organisations more effective, as well as to benefit wider society (ESRC, 2017). The W3C online archives are one example of such types of data.

[3] The email archive can be found at https://lists.w3.org/Archives/Public/public-prov-wg/.

Another advantage of this dataset it that it is publicly accessible, in contrast to data from social network platforms which are typically not publicly available to researchers outside the respective social network company, or only partly made available, as data is the competitive advantage and the revenue source for these companies (Barbieri et al., 2013; Marres, 2017, p. 17). Furthermore, this kind of collective collaboration data allows one to extend the analysis of online social influence beyond the node-to-node interaction perspective that focuses soleley on individual-level actions and interactions (with the related limitations of such approaches, e.g. Tufekci, 2014) to also encompass collective-level analyses of collaboratively produced outcomes (the latter will be presented in Chapters 6 and 7). Finally, an important advantage of using the W3C Provenance Working Group data is that the context and rationale of the group's efforts are published (e.g. in Moreau et al., 2015), and the public dataset contains outcomes that are justifiably meaningful and important in that context (these outcomes are used in Chapters 6 and 7, as they lie outside the domain of the email communications data). In contrast, the contagion-based paradigm typically assumes that responses in the domain of online social communications are an appropriate outcome signifying adoption, without proper justification and context (Freelon, 2014; Tufekci, 2014), often because of lack of access to data on the actual outcomes of interest (e.g. actual purchases of products, or actual adoptions of opinions or beliefs, rather than just online mentions of or responses to them, as noted in Watts, 2011, Chapter 4).

Some of the challenges that the W3C PROV email dataset exhibits, in terms of studying online node-to-node (person-to-person) email communications, are briefly noted below:

- Some of the In-Reply-To fields in the SMTP email headers, used for establishing which email is a reply to which other email, are wrong. This happens for around 10-15% of the emails, as the email client used may truncate or omit this information. Still, this only affects a small proportion of emails, and imperfections like this are common in real-world data.

- The heavy-tailed distribution of thread sizes (which will be discussed) means that too few threads contain enough messages, while the majority of threads have too few messages to conduct other kinds of statistical analysis, like analysing the shape of the cascade through curve fitting, in order to see whether the topology of communications follows a pattern that might tell us something about the nature of the interaction (under the contagion paradigm's notion of influence-as-response).

Overall, these limitations are not severe, so they do not significantly affect the usefulness of the dataset, for the analysis presented here.

The preprocessing steps that needed to be taken in order to prepare the data for analysis of influence patterns, as well as the tools used for data preparation, processing and analysis are presented in Appendix A.

3.2.2 Key findings

Social influence on the Web is typically conceptualised as response to a message, in empirical studies under the contagion paradigm, and studied in terms of the quantity and structural properties of response *chains* stemming from any given message (also known as response *cascades* or response *trees*), and in terms of the structural properties of the social network among participants (if captured in the data), and/or of the response network among participants.

Therefore, the analysis in this section first considers the size of response cascades, and next investigates the topological patterns in the network of responses between individuals. The goal is to answer the question: as the Working Group's interactions unfold, can the topological structure that emerges from the online communications between the participants tell us anything about the social influence of online communications?

In the W3C Provenance Working Group, email was the online communications channel that was used: members emailed the Provenance Working Group mailing list, and other members could then respond to those emails, forming email discussion threads. In this context, answering the above question using the concepts and methods of the contagion paradigm means answering questions like how frequent or rare is it for messages to receive a lot of responses, which people receive and send the most emails, which people receive the most responses ('influentials'), whether people engage equally with everyone else, or whether they engage more with some people than with others. And ultimately, answering the question of whether the answers to the above questions can tell us anything about the social influence of online communications, in the way that social influence has historically been understood in the social sciences (and in everyday parlance), as the causal effect of online communications on an outcome of interest (per the discussions and definitions in Chapters 1.1, 2.3, and 2.6.2, as stated at the start of this chapter).

That is, this analysis is based on the assumptions of the contagion-based paradigm, i.e. the assumptions that causes of outcomes other than the social influence of online communications can be ignored, and that outcomes (responses to a message) are attributable to the sender of the message, and that response can be safely assumed to represent an outcome of adoption, or at least of attention and engagement, i.e. that someone was 'influenced' by the sender to at least pay attention and take the time to respond. After performing the analyses with the methods and practices that follow from these assumptions and concepts, this section considers whether the results allow one to make claims about the social influence of online communications in terms of how the meaning of 'social influence' has historically been established in the social sciences.

3.2.2.1 Response cascades

Let us begin by exploring the properties of ***response cascades***, that is, of chains responses that stem from any given message. Visually, an example is shown in Figure 3.1. A cascade is essentially a ***tree structure***, where the root is the first email of the thread (email A), and from it stems the first wave of responses (email B, email C). Email D is a response to email C, so it stems from it. The ***size*** of a cascade is the number of responses in it. So, in Figure 3.1, the tree contains three responses, so the ***size*** of this cascade is three.

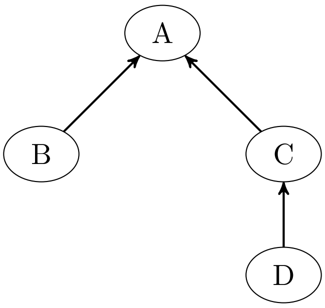


Figure 3.1: Example reply cascade

Cascade size, i.e. the number of responses, is one of the most common measures of online social influence in the contagion-based literature (e.g. Bakshy et al., 2011; Cha et al., 2010; Chen, 2006, and as noted in Ackland, 2013; Freelon, 2014; Tufekci, 2014), so this is an essential feature to study under this paradigm. To gain an understanding of the properties of cascade sizes, some descriptive statistics are calculated, shown in Table 3.1.

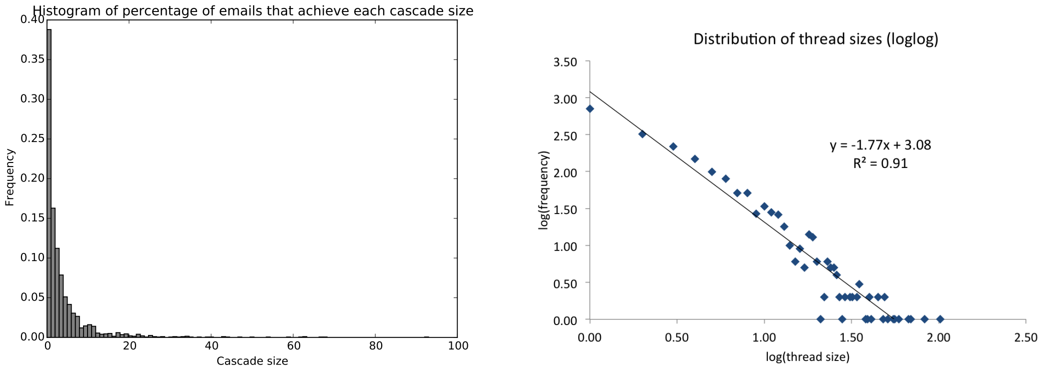
Table 3.1: Descriptive statistics for cascade sizes

n	1929
min	0
max	101
mean	3.57
median	1
variance	50.04
90th percentile	9
95th percentile	15

There are a total of 1,929 threads, with the minimum number of responses in any thread being 0, and the maximum 101. But how common are large threads, i.e. how common is it to have very many responses? The mean number of responses in a thread is 3.57, but as this is sensitive to outliers, the median is also calculated, which is 1. Therefore, at least half the threads receive only a single response. Indeed, 90% of threads receive at most 9 replies, while 95% receive at most 15. Therefore, even modestly large threads with more that 15 responses are very rare.

It is worth comparing these numbers to other online interaction settings from the literature, as a way of situating them in the landscape of studies of social influence online. In the study of URL reshare cascades on Twitter in Bakshy et al. (2011), the average cascade size is 1.14, less than a third the average cascade size in of 3.57 this email dataset. So, in this Working Group, which is a professional collaboration setting focused towards a goal, people are much more engaged in the email discussion, compared to the unstructured casual setting of Twitter interactions.

Next, the distribution of cascade sizes is examined. Figure 3.2(a) shows it follows a heavy-tailed distribution. Attempting to fit a power law distribution (with coefficient 1.77), as shown in the log-log scatter plot with the straight line of best fit of Figure 3.2(b), results in a good fit, as shown by the large R^2 of 91% (albeit with a worse fit for large thread sizes). This means that the probability of measuring a particular cascade size value tends to vary inversely as a power of that value. This is further confirmation that the vast majority of emails do not generate large cascades, and that a minority of emails receive the majority of responses. This result is as expected, as power law patterns of this form, with an exponent of around 2, are also common in online interactions on social media (e.g. Bakshy et al., 2011, for cascades on Twitter; Cheng et al., 2014, for cascades on Facebook) and over email (e.g. cascades of forwarded product recommendations and discounts over email, in Leskovec et al., 2007; cascades of forwarded petitions in Liben-Nowell and Kleinberg, 2008).



(a) Normalised histogram of cascade sizes

(b) Distribution of cascade sizes (loglog)

FIGURE 3.2: Cascade sizes

3.2.2.2 Predicting roles using response topology: the ‘influentials’

This section considers the question of whether the characteristics of the topology of the online (email) response network can offer any statistical predictive power, in terms of helping statistically predict whether a given group member has a formal role as a group chair, document editor or document contributor. That is, the question is whether the patterns of email participation group members exhibit are indicative of the kind of role they have in the group. This kind of examination, ranking people based on measures of their social network topology and investigating how this ranking may be related with different roles of the people involved, has not been uncommon in contagion-based studies of online social influence (e.g. Cha et al., 2010; Dubois and Gaffney, 2014).

The hypothesis that this section aims to test is that topological characteristics of the email response network will have some predictive power in terms of helping statistically predict (classify) the roles of group members. That is, the hypothesis is that the people who are active in email discussions will likely also have a formal role in the workings of the group (group chairs) and/or in the construction of the specification documents (document editors and contributors, where these roles are indicative of one's contribution to the documents).

This section begins with an exploratory visualisation of the email network topology, to help offer an initial picture of whether there are any observable characteristics in how group members are interconnected based on how much they respond to each other's emails. Having established some observable characteristics, this section next tests whether group members' network centrality, a very commonly used network characteristic in contagion-based studies of online social influence, is a statistically predictive feature of their formal role in the group, i.e. a feature that is helpful in a classification task that aims to predict a member's role in the group.

Let us first visually explore the characteristics of the topology of the email response network: whether group members interact (through email) evenly with everyone else, or whether they interact more with some members than with others. To perform an exploratory visual investigation of this, based on the frequency of email communication between each pair of members, a social interaction network is constructed. This is shown in Figure 3.3.

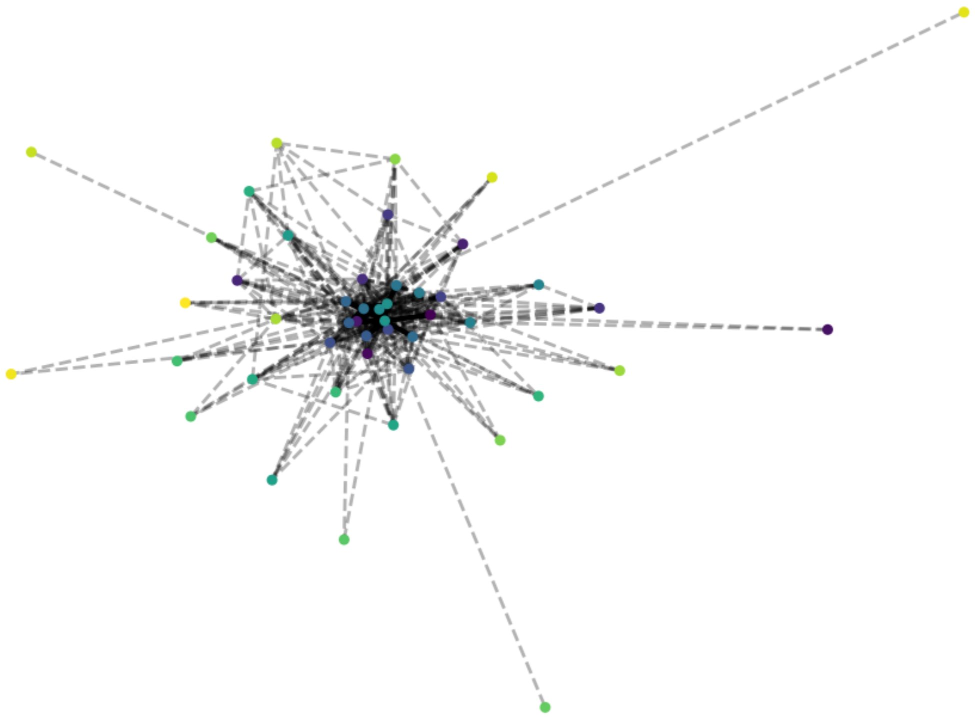
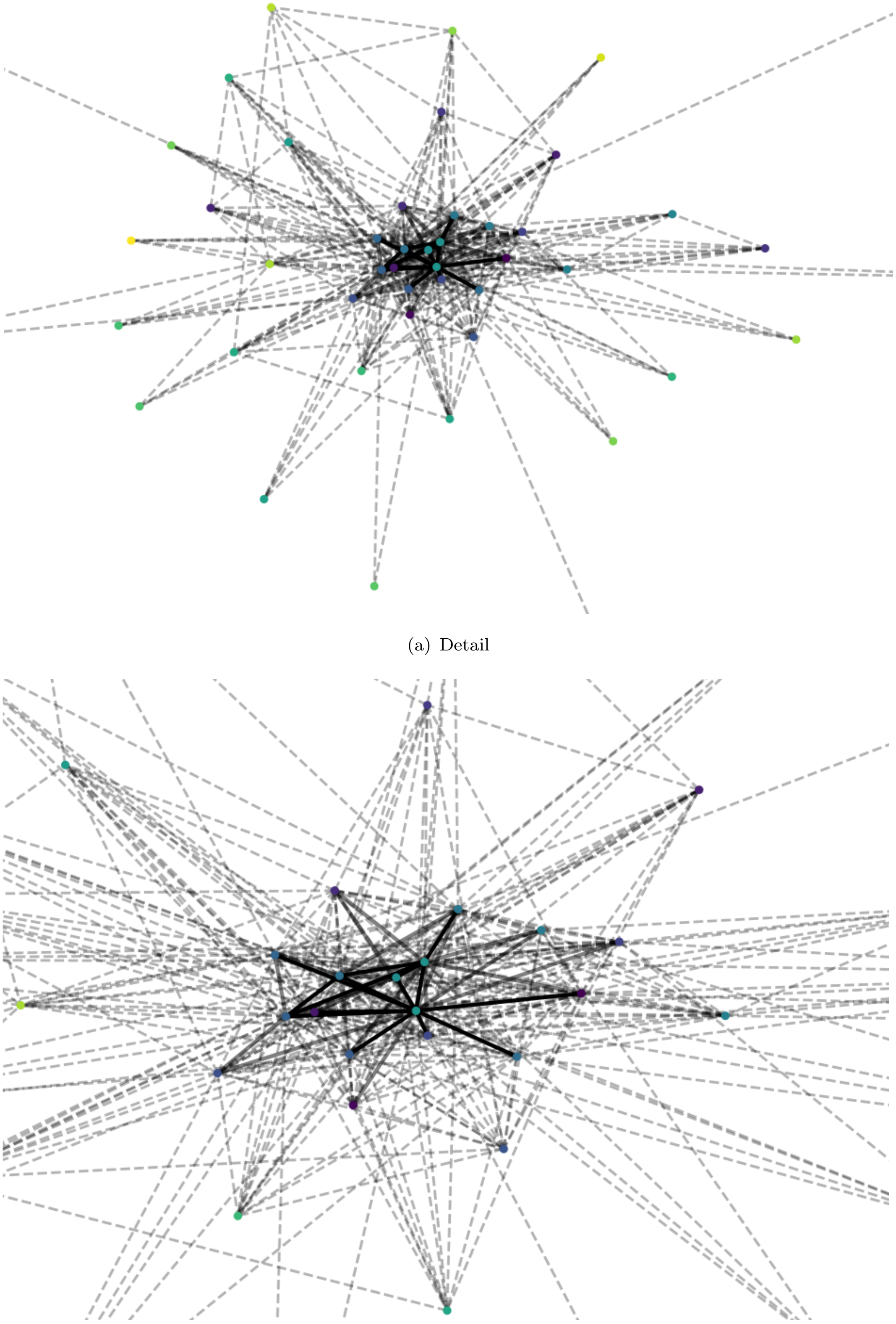


Figure 3.3: Email communications network

In Figure 3.3, nodes represent group members, and there is an (undirected) edge between any pair of members if they have ever corresponded, i.e. if either of them ever replied to an email sent by the other. This information comes from the In-Reply-To field in the SMTP header of each email.

Further, edge styles indicate the frequency of correspondence, in terms of how many replies (in either direction) have been exchanged between each pair of people: a solid black edge means this pair has corresponded at least 100 times, a semi-transparent dark grey edge means this pair has corresponded at least 40 times and fewer than 100 times, and a lighter semi-transparent grey dashed edge means this pair has corresponded fewer than 40 times.

Edge colours reflect the weighted degree of that node, i.e. the total weight of all edges adjacent to that node, which represents the total number of replies this node has exchanged (in any direction) with all other nodes. Yellow colour stands for minimum weighted degree and dark purple stands for maximum weighted degree, so, from minimum to maximum weighted degree, colours go from yellow to green to blue to purple.



(a) Detail

(b) Core

FIGURE 3.4: Detail of email communications network

Zooming into the heavily interconnected component, we obtain Figure 3.4. From the visualisations in Figures 3.3 and 3.4, one may visually discern a core-periphery structural pattern: there is a heavily interconnected component in the centre, connected by strong edges (solid black), and more peripheral nodes connected more sparsely (fewer edges) and weakly (dashed grey edges) to this component and to each other. Further, in Figures 3.3 and 3.4, one may also observe that nodes in the periphery of the network often have yellow and light green colours, meaning that they tend to have low weighted degree, i.e. they communicate little with others, while nodes in the centre or 'core' of the network often have dark green, blue, or purple colour, meaning that they have high weighted degree, i.e. they communicate a lot with others. Indeed, the nodes in the strongly interconnected core cluster, which are connected by the solid black edges, have dark green, blue and purple colours, meaning that they have high weighted degree, i.e. each of them is a relatively heavy communicator, sending and receiving many replies. That is, the strongly interconnected nodes are also heavy email communicators overall.

Therefore, from this exploratory visualisation, it is possible to observe that not everyone is equally active in the emails; rather, some people send and receive many more replies than others: there is variation in node colours, which represent weighted degree. One may also observe that not everyone interacts evenly with everyone else, but rather, there is a strongly interconnected core cluster (connected by solid black edges), and more sparsely connected nodes in the periphery (light dashed edges). This sort of pattern is common in online interaction settings (e.g. Kitsak et al., 2010, Gomez Rodriguez et al., 2010; González-Bailón et al., 2011).

Having inspected the interpersonal tie structure in the visualisation above, we now turn to the question of whether the people that have many strong ties to many others might be the ones credited as contributors or editors on the specification documents, and/or the chairs of the group. Therefore, this question attempts to examine the association between 'influentials' as 'well-connected' individuals in the interaction network (individuals that receive a lot of responses in online communication settings, per the contagion paradigm), with the formal roles participants may have had in shaping the contents of the actual outcomes of the Working Group (editors and contributors of the documents) and in shaping the operation of the Group itself (the chairs).

In order to investigate this, the relationship is examined between one's centrality values in the online response graph and whether s/he is a contributor or an editor to the documents, or a chair of the group. (It is noted that both chairs were also editors and/or contributors to documents, so these two roles overlap for the chairs.) This kind of investigation is common in empirical studies under the contagion based paradigm (e.g. Cha et al., 2010; Dubois and Gaffney, 2014; Ghosh and Lerman, 2014), where people's various centrality values in the social network (in the followership or friendship network) are associated with the number of responses their posts obtain (often for different kinds of responses, e.g. retweets and mentions), in an attempt to assess the extent to which centrality metrics reflect, or correlate with, 'influence' in terms of interaction. The difference is that here there is no social network data (friendships or followerships), so the response network is used for measuring centralities, and those are compared against formal roles, which exist in this setting but are not as common in social media studies – however Cha et al. (2010) and Dubois and Gaffney (2014) do something similar, where they classify social media users into news accounts, celebrities, politicians, etc., and investigate which centrality and response metrics rank which types of user at the top.

The centrality measures used here are the *degree*, and the *in-degree* (very commonly used in the literature), and for each an *unweighted* and a *weighted* variant is calculated. A person's ***(unweighted) degree centrality*** equals the total number of (undirected) connections they have to everyone else, i.e. how many people they have communicated with in the emails, while a person's ***weighted degree centrality*** is the sum of the weights of all the connections (edges) they have to everyone else, i.e. how many *times* they have ever communicated with anyone. A person's ***weighted and unweighted in-degree centrality*** is the same except that only edges pointing to that person are considered, i.e. only *responses to* that person. For example, if Alice has only communicated with Bob and Cathy, having sent 4 emails and received 3 emails from Bob, and having sent 5 emails and received 4 emails from Cathy, then Alice's (unweighted) degree is 2 (number of people Alice has corresponded with, i.e. has sent emails to *or* received emails from), and her weighted degree is $4 + 3 + 5 + 4 = 16$ (the total number of emails sent and received by Alice). In terms of in-degree, Alice's (unweighted) in-degree is 2 (the number of people who responded to her), and her weighted in-degree is $3 + 4 = 7$ (the total number of responses she received).

These values are computed for every person in this Working Group, out of the 47 members who were active in the emails. Then, members are ranked by each of these centrality values, with the person with largest value at the top. Next, the top k individuals are selected from each ranked list, where k is the number of people in the target set of people that are known to have held formal roles in this Working Group, membership to which we want to predict. Three variations of this analysis are performed, one for each target set of roles: chairs or editors or contributors to Recommendation documents; chairs or editors or contributors to all documents (Recommendations and Notes); chairs.

This task is framed as a binary classification task, where the group members are the items to be ranked. The items ranked in the top k positions by the given centrality measure are classified in the 'Positive' class (of size p), i.e. considered to have been labelled by the centrality-based classification process as having a formal role, and the members ranked below the top k are classified in the 'Negative' class (size n), i.e. considered to have been labelled by the centrality-based classification process as not having a formal role. Then, since our dataset contains ground truth information for which members had which formal role(s) (chair, editor or contributor to a Recommendation document, editor or contributor to any document) and which did not, it is possible to calculate performance metrics for how well each centrality-based classification task classifies people in terms of their roles, and compare each to a random baseline, to evaluate whether and how much each centrality-based measure improves the prediction of roles compared to if one was selecting members purely randomly.

The standard performance metrics for classification tasks are used, *precision*, *recall* and *accuracy*. These are defined in terms of the number of items that were classified in true and false positive and negative classifications, which are defined as follows:

- Positives (p): the number of items assigned to the Positive class
- Negatives (n): the number of items assigned to the Negative class
- True positives (tp): the number of items correctly classified as being in the Positive class, i.e. items that were actually positive and were classified as positive
- False positives (tp): the number of items incorrectly classified as being in the Positive class, i.e. items that were actually negative but were misclassified as positive
- True negatives (tn): the number of items correctly classified as being in the Negative class, i.e. items that were actually negative and were classified as negative
- False negatives (fn): the number of items incorrectly classified as being in the Negative class, i.e. items that were actually positive but were classified as negative

Based on these four types of classification and misclassification, the precision, recall, and accuracy rates are defined as follows:

- Precision = $\frac{tp}{tp+fp} = \frac{tp}{p}$
- Recall = $\frac{tp}{tp+fn}$
- Accuracy = $\frac{tp+tn}{p+n}$

Precision shows the percentage of correct positive classifications: out of those items labelled positive, how many were actually positive. Recall shows how many correct positive classifications were made, out of all classifications that should have been positives (true positives and false negatives). Finally, accuracy shows how many correct classifications were made (true positives and true negatives), out of all items (items classified as positives and as negatives).

In the classification tasks here, by design, the number of items classified as positive (p) is the same as the number of items known to actually be positive (these will have been classified as true positives or false negatives, so this will correspond to $tp + fn$ items), k, so $p = tp + fn = k$, so here precision will equal recall. Hence, in the results tables below, only precision will be reported and its value is the same as the value of recall.

For each target set of k people with formal roles, a baseline scenario precision and accuracy is also calculated as follows. The baseline scenario represents the case where k items are chosen at random from the 47 group members. This is simulated computationally, using a pseudo-random number generator (Python's Random library[4]), to generate a random integer in the range [1, 47]. This yields a random integer, r, and the rth name in the list of all 47 group members is removed from that list, and added to the Positive class. Then, for the remaining 46 members, a random number is selected in the range [1, 46], and so on, until the Positive class contains k items, at which point the rest of the items not in the Positive class are put in the Negative class. This completes the classification task, and precision and accuracy rates are calculated as above. This simulated classification procedure is repeated 100 times, storing the precision and accuracy rate of each procedure, and then the 100 precision and accuracy values are averaged, yielding the average baseline precision and accuracy rate. This is done for each of the three target sets of k people with formal roles.

So, overall, for each of the three sets of formal roles, we investigate whether weighted and unweighted degree and in-degree centralities are good predictors of role in the respective classification task. The results are shown in Tables 3.2, 3.3, and 3.4, with percentages rounded to integer percentage points. For each of the three target sets, the precision of each centrality-based classification is compared to the precision of the respective baseline classification using the difference *(centrality precision - baseline precision)*, and similarly for the accuracy, for each of the four centrality metrics. This improvement measure is shown in the '*Improvement*' columns in each table.

TABLE 3.2: Chairs, or editors or contributors to all documents (k = 29)

Evaluation metric	Baseline	Centralities	Degree (*Improvement*)	In-degree (*Improvement*)
Precision	61%	Unweighted	83% (*22%*)	83% (*22%*)
		Weighted	86% (*25%*)	86% (*25%*)
Accuracy	52%	Unweighted	79% (*27%*)	79% (*27%*)
		Weighted	83% (*31%*)	83% (*31%*)

[4] https://docs.python.org/3/library/random.html

Table 3.3: Chairs or editors or contributors to Recommendation documents ($k = 23$)

Evaluation metric	Baseline	Centralities	Degree (*Improvement*)	In-degree (*Improvement*)
Precision	49%	Unweighted	87% (*38%*)	78% (*29%*)
		Weighted	83% (*34%*)	83% (*34%*)
Accuracy	50%	Unweighted	87% (*37%*)	79% (*29%*)
		Weighted	83% (*33%*)	83% (*33%*)

Table 3.4: Chairs ($k = 2$)

Evaluation metric	Baseline	Centralities	Degree (*Improvement*)	In-degree (*Improvement*)
Precision	4%	Unweighted	100% (*96%*)	50% (*46%*)
		Weighted	50% (*46%*)	50% (*46%*)
Accuracy	92%	Unweighted	96% (*4%*)	96 % (*4%*)
		Weighted	100% (*8%*)	96% (*4%*)

In Table 3.2, the goal is to predict whether a person is one of the 29 chairs, editors and contributors to all documents (Recommendations and Notes documents), $k = 29$. The baseline of random classification has 61% precision. Classifying based on (unweighted) degree rank has precision 83%, and using weighted degree gives 86% precision, so these two offer an improvement of 22-25 percentage points compared to the baseline's precision level. Using the respective in-degree rankings yields the same levels of precision respectively. In terms of accuracy, all centrality-based classifications also improve upon the baseline precision of 52% by 27-31 percentage points, with unweighted degree-based and unweighted indegree-based classification both yielding an accuracy improvement of 27 percentage points and weighted degree-based and weighted indegree -based classification both having an accuracy improvement of 31 percentage points. Overall, in all cases, using centrality-based rankings for classification of roles yields substantial precision and accuracy improvements over the purely random classification.

In Table 3.3, the goal is to predict whether a person is one of the 23 chairs or editors or contributors to any of the *Recommendation* documents only, $k = 23$. Here also, in all cases, using centrality-based rankings for classification of roles yields substantial performance improvements over the purely random classification. In terms of precision, the baseline of random classification has an average precision of 49%, whereas classifying based on (unweighted) degree rank has an 87% precision, hence improving upon the baseline by 38 percentage points. Classifying based on weighted degree yields a similar but slightly lower precision of 83%, a 34 percentage point improvement on the baseline's precision. Classifying based on unweighted in-degree gives the lowest precision of all centrality-based classifications, at 78%, which still offers a large (29 percentage point) improvement on the baseline. The precision of unweighted in-degree is the same as that for the unweighted degree-based classification. The numbers for accuracy are nearly identical (within one percentage point), the baseline having 50% accuracy, and all centrality-based classification accuracy values being the same as the respective precision values, except for the accuracy of the unweighted in-degree-based classification, which has an accuracy of one percentage point higher than its precision.

Finally, in Table 3.4, the goal is to predict whether a person is one of the 2 chairs, hence $k = 2$. The baseline of purely random classification has a extremely low precision of 4%. That is because the purely random baseline classifications, in 92 out of the 100 simulated runs, have a precision of 0% (it classifies neither of the actual chairs in the Positive class), and in only 8 out of 100 simulated runs it has a precision of 50% (it classifies only one of the actual chairs in the Positive class). On the other hand, unweighted degree-based classification has 100% precision, and weighted degree-based classification has 50% precision. In the latter case, only one out of the two people ranked in the top two was a chair, as the other chair was ranked third. Since the baseline precision is so extremely low, selecting by degree or even by weighted degree offer very big improvements on precision, by 96 and 46 percentage points respectively.

In Table 3.4, in terms of accuracy, the baseline classifier seems to perform relatively well, with 92% accuracy. This is because, compared to precision, accuracy also includes the true negatives, and in this analysis where the target set size ($k = 2$) is so small compared to the population size (47 group member, i.e. items), it is expected that the large majority of items will be classified in the Negative class. That is, the tn number in the numerator will always be large, even when the tp is 0. In more detail, the worst-case for a purely random classification is to assign neither of the actual chairs in the Positive class ($tp = 0$ instead of the correct 2) and hence to have all but two non-chairs assigned to the Negative class ($tn = 43$ instead of the correct 45), which would yield a worst-case accuracy of $(0+43)/47 = 91.489\%$. The best-case accuracy for random classification is to classify the chairs in the Positive class ($tp = 2$) and the non-chairs in the Negative class ($tn = 45$), yielding an accuracy of $(2 + 45)/47 = 47/47 = 100\%$. Performing 100 random classifications, we obtain here an accuracy of 92% (91.829% before rounding), which is closer to the worst-case than the best-case scenario for accuracy, and indeed, as reported above, 92 out of 100 baseline runs classified neither of the actual chairs correctly, with only 8 having classified one of the chairs correctly. Hence, the accuracy metric is not particularly informative in this analysis, especially compared to the precision metric, but it is reported for completeness. Given this, the centrality-based classifications improve upon the baseline by 4 percentage points (96% accuracy), or, at best, for the weighted degree-based classification which has 100% accuracy, by 8 percentage points.

In conclusion, this analysis shows that there is some limited differentiation between the in-degree and the degree metrics, in terms of precision and accuracy of predicting (classifying) group participants' roles. In all tables (Tables 3.2 - 3.4), the in-degree metrics (reflecting how many responses one gets, which is commonly taken to reflect whether one is an 'influential') perform as well as, or somewhat worse than, the degree metrics (reflecting how many responses one gets and how many messages they send themselves, i.e. both response and participation), in terms of helping predict (classify) participants' roles.

3.2.2.3 Returning to the matter of causal claims about the influence of online communications

On the whole, the above analysis of Section 3.2.2.2 confirms the hypothesis that the topological structure of online (email) communications does have 'predictive' (associational) power in terms of helping determine group members' roles. This means that people who were active in email discussion (both in terms of sending and receiving responses) were also likely to be credited as document contributors or editors (the latter roles being indicative of one's contribution a document). In addition, the chairs seem to have been very heavily involved both in email discussions and in contributing to documents. That is, overall, people seemed to adopt similarly active roles both in the email communications and in contributing to the specification documents, in this particular setting.

However, this pattern of association between topological patterns of email communications and roles does not imply causation. That is, we now step outside the contagion paradigm, in order to answer the overarching question of this section (Section 3.2), of what the empirical application of the contagion paradigm can and cannot tell us about the social influence (causal effect) of online communications specifically.

While the patterns of statistical ('predictive') association resulting from the above analysis are interesting descriptive findings, they do not constitute evidence of causation, i.e. of the causal effect (influence) of one's online communications on their formal role in the group.

Based on them, one does not know the direction (or source of causation), e.g. one cannot claim that people obtained formal credit as an editor or contributor because they were active or received a lot of responses (high centralities) in the emails, nor can one claim the opposite, that people were active and/or received a lot of responses in the emails *because* they had a formal role in the group. Either might in theory have been the case (or neither, e.g. the hidden common cause case discussed below) but based on this data and analysis, no such claims can be made.[5]

[5] In fact, in this Group, editors were decided up front, in an appointment 'process'. But in the absence of this knowledge, based only on the email communications data and on the editor names in the document credits, and the resulting association of email communication patterns with roles, no causal claims can be made.

Indeed, the associations discovered above may be due to other types of hidden causal paths, e.g. a hidden common cause behind both having a formal role and having high degree and in-degree in emails. For example, one hidden common cause of a person A both having high centrality in the email networks and being credited as an editor or contributor or chair (having a formal role) might be that this person is an expert in the domain this group works on (i.e. the domain of Provenance), hence receiving a lot of responses, contributing a lot to the emails, *and* having a formal role in the Group. Another possible common cause is that this a person might not have particularly exceptional expertise, but might be highly motivated to contribute to this particular effort, therefore sending and receiving a lot of emails *and* having a formal role in the Group. Or, it might be that there are certain topics around Provenance that are of particular interest to this group, while others topics are not considered as interesting or relevant, so it might be that someone has a lot to contribute about a specific topic that is of interest to the group, hence being very active in the emails and also having a formal role.

But based on the data and methods of the above analysis, one cannot know what were the causes behind the observed associations.

Similarly, one cannot claim that Group members highly ranked in terms of the degree and in-degree metrics were 'influentials', as there is no evidence to suggest that a) email replies were caused by the email's sender and not by other causes (e.g. a particular argument itself being persuasive, or provocative; the general topic of a given message being particularly timely; pre-existing shared interest or expertise in the given topic on behalf of the sender and the responders), and b) responding to an email does not necessarily mean having been successfully influenced, i.e. agreeing, or endorsing, or adopting, what is being proposed – hence, responding is not meaningful as an outcome.

Overall, if one is interested in answering a quantitative research question of whether people in this group tended to have (due to whatever causal factors) similarly active roles both in the online discussion (measured in terms of the topological properties of the email communications network) and in terms of contributing to documents or to the group's process (indicated by having a formal role), then the above analysis is appropriate, and yields an affirmative answer to this question. Similarly, the above analysis is extremely useful if one wants to answer a research question framed in the engineering paradigm of 'signal versus noise', that is, the question of whether the topological properties of the network of one's email communications can be used as a signal that a person likely had a formal role in this group, or the question of whether formal roles in the group can serve as a signal that one may have high centrality in the email communications network. The analysis performed above is appropriate to answer these questions, and yields affirmative answers to both. All these are associational questions.

However, these associational questions are entirely distinct from the causal question that is of interest in this thesis, which has to do with measuring the causal effect (influence) of online communications on outcomes of interest. The former questions are associational, and involve specifically the topology of the email communications network and people's formal roles in the group, while the latter is a causal question, about the causal effect of online communications on an outcome. And, as discussed above, while the above analysis is perfectly appropriate for answering the former types of (associational) questions, it cannot answer the latter, causal, question, due to the limitations presented above (direction of causation not known, other possibly confounding common causes not measured, responding to person A's email cannot be automatically attributed to A nor be interpreted as agreement or endorsement).

So, conducting analyses such as the above is appropriate, as long as those analyses are in response to associational research questions and no claims are made about 'influentials', or about the influence (causal effect) of online communications (or about whether this influence operates like a contagious disease). In order for such causal claims to be made, and to measure the influence of online communications, one needs to *first* address the conceptual and methodological limitations discussed above, using causal concepts and methods. That is, as discussed at the start of this chapter, if one wants to use these methods to make causal claims (about the influence of online communications), one must first use causal methods in order to ensure their adoption or endorsement outcomes do indeed unambiguously mean adoption or endorsement (email responses do not, as described above and in Section 3.1), to represent their causal assumptions in a graphical causal model, and to use this model and the backdoor criterion (Definition 2.6.4) in order to determine which variables are possible confounders and to determine whether they introduce confounding bias to the estimates of the influence of online communications, and if so, to measure and adjust for these confounders to remove this bias, as much as possible, in an effort to recover the unbiased estimate of the influence of online communications.

Again, these limitations only apply if one wants to make causal claims about the influence of online communications; otherwise, if one wants to answer purely associational questions such as the ones described above, these limitations are not relevant and the methods used in the analysis above are perfectly adequate.

3.2.3 Summary

In summary, this section attempted to empirically assess whether the contagion-based paradigm can offer any insights on the social influence of online communications. That is, given that claims about the social influence of online communications are causal claims, this section aimed to investigate whether the contagion paradigm allows one to make such claims, i.e. whether it allows one to go beyond associational, descriptive claims about patterns of online response, and how those might be associated with various characteristics of an interaction settings, towards causal claims about the role of the social influence of online communications on outcomes of interest.

To that end, this section used this paradigm's most common measures of online social influence (number and network properties of responses), and it used primarily the kinds of online data this paradigm usually considers (online communications data; in this thesis, email communications from the public W3C Provenance Working Group archives).

Overall, while using the methods and conceptualisations of the contagion-based paradigm yields interesting and valuable insights, these insights do not allow for any inferences about the social influence of online communications. These associational insights have value as descriptive findings, but cannot offer any evidence about whether the social influence of online communications, or other causes, were behind the observed patterns.

In contrast, Chapters 6 and 7 will demonstrate how using causal methods can allow one to make causal claims in this setting, about the social influence of the emails, versus the effects of other factors, in shaping the actual outcomes of interest in this setting, i.e. the contents of the specification documents.

3.3 Summary

As an initial contribution, this chapter has presented a critique of the contagion-based paradigm for understanding social influence online. This critique is two-fold, being comprised of an analytical critique, and of an empirical critique which is based on a the empirical analysis of a real-world dataset.

The analytical critique notes how the assumptions, concepts, methods, and data used in the contagion-based paradigm are generally inadequate to warrant claims on the social influence of online communications on observed outcomes. This includes considering how the contagion-based paradigm often conflates the concepts of social influence with attention and response, defining it in an ad-hoc manner in terms of observed and easily measurable online responses (Ackland, 2013; Freelon, 2014; Tufekci, 2014), how these concepts should be distinguished and how social influence can be defined.

Compared to existing criticisms of the contagion paradigm, the analytical critique presented here contributes a more comprehensive classification and critique of the four key problems of the contagion-based approach, in the context of online social influence specifically. It discusses the problems of applying the analogy and language of the contagious spread of diseases to the setting of social influence, highlights the unfoundedness of assumptions on the causes and on the meaning of easily measured outcomes, and stresses how these assumptions have been taken as self-evident facts, rather than being empirically tested as they should have been. Overall, this analytical critique demonstrates how the methods and concepts (contagion-based models) and the data (ambiguous outcomes, consideration of only one out of many causes) used in the contagion-based paradigm are insufficient to make claims on the social influence of online communications.

In summary, the above four classes of problems of the contagion-based paradigm highlight the key limitations of this paradigm for making claims on social influence online. The analytical critique has shown the problems of applying the analogy and language of of the contagious spread of diseases to the setting of social influence, highlighted the unfoundedness of assumptions on the causes and on the meaning of easily measured outcomes, and discussed how these assumptions have been taken as self-evident facts, rather than empirically tested.

The empirical part of this critique, in Section 3.2, has then demonstrated how these limitations manifest if one attempts to empirically apply the concepts and methods of the contagion-based paradigm to a real-world setting of collaboration and online communication. It has illustrated how far one can go using this paradigm in terms of making claims on social influence online: can one go beyond associational claims on patterns of response, and how those might be associated with people's roles, to causal claims about what the social influence of online communications was on the observed outcomes? While interesting and valuable insights are produced with respect to patterns of online interaction and response, and associations of those with formal roles in the given setting, these findings do not warrant any claims, beyond the associational, about the existence or role of the social influence (i.e. the causal effects) of online communications. Rather, these associational insights remain useful as descriptive findings only.

Overall, the critique in this chapter has discussed analytically and demonstrated empirically how the methods and concepts (contagion-based models) and the data (ambiguous outcomes, consideration of only one out of many causes) used in the contagion-based paradigm are insufficient for making claims on the social influence of online communications.

Given the limitations presented in this critique, Chapters 4 to 7 will present a causal approach that can successfully address these limitations and allow investigators make more robust claims on the social influence of online communications.

Chapter 4

Abstract Causal Framework for online social influence

This far, this thesis presented the way in which the social influence of online interactions (online social influence, for short) is commonly understood and analysed based on observational data of online interactions per the contagion-based paradigm, in Chapter 2. Given this background, and building upon the critique of the contagion-based paradigm from Chapter 3, this chapter will present a causal conceptual and methodological framework for conceptualising and measuring online social influence. This framework is named *the Abstract Causal Framework (ACF).*

This is a conceptual and methodological framework in the sense that it is composed of a set of conceptual and methodological principles, with the goal of addressing the problem of conceptualising and measuring online social influence. It is 'abstract' (i.e. general) in the sense that the principles it proposes are intended to be generic, and not specific to individual-level or collective-level analyses only, but rather applicable to both kinds of analysis. Further, it is independent of particular settings or of the particular features of any one mode of online communication, and applicable in principle to any context where there are online social communications whose influence on an outcome of interest one wants to measure.

When applying the ACF to a particular setting, the framework gives the investigator the flexibility to apply (or instantiate) it in the particular context of interest, and its particular features, as they see most appropriate.

The ACF contributes a way for addressing the limitations of the contagion-based paradigm, in terms of conceptualising, defining and measuring the social influence of online communications, in a manner aligned with how social influence has historically been understood in the social sciences as a causal concept. The ACF is made up of a set of generic and flexible principles to be taken into account when applying (instantiating) this framework to specific real-world settings, in order to measure, qualify and contextualise the influence of online social communications on outcomes of interest, while appropriately accounting for other relevant causes, based on observational digital trace data. The ACF is applicable not only when performing individual-level analyses of online social influence, but also when performing collective-level analyses, which have received relatively little attention in the contagion-based literature.

The ACF that will be proposed in this chapter is composed of a set of principles, addressing the following core topics:

1. Having a clear definition of social influence, aligned with how social influence has historically been understood in the social sciences as a causal concept;
2. Distinguishing outcomes from causes (not conflating observed outcomes with a specific cause);
3. Taking into consideration any other relevant causes behind the observed outcome of interest;
4. Applying causal methods for the analysis of online social influence (this involves causal modelling, identification of causal effects and assessing confounding bias, estimation of causal effects, and evaluation the fit of the causal model to the empirical data).

Each of these four topics will be discussed in turn, in Sections 4.1-4.4 of this chapter. Then, they will be used to formulate the principles of the ACF, i.e. to define the ACF, in the final summary section of this chapter (Section 4.5).

The applicability of the ACF to different real-world settings, for individual-level and collective-level analyses, will be demonstrated in more detail in Chapters 5, 6 and 7.

4.1 Defining online social influence

The first important step that is necessary when attempting to understand and analyse online social influence is to have a clear definition of it. Based on the definitions of influence and social influence in Chapters 2.3 and 2.6.2, which demonstrate the causal nature of the concept of ‘influence’, and given the shortcomings of the contagion-based paradigm’s conceptualisation of social influence (Chapters 2.5 and 3), this section describes how the terms ‘influence’, ‘social influence’, and ‘online social influence’ (short for ‘the social influence of online communications’) will be used in the analyses presented in this thesis.

As the focus of this thesis is the influence of online social communications on outcomes of interest, i.e. online social influence, in order to clarify the meaning of this term, one must clarify the meaning of its components. Hence, this section considers 'influence', 'social influence', and finally 'online social influence'.

Definition 4.1.1. Influence. The phrase 'A influences (the occurrence of) event B' is used in this thesis to mean 'A is a cause of B', in a manner that is not forceful or coercive, e.g. A does not cause B to happen through threats or coercion, there is free will involved in whether B occurs or not (in contrast to the related concept of power, which involves force or coercion, per Dahl, 1957). It is not required that A intended to influence B. A represents one or more events, such as a single or multiple actions, behaviours, decisions, statements, or messages, attributed to a person or group. By extension, A might represent that person or that group itself, but it is argued here that it is better to specify those particular actions, behaviours, or other events (attributed to that person or group) that are relevant to the occurrence of B. B represents one or more events, similarly to A. A may be one of several causes of B, it does not have to be the only cause of B. Here, B is the *outcome*, and A is a *source* of causation for this outcome. So, this is how the verb 'to influence' is used in this thesis. The noun 'influence', for the above example, means 'the causal effect of A on B', i.e. the role that A had in causing the occurrence of B (Pearl, 2009a, Thomas, 2013, as discussed in Chapters 2.3 and 2.6.2). If it is known who produces outcome B, then the phrase 'A influences C to do B' can be used, where C can be a person or a group of people, that freely chose to do B (or perhaps subconsciously did B, as influence applies also in such contexts), i.e. C was not forced or coerced to do B. For instance, B may be an action or decision taken by C. This usage of the term 'influence' follows the way the term has been traditionally understood as a form of causation, as a directional causal effect (Thomas, 2013; Pearl, 2009b, and as discussed in Chapters 2.3 and 2.6.2).

Definition 4.1.2. Social influence. Given that 'influence' is meant in the above manner of Definition 4.1.1, the term 'social influence' denotes influence that happens in a setting (or context) of *social* interaction or communication. So it is a qualifier for influence, to denote that one is not studying e.g. the causal effect of a given medicine on patients (medical context), or the causal effect of a new education policy enforced by a government on the standardised test scores for schools in an area (policy context) - rather, one is studying the effects of some kind of social interaction or communication (A), whether formal or informal, professional or casual, produced by an individual or a group of people, on an outcome (B) of interest. The outcome B might, or might not, also be an instance of social interaction or communication (e.g. a written message). The agent C producing the outcome may or may not be specified.

Definition 4.1.3. Online social influence, or, the social influence of online communications. Based on these definitions of 'influence' and 'social influence', the term 'online social influence', used in this thesis as shorthand for 'the social influence of online communications', is then used to denote the social influence of online communications. These are communications occurring in an online setting of social interaction or communication, such as one (or more) social media platforms, Web forum, blog, email, or other Internet-mediated communication. Here, one is studying the effects of some kind of online social communication or interaction (A), such as a message posted online, or a group of messages, on an outcome (B) of interest, which may or may not itself be an example of social interaction or communication, online or offline. As in the above definitions, the agent C producing the outcome may or may not be specified. Therefore, 'online social influence' is used to denote the causal effect (the influence) of online communications (e.g. one or more online messages) on an (online or offline) outcome of interest. The focus of this thesis is on text-based online communications, but of course online social influence in general also applies to other forms of online communication (e.g. video posts, audio posts).

It is noted here that, even though in common parlance one may say 'cause and effect' where 'effect' means the result or outcome of that particular cause, in this thesis, as in the causal literature, the terms 'effect' and 'outcome' are used differently. The term 'cause' here means the same thing as in the phrase 'cause and effect'. However, instead of 'effect,' this thesis uses the term ***'outcome'***, and the word ***'effect'*** instead means 'the importance of the role that this cause played in the occurrence of this outcome' or 'the amount of impact, or influence, this cause had on the occurrence of this outcome'. Hence, in a causal diagram of the form $X \to Y$, Y is the outcome ('effect' in common parlance, but not in this thesis and in the causal literature), X is a possible cause of Y, and the arrow from X to Y represents the (causal) effect, or the impact, or the influence, of X on Y.

4.2 Distinguishing cause from outcome

When one is interested in studying online social influence (the social influence of online communications) in an observational setting, whether theoretical or empirical, one must apply clear definitions of influence, social influence, and online social influence (Definitions 4.1.1, 4.1.2, and 4.1.3) .

In doing this, it is important to not conflate the occurrence of an outcome with evidence that one specific event caused this outcome (i.e. with evidence of social influence from specific online communications of interest), as has been done in the contagion-based paradigm, by either defining the social influence of a given set of communications as an observed outcome (e.g. in Ghosh and Lerman, 2010b) or assuming that an observed outcome can be interpreted as evidence of social influence from one particular person's, or group's, online communications (e.g. as in Barbieri et al., 2013; Cha et al., 2010; Dubois and Gaffney, 2014). That is, when investigating the social influence of certain online communications (e.g. online communications produced by a given person or group, or online communications related to a keyword or hashtag) on a given outcome, one must not conflate observing the outcome with evidence that its only cause was (its only source of influence was) the social communications produced by the person or group of interest.

Therefore, one needs to distinguish between what observed events (e.g. actions) in one's dataset (or in theoretical models), qualify as outcomes, and what events, or people, qualify as possible causes of these outcomes.

In terms of determining outcomes in an empirical setting, one needs to choose observed events that reflect the outcome of interest for the given research question (e.g. was product X purchased? Was belief Y adopted?). If this is not directly recorded in the dataset available to the investigator, and one needs to use proxy outcomes instead (as is commonly the case, e.g. in social media studies, as discussed in Freelon, 2014; Watts, 2011), then the proxy outcomes chosen need to be accurate reflections of the outcome of interest, as much as possible, and need to be unambiguous (e.g. retweeting something does not mean adoption or endorsement of it, as discussed in Chapters 2.3 and 2.6.2). For example, if one finds that the only observed outcomes they have available in the dataset cannot indicate adoption or endorsement of a belief, product, or other type of focal item of interest, and they can at best indicate some level of interest or attention, then this changes the claims on influence that can be made. In this case, a research question of, for example, 'was belief Y adopted?' changes to 'was their interest shown, or attention paid to, belief Y?' That is, the kind and strength of the outcome (e.g. endorsement versus just interest or attention) affects the kind and strength of claims on can make on social influence (this is discussed further in Chapter 5). That is because, as noted in the previous section, influence needs a source (A) and an outcome (B); it is defined with reference to those two components.

Once appropriate outcomes have been established, one must then determine what observed events (or people), qualify as possible causes of that outcome: this includes observed events (e.g. actions) by the person or group of interest, whose influence on the outcome the investigator wants to study. However, outcomes are generally affected by more than one cause, so other relevant causes must also be considered. This issue is considered next.

4.3 Considering other possible causes of outcomes

In an empirical or a theoretical study, besides one particular source of social influence on the given outcome, that one is interested in studying, one also needs to consider other causes of that outcome. That is, in the case where the source of interest is online communications, where one is interested in studying the social influence of online communications on an observed outcome, one should also consider other causes of that outcome. These other causes may of course include causes from the social environment, e.g. offline face-to-face discussions with one's friends, or broader trends and practices in society. In real life, for an event to happen, or for a person to form an opinion, generally a multitude of factors play a role, and the online communications of a specific person, group, or relating to a specific event, is only one of those influencing factors (e.g. Aral et al., 2009; Bakshy et al., 2011; Mason et al., 2007; Watts and Dodds, 2007; Watts, 2011, Chapter 4).

Therefore, one must also determine other relevant causes of the given outcome, other than social influence from a given source (i.e. a given causal, or influencing, factor, such as online communications). For these causes, one also needs to determine whether they may cause not only the outcome but also the causal factor of interest: such common causes introduce confounding bias to one's estimate of the magnitude of social influence from the source of interest on the outcome, so they must be measured and adjusted for (as described in Chapter 2.6.3.3).

That is, not only is it useful to consider other relevant causes as a way to contextualise the role of the influencing factor of interest (A) in the broader landscape of causal factors, but it is necessary to consider and account for such causes (to the extent possible) if they are common causes both behind the outcome and behind the influencing factor of interest. This can be done using causal methods, as described in Chapter 2.6.4.2, following the steps outlined in the next section.

It is often not possible, in empirical studies of real-world settings, to measure all possible causes of a given outcome, and hence to measure all possible common causes and remove all confounding bias, as there is often a multitude of causal factors that affect outcomes in real-world human interactions and decisions. Still, it is important to acknowledge this, and to appropriately contextualise and qualify any empirical claims on the social influence of a given cause (e.g. online communications) versus the influence of other (social an non-social) causes, by trying to measure and adjust for as many confounding causes as possible.

4.4 Applying causal methods

Given the causal nature of the concept of influence, causal methods allow one to reason about and measure the influence of online communications on outcomes, by appropriately accounting for other causes, particularly causes that introduce confounding bias (as presented in Chapter 2.6). This involves the following sequence of steps: causal modelling, identification of causal effects, estimation of causal effects, and evaluation of the fit of the causal model to the data. Each step is next discussed in turn.

Causal modelling. As explained in Chapter 2.6.3, graphical causal models allow one to represent causal relationships in a directed acyclic graph (DAG). This allows one to encode their causal assumptions visually, so that glancing at the structure of the causal graph enables them to determine what inferences can be made about the causal relationships among the variables studied. As the information encoded in graphical causal models can equivalently be encoded in Structural Equation Models, an investigator may also, or alternatively, express causal relationships in the form of structural equation functions (Chatper 2.6.3.2), if the structural equation format is deemed to be helpful or more intuitive. For example, in the case of parametric modelling, where an investigator has evidence that the functional relationship between variables takes a particular form (e.g. linear relations), writing out the causal relationships in structural equation form would be useful. On the other hand, if one is interested in nonparametric modelling, and since nonparametric structural equations can be directly read off the causal diagram, the structural equation format might not add extra value.

Identification of causal effects. Based on the rules of causal inference using the structure of graphical causal models, one can identify confounding causes and appropriately adjust for them, using the graphical backdoor criterion (Definition 2.6.4 and Equation 2.3), in order to remove confounding bias as much as possible from estimates of the influence of the factor of interest (A) on the outcome (B), as described in Chapter 2.6.3.3. As confounding is often present in real-world problems, addressing the identification problem (i.e. removing confounding) per this procedure must be done before proceeding to the estimation of causal effects. That is because addressing the identification problem results in a deconfounding set of variables, representing those variables that should be adjusted for during estimation, in order to remove confounding bias as much as possible from the causal estimate of interest. Skipping the identification step, and proceeding straight to estimation by adjusting for an arbitrary set of variables (as is the case in some causal studies), without first having established whether those variables are in the deconfounding set or not (through the causal DAG and the backdoor criterion), is not guaranteed to reduce confounding bias, and it might even increase it (as discussed in Pearl, 2009b, and in Chapter 2.6.4 of this thesis).

Estimation of causal effects. In order to obtain an estimate for the magnitude of influence from a given factor interest, one needs to first determine whether the dataset contains the relevant values of the respective variables, for the quantity to be estimable. That is, commonly estimators describe the *change* in the outcome (B) when the value of causal factor of interest (A) changes from one level to another, e.g. from 0 to 1. Therefore, the investigator must ensure the variable representing A in the dataset does take on both values (0 and 1), in order for the causal effects of A in B to be estimable. In order to measure this change, one may choose estimands that calculate a difference, or a ratio, for example. Depending on how many confounders are in the deconfounding set, one may decide to use approximate estimation methods for the calculation to be tractable, as appropriate, e.g. linear estimation, or if the deconfounding set is not prohibitively large, one may choose nonparametric estimation (per Chapter 2.6.4). As noted above, these estimation methods are not deconfounding strategies: rather, one must first draw the graphical causal model and apply the backdoor criterion, in order to determine which variables are confounders, and what the admissible (deconfounding) set of variables is that should be adjusted for.

Evaluation of the fit of the causal model to the data. The investigator should also evaluate the fit of the chosen graphical causal model to the empirical data, versus the fit of any alternative or competing models. This can be done by examining whether any models violate any of the statistical dependencies and independencies among the variables in the empirical data, and to thus find the model that is most consistent with the given dataset (as described in Chapter 2.6.5).

4.5 Summary: the ACF

Building upon the critique presented in Chapter 3, this chapter has presented the key topics to be addressed by the Abstract Causal Framework (ACF), a causal conceptual and methodological framework for conceptualising and measuring the social influence of online communications based on observational digital trace data.

Having discussed these four topics in turn, each topic (and its key features) is employed next to formulate a principle for the ACF. Hence, the ACF is presented below. It is composed of the following four conceptual and methodological principles, to be followed when applying (instantiating) this framework to a given real-world setting:

Definition 4.5.1. ACF1. *Having a clear definition of online social influence, in accordance with how the term has historically been understood in the social sciences as a causal concept.* This principle requires having clear definitions of the concepts of influence, social influence, and online social influence, by using and tailoring Definitions 4.1.1 - 4.1.3 to the particular setting under study. It also requires clearly specifying the *source* of social influence (a specific set of online communications), and the *outcome* of social influence, and optionally, the agent (person or group) who was influenced by the source to produce the outcome. This principle clarifies that the concept of influence concerns cases where no force or coercion was used in order to achieve the outcome.

Definition 4.5.2. ACF2. *Distinguishing outcomes from causes (not conflating observed outcomes with a specific cause).* Per principle ACF1 (Definition 4.5.1), influence is defined with reference to two components: the source of influence (the cause), and the outcome of the influence. So this principle (ACF2) requires that an investigator identifies appropriate outcomes, and causes (sources of influence), and does not conflate the two. It requires that the investigator does not assume that an observed outcome is necessarily a manifestation of the social influence of specific online communications , i.e. that one does not conflate an observed outcome with evidence that its only cause (its only source of influence) was a specific set of online communications, as there are often many possible causes that influence the occurrence of an outcome. It also requires that an investigator selects appropriate outcomes, that unambiguously mean adoption or endorsement.

Definition 4.5.3. ACF3. *Taking into consideration any other relevant causes behind the observed outcome of interest.* For real-world outcomes, it is often the case that online communications are only one of many influencing factors (causes) that affect their occurrence. Therefore, this principle requires that an investigator also consider other relevant causes of the given outcome. For these causes, it is required that the investigator also determine whether they might be common causes behind both the outcome and the online communications of interest, as such common causes can introduce confounding bias to estimates of the social influence of online communications on the outcome, and so they should be measured and adjusted for. It is often not possible in practice to measure all possible causes of a given outcome, so it this principle requires that one measure as many confounding as possible, and that one acknowledge the possible existence of any confounding causes that could not be measured, in order to appropriately contextualise and qualify claims about the influence of online communications on the outcome of interest.

Definition 4.5.4. ACF4. *Applying causal methods for the analysis of online social influence.* Given the causal nature of the concept of influence (principle ACF1, Definition 4.5.1), and the need to consider and adjust for other causes (principle ACF3, Definition 4.5.3), causal methods allow one to measure the social influence of online communications, and to reason about and appropriately adjust for other causes. This principle requires using causal methods, and performing the following steps: graphical causal modelling; identification of causal effects and assessing confounding bias using the graphical causal model; estimation of causal effects (if the dataset contains all required values of the relevant variables, in order for effects to be estimable from the data); and evaluation of the fit of the causal model to the empirical data. It requires that the estimation steps happens only after the identification step (and that the identification step is performed after the causal modelling step), as the identification step is needed to inform whether estimation can occur, and if so, which causal variables should and should no be adjusted for in the estimation step.

Overall, the ACF proposes a generic and flexible way for addressing the limitations of the contagion-based paradigm, by conceptualising, defining and measuring the social influence of online communications in a manner aligned with how social influence has historically been understood in the social sciences as a causal concept. The ACF is made up of a set of principles, to be taken into account when applying (instantiating) this framework to specific real-world settings, in order to measure, qualify and contextualise the influence of online social communications on outcomes of interest, while appropriately accounting for other relevant causes, based on observational digital trace data. The ACF is applicable not only when performing individual-level analyses of online social influence, but also when performing collective-level analyses, which have received relatively little attention in the contagion-based literature. In addition, the principles that the ACF proposes are intended to be generic, as they are independent of particular settings or of the particular features of any one mode of online communication, and applicable to any context where there are online social communications whose influence on an outcome of interest one wants to measure. This gives an investigator the flexibility to instantiate the ACF in the particular context of interest, and its particular features, as they deem most appropriate.

In the next chapters, this thesis will present two instantiations of the framework proposed here, for analysing the influence of online communications on outcomes. Chapter 5 presents an instantiation for individual-level analysis, comprised of a theoretical rather than empirical, analysis. So, this analysis goes as far as the identification (deconfounding) problem (causal estimation, and evaluation of the fit of the causal model to data, are not applicable here). Chapters 6 and 7 present an instantiation for collective-level analysis, using real-world empirical data, performing all steps of the causal method of the framework (including causal estimation and evaluation of the model's fit to data).

These chapters will demonstrate how the abstract causal framework proposed here can be applied in practice in different real-world settings, how it can address the limitations of the contagion-based paradigm, in order to obtain more robust insights on the influence of online social interactions on outcomes, and to theoretically and empirically test the existence and extent of confounding (which the contagion-based paradigm assumes is non-existent, without testing).

Chapter 5

Individual-level Causal Framework

This chapter presents the *Individual-level Causal Framework (ICF)* for analysing the social influence of online communications (or, for short, online social influence). This is the instantiation of the abstract framework (ACF) of Chapter 4 for settings of individual-level decision making and action. That is, the ICF applies to cases where the outcome of interest represents the decision or action of *one individual* (not a collective decision, action, or work - these are addressed in the Collective-level Causal Framework of Chapters 6 and 7).

As mentioned, most of the work in this chapter was published as a poster paper in Liotsiou et al. (2016), which received the Best Poster Award for the accompanying poster.

For the problem of analysing and measuring the social influence (causal effect) of an event or action of interest (specifically here, of particular online communications of interest) on an individual's observed action or decision (outcome), the framework proposed here covers the space of other types of causes that may lead to the given outcome (namely, similarity of personal traits, traits of the focal item, and external circumstances). It systematically considers and formally addresses (using the rules of graphical causal models) these other causes and any confounding bias they may introduce to estimates of social influence from online communications, and also considers how different combinations of these causes can lead to different qualities in the observed outcome. In contrast, previous studies of the social influence of online communications have generally identified and discussed only a subset of these other causes at a time (e.g. particularly social influence versus homophily in Aral et al., 2009; Shalizi and Thomas, 2011; Watts, 2007; also social influence versus susceptibility in Aral et al., 2009; social influence versus congitive factors in Lerman, 2016; the issue of interpretation of observed outcomes in Freelon, 2014;

Tufekci, 2014). In contrast, and in addition to what existing studies have done, the ICF proposed here makes the following contributions:

1. It systematically expands the discussion about the social influence of online communications, from online social influence versus homophily, or versus susceptibility, or versus cognitive factors, to online social influence versus all other types of causes, and offers a classification, or a partitioning, of other causes into classes, while allowing for flexible classification of any specific cause under these classes. This identification and classification of types of causes that may directly affect an individual's actions or decisions is based on established results from the social sciences literature (sociology, psychology, social psychology, cognitive science, social neuroscience, behavioural science, management and marketing).

2. It further expands the discussion about the social influence of online communications, from focusing on data where the social ties between people are known, to data where social ties are not necessarily known, hence covering not only social influence from the communications of one's known social ties but also from people who may not be in one's social network. This is again supported by established findings from the social sciences.

3. It uses causal methodology in order to systematically and formally reason about these classes of causes, and to address any bias they may introduce to estimates of the social influence of online communications, by using the formal rules of graphical causal models, in order to:

    - Examine whether these classes of causes introduce confounding bias to the estimate of the social influence of online communications (and finds that indeed the estimate of the social influence of online communications is confounded with each of them), and
    - Determine what is the minimal deconfounding set of these causes that must be measured and adjusted for, in order to recover an unbiased estimate of the social influence of online communications on the outcome (finding that this minimal set contains causes from each of the causal classes, i.e. that each causal class must be measured and adjusted for).

4. It considers, and classifies, a range of qualitative issues related to how different combinations of the social influence of particular communications or events and the influence of these other types of (social and non-social) causes can lead to different qualities in the observed outcome.

5. It performs a demonstration of the above contributions of the ICF, using previous studies of online settings (but also of offline and of mixed settings), illustrating the usefulness of these contributions, and the versatility of their applicability, across a range of real-world practical settings of social interaction and communication.

In more detail, as discussed in the previous chapters, while one may claim that responses to online messages (such as reshares, mentions, or comments) represent the levels of attention or interest that a given message has received (Ackland, 2013; Watts, 2007), it is far from straightforward to infer the *meaning* or the *causes* behind such observed actions. Indeed, Anagnostopoulos et al. (2008) and Bakshy et al. (2011) recognize that this approach yields an overestimate of the social influence of social media communications. Moreover, it has been acknowledged that the ideal way to make causal claims in empirical settings is to use controlled experiments, but this can often be difficult or infeasible in practice (Anagnostopoulos et al., 2008; Shalizi and Thomas, 2011; Sharma et al., 2015; Spirtes, 2010).

As noted in Chapter 1.1, the difficulty in estimating the social influence of online communications based on non-experimental, observational data is that online communications constitute only one of many possible causes (influencing factors) behind a an observed outcome. Rather than an observed outcome occurring due to specific online communications, there may be other unobserved common causes behind both the observed observed outcome and those online communications. Observationally determining that a given outcome is due to (i.e. its occurrence was influenced by) online communications rather than any one of the other causes, or a mix of many of these causes, is known to be a very difficult problem (Anagnostopoulos et al., 2008; Aral et al., 2009; Bakshy et al., 2011; Shalizi and Thomas, 2011; Sharma and Cosley, 2016; Sharma et al., 2015).

This chapter focuses on the questions of: Why does a person take a given observed action (the outcome of interest)? That is, what are the underlying causes and the mechanism that determine whether this person takes this action? If one were to intervene upon a causal factor, e.g. recruiting an ‘influential’ to endorse a product or a healthy behaviour on social media (Bakshy et al., 2011), would this social media message bring about adoption outcomes, i.e. cause the people exposed to it to buy this product or adopt that healthy behaviour? These are questions typical of *causal* inference (Spirtes, 2010).[1]

This chapter proposes an individual-level causal framework (ICF) for the social influence of online communications, expanding upon the work on distinguishing social influence and homophily from Shalizi and Thomas (2011). This framework is used to show that the social influence of online communications is confounded with causes related to personal similarity, traits of the focal item, and external circumstances. The chapter then describes how this ICF enables an investigator to systematically evaluate, improve and qualify causal claims on the social influence of online communications versus the influence of each of the other types of possible causes, focusing on observational data from online social settings. The discussions in this chapter will likely also apply to cases where the source of social influence is not online communications, e.g. offline social influence,

[1] As opposed to inference based on *statistical* prediction methods (Barnett et al., 2009; Diebold, 1998, 2015; Eichler, 2012, 2013; Hlaváčková-Schindler et al., 2007; Spirtes, 2010; Pearl, 2009b, p. 39), which have been used elsewhere in the literature (e.g. in Borge-Holthoefer et al., 2016; Chikhaoui et al., 2015; Runge, 2015).

i.e. to social influence from any social factor. However, as the focus of this thesis is the social influence of online communications, the discussions in this chapter will focus on analysing thet social influence of online communications on outcomes of interest.

The ICF merges computational methods with causal assumptions rooted in established findings from a range of social science studies, offering a promising way to address the need for interdisciplinary common methodological ground in the nascent field of computational social science (Counts et al., 2014; Lazer et al., 2009; Mason et al., 2014; Wallach, 2016). The focus here is limited to building this theoretical framework, and to performing a theoretical demonstration and evaluation of its merits using previous studies of online datasets. A full empirical application to, and validation of the framework on, an online dataset that can adequately capture the confounding causes (typically left at least partly unobserved in online social datasets) is one possible future work direction.

The following four sections apply the ACF to individual-level settings (principles ACF1-ACF4, from Definitions 4.5.1 - 4.5.4).

5.1 Defining individual-level online social influence

Per principle ACF1 (Definition 4.5.1), from the definitions of social influence in Chapter 4.1 (Definitions 4.1.1 - 4.1.3), this chapter uses the definition where the person C who is responsible for the outcome B is known, that is, the definition 'A influences C to do B', where A is an event or a decision or action taken by a person or group, B is the outcome of interest (e.g. an action or decision taken by C, or an opinion or belief adopted by them), and C is an individual who freely chooses to do B, i.e. C is not forced or coerced to do B by A.

For the sake of continuity with the notation used in Shalizi and Thomas (2011), upon which the ICF builds, from now on instead of person B we will talk about person i, instead of person (or group) A we will talk about person (or group) j, and instead of outcome Y we will talk about action Y. The source of social influence under study is j's online message about their performance of action Y at a time $t-1$, written as $Y_{j,t-1}$, while the outcome is i's performance of an action Y', denoting agreement with, endorsement of, or adoption of j's action Y, at a later time t, written as $Y'_{i,t}$.

Therefore, for the purposes of the ICF, the social influence of specific online communications on an outcome can be defined (similarly to Shalizi and Thomas, 2011) as the phenomenon where a person's adoption, endorsement or agreement with of a focal item (the *outcome*, where the focal item might be, for example, an action, behaviour, opinion, idea, or belief) is caused by another person's (or group's) observed online communications about that focal item: a person i may perform (at time t) an action Y' in agreement with a message Y that person j posted earlier (at time $t-1$, $Y_{j,t-1}$) because j's message about the focal item (e.g. an action, behaviour, idea or belief) was so inspirational, persuasive, or impressive (e.g. j making a persuasive argument based on domain expertise) that i was convinced or became inclined to adopt that focal item at a later time t, denoted as $Y'_{i,t}$. This definition is in line with the discussions on the notion of social influence from Chapters 4.1, 2.3 and 4.1, and, as discussed there, only cases where i has free choice are considered, i.e. where j does not force i to perform the given action.

As the focus of the ICF is online social influence, the focus here is on cases where $Y_{j,t-1}$ represents an online message. The outcome $Y'_{i,t}$ may also be online, or it may be offline.

5.2 Distinguishing causes from outcomes

Per principle ACF2 (Definition 4.5.2) of Chapter 4, it is important not to conflate the outcome with the source of social influence that one is interested in studying, as an outcome may be due to several causal factors. Here, as stated in the previous section, the outcome (denoted as $Y'_{i,t}$) represents an individual's (i's) adoption of, endorsement of, or agreement with person j's earlier message about a focal item ($Y_{j,t-1}$), at a time t.

Causes of this outcome include any circumstances or events that existed at, or prior to, time $t-1$, that may have caused person i to do Y' at a later time t. One of these circumstances or events may be person j's message, Y at time $t-1$, denoted $Y_{j,t-1}$, i.e. the outcome $Y'_{i,t}$ may be due to social influence from the online message $Y_{j,t-1}$. However, the outcome may be due to other factors, whether offline or online, whether within or outside person i, such as personal traits of i (e.g. his/her interests, or prior beliefs), or external circumstances (e.g. current fashions, or social norms, or the economic climate), or the traits of the focal item itself (e.g. if it is an opinion that is very provocative or anger-inducing), or may be due to a mix of factors.

Therefore, observing $Y_{j,t-1}$ and $Y'_{i,t}$ does not warrant making the assumption that the latter is due to social influence from the former (this would be an example of the post-hoc ergo propter-hoc fallacy, discussed in Chapter 3.1.2). The outcome may be due to other causes, and indeed, both observed events, i.e. both the outcome ($Y'_{i,t}$) and the cause of interest ($Y_{j,t-1}$), may have happened independently (no social influence from the latter on the outcome) due to shared causes (e.g. i and j having shared interests or beliefs, or living under the same, or similar, social or economic circumstances, etc). These other causes are next considered in more detail, and whether they may be common causes between the two observed events (as the latter case would mean confounding bias into the estimate of the influence of $Y_{j,t-1}$ on $Y'_{i,t}$).

5.3 Other possible causes of outcomes

In accordance with principle ACF3 (Definition 4.5.3), the importance of considering causes of the outcome, $Y'_{i,t}$, other than social influence from the source of interest (the online message) $Y_{j,t-1}$, is pointed out in this section. Particularly any causes of $Y'_{i,t}$ that may also be causes of $Y_{j,t-1}$ are of interest, as these would introduce confounding into the estimate of the social influence (effect) of $Y_{j,t-1}$ on $Y'_{i,t}$. Such causes will be the focus of this chapter, starting with homophily, which has been established (in Shalizi and Thomas, 2011) to be a confounding cause in studies of the social influence of online communications ($Y_{j,t-1}$) on outcomes ($Y'_{i,t}$). This chapter then expands beyond homophily, which still concerns only the two people j and i, who are assumed to be connected by a social tie (e.g. friendship) $A_{i,j,}$, to also include cases where j and i may not be connected by a social tie but may still have some personal traits in common, and to also include factors outside people j and i, such as traits of the focal item itself, and circumstances in the broader external environment (e.g. current social trends). This section examines these additional types of causes of the outcome $Y'_{i,t}$ and how they too may be common causes behind both the outcome and the cause $Y_{j,t-1}$, hence introducing confounding bias on the effects (social influence) of the latter on the former.

5.4 Causal methodology

Per principle ACF4 (Definition 4.5.4), this chapter uses graphical causal models, and in particular the backdoor equation (Equation 2.3), to reason about the identifiability of online social influence in the presence of different types of confounding causes (i.e. the presence or absence of confounding bias in estimates of the social influence of $Y_{j,t-1}$ on $Y'_{i,t}$, due to unobserved common causes). That is, this chapter focuses on the identification problem for online social influence in the context of individual outcomes, and not on estimation of influence from empirical data.

Graphical causal models are used to reason about possible confounding of the social influence of online communications with the effects other causes, when working with observational data. We begin with the causal model presented in Shalizi and Thomas (2011), which is then simplified slightly without affecting its results with respect to confounding. This causal model is then adjusted such that it can model confounding even in the absence of a social tie. Next, a similar causal model is constructed, which shows how the social influence of online communications is confounded with the effects of similarity in personal traits, with intrinsic traits of the focal item, and with shared external circumstances. Finally, these separate models are put together into a single graphical causal model which shows the causal relations between causes and outcomes, and makes visible which variables should be measured and adjusted for to remove confounding from estimates of the social influence of online communications.

In order to minimise cluttering in the notation and in the causal graphs that will be used in this chapter, all variables that do not have a time subscript should be assumed to have the subscript t, i.e. to represent the state of that variable at time t, the time when outcome $Y'_{i,t}$ occurs, so that they represent the most recent state of things when $Y'_{i,t}$ occurred. If an investigator desires to include previous states of any variable, this can be done by adding one or more previous instances of that variable, with their time subscripts ($t-1$ or earlier, e.g. $t-2$, $t-3$ and so on), along with the relevant causal factors, and their time subscripts, that influenced those previous instances of the variable.

5.5 Online social influence is confounded with homophily

The discussion of the social influence of online communications versus the effects of other causes begins by presenting the graphical causal model used in Shalizi and Thomas (2011), which demonstrates that the phenomena of homophily (the tendency of people to form social ties with people similar to them, i.e. personal similarity causing tie formation) and of behaviour adoption due to social influence from friends are confounded in observational social network data. We follow their notation for consistency:

- Symbols X_k and Z_k denote sets of random variables representing, respectively, the unobserved and observed personal traits of person k. Each of those may be discrete or continuous;
- $A_{k,l}$ is an observed variable, for simplicity in this case assumed to be binary, with value 1 if person k considers person l to be a 'friend', and with value 0 otherwise;
- $Y_{k,t}$ is an observed response variable, denoting whether person k performs action Y at a time t, and may be discrete or continuous.

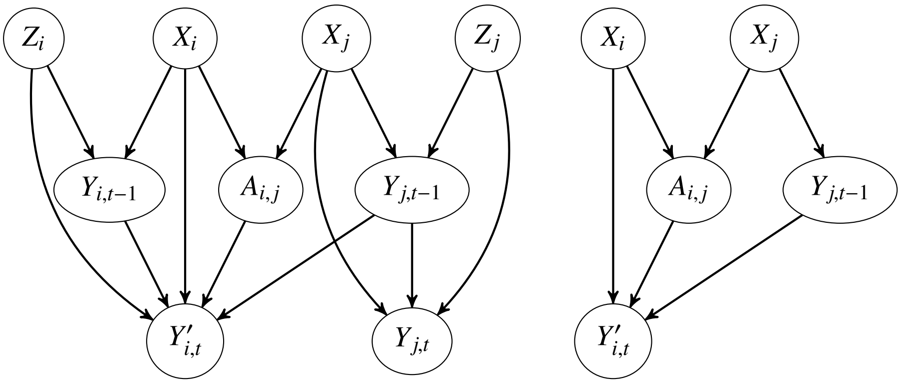
For simplicity, it is assumed that time progresses in discrete steps (although this is not essential, as stated in Shalizi and Thomas, 2011). It is also assumed that there is *latent* homophily in this system, hence whether two people are friends, i.e. whether $A_{i,j} = 1$, depends causally on their latent personality traits X_i and X_j (as well as on their observed personality traits, Z_i and Z_j). The model is shown in Figure 5.1(a).

The goal here is to estimate the social influence, i.e. the *causal effect*, of person j's online communications Y about a focal item at time $t-1$, $Y_{j,t-1}$, on person i's subsequent adoption (Y') of the same focal item at time t, $Y'_{i,t}$, represented by the arrow $Y_{j,t-1} \rightarrow Y_{i,t}$: person i adopts (Y') the focal item of interest because person j's earlier online communications about that focal item inspired them.[2] Latent homophily introduces a

[2] In Shalizi and Thomas (2011), it is assumed that one can be directly socially influenced only by those people she considers her 'friends' ($A_{i,j} = 1$), and not by anyone else.

backdoor path between $Y'_{i,t}$ and $Y_{i,t-1}$ through the latent X_i and X_j: $Y_{i,t} \leftarrow X_i \rightarrow A_{i,j} \leftarrow X_j \rightarrow Y_{j,t-1}$, i.e. the latent X_i and X_j are in the deconfounding set, thus online social influence (the causal effect of $Y_{j,t-1}$ on $Y'_{i,t}$) is confounded with homophily. So X_i and X_j should be measured and adjusted for, to retrieve the pure causal effect of $Y_{j,t-1}$ on $Y'_{i,t}$.

Before moving on to apply this type of modelling to show how the social influence of online communications is confounded with the effects of other causes, the causal model is first simplified for ease of examination of paths. As Shalizi and Thomas (2011) say, the assumption that $Y'_{i,t-1}$ has a direct causal effect on $Y'_{i,t}$ can be dropped without affecting the results of the investigation. Therefore, $Y'_{i,t-1}$ is removed, and, similarly for j, $Y_{j,t}$ is removed - since we are interested in examining the causes behind why i did Y at time t, $Y_{j,t}$ is not relevant.[3] In addition, since the observed personal traits Z_i and Z_j do not play a role in either introducing or removing confounding in this model or in our next models, we also remove those, and assume that all personality traits are unobserved, hence represented by X_i and X_j - indeed, usually there is no, or insufficient, data on users' personal traits in observational online social network studies. This simplification yields the model in Figure 5.1(b).



(a) Causal model from Shalizi and Thomas, 2011 (b) Simplified version

FIGURE 5.1: Graphical causal model from Shalizi and Thomas, 2011 (a), and simplified version (b)

5.6 Social influence is confounded with more than homophily

This section presents the graphical causal models that show how the social influence of online communications is confounded not only with homophily, but also with each of the following types of causes: similarity in personality traits, focal item traits, and external circumstances. It is noted that all confounding cases are due to structurally equivalent backdoor paths of the form presented in Figure 2.1 - each could essentially be regarded

[3] $Y'_{i,t-1}$ might represent a plausible and relevant kind of cause, e.g. that i adopts the given focal item at time t because i did so in the past, at time $t-1$, and was happy with the results, or out of habit from having done it previously at time $t-1$. However, this previous happiness or habit may best be included in X_i as a variable representing an interest in Y.

as a common cause: person-internal (personal traits), item-internal, or circumstances external to the person and the item.

5.6.1 Confounding with similarity in personal traits

Two people i and j may *independently* adopt or agree with a focal item because they share one or more personal traits, such as interests, values, beliefs, opinions, needs, desires, personality profile, or demographic characteristics, like age, race, gender, social class, economic circumstances (Aral et al., 2009; Aral and Walker, 2012; Bakshy et al., 2012; Mason et al., 2007; Sperber, 1996). For instance, two people may each independently post about political news on Twitter, because they each have an active interest in politics. Such person-related traits may also include more temporary circumstances or characteristics, such as a person's current cognitive load or attention span (Fogg, 2009; Lerman, 2016), which, for example, may be the cause why two people i and j do not comment about something on Twitter, or do not engage in a new behaviour, or why two people i and j only comment on a political news post on Twitter that appeared at the top of their feeds, due to Twitter's algorithm, and do not comment, or even see, similar posts that are much further down in their feed, or individual's current financial circumstances (Fogg, 2009), which affect whether they will be able to engage in an expensive activity, or buy an expensive product.

It is noted that personal traits are considered here at a given instant in time, i.e. like a snapshot of one's personal traits at the time point of interest. These personal traits at the given time point may have been developed due to both the person's 'nature' and 'nurture', i.e. both due to the individual's innate personality traits, psychological tendencies, genes, talents, and so on, and due to the external environment they have experienced through their life up to that point, and due to social influence from others. But, at the time point of interest, we 'freeze' the personal traits to what they are, and the past causal factors that shaped them are out of scope for the purposes of this analysis.

To show how a shared personality trait may be a cause behind i and j independently adopting or agreeing with a focal item ($Y_{j,t-1}$ and $Y'_{i,t}$), we now replace the previous latent personal trait variables X_i and X_j with W, representing the latent shared traits between i and j (i.e. W is the intersection of sets X_i and X_j), and W_i, i's remaining latent traits that j does not share, and respectively W_j for j's latent traits that i does not share. This produces the model of Figure 5.2, which shows that $Z = \{W\}$ is the deconfounding set on which to perform backdoor adjustment.

Therefore, in practice, whenever a person i will adopt a focal item ($Y'_{i,t}$) does not only depend on observing others' online communication about their adoption of that focal item, $Y_{j,t-1}$, but also on their own personal traits and circumstances.

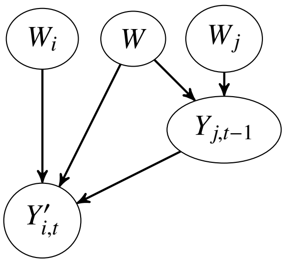


Figure 5.2: Graphical causal model for the social influence of online communications versus the influence of similarity in personality traits

The importance of one's interests, personality, and other traits in whether they will adopt or not a proposed focal item (e.g. behaviour, opinion, action, or product) has long been known in advertising and marketing, where the practice or targeting, customisation, or personalisation is widespread: this involves customising an advertisement or other message according to the target audience's characteristics, to enhance the probability of adoption of the focal item (e.g. purchase of the advertised product). Given the proliferation of Big Data in recent years, personalisation has become extremely widespread in many businesses (e.g. Shriber, 2017), including of course in the Web domain, where it is common practice across major Web companies, for example being used in advertisements and sponsored posts on Facebook (e.g. Faggella, 2016), Twitter, Instagram, and on Google search, in Amazon product recommendations, and in Netflix movie recommendations (Wilson, 2017). These companies leverage the vast amounts of user data they collect in order to raise the advertising revenue on which their operation relies. This practice of capturing personal data for use in personalisation and targeted advertising has become so widespread that it has raised ethical and privacy concerns, highlighting the crucial role of ethical oversight in determining whether and how such data should be recorded, stored, and used (e.g. Crawford et al., 2014; Mittelstadt, 2016; O'Neil, 2017; Ugander, 2017; Wilson, 2017).

In the academic community, targeting and personalisation have long been active areas of research. In the field of computer science alone, there are several long-established academic publication venues devoted to personalisation (e.g. the ACM-sponsored conference on User Modelling, Adaptation and Personalization, which developed from the User Modelling conference which started in 1986;[4] the ACM conference on Recommender Systems[5]).

Indeed, in the field of psychology, it has been found that people often tend to discard information or facts that go against their own beliefs - this is often called *confirmation bias* (e.g. Gregg et al., 2017). In the context of politics it has been found that fact-checks and corrections of false or unsubstantiated information are often ineffective, and can even backfire, when they run counter to the audience's predispositions (Nyhan and

[4] http://www.um.org/umap2018/sponsors/
[5] https://recsys.acm.org/recsys18/

Reifler, 2010; Nyhan et al., 2017). Similarly, the cognitive linguist George Lakoff talks about the finding that people think in terms of typically unconscious structures called 'frames', which encode how a person views the world (Lakoff, 2010). Lakoff discusses how these frames can be so powerful that 'If the facts don't fit the frames, the frames stay and the facts are ignored' (Lakoff, 2004), and has analysed how framing affects political discourse, as well as the debate around global warming and the environment (Lakoff, 2004, 2010). That is, the importance of personal traits is established in computer science as well as in several social scientific disciplines.

A note on homophily and on the existence of ties It is worth clarifying here that homophily is not a confounding *cause* with respect to estimating the social influence of online communications, as it is not a cause but a *separate social phenomenon*, whereby two people form a tie because they have similar personal traits. So, homophily is a phenomenon where the outcome is tie formation, and the cause is similarity in personal traits. In the study of the social influence of online communications, the confounding cause in this regard is shared or similar personal traits, not homophily. This is also made clear in Shalizi and Thomas (2011), even though the title is about homophily versus contagion ('contagion' being the word they use to denote the social influence of a given action or event on an outcome).

Further, the ICF expands beyond cases where there is known to be a social tie between the source of influence (j, the person or group whose online communications may influence the outcome performed by the individual of interest) and the individual of interest (i), to cases where one may not have data on existence of ties, or where one may know that a tie does not exist between the two people studied. That is because, even in the absence of a social tie, the common cause behind two observed actions may still be similarity in personal traits (Fogg, 2002; Kelman, 1961; Sperber, 1996).

As discussed, the literature on online social influence has tended to use the contagion analogy for influence, which assumes that adoptions of a focal item can only 'spread' through social influence to someone from one's immediate social network ties (their 'neighbours'), hence focusing on node-to-node models (as noted in Tufekci, 2014, and also in Chapters 2.5.2 and 3), and to use online communications data, most recently from online social network platforms. Therefore, given this focus on individuals between whom social ties are known to exist, the literature has tended to focus on social influence from online communications versus homophily (e.g. Aral et al., 2009; Shalizi and Thomas, 2011), rather than social influence from online communications versus similarity in personal traits (regardless of presence or absence of a social tie). In contrast, the ICF is not limited by this contagion analogy, and is applicable to both cases where social ties are known to be present, as well as to cases where social ties are unknown, or known to be absent.

5.6.2 Confounding with focal item traits

In the social psychology, management, and marketing literature (Berger, 2013; Kilduff et al., 2010), it has been established that certain features can be 'engineered into' a *focal item* (e.g. a belief, an idea, a product) to entice people to reshare a message concerning it with others, making it 'go viral' and potentially increasing sales or adoption rates. An important type among them is features that invoke emotional arousal, specifically *activating* emotions such as excitement or anger, as these have been found to increase the chances that the viewer will then reshare or discuss the message about this focal item, or even adopt the focal item discussed in that message. Hence, investigators should account for such relevant features, as well as other more general features (e.g. the price of a product; the effort or risk associated with a behaviour (Centola and Macy, 2007) that play a causal role in a person's reaction to a focal item. This has also been acknowledged in some computer science studies of social influence on social media, for example in Aral and Walker (2012) and in Bakshy et al. (2011).

Similarly to the case of confounding with personal traits, Figure 5.2, Figure 5.3 shows that variable F, representing the focal item traits, lies on a backdoor path $Y_{i,t} \leftarrow F \rightarrow Y_{j,t-1}$. Hence, the deconfounding set to be backdoor adjusted is $Z = \{F\}$.

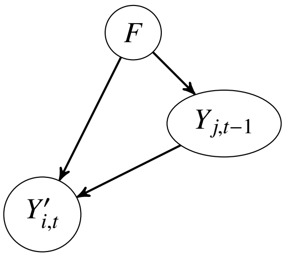


FIGURE 5.3: Graphical causal model for the social influence of online communications versus the influence of similarity in focal item traits

5.6.3 Confounding with external circumstances

External circumstances may be the common cause why two people may *independently* adopt, endorse or agree with a focal item. For example, users i and j may post the same video or URL on social media because it relates to an important current news item, or a popular trend, that they both are aware of. External circumstances encompass factors from the external environment (e.g. a current news item, a trend or a currently popular belief or attitude, a new law, a natural disaster), i.e. factors outside the personal traits of person i and j, and outside the traits of the focal item (e.g. Anagnostopoulos et al., 2008; Bakshy et al., 2011; Ogburn, 2017).

Similarly to Figure 5.3, in Figure 5.4 variable U represents the external common cause (e.g. a shocking news item), and the backdoor path $Y_{i,t} \leftarrow U \rightarrow Y_{j,t-1}$ introduces confounding. Hence $Z = \{U\}$ is the deconfounding set that should be backdoor adjusted.

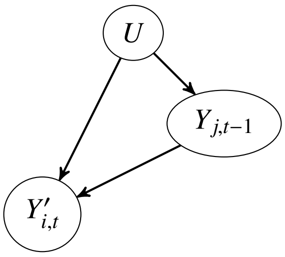


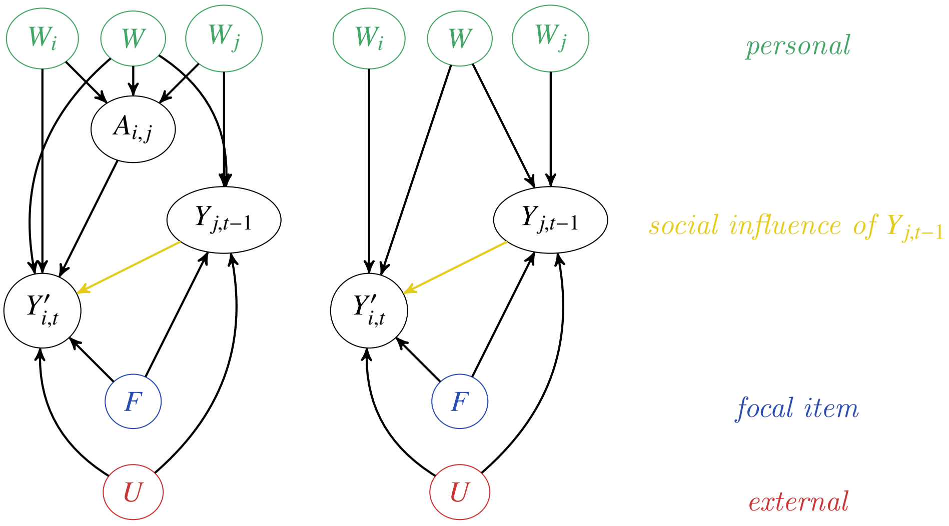
Figure 5.4: Graphical causal model for the social influence of online communications versus the influence of similarity in external circumstances

External circumstances, or contextual circumstances, or shared environment, as they are sometimes called, are also acknowledged as confounders in Lyons (2011) (in the context of discussing the problems of the contagion-based methods used and claims made by Christakis and Fowler in Cacioppo et al., 2009; Christakis and Fowler, 2007, 2008; Fowler and Christakis, 2008), and in the context of the challenges of estimating 'contagion effects' from observational data in Ogburn (2017). Their importance is also noted in a discussion of the uses of Big Data in the social and cultural sciences. For instance, Wagner-Pacifici et al. (2015, p. 7) discusses 'a variety of broader social conditions', using as examples 'current events or other "opportunity structures"', and noting that these are part of 'multiple causal pathways toward an outcome'. In the context of marketing, in Watts (2007, p. 207), external circumstances are mentioned anecdotally, in a case study about the widespread adoption of Hush Puppies shoes. This phenomenon was attributed by some to 'hipsters' being the 'influentials' behind this widespread trend, but Watts (2007, p. 207) points out that hipsters were not necessarily the cause, but instead people may have started buying Hush Puppies shoes in response to the same broader environmental factors (in society, in fashion) as the hipsters. In Fogg (2009), discussing persuasive technologies, external factors such as social norms are also recognised as an important aspect of whether one adopts a focal behaviour or idea. From the domain of psychology and the study of health risk behaviours (particularly smoking in adolescents), Liu et al. (2017) highlight the importance of 'accounting for cultural variables'. Finally, from the domain of cognitive neuroscience and psychology, it is noted that whether acceptance of a focal behaviour occurs depends on the receiver's valuation of external social factors (as well as on personal factors), in Falk and Scholz (2018).

5.6.4 Putting it all together: the influence of online communications, personal similarity, focal item traits, external circumstances

All the above graphical causal models are now put together, to show the full picture of all types of causes that affect person i's outcome Y' at time t, and how these, if left unobserved and unadjusted for, introduce confounding bias into the estimate of the social influence of person j's online communications Y at time $t-1$ ($Y_{j,t-1}$) on person i's individual-level outcome of interest, $Y'_{i,t}$.

Keeping the same notation, two models are presented here: one with a social tie variable $A_{i,j}$ in Figure 5.5(a), for cases where it is known that there is a social tie between i and j (i knows j or knows of j), and one without a social tie variable $A_{i,j}$ in Figure 5.5(b), for cases where either it is not known whether there is a social tie between i and j (social tie data is missing from the investigator's dataset), or where it is known that there is no social tie between i and j (i does not know, and does not know of, j). The phrase 'i knows of j' is used here to include cases where i may have not met j in person, but where i is still aware of who j is and has an opinion of j (whether positive, negative, or neutral).



(a) Full model with social tie (b) Full model without social tie (c) Context

FIGURE 5.5: Full graphical causal models for the social influence of online communications versus the influence of other causes, with social tie variable $A_{i,j}$ (a), and without social tie (b), with the legend (c) on the right showing the context of each latent causal variable and of the causal relation of social influence

In Figure 5.5(a), the social tie variable $A_{i,j}$ may take values in the range [-1, 1], where $A_{i,j} > 0$ indicates a positive social tie, i.e. i knows (of) j and thinks positively of j, $A_{i,j} < 0$ indicates a negative social tie, i.e. i knows (of) j and thinks negatively of j, and $A_{i,j} = 0$ indicates a neutral social tie, i.e. i knows (of) j and has a neutral opinion of j. How to most appropriately measure the quality of the tie (positive, negative, neutral), and at what granularity (e.g. discrete values, real vales) is up to the investigator to decide based on the context under study.

Having two separate models, one with and one without a social tie, is useful because it helps minimise any ambiguity around what $A_{i,j}$ means and how it should be interpreted. That is, it is chosen to have two separate models, rather than having only the model of Figure 5.5(a) and having the following three possible interpretations for $A_{i,j} = 0$: 1) "a tie exists between i and j and i has a neutral opinion of j"; 2) "it is not known whether a tie exists between i and j (missing data for social ties)"; and 3) "it is known that a tie does not exist between i and j". These three interpretations represent distinct cases, and hence should be treated separately at the modelling stage, to avoid confusion and ambiguity around the meaning of $A_{i,j} = 0$.

In more detail, let us compare the meaning of $A_{i,j} = 0$ in Figure 5.5(a) to the meaning of each of the two cases for which Figure 5.5(b) should be used (namely, the case of not knowing whether a tie exists, and the case of knowing that a tie does not exist). $A_{i,j} = 0$ in Figure 5.5(a) means that there is a tie between i and j, and moreover this tie is neutral, which is not the same as not knowing whether a tie exists between i and j (missing data). And, the $A_{i,j} = 0$ in Figure 5.5(a) which means that there is a tie and it is neutral, is also not the same as knowing that there is no tie between i and j, i.e. knowing that j is a stranger to i and i does not even know of j. The latter case means that i has not even had the chance to meet j or hear of j. But if i known (of) j (establishment of social tie, $A_{i,j}$), this does not guarantee that i would have formed a neutral opinion of j ($A_{i,j} = 0$); rather, i could have formed any opinion on j (positive, negative, or neutral). So, we just do not know what opinion i would have, hypothetically, formed of j, had i known (of) j. Therefore, we cannot equate the case where the data shows that i does not even know (of) j with the case where the data shows that i knows (of) j and has a neutral opinion of j.

In addition, Figure 5.5(b) can be used to represent two kinds of cases, as discussed above: a case of having no social tie data for a given pair of individuals (not knowing whether there is a tie between them), and a case of knowing that there is no tie between two individuals. When using this model, an investigator should state which of theses two cases this model represents, depending on which of these cases applies to the investigator's dataset. That is, any given instantiation of this model should represent only one of these two cases, and it should be clearly stated which case it represents, to avoid ambiguity. If a given dataset contains data on more than two individuals, and contains instances of both these cases, and also instances where it is known that a social tie does exist between some pairs of people, then an investigator should use two separate instances of the model of Figure 5.5(b) (one instance for each pair of individuals for whom there is no data on whether there is a tie between them, and one instance for each pair of individuals where the data shows there is no tie between them), and one instance of the model of Figure 5.5(a) for each pair of individuals where the data shows that there exists a tie between them.

In terms of the causal factors affecting $A_{i,j}$ in Figure 5.5(a), given the separation of personal traits into those that both people have in common (W) and those they do not (W_i, W_j), it is assumed that the quality of the social tie (positive, negative neutral) depends on the personal traits i and j have in common and on those they do not have in common. For instance, whether a social tie is positive (i thinks positively of j, whom he knows (of), $A_{i,j} > 1$) depends on having enough things in common (W), and also on not having too many differences in personality (e.g. to the extent that one cannot tolerate or is offended by the other's value system) – hence, besides W, it is assumed that W_i and

W_j also causally affect whether a social tie is positive (i.e. whether $A_{i,j} > 0$), or negative ($A_{i,j} < 0$, the differences in personality or values outweigh the similarities) or neutral ($A_{i,j} = 0$, similarities and differences in personal traits balance each other out).

Overall, in terms of removing deconfouding bias in both models above, the minimal deconfounding set for Figure 5.5(b) is $Z = \{F, U, W\}$, and for Figure 5.5(a) it is $Z' = \{F, U, W, W_j\}$. W_i could be in Z' but it is redundant, due to the assumed asymmetry of A_{ij} (for consistency with Shalizi and Thomas, 2011); if there was an edge $A_{ij} \rightarrow Y_{j,t-1}$ then W_i would have to be in the minimal confounding set. This asymmetry of the social tie applies, for example, to online social network settings with asymmetric ties, i.e. followership (e.g. Twitter, Instagram) rather than friendship (e.g. Facebook).

Hence, in order to retrieve the purely causal effect of $Y_{j,t-1}$ on $Y'_{i,t}$, an investigator must implement the deconfounding strategy from Chapter 2.6.3.3, which, as per Definition 2.6.6, is to: 1) Select a large random sample from the population of interest, 2) For every individual i in the sample, measure $Y_{j,t-1}$, $Y'_{i,t}$, and all variables in Z, and 3) Adjust for Z by partitioning the sample into groups that are homogeneous relative to Z, assess the effect of $Y_{j,t-1}$ on $Y'_{i,t}$ in each homogeneous group, and then average the results, as per the backdoor formula of Equation 2.3. Crucially, all variables in the appropriate minimal deconfounding set must be measured for *every* individual in the random sample, and adjusted for as per Equation 2.3.

One can see that each full model in Figure 5.5 presents a complex picture, with many factors playing a role in i's outcome Y'. Indeed, as shall be discussed in the following two sections, it is known in the social sciences that social influence from one source (one set of online communications alone) is seldom enough to ensure an outcome $Y'_{i,t}$ - rather, a beneficial combination of social influence from online communications and of influence from all the other causal factors is often needed to ensure outcome $Y'_{i,t}$ (or to ensure its probability of occurring is as high as possible).

For example, in the context of modelling and measuring social influence, Mason et al. (2007) advocate accounting for causes other than the social influence of particular communications, saying: ‘models should attempt to integrate social influence with other effects on individual decisions rather than to be models solely of social influence that assume people have no other nonsocial reasons to hold one opinion or another’.

In addition, studies of *online* social influence also highlight the importance of accouting for other causal factors. For instance, regarding the social influence of social media communications, Aral and Walker (2012) state that ‘recipient selection and message content may be important aspects of influence and should therefore be estimated in future experiments’, i.e. they consider personal traits (of the recipient) and focal item traits (F) or traits of the message $Y_{j,t-1}$ itself (depending on what is meant by ‘message content’). They also note the importance of the focal item traits (the ‘product, behavior, or idea [that] is diffusing’). And in the context of political participation and mobilisation online, in Margetts et al. (2015), the authors 'show how different personality types react to social influences and identify which types of people are willing to participate at an early stage in a mobilization when there are few supporters or signals of viability', i.e. they account for personal traits (W, W_i, W_j), sources of social influence (e.g. $Y_{j,t-1}$), and broader external circumstances (U, from the social realm in this case, representing that there are few supporters of the mobilisation), and/or focal item traits (F, representing the signals of viability - alternatively, one could potentially model this as part of U if preferred).

Furthermore, from the domain of psychology and social neuroscience, Falk and Scholz (2018) highlight the importance of social external factors, and personal views, alongside social influence from a person's observed message, $Y_{j,t-1}$, in determining whether an individual will indeed decide to adopt that focal item ($Y'_{i,t}$). As a final example, from the area of behavioural science, Lee and Leets (2002) performed a longitudinal study on the influence of the online presence of hate groups (specifically, white supremacist webpages) on adolescents, and found that personal traits (predisposition), the type of narrative (i.e. the qualities of $Y_{j,t-1}$), and the message explicitness (i.e. a focal item trait) affected whether, and the extent to which, a participant was persuaded by the given message (whether $Y_{i,t} = 1$) and how this outcome changed over time, while acknowledging the broader external environment as another important causal factor of the outcome.

5.7 Qualitative considerations: the impact of causal factor characteristics on the nature of observed outcomes

This section aims to shed some further light on the question of what kinds of causal circumstances are needed for a given outcome to occur (e.g. for person to adopt a focal action or behaviour). In the empirical and the theoretical literature (Berger, 2013; Katz and Lazarsfeld, 1955; Kelman, 1961; Sperber, 1996; Watts, 2011) it has been widely acknowledged that no person is a clean slate, and no situation is 'neutral', therefore the social influence of online communications does not operate in a vacuum, and on its own is rarely sufficient to ensure an individual i adopts a new behaviour ($Y'_{i,t}$) (and hence to influence many individuals i, e.g. so many individuals as to ensure that the product this 'influencer' person is endorsing will sell out). As discussed above, a single 'well-connected' person j alone is generally not enough to reliably influence others i to act a certain way; rather, a combination of compatible personal traits (W and W_j), a focal item with appropriate features F, and beneficial external conditions U are also needed.

Therefore, this section synthesizes and classifies some important qualitative aspects of how different combinations of causal factors may lead to different qualities in the final observed outcome. These qualitative aspects affect the extent and nature of claims one can make about the social influence of online communications, and hence should be measured, e.g. by recording more details of the causal process that led to the outcome than is common in observational social network datasets, in a privacy-respecting manner, or (e.g. to avoid making the process intrusive for participants) through interviews, or through a combination of methods.

Magnitude, direction, and duration of the outcome. As noted in the ACF (Chapter 4) and in the critique (Chapter 3), the outcome of interest should unambiguously denote adoption of, agreement with, or endorsement of the focal item. In addition to this requirement, instead of modelling the outcome $Y'_{i,t}$ as a binary variable (adoption or non-adoption), it could instead have a magnitude, duration, and direction. The magnitude would represent the intensity of i's adoption, Y, from time t onwards – whether this adoption is only superficial (small magnitude) or serious and incorporated into their value system (large magnitude), while duration would capture how short-lived or long-lasting this adoption is (Cebrian et al., 2016; Kelman, 1961). The direction would capture whether i acts as per j's message (i.e. following what the message proposes) with respect to the focal item (positive direction), or whether i does the opposite of what the message proposes (negative direction), e.g. because j's way of engagement with the focal item was against i's values, or whether i does not take any substantive action in relation to the focal item discussed in the message, e.g. out of loss of interest in the message (Alshamsi et al., 2015). For example, in (Mason et al., 2007, p. 293-4) it is pointed out that social influence is not always assimilative, but it may instead be contrastive, with reference to theories of identity from social psychology (e.g. one might be motivated to do the opposite from what his/her peers do, in order to stand out, in an effort to formulate a more original or unique identity). The duration of the outcome over time, in particular whether the influence of a message persists, increases, or decays over time is also studied in Lee and Leets (2002). As a practical example of how such aspects might be captured and measured, Facebook's addition of specific reaction buttons for love, anger, etc. to the 'Like' button, which was previously used to express any type of reaction (Greenberg, 2016), is one approach to capturing direction.

Normative versus informational social influence. A person may adopt a behaviour or take an action not because they find the traits of that behaviour or action (F) inherently worthwhile, but rather because they want to please or feel accepted by someone they know (j, $A_{i,j}$) or by a wider social group (U). In Deutsch and Gerard (1955), the former type is termed *informational influence*, and the latter *normative social influence*, as discussed in Kelman (1961) and in Mason et al. (2007). Which type of social influence occurs in a given case depends on all the causal variables. For instance, Falk and Scholz (2018) discuss evidence that 'self-related and social considerations are two key inputs to the value calculation' (i.e. to an individual i's calculation of the value, or benefit, they would gain if they adopted a focal action at time t, $Y_{i,t}$, having observed another person j's communications advertising the same action earlier, $Y_{j,t-1}$), such as bonding with person j. It is also noted that 'People share when they believe that information is [...] valuable to the way that others will see them', and that motivations for sharing information also correspond to 'central human goals of [...] holding a positive image of the self'.

Generalizability of observed outcomes. Often, investigators use observational social media data capturing the levels of online interest in a product or behaviour as proxies for estimating a different outcome like product sales or adoption of that behaviour. However, it has been shown that the levels of interest on social media may not translate to actual purchases or behaviour change (Cebrian et al., 2016). For example, Berger (2013, p. 196) discusses the case of the online Evian water advertisement which was wildly popular online, being declared 'the most viewed online advertisement in history' by Guinness World Records, but did not increase sales. That is because the causal factors in the two cases are very different: in the latter case, factors that do not apply in online discussions, for instance the price, qualities, effort and/or risk associated with this product or behaviour, F, and society's views of adopting it, U, come into play. Therefore, when using data from online social networks as proxies for outcomes of interest, the underlying causal factors should be adequately similar, in order for the findings and claims with respect to the proxy to also generalise to the actual outcomes of interest. For example, this is recognised in Bakshy et al. (2011), where the outcomes studied are Twitter posts containing a given URL. Here, the authors are careful to note that based on such outcomes one cannot make claims on whether the opinion related to the URL is adopted, or whether a product related to the URL is purchased, by stating that 'influencing another individual to pass along a piece of information does not necessarily imply any other kind of influence, such as influencing their purchasing behavior or political opinion'.

Changing deep rooted behaviours: identity, effort and risk. It has been claimed that the social influence of the communications and actions one is exposed to drives behaviours as diverse as sharing a message with friends, purchasing decisions, smoking habits, happiness levels and divorce (Berger, 2013; Christakis and Fowler, 2007, 2013; McDermott et al., 2013). However, some behaviours (e.g. quitting or restarting smoking, becoming happier, or getting a divorce) are much more deeply rooted in a person's identity, psychology or worldview (X_i plays a stronger role), are more difficult to change (F), and carry more risk in terms of social acceptance (U) (Berger, 2013; Centola and Macy, 2007), than other actions (e.g. re-sharing some information on social media, or choosing which brand of bottled water to buy). Therefore, appropriate consideration must be given to relevant causal factors and confounders, depending on how deeply connected the desired focal item might be in relation to one's personal traits – for instance, one's worldview, values, and identity may not be a very important causal factor in determining which brand of toothpaste to buy, but these are likely to be much more important causal factors when deciding which politician to vote for, or whether they should get a divorce.

5.8 Demonstration and evaluation using previous studies

This section demonstrates how the ICF, including its qualitative considerations, can help investigators position their findings within the full causal picture for the social influence of online communications, assess the extent and types of causal claims on online social influence their data allows them to make, and determine what causal variables should next be measured and adjusted for in order to make more robust causal claims.

This section examines examples of studies that actively try to capture the social influence (causal effects) of online communications (or in some cases offline communications, or other social factors) by reducing the effects of confounders, using quantitative and/or qualitative methodologies, in research settings involving one or more of the disciplines of sociology, social psychology, marketing, and computer science. That is, these studies tend to be based on more data, i.e. data that also includes measurements of some confounders (and hence allows one to make some causal claims), than what is typically done in contagion-based empirical studies of online social influence, which generally do not have data on confounders in the online communications datasets they are based on.

The ICF is employed here in order to examine how these studies lay out potential avenues, as well as expose caveats, for future attempts at measuring and adjusting for confounders and at capturing the qualitative aspects of the social influence of online communications, with a focus on online social settings.

In Aral and Walker (2012), a controlled experiment on Facebook is performed, with the focal item being a Facebook app about films. It is randomized which friends i of j see messages $Y_{j,t-1}$ declaring j's use of this app, aiming to measure the social influence of the messages posted by one's Facebook friends versus susceptibility (i's tendency to imitate $Y_{j,t-1}$ by also downloading the focal item). It is assumed that randomly choosing the subjects i who will be exposed to $Y_{j,t-1}$ will suffice to control for homophily (similarity W among friends i and j linked through $A_{i,j}$) and for exposure to common external causes (U). Hence, it is assumed that whenever an exposed person i also downloads the app ($Y'_{i,t}$) the only cause must be social influence from message $Y_{j,t-1}$ ($Y_{j,t-1} \rightarrow Y_{i,t}$). However, since the alternative causes have not been measured, they may continue to introduce confounding – for instance, it might have been that all people who also downloaded the app did so because they themselves had an interest (W) in films, and all the people who did not download it did so because they had no particular interest in films. Therefore, the cause might rather have been a common personal trait W – one cannot know whether the cause was the Facebook posts of one's friends regarding this app or another cause, until one has measured and adjusted for the confounders for every person i in the sample.

In addition, it is noted here that the action of downloading the app is a relatively weak outcome in itself, as it denotes only a tentative form of adoption, without measuring whether this downloaded app is ever actually used, or if it is used for a long period of time.

Taking steps to observationally measure personality traits for each participant, Aral et al. (2009) use an observational dataset containing many personal traits (X_i, X_j) for each pair of users, in an attempt to disentangle homophily from the social influence of online communications. Still, as explained in Shalizi and Thomas (2011), there may still remain some latent personal similarity (W) which affects behaviour adoption $(Y'_{i,t})$. Moreover, it is noted here that the confounders relating to the focal item traits (F), and to external common cause (U) remain latent. Still, this study shows one way of observationally measuring X_i and X_j to some extent.

In an online randomized experiment, Salganik et al. (2006) manage to measure some confounders and obtain a relatively close estimate of the causal effect of the aggregate social influence (from a number denoting how many times a song was downloaded by others) on users' choices of whether to download a given song (focal item). It is randomized which users i are exposed to aggregate social influence (total number of downloads a song has received, $\sum_j Y_{j,t-1}$, where the identities of users j are not displayed). To reduce the effect of external common cause U, special care is taken (including conducting surveys) to ensure the displayed songs and artists are virtually unknown. The songs are kept the same (F constant) while some participant groups see the number of downloads for each song and other groups do not. However, as W has not been measured, and neither has F (e.g. song genre), a small possibility remains that the same song might have been downloaded more in a so-called *social influence group* than in a neutral one not because of social influence from the displayed download count, but rather because that group contained more participants who were fans (W) of that song's genre (F). Therefore, some confounding due to latent W and F might remain, so these should be measured and adjusted for. Still, this study offers a good example of a significant and detailed effort to reduce confounding bias from U while experimentally controlling F.

In Sharma et al. (2015), observational data is used to study the causal effect of Amazon recommendations of the form 'Customers who bought this [product A] also bought [that product B]' on the views of product B (the focal item). Again, customer i cannot see the identities of customers j who bought both products. The investigators attempt to control for F to an extent, by studying many different product categories, and try to ensure that external causes U are held constant as much as possible. They also investigate the effect of the type of users i they have studied (X_i) on the causal effect of the recommendation. In qualitative terms, they recognize that a user's clicking on a recommendation might be due to convenience rather than the persuasive qualities of this particular recommendation. Overall, they caution that their results are still an upper bound for the causal effect (influence) of Amazon recommendations, but a stricter one than under naive assumptions, and acknowledge that their results may not readily generalize to the average Amazon user, or to all Amazon product categories, or to other recommendation settings. In addition, it is noted here that a view of a product is not an especially meaningful outcome, as it denotes some potential interest in or curiosity about the product, but it is still far from denoting adoption (purchase) of the product.

An example of how qualities of outcomes can be measured at a fine granularity and over time is presented in Alshamsi et al. (2015). Here, the social influence from one participants' emotional state on another's (effect of $Y_{j,t-1}$ on $Y'_{i,t}$), in the setting of offline face-to-face interactions, is measured using a mixed methodology of infrared sociometric sensors (badges) and questionnaires. The authors measure here many 'directions' of outcomes: not just mimetic (termed 'attraction'), but also neutral or negative (termed 'inertia, repulsion and push') at three points per day. They also measure fixed personality traits of the participants, X_i and X_j, but do not measure other confounders, and are careful to clarify that their social influence claims are correlational, not causal.

The offline controlled experiments in Kelman (1961) offer useful examples of how to design experiments, control for some confounders, and use varied types of questionnaires, and how to measure the ways in which the combination of causal circumstances $(U, F, Y_{j,t-1})$ affect the nature of the resulting outcome $Y'_{i,t}$. Here, the goal is to empirically evaluate how different combinations of causal circumstances (particularly $Y_{j,t-1}$, U) lead to different types of outcomes (termed 'compliance, identification and internalization', each denoting a stronger kind of adoption of the focal item than the previous one). Still, the broader external environment U (e.g. popular attitudes relevant to the topic of the focal message) and the participants' personal views (W) remain unmeasured and so may introduce confounding. Experimentally, the core of the argument (F) is kept the same, but the way it is framed $(Y_{j,t-1})$ is varied. To measure the 'magnitude' of the outcome, i.e. extent to which it was internalised and incorporated into i's worldview and value system, and its duration, questionnaires are used which include open-ended questions, both soon after exposure to $Y_{j,t-1}$ and some weeks after.

In summary, this section has demonstrated how the ICF, including its qualitative considerations, can be used to help one position, assess and improve the claims they can make on the social influence of online communcations by ensuring they measure all relevant confounders as much as possible and adjust for them. To demonstrate how this might be achieved in practice, the discussion in this section assessed the merits of practical attempts at reducing confounding and at accounting for qualitative aspects, both in observational and experimental settings, whether in online, offline or in mixed online-offline setups, covering quantitative and qualitative methods.

5.9 Summary

Overall, this chapter has proposed an individual-level instantiation of the ACF, for assessing the social influence (causal effect) of online communications on individual-level outcomes, covering the space of other types of causes that may lead to an observed action (outcome), namely similarity of personal traits, traits of the focal item, and external circumstances.

This Individual-level Causal Framework (ICF) makes the following contributions: it covers the space of types causes other than online communications that may be behind observed outcomes, and offers a flexible classification scheme for them; using causal methods, it finds that the social influence of online communications is confounded with the effects of each of these types of causes, hence each of them should be measured and accounted for; it considers and classifies key qualitative aspects of how different combinations of causes can lead to different qualities in the outcome. The usefulness and versatility of these contributions are demonstrated in an analysis using previous studies of online datasets from different real-world settings.

That is, methodologically and conceptually, the ICF uses the formal rules of graphical causal models, drawing upon robust causal assumptions about what types of causes might directly affect an individual's actions, which stem from well-established results from the social sciences literature. In merging computational rules with social science-based causal assumptions, this framework offers a promising interdisciplinary methodology of the type that is much-needed in computational social science. Drawing from social and computational disciplines, this framework then presented some important characteristics of the observed outcomes and the causal variables, which affect the nature, form and extent of the claims one can make on the social influence of online communications.

In more detail, given that previous studies of online social influence have generally discussed causes other than online social influence sporadically, with each study only addressing a subset of alternative causes, the ICF proposed here attempts instead to take a more comprehensive view of alternative causal factors, and to consider and address them systematically and formally. That is, in addition to what existing studies have done, the ICF proposed here contributes the following:

- It expands the discussion on online social influence, from online social influence versus homophily, or versus susceptibility, or versus cognitive factors, to the social influence of online communications versus the influence of all other types of causes, by offering a classification, or a partitioning, of other causes into classes, while allowing for flexible classification of any specific cause under these classes. This identification and classification of types of causes that might directly affect an individual's actions or decisions is based on well-established results from the social sciences literature (sociology, psychology, social psychology, cognitive science, social neuroscience, behavioural science, management and marketing).

- It further expands the discussion on online social influence, from focusing on data where the social ties between people are known, to data where social ties are not necessarily known, hence covering not only social influence from one's known social ties but also from people who may not be in one's social network. This is again supported by established findings from the social sciences.
- It uses causal methodology in order to systematically and formally reason about these classes of causes (alternative classes of causes), and to address any bias they may introduce, by using the formal rules of graphical causal models, and it:
  - Examines whether the alternative classes of causes introduce confounding bias, and it finds that online social influence is confounded with each one of these alternative classes of causes, and
  - Determines what is the minimal deconfounding set of these alternative causes that must be measured and adjusted for, in order to recover an unbiased estimate of the social influence of online communications on the outcome, both for cases where a social tie exists and for cases where it does not exist or its existence is unknown. It finds that the minimal deconfounding set in both cases contains variables from all of the alternative classes of causes.
- It considers a range of qualitative issues related to how different combinations of the social influence of online communications and the influence of these other types of causes can lead to different qualities in the observed outcome.

As a final contribution of the ICF, it has been demonstrated it may be applied in practice, by using it to evaluate the robustness of online social influence estimates (how much confounding has been successfully adjusted for, how much still remains, and what qualitative aspects have been examined) from a set of diverse social influence studies from the social science and computer science literature that employed a varied range of quantitative and/or qualitative methodologies.

Chapter 6

Collective-level Causal Framework, and empirical implementation

This chapter begins by proposing the *Collective-level Causal Framework (CCF)* in Section 6.1, i.e. the instantiation of the abstract framework (ACF) proposed in Chapter 4 for settings of collective action, i.e. where the outcome represents (the result of) a collective decision or action (rather than a single individual's decision or action, which was the focus of the ICF of Chapter 5). The CCF proposed here presents a set of principles for how the influence of online communications can be conceptualised and measured at the collective level specifically (something which has received little attention in the literature). It tailors the general principles of the ACF to collective-level analysis, by addressing issues that are specific to collective-level analyses, including: mapping any variables captured at the individual-level to collective-level variables; determining the appropriate unit of analysis (which is no longer individual people, as it was for the ICF); and offering a flexible classification scheme for possible confounding causes (causes internal to the collective setting, causes external to the collective setting, traits of the focal item). The CCF constitutes a contribution applicable to a wide range of collective settings, i.e. any setting where there is a record of collective outcomes and of online communications. It is generic and flexible, so as to be applicable to any setting where outcomes are produced collectively and where there is an online communications component whose influence on the collective outcome one wants to measure. Such settings may range from professional collaborations and formal projects, to the newer but increasingly ubiquitous kinds of projects in the areas of crowdsourcing and citizen science.

As discussed in Chapter 2.2, studies of online social influence have focused on analysing individual-level outcomes, and collective outcomes have received relatively little attention, even though settings of collectively produced outcomes are also very common (e.g. organisational or other professional settings, crowdsourcing projects, citizen science projects). At the same time, collective-level analyses require a different approach than individual-level analyses, as in the former case node-to-node (individual-to-individual) modelling is often not as useful, and choices need to be made in how any individually-captured variables (causes and/or outcomes) should be aggregated so as to be modelled at the collective level. Also, in collective-level analyses, there is no need to make the often unrealistic assumption that the actions and decisions of individuals are independent, which is frequently made in individual-level analyses in order to simplify the analysis and make it more tractable, as in collective-level analyses, actions and decisions are considered collectively, as a whole.

After first presenting the CCF in Section 6.1, in the following sections, this chapter proceeds to present the empirical application of this collective framework to a specific setting of collective action and collaboration, using publicly available real-world observational data from the W3C Provenance Working Group. This demonstrates the flexibility and real-world applicability of the CCF, showing how the CCF can be applied empirically, to a real-world setting of collectively-produced outcomes, using public data, covering concepts (e.g. outcomes versus causes), causal modelling, variable extraction from the data, and causal estimation formulae implementation. First, Section 6.2 presents the nature and context of this Working Group and describing the dataset. Then, Section 6.3 describes the graphical causal model for the causes and outcomes that will be studied in this setting, and then proceeds to describe the empirical design and implementation used for the application of the collective causal framework to this setting in Section 6.4.

The empirical findings and discussion from this empirical application of the CCF are presented in the next chapter, Chapter 7.

6.1 Collective-level Causal Framework

This section instantiates the abstract causal framework (ACF) presented in Chapter 4, for settings where outcomes are produced collectively, into a Collective Causal Framework (CCF). First, the applicability of this CCF to various collective settings is discussed. Then, the instantiation of each component of the abstract framework for collective settings is addressed in turn.

6.1.1 Applicability of the Collective-level Framework to different settings

This collective causal framework (CCF) is applicable to a wide range of settings of collectively-produced outcomes which involve some form of Web-mediated communication. In this thesis, the CCF is applied to a setting of professional collaboration, the

W3C Provenance Working Group, using public data that records the email communications of the group and its deliverables (specification documents) over time. As the W3C's website contains public archives of many other such Working Groups, the CCF and causal analysis presented in this thesis are readily applicable to these other Working Groups, as their email archives and document drafts are also publicly available and in the same format as the Provenance Working Group data analysed in this thesis.

The CCF can be applied, in a similar manner to the application presented here, to any organisational setting or project for which there is a similar digital record of online communications (e.g. emails, or conversations on other channels, such as Slack which is becoming increasingly popular for workplace messaging (Manjoo, 2015)) and of the collectively-produced outcomes and their drafts over time (e.g. text documents such as reports, spreadsheets, slides, or any other deliverable depending on the context of each setting). As such digital traces of interactions are increasingly available in organisations and professional projects (with email in particular being the standard form of professional communication, for many years now), the CCF can potentially by applied to a wide range of organisational settings, in addition to its applicability to the data of any other W3C Working Group mentioned.

In addition, alongside more traditional forms of formal projects and organisations, there is the growing field of crowdsourcing and citizen science projects, which are becoming more and more ubiquitous. Such projects often generate digital traces both of deliverables and of communications over time, hence being another promising area of application of the proposed CCF.

As discussed in Chapter 2.2, crowdsourcing is a sourcing model whereby individuals or organisations use contributions from Internet users to obtain needed services or ideas (Estellés-Arolas and González-Ladrón-de Guevara, 2012; Taeihagh, 2017). Citizen science, also known as crowdsourced science, is one type of crowdsourcing where members of the public help in scientific research by completing specific tasks (Bonney et al., 2009; Silvertown, 2009), with Wikipedia being perhaps the best known crowdsourcing platform. There are other crowdsourcing marketplaces (such as Amazon Mechanical Turk[1]) that connect crowdsourcing workers with tasks commissioned by companies or other institutions, however crowdsourcing workers work independently and usually cannot directly communicate with each other on such marketplaces.

Citizen science and crowdsourcing platforms often have an online communication channel for participants, and may also record outcomes over time. Therefore, this is another type of domain with exactly the kinds of data that the CCF is applicable to, where one could study the influence of online social communications on the quality or nature of collective outcomes produced, over time, and across contexts (e.g. sub-groups) as applicable. And indeed, such citizen science platforms have already attracted the interest of researchers,

[1] https://www.mturk.com/mturk/welcome

with existing studies on e.g. on how the social interaction functionality provided on crowdsourcing platforms may be associated with the quality of the collective outcomes produced (Tinati et al., 2015, 2016). In any particular setting, domain knowledge will help determine what actions and outcomes are most meaningful to study, and the CCF is flexible enough to accommodate this.

Overall, the CCF is applicable to a wide range of Web-mediated settings of collectively produced outcomes, from traditional organisational records such as emails and text document deliverables, to newer forms of collective production of outcomes through crowdsourcing or citizen science which are growing in popularity. The application of the CCF would allow an investigator to assess the causal effect of online communications on the outcome of interest, while accounting for confounding bias from other factors (e.g previous outcomes), and to compare its importance to the importance of other causal factors, in order to determine what are the strongest factors that affect the quality (or other properties) of the produced collective outcomes. Obtaining such an understanding could help improve productivity, outcome quality, or collaboration in a variety of Web-mediated collective settings.

6.1.2 Distinguishing outcomes from causes

Per the ACF of Chapter 4, one of the first things to determine is, out of the observable actions in the data from this setting, what are meaningful collective outcomes in this setting? Which of the observable actions or events should be considered outcomes, and which should be causes of the outcomes? This is important in order to determine what the outcome variables and what the causal variables one should study, and to ensure that the variables chosen to measure the outcome do indeed measure the actual outcome, or end-goal, of the given collective setting, as much as possible, rather than only measuring one of several possible causes of the actual outcome.

Care is needed in distinguishing which variables can serve as causes and which as outcomes, as in literature it is often the case that actions are taken as meaningful outcomes where at best they may be potential causes, or often ambiguous symptoms, of the implied meaningful outcome of interest. That is, it is common in the literature to interpret high volumes of online discussion about an idea as evidence of widespread *adoption* of that idea, where in fact actual adoption (the outcome) has not been measured, and online discussions could only be one possible causal factor of adoption, among others, like people's prior belief, and the environment in which they live. For example, in González-Bailón et al. (2011), the use of a protest-related hashtag on Twitter is assumed to indicate having joined the protest, when there is no evidence that all users of the hashtag are actually at the protest, or that they have declared they support the protest's premises; rather, they may be merely discussing neutrally, or even against, the protest. Or, in Ghosh and Lerman (2010b), news articles on the online platform Digg that are up-voted are assumed to be ‘influential’, but the action of up-vote is not evidence of an outcome of adoption or of belief in the topic or ideas of that news item, and at best it might indicate some level interest in that news item.

6.1.3 Collective outcomes

As discussed in Chapter 4, the investigator must first identify the outcomes of interest. In settings of collectively-produced outcomes, we are interested in considering outcomes at the collective level. Therefore, the investigator must begin by identifying the collective outcome, or outcomes, that are most important for their research goals, depending also on the nature of the setting under study.

That is, a meaningful outcome should be, or relate to, the actual output or end-goal in the collective context studied. For example, when studying records of online communication and of work-in-progress deliverables, in the context of a professional project, the outcome of interest may be the nature or contents of the group’s deliverable itself, or, more specifically, one or more features of this deliverable. In crowdsourcing cases, if considering Wikipedia as an example of a crowdsourced setting, the outcome is the article produced, so the outcome variable studied should relate to the contents of the Wikipedia article (or articles) studied, rather than the contributor’s online conversations while they are contributing to that article. Similarly, in a citizen science project, the end-goal is for each participant to help in completing an overall task (e.g. in an astronomy project, annotating sections of the night sky), so an appropriate outcome variable could be the contents, quality, or some other aspect of the final collectively-completed task; rather than what the participants talk about in their online conversations.

Contexts that are not explicitly collaborative are also relevant here, i.e. settings where people do not necessarily work with each other to achieve a common goal, but where the outcome is ‘collective’ nonetheless. For instance, one may want to study whether enough people sign an online petition, after there has been a lot of online discussion about that petition – the decision whether to sign is individual, not collective, but whether a threshold of signatures is reached (as a measure of the petition’s success) is a collective outcome. The online discussion is one possible cause of the petition’s success or failure, not the end-goal itself. Similarly, one may want to study whether a given political candidate gets elected, given the relevant online discussions about them, among other causal factors – again, the online discussions and their features (e.g. volume, sentiment) are one causal factor, not the end-goal itself.

Having clarified which actions or events can be considered causes versus which can be considered meaningful outcomes, one next step is to determine at what level the collective causes and outcomes should be measured, or how individual-level variables should be aggregated into collective ones. For example, should variables be aggregated over all people involved, or only over some specific subset of those people? Are people naturally split into sub-groups in the setting under study (as in the Working Group under study here, where each document has its own group of people working on it)? Similarly, if data is available on online communications and outcomes over time (as is the case for the Working Group data studied in this thesis), the investigator should determine if and how the analysis (and variables) should be split into time periods. For instance, for the Working Group data studied here, the duration of each document's lifetime will be split into intervals between subsequent drafts (inter-draft intervals), and a causal analysis will be performed for each of these drafts, with the final draft of each interval being the initial draft of the next interval. Once these aspects have been determined, what quantity should be used to aggregate individual-level actions into a single collective one? For instance, should a collective variable be the average of individual-level variables, or their sum, or some other quantity?

6.1.4 The social influence of online communications on collective outcomes

Having established what the collective outcomes of interest are, the investigator must next turn their attention to measuring the effect of online communications on those outcomes. It must be determined what aspect (or feature) of online communications can be measured (e.g. their volume, duration, number of participants, sentiment expressed).

Similarly to the outcome variable(s), for the online communications causal variables (or variable) one must determine whether they are measurable at the collective level in the data, or how to aggregate them if they are captured in the data at an individual level rather than at the collective level, and if data is available on online communications and outcomes over time, how these variables will be measured over time.

In addition, if more that one channels of online communication are used and recorded in the available data, the investigator must determine whether the effects of each channel will be studied separately, or whether and how data from different channels can be integrated (as different communications channels or platforms allow users to take different types of actions, and impose different constraints) in order for the effects of all online communications channels to be analysed as one.

6.1.5 Other causes of collective outcomes

In addition to the causal factor represented by online communications, one must consider other causal factors that may also affect and shape the outcome(s) of interest, as discussed in Chapter 4. Taking such factors into consideration is vital, in order to then examine whether leaving them unmeasured may result in bias being introduced into the estimate of the effects of online communications on outcomes.

In this thesis, the CCF will be applied exactly in this way, to analyse online communications and their influence on collective outcomes, and particularly whether this influence can safely be studied in isolation, or whether there are confounding causal factors that introduce bias to estimates of this influence if left unmeasured, and what is the extent of confounding bias introduced by these factors, as well as whether the effects of these factors are stronger or weaker than the effects of online communications on the outcome.

In general, these other causes may lie within the group studied (e.g. they may be the previous outcomes, if data at previous time points is available, as it is in this thesis), or they may lie outside the boundaries of the group, in the external environment, i.e. be external developments that may affect the group's deliverables. For example, in the context of a professional collaboration setting, where the deliverable is a product, a scenario where a competitor launches a similar product to the one a team is working on, or where new legislation is introduced which affects the requirements the end product or deliverable must satisfy.

Again, as above, the investigator must determine at what level to measure these additional causal factors, how to aggregate them if they are captured in the data at an individual level rather than at the collective level, and if the data captures online communications and outcomes over time, how these variables will be measured over time.

6.1.6 Causal modelling

Having established what the outcome variable is, what the causal variable for the online communications is, and what other causal factors may be relevant, one must now put all those together in a causal model. As discussed in Chapter 2.6 and Chapter 4, a graphical causal model, in the form of a directed acyclic graph (DAG), helps an investigator to visually determine by looking at the paths in the DAG whether the causal effect of of online communications on the collective outome is identifiable or not, i.e. whether there are any unmeasured confounders or not, and to determine the minimal set of such confounders that it is sufficient to adjust for in order to recover the unbiased estimate (i.e. to *identify*) of the causal effect of online communications on the outcome.

Therefore, the investigator can now represent the outcome and causal variables as nodes in the causal DAG, with arrows pointing from causes to their outcomes. Then, confounding variables can be identified by looking for *backdoor paths*, as explained in Chapter 2.6, and one can then find the minimal set of variables that can block this path, and hence remove confounding bias.

In practice, it may be that not all possible confounders can be measured, as in real world social settings there is likely a myriad of factors that may be a common cause of observed actions, and it is normally not possible to measure all of them. Instead, what can be done is to try and pick apart the layers of bias one by one, and to see if any of the confounding bias can be removed, i.e. if any of the confounding variables can be measured and adjusted for, while acknowledging that confounding from other unmeasured sources might remain.

This enables one to then establish how much confounding bias each confounder introduces, as adjusting for confounders can be computationally expensive (the state space increases exponentially with the number of confounders, as noted in Bengio and Bengio, 2000; Pearl, 2009a), to determine whether any of those confounders introduced only negligible bias and can hence be safely ignored from the analysis, or whether any introduce very large bias and should be accounted for.

Particularly for social influence online, it is commonly assumed (as discussed in Chapter 3), without empirical testing, that the influence of online communications on outcomes can be studied in isolation, without accounting for any causal factors. So using this CCF can enable an investigator to *empirically test* whether this claim holds, in the given collective setting under study. In this thesis, as we shall see, it is found that this claim does not hold, in the collective setting of the W3C Provenance Working Group: one confounder (the previous outcome, i.e. the previous document draft) introduces very large bias to the causal estimates of social influence on the outcome (the next draft of the document). In fact, ignoring this causal factor would mean missing out on a causal factor much stronger and much more stable over time than the online conversations, as it is found that the confounder (previous outcome) has itself a much larger causal effect on the next outcome than the online email conversations, which also tends to increase over time unlike the effects of the email conversations that do not display a pattern as steady. All this will be discussed in Chapter 7.1. So, this CCF enables an investigator to pick apart the effects of confounders, determine whether the amount of confounding bias is negligible or not, and to compare the effects of different causal factors on the outcome to determine what the strongest factors may be.

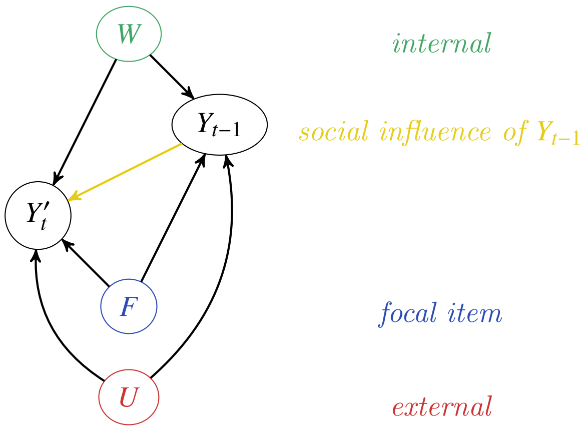
Again, as discussed, the investigator must determine at what level to measure these additional causal factors, how to aggregate them if they are captured in the data at the individual level rather than at the collective level. If the data covers online communications and outcomes over time, the investigator must determine if and how the data might be split into time intervals and how variables will be measured over time.

One aspect of the modelling and the analysis that is worth noting is the choice of the units of analysis, or entries, or datapoints, to which the variables pertain, and over which one may be calculating effects (e.g. average effects over all units). For instance, for the W3C Provenance Working Group in this thesis, the units of analysis are terms (words), so each causal variable and outcome has a value for each terms, out of a body of many terms (that is, out of a *term corpus*), at each time interval, and the effects obtained represent effects over all terms on average, or for the average term. I.e. the analysis in this thesis is about the factors that affect whether a word becomes part of the formal vocabulary of terms, and of the language used in the narrative of the documents, produced by this Group (this shall be discussed in more detail in Section 6.4.1).

So, the units of analysis in this chapter and in Chapter 7 will not be individual people, as is common when in individual-level analyses of online social influence, and as was the case with the ICF in Chapter 5. That is, in general, studies about online (and offline) social influence tend to be about how communications about a specific topic affect a person's beliefs or actions with regards to this topic, for the average person studied (e.g. Aral et al., 2009; Berger, 2013; Cha et al., 2010; Christakis and Fowler, 2013; Ghosh and Lerman, 2010b; Kempe et al., 2003; Kwak et al., 2010; Shalizi and Thomas, 2011). However, this is not the case for the collective-level causal analysis here. Similarly, in citizen science, the units of collective-level analysis might be the overall tasks to be completed, and how properties of online communications affect the quality of the average completed task. On Wikipedia, the units may be individual articles, and how properties of online communications affect the quality, contents, or other aspect of the produced articles, on average. For example, in a different setting, a study on estimating the effect of aggregate social signals on music downloads (Salganik et al., 2006), the units of analysis were the individual songs (the social signal displayed to each user was a collective social signal, specifically the total number of previous downloads of each song by others).

As mentioned in the previous sub-section, when measuring the effect of a group's online communications on a collectively-produced outcome, possible confounding causes may include factors internal to the collective setting (e.g. the organisation's or group's previous outcomes, the internal structure, the internal culture; e.g. Barnett and Carroll, 1995; Hannan and Freeman, 1984; Morgan et al., 1997), factors external to the collective setting (e.g. social norms and trends, the economic, political, or technological environment, the broader status quo in the broader field or professional sector in which the organisation or group belongs, another organisation's competing product if studying a commercial setting; e.g. see Barnett and Carroll, 1995; Davenport and Prusak, 1997), as well as traits of the focal item (the unit of analysis) under study (e.g. its perceived risk of adoption, its perceived value, quality, or novelty). This classification of confounders follows a similar rationale to the respective classification for the ICF in Chapter 5, and the resulting graphical causal model is depicted in Figure 6.1, which uses similar symbols and colours as Figure 5.5 from the ICF of Chapter 5, for continuity and ease of comparison.

That is, in Figure 6.1, the collective outcome at time t is represented by Y'_t, while the online communications at time $t-1$ (whose influence on the outcome one wants to estimate) are represented by Y_{t-1}. That is, the same symbols are used as for the ICF (Figure 5.5 of Chapter 5), but without person-specific subscripts (i, j) here, as here we are not interested in individual-level outcomes or individual-level online communications, but rather on collective-level variables. Similarly, symbol U denotes here variables external



(a) Collective-level graphical causal model

(b) Context

Figure 6.1: Graphical causal model for collective-level analyses (a), with the legend (b) on the right showing the context of each latent causal variable and of the causal relation of social influence

to the collective setting being studied, where in Chapter 5 it represented factors external to the two individuals (i, j) being studied. Symbol W here represents factors internal to the collective setting, where in Chapter 5 it represented factors internal to the two individuals. Finally, symbol F, both here and in Chapter 5, represents the traits of the focal item – only here the focal item is the unit of analysis (as discussed above), rather than an individual (j) being the unit of analysis, which was the case in the ICF of Chapter 5.

6.1.7 Causal estimation

Per Chapter 4, in addition to the causal DAG model, the investigator must decide what kind of causal estimation process will be used. One may choose to estimate causal effects nonparametrically, e.g. using the Average Causal Effect formula in Equation 2.5, or parametrically. In the latter case, a common choice is to assume a linear relationship between outcome and causes, with 0-mean Gaussian and independent errors (Pearl et al., 2016). However, such assumption are strong and may not always hold for any given dataset, so it is important that it is tested that the the dataset does not violate any of the parametric assumptions of a linear (or other) model.

If a nonparametric approach is chosen, one does not have to use the difference-based ACE, but could use ratio-based estimators, or other measure, as desired and as appropriate (Morgan and Winship, 2014; Pearl et al., 2016). Nonparametric estimation has the advantage of not requiring any assumptions on the form of the relation (e.g. linear) between the outcome and the causes. However, it may be intractable to estimate if there are many confounders to adjust for; in such cases, parametric estimation may be necessary (Pearl, 2009a). Other estimation methods involve matching methods and propensity-score based methods, noting that one must still first measure and account for confounders when using those, as they are merely efficient estimator methods and not methods for circumventing confounding (Pearl, 2009a), as discussed in Chapter 2.6.

For causal estimation in general, including for the nonparametric ACE formula (Equation 2.5), the causal variables are expected to be binary, so one must binarise the causal variables that are not already binary, using a threshold. In the analysis in this thesis, the median is used as the binarisation threshold for each variable, as the mean is sensitive to extreme values. Using binary causal variables is common in causal analyses, also when other estimation methods are used (e.g. matching, in Tsapeli and Musolesi, 2015), as causal analyses can be thought of as simulated experiments where a 'treatment' (causal variable) is present in (or administered to) some subjects and absent in others (treatment and control groups).

For the collective empirical analysis in this thesis, non parametric estimation is tractable, because even though there are many people involved in the group, all variables are collective (rather than having one causal variable per person), and there are relatively few confounders to adjust for, so there is no need to resort to e.g. a linear parametric approximation. In general, in collective settings where individuals' actions are aggregated into collective causal factors, and especially in settings where there are relatively few collective-level confounders to adjust for, nonparametric methods may be more applicable (i.e. more tractable) than in individual-level analyses with many participants where one may need to adjust for confounding from each individual person.

6.2 The dataset: W3C Provenance Working Group online archives

The dataset used in the empirical analysis of this thesis comes from the public online archives of the World Wide Web Consortium (W3C) Provenance Working Group.[2]

As discussed, this dataset offers the advantage that the outcomes of interest are captured in the data, rather than assuming unmeasured adoption outcomes can be inferred from proxies of questionable reliability, as is commonly the case in the contagion-based paradigm literature. An additional advantage of this dataset is that outcomes are captured over time, and also across contexts (different sub-groups of people working on separate deliverables), enabling one to investigate how the influence of online communications may vary over time, and across different contexts, dimensions which are often not considered in studies of the social influence of online communications. In addition, W3C Working Groups are settings of international collaboration, and lead to the production of Web standards that can shape Web practices and usage worldwide. Therefore, it is

[2] https://www.w3.org/2011/prov/

interesting to investigate the causal factors that shaped the nature and content of these global-reaching standards.

In more detail, the World Wide Web Consortium (W3C) is ‘an international community where Member organizations, a full-time staff, and the public work together to develop Web standards’.[3] The W3C Provenance Working Group represents one such standarisation effort, which defined a standard for provenance information interchange on the Web, and took place in 2011-2013. One of the main goals of the Working Group was to develop a common vocabulary or language for provenance on the Web (Moreau et al., 2015).

As described in Section 3.2.1, such Working Groups are settings of collective decision-making and collaboration, where domain experts in a topic area participate in order to produce a set of standardisation documents, in this case on the topic of provenance on the Web. Participants in the Provenance Working Group included academics, industry professionals, and staff of the W3C. The Working Group had an internal organisational structure, where two members were the co-chairs of the group, two members were W3C staff contacts, some members were editors of or contributors of the standardisation documents produced, and the rest where members with no additional formal ‘position’. The Working Group had 57 members in total.[4]

The Provenance Working Group had its own mailing list for its participants to discuss with each other, and the W3C has made the archives of the emails sent to this mailing list publicly available online.[5] The email archives contain 8,819 emails, making up 2,017 threads.

The standardisation documents produced by the group, and their published draft versions over time are also publicly available online.[6] The group named the standard they produced ‘PROV’, and it is comprised of the following twelve specification documents: AQ (Access and Query), Constraints, DC (Dublin Core), DM (Data Model), Dictionary, Links, N (Notation), Ontology, Primer, Overview, Sem (Semantics), and XML. Of those, four documents (Constraints, DM, N, Ontology) are full Recommendations, and the remaining are Notes. Work did not start at the same time for all documents, and not everyone was involved in all documents. Each document had one or more editors, and usually one or more contributors (their names and affiliations are listed at the top of each document page, for each draft). The convention adopted in this Working Group was that only document editors could directly modify the document drafts. Hence, even if several others contributed ideas, members’ individual contributions cannot be traced directly to the document drafts. The contributors and editors of some documents varied

[3] From the W3C’s official webpage: https://www.w3.org/Consortium/

[4] As per https://www.w3.org/TR/prov-overview/#acknowledgements

[5] At https://lists.w3.org/Archives/Public/public-prov-wg/

[6] The URLs of each document can be found at https://www.w3.org/2011/prov/wiki/Main_Page#Specifications. On the webpage of each document, the previous versions can be accessed by clicking the ‘Previous version’ link.

over time, across drafts. The chairs of the whole Working Group remained the same throughout.

In addition to the Provenance Working Group's documents, this analysis also considers the Recommendations of the PROV Charter document, published in 2010 by the Provenance Incubator Group, which outlined a skeleton of topics intended to be covered by the Working Group.[7] The editors and contributors Incubator Group credited in the Charter were all also members of the Working Group except one person. The concepts and topics proposed in the Charter could be addressed across all the documents produced by the group, hence each document contains some elements of the Charter, so the Charter is included in the analysis of each document. It is also noted that the Charter (or, rather, the part of the Charter that is included in the dataset) is much shorter in length than any of the document drafts produced by the Working Group. So, the Charter has a much broader and less detailed scope than each of the documents, and it is also much shorter than any of the documents. It is worth noting the different nature of the Incubator Charter versus the Working Group documents, particularly in the context of one of the important goals of the Working Group which was to develop a common vocabulary or language for Provenance. The Charter is only a short skeleton covering a breadth of key concepts and topics, while each of the Working Group's documents only covers one narrow topic area, and it covers it at length, likely including many concepts and specific terms, which will not be in the much shorter and broader-scope charter. This is important as the causal analysis will focus on the factors shaping whether a term becomes part of the commonly used vocabulary of each document as its drafts develop over time, starting with the Charter, so it is likely that topic-specific terms that appear frequently in the vocabulary used in the topic-specific and lengthy document drafts may not appear much in the more skeletal and broader-scope Charter document.

6.3 Graphical causal model

This section presents the causal diagram that will be used in the causal analysis of the W3C Provenance Working Group online archives, where the online communications take place over email, and the deliverables are documents, whose drafts over time are recorded in the dataset. The goal is to analyse the factors affecting whether a term becomes part of the formal vocabulary defined by the Working Group, and of the language and exposition style used to define and explain that vocabulary, with the focus being determining whether the effects of online communications on this outcome are unconfounded by other causes (as is assumed, without empirical testing, in the contagion-based paradigm for online influence), and to compare the relative importance of the various causal factors.

[7] This can be found at https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/#Proposed_Charter_for_a_Provenance_Interchange_Working_Group, and only paragraph 4, titled 'Provenance Concepts' is used for the analysis, as that is the part of the Charter that outlines the concepts and topics to be addressed by the Working Group and that is most relevant to the Working Group's documents.

For the causal analysis of the effects of online communications (emails) on outcomes (documents), accounting for other causal factors (previous drafts of documents), the causal diagram of Figure 6.2 is used. It represents one inter-draft interval, and the variable symbols represent the following variables:

- I: Presence of a given term in the top most frequent terms of the previous draft, i.e. of the *initial* draft in the context of this interval;
- F: Presence of a given term in the top most frequent terms of the current draft, i.e. of the *final* draft in the context of this interval;
- S: Sentiment in the email communications relating to a given term, between the publication times of I and F;
- P: Participation in the email communications relating to a given term, between the publication times of I and F.

That is, in relation to the general causal DAG of the CCF, shown in Figure 6.1 of Section 6.1, variable F here is an instantiation of variable Y'_t in Figure 6.1, variables S, P stand for (features of) the online communications represented by Y_{t-1} in Figure 6.1, while variable I can be considered part of the class of factors internal to the collective setting, represented by symbol W in Figure 6.1.In terms of the other confounders in Figure 6.1, there do not seem to be any external factors (U in Figure 6.1) relevant to the efforts of this Working Group that changed during the lifetime of the Group. Regarding focal item traits (F in Figure 6.1), in this context there do not seem to be any particular traits that one may extract or infer from the data. Besides, the goal here is not to cover all possible confounding causes; rather, the aim is to test the assumption of the contagion-based paradigm that there are no other causes one should account for when measuring the influence of online communications – it suffices to finding substantial confounding bias from one cause to refute this.

Details for how the top frequent terms are determined are provided in the Section 6.4 and Appendix C.

In Figure 6.2, because variables S and P (and S_{-1}, P_{-1}) have the same incoming arrows and the same outgoing arrows, and have no arrows to or from each other, they are shown together in the same node to avoid unnecessarily complicating the DAG, but they are treated as separate variables in the analysis. In the DAG, time flows from top to bottom. As each inter-draft interval is analysed separately, the focus here is on the interval starting at the draft corresponding to I and ending at the draft corresponding to F, with S, P corresponding to the emails between them. S_{-1}, P_{-1} correspond to the emails in the interval immediately preceding this one, and the DAG extends into the past and into the future similarly, for each prior and future inter-draft interval.

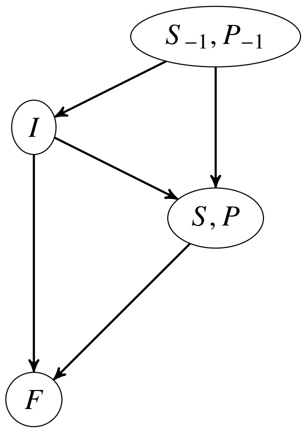


Figure 6.2: Causal DAG for the W3C Provenance Working Group

The main focus of the analysis here is the causal effect of S,P (email sentiment, participation) in the email communications on the outcome (current draft, F). The causal diagram shows that this is confounded, due to I: there is a backdoor path, that can be blocked on I. That is, I is a common cause of S, P ($I \rightarrow S, P$) and of F ($I \rightarrow F$), hence I introduces confounding bias on the estimates of the effects of S and P on F. Measuring and adjusting for I, i.e. blocking on I, blocks this backdoor path, and removes the corresponding confounding bias from the causal estimate of S and P on F. Therefore, the minimal deconfounding set, Z, here is $Z = \{I\}$.

Next, one needs to determine the extent of this confounding bias – can it be ignored, i.e. is its magnitude negligible, or not? This is important because the contagion-based paradigm for the influence of online communications assumes, without empirically testing, that any other causal variables can be safely ignored.

Once the extent of confounding I introduces on the estimates of the effect of S and P on F has been studied, it is also of interest to study the effect of I on F, i.e. to study the confounder I itself. For that effect to be identifiable, the backdoor path between I and F must be blocked. That is, the path $I \leftarrow S_{-1}, P_{-1} \rightarrow S, P \rightarrow F$ must be blocked. For that, the minimal deconfounding set, Z, to block on consists of S_{-1} and P_{-1}: $Z = \{S_{-1}, P_{-1}\}$.

It is noted that there may also exist other unmeasured confounders, which are not explicitly modelled in the DAG to avoid cluttering and confusion. That is, there may be one or more hidden common causes behind any or all of S, P, I, F. For instance, at some point, consensus may have been reached between editors and contributors of a document on which kinds of terms should and should not be discussed extensively and should and should not appear extensively in the documents, i.e. some kind of consensus limiting the scope of the email discussions and of the documents. This hidden consensus may justify any observations of terms that have all or many of of S, P, I, F with values 1 (high), as those may be terms that have been decided to be in scope, and may similarly justify any terms that have all (or many of) S, P, I, F with values 0 (low), as those may be terms that have been decided to be out of scope, i.e. not worth discussing in extensively or mentioning extensively in the documents. Based on the dataset used, and the automated analysis conducted, no evidence of such a scenario has been produced, and indeed one could argue that any member of the Working Group was in principle free to email the others about anything they deemed worth discussing, as extensively as they wished, and to express their own perspective on matters. Still, without further investigation, this scenario cannot be ruled out either. As another example, one kind of hidden common cause might exist in instances where the Group agreed upon a deadline by which discussions on a given topic should conclude. In such instances, the agreed deadline might be a confounder. However, this time restriction did not generally impose any additional limits on how much that topic should be discussed, and did not predetermine the outcome of the discussion in terms of whether the topic would or would not end up featuring heavily in the produced documents. Therefore, without further investigation, it is not clear whether such deadlines might be hidden common causes behind the volume or sentiment or email discussions and inclusion or exclusion of a topic or term from the documents. Overall, it is acknowledged that hidden common causes such as the above, or other kinds of hidden common causes, may still exist.

This limitation is not specific to this dataset only; rather, in general, in analyses of real-life settings of human interactions one cannot ever measure all possible common causes of the involved variables. The focus here is testing the assumption of the contagion paradigm for online social influence that the influence of online communications can safely be measured in isolation, without regard or adjustment for other causal factors. For this assumption to be refuted, it is sufficient to find one causal factor that introduces non-negligible confounding bias to the estimates of the influence (causal effects) of online interactions. Therefore, the focus here is on removing bias from one known and measured confounder, and on then extending the analysis to also studying the effects of this confounding causal factor on the outcome of interest.

It is assumed that there can not be any causal arrows between S and P in any inter-draft interval. This is sensible, as both features pertain to the exact same emails: a value for sentiment and participation is calculated for each email, and then those are aggregated for all emails. It is assumed that the length (participation) of any given email does not cause the sentiment of the same email, and that the sentiment of a given email does not cause its length. That is, it is assumed that one might write a long email (high participation value) either because they want to say something positive, or because they want to say something negative, or because they want to say something in neutral sentiment that they find interesting, or worthwhile. Similarly, one might want to say something very negative and do so in a short email, and so on. So it is not obvious why one kind of sentiment should cause (or be caused by) one level of participation, and another kind of sentiment should cause (or be caused by) a different level participation.[8]

[8] As we shall see in the results of the analysis (Section 7.1), the effects of the S and P variables used here are correlated, but this is because the machine learning classification algorithm used for the S measure tends to give higher sentiment values for longer emails (i.e. for higher values of P), even if

In addition, the analysis does not consider the effects of emails on subsequent emails, i.e. the email domain is not considered to be the outcome domain at all. That is because the most important and consequential outcomes the group produced were the documents themselves, and the data for the documents (and their evolution over time) is publicly available. So, even though it is common in studies of the influence of online communications (e.g. on social media) for the outcomes to be in the same domain as the causes (e.g. outcomes that are recorded on social media, such as pressing the 'Like' button, or commenting, or re-sharing a post, or using a URL or a hashtag in a post or message), which are often used as proxies for outcomes for which there is no data available to the investigators (e.g. proxies for adoption or endorsement), since there *is* data on the *actual* outcomes of interest in the setting under study here, this data is used.

Participation is chosen as a measure of the nature of the online communications, in order to determine whether terms that are discussed the most are also the ones that will feature most prominently in the documents produced. Rather than counting only the number of emails, the length of the emails is measured (number of words per email), in order to include information on how much discussion, how many words, went into talking about each term. Such quantity-based, volume-based measures are common in the literature on online social influence. Common measures include the volume of response to a user's posts on social media (e.g. Bakshy et al., 2011; Cha et al., 2010; Ghosh and Lerman, 2010b; Kitsak et al., 2010; Kwak et al., 2010), or the amount of participation in conversations about a topic (signalled by the inclusion of the relevant topic keyword, e.g. a URL or hasthag, in one's social media posts, e.g. Bakshy et al., 2011; González-Bailón et al., 2011). For instance, it is common for studies of Twitter or other social media platforms to focus on studying people and/or topics with the most responses on the premise (or assumption) that high levels of response make those people or topics 'influential'. Here, we test whether those high-online-participation topics are in fact influential with respect to the outcome of interest (featuring frequently in the documents produced by the Group). That is, care is taken to avoid the begging-the-question or circular reasoning fallacy, i.e. the fallacy of assuming that which is to be proven (i.e. that high online participation causes a topic to be more likely to be adopted, i.e. appear frequently, in the final outcome) rather than proving it. Instead, this thesis actually tests empirically whether high email participation around a topic (a term, here) causes it to be more successful in being frequently used in the documents produced by the Group, and tests the extent to which online participation is the root cause of that outcome rather than other (confounding) causes (by appropriately adjusting for measured confounders). As discussed, this is an important issue to test, and testing such claims empirically, rather than assuming them, has been repeatedly called for in the literature (Ackland, 2013; Bright et al., 2017; Freelon, 2014; Tufekci, 2014). Moreover, Bright et al. (2017)

the content is neutral. That is, this is a limitation of that particular classification algorithm, as will be discussed in Section 6.4, not a limitation of the causal model.

stress the importance of causal empirical testing of the effect of the volume of online communications on actual outcomes (in their case, in the context of political campaigns). They note how there is still little systematic empirical evidence that heavy social media campaigning (heavy posting on online social media by politicians) has actual impact on campaign outcomes (whether a politician is elected), and how even then it is difficult to control for hidden common causes of both social media use by the candidate and election of the candidate, and that strong causal claims are difficult to make.

Sentiment, then, is chosen here as more qualitative complementary measure to the volume-oriented Participation measure, as an additional way of measuring the contents of the emails, so as to measure not only how much something is discussed, but also how positive or negative are the attitudes expressed in those online discussions. As shall be discussed in the next section, Sentiment is calculated using a neural network.

6.4 Empirical design and implementation

This section presents the design and implementation details of the causal analysis procedure of the W3C Provenance Working Group archives. It first describes how the data was prepared and cleaned, and how features were extracted, both for the document drafts data and for the mailing list data. It then proceeds to describe what causal estimation methods and measures were used, and the properties of those causal measures.

6.4.1 Data preparation and feature extraction

The W3C Working Group dataset used in this thesis can be thought of as consisting of two modalities: the online communications modality (emails) and the formal documents modality (document drafts). The former modality could be considered as the online 'talk' domain, and the latter modality as the formal 'action' domain, as the formal outcomes (document drafts) lie in the latter modality.

In the causal analysis of this dataset, the main goal will be the estimation of the causal effects of the online talk domain (emails) on the actual formal outcomes of the group (documents), over time, as explained in the causal model discussion of Subsection 6.3. As discussed, the features of the online communications that will be studied are the overall sentiment and participation, while, for the documents produced, the top frequent terms across drafts will be used. Table 6.1 presents a summary of the data and the features, or causal factors, that will be studied.

One of the first steps of the analysis is to determine a body (*corpus*) of words (or terms) of interest, which will be the units of the analysis here: how is the presence (or absence)

Table 6.1: Summary of Data and Features (Causal Factors)

Modality	Description	Features (Causal Factors)	Measures
Emails	Approx. 9,000 emails, 1,200 threads	Attitude towards term discussed	Sentiment (Pre-trained Deep Learning classifier, probability that sentiment is positive)
		Extent of participation, discussion, debate	Participation (Total number of words contributed per person per term)
Specification documents	12 documents, each approx. 2 editors, approx. 10 contributors	Contents, nature of the documents	Top most prominent (frequent) terms per document

of each term in the documents shaped by how it was spoken about in the online email communications.

Words (terms) were chosen as the unit of analysis given the context and goal of this Working Group: one main goal was to develop a *vocabulary* or *ontology* (Moreau et al., 2015), which are terms from the Semantic Web technologies research area of Computer Science. Per the W3C, 'On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern' and 'There is no clear division between what is referred to as "vocabularies" and "ontologies". The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense.'[9] In more detail, for this particular Working Group, as stated in Moreau et al. (2015, p. 241), 'The overarching approach adopted by the Provenance Working Group is to consider an (extensible) core provenance language that allows any provenance model to be translated into such a *lingua franca* and exchanged between systems'.

Therefore, part of the goal of this analysis is to study whether a given word succeeded in becoming part of the formal language, of the formal vocabulary of terms (concepts and relations) developed by this Working Group, in terms of being used frequently in the produced standardisation documents. In addition to the formal language that is part of the formal vocabulary or ontology defined by this Group, the term corpus in this thesis also includes any other words used frequently in the documents, i.e. in the narrative and exposition style the Group uses to talk about and explain the produced ontology, e.g. words employed in frequently used examples, or words that are not in the ontology but are employed frequently to define or explain the meaning of formal concepts and relationships that are in the ontology.

To help illustrate this, Appendix C provides more information on the content and goals of the provenance standard produced by this Group. It also shows which words in the term corpus of two of the Recommendation documents (DM and Constraints) represent

[9] From https://www.w3.org/standards/semanticweb/ontology, accessed 13 April 2018.

elements of the formal vocabulary defined by the Group, and which represent elements that are outside the formal language and instead related to the exposition and narrative (e.g. definitions and explanations used for these formal elements, words used to refer to the Group and its organisation, words used in frequently employed examples).

Hence, when using the words ‘term’ and ‘term corpus’ in this thesis, it is not in the sense of the W3C’s use of ‘term’ cited above (a formally defined entity or relation that is part of a Semantic Web vocabulary or ontology), but rather in the sense of frequent words that are employed in the text of the Provenance Working Group’s documents. Hence, the analysis in this thesis is about how email conversations affected whether a word became part of the formal vocabulary or of the language used in the narrative of the documents produced by this Group.

Frequent words for the term corpus are only extracted after first having removed non-informative sections (e.g. acknowledgements, bibliography, table of contents) and numbers, lemmatized the remaining words, and removed all stop words (English function words with no particular inherent meaning) from the document drafts, as described in Appendix B.1. Using the top most frequently occurring terms (words) in the documents also serves to distil the contents and nature of the documents, to separate signal and noise, and to reduce the dimensionality of the documents.

As there is no external standard of domain knowledge to determine the body (or corpus) of terms that were important at each stage (draft) of each document and should be studied, this term corpus is constructed based on the dataset itself. As each document has several drafts, and each of those drafts is studied as an outcome, when studying each draft, the terms corpus is made up of all terms that were in the most frequent terms of any of the drafts of this document up to and including the current draft *and* that also appeared in at least one email subject line out of all emails sent up to the publication of the current draft. Therefore, for every draft studied as an outcome, the terms in the relevant term corpus have been prominent in at least one of the drafts published this far and been in the subject line of at least one of the emails sent this far. Considering also the emails serves as a cross-check that any term that was frequent in the drafts was not just draft-specific noise, but was actually also part of the topic of the group’s email discussion in at least one instance. So, the unit of analysis here is a *term*, with the outcome being whether a term appears frequently in the next draft of a given document or not, over all terms on average.

That is, terms (words) are not the cause, but just the unit of analysis, just an organisational unit for analysing the social influence of online communications (emails) on the contents of the document drafts. So, for measuring the social influence of online communications in the context of the W3C Provenance Working Group, the analysis is organised around terms: for the average term, how do online communications (particularly, two measured characteristics of them: sentiment and participation) affect whether the average term will feature frequently in the document drafts?

It is worth noting that, in the construction of the term corpus, a relevant issue is the phenomenon of selection bias, in the sense of bias that is introduced in an analysis of the occurrence of an outcome because an investigator is selecting only cases where the outcome has occurred (this is often called 'selecting on the dependent', i.e. filtering the data based on the values of the dependent, or outcome, variable; e.g. see the discussion in Tufekci, 2014). In the term corpus construction design described above, terms are chosen such that they have appeared in at least one previous draft ($F = 1$ in this or in a previous inter-draft interval). This design means that, at any inter-draft interval, the corpus will include both terms prominent in, and terms not prominent in, the draft at the beginning and at the end of this interval (I and F will have 0 values and 1 values). That is, this design does not exclude from study terms that are absent from the top prominent terms of drafts; rather, it includes terms that are absent and that are present in the set of prominent terms, and imposes only a minimal constraint, that the terms in the corpus have been present in the set of prominent terms for at least *one draft* up to the current time point. Therefore, this design does not limit the values that the relevant variables can take, in contrast to what happens in selection bias (where the outcome variable only ever takes one value; in our case, this would mean $F = 1$ for all terms in the corpus, for all inter-draft intervals). Further, given that there is no body of external domain knowledge based one which one might populate the term corpus, this design is a sensible and meaningful way to construct a term corpus, while keeping in mind the issue of selection bias, and minimising its presence as much as possible, by requiring that F=1 only once, for one draft up to the current time point, and by ensuring that, at every inter-draft interval, the corpus contains terms that may be absent or present from the two drafts at either end of the current inter-draft interval. This is a meaningful design for the constructing a term corpus, as it reflects the rationale where, as the Working Group's procedures are unfolding, one wants to keep track of the terms that were ever prominent, to see if they will continue to be prominent in future drafts, or if they will be replaced by new terms which may become prominent at later stages.

6.4.1.1 Different designs for constructing the term corpus

As discussed above, the term corpus is a set of terms that are deemed important to be studied, in order to understand the factors affecting whether the average term becomes part of the common vocabulary and narrative used by this Working Group (i.e. whether a term is used frequently in the documents produced by the Group), and how this varies over time. As there is no external agreed-upon standard or benchmark for determining which terms are important in this context and should be analysed over time, other than the contents of the document drafts themselves, prominence in the document drafts themselves is used as an indicator that a term was, at least at one point, considered important by the Working Group, and hence should be studied. So, the term corpus is constructed based on the frequency of terms in document drafts.

In addition to how the term corpus is constructed in the main analysis (to contain the top frequent terms from the drafts up to the draft at the end of the current time interval), this thesis also presents, in Section 7.3 and Appendix E, the results of two further analyses, each using a concept corpus constructed differently, as follows. Subsection E.1 presents an implementation design where the corpus used at each time interval contains the top frequent terms from *all* drafts (including future drafts) of a given document, rather than containing only the top frequent terms from the drafts only up to the draft at the end of the current time interval. Then, subsection E.2 presents the findings from another implementation design, where the corpus at each interval is real-time like the original corpus, i.e. it only contains top terms from the drafts up to the current interval, but now it contains more of the top frequent terms, hence capturing more of the contents of each draft.

Further details on how these two additional corpora are constructed are presented in Appendix D.

6.4.2 Feature extraction from the emails and documents

Following the construction of the term corpus, for each of the words in the term corpus, the following steps are taken:

In the document drafts domain. For each specific draft, recording which of these words appear in it, and which do not, to obtain values for which terms are most frequent in each draft (variables I and F, for each inter-draft interval);

In the email domain. Searching for the occurrence of each word in the thread subject lines, to obtain relevant email conversations for each word, and from all those email conversations aggregating the overall sentiment and participation levels (variables S and P) for each inter-draft time interval.

The details on how the text of the document drafts and the emails is cleaned and processed in order to then extract the features (variables) of interest are described in Appendix B. There, it is first described how document drafts are processed in order to extract frequent terms from them (Algorithm 1). It is then described how the email archives are processed and how sentiment and participation values for each term discussed are extracted (Algorithm 2) Finally it is outlined how everything is put together to causally analyse how sentiment and participation from emails, as well as the contents of the previous draft, affect the contents ofthe next draft, for each draft of each document (Algorithm 3).

6.4.3 Causal formulae implementation

This section discusses how the formulae for estimating causal and selection effects were implemented in practice, as well as some properties of these formulae that will be useful later on in the thesis.

As explained in Section 2.6, nonparametric estimation of causal (and selection) effects is used in this thesis. The Average Causal Effect (ACE) formula is used, specifically, Equation 2.7, as the outcomes and the causal factors are all binary. And similarly for the selection effect, the corresponding Average Selection Effect (ASE) formula is used (Equation 2.8).

The probabilities that feature in those estimation formulae represent frequencies in the dataset. That is, a fully data-driven, frequentist approach is used, and not a Bayesian one (as causal estimation can be done under either paradigm, since the Bayesian or frequentist debate is orthogonal to the usage of causal methodology, per Pearl, 2009a). In such a context where probabilities are represented by frequencies in the dataset, conditioning (in conditional probabilities) can be thought of as *filtering* the dataset based on the value of one or more variables (Pearl et al., 2016).

Therefore, the probabilities of the estimation formulae are calculated in practice by counting the frequencies of different value combinations among the variables, across all the entries (terms) in the dataset.

Hence, probabilities are implemented using their standard formal definitions:

Probability. The probability of a variable Y having value y, $P(Y = y)$, is implemented by counting entries in the data for which variable Y has value y, and dividing that count by the total number of entries (i.e. of terms in the corpus), n, hence obtaining the frequency of variable Y having value y.

Joint probability. Using the above calculation for the probability of a variable Y having value y, the joint probability of two variables Y and V having values y and v respectively, $P(Y = y, V = v)$, is implemented by counting entries in the data for which variable Y has value y and also variable V has value v, and dividing that count by the total number of entries, n. This extends similarly to the joint probability of more than two variables.

Conditional probability. Finally, the conditional probability $P(Y = y|X = x)$ is computed using the above two probability implementations as building blocks, as, per the standard formal definition of conditional probabilities, $P(Y = y|X = x) = P(Y = y, X = x)/P(X = x)$, for $P(X = x) > 0$.

The above conditional probability formula is then the building block for calculating both selection effects and causal effects:

Selection effects. Selection effects are calculated using the Average Selection Effect (ASE) formula of Equation 2.8: $ASE = P(Y = 1|X = 1) - P(Y = 1|X = 0)$.

Causal effects. Causal effects are calculated using the Average Causal Effect (ACE) formula of Equation 2.7: $ACE = P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$. The ACE is expressed in terms of standard statistical probabilities, without the *do* operator, by using the backdoor formula of Equation 2.3: $P(Y = y|do(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$, where Z is the deconfounding set.

In the context of the W3C Provenance Working Group, we are interested in estimating the influence of online communications (emails) on the contents on the produced documents, i.e. for a given document and a given inter-draft interval, how did the emails (specifically, the sentiment and participation volume in emails) affect whether the average term would feature frequently in the next document draft? Are there confounders, and if so, can they be ingnored or do they actually introduce substantial confounding bias?

Terms (words) are used as the unit of analysis, over which we are averaging effects. Therefore, the outcome of interest (Y) is whether a term will feature frequently in the next draft, i.e. the final draft of this inter-draft interval (F). The causal factor of interest (X), whose influence (causal effect) on the outcome we want to estimate, is the features of the emails, let's say here the Participation feature (P). For this causal effect, to remove confounding from backdoor paths, one must adjust for I, so the deconfounding set (Z) is $Z = \{I\}$ (as established in Section 6.3). All these variables are binary.

As a toy example of how these effects are calculated, consider Table 6.2, depicting a simple hypothetical scenario, where there are 25 terms (named here for simplicity after letters of the alphabet), with a value for each variable (P, I, F) being recorded for each.

To obtain the average causal effect of P on F, one must calculate the ACE, which here is $P(F = 1|do(P = 1)) - P(F = 1|do(P = 0))$, and per the backdoor equation, with $Z = \{I\}$ here, this expands to $P(F = 1|do(P = 1)) = \sum_{i=\{0,1\}} P(F = 1|P = 1, I = i)P(I = i)$. Similarly, the average selection effect (ASE), which we must compare to the average causal effect in order to assess the extent of confounding bias, is $P(F = 1|P = 1) - P(F = 1|P = 0)$.

Both ACE and the ASE are comprised of two terms - for each of these effects, let us calculate each of the two terms in this expansion. The average causal effect of P on F is shown in Table 6.3.

Table 6.2: Toy example, for the effect of P on F

	X	Z	Y
Term	P	I	F
a	0	0	0
b	0	0	1
c	0	0	1
d	0	1	0
e	0	1	0
f	0	1	0
g	0	1	1
h	0	1	1
i	0	1	1
j	0	1	1
k	1	0	0
l	1	0	0
m	1	0	1
n	1	0	1
o	1	1	0
p	1	1	0
q	1	1	0
r	1	1	1
s	1	1	1
t	1	1	1
u	1	1	1
v	1	1	1
w	1	1	1
x	1	1	1
y	1	1	1

Table 6.3: Toy example, Average Causal Effect of P on F

p, i	$P(F = 1 \mid P = 1, I = i)$	$P(I = i)$	$P(F = 1 \mid P = 1, I = i)P(I = i)$
1,0	50%	28%	14%
1,1	73%	72%	52%
			Total: 66%
p, i	$P(F = 1 \mid P = 0, I = i)$	$P(I = i)$	$P(F = 1 \mid P = 0, I = i)P(I = i)$
0,0	67%	28%	19%
0,1	57%	72%	41%
			Total: 60%
			Total: 6%

The top half of Table 6.3 shows the calculations for the first term of the ACE ($P(F = 1|P = 0, I = i)P(I = i)$), and the bottom half of the table shows the caclulations for the second term ($P(F = 1|P = 0, I = i)P(I = i)$). Each of those terms is a product of two further terms, and those are shown first, for each value p of P and i of I. For example, the first value shows that, for $p = 1$ and $i = 0$, $P(F = 1|P = 1, I = 0)$ is 50%, as, in Table 6.2, there are 4 terms for which $P = 1$ and $I = 0$ (terms ‘k’, ‘l’, ‘m’, ‘n’), and half of those (‘m’ and ‘n’) have $F = 1$. Then, $P(I = 0)$ is 28% here, as there are 7 out of 25 terms for which $I = 0$ (terms ‘a’, ‘b’, ‘c’, ‘k’, ‘l’, ‘m’, ‘n’). Multiplying these two numbers, we obtain the value of the first term of the first term of causal effect, which is 14%. The same procedure is performed for the next line with $p, i = 1, 1$, obtaining 52%. These two numbers are added, to obtain the first term of the causal effect, 66%. The same procedure is repeated in the bottom half of the table, yielding a value of 60% for the second term. Subtracting the second from the first term, we obtain a causal effect of 6%. This is a positive causal effect (greater than zero), meaning that a change in P (email participation) from 0 to 1 (i.e. from low to high) causes on average a 6% increase in the probability that $F = 1$ (that a term would feature frequently in the final draft of this interval).

Table 6.4 shows the respective calculation procedure for the average selection effect (ASE) of P on F.

Table 6.4: Toy example, Average Selection Effect of P on F

p	$P(F = 1 \mid P = 1)$
1	87%
p	$P(F = 1 \mid P = 0)$
0	60%
	Total: 27%

Table 6.4 shows that the first term of the ASE, $P(F = 1|P = 1)$, is 87%, as, in Table 6.2, out of the 15 entries with $P = 1$, 13 also have $F = 1$. The second term of the ASE, $P(F = 1|P = 0)$ is calculated similarly, and the difference fo the two terms gives an ASE of 27%. Therefore, this ASE of 27% is much larger than the ACE of 6% (4.5 times larger), overestimating the effect of P on F. The discrepancy between the causal and the selection effect tells us how much confounding there is due to the confounders identified and adjusted for in the causal effect (in the above example, variable I).

As a measure of discrepancy between causal and selection effect, i.e. of bias due to unmeasured confounding in the selection effect, the absolute relative difference (ARD) will be used in this thesis, defined in Equation 6.1.

$$ARD = \left| \frac{\text{causal effect} - \text{selection effect}}{\text{causal effect}} \right| \tag{6.1}$$

We call this quantity absolute relative *difference* rather than absolute relative *error*, as the word ‘error’ is conventionally employed in statistics to denote the discrepancy or deviation of a (usually parametric) model from the true data, whereas here we are assessing the discrepancy of a causal effect based on a nonparametric model from a selection effect also based on the same nonparametric model. In other words, we are assessing the discrepancy of a model which accounts for confounders, yielding the ACE, from the same model but disregarding confounders, yielding the ASE (the ASE would be equal to the ACE if there was no confounding bias present). Otherwise, we use the standard practice of making this comparison in relative terms (i.e. not just a difference, but a difference divided by the causal effect, i.e. made relative to the causal effect), and we use the absolute value of this ratio because we are not interested in the direction of the discrepancy (positive or negative) but only in its magnitude, and indeed because we will be calculating descriptive statistics over those magnitudes we do not want them to cancel out due to them having positive or negative signs.

So, for the toy example in Tables 6.2, 6.3 and 6.4, the ARD is 344%, a very large number, indicating that the amount of confounding due to variable I is not negligible but quite substantial, so one should adjust for I as it does introduce large confounding bias into the causal estimate of the influence of P on F.

As a side note, it is noted here that the selection effect, $P(F = 1|I = 1) - P(F = 1|I = 0)$, is equal to $P(F = 0|I = 0) - P(F = 0|I = 1)$, provided that $P(I = 1) > 0$ and $P(I = 0) > 0$. The same holds for the respective causal effect.

This observation is mentioned as it may be helpful or more intuitive sometimes, when discussing causal and selection effects, and their properties when different types of term corpora are used, to think in terms of $P(F = 0)$ (probability that a term is *not* prominent in the next draft, which appears in the latter formulation) rather than in terms of $P(F = 1)$ (which appears in the original formulation of causal and selection effects).

To prove this, let us compare the default selection effect expression, $P(F = 1|I = 1)-P(F = 1|I = 0)$, and the other expression in terms of $F = 0$, $P(F = 0|I = 0) - P(F = 0|I = 1)$.

For simplicity brevity of exposition, let us assign a variable to each term of each expression:

- $a = P(F = 1|I = 1)$
- $b = P(F = 1|I = 0)$
- $c = P(F = 0|I = 0)$
- $d = P(F = 0|I = 1)$

So then the selection effect becomes $a - b$, and the goal is to test whether this equals the expression $c - d$. For both expressions, let us add b and d.

Then we obtain, for the selection effect:
$a + d = P(F = 0|I = 1) + P(F = 0|I = 1) =$
$= P(F = 0, I = 1)/P(I = 1) + P(F = 0, I = 1)/P(I = 1) =$
$= (P(F = 0, I = 1) + P(F = 1, I = 1))/P(I = 1) =$
$= P(I = 1)/P(I = 1) = 1$, for $P(I = 1) > 0$.

And the $c - d$ expression becomes:
$c + b = P(F = 0|I = 0)P(F = 1|I = 0) =$
$= (P(F = 0, I = 0) + P(F = 1, I = 0))/P(I = 0) =$
$= P(I = 0)/P(I = 0) = 1$, for $P(I = 0) > 0$.

So, the two expressions are both equal to 1, and are equal to each other, provided that $P(I = 1) > 0$ and $P(I = 0) > 0$. So, either expression can be used as the selection effect.

Similarly, the same holds for the causal effect:
$\sum_z P(F = 1|I = 1, Z = z)P(Z = z) - \sum_z P(F = 1|I = 0, Z = z)P(Z = z)=$
$= \sum_z P(F = 0|I = 0, Z = z)P(Z = z) - \sum_z P(F = 0|I = 1, Z = z)P(Z = z)$.

Finally, as a brief illustration of how these causal and selection effects behave depending on the values present in the dataset, it is noted that, for a dataset where the top frequent terms in the drafts change over time, where at every stage some prominent ones may be abandoned (no longer frequently mentioned in the drafts), some new ones becoming prominent, and some staying the same over time, the causal and selection effects of I (the contents of the previous draft) on F (the contents of the current draft) are expected to take any values in $[0, 1]$. If one imagines an extreme case, where the terms that are most frequent in the beginning stay exactly the same throughout, i.e. if the working group did not drop any of the prominent terms and did not introduce any new ones along the way, then at every inter-draft interval $I = 1$ and $F = 1$ for all terms, and none of the other possible value combinations every appear in the dataset. (This does not actually occur for the dataset in this thesis.) This would mean that the selection effect would now be $P(F = 1|I = 1) - P(F = 1|I = 0)$, where the first term equals 1 and the second terms is undefined (as its denominator, $P(I = 0)$, equals 0) – or it could be treated as being 0 as is sometimes done in practice. If treated as a 0, the selection effect would be equal to 1. At the same time, for the alternative (and equivalent) formulation of the selection effect, $P(F = 0|I = 0) - P(F = 0|I = 1)$, the second term is 0, while the first term is undefined (denominator is 0) – or it could be treated as being 0 as is sometimes done in practice. However, if the undefined terms in both formulations of the selection effect are treated as 0s, then the alternative formulation of the selection effect would have a value of 1, which is not equal to the value of the default formulation of the causal effect (0), but is instead equal to its complement. Therefore, in such cases, it is better to not assign values of 0 to undefined quantities. And indeed, in the proof of equality of the two forms of the selection effect, above, the requirement was that all denominators are greater than zero. So, in such an extreme case where prominent terms remain exactly the same throughout, the selection effect would be undefined. And the same would hold for the causal effect (as it also depends on the same conditional probability, marginalised over confounders Z which do not alter the above logic). So, this example of an extreme case where only one value combination is present (both F and I are 1) serves as a demonstration that, in order for causal (and selection) effects to be estimable, the variables in the dataset must take on an appropriate range of values (in this example, there must also be some cases where $I = 0$, otherwise the effects are undefined).

6.5 Summary

This chapter has presented the Collective-level Causal Framework (CCF), the collective-level instantiation of the abstract causal framework (ACF) of Chapter 4, in Section 6.1. The CCF presents a set of principles for how the influence of online communications can be conceptualised and measured at the collective level specifically (something which has received little attention in the literature). It tailors the general principles of the ACF to collective-level analysis, by addressing issues that are specific to collective-level analyses, including: mapping any variables captured at the individual-level to collective-level variables; determining the appropriate unit of analysis (which is no longer individual people, as it was for the ICF); and offering a flexible classification scheme for possible confounding causes (causes internal to the collective setting, causes external to the collective setting, traits of the focal item). The CCF constitutes a contribution applicable to a wide range of collective settings, i.e. any setting where there is a record of collective outcomes and of online communications. It is generic and flexible, so as to be applicable to any setting where outcomes are produced collectively and where there is an online communications component whose influence on the collective outcome one wants to measure. Such settings may range from professional collaborations and formal projects, to the newer but increasingly ubiquitous kinds of projects in the areas of crowdsourcing and citizen science. Next, Sections 6.2 to 6.4 present a practical demonstration of the value and real-world applicability of the CCF, showing how it can be applied empirically, to a real-world setting of collectively-produced outcomes. This demonstration covers concepts (e.g. outcomes versus causes), causal modelling, variable extraction from the data, and causal formulae implementation in the particular setting under study.

In more detail, from Section 6.2 and onwards, this chapter provided details on how this CCF is empirically applied to the data of the W3C Provenance Working Group archives. This is a publicly available real-world observational (non-experimental) dataset, recording a specific setting of collective action and collaboration. Section 6.2 presented the nature and context of this W3C Provenance Working Group and described the dataset. Next, Section 6.3 described the graphical causal model for the causes and outcomes that will be studied in this setting, and then Section 6.4 presented the empirical design and implementation used for the application of the CCF to this setting. As mentioned, these sections serve as a demonstration on how the CCF can be applied in practice, in a real-world setting of collaboration that had an online communications component and that produced collective outcomes.

The results, discussion, and evaluation of this implementation of the CCF to the analysis of the Provenance Working Group archives will be presented next, in Chapter 7.

Chapter 7

Collective-level causal framework: findings

Chapter 6 presented the Collective-level Causal Framework (CCF), the W3C Provenance Working Group dataset, and described the implementation of how the CCF is applied to that dataset.

This chapter presents the findings of that empirical application of the CCF to the W3C Provenance Working Group archives. Section 7.1 presents and discusses the findings obtained using this causal framework, and Section 7.2 proceeds to an evaluation of the fit of the causal model (Figure 6.2) to the data. Next, Section 7.3 discusses results when some design and implementation choices are varied and compares those to the main findings. The chapter concludes with a summary of the key findings in Section 7.4.

As will be discussed in this chapter, the CCF and its empirical application contribute the empirical finding that the assumption of the contagion-based paradigm that, in analysing the social influence of online communications on outcomes, other causes can safely be ignored, does not necessarily hold.

That is, in the particular setting studied here (the W3C Provenance Working Group archives), this assumption of the contagion-based paradigm that causal factors other than online communications can be ignored, when measuring the influence of online communications on outcomes, is found to not hold, per the findings listed below:

The social influence of online communications is significantly confounded The (untested) assumption made in the contagion-based paradigm, that, when measuring the social influence of online communications, other factors can be ignored, does not hold - evidence is found that the social influence of online communications is confounded with the effects of other causes (specifically, of previous outcomes), and this confounding bias is far from negligible (mean bias across contexts at 59%, mean bias of per-context means at 400%). Therefore, to avoid estimates being substantially biased thus, one cannot ignore other factors, but must measure and adjust for them.

The magnitude of online social influence is relatively small. It is found that the magnitude of the social influence of online communications on the outcome is relatively small (mostly lower than 30%, and is nowhere near 100% at any point), compared to the effect of other causal factors (specifically, the previous outcome, whose effect magnitude is mostly at above 50%, and near 100% at later stages of the Group's lifetime). This causal factor is also one that introduces confounding bias to estimates of the social influence of online communications, so assuming that one can ignore it (as per the contagion-based paradigm) not only introduces bias to the estimate of the social influence (causal effect) of online communications on the outcome, but it also leads to completely missing a cause that is more important than online communications in influencing the outcome.

Ignoring confounding causes leads to very different conclusions. Ignoring confounding causes leads to different patterns of the magnitude of social influence from online communications on the outcome over time, than when accounting for confounding causes. That is, the causal effect (accounting for confounders) of email communications on the outcome tends to be smaller in later stages than it was in early stages, however the selection effect (ignoring confounders) tends to be higher in later stages than it was in early stages. This serves as further evidence (in addition to the first point above) that confounding causes cannot be ignored, as ignoring them would yield very different patterns for the social influence (causal effect) of online communications over time.

Time matters: variation of causal effects over time. The social influence of online communications, and the effect of the previous outcome, on the next outcome, both vary over time: the latter decreases over time, while the former increases, hence this collaborative setting displays a pattern of stabilisation: as time passes, the previous outcome becomes more and more important in shaping the next outcome. This happens for the majority of contexts (deliverables, or sub-groups).

Context matters: variation of causal effects across contexts. The social influence of online communications, and the effect of the previous outcome, on the next outcome, both vary across sub-groups (each working on a separate document deliverable): not all sub-groups exhibit the same effect magnitudes, nor the same effect patterns over time.

Better fit to data. The causal model proposed here better fits the empirical data than the contagion-based paradigm's (implied) model, which assumes other causes do not introduce confounding bias and can be ignored, and which is found to violate dependencies and independencies in the data.

That is, by also paying some attention to the confounder itself (the previous outcome), rather than ignoring it per the contagion-based paradigm, it has been found that the confounder is, in this setting, much stronger *both* as a statistical predictor *and* as a causal factor of the outcome, that its effects on the outcome develop much more steadily over time, and that its effects are much more robust to confounding, compared to online communications. Therefore, in this setting, it is found that the social influence of online communications is not only substantially confounded, but it is much weaker, and much less steadily evolving over time, and much less robust to confounding, than previous outcomes. Hence, this chapter sheds light on a very important factor, that has been a blind spot in many contagion-based studies of how online social interactions affect outcomes, and which in the setting considered here is a much stronger causal factor of outcomes than online social influence itself.

Therefore, these findings show that, when attempting to measure the social influence of online communications on outcomes, one should not assume that other possible causes of the outcome can be safely ignored, as it may be that other causes introduce substantial confounding to the estimate of the social influence of online communications on outcomes, and it might even be that these other causes are stronger in influencing the outcome than online communications, and the effects of these causes might even be much more robust to confounding and more steadily evolving over time than the effects of online communications.

The finding that other causes cannot be ignored, because they may introduce confounding, is in line with other individual-level theoretical and empirical studies, which have found that accounting for factors that are usually ignored can reduce confounding bias in estimates of the social influence of online communications (e.g. Aral et al., 2009; Aral and Walker, 2012; Eckles and Bakshy, 2017; Shalizi and Thomas, 2011), as shall be discussed in this chapter. Therefore, the findings here offer additional evidence, for the problem of measuring the social influence of online communications on outcomes, from a setting of a different kind than what the literature has focused on (a setting with collective outcomes, rather than a setting with individual outcomes), that this assumption of the contagion-based paradigm may not hold. This chapter also goes beyond this, by also investigating the effects of confounding causes themselves, and their evolution over time, and comparing them to the influence of online communications.

An additional contribution is that the findings of the empirical application of the CCF presented in this chapter (given the implementation details presented in Chapter 6) also serves to demonstrate how the CCF can be employed and adapted to settings of Web-mediated or Web-assisted collective action, and how it enables one to measure and compare the social influence of online communications, versus the effects of other causes, on the outcome of interest, and to test the assumption of the contagion-based paradigm that causal factors other than online communications can be ignored, over time and across contexts, based on observational digital trace data.

7.1 Results and discussion

This section presents the findings obtained from the application of the Collective-level Causal Framework (CCF) to the public archives of the W3C Provenance Working Group, using the causal model (Figure 6.2) and implementation described in Chapter 6. As discussed, in this setting, the online communications are emails, and two of their features are studied: Participation and Sentiment. The outcomes are documents, and the feature of theirs that is measured is which terms are prominent or not in the formal vocabulary and in the language and exposition used in these documents. There are twelve documents produced by the Group, and each draft of each document is captured in the dataset, as well as the emails from the beginning to the end of the project's lifetime. Therefore, the analysis of the social influence of online communications on the outcome in this setting translates to measuring, for every document, at every inter-draft interval, for the average term, the influence (causal effect) of email Participation and email Sentiment in emails about that term on whether this term will feature heavily on the next document draft. In this chapter, the causal DAG used is the one presented in Figure 6.2 of Chapter 6.3.

The main goal of this section is to empirically test the assumption of the contagion-based paradigm that, when measuring the social influence of online communications on an outcome of interest, one may safely ignore any other factors. For this assumption to hold, it must be that the measurement, i.e. the estimate, of the social influence of online communications on an outcome of interest is not biased by ignoring other factors. Therefore, to test this assumption, this section tests whether ignoring other factors introduces any bias, specifically confounding bias, to the estimate of the social influence of online communications on the outcome of interest. As the dataset covers outcomes and online communications over time, and there are twelve different groups of people working on outcomes (documents) in parallel (different contexts), this analysis is also performed over time and across contexts. This is presented in Section 7.1.1.

Further, it is examined whether this confounding bias is small, in which case one might still argue that the contagion-based paradigm's assumption approximately stands, if one is willing to tolerate a small level of bias in order to not have to worry about measuring other factors, or whether instead the confounding bias is large. This analysis is also presented in Section 7.1.1. In addition, these measurements are taken over time, testing whether causal effect magnitudes remain the same or change over time, and whether it is the effects of online communications or of other causes that are more steadily evolving over time. These measurements are also taken across contexts, testing whether causal and selection effects magnitudes, confounding bias, and patterns over time are the same or vary across contexts

In addition to testing this assumption, this chapter goes further, to test the hypothesis that, even if there is confounding bias in the estimate of the social influence of online communications on the outcome, online communications are the strongest cause, with other causal factors having much smaller effects on the outcome. If that hypothesis holds, then one might argue that a weaker, or a modified, version of the dominant paradigm's above assumption holds: even if other causes do introduce confounding bias, their own effects are much weaker, so one need not worry about measuring their effects on the outcome – even if biased, the most important cause of the outcome is still online communications. In order to test this hypothesis, this section measures the effects of other causes (specifically, previous outcomes) on the outcome of interest. It also measures the amount of confounding bias this estimate suffers from, to examine whether this effect of other causes is more, or less, confounded than the effect of online communications on the outcome. This analysis is shown in Section 7.1.2.

Both Section 7.1.1 and Section 7.1.2 perform the above analyses over time, testing whether causal effect magnitudes remain the same or change over time, and whether it is the effects of online communications or of other causes that are more steadily evolving over time.The analyses are also performed across contexts, testing whether causal and selection effects magnitudes, confounding bias, and patterns over time are the same or vary across contexts.

For each hypothesis and assumption test above, each finding is presented in turn. Section 7.1.1 presents the causal analysis of the influence of online communications on document drafts (the outcomes), the first finding being that the social influence of online communications actually is confounded with another cause, specifically previous outcomes, in the collective setting studied here. It is also found that this confounding bias is in general quite large. This is an important finding, as it empirically examines the validity of the untested assumption of the contagion-based paradigm that when estimating the effects of online communications on outcomes any other causal factors can safely be ignored, and it shows that this assumption does not hold in this setting. This section also presents the findings that the effect of online communications varies over time and across contexts (documents, or sub groups).

Next, Section 7.1.2 presents the causal analysis of the effects of each document draft on the next draft (i.e. of each previous outcome on the next outcome), and compares those effects to the effects of online communications. It is shown that the former effects are stronger, more consistent over time, and more robust to confounding bias, than the latter. This finding adds further evidence to the perils of assuming that causal factors other than online communications can be ignored. This section also presents the findings that the effect of the previous draft on the next varies over time and across contexts (documents, or sub groups).

All the analyses presented in this section use the below measures and notation.

In each analysis, there is a figure containing one plot for each document, where each document represents a separate context, or sub-group, as the documents produced by this Working Group each pertained to a separate topic, and each had a group of editors and contributors working on it. In all the plots, the x-axis represents time, with each time-point denoting an inter-draft interval, i.e. the time interval between the publication of one draft and the next, for all drafts of the document concerned. The first interval on the x-axis (numbered 0) denotes the interval between the publication of the PROV charter document by the Provenance Incubator Group and the publication of the first document draft by the Provenance Working Group, for every given Provenance Working Group document. The y-axis in all plots shows the size of the causal and of the selection effects, ranging from 0 to 1. The causal effect is the causal quantity, which accounts for the relevant confounding cause using the backdoor forumla, and the selection effect, which does not account for any confounding causes, is a statistical, not a causal, quantity. The causal effect is represented by red bars, while the selection effect is represented by yellow bars in the plots. As presented in the Background Chapter (Chapter 2), confounding is present when the causal effect (red bars) does not equal selection effect (yellow bars).

As a measure of the causal effect, the Average Causal Effect (ACE) is used, as described in the Background Chapter (Chapter 2, Equation 2.7), that is, the formula $E(Y = 1|do(X = 1)) - E(Y = 1|do(X = 0))$, where X is the causal factor of interest, Y is the outcome of interest, and since Y is binary in this thesis, this formula is equal to $P(Y = 1|do(X = 1)) - P(Y = 1|do(X = 0))$.

7.1.1 The social influence of online communications is confounded with the effect of the previous outcome

This section examines the untested assumption of the contagion-based paradigm that one can safely ignore other causes of observed outcomes when studying the effect that social influence from online interaction has on those outcomes. This assumption is the hypothesis to be tested. Here, it is investigated whether social influence from online interpersonal interactions is confounded with the effects of any other causes of the outcome of interest.

Per the causal graph (Figure 6.2) proposed in Section 6.3, the contents of the previous draft (variable I) may introduce confounding bias to estimates of the influence of online communications (quantified using the Sentiment, S, and Participation volume, P, in the Group's email communications) on the next draft (variable F).

If it is found that there exists non-negligible confounding bias due to other causal factors, then the above assumption of the contagion-based paradigm does not hold, and one should instead account for and appropriately adjust for the confounding causal factor(s) in order to remove this confounding bias from the estimate of the effect of online social influence on the outcome.

In order to estimate the social influence from online communications on the outcome of interest, the overall Participation levels (P) and the overall Sentiment levels (S) in the emails are used as measures of the contents of the online email communications. The outcome of interest is each draft of each document produced by the Working Group. Therefore, in this section, the effect (or influence) of the overall Participation and of the overall Sentiment in the email communications on the contents of the next draft of a given document are measured. Specifically, we compare causal effects (which accounts appropriately for the confounding variable representing the contents of the previous draft of the same document) to selection effects (which ignore any confounding variables), to determine the presence and magnitude of confounding bias in the estimate of the effect of online social influence on outcomes.

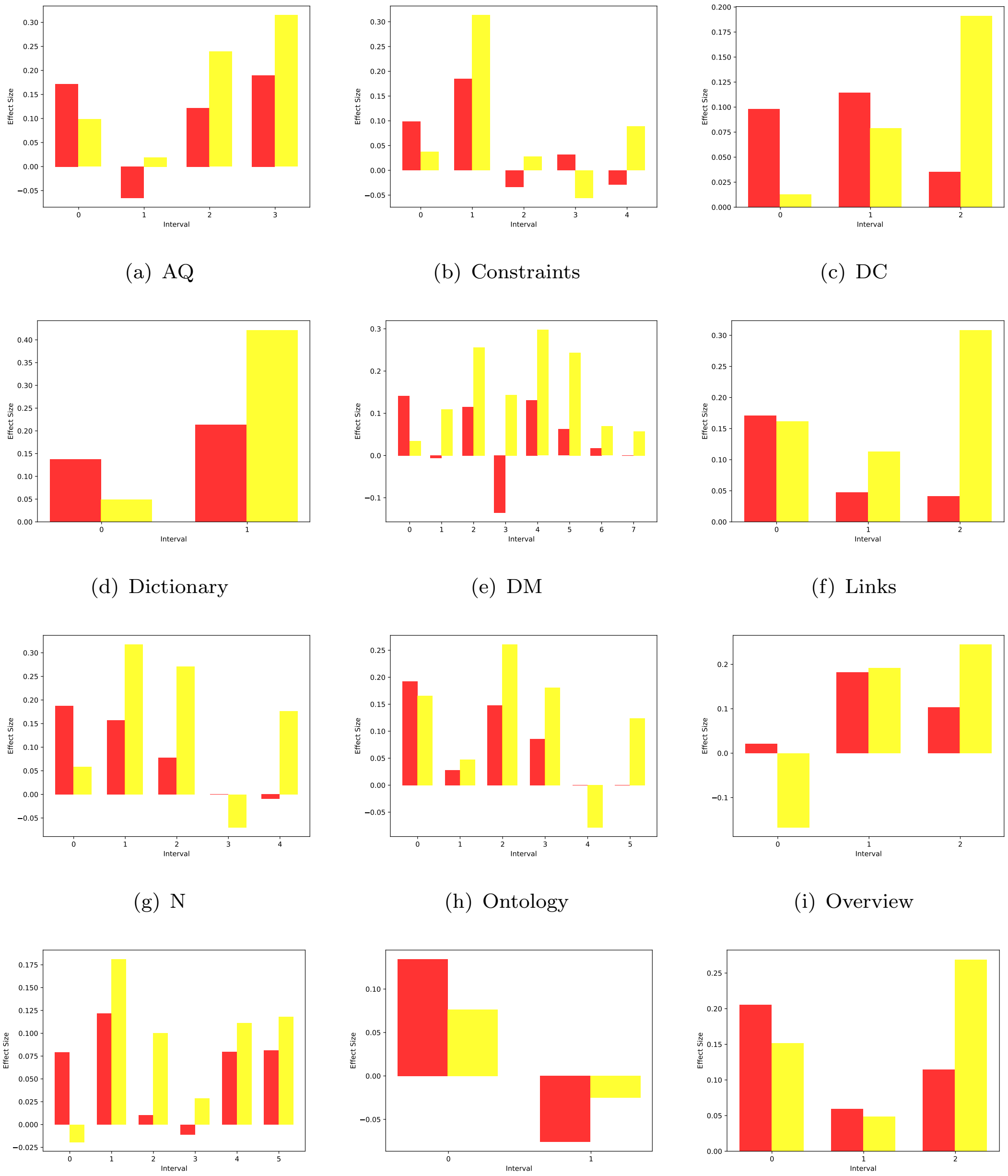
We begin with analysing the effects of the Participation variable, to measure the confounding bias in the estimate of its effects on the document drafts, and we then analyse the Sentiment variable and the confounding bias for it.

7.1.1.1 Participation volume in email conversations

For Participation, the causal and selection effects for each inter-draft interval (slice) of each document are plotted in Figure 7.1.

Here, the causal effect represents the effect of Participation volume (in the email conversations since the previous drafts) on the contents of the next draft, adjusting for the contents of the previous draft, as per the backdoor formula (Equation 2.3): $P(F = 1|do(P = 1)) - P(F = 1|do(P = 0)) = \sum_z P(F = 1|P = 1, Z = z)P(Z = z) - \sum_z P(F = 1|P = 0, Z = z)P(Z = z)$, where $Z = \{I\}$. Here, F stands for the contents of the current draft ($f$inal draft, at the end of this time interval), I stands for the contents of the previous draft ($i$nitial draft, at the start of this time interval), P stands for $p$articipation volume in the email communications during this time interval. These are binary variables, taking values as follows: $I, F, P \in \{0, 1\}$, as described in more detail in Section 6.4.

The selection effect represents the effect of Participation on the next draft, without accounting for the previous draft, calculated using only standard conditional probabilities (as this is not a causal quantity, but an associational one), with the formula: $P(F = 1|P = 1) - P(F = 1|P = 0)$.



(a) AQ (b) Constraints (c) DC

(d) Dictionary (e) DM (f) Links

(g) N (h) Ontology (i) Overview

(j) Primer (k) Sem (l) XML

Figure 7.1: Effects of Participation in emails on the current draft: the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; causal and selection effects are not always positive, they are sometimes negative; causal effects never exceed the 25% mark, while selection effects can go as high as 30% or 40%; selection effects often peak later than causal effects, while causal effects are often smaller in later intervals than in early intervals; the magnitude of confounding bias and the temporal patterns vary across documents.

From the plots of Figure 7.1, one can see that the causal effect is not equal to, and generally not even very close to, the selection effect, and the difference between the two quantities (i.e. the confounding bias) can often be quite large. Moreover, in some cases the causal effect is larger than the selection effect, while in other cases the selection effect is larger than the causal effect. To further investigate and summarise the degree of divergence between the causal and selection effects (i.e. the magnitude of confounding bias), Table 7.1 shows descriptive statistics of the confounding bias, calculated using the absolute relative difference (ARD) formula (Equation 6.1) between the causal and the selection effect, at each data-point (inter-draft time interval) for each document. The descriptives are presented first for each document separately, then for all documents together in the 'All documents' row (i.e. using all datapoints of all documents together), and the table finally shows the mean and the median of each descriptive statistic across all documents. These last two rows present descriptives of descriptives, as an additional measure to the 'All documents' row descriptives. That is because, given that each document covers a different topic and has a dedicated team of members (editors and contributors), it represents a different context thematically and in terms of collaborators, so it makes sense to first calculate descriptives for each context, and to then summarize those context-level descriptives, using here the mean and the median. This complements the 'All documents' row, which bundles together all datapoints from all documents, i.e. assumes that all datapoints can be taken together regardless of which document (context) they belong to (which quantitatively makes sense as the plots in Figure 7.1 show the effects in all documents are in the same order of magnitude and in similar ranges), and calculates descriptives on this raw data.

Table 7.1: Participation: Descriptives of confounding bias (ARD) across documents

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	43%	129%	83%	81%	11%	32%
Constraints	62%	409%	201%	183%	172%	131%
DC	31%	439%	185%	87%	326%	180%
Dictionary	64%	97%	80%	80%	3%	16%
DM	76%	1771%	413%	206%	3138%	560%
Links	6%	650%	265%	138%	773%	278%
N	69%	1942%	590%	174%	6139%	784%
Ontology	14%	11460%	2346%	76%	207659%	4557%
Overview	5%	907%	349%	136%	1581%	398%
Primer	40%	857%	247%	87%	872%	295%
Sem	43%	67%	55%	55%	1%	12%
XML	18%	134%	59%	26%	28%	53%
All documents	5%	11460%	482%	102%	27745%	1666%
Descriptives of descriptives						
Mean	39%	1572%	406%	111%	18392%	608%
Median	41%	545%	224%	87%	549%	229%

Overall, from the plots in Figure 7.1 and the descriptive statistics of confounding bias in Table 7.1, one reaches the following conclusions:

Confounding In Table 7.1, one sees that for Participation, the mean confounding across all documents is at 482%, which is affected by extremely high-valued datapoints, as the median is at 102%. Even though this median value is much lower than the average value for confounding, it still represents a confounding bias magnitude that is far from negligible and that is quite large. At the same time, the mean of means is at 406%, while the median of medians is at 87%. So, the choice of summary descriptive, mean or median, affects the conclusions reached about the extent of confounding bias. Still, for the purposes here of testing whether the magnitude of confounding is negligible, whether one looks at the median or the mean one may conclude that confounding is not negligible, but is instead rather large. In the plots of Figure 7.1, the height of the selection effect bar is often quite different to that of the causal effect bar at the same inter-draft interval (slice), so indeed the confounding bias is easily observable from the plots themselves.

Effect sign In the plots of Figure 7.1 one sees that causal and selection effects are not always positive, but are sometimes negative; when this happens, it means that either the average term is prominent in the next draft but there was low participation about it in the emails, or that the terms had on average high participation in the emails but were not prominent in the next draft. Either scenario may occur, and one might speculate on a range of possible reasons: for instance, because of other unobserved causes, such as intrinsic properties of the topic the term pertains to; or because consensus was reached without having devoted a lot of words to email discussions on average; or participants may have discussed a lot offline and not over email; or maybe on average terms were decided to not be worth mentioning prominently in the draft even though they were talked about or debated a lot in the online email communications. (Examining these scenarios further is outside the scope of the investigation in this thesis.)

Effect magnitude In terms of the magnitude of the effects, in Figure 7.1 one sees that causal effects never exceed the 25% mark on the y-axis. However, the selection effects do sometimes exceed the 25% mark (e.g. Figure 7.1(e) at interval 4, Figure 7.1(g) at interval 2, Figure 7.1(h) at interval 2, Figure 7.1(l) at interval 2) and can go as high as 30% (Figure 7.1(a) at interval 3, Figure 7.1(b) at interval 1, Figure 7.1(f) at interval 2, Figure 7.1(g) at interval 1) or 40% (Figure 7.1(d) at interval 1). Not only is the maximum of the causal effect lower than that of the selection effect (meaning that the selection effect overestimates the causal effect of Participation because the former captures spurious associations from the effects of the previous draft, I), but, as we shall see in the next section, the overall magnitude of the causal effect of Participation is lower than that of the confounding cause (contents of the previous draft, variable I). That is, not only does the previous draft, I, introduce large confounding to the estimate of the effects of email Participation, but, as we shall see, the previous draft itself is a stronger cause of the next draft's contents than email Participation.

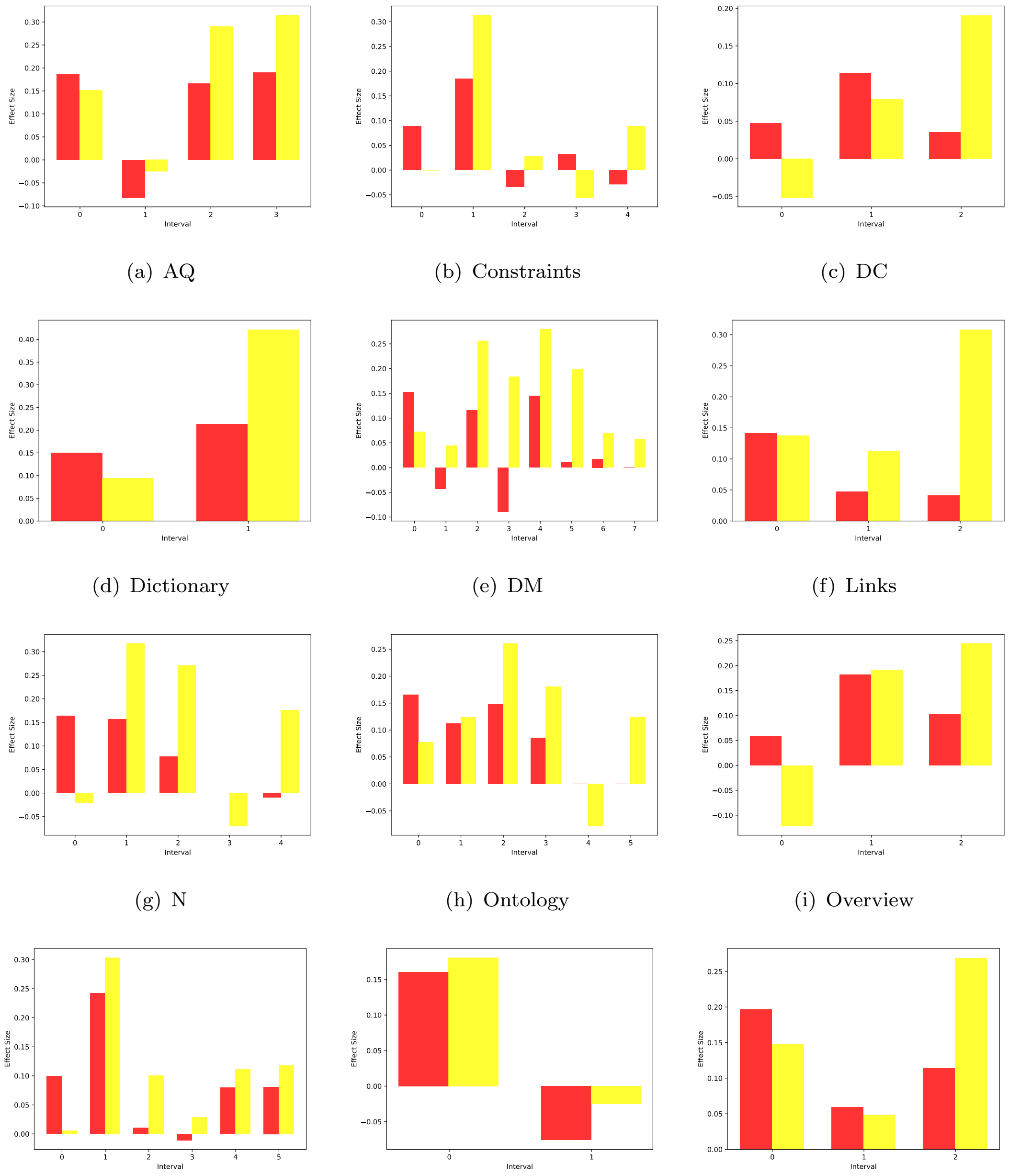
Evolution over time: causal versus selection effects For the causal effects, the effect magnitudes in the last few slices are often smaller than in the initial few slices, with the exception of AQ (Figure 7.1(a)), Dictionary (Figure 7.1(d)), and to a degree Primer (Figure 7.1(j)). Indeed, the peak is often at the first or second time slice. However, the selection effects do not follow this pattern; rather, their peak tends to be much later (e.g. Figures 7.1(f), 7.1(h), 7.1(i)), and/or the effects are often larger towards the end than towards the beginning (e.g. Figure 7.1(l)), or they continue to be large while the causal effects have gotten smaller (e.g. Figures 7.1(c), 7.1(e), 7.1(g), 7.1(h)). Therefore, if one were to only look at the evolution over time of the (biased) selection effect, they would often reach quite different conclusions compared to the evolution over time of the causal effect. In addition, the standard deviation values for each document in Table 7.1 can be quite large for some documents, for example Ontology exhibits the largest standard deviation value at 4,557%, with several other documents having 3-digit standard deviation values, showing there is large variability over time for many documents (although not for all, e.g. not for Sem).

Patterns across documents From Table 7.1 one can see that the magnitude of bias varies across documents. For instance, if one taked the median bias across documents, one sees that DM has the highest median bias at 206%, followed by Constraints at 183% and N at 174%, while the lowest median bias is 26% for XML. It is noted still that even a median bias of as low as 26%, in the case of XML, is not negligible, and indeed if one looks at the plot for XML in Figure 7.1(l), the discrepancy between causal and selection effect is obsrvable at all intervals, and it is particularly large at the last interval. Overall, given that the highest and lowest median bias figures (206% versus 26%) differ by a factor of almost 8, and that the median values across documents take a range of values, the overall level of bias can differ quite broadly across documents (contexts). In addition, temporal patterns vary across documents, in some cases exhibiting opposite tendencies - for instance, the causal effect gets smaller over time for Links (Figure 7.1(f)), but it increases for Dictionary (Figure 7.1(d)) (although Dictionary only offers two time points, hence there is not an evolution that one can observe over a long time period).

7.1.1.2 Sentiment in email conversations

Next, the effects of the overall Sentiment expressed in email conversations on the contents of the next draft are considered, adjusting, in the causal effect calculations, for the contents of the previous draft, as that variable may introduce confounding bias, according to the causal model (Figure 6.2). The aim is again to investigate whether such a confounding bias is indeed present, and if so whether its magnitude is negligible or not. Here, the causal effect represents the effect of Sentiment (variable S) in the email conversations since the previous draft on the contents of the next draft (variable F), adjusting for the contents of the previous draft (variable I), as per the backdoor formula of Equation 2.3, as follows: $P(F = 1|do(S = 1)) - P(F = 1|do(S = 0)) = \sum_z P(F = 1|S = 1, Z = z)P(Z = z) - \sum_z P(F = 1|S = 0, Z = z)P(Z = z)$, where $Z = \{I\}$. These variables again take values as follows: $I, F, S \in \{0, 1\}$.

The plots of causal and selection effects for each inter-draft interval (slice) of each document are shown in Figure 7.2 for Sentiment.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure 7.2: Effects of Sentiment in emails on the current draft: similarly to Figure 7.1 (for the effects of email Participation), the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; causal and selection effects are often positive but sometimes negative; causal effects never exceed the 25% mark, while selection effects can go as high as 30% or 40%; selection effects often peak later than causal effects, while causal effects are often smaller in later intervals than in early intervals; the magnitude of confounding bias and the temporal patterns vary across documents.

The plots for Sentiment in Figure 7.2 look very similar to those for Participation in Figure 7.1. Figure 7.2 shows that the causal effect is always unequal to, and generally not very close to, the selection effect, across documents. Hence, one can readily observe the presence of confounding bias due to the contents of the previous draft which the selection effect ignores. In addtion, in some cases, the causal effect is bigger than the selection effect, in other cases the opposite holds. To summarize the magnitude of confounding bias present for each document and across documents, let us look at the descriptives of confounding bias (measured using the ARD formula, of Equation 6.1), presented in Table 7.2, which is in the same format as Table 7.1.

Table 7.2: Sentiment: Descriptives of confounding bias (ARD) across documents

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	19%	75%	57%	68%	5%	23%
Constraints	69%	409%	208%	183%	154%	124%
DC	31%	439%	226%	210%	278%	167%
Dictionary	37%	97%	67%	67%	9%	30%
DM	53%	1640%	387%	202%	2698%	519%
Links	3%	650%	264%	138%	778%	279%
N	102%	1942%	601%	179%	6028%	776%
Ontology	9%	11460%	2342%	76%	207846%	4559%
Overview	5%	310%	150%	136%	155%	125%
Primer	25%	857%	238%	70%	901%	300%
Sem	12%	67%	40%	40%	8%	27%
XML	18%	134%	59%	25%	28%	53%
All documents	3%	11460%	465%	101%	27662%	1663%
Descriptives of descriptives						
Mean	32%	1507%	387%	116%	18241%	582%
Median	22%	424%	217%	106%	217%	146%

Overall, the plots in Figure 7.2 and the descriptives in Table 7.2 lead to the following conclusions:

Confounding In Table 7.2, one sees that for Sentiment, the mean confounding across all documents is at 465%, which however is affected by extremely high-valued data-points, as the median is at 101%, which even though lower is still quite a high magnitude for confounding. At the same time, the mean of means is at 387%, while the median of medians is at 106%. So, one notes that the choice of summary descriptive, mean or median, affects conclusions about the extent of confounding bias. Still, for the purposes of testing whether the magnitude of confounding is negligible, both in the case of the mean and in the case of the median one may conclude that confounding is not negligible, but is instead rather large. In the plots of Figure 7.2, the heights of the selection effect bars are often quite different to those of the causal effect bar at the same inter-draft interval(slice), so indeed there is observable confounding. In terms of other descriptives, one notes also that the maximum confounding bias values encountered at any time interval in each document are quite large, ranging from 67% (for Sem) to the extremely large value of 11,460% (for Ontology), with 3-digit values being common for other documents (Constraints, DC, Links, Overview, Primer, XML) and some 4-digit values too (DM, N). Indeed, the median of the maximum values encountered in many documents is at 424%, which is very high, and means that, in the worst case (as this concerns the maximum bias per document), ignoring the causal effect and only looking at the selection effect would lead to an estimate that is off by 424% on average across all documents.

Effect sign Similarly to the case of the Participation variable, in Figure 7.2 causal and selection effects are not always positive, but are sometimes negative. When this happens, it means that, on average either terms with high sentiment were not prominent in the next draft, or that on average terms with low sentiment were prominent in the next draft. Either scenario may occur, for a range of unobserved causes (e.g. agreement was reached for terms to be prominent without strong emotions being expressed; certain topics were considered of particular interest by the participants from the beginning, so relevant terms were included in the next draft without strong emotions being expressed; strong emotions were expressed offline, outside the email records), as explained earlier in the analysis of Participation.

Effect magnitude In terms of the magnitude of the effects, Figure 7.2 shows that causal effects never exceed the 25% mark on the y-axis. However, the selection effects do sometimes exceed the 25% mark (e.g. Figure 7.2(a) at interval 2, Figure 7.2(e) at interval 4, Figure 7.2(g) at interval 2, Figure 7.2(h) at interval 2) and can go as high as 30% (Figure 7.2(a) at interval 3, Figure 7.2(b) at interval 1, Figure 7.2(f) at interval 2, Figure 7.2(g) at interval 1) or 40% (Figure 7.2(d) at interval 1). This was also the case for the Participation variable. Not only is the maximum of the causal effect lower than that of the selection effect (meaning that the selection effect overestimates the causal effect of Sentiment because the former captures spurious associations from the effects of the previous draft, $\boldsymbol{I}$), but, as we shall see in the next section, the overall magnitude of the causal effect of Sentiment is lower than that of the confounding cause (contents of the previous draft, variable $\boldsymbol{I}$). That is, not only is the causal effect of Sentiment smaller than the (biased) selection effect of Participation ($\boldsymbol{P}$), but, as we shall see, it is also smaller than the causal effect of the previous draft ($\boldsymbol{I}$). This means that, not only does the previous draft, $\boldsymbol{I}$, introduce large confounding to the estimate of the effects of email Sentiment, but, as we shall discuss, the previous draft itself is a stronger cause of the next draft's contents than email Participation.

Evolution over time: causal versus selection effects For the causal effects, the effect magnitudes in the last few slices are often smaller than in the initial few slices, with the exception of AQ (Figure 7.2(a)), Dictionary (Figure 7.2(d)), and to a degree Primer (Figure 7.2(j)). Indeed, the peak is often at the first or second time slice. However, the selection effects do not follow this pattern; rather, their peak tends to be much later (e.g. DC in Figure 7.2(c), Links in Figure 7.2(f), XML in Figure 7.2(l)), and/or they are often larger towards the end than towards the beginning (e.g. AQ in Figure 7.2(a), DC in Figure 7.2(c), Dictionary in Figure 7.2(d), DM in Figure 7.2(e), Links in Figure 7.2(f), Overview in Figure 7.2(i), XML in Figure 7.2(l)), or they continue to be large while the causal effects have gotten smaller (DM in Figure 7.2(e), Links in Figure 7.2(f), N in Figure 7.2(g), Ontology in Figure 7.2(h)). Therefore, if one were to only look at the evolution over time of the (biased) selection effect, they would often reach quite different conclusions compared to the evolution over time of the causal effect. In addition, Table 7.2 shows that there is large variability over time for each document: variance and standard deviation values are quite large, with standard deviation having 3-digit values or larger for most documents (Constraints, DC, DM, Links, N, Ontology, Overview, Primer) with the median standard deviation being 146%. The minimum standard deviation of 23% for AQ is itself not negligible. This shows that time matters; there is considerable variability in the extent of bias according across different intervals of the lifetime of any given document.

Patterns across documents Table 7.2 shows that the magnitude of bias varies across documents. For instance, if we take the median bias across documents, we see that DC has the highest median bias at 210%, followed by DM at 202% and Constraints at 183%. At the same time, while the lowest median bias is much lower, at 25% for XML, it is still not negligible. Overall, given that the highest and lowest median bias figures (210% versus 25%) differ by a factor of 8.4, so that the median values across documents take a broad range of values, the overall level of bias can differ quite broadly across documents (contexts). In addition, temporal patterns also vary across documents; for instance, the causal effect gets smaller over time for Links (Figure 6.1(g)), but it increases for Dictionary (Figure 6.1(e)) (although Dictionary only offers two time points, hence there is not an evolution that we can observe over a long time period).

7.1.2 The influence of online communications versus the effects of previous outcomes

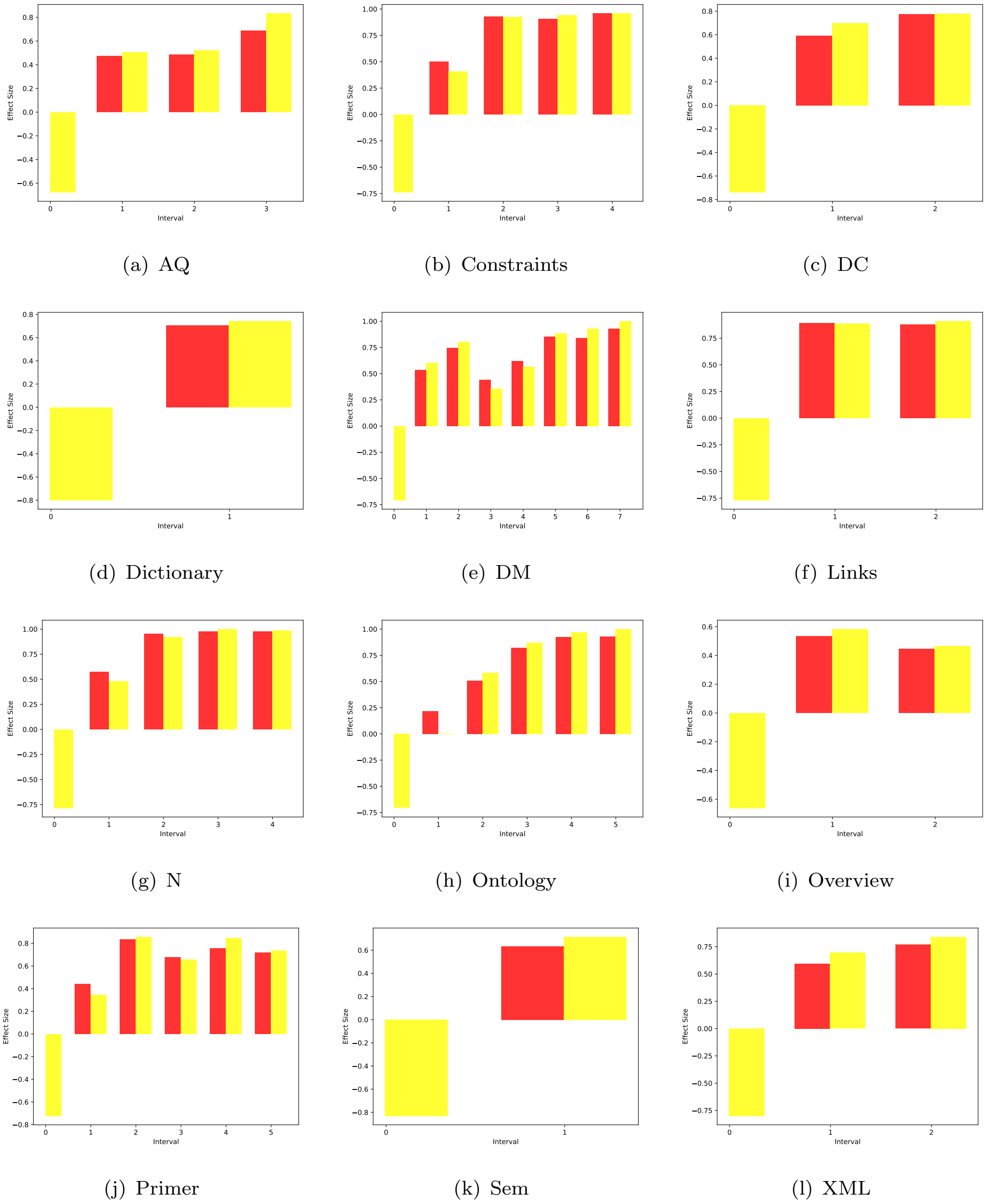
In addition to the finding above that the contents of the previous draft are a confounding cause, i.e. they introduce confounding bias to the estimate of the influence of online communications on the next draft, we now proceed to study in more detail the effect of the previous draft itself on the next draft. This analysis includes investigating the extent to which the contents of online communications (the Sentiment and Participation variables, S and P) introduce confounding bias to the estimate of the effect of the previous draft (variable I) on the next (variable F), as the causal model (Figure 6.2 in Section 6.3) suggests.

This section begins by analysing the effects of I on F, similarly to the analyses of the effects of P and S on F in the previous sections, for each inter-draft interval of each document. It then proceeds to compare the causal effects of I on F to the causal effects of S and P on F, to investigate the relative importance of online communications versus that of the previous draft, in shaping the next draft, at each interval. This comparison is important because the contagion-based paradigm focuses on the influence of online social communications on outcomes, while largely ignoring the role played by other causes such as previous outcomes.

7.1.2.1 Effects of previous outcomes

For the effects of the previous outcome, that is, the previous draft (variable I), on the next one (variable F), the causal effects and selection effects for each inter-draft interval of each document are plotted in Figure 7.3.

Here, the causal effect represents the effect of the contents (most frequent words) of a given draft (I) on the contents of the next draft (F), adjusting for the contents (overall Sentiment, Participation) of the email conversations pertaining to these contents that happened in the interval *before* the publication of the given draft (S_{-1}, P_{-1}) as per the causal graph of Figure 6.2. It is calculated per the backdoor formula (Formula 2.3): $P(F = 1|do(I = 1)) - P(F = 1|do(I = 0)) = \sum_z P(F = 1|I = 1, Z = z)P(Z = z) - \sum_z P(F = 1|I = 0, Z = z)P(Z = z)$, where $Z = \{S_{-1}, P_{-1}\}$.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

FIGURE 7.3: Effects of the previous draft on the current draft: the causal effect (red) is generally not exactly equal, but still relatively close, to the selection effect (red), across documents, meaning that the presence of some confounding bias can be observed, but that bias is not as large as in Figures 7.1 (for Participation) or 7.2 (for Sentiment); the causal effect is often, but not always, smaller than the selection effect; both causal and selection effects are always positive, except at Interval 0 (which will be discussed), for all documents; effects are generally large, compared to Figures 7.1 or 7.2, with values that can exceed 60% particularly in later intervals; causal and selection effects generally grow, in sync, over time (with the exception of some temporary drops in some cases); across documents, the magnitude of confounding bias can vary, but temporal patterns are often very similar.

In each plot of Figure 7.3, for the first inter-draft interval (numbered 0), corresponding to the interval between the charter document and the first Working Group draft, there is only a selection effect, as there is no recorded email discussion prior to the charter from which to extract and adjust for Sentiment and Participation values (for S_{-1} and P_{-1}). As shall be discussed, all these selection effects are negative.

These plots show that the causal effect is generally not exactly equal, but is still relatively close, to the selection effect, across documents. Hence, one can observe the presence of some confounding bias due to the contents of the email conversations (prior Sentiment and Participation variables, S_{-1} and P_{-1}) which the selection effect ignores. In addition, in many cases, the causal effect is smaller than the selection effect, but in some cases the opposite holds.

To summarize the magnitude of confounding bias present for each document and across documents, let us look at the descriptives of confounding bias (measured using the ARD formula, of Equation 6.1), presented in Table 7.3. As in the previous sections, the descriptives are presented first for each document separately, then for all documents together in the 'All documents' row (i.e. using all datapoints of all documents together).

Table 7.3: Previous draft: Descriptives of confounding bias (ARD) across documents

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	6%	21%	12%	7%	0%	7%
Constraints	0%	19%	6%	2%	1%	8%
DC	1%	18%	9%	9%	1%	9%
Dictionary	5%	5%	5%	5%	0%	0%
DM	4%	20%	10%	9%	0%	5%
Links	0%	4%	2%	2%	0%	2%
N	1%	16%	6%	3%	0%	6%
Ontology	5%	99%	26%	7%	13%	36%
Overview	4%	9%	7%	7%	0%	3%
Primer	2%	21%	8%	3%	1%	7%
Sem	13%	13%	13%	13%	0%	0%
XML	9%	17%	13%	13%	0%	4%
All documents	0%	99%	11%	7%	2%	16%

Descriptives of descriptives						
Mean	4%	22%	10%	7%	1%	7%
Median	4%	17%	9%	7%	0%	5%

Overall, the plots in Figure 7.3 and the descriptives in Table 7.3 lead to the following conclusions:

Confounding bias Figure 7.3 shows that there is some confounding bias, but not to the extent that the (biased) selection effect would yield different conclusions for the evolution of the effect over time compared to the (unbiased) causal effect, as both effects generally increase or decrease together. Indeed, in Table 7.3 one sees that the mean bias across all documents is quite low, at 11%, the median being only 7%, and similarly the mean of means is 10% and the median of medians 7%. The median confounding bias across documents ranges from as low as 2% (Constraints, Links) to at most 13% (Sem, XML), so the amount of bias here is small and one might argue that on the whole it is negligible.

Effect magnitude In Figure 7.3, the causal and selection effect size, especially in the last few time intervals, is quite large, taking values higher than 60% for some documents (e.g. Links, Primer, in Figures 7.3(f), 7.3(j); and for the last interval of AQ, DC, Dictionary, and XML, in Figures 7.3(a), 7.3(c), 7.3(d), and 7.3(l)) and indeed higher than 80% for the four Recommendation documents (Constraints, DM, N, Ontology, in Figures 7.3(b), 7.3(e), 7.3(g), 7.3(h)). This indicates that, in the later stages of the documents' lifetime, the contents of the document tend to stabilise and largely remain the same (in terms of the words that feature most prominently in the documents). As an aside, it is noted also that, for DM (Figure 7.3(e)), there is a drop in both causal and selection effect at interval 3, which corresponds to the time when the group decided that DM should be split into more documents: DM (Data Model), Constraints, and N (Notation), and that was when the latter three documents were first created.[1] Therefore, this group decision represents a change in the nature and scope of DM, which was reflected in a change in the most prominent terms that featured in DM, and is hence reflected here in the plots (in that the draft prior to this split is not such a good predictor, in the case of the selection effect, or strong causal factor, in the case of the causal effect, for the contents of the post-split draft). After that point, the upwards trend resumes for both effects.

Effect sign For all documents, both selection and causal effects are always positive, except for the selection effects in the first inter-draft interval: this interval represents the time between the Incubator Group (charter document) and the Working Group (specification documents). This negative sign is because, in the first interval, by construction of the term corpus, there are no terms that are not in the charter and also not in the first draft; that is, there are no terms for which both $I = 0$ and $F = 0$. Rather, for all terms in the first interval, at least one of I, F must equal 1. This means that the first term of the selection effect, $P(F = 0|I = 0)$ equals 0, so the selection effect $P(F = 0|I = 0) - P(F = 0|I = 1)$ is less than or equal to zero. The maximum value it can have is zero. (This does not apply to later time intervals, as it is possible for those to contain terms for which both $I = 0$ and $F = 0$: these would be terms that appeared in a previous draft, but no longer

[1] As recorded at https://www.w3.org/2011/prov/meeting/2012-04-19, under 'Resolution 1'. (It is possible that any other such 'breaks' in the upward trends of the effects over time, in other documents - e.g. interval 2 in Overview, interval 3 and 5 in Primer, are also due to such drastic decisions, which may or may not have been recored in the public archives of the Working Group.)

appear in the initial or the final draft for this interval). In this sense, the first interval should be considered a special case, in the analysis of the effects of I on F with this term corpus. Further, the Incubator Group (charter) and the Working Group were different contexts (as discussed in the beginning of the chapter), so it seems the contents of the charter are not good predictors of the contents of the first draft of the working group's documents, to the extent that a term is likely to be not-prominent in the first draft of the working group when it was prominent in the charter. Indeed, inspection of the top frequent terms of the Charter versus of the document drafts indicates that there is little overlap in the most frequent terms of the document drafts and the most frequent terms of the charter. Given the different contexts of the charter and the documents, the charter being only an initial skeleton of themes and topics intended to be addressed by the Working Group, and, on the other hand, with each document only pertaining to one theme and fleshing out the charter's original skeleton of topics with particular details, this limited overlap is not particularly surprising.[2] Other that these first intervals, selection effects and causal effects are always positive. This means that, whether adjusting for previous Sentiment and Participation from email communications or not, on average, having a concept feature prominently in the previous draft will consistently lead to higher chances (both causally and statistically) of it appearing prominently in the next draft.

Evolution over time In Figure 7.3, we see that both effects generally evolve in sync with each other, and in a consistent manner over time: they monotonically increase over time, hence one could say that the group exhibits a pattern of stabilisation: as time passes, it becomes more and more likely that what is already prominent in the previous draft will also be prominent in the next draft (both causally and statistically). This happens for the majority of documents (with the exception of Overview, and of the temporary drops observed in DM and Primer, as discussed above). In addition, Table 7.3 shows that the variance and standard deviation of bias within each document (representing the variability of bias over the various time points) are quite small, with the variance being at 0% or 1% and the standard deviation in the single digits, with the exception of the Ontology document (13% variance, 36% standard deviation). The median of variances is at 0% while the median of standard deviations is at 5%. Again, this shows that bias is quite small here, and it does not fluctuate particularly over time, so whether one looks at the evolution of the selection effects over time, or of the causal effects over time, they

[2] For instance, the first draft of the DM document contains in its top most frequent terms many concrete formal terms, including core concepts (or early versions of them) of the final PROV standards, like the terms `entity, identifier, thing`, `process, use`, `attribute, qualifier`, `execution`, `identify`, `value`, `activity, assertion, derivation`, `wasderivedfrom`, `wascomplementof`, `generation`. On the other hand, the charter tends to feature prominently more high-level, less detailed, words, and words relating to the goals of the Working Group, like `language, group, work` (due to many mentions of 'Working Group' and 'Working Draft'), `model`, `core`, `propose`, `vocabulary`.

would not reach significantly different conclusions about the temporal patterns present.

Patterns across documents In terms of how bias varies across documents, Table 7.3 shows the median bias ranges from 2% (Constraints) to as large as 30% (AQ), which corresponds to a difference by a factor of 15. So some documents have more confounding bias than others (many documents have 2-digit median bias), but, as selection and causal effects follow the same overall temporal patterns, it could be argued that this kind of confounding is not very important if one is willing to sacrifice some accuracy for the computational 'ease' of ignoring (i.e. not adjusting for) confounders. In Figure 7.3, in general, causal and selection effects increase over time for all documents, with the exception of Overview (Figure 7.3(i)). So there is not very much variability across most documents in terms of temporal patterns. For the DM document, in Figure 7.3(e), there is a drop in the causal and the selection effect, at inter-draft interval 3. At that time, the group had decided to remove some content from DM and instead put it in new documents (Constraints, Dictionary, N), so this drop reflects that change of context for DM, as DM's scope has changed, and its focus has now narrowed. From this point onward, both effects start increasing again. Similarly, for the first inter-draft interval, corresponding to the interval between the charter and the first Working Group draft, the selection effect is negative, for all documents. This is to be expected due to the nature of the contents of the term corpus, as explained in Section 6.4: at the first interval, all terms in the corpus are in the top terms of either the charter or the first draft, or both. As there are no previous drafts before the charter, there are no terms that are not prominent in the charter, not prominent in the first draft, but that are in the corpus because they featured prominently in previous drafts. This means that $P(F = 0|I = 0) = 0$, and as this is the first term of the selection effect, $P(F = 0|I = 0) - P(F = 0|I = 1)$, the selection effect must be less than or equal to zero (equal to zero when $P(F = 0|I = 1) = 0$). To look at this another way, by manually inspecting the contents of the charter, and the terms that the algorithms identify as most prominent in it, one sees that the charter's contents are very different to the contents of the subsequent document drafts, and there is very little overlap between the top frequent terms in the charter and the top frequent terms in any of the drafts. So, this negative selection effect can be interpreted as a reflection of a sort of change in context, from the Incubator Group effort that created the charter, to the actual Working Group effort.

7.1.2.2 Comparison: Influence of online interactions versus effects of previous outcomes

The patterns observed in the effects (influence) of online social interactions (variables S, P, for the sentiment and participation in emails) on the outcome (document draft), are now compared to the effects of the previous outcome (previous draft, variable I) on the outcome.

In terms of the amount of confounding bias present in each case, Table 7.3 shows that, for the previous draft (I), the confounding bias is much smaller than for the email communications (S, P, shown in Tables 7.2 and 7.1, respectively): for I, the median bias across all documents, and the median of per-document-medians, is at only 7%, whereas for S and P those values are at 101% and 102% (median across all documents), and at 106% and 87% (median of per-document medians). So, the median confounding bias present in the effect of email communications (S and P) is more than 12 times larger than the median bias in the effect of the previous draft (I), as $87\%/7\% > 12$.

Therefore, the selection effects of the previous draft are much more robust to confounding bias than the selection effects (influence) of email communications. And as these variables are each other's confounding cause, this means that the selection effects of the previous draft are much more robust to any confounding introduced due to the effects of email communications, than the selection effects of email communications are to the bias introduced by the effects of the previous draft. This means that, if an investigator is considering allowing some bias in exchange for the simplicity of not adjusting for confounders in effect calculations (by using selection effects and not causal effects), it is relatively safe to go ahead with this strategy if calculating the effects of each draft on the next (as median inaccuracy due to bias is only 7%), but it is not safe to do this for the effects of emails (sentiment and participation variables) on the next draft, as the median bias there would be very high, at 87% and 106% respectively. In other words, in this empirical setting, we have found that previous outcomes *are* robust to any confounding from online discussions, but online discussions are *not* robust to confounding from previous outcomes.

In terms of the properties of the effects themselves, both the causal and the selection effects of the previous draft on the current one (effects of I on F) are generally positive, large, and exhibit a relatively stable temporal pattern where they monotonically increase over time (Figure 7.3). This is in contrast to the effects of sentiment and participation in the emails on the current draft (effects of S and P on F), which are not always positive, are quite small, and their temporal evolution exhibits a much less consistent or 'smooth' pattern - some documents do not display a consistent pattern, whereas in some others (Constraints, DM, Links, N, Sem, and to an extend also DC, Ontology, and Primer, in Figures 7.2 and 7.1) the effects in the late stages are smaller than in the first stages. This means that, the previous draft is a positive and a strong causal factor and statistical predictor, of the current draft (positive, large causal and selection effects), which gets stronger and stronger as documents evolve over time (increase in effect size over time). However, the features of the email archives are not strong causally or statistically (small magnitudes), they do not always have a positive effect (negative signs for some effects), and as the documents evolve over time, they do not get any stronger - indeed, in several

Table 7.4: Effects of I, P, S for DM, N, Constraints

Interval	DM			N			Constraints		
	I	P	S	I	P	S	I	P	S
0	-	14%	15%	-	19%	16%	-	10%	9%
1	54%	-1%	-4%	57%	16%	16%	50%	18%	18%
2	74%	12%	12%	95%	8%	8%	93%	-3%	-3%
3	44%	-14%	-9%	97%	0%	0%	90%	3%	3%
4	62%	13%	14%	97%	-1%	-1%	96%	-3%	-3%
5	85%	6%	1%						
6	84%	2%	2%						
7	93%	0%	0%						
Descriptives									
Minimum	44%	-14%	-9%	57%	-1%	-1%	50%	-3%	-3%
Maximum	93%	14%	15%	97%	19%	16%	96%	18%	18%
Mean	71%	4%	4%	87%	8%	8%	82%	5%	5%
Median	74%	4%	1%	96%	8%	8%	92%	3%	3%

cases they become weaker. This observation provides further evidence why previous outcomes (drafts) should not be ignored, as is done in the contagion-based paradigm, when studying the factors that shape outcomes over time: not only do they introduce large confounding bias to the estimates of the influence of online communications on outcomes, but they themselves constitute stronger causal factors and stronger statistical predictors of the outcomes than the email communications, while evolving more steadily over time and not being particularly affected by confounding bias.

To illustrate a little further how wide the gap is between how strong a causal factor the previous outcome is, and how much weaker the online conversations are, Table 7.4 presents a comparison of the magnitude of causal effects of previous outcomes (I) versus of the email communications (S, P). For economy of space, the table only includes three of the twelve documents, specifically three of the core Recommendation documents: DM, N, and Constraints.

Table 7.4 shows that for all documents presented, the effects of I are so much larger than those of S and P that not only do their magnitudes not overlap at all, but the minimum value of I is much greater than the maximum values of S and P: for DM, the minimum of I is 44% which is around 3 times larger than the maximum of S and P at 14-15%; for N, the minimum of I is 57% which is 3.5 times larger than the maximum of S and P at 16%; and for Constraints, the minimum of I at 50% is 2.8 times larger than the maximum of S and P at 18%. Moreover, the maximum attained by I is 93% for DM, 97% for N, and 96% for Constraints, values that are 5 to 6 times larger than the respective maximum values of S and P for each document. Finally, if one compares median causal effects, in the case of DM the median of I is at least 18.5 times the median of S and P (74% vs. 4%), and similarly for N the median of I is 12 times the median S

and P (96% vs. 8%), while for Constraints the median of I is more than 30 times the median of S and P (92% vs. 3%)

So, in addition to establishing that previous outcomes cannot be assumed ignorable, because they introduce large confounding bias to online social influence estimates, this chapter has dug deeper into the role of this confounding variable, and found that it is actually much more robust to bias, and has a large, positive, and monotonically increasing causal effect on the outcomes over time, that can reach magnitudes higher than 75% in later stages of the document's lifetimes. That is, it is also found that what is talked about most positively (S) and most frequently (P) in the emails matters little, and in several cases matters even less as time passes, while the importance of what is already committed to and prominent in the drafts (I) is, in contrast, large and generally increases fairly steadily over time.

Furthermore, the finding that previous outcomes cannot be ignored as they introduce confounding bias to estimates of the influence of online communications on final outcomes is in line with other studies of how accounting for factors that are often ignored can reduce bias in estimates of social influence, in different settings of online interaction than what is considered here. For example, Aral et al. (2009) and Shalizi and Thomas (2011) show how adjusting for homophily (or for shared personal traits among people with known social ties between them) reduces confounding bias in estimates of social influence, using individual-level outcomes, empirically and using simulations, respectively. In addition, in Aral and Walker (2012) it is found that adjusting for what they define as susceptibility reduces bias in estimates of online social influence, for individual outcomes. Eckles and Bakshy (2017) show that previous outcomes related to the outcome of interest reduce confounding bias in estimates of online social influence, using Facebook data and individual actions as outcomes, and having performed controlled experiments rather than using observational data. Particularly the findings in Eckles and Bakshy (2017) are of most relevance, as they also found, albeit when studying individual actions, that adjusting for prior behaviours 'closely related' to the behaviour (outcome) of interest greatly reduces bias in the estimates of online social influence (peer effects), in their case by 91%. This is very similar to the findings here that adjusting for previous outcomes greatly reduces bias in the estimate of online social influence (from emails), the mean reduction being 406% and the median reduction being 87%.

7.1.2.3 Additional interpretations: inertia and materiality

This section briefly considers some additional interpretations of the finding that the causal effect of the previous draft on the next is very large, especially at the later stages. These interpretations are discussed with reference to the notions of 'inertia' (used in sociology, psychology, and management), and of 'materiality' (from the discipline of sociology).

Inertia. The finding that the causal effect of the previous draft on the next is very large, particularly towards later stages, i.e. that draft contents changed less and less as time passed, might be interpreted as a form of stabilisation, or even of inertia. In Barnett and Carroll (1995, p. 2), a sociological study, the term 'inertia' is defined as the 'logical converse' of the 'phenomenon of organizational change'. It is discussed how research has been conducted on internal organizational change with respect to organizational age, noting the so-called *structural inertia theory* of Hannan and Freeman (1984) as one of the most established works in this area. This theory 'asserts that organizations become increasingly inert over time as procedures, roles, and structures become well established' and 'this implies that the likelihood of organizational change decreases with an organization's age – a prediction that has received empirical support' (Barnett and Carroll, 1995, p. 5). Therefore, this finding is similar to the finding in this chapter about the later-stage stabilisation of document draft contents. However, the focus in the research field discussed in Barnett and Carroll (1995) is largely on change in organizational structure and processes, and not so much on change in the outputs produced by the organisation (which is the focus of the empirical analysis in this chapter).

In addition, as discussed in Alós-Ferrer et al. (2016), in sociology and psychology, the concept of inertia has been used to describe phenomena that relate to 'the tendency to maintain the status-quo', and phenomena of 'resistance to change or the (excess) stability of relationships in societies or social groups'. Similarly, from the discipline of organisation management, Huff et al. (1992) define inertia as 'commitment to current strategy', while Barr et al. (1992) define it as 'pressure for maintaining the status quo'.

While the notions of '*tendency* to maintain the status quo' or of '*commitment* to current strategy' may apply as a characterisation of the later-stage stabilisation observed in the document drafts of this W3C Working Group, it is unclear whether the notion of *resistance* to *change* and *excess* stability, or the notion of *pressure* for maintaining the status quo might apply. The former notions are arguably neutral, descriptive, while the latter notions talk about change, resistance, excess, or pressure, of which we do not have evidence in the data used and analysis conducted in this chapter. That is, based on this chapter's above analysis of the Working Group, there is no evidence that the observed tendency towards stabilisation in the document draft contents was due resistance to change, or that the stability was excessive, or that it was due to pressure. It is outside the scope of this chapter to investigate whether there were any attempts to significantly change the drafts in later stages, and whether those attempts at change were resisted, and whether this was excessive, and/or whether there were pressures to maintain the status quo, and whether this was the reason why the drafts' contents remained mostly stable in later stages, or whether the observed stability was due to other reasons. For example, this observed stabilisation might instead be a positive phenomenon, signifying that the Group succeeded in maintaining its focus and continuity, and in staying within the goals and scope it had defined at the earlier stages of its lifetime, with respect to what content should and should not be in the produced documents.

Further to the above social scientific studies about inertia, Paulus and Nijstad (2003, Chapter 8, p. 166-167) discuss inertia in the context of group collaboration efforts that involve online and offline communication. The specific setting is group brainstorming tasks, where the goal was to maximise the number of unique and novel ideas produced, and it is discussed how the effects of participants' cognitive inertia (which was undesirable in this context) can be mitigated, e.g. by involving participants in multiple simultaneous online dialogues. However, the goal of the W3C Provenance Working Group was not to maximise the number of unique ideas, but rather to produce a formal and coherent vocabulary on the topic of Provenance specifically, therefore, later-stage stabilisation of document contents is not necessarily an undesirable phenomenon to be mitigated here.

Overall, insofar as the term 'inertia' is used in a negative sense, to denote ('excessive' and/or undesirable) 'resistance' to change, or 'pressure' to maintain the status quo, and given that it often refers to organisational structures or social relationships rather than organisational outputs, it is not clear that it would be appropriate to characterize the observed later-stage stabilisation of the document drafts produced by this W3C Working Group as 'inertia' in that sense. However, under the more neutral sense of 'inertia' as 'the tendency to maintain the status-quo' or as 'commitment to current strategy', this term might be an appropriate characterisation.

Materiality. The area of sociology studying *materiality*, i.e. what material artefacts represent in terms of social interactions and processes (e.g. Carlile and Langley, 2013; Dale, 2005), offers another frame in which to examine the finding that the previous document drafts have a large effect on the next drafts. In this frame of thinking, the finding that previous document drafts have large effects on the next drafts can serve to draw one's attention to how it is also the documents themselves, rather than only the email conversations, that carry in them social interactions, collective decisions, and the embedded consensus of the Group accumulated over time, and how these documents also shape the context and structures in which the Group operates.

That is, it is not only the emails but also the documents that are collectively and socially constructed, and that embody the collective decisions and collaborative efforts of the Group. The fact that, as time passes, the previous draft becomes a stronger and stronger cause that shapes the current draft should perhaps not be interpreted merely as inertia, i.e. as the Group becoming less flexible over time, and less willing to pick apart what they have spent so long building as the final deadline draws nearer and nearer. While this might be true, one could also consider that the document drafts are not something other to social interactions; they are a different kind of embodiment (i.e. different from the emails) of social interactions, storing the accumulated efforts and decisions the Group has made collectively over time.

Even though one might more readily think of the emails as representing social communication and real-time and dynamic collaboration, while thinking of the published document drafts as formal and static objects (digital files), the documents drafts too embody the collective interactions and decisions of the Group over time. They too were socially constructed, by the Group, and they also affect the next wave of what the Group will construct.

This field of how material artefacts (the document drafts here) are much more than objects has long been studied in sociology, often in organisational contexts, on topics such as social materiality and embodiment, interrogating how material objects are representations of collective memory and culture, and how they are also constitutive of and affect social relations (Carlile and Langley, 2013; Dale, 2005; the culture metaphor for organisations discussed in Morgan et al., 1997).

This area has also been studied in theoretical fields of sociology like the field of Actor Network Theory (from the discipline of Science and Technology Studies), where it is investigated how non-human agents, i.e. objects and artefacts, could be thought of as having 'agency' and how they can shape social processes (Latour, 2005; Sismondo, 2009).

7.2 Evaluation of the model

This section is about determining whether the causal model proposed in this thesis or the causal model *implied* by the contagion-based paradigm better fits the empirical data of the W3C Provenance Working Group. The causal model proposed in this thesis (Figure 6.2) includes a backdoor path between email Participation (P), and Sentiment (S) and the contents of the next draft (F), through the contents of the previous draft (I). This backdoor path introduces confounding bias to the estimate of the influence of P (and of S) on F. In contrast, the contagion-based paradigm assumes that one can safely ignore other causal factors, without the estimates of online social influence being affected by confounding bias – hence, its causal model includes no such backdoor path.

That is, the contagion paradigm's untested assumption that other causes can be ignored, that they need not be accounted for, when measuring the influence of online communications on outcomes, implies that not accounting for other causes will not introduce bias to the estimate of the influence of online communications. This assumption about the relationship between online communications and other causes can be encoded in a graphical causal model, in which there are no backdoor paths between online communications and the outcome of interest. As noted in the critique (Chapter 3), the problem is exactly that the contagion-based paradigm makes such unwarranted and untested causal assumptions, and makes causal claims about the influence of online communications based on them. The problem with this assumption is that it is unfounded, as it is does not follow from relevant theory nor from empirical evidence on the influence of online (or offline) communications on outcomes.

This thesis tests this assumption empirically, using the W3C Provenance Working Group setting as a case study. It tests the assumption both in terms of the causal effect estimates and confounding bias estimates (where the selection effect reflects the contagion paradigm's estimate which ignores confounders) that result from the application of the proposed graphical causal model of this thesis to the data, in Chapter 7, and additionally, in this chapter, in terms of testing the fit of the causal model proposed in this thesis (with backdoor paths) to this empirical dataset versus the fit of the contagion-based paradigm's causal model (no backdoor paths).[3]

Given the above, this section tests whether it is the causal model proposed in this thesis, or the contagion-based paradigm's implied causal model, that better fits the empircal data of the W3C Provenance Working Group. As presented in Section 2.6.5 of the Background chapter, causal models have testable implications in the data sets they generate. There exists a testing procedure, which uses the d-separation criterion, and allows one to make local tests about particular variables in a given DAG nonparametrically, to determine the fit of a proposed model to the empirical data. As the focus here is testing local differences between causal models, and more specifically the existence or not of backdoor paths between S,P and F, and since the models in this thesis are non parametric (we are not assuming specific relationships, e.g. linear, between variables and specific error distributions), the d-separation tests are appropriate here.

Recalling the relationship between the graphical d-separation criterion (Definition 2.6.2) and for conditional independence (Definition 2.6.3), i.e. that the d-separation criterion states the following: 'If a set Z blocks every path between two nodes X and Y, then X and Y are d-separated, conditional on Z, and thus are independent conditional on Z', this means that one can test for the existence of blocked paths in the graphical causal models, by performing standard statistical conditional independence tests on the

[3] This assumption that other causes can be ignored was also tested more theoretically, in Chapter 5, at the level of the graphical causal model, and drawing upon established findings from the social sciences, for individual-level analyses. Another key unwarranted assumption of the contagion-based paradigm, that online response to some online communications must mean adoption or endorsement of or agreement with the focal item of interest (e.g. a specific topic or idea) discussed in those online communications, was tested and shown to be unfounded conceptually and theoretically in Chapters 3 and 5. Based on this, Chapters 6 and 7 refrained from assuming that online responses are necessarily a meaningful outcome of adoption of or agreement with the focal item discussed in the online communications, and instead used an outcome that was more meaningful in terms of signifying adoption or non-adoption in the context studied (representing the adoption or non-adoption of terms in the frequently used vocabulary, language and exposition style of document drafts). Overall, as noted in the critique (Chapter 3), if one does not want to, or is not able to, measure confounding causes and measure appropriate outcomes denoting adoption or non-adoption of the focal item of interest, then one can only make descriptive (not causal) claims, about the associations (not the causal effects) between online communications and the outcomes measured.

empirical data (Pearl et al., 2016). In this manner, one can assess which of the two causal model candidates (the contagion-based paradigm's one versus the one proposed in this thesis) obeys and which violates these tests, hence which graphical causal model fits the data and which does not, respectively.

Figure 7.4 shows the two competing causal model candidates. Each sub-figure there shows two inter-draft intervals: the one we focus on is the interval between I and F, with the email Sentiment and Participation during this interval being represented by S, P, but the previous interval is also shown, which starts with the draft whose contents are represented by I_{-1} and ends with the draft represented by I, with the email Sentiment and Participation during that interval being represented by S_{-1}, P_{-1}. Therefore, the backdoor paths discussed here will be discussed with reference to the variables of the main interval (I, F, S, P).

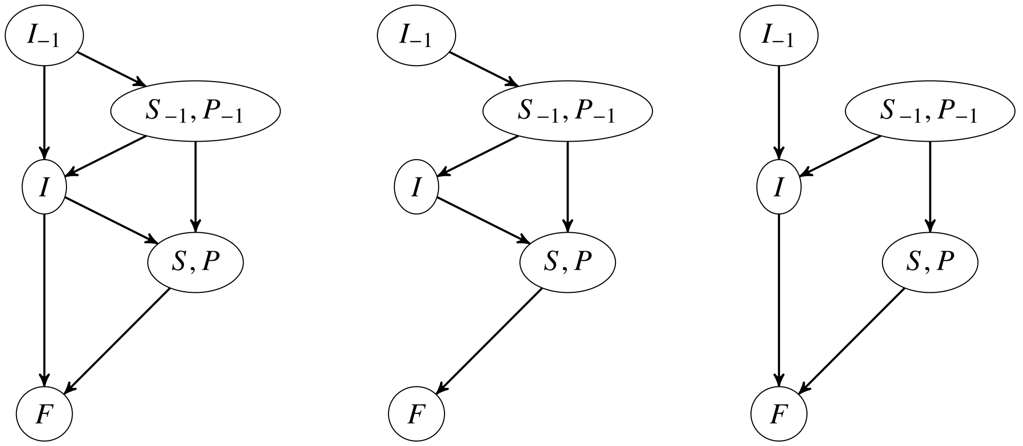
Figure 7.4(a) shows the causal model proposed in this thesis, whereby I is a common cause of S, P ($I \rightarrow S, P$) and of F ($I \rightarrow F$), hence I introduces confounding bias on the estimates of the effects of S and P on F. That is, there is a backdoor path between S, P and F, through I. On the contrary, the contagion-based paradigm assumes that there are no backdoor paths between S, P and F, through any variable, including I. For the latter to hold, it is sufficient that either of the following conditions holds:

- for all inter-draft intervals, there is no arrow $I \rightarrow F$, as in Figure 7.4(b) [4], which leads to Test 1, below, or
- for all inter-draft intervals, there is no arrow $I \rightarrow S, P$, as in Figure 7.4(c) [5], which leads to Tests 2 and 3, below.

The three independence tests that follow from the structure of the competing causal DAGs, when the d-separation criterion is employed, are presented below. Where the S variable (Sentiment) is used in a test, the same test applies and is performed for variable P (Participation) as well.

[4] meaning that, for the previous interval there is also no arrow $I_{-1} \rightarrow I$

[5] meaning that, for the previous interval there is also no arrow $I_{-1} \rightarrow S, P_{-1}$



(a) Proposed Causal DAG

(b) contagion-based paradigm's DAG

(c) contagion-based paradigm's DAG

FIGURE 7.4: Proposed DAG (a) vs. contagion-based paradigm's Implied DAGs (b, c)

Test 1 $F \perp\!\!\!\perp I|S$, which implies that $P(F|S) = P(F|S, I)$, for all values of F, S, I. This does not hold for Figure 7.4(a): conditioning on S does not make F and I independent, because there is also a directed path from I to F. No path from I to F exists in Figure 7.4(b), so this test holds for that causal model.

Test 2 $S \perp\!\!\!\perp I$, which implies that $P(S|I) = P(S)$, for all values of S, I. This does not hold for Figure 7.4(a), as there is an arrow from I to S. In Figure 7.4(c), there is no arrow from I to S, but there is a path between them through S_{-1} ($I \leftarrow S_{-1} \rightarrow S$). So, for this test to hold, there must also not be a path from I to S via S_{-1}, which yields Test 3 below. That is, in Figure 7.4(c), if there is no path from I to S via S_{-1}, Test 2 holds; otherwise Test 2 does not hold, so, to differentiate between the testable implications of the backdoor path in Figure 7.4(c) and the testable implications of Figure 7.4(a), one must also perform Test 3.

Test 3 $S \perp\!\!\!\perp I|S_{-1}$, which implies that $P(S|S_{-1}) = P(S|S_{-1}, I)$, for all values of S, S_{-1}, I. This does not hold for Figure 7.4(a), as there is an arrow from I to S. However, this does hold for Figure 7.4(c), as S_{-1} blocks all paths between I and S (the path through F, $I \rightarrow F \leftarrow S$, does not count, since it does not satisfy the blocking criterion invoked by the d-separation criterion, because F is a collider in this path).

Each of these three tests expresses an *equality* between two probabilities. For each of these tests, if the equality holds, then the contagion-based paradigm holds, i.e. there is no backdoor path through I for the effect of P (or S) on F, hence one should not adjust for I when calculating the social influence (causal effect) of emails (as reflected in the Participation and Sentiment levels) on the produced outcomes (draft contents). If however the equality does not hold, then the test does not hold, as there is a confounding path, and one should adjust for I, per the causal model proposed in this thesis, and so the contagion paradigm's assumption and causal model(s) does not hold. It is noted that for a test to hold, the respective equality must hold for *all values* of all the involved variables; if even one value combination violates the equality, then the test is considered refuted.

These tests are performed for every document, for every inter-draft interval, and for all possible values of all involved variables. For all cases, it is found that none of the tests hold, i.e. it is found that the causal model proposed in this thesis (Figure 7.4(a)) better fits the empirical data than the contagion-based paradigm's implied causal model which assumes there are no confounding paths (Figures 7.4(b), 7.4(c)), as the latter violates dependencies/ independencies in the data.

For example, the results of performing these tests is shown in Table 7.5, for the DM document, for inter-draft interval 2. This table, and the following tables of test results in this section, can be read as follows: for Tests 1 and 3, the first three columns, and for Test 2, the first two columns, show all value combinations for the involved variables, with the outcome variable in the first column. The value for each variable stays the same in the rows below, until there is a horizontal line, below which the value changes. For instance, in Test 1 of Table 7.5, the second row (excluding the header row) represents the subset of data for which $f = 0$ and $s = 0$ and $i = 0$, while the fifth row represents the subset of data for which $f = 0$ and $s = 1$ and $i = 0$. The next column (i.e. the third-from-last column) shows the value of a particular probability, for which the value of the variable of the previous column (i) is irrelevant, hence the values in this column are printed one row higher than the values of the previous column (i). For example, the first row of Test 1 in Table 7.5, at column $P(F = f|S = s)$, shows the value 0.60, which is the value of $P(F = 0|S = 0)$; similarly, the fourth row at this column shows that $P(F = 0|S = 1) = 0.34$. The third-from-last column is compared to the last column, in the second-to-last column, as per the corresponding Test, and as these two quantities are unequal for all Tests, the direction of inequality is shown in the 'sign' column, in the second-from-last column. For example, in Test 1 of Table 7.5, the third row for the last three columns means that $0.6 > 0.16$, where 0.6 is the value of $P(F = 0|S = 0)$ and 0.16 is the value of $P(F = 0|S = 0, I = 1)$.

Table 7.5: Tests of fit of model to data: DM, Interval 2, Sentiment

Test 1						Test 2					Test 3					
f	s	i	$P(F = f \| S = s)$	sign	$P(F = f \| S = s, I = i)$	s	i	$P(S = s)$	sign	$P(S = s \| I = i)$	s	s_{-1}	i	$P(S = s \| S_{-1} = s_{-1})$	sign	$P(S = s \| S_{-1} = s_{-1}, I = i)$
0	0		0.6			0		0.64			0	0		0.84		
		0		<	0.91		0		<	0.71			0		<	0.87
		1		>	0.16		1		>	0.55			1		>	0.81
	1		0.34			1		0.36				1		0.41		
		0		<	0.83		0		>	0.29			0		<	0.5
		1		>	0		1		<	0.45			1		>	0.34
1	0		0.4								1	0		0.16		
		0		>	0.09								0		>	0.13
		1		<	0.84								1		<	0.19
	1		0.66									1		0.59		
		0		>	0.17								0		>	0.5
		1		<	1								1		<	0.66

As an interesting aside, it is noted that, in the test results for Table 7.5, and for the test results of many other documents and intervals (although not for all), not only do the tests not hold, i.e. the equality does not hold, but the inequalities follow a pattern, described in Table 7.6.

Table 7.6: Frequently observed pattern of inequalities in the three Tests

In Test 1:	Knowing also that I=1 increases the chances that F=1, and reduces the chances that F=0, while knowing that I=0 enhances the chances that F=0, and reduces the chances that F=1; and, similarly
In Tests 2, 3:	Knowing that I=1 increases the chances that S=1 and reduces the chances that S=0, and vice versa when I=0.

The pattern of Table 7.6 is illustrated in column 'sign' of each test in Table 7.5, showing the direction of inequality in each test, for each combination of values. Again, this pattern occurs for many documents and intervals, but does not necessarily occur for all intervals of all documents. Overall, the important result is the contagion-based paradigm's causal model violates all tests (there are inequalities, not equalities), hence the contagion-based paradigm assumption that there is not confounding does not hold, therefore the contagion-based paradigm's causal model does not hold; rather, the causal model proposed in this thesis holds.

Similarly, Table 7.7 shows the three tests performed for the P (Participation) variable this time, for the same document and interval. The patterns are exactly the same as those for the S (Sentiment) variable: there are no equalities (that is, none of the tests hold, hence the contagion-based paradigm's causal model does not hold, and the proposed causal model does), and the inequalities follow the pattern described in Table 7.6.

Table 7.7: Tests of fit of model to data: DM, Inteval 2, Participation

Test 1						Test 2					Test 3					
f	p	i	$P(F = f\|$ $P = p)$	sign	$P(F = f\|$ $P = p, I = i)$	p	i	$P(P = p)$	sign	$P(P = p\|$ $I = i)$	p	p_{-1}	i	$P(P = p\|$ $P_{-1} = p_{-1})$	sign	$P(P = p\|)$ $P_{-1} = p_{-1}, I = i)$
0	0		0.52			0		0.52			0	0		0.84		
		0		<	0.8		0		<	0.92			0		<	0.87
		1		>	0.24		1		>	0.79			1		>	0.81
	1		0.41			1		0.48				1		0.41		
		0		<	0.9		0		>	0.08			0		<	0.5
		1		>	0.16		1		<	0.21			1		>	0.34
1	0		0.48								1	0		0.16		
		0		>	0.2								0		>	0.13
		1		<	0.76								1		<	0.19
	1		0.59									1		0.59		
		0		>	0.1								0		>	0.5
		1		<	0.84								1		<	0.66

The same conclusion (that the contagion-based model violates the independencies in the data) holds across time intervals - for example, Table 7.8 shows the results of the three tests when applied to the fifth interval of the DM document, using the Sentiment variable.

Table 7.8: Tests of fit of model to data: DM, Interval 5, Sentiment

Test 1						Test 2					Test 3					
f	s	i	$P(F = f\|$ $S = s)$	sign	$P(F = f\|$ $S = s, I = i)$	s	i	$P(S = s)$	sign	$P(S = s\|$ $I = i)$	s	s_{-1}	i	$P(S = s\|$ $S_{-1} = s_{-1})$	sign	$P(S = s\|)$ $S_{-1} = s_{-1}, I = i)$
0	0		0.59			0		0.85			0	0		0.95		
		0		<	0.95		0		<	0.9			0		<	0.96
		1		>	0.09		1		>	0.8			1		>	0.92
	1		0.39			1		0.15				1		0.75		
		0		<	1		0		>	0.1			0		<	0.78
		1		>	0		1		<	0.2			1		>	0.73
1	0		0.41								1	0		0.05		
		0		>	0.05								0		>	0.04
		1		<	0.91								1		<	0.08
	1		0.61									1		0. 25		
		0		>	0								0		>	0.22
		1		<	1								1		<	0.27

These inequalities in all tests (denoting that the contagion-based paradigms' model can be rejected, and the proposed model in this thesis can be accepted) also hold across documents. For example, Table 7.9 presents the results for the three tests for the Constraints document, for Interval 3, using the Sentiment variable.

Table 7.9: Tests of fit of model to data: Constraints, Interval 3, Sentiment

Test 1						Test 2					Test 3					
f	s	i	$P(F = f\|$ $S = s)$	sign	$P(F = f\|$ $S = s, I = i)$	s	i	$P(S = s)$	sign	$P(S = s\|$ $I = i)$	s	s_{-1}	i	$P(S = s\|$ $S_{-1} = s_{-1})$	sign	$P(S = s\|)$ $S_{-1} = s_{-1}, I = i)$
0	0		0.59			0		0.77			0	0		0.905		
		0		<	1		0		<	0.74			0		>	0.906
		1		>	0.07		1		>	0.80			1		<	0.904
	1		0.64			1		0.23				1		0.44		
		0		<	1		0		>	0.26			0		>	0.30
		1		>	0		1		<	0.20			1		<	0.62
1	0		0.41								1	0		0.095		
		0		>	0								0		<	0.096
		1		<	0.93								1		>	0.094
	1		0.36									1		0. 56		
		0		>	0								0		<	0.70
		1		<	1								1		>	0.38

In Table 7.9, Test 3 again does not hold, as $P(S = s|S_{-1} = s_{-1})$ and $P(S = s|S_{-1} = s_{-1}, I = i)$ are again not equal. So the contagion-based paradigm's model violates this test, while the proposed causal paradigm of this thesis satisfies this test. Moreover, the inequalities here are in the opposite direction than in the previous tables for Test 3, i.e. in the opposite direction than in the frequent pattern of inequalities that has been described in Table 7.6.

This case can therefore serve as a counter-example to the common pattern of inequalities described above. So, what this means is that, here, I has a negative effect on S, for this document and interval. If we look at the bottom half of Test 3 in Table 7.9 (the half for which the outcome S equals 1), the pattern of inequalities here shows that, when $S_{-1} = 1$, $P(S = 1) = 0.56$ ($P(S = s|S_{-1} = s_{-1} = 0.56$). But, out of those cases where $S_{-1} = 1$, for the cases where $I = 1$, the chance that $S = 1$ is actually lower than $P(S = s|S_{-1} = s_{-1})$, at 0.38 (< 0.56); while for the cases where $I = 0$, the chance is higher, at 0.70 (> 0.56). And the same pattern holds when $S_{-1} = 0$. So this means that, no matter how positively a term was previously talked about (i.e. regardless of the value of S_{-1}), a term is more likely to be talked about positively (S=1) if it was *not* prominent in the previous draft (I=0), than if it was prominent in the previous draft (I=1), for this interval of this document. This pattern is the opposite of the pattern of Table 7.6, where the latter holds in all other Test 3 cases presented in this section.

In addition, the Test 3 results of Table 7.9 also offer an example where inequalities may have a smaller than usual margin: here, in the cases where $s_{-1} = 0$, we need three decimal digits to observe the inequality (as opposed to all other cases presented in this section, where two decimal digits where sufficient), and the respective quantities only differ by 0.001, i.e. by 0.1%. This still counts as an inequality. Indeed, it is noted again that, for the test to pass, all the qualities must hold, for all value combinations of all the involved variables; if even one value combination violates equality, this is considered to refute the test.

These three tests have been performed for all intervals of all documents, for both Sentiment and Participation. For economy of space, those results are presented in Appendix F.

Overall, the results of the tests in this section show that the contagion-based paradigm's assumption that there is not confounding does not hold (the contagion-based paradigm's causal model violates all relevant tests: there are inequalities, not equalities, in the tests). Therefore the contagion-based paradigm's causal model does not hold; rather, the causal model proposed in this thesis holds. That is, the contagion-based paradigm's causal model (specifically the no-confounding assumption encoded in it) violates the dependency and independency relations in the empirical data, while the causal model proposed in this thesis obeys the dependency and independency relations in the empirical data, and is hence a better fit to the empirical data.

7.3 Results under different implementation choices

Further to the analysis and findings presented in Section 7.1, which was based on the implementation presented in Section 6.4, the present section investigates how much the findings would vary under different design choices (in terms of the construction of the term corpus). It is found that the results and conclusions of Section 6.4 are robust to the variations in implementation presented here: the conclusions and patterns obtained in the present Section are the same as those in Section 6.4.

That is, both the alternative design choices investigated lead to the same overall conclusions as the main implementation whose results were presented in Section 7.1: the amount of confounding bias present when estimating the effects of email conversations on outcomes is not negligible, but is rather quite large, while the characteristics and patterns (over time and across contexts) of the effects of the emails on outcomes, and of the effects of previous outcomes on next outcomes, are overall the same. Hence, the findings of Section 7.1 are robust to these design variations.

In more detail, to additional kinds of term corpora are constructed: a retrospective corpus of frequent terms, and an expanded corpus. Details on how these are constructed, as well as the resulting analyses and findings when using each, are presented in Appendix E.

7.4 Summary

Overall, this chapter has shown how the ACF can be instantiated to measure the social influence of online communications at the collective level (something that has received very little attention in the literature, which has focused instead on individual-level analyses), resulting in the CCF, using public data from the W3C Provenance Working Group as a case study.

The CCF and its empirical application here contribute the empirical finding that the assumption of the contagion-based paradigm that, in analysing the social influence of online communications on outcomes, other causes can safely be ignored, does not necessarily hold. That is, in the particular setting studied here (the W3C Provenance Working Group archives), this assumption of the contagion-based paradigm that causal factors other than online communications can be ignored, when measuring the influence of online communications on outcomes, is found to not hold. By using the CCF, it has been have found that there is large confounding bias in the effect (social influence) of online communications, due to previous outcomes, hence one cannot assume that other relevant causes (previous outcomes, in this case) can be ignored and study the influence of online communications on outcomes in isolation, as is commonly done in the contagion-based paradigm for social influence online. Rather, in this setting, previous outcomes introduce large confounding bias (spurious associations) to the effect of online communications on the Working Group's outcomes. Hence, in general, other relevant causal factors (such as previous outcomes in this setting) should not be ignored, but rather they should be measured and adjusted for appropriately, otherwise one may reach very different (and biased) conclusions (as actually happens in the setting studied here).

In more detail, the CCF and its application to the W3C Provenance Working Group make several contributions, substantiated by the following findings:

Social influence from online communications is severely confounded. The social influence of online communications is confounded with the effects of other causes (previous outcomes), and the magnitude of confounding bias is far from negligible. This is contrary to the assumption in the contagion-based paradigm that any other causes can be ignored. (Selection effects and causal effects of email sentiment, participation (S,P) on draft contents are different, i.e. red bars vs. yellow bars differ, the mean being 406%, with the median difference being 87%.)

The social influence (causal effect) of online communications is relatively small. The social influence (causal effect) of online communications on the actual outcome of interest is relatively small, as there is another cause (contents of previous draft) that has a much stronger effect on the outcome than social influence from online communications. (That is, the effect of previous draft's contents on the next draft's contents is not only greater than zero, but is actually quite large, mostly above 50%, and near 100% in later stages of documents' lifetimes – this is much larger than effect of S, P on draft contents, which is mostly lower than 30%, and nowhere near 100% at any point.) This cause is also a confounding cause, so assuming that one can ignore it (as per the contagion-based paradigm) only introduces bias to the causal effect of online communications (per the first point above), but it also leads to missing out on a cause that is more important than online communications in influencing the outcome.

Ignoring confounding causes leads to very different conclusions. Ignoring confounding causes leads to very different patterns of the social influence (causal effect) of online communications on the outcome over time, than when accounting for confounding causes. That is, the causal effect (adjusting for confounders) of email communications on the outcome tends to be smaller in later stages than it was in early stages, however the selection effect (ignoring confounders) tends to be higher in later stages than it was in early stages. This is further evidence (in addition to the first point above) that confounding causes cannot be ignored, as ignoring them would yield very different patterns for the the social influence (causal effect) of online communications over time.

Time matters: variation of causal effects over time. The social influence (causal effect) of online communications, and the causal effect of the contents of the previous document, on the outcome, both vary over time: the latter tends to decrease over time, while the former increases, hence the group displays a pattern of stabilisation: as time passes, what is already in the previous draft becomes more and more important in what will be in the next draft. This happens for the majority of documents (subgroups). This is in line with the findings, in the area of organisational sociology, that 'the likelihood of organizational change decreases with an organization's age' (Barnett and Carroll, 1995, p. 5, discussed in Section 7.1.2.3). The variation of specifically the influence of *online* communications on produced outcomes over time has received relatively little attention in the literature.

Context matters: variation of causal effects across contexts. The social influence (causal effect) of online communications, and the causal effect of the contents of the previous document, on the outcome, both vary across documents not all subgroups (each producing a separate document) exhibit the same effect magnitudes, nor the same effect patterns over time. Variation of the influence of online communications across sub-contexts has received little attention in the literature, but this finding indicates that context matters.

Better fit to data. The causal model proposed in this thesis better fits the empirical data, compared to the contagion-based paradigm's (implied) model, which violates dependencies and independencies in the data. This was shown in Section 7.2.

Robustness to implementation choices The above findings on bias levels, causal effect size and patterns over time and across contexts, are robust to variations in the implementation design: Section 7.3 shows that the same conclusions are reached when using a retrospective corpus of frequent terms, and when using an expanded corpus of important terms.

The above empirical contributions demonstrate how the proposed causal framework can be used to empirically examine unjustified assumptions taken as self-evident by the contagion-based paradigm, when analysing observational digital traces of Web-mediated interactions. It has been shown that, in the above case study, these assumptions on which the contagion-based paradigm relies do not hold: there is at least one causal factor (previous outcomes) that introduces large confounding to the estimates of the influence of online social interactions on outcomes.

As discussed in Subsection 7.1.2.2, this result is in line with other studies of how accounting for factors that are often ignored can reduce bias in estimates of social influence, in different settings of online interaction than what was considered here (e.g. Aral et al., 2009; Shalizi and Thomas, 2011, on adjusting for homophily to reduce confounding bias in estimates of social influence; Aral and Walker, 2012 on adjusting for what they call 'susceptibility' to reduce bias in estimates of online social influence). The findings in Eckles and Bakshy, 2017 are especially relevant, as they also found (albeit when studying individual actions, on Facebook, using controlled experiments rather than observational data), that adjusting for prior behaviours 'closely related' to the behaviour (outcome) of interest greatly reduces bias in the estimates of online social influence (peer effects), in their case by 91%. This is similar to the finding here that adjusting for previous outcomes greatly reduces bias in the estimate of online social influence (from emails), the mean reduction here being 406% and the median reduction being 87%.

In addition, by also paying some attention to the confounder itself, it has been found that the confounder is, in this setting, much stronger *both* as a statistical predictor *and* as a causal factor of the outcome, that develops much more steadily over time, and is much more robust to confounding, compared to online social interactions. Therefore, the social influence from online communications is not only confounded, but it is much weaker, and much less steadily evolving over time, and much less robust to confounding, than previous outcomes. Hence, this chapter has shed light on a very important factor, that has been a blind spot in many studies of how social interactions affect outcomes, and which in the setting considered here is a much stronger causal factor of outcomes than online social influence itself. This shows that, contrary to the contagion-based paradigm's assumptions, causes other than online communications cannot be safely ignored – not only may there be other causes that introduce confounding to the estimate of the social influence of online communications on outcomes, with this confounding being large, but also these other causes may even be stronger causes than online communications, and their effects might even be more robust to confounding and more steadily evolving over time (as happens in the setting studied here) than the effects of online communications.

An additional contribution is that the findings of the empirical application of the CCF presented in this chapter (given the implementation details presented in Chapter 6) offer a demonstration of how the CCF proposed in this thesis can be employed in practice, and adapted to settings of Web-mediated or Web-assisted collaboration, and how it enables one to measure and compare the social influence of online communications, versus the effects of other causes, on the outcome of interest, and to test the assumption of the contagion-based paradigm that causal factors other than online communications can be ignored, over time and across contexts, based on observational digital trace data.

Chapter 8

Conclusions and future directions

This chapter begins with an overview of the problem studied in this thesis and of the motivation and context for that problem, of the research questions on which this thesis focuses, and of the thesis statement. It next proceeds to summarize the key contributions and results of this thesis. Following this, the impact and applicability of these contributions and results are discussed. The chapter next presents future directions, and finally closes with final remarks and conclusions.

The focus of this thesis has been the problem of understanding and analysing the social influence of Web-mediated communications (or, for short, online social influence). To that end, it has proposed a causal conceptual and methodological framework for conceptualising, measuring and qualifying the social influence of Web-mediated interactions, using observational (non-experimental) digital data.

This is an important problem, as it has been claimed that social influence drives the spread of behaviours and phenomena as diverse as obesity, loneliness, getting divorced, and political participation, along social ties, in a process analogous to the contagious spread of viruses (e.g. Cacioppo et al., 2009; Christakis and Fowler, 2007; Domingos and Richardson, 2001; Easley and Kleinberg, 2010; Kempe et al., 2005; Nickerson, 2008), to the extent that ensuring a select few trend-setting ‘influentials’ adopt a behaviour would suffice to lead a large population to follow their example and also adopt this behaviour. This contagion-based paradigm for understanding social influence has been very widely used in the literature, including in the literature on online social influence, i.e. on the social influence of Web-mediated communications. If social influence does indeed operate in this contagious manner, then harnessing its power could bring immense benefits to areas like marketing, public policy, and public health interventions. Then, it would be sufficient for one to target a message about a desirable behaviour only at a select few ‘influentials’ in order for a large population to also adopt that behaviour.

However, such claims that social influence (online or offline) operates like a contagious virus, and the practice of interpreting the spread of information, ideas, opinions, actions, or behaviours as evidence of social influence from a particular source, has been critiqued and found to be limited (e.g. Alshamsi et al., 2015; Freelon, 2014; Lerman, 2016; Marres, 2017; Mason et al., 2007; Tufekci, 2014; Shalizi and Thomas, 2011; Watts, 2007). More details on the background of the problem of measuring social influence online from observational data are given in Chapter 2.

Therefore, this thesis set out to address the following research questions:

How can the social influence of Web-mediated human communications be measured, using 'found' observational (non-experimental) digital trace data? What kinds of methods and what kinds of data are needed to measure it?

In order to answer these research questions, and given that the contagion-based paradigm for measuring online social influence using observational data has been criticised for making untested claims based on inadequate data and methods, the thesis statement was formulated as shown below:

This thesis challenges the contagion-based paradigm for understanding and analysing the social influence of Web-mediated communications, and proposes an alternative conceptual and methodological framework. The proposed causal conceptual and methodological framework can more accurately conceptualise, measure and qualify the social influence of Web-mediated communications, while accounting for other relevant causes (social or non-social). It can do this at the individual and at the collective level, based on observational digital trace data.

That is, this thesis proposes a causal conceptual and methodological framework which can conceptualise and measure the social influence of online communications more accurately than the contagion-based paradigm, as it addresses the core limitations of the contagion paradigm. By applying this proposed framework at the individual level, this thesis finds that one should measure and account for, in addition to social influence from online communications, the effects of three other types of causes: personal traits, focal item traits, and external circumstances. And by applying the proposed framework at the collective level, it is found in this thesis that one should measure and account for, in addition to the social influence of online communications, the effects of causal factors such as previous outcomes. Hence, this thesis provides evidence that other causes should not be ignored in the analysis of social influence online, contrary to the untested assumption of the contagion-based paradigm. Rather, since it is found that these causes introduce confounding bias to estimates of online social influence, they should be measured and adjusted for appropriately, so as to remove confounding bias from estimates of online social influence, as much as possible. The framework proposed and applied in this thesis enables one to determine what kinds of claims can and cannot be made based on the available data (i.e. which confounding causal factors are and are not measured in the data), and to assess what additional data would be needed in order to strengthen and/or qualify those claims.

In more detail, the framework proposed in this thesis is based on causality theory and is also informed by the social sciences, constituting a methodological contribution of the kind that is much needed in the emergent interdisciplinary area of computational social science (Counts et al., 2014; Mason et al., 2014; Wallach, 2016). It is demonstrated theoretically and empirically how this framework can successfully address the core limitations of the contagion-based paradigm, and enable researchers to systematically disentangle, measure and qualify the social influence of online interactions, versus those of other causes, at the individual and at the collective level.

8.1 Summary of contributions and results

The above thesis statement is substantiated through the contributions summarised below. In brief, this thesis first contributes an analytical and empirical critique of the contagion-based paradigm. Based on this, the thesis proceeds to formulate an abstract causal conceptual and methodological framework for understanding and measuring the social influence of online communications. Next, the thesis applies this framework, first at the individual and then at the collective level. In both cases, it is found that the assumption of the contagion-based paradigm that the social influence of online communications can be measured without taking other causes into account does not hold. Rather, it is found that other causes can introduce large confounding bias to estimates of the social influence of online communications, and should therefore be measured and appropriately adjusted for, using causal concepts and methods.

For all of the empirical analyses in this thesis, real-world observational digital trace data from the public online archives of a World Wide Web Consortium (W3C) Working Group are used, as a case study, specifically from the Provenance Working Group.

8.1.1 Critique of the contagion-based paradigm for online social influence

Having reviewed the relevant background of how social influence is conceptualised, analysed and measured in the literature, particularly in the context of Web-mediated interactions, in Chapter 2, Chapter 3 presented a critique of the contagion-based paradigm for the social influence of online communications. This critique is made up of two parts: an analytical critique, and an empirical critique through an empirical real-world worked example. This critique makes the following contributions:

Analytical critique of the contagion-based paradigm. The various limitations of the contagion-based paradigm for social influence that have already been noted in the literature, while very valuable, are either generic limitations of analysis practices of social Big Data in general (e.g. Freelon, 2014; Marres, 2017; Tufekci, 2014), or, in studies critiquing the contagion-based analysis practices of online social influence specifically, limitations are covered sporadically and partially, with only a limited subset of limitations being covered by any given study (e.g. Bakshy et al., 2011; Shalizi and Thomas, 2011; Watts, 2007). Therefore, the analytical critique presented in this thesis contributes a new, and more comprehensive, critical analysis and classification of the key conceptual and methodological limitations of the contagion-based paradigm for online social influence. It attempts to systematically capture, classify and analyse what this thesis argues are the key limitations of the contagion-based paradigm, for the problem of understanding and measuring the social influence of online communications. The key problems of the contagion-based paradigm are classified into the following four groups:

Language. In the contagion-based paradigm, several empirical studies define influence as an easily measured social media platform-specific action. However, this kind of definition is not aligned with how influence has historically been understood in the social sciences or in everyday parlance. In addition, this paradigm frequently uses vague, rather than concrete, language, about the 'spread' (rather than the *adoption*, which is implied) of 'ideas' (as a blanket term, even for things such as products or news items that are not ideas).

Assumed cause of outcomes. The contagion-based paradigm assumes that the only possible cause of observed outcomes is social influence from someone in one's immediate or extended olnine social network on the given online communication setting being studied (e.g. social media setting, email setting). In empirical studies using this assumption, one common problematic consequence of this thinking is the *post hoc ergo propter hoc* ('after this, therefore because of this') fallacy.

Assumed meaning of outcomes. Observed outcomes (actions or behaviours) in this paradigm are assumed to signify adoption of or agreement with the topic (e.g. an idea, action, opinion, behaviour) being discussed in online communications, when actually these outcomes are ambiguous and can be interpreted in several different, and often opposing, ways.

Untested assumptions. The assumptions made in this paradigm, often cloaked in vague or ill-suited language (as per the first point above), are frequently stated as self-evident, as facts of nature, when in reality they are assumptions, the validity of which should be tested empirically or otherwise (e.g. theoretically, drawing upon domain expertise or findings). This kind of practice often leads to the *begging the question*, or *circular reasoning*, fallacy in this paradigm, whereby that which is to be empirically proven (e.g. whether someone is an 'influential') is assumed as a given. One important consequence of this problematic reasoning is the issue of what is an 'influential' –

in the usage of this term in the literature, the defining characteristic of being an 'influential' (consistently succeeding in persuading others to adopt what they propose) has been conflated with one of the traits (being well-connected socially) that are often associated with, but not defining of, such people.

Empirical critique through a worked real-world example. Given the key limitations of the contagion-based paradigm, raised in the analytical critique, the empirical critique demonstrates how these limitations manifest if one attempts to empirically apply the concepts and methods of the contagion-based paradigm to a real-world dataset. To assess whether, and to what extent, the methods of the contagion-based paradigm can offer any insights on the social influence (i.e. the causal effect) of online communications on outcomes of interest, this paradigm's most common measures of online social influence (number and network properties of online communications) are employed, using primarily the kinds of online data this paradigm usually considers (online communications data; in this thesis, email communications from the public W3C Provenance Working Group archives). It is found that, while interesting and valuable insights are produced (e.g. the association of the topology of the response network with people's formal roles), these do not allow for any inferences about the social influence of online communications on outcomes of interest. Rather, these insights remain useful as descriptive findings only.

8.1.2 Abstract Causal Framework for the social influence of online communications

Building upon the critique, Chapter 4 develops an abstract causal conceptual and methodological framework (the Abstract Causal Framework, ACF) for conceptualising and measuring the social influence of text-based online social communications, based on observational digital trace data. The ACF can address the key limitations of the contagion-based paradigm, in a manner that is 'abstract' (general), in the sense of not being specific to individual-level or collective-level analyses only, but rather being applicable to both. The ACF is also flexible, as it is can be applied to any setting where the social influence of online communications is to be measured, and is not limited to any one kind of online communications. This abstract framework is comprised of the following conceptual and methodological principles:

1. Having a clear definition of social influence, in line with how it has been understood in the social sciences;
2. Distinguishing outcomes from causes (not conflating observed outcomes with a specific cause);

3. Taking into consideration any other relevant causes behind the observed outcome of interest;

4. Applying causal methods for the analysis and measurement of online social influence (this involves causal modelling, identification of causal effects and assessing confounding bias, estimation of causal effects, and evaluation the fit of the causal model to the empirical data).

The applicability of the ACF to different real-world settings, for individual-level and collective-level analyses, is demonstrated by instantiating it at the individual level (resulting in the Individual-level Causal Framework, ICF) and the collective level (resulting in the Collective-level Causal Framework, CCF).

8.1.3 Individual-level Causal Framework

In Chapter 5, an individual-level instantiation of the ACF is developed, resulting in the Individual-level Causal Framework (ICF). Using graphical causal models, this instantiation shows that the social influence of olnine communications is confounded with the effects (influence) of an expanded list of causal factors than previously discussed in the literature. Moreover, it proposes a set of qualitative characteristics that affect the kinds of claims that can be made about the social influence of online communications. The usefulness and versatility of these contributions is also demonstrated, by applying this individual-level framework to a variety of common online (and some mixed, and some offline) social settings using previous studies.

As mentioned, most of Chapter 5 was published as a poster paper in *Dimitra Liotsiou, Luc Moreau, and Susan Halford. Social influence: From contagion to a richer causal understanding. In* International Conference on Social Informatics, *volume 10047, pages 116-132. Springer, 2016.* (Liotsiou et al., 2016), which was honoured with the Best Poster Award for the accompanying poster.

In more detail, for the problem of analysing and measuring the social influence (causal effect) of some online communications of interest on an individual's observed action or decision (outcome), the framework proposed here covers the space of other types of causes that may lead to the given outcome (namely, similarity of personal traits, traits of the focal item, and external circumstances). It systematically considers and formally addresses (using the rules of graphical causal models) these other causes and any confounding bias they may introduce to estimates of the social influence of the online communications of interest, and also considers how different combinations of these causes can lead to different qualities in the observed outcome. In contrast, previous studies of the social influence of online communications have generally identified and discussed only a subset of these other causes at a time (online social influence versus homophily in Aral et al., 2009; Shalizi and Thomas, 2011; Watts, 2007; versus susceptibility in Aral et al., 2009; versus congitive factors in Lerman, 2016; the issue of interpretation of observed outcomes in Freelon, 2014; Tufekci, 2014). In contrast, and in addition to what existing studies have contributed, the ICF proposed here contributes the following:

- It expands the discussion about the social influence of online communications, from online social influence versus homophily, or versus susceptibility, or versus cognitive factors, to the social influence of online communications versus the influence of all other types of causes, in order to cover the space of other types of causes behind observed behaviours, and offers a classification, or a partitioning, of these causes into classes, while allowing for flexible classification of any specific cause under these classes. This identification and classification of types of causes that might directly affect an individual's actions or decisions is based on established results from the social sciences literature (sociology, psychology, social psychology, cognitive science, social neuroscience, behavioural science, management and marketing).
- It further expands the discussion about the social influence of online communications, from focusing on data where the social ties between people are known, to data where social ties are not necessarily known. Hence, the ICF covers not only the social influence from the online communications of one's known social ties but also from people who may not be in one's social network. This is again based on established findings from the social sciences.
- It uses causal methodology in order to systematically and formally reason about this expanded list of classes (or types) of causes, and to address any bias they may introduce to estimates of the social influence of online communications. This is done by using the formal rules of graphical causal models, in order to:
  - Examine whether these classes of causes (in this expanded and more comprehensive list of causal factors) introduce confounding bias to the estimate of the social influence of online communications, and finds that indeed the estimate of the social influence of online communications is confounded with the effects of each of them; and
  - Determine what is the minimal deconfounding set of these causes that must be measured and adjusted for, in order to recover an unbiased estimate of the social influence of the online communications of interest on the outcome, and finds that this minimal set contains causes from each of the causal classes, i.e. that each causal class must be measured and adjusted for.
- It offers a classification of a range of important qualitative issues related to how different combinations of the social influence of online communications and the effects of these other types of causes can lead to different qualities in the observed outcome, drawing upon social and computational disciplines.

- It performs a demonstration of the above contributions of the ICF, using previous studies of online settings (but also of some offline and some mixed online/offline settings), illustrating the usefulness of these contributions, and the versatility of their applicability, across a variety of real-world practical settings of social interaction and communication.

8.1.4 Collective-level Causal Framework

In addition to the individual-level analysis of online social influence (per the ICF), this thesis has also covered the analysis of the social influence of online communications at the collective level, which is relatively rare in the literature. It is demonstrated how the abstract causal framework (ACF) can be instantiated for collective-level analysis, resulting in the Collective-level Causal Framework (CCF), in Chapter 6. The value and flexibility of the CCF is also demonstrated by showing how it can be applied in practice, using data from a real-world collective setting.

The CCF contributes a set of principles for how the influence of online communications can be conceptualised and measured at the collective level specifically (something which has received little attention in the literature). It tailors the general principles of the ACF to collective-level analysis, by addressing issues that are specific to collective-level analyses, including:

- mapping, modelling, and aggregating any variables captured at the individual-level to collective-level variables;
- determining the appropriate unit of analysis (which is no longer individual people, as it was for the ICF);
- offering a flexible classification scheme for possible confounding causes (causes internal to the collective setting, causes external to the collective setting, traits of the focal item).

The CCF constitutes a generic and flexible contribution, applicable to a wide range of collective settings, i.e. any setting where there is a record of collectively-produced outcomes and of online communications. Such settings may range from professional collaborations and formal projects, to the newer but increasingly ubiquitous kinds of projects in the areas of crowdsourcing and citizen science.

A demonstration of the flexibility and real-world applicability of the CCF is also presented, by showing how the CCF can be applied empirically to a real-world setting of collectively-produced outcomes, using public data, from concepts (e.g. outcomes versus causes), to causal modelling, to variable extraction from the data, and finally to the implementation of causal estimation formulae.

8.1.5 Empirical findings using the Collective-level Causal Framework

The CCF is finally applied to a real-world setting of collectively-produced outcomes (in Chapter 7). It is shown how the CCF can address the limitations of the contagion-based paradigm, and how it enables one to empirically test the untested assumptions of the contagion-based paradigm for the social influence of online communications (as has been called for in the literature, e.g. in Mason et al., 2007; Freelon, 2014; Tufekci, 2014), at the collective level.

The core contribution of the empirical application of the CCF is that the untested assumption of the contagion-based paradigm, that causes other than online communications can be ignored when measuring the social influence of online communications on outcomes, are empirically tested and are found to not hold, in the real-world setting analysed here. This means that causal factors other than social influence from online communications cannot be ignored, but rather should be measured and accounted for, as they may introduce substantial confounding bias into the estimate of online social influence. This is in line with what has been found in the literature for other settings of online interactions (e.g. Aral et al., 2009; Eckles and Bakshy, 2017; Sharma et al., 2015) and in simulations (e.g. Shalizi and Thomas, 2011), and, in addition, this result is established in this thesis over time and across sub-contexts in the setting studied. Another important contribution is the estimation and comparison of the patterns of the social influence of online communications, versus the effects of other causes, over time and across contexts, which has not been done in the contagion-based paradigm literature. The above contributions are substantiated through the empirical application of the CCF to the setting of the W3C Provenance Working Group, and are listed in more detail below:

The social influence of online communications is significantly confounded. The untested assumption made in the contagion-based paradigm, that, when measuring the social influence of online communications, other factors can be ignored, does not hold in this setting - evidence is found that the social influence of online communications is confounded with the effects of other causes, and this confounding bias is far from negligible (mean bias across contexts at 59%, mean bias of per-context means at 400%).

The magnitude of online social influence is relatively small. It is found that the magnitude of the social influence of online communications on the outcome is relatively small (mostly lower than 30%, and is nowhere near 100% at any point), compared to the effect of other causal factors (specifically, the previous outcome, whose effect magnitude is mostly at above 50%, and near 100% at later stages of the Group's lifetime). This causal factor is also one that introduces confounding bias to estimates of social influence of online communications, so assuming that one can ignore it (as per the contagion-based paradigm) not only introduces bias to the estimate of the causal effect (influence) of online communications, but it also leads to missing a cause that is more important than online communications in shaping the outcome.

Ignoring confounding causes leads to very different conclusions. Ignoring confounding causes leads to very different patterns of the magnitude of social influence (i.e. the causal effect) of online communications on the outcome over time, than when accounting for confounding causes. That is, the causal effect (adjusting for confounders) of email communications on the outcome tends to be smaller in later stages than it was in early stages, however the selection effect (ignoring confounders) tends to be higher in later stages than it was in early stages. This is further evidence (in addition to the first point above) that confounding causes cannot be ignored, as ignoring them would yield very different patterns for the causal effect of online communications over time.

Time matters: variation of causal effects over time. The social influence of online communications, and the effect of the previous outcome, on the next outcome, both vary over time: the latter decreases over time, while the former increases, hence this collaborative setting displays a pattern of stabilisation: as time passes, the previous outcome becomes more and more important in shaping the next outcome. This happens for the majority of contexts (sub-groups working on separate document deliverables). This finding is in line with the findings, from the area of organisational sociology, that 'the likelihood of organizational change decreases with an organization's age' (Barnett and Carroll, 1995, p. 5). The variation of specifically the influence of *online* communications on produced outcomes over time has received relatively little attention.

Context matters: variation of causal effects across contexts. The social influence of online communications, and the effect of the previous outcome, on the next outcome, both vary across sub-groups (each working on a separate deliverable): not all sub-groups exhibit the same effect magnitudes, nor the same effect patterns over time. Variation of the influence of online communications across sub-contexts has received little attention in the literature, but this finding demonstrates that context matters.

Better fit to data. The causal model proposed here better fits the empirical data than the contagion-based paradigm's (implied) model, which assumes other causes do not introduce confounding bias and can be ignored, and which is found to violate dependencies and independencies in the data.

The W3C Provenance Working Group dataset that was used for the above empirical application of the CCF was chosen because it offers the advantage that the outcomes of interest are captured in the data, rather than assuming unmeasured outcomes can be inferred from proxies of questionable reliability, as is common in the contagion-based paradigm. An additional advantage of this dataset is that outcomes are captured over time, and also across contexts (different sub-groups of people working on separate deliverables, i.e. on separate documents), enabling one to investigate how the social influence from online communications may vary over time, and across different contexts, dimensions which are often not considered in contagion-based paradigm studies of social influence. Moreover, W3C Working Groups are settings of international collaboration, and lead to the production of Web standards that can shape Web practices and usage worldwide. Hence, it is interesting to investigate the causal factors that shaped the nature and content of these global-reaching standards.

Based on this empirical application of the CCF, it is shown that the assumptions on which the contagion-based paradigm relies may not hold, in line with findings in other settings of online interaction – in this setting, is is found that causes other than online communications should not be ignored, as ignoring other causes would in fact introduce large confounding bias to the estimate of the social influence of online communications on outcomes. Furthermore, it is shown that the effects of these confounding causes themselves on the outcomes of interest may even be stronger than the influence of online communications, and they may also be more robust to bias, and may even follow a much steadier pattern over time and across contexts than the influence of online communications (as happens in this setting).

The above empirical contributions demonstrate how the CCF can be used in practice, to empirically test unjustified assumptions taken as self-evident in the contagion-based paradigm, based on observational digital trace data. This empirical application also demonstrates how the CCF can be employed and adapted to settings of Web-mediated or Web-assisted collective action, and how it enables one to measure and compare the social influence of online communications, versus the effects of other causes, on the outcome of interest, over time and across contexts, based on observational digital trace data.

8.2 Impact and applicability of contributions and results

The contributions of this thesis have an important impact on the field of understanding and analysing the social influence of online interactions, using observational data, as they show how to address the core pitfalls of the contagion-based paradigm, in order to more accurately conceptualise, measure and qualify the effects of social influence from Web-mediated interactions, while accounting for other relevant causes, both at the individual and at the collective level.

In addition, the contributions, including the theoretical and empirical results, of this thesis have a positive impact on, and make a contribution to, the broader and emergent interdisciplinary field of computational social science, for which there have been calls for new methods to be developed for systematically combining the social and the computational sciences (Counts et al., 2014; Lazer et al., 2009; Mason et al., 2014; Wallach, 2016). The causal conceptual and methodological framework proposed in this thesis merges computational methods with conceptualisations and causal assumptions rooted in established social science findings, offering a promising way to address this need in computational social science.

In more detail, the critique, the abstract causal framework (ACF) and its instantiations (ICF, CCF) proposed here are applicable to a wide range of settings where one is interested in studying the influence of online social communications on an outcome of interest.

The points raised in the critique, and the principles of the abstract causal framework (ACF) are general and flexible enough to apply to any setting where one is interested in studying the influence of online social communications on an outcome of interest, whether at the individual or at the collective level. That is, these are applicable to any setting for which there are observational digital traces of text-based online communications and of outcomes.

The individual-level instantiation of the framework (ICF) can be applied to any kind of online communications channel, including email and social media, as has already been demonstrated to an extent through the discussions in Chapter 5 (and, more broadly, in Chapter 3). Indeed, it could potentially be applied to any kind of individual-level decision making which includes consideration of social influence from specific online communications, in online and mixed online/offline settings. The classes of causes (personal traits, focal item traits, external circumstances) are flexible enough to allow an investigator to classify a specific cause as they deem most appropriate for the context they are studying.

The collective-level instantiation of the framework (CCF) can be applied, in a similar manner to its application in this thesis to the W3C Provenance Working Group mailing list and documents, to any organisational setting or project for which there is a similar digital text-based record of online communications, and of the outcomes (project deliverables) and their drafts over time (as discussed in Chapter 6.1.1). As such digital traces of communications and of deliverables are increasingly available in organisations and professional projects, the CCF can potentially by applied to a wide range of settings, in addition to its applicability to the data of any other W3C Working Group. Moreover, alongside more traditional forms of formal projects and organisations, there is the growing field of crowdsourcing and citizen science, whereby members of the public volunteer to participate in open collaborative scientific projects hosted on online platforms, completing tasks in order to achieve an overarching scientific goal (e.g. see Bonney et al., 2009; Silvertown, 2009). Such crowdsourcing and citizen science platforms often have an online communication channel for participants, and may also record outcomes over time. Therefore, this is another type of domain with exactly the kinds of data that the CCF is applicable to, where one could study the effects of social influence from online communications on the outcomes produced at the collective level, over time, and across contexts (e.g. sub-groups) as applicable. And indeed, such citizen science platforms have already attracted the interest of researchers, with existing studies on e.g. the social dynamics and interactions on these platforms, or on the social and other factors that may affect the quality of the collective outcomes produced (Tinati et al., 2015, 2016). In any particular setting, domain knowledge will help determine what actions and outcomes are most meaningful to study, and the CCF is flexible enough to accommodate this.

8.3 Future directions

This section discusses some possible future directions for the causal conceptual and methodological framework proposed in this thesis, first for the Individual-level Causal Framework (ICF) of Chapter 5, and then for the Collective-level Causal Framework (CCF) of Chapters 6, 7.

As discussed, the datasets typically used to measure the social influence of online communications do not adequately capture many, or often any, relevant confounding causes, nor do they typically capture outcomes that unambiguously denote agreement, adoption, or endorsement. So, in future work, in order to make more robust causal claims about the social influence of online communications, it would be worth empirically applying the ICF to datasets that are detailed enough in capturing appropriate outcomes and relevant confounding causes (personal traits, focal item traits, external circumstances) as much as possible. Measurements of confounders and appropriate outcomes could be obtained either by attempting to extract more data from online communication dataset (e.g. if studying social media communications, using a user's profile information and previous posts over a long time, as a potential proxy of their personal traits), or by augmenting online communication datasets with survey questions or interviews of the social media users in the dataset, whether performed digitally or offline. How to best obtain data on confounding variables, with minimal intrusion to the subjects studied, and with appropriate ethical oversight (e.g. in terms of privacy and informed consent), will likely depend on the setting and research questions at hand.

Further, for the ICF (but also for the CCF), it would be worth investigating how to harness social science expertise and domain expertise to devise systematic methods for identifying which specific causal variables for each type of cause are relevant in a given setting and should be measured, and how this may vary across different settings. In addition, since the observed outcome (whose causes we aim to estimate) reflects a possibly subjective decision made by a specific person (or group, in collective-level analysis), this person's choice and interpretation of relevant causes might differ from the investigator's, so it may be worth accounting for this potential difference using social science expertise (e.g. from social psychology).

In terms of the CCF, it would be worth attempting to measure more confounding causes, in order to remove more confounding bias from the estimate of the influence of online communications on outcomes. For instance, in the W3C Provenance Working Group CCF analysis performed in this thesis, the confounding bias of one type of confounder (the previous outcome) is considered, to test whether the assumption of the contagion-based paradigm that other causes can be ignored holds. Additional confounding causal factors might include additional channels or modes of communication or interaction, such as data on various votes during the Group's lifetime, or minutes of meetings or teleconferences. However, such communication channels are not as straight-forward to automatically parse and interpret as the email archives, so care would be needed in determining how to most appropriately make use of such data, and domain knowledge would be particularly helpful in this process. Furthermore, data of offline communications, or in general of participants' views on important factors that shaped the outcomes, could be obtained – for example, by designing and deploying appropriate interviews or surveys, or, in cases where the analysis is performed during the lifetime of a collective effort, having an organisational ethnographer present during the workings of the group could also be useful.

In addition, in the case of the W3C Provenance Working Group studied here, but also in any other similar setting, it may be interesting to investigate whether one's formal role, or one's professional sector, affected whether the content of one's emails would have a strong or a weak effect on the produced document's content. So it may be interesting to study the social influence of the online communications of different sub-groups of people on the collective outcomes, where participants would be grouped around roles, e.g. the chairs, versus of the document editors, versus of the document contributors, versus of the remaining members of the Working Group. Similarly, one might group participants by professional sector, e.g. industry professionals versus academics, and explore and compare the social influence of each of these groups' online communications on the collective outcomes produced. At an even finer granularity, it may be interesting to study the social influence of individuals on collective outcomes. However,when there are many participants involved (e.g. more than sixty individuals involved in the Provenance Working Group's email conversations), this may be impractical for nonparametric estimation (particularly if many-valued variables are used for email Sentiment and Participation, rather than binarised variables), hence a parametric causal estimation method might be needed, such as a linear model (and an assessment of whether the relevant parametric assumptions are applicable in the given setting). Moreover, it is unlikely that individual participants, or groups of participants, acted independently of each other, i.e. the Sentiment and Participation in the emails of each person or group is likely not independent of those of other groups, hence care is needed to consider such dependencies (called *interactions*) when specifying the model. Determining how and to what extent one may be able to disentangle and measure causal effects in the presence of such interactions is still an open research problem (e.g. see Ogburn, 2017).

In terms of calculating the Sentiment expressed in each email of the archives, this thesis used an open-source neural network classifier, trained on a large public dataset of labelled movie reviews from IMDB. This approach has certain limitations (as is also noted in Appendix B.2), relating to the sequence length limits used and the trade-off between such a limit and the speed of training the neural algorithm, and to the generalisability of the training on movie reviews to the context of emails, both in terms of context and in terms of the open-source code only using highly polar reviews. As discussed, these limitations of the Sentiment measure used in this thesis are not especially problematic for the overall aims and scope of this thesis, as Sentiment is only used as a complementary, and more qualitative, measure of the contents of the emails, in addition to the more quantitative, volume-oriented, Participation measure. Investigating how and to what extent these limitations might be addressed is one possible future direction. Therefore, one possible future direction may involve exploring the best trade-off of training time versus information content (i.e. attempting to identify the maximum number of words of all emails that can be retained, without severely impacting the run-time of the training and application of the model). Another possible direction would be to attempt to account more for context of the online communications in the given setting (emails here), by obtaining labels for them. This is a challenging problem, as one commonly method used to obtain labels for large datasets of text (9000 emails, here), crowdsourcing, is likely to not yield reliable results when the content to be labelled is highly technical, jargon filled, and nuanced with notions from a very specific and complex area of expertise (as is the case for the Provenance Working Group's emails).

Regarding the construction of the term corpus, in the analysis in this thesis, the term corpus was made up of single words from the documents and emails. This could be extended to also include 2-word phrases, or 3-word phrases, or n-word phrases (called 'n-grams', in the field of Natural Language Processing), if one deems that it is also worth including frequently-occurring phrases in the analysis, in the given setting. In addition, when extracting the emails that are relevant to each term in the corpus, this thesis only considered the subject line of each email thread: if the given term from the corpus appeared in it, then all the emails in that thread were considered relevant to the given term. Instead of this binary indicator of relevance, one might want to use degrees of relevance, by considering an email relevant if a term appears more than a specific number of times (a cut-off threshold) in the body and subject of the email. This cut-off threshold would determine the number of term occurrences (absolute or relative to the length of the email) above which an email qualifies as relevant to a term. Care would then be needed in determining an appropriate cut-off threshold. Finally, another way of constructing the term corpus could be to account for the views of the Group's participants, e.g. by surveying participants, and aggregating their answers, about which topics, and associated terms, they think of as important, at each stage (draft) of each document. As a starting point, one could consider Moreau et al. (2015), written by the Group's chairs and three document editors, which offers a retrospective analysis of the key requirements that shaped the produced document drafts. Such methods that account for the views of the participants could also be used as a way of evaluating the top terms identified for each inter-draft interval of each document by the automated method used in this thesis.

8.4 Final remarks and conclusions

In conclusion, this thesis set out to answer the questions of whether, and how, the social influence of Web-mediated human communications can be measured, based on observational (non-experimental) digital trace data.

It has successfully answered these questions, by challenging the popular but flawed contagion-based paradigm for analysing the social influence of Web-mediated communications, and proposing an alternative causal conceptual and methodological framework. This proposed framework can more accurately conceptualise, measure and qualify the social influence of Web-mediated textual communications, while accounting for other relevant causes. It can do this at the individual and at the collective level, based on observational digital trace data.

That is, this thesis, starting with a critique of the contagion-based paradigm, has proposed a causal conceptual and methodological framework (ACF), with two instantiations, at the individual-level (ICF) and the collective level (CCF). This framework can conceptualise and measure the social influence of online communications more accurately than the contagion-based paradigm, as it addresses the core limitations of the contagion paradigm. By applying the proposed framework, this thesis finds that one should measure and account for, in addition to social influence from online communications, an expanded list of three types of causes (external factors, traits of focal item, personal traits), in individual-level analyses, and causal factors such as previous outcomes, in collective-level analyses.

Therefore, this thesis has provided evidence that, contrary to the untested assumption of the contagion-based paradigm, these causal factors should not be ignored, as they can introduce substantial confounding bias to estimates of social influence from online communications. Further, the effects of these causal factors on the outcome may even be stronger that those of online communications, more robust to confounding, and more steadily evolving over time. So, other causes should instead be measured and adjusted for appropriately, in order to remove confounding bias from estimates of online social influence, as much as possible, and to not miss causes which may themselves have a stronger impact on the outcomes of interest.

The critique and causal conceptual and methodological framework proposed and applied in this thesis enables investigators to critically think about and determine what kinds of claims can and cannot be made based on the available data (including which confounding causal factors are and are not measured in the data), and to assess what additional data might be needed in order to strengthen and/or qualify those claims.

Overall, the critique and the causal framework proposed in this thesis, based on causality theory and informed by the social sciences, constitute a methodological contribution of the type that is much needed in the emergent interdisciplinary area of computational social science. This thesis has demonstrated theoretically and empirically how the proposed framework can successfully address the core limitations of the contagion-based paradigm, and how it enables researchers to systematically disentangle, measure and qualify the social influence of online communications, versus the effects of other causes, and address the issue of confounding bias, over time and across contexts, at the individual and at the collective level, based on observational digital trace data.

Appendix A

Data preparation for the empirical critique

This appendix presents the preprocessing steps that needed to be taken in order to prepare the data for analysis of influence patterns, as well as the tools used for data preparation, processing and analysis, for Chapter 3.2.

The internal structure of the public directory in which the W3C Provenance Working Group emails are stored is the following: there is one directory for each month, and within each month directory the emails that were sent in that month are stored, ordered by the day and time they were sent, as HTML files. The file name for each HTML email file starts at `0000.html` for the first email sent in that month, with `0001.html` being the next email sent in that month, and so on, until the last email sent in that month having the largest number as its filename.

At the bottom of each HTML email file it is stated that these email archives were generated with the Hypermail program.[1] To the best of our knowledge, there is no API to parse Hypermail-generated email archives and extract useful fields, so a new parser script was written to extract information of interest from these emails. (Indeed, it seems that many, and possibly all, other public W3C WG email archives were generated by HyperMail,[2] so assuming HyperMail was deployed in the same way in the other public W3C Working Group email archives, the parser written for this analysis should also work for the other public W3C Working Group email archives.)

For each email, the following attributes were determined to be of interest for the analysis of online social influence here, and so the parser extracts a value for each of these attributes:

[1] 'Hypermail is a free (GPL) program to convert email from Unix mbox format to html' (http://www.hypermail-project.org/)

[2] See http://www.w3.org/Search/Mail/Devel

- Author name and email address,
- Subject,
- Date, and isoSent date-time, a time zone-adjusted time,
- Message ID (MID), the unique identifier of this email message,
- To recipient(s), the recipients in the 'To' field of the email,
- CC recipient(s), the recipients in the 'CC' field of the email,
- In-Reply-To message with message ID, the MID of the email to which this email replies,
- Maybe-in-Reply-To message, with message ID, the MID of the email to which this email 'maybe' replies (see below for what 'maybe' replies are),
- Next in thread message, with message ID, the MID of the next message in this thread,
- Replies to this message (message IDs), the message IDs of the emails that were sent in reply,
- Maybe-replies to this message (message IDs), the message IDs of the emails that were sent in 'maybe' reply (see below for what 'maybe' replies are),
- Message body, the contents of the email message.

These attributes were recorded in a semi-structured way in the HTML file for each email, and regular expressions were used to extract the values of these fields for each email.

Note that 'Maybe-in-Reply-To' and 'Maybe-replies' differ from their non-'maybe' counterparts in the following way: Hypermail deduces the non-maybe replies by using the 'In-Reply-To:' email header if it is available otherwise it uses the 'References:' header. If neither of these is available, it searches for previous emails with the same subject header, and matches are listed as 'maybe' replies. However, if the 'linkquotes' option is used when running the Hypermail program, these rules are overridden in favour of also looking at the text being quoted in the email body.[3]

As it is not known whether Hypermail was run with the 'linkquotes' option or not, one cannot be sure exactly how the Replies and Maybe-replies were inferred. By manual inspection of the Maybe-replies in a sample of emails, it looks like Maybe replies are indeed linked to the emails they are quoting in the email body, so they can be reliably

[3] For more details see the relevant section in the Hypermail documentation at http://www.hypermail-project.org/hypermail-faq.html#17, although a full description of the algorithm is not included.

considered proper Replies. So in the analysis of email replies in this setting, Maybe-replies as treated in the same way as normal replies.

Each email file also contains a link 'Next' and 'Previous', linking to the next and previous emails in that directory (i.e. linking to the chronologically next and previous emails sent during that month,regardless of sender, subject line, or thread). For example, email file `0021.html` in the `2012Dec` directory has `0022.html` as 'Next' and `0020.html` as 'Previous', as email files are numbered based on the date and time when they were sent. However, these relations are not semantically significant in the analysis of influence as response, so these links are not relevant here and are not considered.

Emails often do not include links to replies that were sent in later months, but the reply email does have a field saying which email it replies to. The HTML file representing each email may not display the replies that the email in fact has: for example, it is common to have the HTML file of email A not showing any replies for A, but in fact the file of an email B showing that B is in reply to email A, with email A being quoted in the body of B, and this being corroborated by the subject lines and contents of the emails. This happens when the reply email B was sent during a different month than A, which means it is archived in a different directory on the online archives. Directory names are in the from '`YYYYMMM`', for instance '`2013Jan`', and each email is represented as an HTML file within the appropriate directory of the month when it was sent. The names of the HTML email files are integers showing the time order in which the emails were sent each month, starting from '`0000.html`'. Emails that reply to each other and were both sent in the same month are linked to each other not through the use of each email's unique message identifier (MID), but rather through their filenames in the directory structure (e.g.'`0012.html`'), taking it as a given that they were sent in the same month hence are stored in the same directory. If two emails were sent in different months, they are stored in different directories, so then they are linked using the unique MID. However this link tends to only appear in the 'In Reply To' field of the reply email B, pointing back to the original email A that was sent at an earlier month, but the original email A does not contain a link to email B in its 'Replies' field. In the representation of the email archives used for the purposes of this analysis, in order to ensure that all emails are correctly linked to all other emails, i.e. they are connected to all their replies, and they are also connected the email they reply to, the unique MID is always used to reference emails when linking them. Hence, a first step was performed, to get all emails, and extract their fields of interest, including the 'Replies' and the 'In Reply To' field. Then, a second step was performed, to link the emails correctly.

In terms of programming tools used for data cleaning and analysis, Python was used for parsing, data preparation, modelling and analysis. The Python NetworkX[4] library was used to represent and to visualise person-to-person email responses, and to process those

[4] https://networkx.github.io/

graphs (e.g. to calculate centralities). The Numpy,[5] Scipy[6] and Matplotlib[7] libraries were used to perform the analysis and visualise results.

[5] http://www.numpy.org/

[6] http://www.scipy.org/

[7] http://matplotlib.org/

Appendix B

Data preparation for the empirical colllective-level causal analysis

This appendix describes the details of how the text of the W3C Provenance Working Group document drafts and emails was cleaned and processed in order to extract the features (variables) of interest, further to Chapter 6.4.

B.1 Processing the document drafts

Algorithm 1 Extract features from documents

```
documents ← [‘AQ’, ‘Constraints’, ‘DC’, ‘Dictionary’, ‘DM’, ‘Links’, ‘N’, ‘Ontology’,
    ‘Overview’, ‘Primer’, ‘Sem’, ‘XML’]
for each document in documents do
    drafts ← getDrafts(document)
    for each draft in drafts do
        draft ← cleanDraft(draft)
        $\text{topTerms}_{\text{document, draft}}$ ← getTopTerms(draft)
    end for
end for
```

Algorithm 1 calls the function `cleanDraft` on each draft of each document. This function cleans the text of the respective draft. That is, the text of each draft is cleaned, in order to keep only the main body of content: the editors and contributors credits from the top are removed, as are the ‘PROV Family of Documents’ and ‘Status of this Document’ sections and the table of contents. The acknowledgements, change logs, and references sections at the bottom of each draft are also removed. The frequently appearing 'Hide All' hyperlinks are removed, and words that are only made up of numbers (e.g. representing publication dates of drafts) are also removed. Then, the remaining contents are lemmatized, using the WordNet Lemmatizer[1], on each word of the contents, first using the verb option (`wnl.lemmatize(word, 'v')`), such that tense inflections or gerunds of the same verb are mapped to the present tense of the same term, and then the resulting word is lemmatized again, without the verb option, such that noun forms of a word, such as singulars and plurals, map to the same word in the singular number.

Once this is done, Algorithm 1 calls the function `getTopTerms(draft)` in order to extract the most frequent terms in the document, which works as follows: as per the Python skicit-learn package user guide for text feature extraction,[2] the `CountVectorizer` is used in order to both remove so-called *stop words* and to extract occurrence frequencies for each words, using the following parameters: `count_vect = CountVectorizer(stop_words= 'english', vocabulary= None, max_features= 100, max_df= 300, min_df= 1, token_pattern= r'\w{3,}')`. Here, the parameter `stop_words = 'english'` means that the built-in stop word list for English is used.[3] *Stop words*, which are known in linguistics as *function words*, denoting words with little lexical meaning that are used to express the grammatical relationships among words in a sentence (such as 'a', 'the', 'is', 'at', 'which', 'on'). The stop words used here are the standard stop words in the Python scikit-learn feature extraction package.[4] The parameter `token_pattern = r'\w{3,}')` means that only words that are three characters long or more are considered, while shorter words are discarded. The `maxfeatures` parameter limits the number of top frequent features (words) to consider, to 100 in this case. The `max_df` parameter means this procedure should ignore the features that appear more than `max_df` times, as such words can be considered context-specific stop words that appear too often and hence have little specific meaning. Similarly, the `min_df` means this procedure should ignore the features that appear less than `min_df` in the text (1, in this case, which is the default value), the rationale being that those features (words) are not important enough as they do not appear often enough; this parameter has the default value of 1, and it does not really make a difference as it means words that occur less than once are ignored.

Using this count vectorizer results to a count of occurrences for all words (that are not stop-words) in the draft. This count is then converted to a frequency by normalising it

[1] that is, `WordNetLemmatizer` from Python's module `nltk.stem.wordnet` of the Natural Language Toolkit (http://www.nltk.org/index.html)

[2] http://scikit-learn.org/stable/modules/feature_extraction.html#common-vectorizer-usage

[3] As documented at http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#sklearn.feature_extraction.text.CountVectorizer.fit_transform

[4] This can be found at https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/feature_extraction/stop_words.py

using the length of the draft. Finally, out of these words, only the words whose frequencies are in the top 80$^{\text{th}}$ percentile of the frequencies for these words are kept, in order to have a relative (i.e. percentile-based, rather than an absolute) frequency threshold for all drafts, which helps account for variations in the length of drafts.

B.2 Processing the email archives

This section describes how the emails in the archive were cleaned and processed. All emails of the relevant editors and contributors are considered.[5]

Algorithm 2 Extract features from emails

```
threads ← getThreads(emailArchives)
for each thread in threads do
    for each email in thread do
        email ← cleanEmail(email)
        $\text{participation}_{\text{email}}$ ← calculateParticipation(email)
        $\text{sentiment}_{\text{email}}$ ← calculateSentiment(email)
    end for
end for
```

Algorithm 2 describes how emails are processed, in order to extract a Sentiment and a Participation value for each email message. First, some pre-processing takes place. The `cleanEmail` function removes from the body of each email any lines that are quotes from previous emails (starting with the '>' symbol).

Then, for Participation, Algorithm 2 uses the `calculateParticipation` function, which returns the number of words in the body of the email. For Sentiment, it uses the function `calculateSentiment`, which employs a machine learning classifier that yields a probability that the sentiment of the email message is positive. When trying to determine the sentiment of text, one simple approach is to perform bag-of-words classification, which treats the text as a set (i.e. 'bag') of words, disregarding their order, and results in the overall sentiment of the text being the aggregate of the sentiment value of each word (Le and Mikolov, 2014). As a way to account for the order of words in text, and hence to improve upon bag-of-words methods, a type of deep neural network has been proposed, the so-called *recurrent neural network (RNN)* which performs well with *sequences* of

[5] Some emails, however, even thought they mentioned a given term in their subject line, were not discussions directly around a topic or term, but rather emails regarding formal procedures of the group, such as calls for document reviews, reviewers' submissions, meeting minutes sharing. These kinds of emails are not separated from the emails representing discussions about topics or terms in this analysis, as they are considered to still be about the given term which they mention in their subject line, even if less directly.

data, hence text (sequences of words) is an application area where they perform well (Graves et al., 2012, Chapter 1). More specifically, there are RNN architectures that improve upon a known limitation of RNNs, namely information that appears early in the input sequence being gradually 'forgotten' as one moves along later and later layers, and hence nearer the output, of the neural network (this is known as the *vanishing gradient problem*: the gradient either goes to zero or blows up exponentially), meaning contextual information from early parts of the sequence plays too little a role in helping analyse later parts of the sequence (Graves et al., 2012, Chapter 4). Long Short-Term Memory (LSTM) architectures of RNNs provide a way to address this problem, by maintaining a longer range memory, which is very useful in problems that benefit from the use of long range contextual information (an overview of supervised machine learning for sequential data, recurrent neural networks, and LSTM networks can be found in Graves et al., 2012). This network is first trained on a pre-existing dataset, know as the Large Movie Review Dataset, or the IMDB dataset,[6] consisting of 25,000 movie reviews (Maas et al., 2011).[7] The neural network model that is obtained after this training is then applied to the email archives, in order to obtain a sentiment value for each email. More specifically, the bidirectional LSTM model training code, and code used to load the IMDB data to train it, can be found at the open source code examples repository.[8]

This approach to measuring sentiment has the following limitations. Firstly, the IMDB dataset is comprised only of highly polar data: it only contains movie reviews with very negative scores (scores lower than or equal to 4 out of 10) or very positive scores (scores greater than or equal to 7 out of 10), with any reviews with scores between those values, i.e. with more lukewarm or neutral scores, being excluded. This means that this highly polar dataset is likely quite dissimilar to the Provenance WG email archives, on which the IMDB-trained model is applied, as no such scoring and exclusion of neutral emails has been done for the emails, However, based on manual inspection, it is likely that several emails will indeed be neutral in terms of expressed sentiment, and hence the model may not be well equipped to deal with such emails. More generally, it is known in the field of machine learning with neural networks that it is difficult for a neural network model trained in one context to perform very well in a different context, as such models tend to not generalise very well to different contexts. Attempting address this, and to establish whether there might be ways of using some labelled subset of the Provenance

[6] Per http://deeplearning.net/tutorial/lstm.html

[7] The dataset can be found at http://ai.stanford.edu/~amaas/data/sentiment/

[8] https://github.com/fchollet/keras-resources, specifically under 'Bidirectional LSTM on the IMDB dataset' at https://github.com/fchollet/keras-resources#working-with-text for using the Python Keras deep leaning package (Chollet et al., 2015): the model code can be found at https://github.com/fchollet/keras/blob/master/examples/imdb_bidirectional_lstm.py, the dictionary mapping words to integers, for converting word sequences of the IMDB data to integer sequences for the LSTM model to use as input, can be found at http://www.iro.umontreal.ca/~lisa/deep/data/imdb.dict.pkl.gz, and the code for loading the IMDB dataset to train the model at https://github.com/fchollet/keras/blob/master/examples/imdb_bidirectional_lstm.py (with the modification that instead of using `max_features = 20000`, instead the maximum value number in the index is used here as the `max_features value)`.

Working Group email data to fine-tune the parameters of the neural network in a way that will make it generalise better to the context of these emails, would be a challenging problem. In order to better adapt the IMDB-trained model to the context of the emails, in future work, it may be worth investigating the inclusion of neutral-labelled reviews in the training set, and whether it would be effective to use a sample of labelled emails in an attempt to fine-tune the model parameters in a way that better accounts for the context of the emails. However, the whole email dataset only consists of around 9,000 emails, so it is much smaller than the training set of 25,000 reviews, while labelling the sentiment in all the emails is much less straightforward than the review labels: not only do the emails not come with labels (like the reviews do with the score out of 10), but even if attempting to crowdsource labels (as 9,000 emails is too large a dataset to label manually), the discussions in the emails are of technical nature, with provenance-specific jargon used, so it would require unusually patient crowdsourcing workers, with adequate relevant technical knowledge, in order for them to be willing to spend time reading through the technical email contents and to determine the sentiment expressed, so as to obtain reliable sentiment labels. Therefore, crowdsourcing labels would require significant time, effort, as well as a monetary budget, as it would require very careful design and monetary incentive planning, to entice workers to not abandon the task due to its technical nature and jargon and to produce sentiment labels of good quality. And even then, the labels may not be reliable, given the nuanced, technical, and jargon filled content of the emails. This problem of obtaining truthful and good quality labels from crowdsourcing workers, under budget constraints, is an important challenge in the field of crowdsourcing research (e.g. Allahbakhsh et al., 2013; Gao et al., 2015; Kamar and Horvitz, 2012.)

Another limitation is that, in the open-source code that trains the neural network model on the IMDB reviews, the default maximum sequence length used is 100 words (setting such a maximum limit helps reduce the dimensionality and speed up the training). This means that sequences (reviews) with fewer than 100 words are padded to 100 words with non-entity values, i.e. values that do not affect the sentiment content, and sequences longer than 100 words are truncated to 100 words. The same truncation limit is used when the trained model is applied to the email dataset. However, some of the emails to which this model is applied will be longer than 100 words, so that extra information will be lost. (Therefore, in future work it may be worth exploring the best trade-off of training time versus information content, i.e.what is the maximum number of words of all emails that can be retained, without severely impacting the run-time of the training and application of the model?)

In addition, as we shall see in section 7.1, the causal effects of Sentiment on the contents of the next document draft are very similar to the effects of Participation (the number of words in the emails) on the contents of the next draft. And indeed, for each term in the term corpus, the values of Sentiment and Participation are often very similar.

From a brief inspection, it seems that the Sentiment model gives more positive values the longer the input text is.

The above limitations are worth noting, but they are not particularly problematic for the scope of this thesis overall: Sentiment is used just as another measure of the emails' contents, in addition to Participation. It is acknowledged that it is not a very sophisticated measure, and its limitations are noted.

In summary, it is noted that the Sentiment measure used here has certain limitations, but these are not especially problematic for the overall aims and scope of this thesis, as Sentiment is only used as a complementary, and more qualitative, measure of the contents of the emails, in addition to the more quantitative, volume-oriented, Participation measure.

B.3 Putting it all together

Algorithm 3 shows how the emails and documents data are put together, in order to obtain values for the causal variables used in the analysis of the effects of the emails and the document drafts on the contents of the next draft, for each inter-draft interval of each document. Lines 7-16 show how I and F are calculated for each interval, based on whether a given term from the term corpus appears in the top terms of the initial and final draft of this interval. Then, the algorithm proceeds to calculate the Sentiment and Participation of emails from all threads that relate to a given term from the corpus. The `getThreadsPerTerm` function at line 21 determines if the emails of a thread a relevant or not to each term in the corpus, based on whether the term appears in the subject of that thread. First, it lemmatizes thread's subject line, using the same procedure as in Algorithm 1, using two passes of the WordNet Lemmatizer, first with the verb option (to map all verb tenses to the present tense) and then without (to map non-verb forms, e.g. noun plurals, to the singular number). Then, if a (lemmatized) term is mentioned in a thread's (lemmatized) subject line, that thread (i.e. all the emails in that thread) is added to the thread that are about that term. (In future work, this function could also consider the body of the emails in the thread, and instead of having an indicator variable, i.e. a binary indication of presence or absence, there could be a score of how many times a terms appears to indicate how relevant an email or thread is, if moving beyond a binary indication is deemed more useful.)

On line 24, Algorithm 3 keeps only the emails that were sent during this inter-draft interval. Next, on line 25, the function `wasSentByEditorContributor` limits the emails considered to those that were sent by an editor or a contributor of the draft at the end of this interval (as this is the draft-in-progress discussed during this interval). As found in Section 3.2.2.2, editors and contributors were the group members that were most active in the email conversations throughout the duration of the group's effort. This choice is

Algorithm 3 Aggregate extracted features, calculate all variables needed, from documents and emails

```
documents ← [‘AQ’, ‘Constraints’, ‘DC’, ‘Dictionary’, ‘DM’, ‘Links’, ‘N’, ‘Ontology’,
    ‘Overview’, ‘Primer’, ‘Sem’, ‘XML’]
for each d in documents do
        intervals ← getIntervals(d)
    for each i in intervals do
        $draft_I$ = getDrafts(document)[i[0]]
        $draft_F$ = getDrafts(document)[i[1]]
        for each t in topTerms do
            if term in getTopTerms($draft_I$) then                    ▷ $I$ variable
                $initial_{t, d, i}$=1
            else
                $initial_{t, d, i}$=0
            end if
            if term in getTopTerms($draft_F$) then                    ▷ $F$ variable
                $final_{t, d, i}$=1
            else
                $final_{t, d, i}$=0
            end if
            threads ← getThreads(emailArchives)
            sentimentValues ← []
            participationValues ← []
            for each thread in threads do
                threadsPerTerm ← getThreadsPerTerm(term)
                if thread in threadsPerTerm[term] then
                    for each email in thread do date = getDateSent(email)
                        if date in interval then
                            if wasSentByEditorContributor(email, interval) then
                                sentimentValues.append($sentiment_{email}$)
                                participationValues.append($participation_{email}$)
                            end if
                        end if
                    end for
                end if
            end for
            $sentiment_{t, d, i}$ ← average(sentimentValues)
            $participation_{t, d, i}$ ← average(participationValues)
        end for
    end for
end for
```

made here because the focus of the causal analysis is to determine the factors affecting the contents of the produced documents, and being formally credited as an editor or contributor of a document is taken as evidence that one did indeed contribute, in some way, to that document - therefore, it is these people that are included in the analysis, and their emails are studied. This is done in an effort to ensure relevance (e.g. not using the emails of someone who was not involved in this document at all, who may have been talking about the same term but for another document), and for dimensionality reduction (for each document, keeping only the data for the people who we know were actually involved in the creation and evolution of that document).

Then, Algorithm 3 uses the Sentiment and Participation values for each email which were calculated in Algorithm 2. However, in order to calculate the effect of a causal factor on an outcome using the nonparametric Average Causal Effect formula (Equation 2.7), the cause and the outcome must be binary. This holds for the outcome variable, F, and for the causal factor I, but not for the causal factors S and P, which are continuous variables. So, S and P need to be binarized, using some threshold. This process is presented in Algorithm 4.

Algorithm 4 Post-processing after feature extraction: Binarize S, P variables

```
 1: documents ← [‘AQ’, ‘Constraints’, ‘DC’, ‘Dictionary’, ‘DM’, ‘Links’, ‘N’, ‘Ontology’,
    ‘Overview’, ‘Primer’, ‘Sem’, ‘XML’]
 2: for d in documents do
 3:     for i in intervals do
 4:         $\text{sentimentAllTerms}_{\text{d, i}} \leftarrow []$
 5:         $\text{participationAllTerms}_{\text{d, i}} \leftarrow []$
 6:         for t in terms do
 7:             $\text{sentimentAllTerms}_{\text{d, i}}\text{.append}(\text{sentiment}_{\text{t, d, i}})$
 8:             $\text{participationAllTerms}_{\text{d, i}}\text{.append}(\text{participation}_{\text{t, d, i}})$
 9:         end for
10:     end for
11: end for
12: for d in documents do
13:     for i in intervals do
14:         $\text{median}_{S\,\text{d, i}} \leftarrow \text{median}(\text{sentiment}_{\text{d, i}})$
15:         $\text{median}_{P\,\text{d, i}} \leftarrow \text{median}(\text{participation}_{\text{d, i}})$
16:         for t in terms do
17:             $\text{sentimentBinary}_{\text{t, d, i}} \leftarrow \text{binarise}(\text{sentiment}_{\text{t, d, i}}, \text{median}_{S\,\text{d, i}})$ ▷ S variable
18:             $\text{participationBinary}_{\text{t, d, i}} \leftarrow \text{binarise}(\text{participation}_{\text{t, d, i}}, \text{median}_{P\,\text{d, i}})$ ▷ P
    variable
19:         end for
20:     end for
21: end for
```

Algorithm 4 first collects all Sentiment (and Participation) values in a list (lines 2-11), in order to derive the binarisation threshold from these values - the mean of all values across all terms is used, for the given interval and the given document draft, in order to account for the distribution of values in each interval of each document. Then, the binarisation happens in the second loop, in lines 12-21, using the `binarise` function, which takes as arguments the median value and the list of all values for Sentiment and for Participation. The `binarise` function maps a value to 0 (low) if it is lower than or equal to the threshold (median), and to 1 (high) otherwise. The final, binary, values for S an P are stored in the `sentimentBinary` and `participationBinary` variables respectively.

Appendix C

Contents of the term corpus of two Recommendation documents

Further to Chapter 6.4.1, this appendix presents the contents of the term corpus for two of the Recommendation documents produced by the W3C Provenance Working Group (the DM and the Constraints documents). The contents of the term corpora of the other documents are similar.

As discussed, the term corpus for a given draft of a given specification document is made up of the words that were most frequently used in all drafts up to and including that draft of that document *and* that also appeared in at least one email subject line out of all emails sent up to the publication of the current draft. Frequent words are extracted after first having removed non-informative sections (e.g. acknowledgements, bibliography, table of contents) and numbers, lemmatized the remaining words, and removed all stop words (English function words with no particular inherent meaning) from the document drafts, as described in Appendix B.1).

As discussed, the goal of the term corpus was to contain words that represent terms that entered the formal vocabulary defined by the Group, and words that reflect the language and exposition that was frequently used to define, explain, and discuss this formal vocabulary. Therefore, this appendix demonstrates how this is captured in the contents of the term corpus, by classifying this term in the corpus as either reflecting an element of the formal vocabulary defined by the Group, or an element of the language and exposition style that was frequently used to define, explain, and discuss this formal vocabulary.

C.1 Background: Provenance and the W3C PROV standard

Before beginning to look into the contents of the term corpus, in order to understand what the terms in the corpus are about, it is important to briefly establish the basic goals of the W3C Provenance Working Group, as well as what 'Provenance' is, and what the specification documents (called 'the PROV family of documents') produced by the Working Group are for. As stated in the abstract of the Overview document produced by this Working Group, 'Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. The PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the Web.' [1]

The PROV model defined in the PROV family of documents exposes which entities, activities and agents were associated with the creation and evolution of a piece of data or thing, i.e. it exposes which elements were involved with how a piece of data or thing came to reach its present state. The core concepts of the PROV model are illustrated in Figure C.1. This Figure illustrates how PROV records (records of Provenance information represented per the PROV standard) can be represented visually as directed graphs, where the nodes represent the agents, activities and entities, and the edges represent the relations between them (e.g. `used, wasDerivedFrom, wasGeneratedBy, wasAssociatedWith, wasAttributedTo`). Each relation can only be applied to specific types of elements (e.g. association has an activity as its subject and an agent as its object, that is an agent `wasAssociatedWith` an activity), as shown in Figure C.1.

In addition to the core PROV terms shown in Figure C.1, another useful collection of important terms and keywords defined in the PROV Recommendations is shown in Table C.1. This is useful as the term corpora we shall see in this appendix contain words from all of these columns. The columns here shows how words representing PROV-DM concepts, in the first column, correspond to words representing PROV-O 'classes and properties', and PROV-N 'productions' (terminology per the original caption in the PROV-DM document), in the other columns, and how these are organised by this Working Groups under thematic 'components' in the final column.

[1] From https://www.w3.org/TR/prov-overview/, accessed 15 April 2018.

[2] Diagram from http://www.w3.org/TR/prov-o/diagrams/starting-points.svg, accessed 15 April 2018.

[3] Table from https://www.w3.org/TR/prov-dm/#prov-dm-to-prov-o-and-prov-n, accessed 15 April 2018.

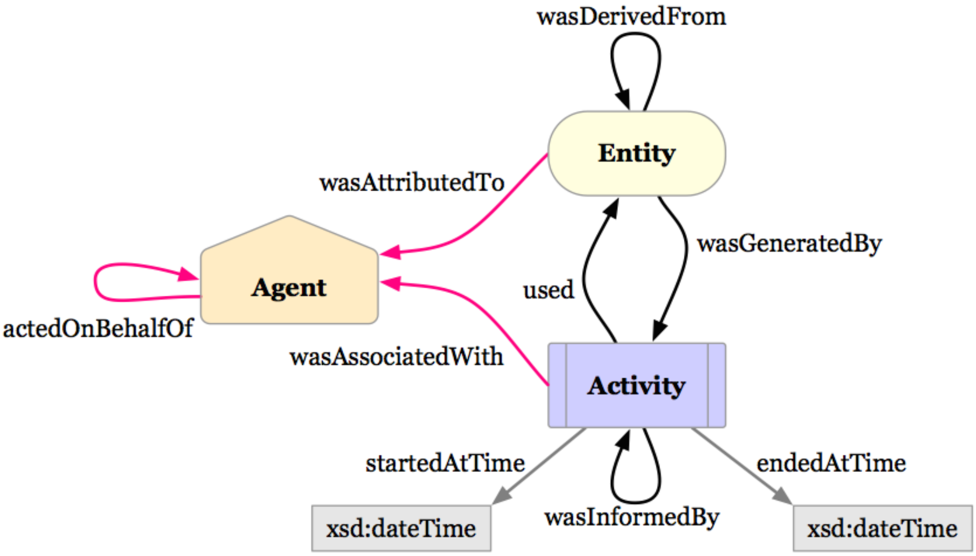


FIGURE C.1: The core PROV elements and relations[2]

Having introduced the goals and terminology of the PROV standard produced by the W3C Provenance Working Group, we next turn to examining the words contained in the term corpus of the PROV-DM and PROV-Constraints Recommendation documents.

Table C.1: Cross-References from PROV-DM to PROV-O and PROV-N[3]

PROV-DM	PROV-O	PROV-N	Component
Entity	Entity	entityExpression	Component 1: Entities/ Activities
Activity	Activity	activityExpression	
Generation	wasGeneratedBy, Generation	generationExpression	
Usage	used, Usage	usageExpression	
Communication	wasInformedBy, Communication	communicationExpression	
Start	wasStartedBy, Start	startExpression	
End	wasEndedBy, End	endExpression	
Invalidation	wasInvalidatedBy, Invalidation	invalidationExpression	
Derivation	wasDerivedFrom, Derivation	derivationExpression	Component 2: Derivations
Revision	wasRevisionOf, Revision	type Revision	
Quotation	wasQuotedFrom, Quotation	type Quotation	
Primary Source	hadPrimarySource, PrimarySource	type PrimarySource	
Agent	Agent	agentExpression	Component 3 Agents, Responsibility Influence
Attribution	wasAttributedTo, Attribution	attributionExpression	
Association	wasAssociatedWith, Association	associationExpression	
Delegation	actedOnBehalfOf, Delegation	delegationExpression	
Plan	Plan	type Plan	
Person	Person	type Person	
Organization	Organization	type Organization	
SoftwareAgent	SoftwareAgent	type SoftwareAgent	
Influence	wasInfluencedBy, Influence	influenceExpression	
Bundle constructor	bundle description	bundle	Component 4: Bundles
Bundle type	Bundle	type Bundle	
Alternate	alternateOf	alternateExpression	Component 5: Alternate
Specialization	specializationOf	specializationExpression	
Collection	Collection	type Collection	Component 6: Collections
EmptyCollection	EmptyCollection	type EmptyCollection	
Membership	hadMember	membershipExpression	

C.2 Classifying words in the term corpus of PROV-DM and PROV-Constraints

This section examines the words in the term corpus of the PROV-DM and of the PROV-Constraints, and classifies each as either reflecting an element of the formal vocabulary defined by the Group, or an element of the language and exposition style that was frequently used to define, explain, and discuss this formal vocabulary.

The term corpus used is the one made up of the top frequent terms of all drafts up to the current draft, and the final draft is taken to be the current draft, so that the corpus contains terms from all drafts, including the Charter document.

The words in the term corpus are colour coded, according to the scheme in Table C.2.

Table C.2: Colour scheme for words in the term corpus

Colour	Meaning
🟦 Blue	Part of the formal vocabulary of terms defined by the Working Group
🟨 Yellow	Part of the language and exposition style

In the colour-based classification scheme shown in Table C.2, words coded in blue are part of the formal vocabulary of terms defined by this Working Group, while words coded in yellow are instead part of the language and exposition style used in the documents produced by this Working Groups. For some terms in the corpus, there is no clear-cut classification in whether they are part of the formal vocabulary or of the language (e.g. terms such as ‘core’, describing a class of formal concepts, the core concepts; this is a meta descriptor of the actual formal concepts, but might also be considered part of the formal terminology) - those terms will be coded in both blue and yellow.

The yellow category, for language and exposition, is assigned to words that relate to any of the following categories:

- Definitions and explanations of formal terms of the vocabulary (including, but not limited to, words shown in Figure C.1 and Table C.1;
- Examples used in the explanations (e.g. ‘bob’, ‘tool’, ‘news’, ‘report’, ‘work’, ‘publication’);
- The group and its organisation (e.g. ‘group’, ‘incubator’, ‘document’, ‘specification’, ‘standard’);
- References to structural elements of a document in the text (e.g. ‘figure’, ‘table’, ‘section’, ‘image’).

Let us first consider the term corpus of the PROV-DM document, and next the corpus of the PROV-Constraints document.

C.2.1 Term corpus for PROV-DM

In this section, the words in the term corpus of the PROV-DM document are classified, per the scheme in Table C.2. This is done in Table C.3.

Table C.3: Classification of words in the PROV-DM term corpus

Term	Code	Term	Code	Term	Code
access	🟨	generation	🟦	recipe	🟨
account	🟨	give	🟨	recommendation	🟨
activity	🟦	grammar	🟦	record	🟦
additional	🟨	group	🟨	refer	🟨
agent	🟦	help	🟨	relate	🟨
alternate	🟦	http	🟨	relation	🟦
annotation	🟦	identifier	🟦	relevant	🟨

application	🟨	identify	🟨	report	🟨
asn	🟦	image	🟨	represent	🟨
asserter	🟦	incubator	🟨	representation	🟦
assertion	🟦	inference	🟨	resource	🟦
associate	🟦	influence	🟦	responsibility	🟦
association	🟦	information	🟨	result	🟨
attribute	🟦	instantaneous	🟦	revision	🟦
attribution	🟦	interval	🟦	role	🟦
bob	🟨	invalidation	🟦	scope	🟦
bundle	🟦	involve	🟨	second	🟦 🟨
characterization	🟦	issue	🟨	section	🟨
characterize	🟦	key	🟦	semantic	🟦
collection	🟦	language	🟨	semantics	🟦
communication	🟦	link	🟦	sense	🟨
community	🟨	location	🟦	set	🟨
component	🟨	match	🟦 🟨	source	🟦
concept	🟦	member	🟦	specialization	🟦
conceptual	🟦	model	🟦 🟨	specific	🟨
constraint	🟦	multiple	🟨	specification	🟦 🟨
control	🟦	namespace	🟦	standard	🟨
core	🟦 🟨	new	🟨	start	🟦
data	🟦 🟨	news	🟨	state	🟦
define	🟨	normative	🟦 🟨	step	🟨
delegation	🟦	notation	🟦	structure	🟨
derivation	🟦	note	🟨	table	🟨
describe	🟨	ontology	🟦	thing	🟦 🟨
description	🟨	optional	🟦	time	🟦
dictionary	🟦	order	🟦	tool	🟨
different	🟨	pair	🟦	type	🟦
do	🟨	plan	🟦	usage	🟦
document	🟦 🟨	prefix	🟦	use	🟦 🟨
draft	🟨	primary	🟦	value	🟦
embed	🟦 🟨	process	🟦	version	🟦
end	🟦	processexecution	🟦	vocabulary	🟦
entity	🟦	production	🟦 🟨	wasassociatedwith	🟦
event	🟦	propose	🟨	wasattributedto	🟦
example	🟨	prov	🟦	wasderivedfrom	🟦
execution	🟦	provenance	🟦	wasgeneratedby	🟦
exist	🟨	provenir	🟦	web	🟦
express	🟨	provide	🟨	work	🟨
extend	🟦 🟨	publication	🟨	world	🟦 🟨

figure	🟨	qualifier	🟦	write	🟨
file	🟨	qualify	🟦	xml	🟦
follow	🟨	query	🟦	xsd	🟦
formal	🟨	rdf	🟦		
generate	🟦	rec	🟨		

In Table C.3, one might further group some entries in the following groups:

Explanatory words. Words such as the below were used frequently in explanations:

production. This word is has a special meaning in this context (hence the blue label), but is not part of the formal vocabulary, and is used often to define or explain notions (hence the yellow label) such as that of 'generation';

thing. This word is used frequently, e.g. in the definitions of Provenance and of the formal 'entity' concept.

Meta-descriptors The following words were not strictly part of the the formal vocabulary (hence the yellow label), but they did have a special meaning in this context and were used to describe types of terms from the formal vocabulary (hence the blue label):

core. Refers to the core terms, or structures, of the PROV standard, shown in Figure C.1;

extend. Refers to the non-core PROV structures, called 'extended structures';

concept. The formal vocabulary defined by the group was made up of terms, which themselves were made of concepts (which this word refers to) and relations;

conceptual. Refers to the conceptual data model defined by the group;

constraints. Constraints have a special meaning in the context of the PROV standard, and indeed are the topic of a whole Recommendation document, PROV-Constraints;

normative. Describes sections that contain formal rules and elements of the standard, as opposed to sections containing more informal information or advice);

world. This word was used often in some of the drafts, as a way to discuss the context that Provenance information my describe, and to discuss how to model the factual world versus different perspectives of the world.

Parts of phrases The following words tended to not occur frequently on their own as much as when they were part of short phrases:

community. This term appears in the phrases 'Web community' or 'community of technical experts' in the Charter;

data. The term ‘data’ is often, but not always (hence the yellow code), part of the phrase ‘data model’, referring to the topic of this document, which was also one of the key deliverables of the Group, hence the blue code;

extend. This term is encountered as part of the phrase ‘extended structures’, referring to the non-core PROV structures;

primary. This word is part of the formal vocabulary as part of the term ‘primary source’;

work. This word is used in the phrase ‘working draft(s)’.

Formal terms only in some non-final drafts. The following words were part of the formal vocabulary at some point(s) but not in the final draft: ‘asn’ (refers to PROV-ASN, the ‘Abstract Syntax Notation’ for PROV), ‘control’ (a type of expression), ‘event’, ‘execution’ (as part of the phrase phrase ‘process execution’, replaced later by the term ‘activity’), ‘grammar’, ‘interval’, ‘link’ (used as part of the term ‘process execution linked derivation expression’ and of the term ‘recipe link’), ‘process’ and ‘processexecution’ (later replaced by the term ‘activity’ and related terms), ‘representation’, ‘resource’, ‘world’.

Examples. The following words appeared frequently in examples employed in the document drafts:

application. This word is not quite used in examples, but rather in frequent discussions in the drafts’ main text on how the PROV standard may be used in practice in various applications, and how it may be instantiated using application-specific information.

bob. Frequently used in several examples in which the person involved is called Bob;

file. Part of a frequently used example which involves recording the Provenance of files;

http. This term comes from the resource identifiers (IRIs) used in various examples in the document drafts;

news. The word ‘news’ is used frequently in examples about recording the Provenance information of news items;

publication. This word is used frequently in examples about recording the Provenance information of the W3C document publication process;

report. This word is used frequently in examples about recording the Provenance information of a report;

tool. This word is used frequently in one non-final draft, in an example about a ‘performance rating tool’;

work. This word is used frequently in one draft (not the final one), in the phrase 'working draft(s)', in an example about recording the Provenance of W3C documents.

References to the group and its organisation. The following terms refer to the Group itself or its organisational procedures: 'document', 'draft', 'group', 'incubator', 'issue', 'rec', 'recommendation', 'specification', 'standard'.

References to document elements. References to document elements in the document's narrative: 'example', 'figure', 'section'.

Overall, Table C.3 shows that most of the terms in the corpus related to the formal vocabulary defined by the group (words with blue label, or blue and yellow label, made up 62% of all words), and many related to the language and narrative used to define, explain and discuss that formal vocabulary (words with yellow label made up the remaining 38%).

Per the discussion above, a few of the yellow labelled words may have not been directly related to the formal vocabulary or its explanation and discussion, but rather to the structure of the documents ('example', 'figure', 'section'), to the group and its processes ('document', 'draft', 'group', 'incubator', 'issue', 'rec', 'recommendation', 'specification', 'standard'), or to words from examples ('bob', 'http'). However, one might argue that some of these words do convey some useful information: the frequent use of the words 'example' and 'figure' suggests that the Group used examples and figures heavily in order to explain the formal vocabulary and how it can be used; the word 'issue' which was frequent in only one draft shows that at that point the Group thought it useful to highlight parts of the proposed vocabulary corresponding unresolved and under review issues in text, but in previous and next drafts it stopped highlighting this; the word 'http' indicates that resource identifiers (IRIs) were used very frequently in examples, i.e. that they are an important part of recording Provenance information in practice using PROV. Therefore, resolving which of these words should be discarded from the corpus would likely benefit from drawing upon context knowledge by surveying Group members. Either way, the words listed here still only represent a small percentage of the corpus's contents (8.9%).

C.2.2 Term corpus for PROV-Constraints

Similarly to above, the words in the term corpus of the PROV-Constraints document are classified, per the scheme in Table C.2. This is done in Table C.4.

Table C.4: Classification of words in the PROV-Constraints term corpus

Term	Code	Term	Code	Term	Code
access	🟨	generation	🟦	relate	🟨
account	🟨	give	🟨	relation	🟦
activity	🟦	group	🟨	relevant	🟨
agent	🟦	help	🟨	report	🟨
alternate	🟦	identifier	🟦	resource	🟦
alternateof	🟦	identify	🟨	result	🟨
apply	🟨	image	🟨	role	🟦
associate	🟦	incubator	🟨	scope	🟨
attribute	🟦	inference	🟨	section	🟨
check	🟦	information	🟨	semantic	🟦
collection	🟦	instance	🟦	semantics	🟦
community	🟨	instantaneous	🟦	sense	🟨
concept	🟦	invalidation	🟦	set	🟦 🟨
conceptual	🟦	involve	🟨	specialization	🟦
constraint	🟦	key	🟦	specification	🟨
core	🟦 🟨	language	🟦 🟨	standard	🟨
data	🟦 🟨	link	🟦 🟨	start	🟦
define	🟨	location	🟨	state	🟦
definition	🟨	mean	🟨	statement	🟦
derivation	🟦	merge	🟦 🟨	subfigure	🟨
description	🟨	model	🟨	term	🟦
different	🟨	multiple	🟨	thing	🟨
do	🟨	normalization	🟦	time	🟦
document	🟦	note	🟨	type	🟦
draft	🟨	ontology	🟦	uniqueness	🟦
embed	🟦 🟨	optional	🟦	usage	🟦
end	🟦	order	🟦	use	🟦 🟨
entity	🟦	parameter	🟦 🟨	valid	🟦
equivalence	🟦	process	🟦	value	🟦
equivalent	🟦	propose	🟨	vocabulary	🟦
event	🟦	prov	🟦	wasassociatedwith	🟦
example	🟨	provenance	🟦	wasderivedfrom	🟦
execution	🟦	provenir	🟦	wasgeneratedby	🟦
exist	🟨	provide	🟨	wasinformedby	🟦
express	🟨	publication	🟨	wasstartedby	🟦
figure	🟨	query	🟦	web	🟦
follow	🟨	rec	🟨	work	🟨
form	🟦	recipe	🟦	world	🟦
formal	🟨	recommendation	🟨	xml	🟦

generate	🟦	refer	🟨	xsd	🟦

In Table C.3, one might further group some entries in the following groups:

Explanatory words. Words such as the below were used frequently in explanations:

apply. This word is used frequently, e.g. to discuss applying formulae to terms, or which relationships apply to which concepts (i.e. as shown in Figure C.1, each relationship, shown as an arrow, is applicable to specific kinds of concepts only, shown at the ends of each arrow), or how to definitions, inferences, and constraints to a PROV instance.

check. Even though not part of the formal vocabulary of this document, this word is core to the definition of the goal of this PROV-Constraints document, and appears in frequent discussions about validity checking, equivalence checking, and constraint checking. As stated in the Abstract of this documents' final draft, the PROV-Constraints document 'defines a subset of PROV instances called valid PROV instances', which 'satisfy certain definitions, inferences, and constraints. These definitions, inferences, and constraints provide a measure of consistency *checking* for provenance and reasoning over provenance.'

define, definition. As this document formally defines various notions and formulae, the words 'define' and 'definition' are frequent.

thing. This word is used frequently, e.g. in the definitions of Provenance and of the formal 'entity' concept.

Meta-descriptors The following words were not strictly part of the the formal vocabulary (hence the yellow label), but they did have a special meaning in this context and were used to describe types of terms from the formal vocabulary (hence the blue label):

core. Refers to the core terms, or structures, of the PROV standard, shown in Figure C.1;

concept. The formal vocabulary defined by the group was made up of terms, which themselves were made of concepts (which this word refers to) and relations;

conceptual. Refers to the conceptual data model defined by the group;

world. This word was used often in some of the drafts, as a way to discuss the context that Provenance information my describe, and to discuss how to model the factual world versus different perspectives of the world.

Parts of phrases. The following words tended to not occur frequently on their own as much as when they were part of short phrases:

community. This term appears in the phrases 'Web community' or 'community of technical experts' in the Charter;

data. The term 'data' is sometimes, but not always (hence the yellow code), part of the phrase 'data model', referring to the topic of this document, which was also one of the key deliverables of the Group, hence the blue code;

Formal terms only in some non-final drafts The following words were part of the formal vocabulary at some point(s) but not in the final draft: 'execution' (as part of the phrase phrase 'process execution', replaced later by the term 'activity'), 'world'.

Examples The following words appeared frequently in examples employed in the document drafts:

application. This word is not quite used in examples, but rather in frequent discussions in the drafts' main text on how the PROV standard may be used in practice in various applications, and how it may be instantiated using application-specific information.

report. This word is used frequently in examples about recording the Provenance information of a report

References to the group and its organisation The following terms refer to the Group itself or its organisational procedures: 'document', 'draft', 'group', 'incubator', 'rec', 'recommendation', 'specification', 'standard'.

References to document elements References to document elements in the document's narrative: 'example', 'figure', 'section'.

It is noted that there are several differences in the term corpus of PROV-Constraints (Table C.4)and of PROV-DM (Table C.3). There are words which were in one corpus, and indeed part of the formal vocabulary of that document (blue label), but are not part of the other document's corpus. For instance, the term 'primary' was in the formal vocabulary of DM (as part of 'primary source'), but not present in the Constraints corpus, while the terms 'event', 'equivalent' and 'equivalence' are part of the formal vocabulary of Constraints but absent from the corpus of DM. Such cases highlight the differences in each document's area of focus.

Overall, Table C.4 shows that most of the terms in the corpus related to the formal vocabulary defined by the group (words labelled blue, or blue and yellow, made up 60% of all words), while many related to the language and narrative used to define, explain and discuss that formal vocabulary (words with yellow label made up the remaining 40%).

As discussed above, and similarly to the discussion for PROV-DM, a few of the yellow labelled words may have not been directly related to the formal vocabulary or its explanation and discussion, but instead to the structure of the documents (‘example’, ‘figure’, ‘section’), to the group and its processes (‘document’, ‘draft’, ‘group’, ‘incubator’, ‘rec’, ‘recommendation’, ‘specification’, ‘standard’). However, as discussed for PROV-DM one might argue that some of these words do convey some useful information: the frequent use of the words ‘example’ and ‘figure’ suggests that the Group used examples and figures heavily in order to explain the formal vocabulary and how it can be used. Hence, resolving which of these words should be discarded from the corpus would likely benefit from drawing upon context knowledge by surveying Group members. In any case, the words listed here still only represent a small percentage (9.2%) of the contents of the corpus.

Appendix D

Additional corpora for the empirical colllective-level causal analysis

This appendix presents details on how two additional term corpora were constructed, in addition to the term corpus used for the main collective-level causal analysis of the influence of online communications in the W3C Provenance Working Group. It discusses the rationale behind the corpora actually used (Section D.1), as well as behind other corpora considered but deemed unsuitable (Sections D.1.1 and D.1.2).

As discussed in Chapter 6.4.1.1, in addition to how the term corpus is constructed in the main analysis (to contain the top frequent terms from the drafts up to the draft at the end of the current time interval), this thesis also presents, in Section 7.3, two further analyses, each using a concept corpus constructed differently. Subsection E.1 presents an implementation design where the corpus used at each time interval contains the top frequent terms from *all* drafts (including future drafts) of a given document (where each of these terms is in at least one email subject line, out of *all* emails), rather than containing only the top frequent terms from the drafts only up to the draft at the end of the current time interval (where each of these terms is in at least one email subject line, out of the emails sent up to the publication time of the draft at the end of the current time interval). Then, subsection E.2 presents the findings from another implementation design, where the corpus at each interval is again real time, as in the original corpus, i.e. it only contains top terms from the drafts up to the current interval (where each of these terms is in at least one email subject line, out of the emails sent up to the publication time of the draft at the end of the current time interval), but now it contains more of the top frequent terms, hence capturing more of the contents of each draft.

Section 7.3 presents results when the term corpus is constructed differently to how it was constructed for the main analysis presented in Section 7.1. Subsection E.1 presents an implementation design where the corpus used at each time interval contains the top frequent terms from *all* drafts (including future drafts) of a given document, rather than containing only the top frequent terms from the drafts only up to the draft at the end of the current time interval. So, in this case, the same corpus is used across all intervals. This is essentially a retrospective corpus, as it requires the document to have reached its final draft, for data from all drafts to be extracted. This is in contrast to the original corpus, which can be constructed, and the analysis performed, in real time, as the Working Group's work is unfolding. Then, subsection E.2 presents the findings from another implementation design, where the corpus at each interval is real time, i.e. it only contains top terms from the drafts up to the current interval (as in the original corpus), but now it contains more of the top frequent terms, hence capturing more of the contents of each draft. The same `CountVectorizer` command as described for the standard corpus above is used for the construction of this lager corpus, with the only difference being that `max_features=200` (rather than `100`) is used for this larger corpus.

D.1 The rationale of the corpus design

The rationale for the real-time corpora is that an investigator is interested, even while the Working Group's project is unfolding, in tracking the terms that have been prominent at any point this far, and how the emails (S and P) and prominence in the previous draft affect future prominence (F) on average across all terms.

The rationale for the retrospective corpus is that an investigator, after the Working Group has ended, is interested in also taking into account, for every draft, retrospective information from future drafts (which terms were prominent in future drafts), i.e. to analyse the evolution over time of *all* terms that were ever prominent.

D.1.1 Why not consider only the terms of the final draft

A corpus made up only of the terms that were prominent in the *final* draft of each document is not used, as this is the classic case of so-called *selection bias*: studying only cases where the final outcome was 'successful'. Rather, it is of interest here to study both which terms 'succeeded' in becoming part of the formal vocabulary and the narrative (i.e. which terms were frequent, or prominent, in the documents) and which were not, and how this changed over time. So, to limit such selection bias, and in the absence of an external standard of which terms are important to study, the study focuses on analysing the terms that were prominent in any draft, and their evolution as they may have been discarded from, introduced into, or kept along in the top prominent (frequent) terms, over time. In addition, restricting the corpus in such a way would lead to an artificial restriction on the possible values of the selection and causal effects, particularly in the case of the last inter-draft interval for each document, as described in the sub-section below. Again, as there is no external standard of which terms were important at each stage of the evolution of each document and should be analysed over time, other than the contents of the document drafts themselves, prominence in the document drafts themselves is used as an indicator that a term was at some point considered important by the Working Group, and therefore should be studied.

D.1.2 Why not consider only the terms in the drafts at either end of the given interval

It is also noted that the corpus is never limited to terms that appear in the top frequent terms in *this interval only*, i.e in the initial or the final draft of this interval only. Rather, top terms from previous drafts (previous intervals) are always included, for all kinds of corpora considered in this thesis. That is because, if the corpus is limited to top terms from the initial or final draft of this interval only, the selection bias is such that it constrains the possible range of causal and selection effect values. (This phenomenon also applies to the above discussed corpus containing only top terms from the final draft, in the last inter-draft interval of each document.) That is, for this kind of corpus, we would have:

- In cases where $I = 0$: F would have to always be 1, because there are no terms with both F and I equal 0 included in the corpus. That is because, for a term to be included in the corpus for the given slice, it must be that the term was in the top terms of the initial or the final draft of this interval (or in both drafts), so, for this term, either $I = 1$, or $F = 1$, or both I and F equal 1. This means that $P(F = 1|I = 0) = 1$ – this is the first term of the selection effect formula.
- In cases where $I = 1$, F could be either 1 or 0 – no restriction due to corpus construction here. So there are no restrictions on the possible values of $P(F = 1|I = 1)$.

So, for the selection effect, $P(F = 1|I = 1) - P(F = 1|I = 0)$, the above two points mean that the second term is always 1, while the first term can have any value in $[0, 1]$, so the selection effect is always less than or equal to 0.[1]

Similarly for the causal effect $\sum_z P(F = 1|I = 1, Z = z)P(Z = z) - \sum_z P(F = 1|I = 0, Z = z)P(Z = z)$, where the second term is 1, and the first term can have any value in $[0, 1]$, and so the causal effect is also negative or 0.

[1] This is also reflected in the alternative form of the selection effect, $P(F = 0|I = 0) - P(F = 0|I = 1)$, where the first term equals zero, and the second term can take any value in $[0, 1]$, so the whole expression is non-positive. And the same hold for the alternative form of the causal effect.

Therefore, artificially restricting the dataset such that it does not contain any terms that have $F = 0$ and $I = 0$, i.e. that are not prominent in either of the drafts at the ends of this interval, leads to skewed effects, and to skewed interpretations. Because excluding cases with $F = 0$ and $I = 0$ makes it seem like whenever $I = 1$ then F may equal 0 or 1, while whenever $I = 0$, F is *guaranteed* to be 1 and it is impossible to have $F = 0$ (as there are no entries with both I and F equal to 0). So I has a negative effect on F (switching I from 0 to 1 means we move from F possibly being 0 to a guarantee that F will certainly be 0; causally this implies that if one wants to have $F = 1$ they should ensure that $I = 0$), hence obtaining negative causal and selection effects for the effects of I on F.

However, there are cases of terms that do not appear in I and also do not appear in F of a given interval, even though they may appear prominently in previous or future drafts, and since excluding them distorts the picture, they should be included in the analysis. That is, since there is evidence that when a term has $I = 0$ in a given interval it may also have $F = 0$, then this evidence should be included in the analysis. In general, if all possible combinations of values are possible in the dataset, then they should appear in the data analysed - since making design choices for the corpus construction that end up excluding certain value combination distorts the effect estimates obtained.

Overall, it is deemed here, in the context of this Working Group, when analysing each inter-draft interval, that a term that may not be prominent at the initial or final draft of this slice is still worth including in the corpus, as long as it did appear in at least one other draft (before or after this interval). That is because it is considered that it is worth including in the analysis terms that may be temporarily absent (e.g. terms that were dropped from the top prominent, perhaps temporarily, perhaps for good, or terms that have not yet been introduced into the top-prominent), at every interval (even in an interval from whose draft that term is absent). Even though each interval (i.e. pair of drafts and the emails in the interval between them) is analysed individually, it is not the aim to limit the scope of the analysis to each interval individually: rather, for each interval, we are interested in including in the analysis terms that are prominent in this interval, as well as terms that are not (but that are prominent in some other drafts, at earlier or later times). This is why restricting the term corpus at each interval to only contain terms that have $F = 1$ or $I = 1$ for this interval is not considered appropriate, and the resulting effects are considered skewed.

D.2 Summary

Overall, constructing a corpus out of terms that were ever prominent, in any draft this far (real-time default corpus, used in the analysis of Section 7.3, and the larger corpus of Section E.2), or in any draft at all, whether this far or in the future (retrospective corpus, in Section E.1) is reasonable, and it does not artificially constrain the value combinations that appear in the analysed data. The only edge case is in the first interval the real-time corpus (both the original and the larger one), where, for the effect of $\boldsymbol{I}$ on $\boldsymbol{F}$, the effects are negative, as there are no previous drafts so there are no terms for which both $\boldsymbol{I}$ and $\boldsymbol{F}$ are zero (i.e. just like the pathology discussed above for the corpus that ignores previous drafts). (Causal effects are not applicable for that interval, as there are no previous email conversations for whose $\boldsymbol{S}$ and $\boldsymbol{P}$ to adjust.)

In future work, it would be possible to also include in the corpus terms that were not prominent in any of the drafts, but it is not clear based on what criteria those terms would be selected – as there is no external agreed-upon standard or benchmark on which terms are important and hence should have their prominence or absence over time analysed (should be in the term corpus), other than the contents of the documents themselves, we have used the data on whether a term was ever prominent in any document draft as an indicator of whether a term was ever considered important. Perhaps one way to do this would be to look at the terms that were prominent in the email discussions, and put those in the corpus, where it may be that several of those were never prominent in any of the document drafts. Another approach would be to interview some of the participants about terms that they considered important that did not make it to the top terms of any of the document drafts.

Appendix E

Results under different corpus designs

Further to the analysis and findings presented in Section 7.1, this appendix presents the results of the causal analysis of the social influence of online communications in the W3C Provenance Working Group archives under two different design choices for the construction of the term corpus than what is used in the main analysis (with that design being described in Chapter

The goal is to investigate how much the findings would vary under different design choices (in terms of the construction of the term corpus). As discussed in Chapter 7.3, it is found that the findings of the main analysis are robust to the variations in implementation presented here: the amount of confounding bias present when estimating the effects of email conversations on outcomes is not negligible, but is rather quite large, while the characteristics and patterns (over time and across contexts) of the effects of the emails on outcomes, and of the effects of previous outcomes on next outcomes, are overall the same

As described in Chapter 6.4, one of the first steps of the analysis is to extract, for every draft of every document, the top (n) frequent terms (or words) from the m drafts up to the current draft. These words make up the *term corpus*. Then for each of the words in the term corpus, the following steps are taken:

1. in the document drafts domain: for each specific draft, recording which of these words appear in it, and which do not, to obtain values for which terms are most frequent in each draft (variables I and F, for each inter-draft interval.)
2. in the email domain: searching for the occurrence of each word in the thread subject lines, to obtain relevant email conversations for each word, and from all those email conversations aggregating the overall sentiment and participation levels (variables S and P) for each inter-draft time interval.

In this appendix, it is this the construction of this term corpus that is different: Subsection E.1 presents an implementation design where the corpus used at each time interval contains the top frequent terms from all drafts of a given document, rather than containing only the top frequent terms from the drafts up to the draft at the end of the current time interval. So, in this case, the same corpus is used across all intervals. This is essentially a retrospective implementation, as it requires the document to have reached its final draft, for data from all drafts to be extracted. Then, Subsection E.2 presents the findings from another implementation design, where the corpus only contains top terms from the two drafts at either end of the given time interval, but now it contains more of the top frequent terms, hence capturing more of the contents of each draft.

The goal in each implementation variant is to study whether it yields any different conclusions, compared to those of Chapter 6.4, in terms of the extent of the confounding bias present, and in terms of the properties of the effects of each variable.

In all sections here, the plots in the figures and statistics in the tables follow the same logic and format as those in Chapter 7.1. Additionally, for each causal variable there are extra tables comparing side-by-side the statistics of effects when using the retrospective corpus versus those obtained when using the real-time corpus in Chapter 7.1. All percentage figures reported in the tables are rounded to the nearest integer.

E.1 Retrospective corpus analysis

In all findings of Chapter 7.1, for analysing the causal (and associational) factors affecting each draft, we only used knowledge up to that point in time, i.e. we only looked at the contents of the drafts up to that point in time. Hence, this kind of analysis can be used ‘on-line’ or in ‘real-time’, during the progress of the collective setting being studied.

In this subsection, a retrospective analysis term corpus will be used; that is, a corpus made up of top words from *all* drafts of a document, not only up to current draft. This is done to investigate whether knowing which terms end up being most prominent in the final drafts helps make better, or different, statistical and causal predictions of the contents of all drafts, over time and across documents. That is, this analysis requires for the collective effort to have ended, hence cannot be conducted while the effort is ongoing - rather, it can only be conducted retrospectively. Therefore, by construction, at the last inter-draft interval of each documents, both corpora will contain the same terms, whereas the corpora will likely be least alike at the beginning (interval labelled 0 in the plots) as the new retrospective corpus will contain several terms that do not actually appear in the drafts until later.

This subsection begins with the analyses of the impact of the characteristics of email conversations (first Participation, then Sentiment) on the contents of each draft, and proceeds with the analysis of the causal and selection effects of the previous outcome (draft) on the current one.

E.1.1 Social influence of email communications

Let us first consider the effects of participation (P) and of sentiment (S) in email communications on the next draft (F).

Participation Figure E.1 shows the causal and selection effects of participation (P) in email communications on the next draft (F), across all documents. It is apparent that effects are not always positive, are generally quite small, and usually causal effects get smaller in later stages of the documents' lifetimes. Moreover, it is apparent that causal effects are often quite different from selection effects, hence there is confounding bias. This bias is quantified in more detail in Table E.1, which contains descriptives for the amount of bias (ARD measure) for each document and aggregated across all documents.

Table E.1: Participation: Descriptives of confounding bias (ARD) across documents, using the retrospective corpus

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	30%	156%	88%	83%	22%	46%
Constraints	45%	409%	198%	183%	181%	134%
DC	4%	439%	159%	33%	393%	198%
Dictionary	59%	97%	78%	78%	4%	19%
DM	24%	641%	251%	208%	333%	183%
Links	3%	650%	258%	119%	795%	282%
N	50%	1942%	590%	183%	6144%	784%
Ontology	17%	11460%	2350%	81%	207483%	4555%
Overview	3%	136%	59%	37%	32%	57%
Primer	35%	1222%	310%	50%	1892%	435%
Sem	42%	67%	54%	54%	2%	13%
XML	22%	134%	78%	79%	21%	46%
All documents	3%	11460%	447%	97%	27496%	1658%
Descriptives of descriptives						
Mean	28%	1446%	373%	99%	18108%	563%
Median	27%	424%	178%	80%	257%	159%

The patterns in Figure E.1 and in Table E.1 here are overall the same as those observed as in the Figure 7.1) and Table 7.1 of Section 7.1, i.e. as follows:

Confounding bias Table E.1 shows that confounding bias is not negligible, but rather quite large: the median of medians is at 80%, with the median across all documents at 97%. For some documents the median is higher than 100%, e.g. for DM it is at 208%, while for N and Constraints it is at 184% (these are three of the core

Recommendation documents). Meanwhile the mean of means is much higher, at 373%, and the mean across all documents taken together is at 447%. These figures are similar to those in Section 7.1.

Effect sign Not all effects are positive, rather there are several cases of negative effects, both for the causal and the selection effects. Therefore, high participation volume for a given term in the emails does not guarantee a high chance of that term being prominent in the next draft, neither causally nor in terms of association, as has been discussed in Section 7.1

Effect magnitude Effects here are relatively small, and follow similar patterns as in Section 7.1, with causal effects here rarely being higher than 30%.

Evolution over time Similarly to Section 7.1, over time, the causal effect of Participation often gets smaller (e.g. Constraints, DC, DM, Links, N, Ontology), which however does not hold for the selection effect. Hence, if one were to only calculate the (biased) selection effect, they would reach quite different conclusions about the temporal evolution of the effects of participation than if they had calculated the causal effect; that is, confounding bias leads to different temporal patterns. Moreover, within each document, there is large variability over time, as per the variance and standard deviation figures in Table E.1, where the median variance is very high at 257% with the median standard deviation at 159%.

Patterns across documents Table E.1 shows how bias varies across documents. For instance, median bias ranges from only 33% for DC, to 208% for DM, numbers that differ widely, by a factor of 6.3, but none of which is negligible. In addition, temporal patterns of the effects themselves vary across documents, in some cases exhibiting opposite tendencies: for instance, the effects in Dictionary increase (although there r only 2 data points), while causal effects decrease in Links, in Figure E.1, similarly to what happens in Figure 7.1.

Tables E.2 and E.3 allow for side-by-side comparisons of bias in descriptives when using the original (real-time) corpus versus when using the retrospective corpus.

Table E.2 shows that the overall median biases across documents are very close, within 5-7 percentage points: the median of median are within 7 percentage points (87% and 80%) while the medians when all documents' datapoints are used together are 5 points apart, at 102% and 97%. So, on the aggregate, very similar levels of bias are present in the results of both implementations. At the individual document level, the median biases are only a few percentage points apart, for the Recommendation documents (Constrains, DM, N, Ontology) and for other documents (AQ, Dictionary, Sem), however for other documents the median biases are very far apart: for Overview, median bias is at 136% in when the real-time corpus is used, but it is only at 37% when the retrospective corpus is use, a very large difference by a factor of 3.7. Similarly, for DC, median bias is at 87%

Table E.2: Participation: Comparison of median confounding bias (ARD) across documents, in real time and retrospectively

	Real-time	Retrospective
Document	Median	Median
AQ	81%	83%
Constraints	183%	183%
DC	87%	33%
Dictionary	80%	78%
DM	206%	208%
Links	138%	119%
N	174%	183%
Ontology	76%	81%
Overview	136%	37%
Primer	87%	50%
Sem	55%	54%
XML	26%	79%
All documents	102%	97%
Median of descriptives		
Median	87%	80%

with the real-time corpus, but much lower at only 33% with the retrospective corpus (2.6 times lower), while for XML the respective values are 3 times apart (26% versus 79%) – so the retrospective corpus does not consistently have lower bias in these cases, it can have lower bias (as in the case of XML, or indeed for AQ, DM, and other cases). In Primer, the discrepancy is by a factor of 1.7. So, even though overall bias levels are very similar, at the individual document level they may differ, although the bias is not negligible in any of these cases (always above 25% here).

Table E.3: Participation: Comparison of median confounding bias (ARD) across descriptives, in real time and retrospectively

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Real time	5%	11460%	482%	102%	27745%	1666%
Retrospective	3%	11460%	447%	97%	27496%	1658%
Medians of descriptives						
Real time	41%	545%	224%	87%	549%	229%
Retrospective	27%	424%	178%	80%	257%	159%

Table E.3 compares the medians of all descriptives across both implementations. In the 'All documents' descriptives, we see that minima are very close, and maxima are identical. The means are close (note means are sensitive to extreme values), and medians are very close, being only 5 percentage points apart. Variances differ by more than 400 percentage points, and standard deviations are very close, differing by only 8. In the medians of descriptives, the minima and maxima are further apart, but within similar orders of magnitude. Importantly, the medians are close, as discussed above, at 80% and 87%, so bias levels are comparable and very similar overall.

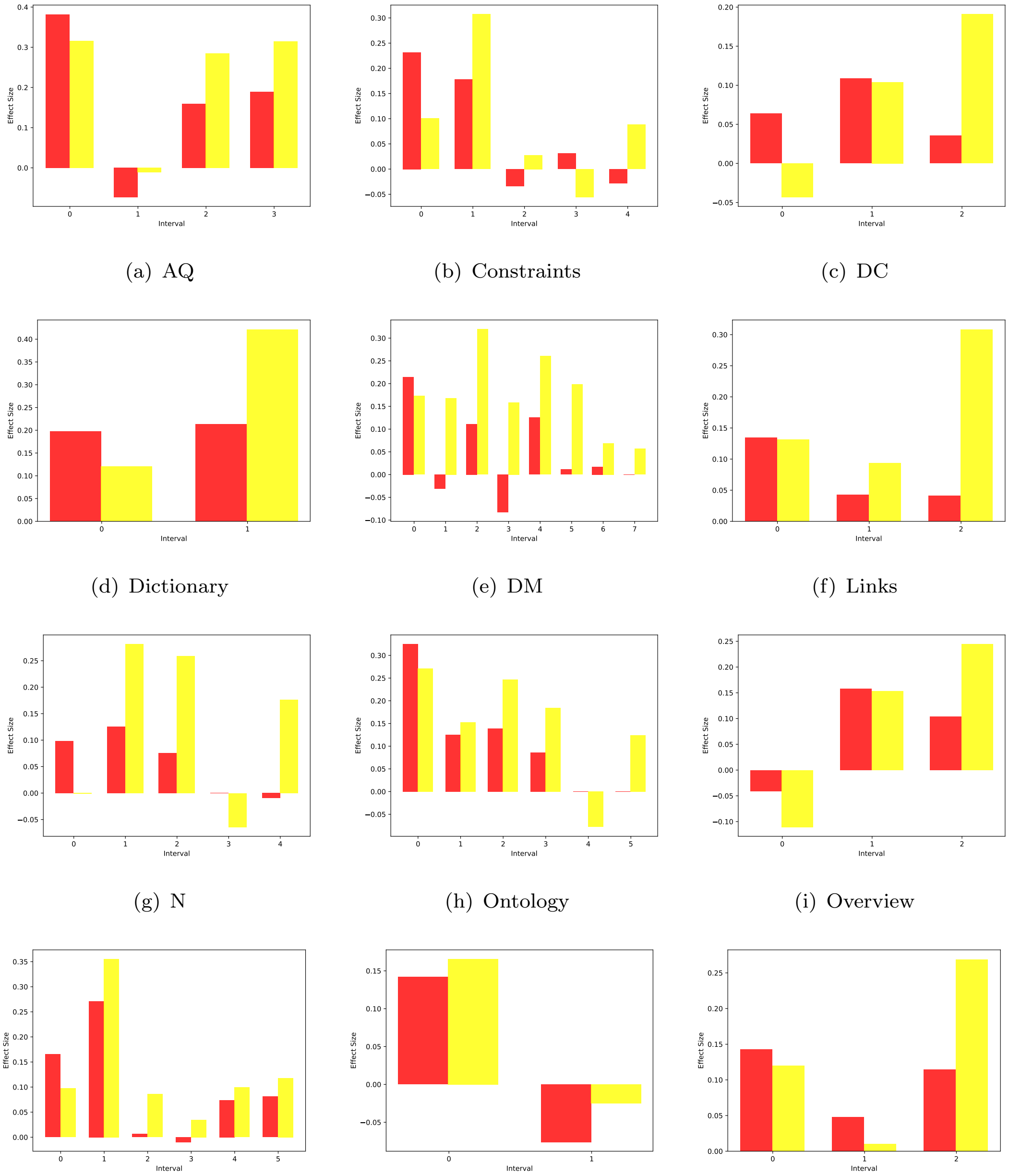
As an aside, it is noted that, for the extra terms that are in the retrospective corpus but not in the original real time corpus (for which $I = 0$ and $F = 0$), the causal effect of P (and of S) on F equals the selection effect. This is derived as follows. The causal effect, as per the backdoor equation (Equation 2.3), expands to

$$P(F = 1|P = 1, I = 1)P(I = 1) + P(F = 1|P = 1, I = 0)P(I = 0)$$
$$-P(F = 1|P = 0, I = 1)P(I = 1) - P(F = 1|P = 0, I = 0)P(I = 0).$$

But since $P(I = 1) = 0$ and $P(I = 0) = 1$ here, the first and third term equal zero, so the causal effect here reduces to $P(F = 1|P = 1, I = 0)P(I = 0) - P(F = 1|P = 0, I = 0)$. And as $P(I = 0) = 1$, this means that $P(P = p, I = 0) = P(P = p)$, so the causal effect becomes $P(F = 1|P = 1) - P(F = 1|P = 0)$, which is the selection effect. So, these terms represent a special case, and for them there is no confounding bias from I in the effect of P on F. I =0 always, so it is essentially a constant, there is no need to adjust for it. It is like a constant in the background. The extent to which the lack of bias from I in these extra terms affects the level of bias in the whole retrospective corpus depends also on the joint distribution of P, I and F in the retrospective corpus.

In summary, although there are some small differences in bias levels for some intervals of some documents, the bias levels are non-negligible, and generally quite high for all documents, for both implementation designs. On the aggregate across documents, confounding bias levels are very close, and high. Patterns of causal and selection effects are also very similar, in terms of sign, effect magnitude, and variation over time and across documents. Overall, it can be said the the effects of email participation on document contents, as well as the level of confounding bias present in them, are generally robust to whether a retrospective or a real-time corpus is used, and the same overall conclusions are reached in terms of their patterns.

Sentiment Moving on to the Sentiment variable, Figure E.2 presents the effects of sentiment (S) in email communications on the next draft (F), for all documents. One can observe that causal and selection effects are generally quite small, very rarely exceeding the 30% mark, they are not always positive, and causal effects are often smaller in later stages of documents' lifetimes than they were in earlier stages. Effect patterns here are in general very similar to those in Figure 7.2 from Section 7.1.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.2: Effects of Sentiment in Online Communications on Current Draft, Using the Retrospective Corpus: Similarly to Figure 7.2 and Figure E.1, the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; causal and selection effects are not always positive, they are sometimes negative; effects are generally relative small, with causal effects never exceeding the 30% mark; causal effects tend to get smaller over time, however selection effects often remain large until later intervals; the magnitude of confounding bias and the temporal patterns show some variation across documents.

It is apparent from Figure E.2 that there is non-negligible confounding bias, since causal and selection effects are frequently quite different in size. Table E.4 quantifies the extent of confounding bias in more detail, using the ARD measure, for all documents. In addition, Table E.5 compares bias levels here versus for the original real-time corpus, for all documents, while Table E.6 compares summary descriptives for the two corpus implementation.

Table E.4: Sentiment: Confounding bias (ARD) across documents, using the retrospective corpus

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	17%	85%	62%	72%	7%	27%
Constraints	57%	409%	200%	183%	174%	132%
DC	4%	439%	204%	168%	321%	179%
Dictionary	39%	97%	68%	68%	8%	29%
DM	19%	1640%	455%	292%	2675%	517%
Links	2%	650%	257%	119%	796%	282%
N	101%	1942%	603%	183%	6010%	775%
Ontology	17%	11460%	2338%	78%	208030%	4561%
Overview	3%	171%	103%	136%	53%	73%
Primer	31%	1222%	306%	43%	1917%	438%
Sem	17%	67%	42%	42%	6%	25%
XML	16%	134%	76%	79%	23%	48%
All documents	2%	11460%	479%	101%	27767%	1666%
Descriptives of descriptives						
Mean	27%	1526%	393%	122%	18335%	591%
Median	17%	424%	202%	99%	247%	155%

Table E.5: Sentiment: Comparison of median confounding bias (ARD) across documents, in real-time and retrospectively

	Real-time	Retrospective
Document	Median	Median
AQ	68%	72%
Constraints	183%	183%
DC	210%	168%
Dictionary	67%	68%
DM	202%	292%
Links	138%	119%
N	179%	183%
Ontology	76%	78%
Overview	136%	136%
Primer	70%	43%
Sem	40%	42%
XML	25%	79%
All documents	101%	101%
Descriptives of descriptives		
Median	106%	99%

Table E.6: Sentiment: Comparison of confounding bias (ARD) across descriptives, in real time and retrospectively

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Real time	3%	11460%	465%	101%	27662%	1663%
Retrospective	2%	11460%	479%	101%	27767%	1666%
Medians of descriptives						
Real time	22%	424%	217%	106%	217%	146%
Retrospective	17%	424%	202%	99%	247%	155%

Together, plots and tables lead to the following conclusions on confounding bias and effect patterns.

Confounding bias Table E.4 shows that confounding bias is not only not negligible, but quite large: the mean over all documents taken together is 479%, and the median is at 101%, while the mean of means is 393% and the median of medians is 99%. Table E.5 shows how these medians of medians bias values are very close to those for the original real-time corpus, and that for each document median bias values are also generally very close, within a few percentage points, although not always (e.g. not for DC, DM, Primer, XML). Even for these exceptions, these patterns are not consistent in direction, i.e. bias in one corpus is not consistently larger than bias in the other). Moreover, even for the exceptions, the confounding bias is not negligible; the minimum value observed is for XML, at 25%, which is still not negligible. Table E.6 shows that all descriptives are in the same order of magnitude, and generally very close, across the two implementations. Overall, the conclusions reached in terms of confounding bias are largely the same in both implementation: bias is not negligible, and in general is quite large.

Effect sign Similarly to the results in Section 7.1, not all effects are positive, as there are several cases of negative effects, both for the causal and the selection effects. Therefore, high sentiment levels for a given term in the emails does not (causally or in terms of association) guarantee a high chance of that term being prominent in the next draft, as discussed in Section 7.1.

Effect magnitude As mentioned, effect magnitudes are relatively small here, as in Section 7.1, and here they sometimes exceed the 30% mark in the first interval (labelled 0), in Figure E.2. There is a slight difference here compared to Figure 7.2 of Section 7.1 where effects did not exceed the 25% mark. This is because the two corpora are at their most different at the first interval: since the retrospective corpus at that interval will include several terms for which both I=0 and F=0 (i.e. terms that are not yet prominent, that are not prominent in the charter nor in the first drafts, but become prominent in later drafts), whereas the real-time corpus at the first interval does not have any terms for which both I=0 and F=0, these extra terms lead to this difference in the effect magnitudes for the two corpora.

Evolution over time Similarly to Section 7.1, over time, the causal effect for sentiment often gets smaller but the selection effect often stays high until later intervals (e.g. Constraints, DM, Links, N, Ontology). Hence, if one were to only calculate the (biased) selection effect, they would reach quite different conclusions about the temporal evolution of the effects of sentiment than if they had calculated the causal effect; that is, confounding bias leads to different temporal patterns. Moreover, within each document, there is large variability over time, as per the variance and standard deviation figures in Table E.4, where the median variance is very high at 247% with the median standard deviation at 155%.

Patterns across documents Table E.4 shows how bias varies across documents. For instance, median bias ranges from 42% for Sem, to 292% for DM, the latter being 6.9 times the former, with no median bias figures being negligible. In addition, temporal patterns of effect sizes vary across documents, in some cases exhibiting opposite tendencies: for instance, the effects in DC and Overview raise then fall, while in XML they fall then raise, in Figure E.2, similarly to what happens in Figure 7.2.

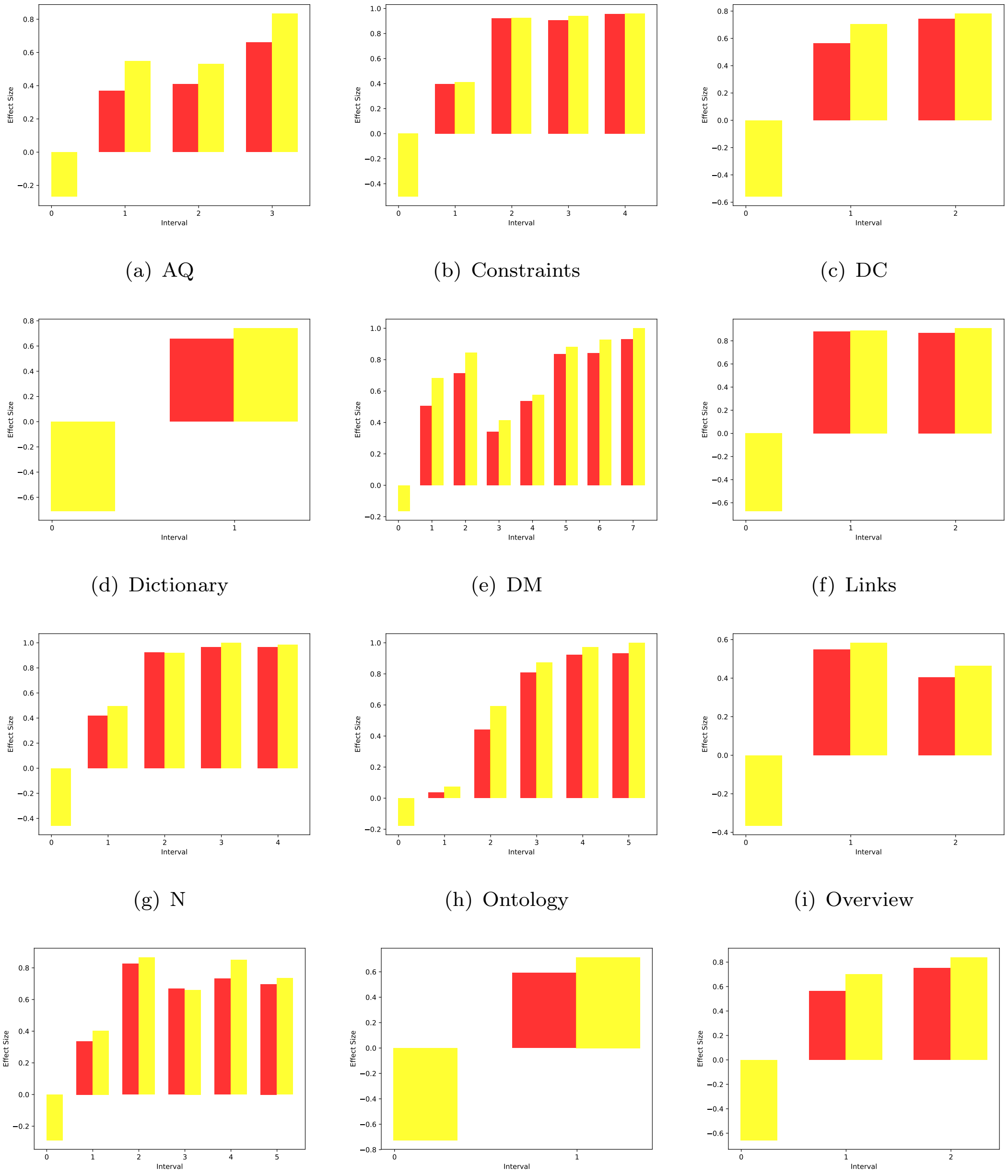
Table E.6 compares the medians of all descriptives across both implementations. In the ‘All documents’ descriptives, we see that minima are very close (3% and 2%), and maxima are identical (11,460%). The means are close (465% and 489%, noting that means are sensitive to extreme values), and medians are identical (at 101%). Variances differ by only 105 percentage points but are both in the 27 hundreds, and standard deviations are very close, differing by only 3 percentage points (1663% and 1663%). In the medians of descriptives, the minima are 5 points apart, while the maxima are again identical. The means, and importantly, the medians are close very close, while variances and standard deviations are also close. So, bias levels are comparable and very similar overall.

As discussed in the analysis of the effects of P on F, for the extra terms that are in the retrospective corpus but not in the original real time corpus (for which $I = 0$ and $F = 0$), the causal effect of S (and of P) on F equals the selection effect. So, these terms represent a special case, and for them there is no confounding bias from I in the effect of P on F. I =0 always, so it is essentially a constant, there is no need to adjust for it. The extent to which the lack of bias from I in these extra terms affects the level of bias in the whole retrospective corpus depends also on the joint distribution of S, I and F in the retrospective corpus.

In summary, although there are some small differences in bias levels for some intervals of some documents, for both implementation designs one may conclude that the bias levels are non-negligible, and generally quite high for all documents. On the aggregate across documents, confounding bias levels are very close, and high. Patterns of causal and selection effects are also very similar, in terms of sign, effect magnitude, and variation over time and across documents. Overall, it can be said the the effects of email sentiment on document contents, as well as the level of confounding bias present in them, are generally robust to whether a retrospective or a real-time corpus is used, and the same conclusions are reached in terms of their patterns.

E.1.2 Effect of previous outcomes

Figure E.3 shows the causal and selection effects of each draft on the next (effects of I on F), for each document. One can observe that causal and selection effects are generally quite large, often above the 60% mark, especially in later stages, they are always positive except for the first interval (which is to be expected due to the nature of the term corpus), and effects tend to increase over documents' lifetimes.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.3: Effects of Previous Draft on Current Draft, Using the Retrospective Corpus: Similarly to Figure 7.3, the causal effect (red) is generally not exactly equal, but still relatively close, to the selection effect (red), across documents, meaning that the presence of some confounding bias can be observed, but that bias is not as large as in Figures E.1 or E.2; the causal effect is often, but not always, smaller than the selection effect; both causal and selection effects are always positive, except at Interval 0 (which will be discussed), for all documents; effects are generally large, compared to Figures E.1 or E.2, with values that can exceed 60% or even 80%, particularly in later intervals; causal and selection effects generally grow, in sync, over time (with the exception of some temporary drops in some cases); across documents, the magnitude of confounding bias can vary, but temporal patterns are often very similar.

Overall, the effect sizes, ranges and temporal patterns in Figure E.3 are very similar to those in 7.3 from Chapter 6.4. It can also be seen from Figure E.3 that there is some confounding bias present, since causal and selection effects are frequently not equal. Table E.7 quantifies the extent of confounding bias in more detail, using the ARD measure, for all documents. Moreover, Table E.8 compares bias levels here versus for the original real-time corpus, for all documents, while Table E.9 compares summary descriptives for the two corpus implementation. Together, these plots and tables lead to the following conclusions on confounding bias and effect patterns.

Table E.7: Previous draft: Descriptives of confounding bias (ARD) across documents, retrospectively

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	26%	49%	35%	30%	1%	10%
Constraints	0%	4%	2%	2%	0%	2%
DC	5%	25%	15%	15%	1%	10%
Dictionary	13%	13%	13%	13%	0%	0%
DM	6%	35%	15%	10%	1%	10%
Links	1%	5%	3%	3%	0%	2%
N	0%	18%	6%	3%	0%	7%
Ontology	5%	98%	31%	8%	12%	35%
Overview	6%	15%	11%	11%	0%	5%
Primer	1%	20%	9%	5%	1%	7%
Sem	21%	21%	21%	21%	0%	0%
XML	12%	24%	18%	18%	0%	6%
All documents	0%	98%	15%	8%	8%	18%
Descriptives of descriptives						
Mean	8%	27%	15%	12%	1%	8%
Median	6%	20%	14%	11%	0%	7%

Confounding bias Figure E.3, similarly to Figure 7.3, shows that there is some confounding bias, but not to the extent that the (biased) selection effect would yield different conclusions for the evolution of the effect over time compared to the (unbiased) causal effect, as both effects generally increase or decrease together. Table E.7 indeed shows the magnitude of confounding bias is generally small: the median of per-document medians is at 11%, with the median of 'all documents' being 8%. These values are very close to the respective for the real-time corpus, which both were at 7%.

Effect magnitude As in Figure 7.3, the causal and selection effects are quite large in Figure E.3. They are particularly high at later stages of the documents' lifetimes, often taking values larger than 60% (AQ, DC, Dictionary, Links, Primer, XML),

Table E.8: Previous draft: Comparison of mean and median confounding bias (ARD) across documents, in real time and retrospectively

	Real-time	Retrospective
Document	Median	Median
AQ	7%	30%
Constraints	2%	2%
DC	9%	15%
Dictionary	5%	13%
DM	9%	10%
Links	2%	3%
N	3%	3%
Ontology	7%	8%
Overview	7%	11%
Primer	3%	5%
Sem	13%	21%
XML	13%	18%
All documents	7%	8%
Descriptives of descriptives		
Median	7%	11%

Table E.9: Previous draft: Comparison of median confounding bias (ARD) across descriptives, in real time and retrospectively

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Real time	0%	99%	11%	7%	2%	16%
Retrospective	0%	98%	15%	8%	8%	18%
Medians of descriptives						
Real time	4%	17%	9%	7%	0%	5%
Retrospective	6%	20%	14%	11%	0%	7%

and in several cases larger than 80% for the four Recommendation documents (Constraints, DM, N, Ontology). And similarly to Figure 7.3, there is a drop in the effects of DM at slice 3, corresponding to the time when DM was split into DM, N and Constraints. Drops shown in Figure E.3 in other documents (Primer interval 3 and 5, Overview interval 2) also mirror those in Figure 7.3.

Effect sign Effect signs are universally positive, except for the selection effects at interval 0 which are negative, as in Figure 7.3. Indeed, here for many documents, the interval 0 selection effects are not as negative as those in Figure 7.3. This is attributable to the fact that, in this hindsight corpus, at every interval there will likely be some terms in the corpus which do not appear in either the initial nor the final draft of that interval. This means that $P(F = 0|I = 0)$ will be non-zero, hence the selection effect, which equals $P(F = 0|I = 0) - P(F = 1|I = 0)$ will have a non-zero first term, which brings up the selection effect, compared to the original real-time corpus, which by definition does not include any terms that do not appear in any of I or F, the $P(F = 0|I = 0)$ will always be zero for the first interval.

Patterns over time Similarly to Section 7.1, over time, the causal and selection effects tend to both get larger, in Figure E.3. In terms of bias, the two effects tend to increase in pace with each other, and their heights are not very different. Hence, if one were to only calculate the (biased) selection effect, one would not reach particularly different conclusions about the temporal evolution of the effects than if they had calculated the causal effect. So, if one is willing to sacrifice some accuracy for the ease of not adjusting for confounders, the conclusions reached would not be dramatically different. Moreover, within each document, there is not much variability in bias over time, as per the variance and standard deviation figures in Table 6.9, where most documents' variance is at 0% or 1%, with the exception of Ontology (12%), while standard deviations are not extremely large either, ranging from 0% to 35%, with a median of standard deviations at 7%.

Patterns across documents Table 7.3 shows how bias varies across documents: median bias ranges from only 2%, for Constraints, to a maximum of 15 times that, at 30%, for AQ. There are several cases where median bias is very low, and one might say negligible, at single-digit values: median bias is only 3% for Links and N, at 5% for Primer, and at 8% for the Ontology document. Still, for the remaining documents bias is much larger, at two-digit levels, and could not be said to be negligible, like AQ (median of 30%), Sem (21%), XML (18%), and DC (15%). The median over all documents is low, at only 8%, with the median of medians at 11%. So, on the aggregate, bias is certainty low, and might perhaps be considered negligible . As Table E.8 shows, the medians over all documents and the medians of medians are very close under both implementations. For most documents, those values are the same or very close. However, for some documents, the median bias values can be different; this is particularly the case for AQ, where the median bias with the real-time corpus is only 7% but with the retrospective corpus it is much higher at 30%, while Dictionary and Sem also have much higher median bias values with the retrospective corpus. In the case of AQ, comparing Figures E.3(a) and Figure 7.3(a) shows that this large difference in bias is due to intervals 1 and 2, where the selection effect is much more different than the causal effect in with the retrospective corpus (Figure E.3(a)) than with the real-time corpus (Figure 7.3(a)). So, while aggregate statistics across documents are very similar across implementations, at the individual document level there are cases where bias levels are quite different. In terms of the temporal patterns of effects, these are mostly consistent, with effects monotonically increasing in most documents, with some cases showing occasional drops (DM, Primer, Overview), as was also the case in Section 7.1. So conclusions about temporal patterns of effects are the same for both implementations designs.

Table E.9 compares the medians of all descriptives across both implementations. In the 'All documents' descriptives, the minima are identical(0%), and maxima are one percentage point apart and both very close to 100% (98% and 99% - these both correspond to the maximum bias observed in the Ontology document, as per Tables E.8 and 7.3). The means are close (11% and 15%, noting that means are sensitive to extreme values), and medians are only one percentage point apart (at 7% and 8%). Variances differ by only 6 percentage points and are both quite low (2% and 8%), and similarly standard deviations are both quite low and very close, differing by only 2 percentage points (16% and 18%). In the medians of descriptives, the minima are only 2 points apart (4% and 6%), while the maxima are only 3 points apart (17% and 20%). The means, and importantly, the medians are very close across both implementations, while median variances are identical at 0% (as mentioned, numbers are rounded to the nearest integer), and median standard deviations are also low and close (5% and 7%). So, bias levels are comparable and very similar overall.

In addition, the following pattern is observed with respect to the magnitude of selection effects in the retrospective corpus, versus in the original real time corpus. In the retrospective corpus, for all time intervals, selection affects are larger than the respective selection effects of the original corpus. That is because $P(F = 0|I = 0)$ is higher in the retrospective corpus, compared to original corpus, as by construction the retrospective corpus contains extra terms which are in the top terms of future drafts. Hence, these extra terms have I, F= 0 (otherwise they would already be in the real-time corpus). So, the addition of these terms in the corpus increases the $P(F = 0)$, the $P(I = 0)$, and the $P(F = 0|I = 0)$. Therefore, the first term of the selection effect, $P(F = 0|I = 0)$, is larger in the retrospective corpus than in the original corpus, while the second term of the selection effect, $P(F = 0|I = 1)$, is the same in both corpora (as the number of terms with I=1 stay the same). So the selection effect is larger (or the same) in the retrospective corpus than in the original corpus.

For the effects of this variable, it seems there is slightly less confounding in the real-time implementation, when not using hindsight, when only considering data up to this time point, not including data from future time intervals. On a per-document basis, this also holds for all documents, except Constraints and N (where the bias values are equal): the median and the mean bias with hindsight (retrospectively) are greater than or equal to the mean and median bias, respectively, without hindsight. Given that we have established that selection effects are higher in the retrospective corpus than in the original corpus, for the bias in the retrospective corpus to be greater than or equal to the bias in the original corpus, the discrepancy between causal effect and selection effect must be wider. So, the causal effect must be smaller (or the same) in the retrospective corpus than in the original corpus. By inspecting the causal and selection effect data, it is found that this is indeed the case: causal effects are either always smaller, or smaller in the first intervals and the same in the last few intervals compared to the causal effects of the original corpus.

It does not follow from the construction of the corpus that the causal effect will always be smaller, as it followed that the selection effect will always be larger. That is because the causal effect also accounts for confounders which are outside the corpus-related variables (S, P from the emails), and it is the distribution of these confounders, as well as the joint distributions of S, P, I and F that determine the causal effects.

From the formula of the causal effect, only the following can be deduced in terms of comparison to the original corpus: all terms that condition on $I = 1$ are the same, as those entries are the same in both corpora. Moreover, all terms of the form $P(F = 1|I = 0, P = p)$ for all p are smaller, as the retrospective corpus has extra terms with I=0, while the number of terms with F=1 is the same (so, constant numerator, larger denominator). The rest of the terms, i.e. $P(P = 0)$ and its complement, may end up being larger or smaller in the retrospective corpus, depending on the distribution of P in the original corpus and i the extra terms added. Therefore, the overall result on whether the causal effect under the retrospective corpus is larger or smaller than that under the original corpus depends on how much larger or smaller $P(P = 0)$ is in the retrospective corpus, and how much smaller the $P(F = 1|I = 0, P = p)$ terms are.

In other words, as the difference between the causal and selection effect formulae is that the causal effect adjusts for confounders (P, S) whereas the selection effect does not, it must be that the distribution of these confounders, and/or its joint distribution with I and F are such that causal effects are lower here than in the original corpus. Indeed, this corpus essentially represents a different, larger dataset: by adding in terms that appear only in later drafts, we also add in the email conversations pertaining to these extra terms, from which to extract the S and P values for these terms. Noting also that S and P are binarized using the median S and P as the threshold, this means that the addition of extra terms, with their S and P values, may change the median, and hence the binarization threshold, and hence alter the probabilities that P and S are high or low, compared to the real-time corpus. By inspecting the data, it is observed that there are indeed cases (e.g. Ontology) when this happens. Therefore, as the two corpora correspond to different datasets being analysed, some discrepancy in the results and in the bias levels is to be expected as a normal consequence of the differences in the data sets. Still, for the purposes here, what matters is that bias levels here are comparable to those in the original implementation.

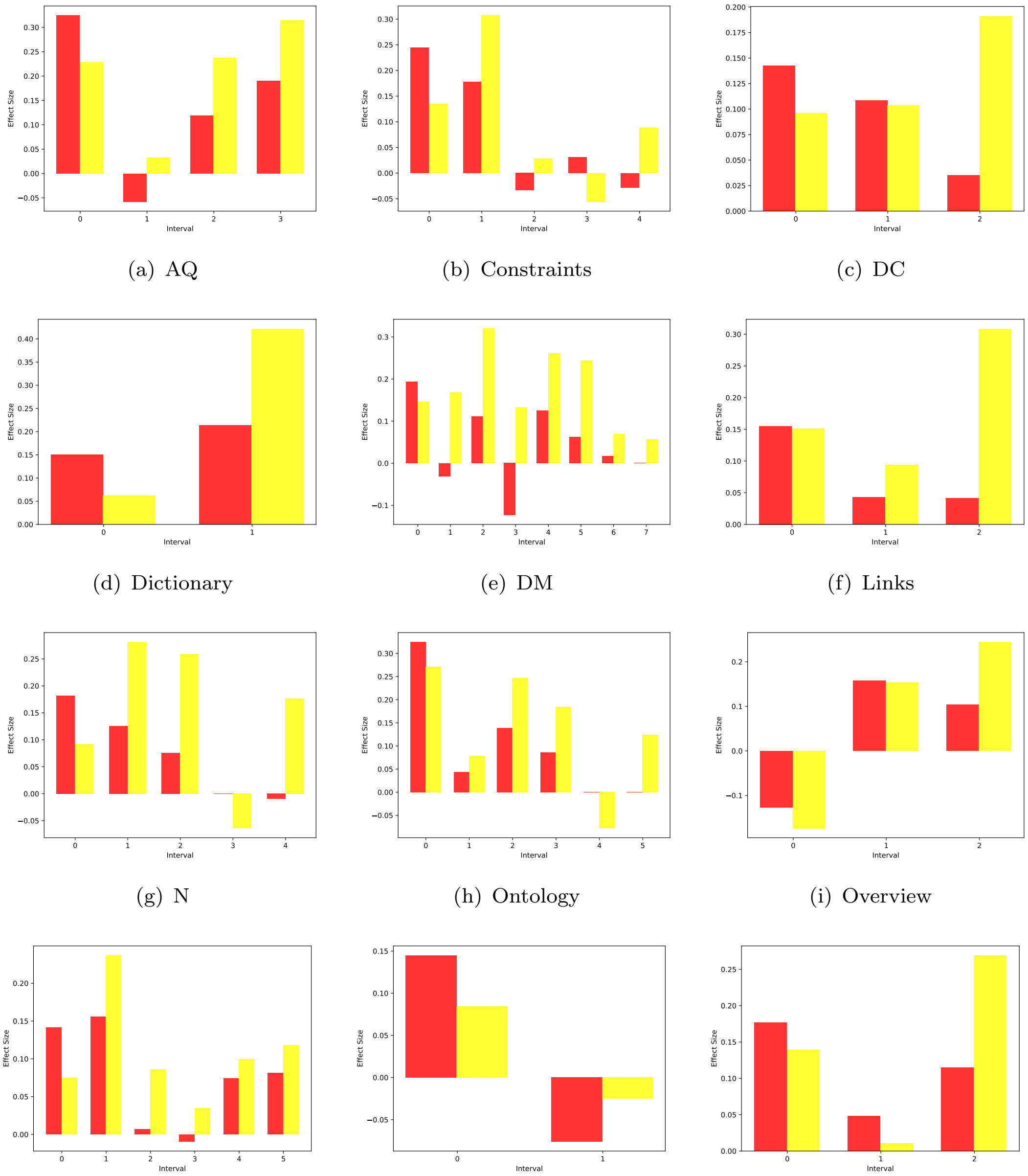
In summary, under the retrospective corpus, although there may be some small differences in bias levels for some intervals of some documents, one may conclude that the bias levels are generally low across documents, and do not particularly affect the conclusions reached about the properties and patterns of causal versus selection effects. This is the same conclusion as was reached under the original real-time corpus implementation design.

Hence, for both types of corpus (retrospective and real-time), one reaches the same conclusions in terms of bias and effect properties: the bias levels are similarly and comparably low, and the causal and selection effects have the same characteristics and patterns (over time, across contexts). So, overall, it can be said the the effects of the previous draft on the next, as well as the level of confounding bias present in them, are generally robust to whether a retrospective or a real-time corpus is used, and the same conclusions are reached in terms of their patterns.

E.1.3 Summary

In summary, bias levels are very similar when using this retrospective corpus versus when using the original, real-time, corpus. As discussed, the properties of the extra terms included in the retrospective corpus may help explain why for Participation (and Sentiment, to an extent) there tends to be less overall bias with the retrospective corpus than with the original corpus: these terms, which have no confounding from from the previous draft (variable I), bring the overall confounding level down. And the mirror of this happens for the bias present in the effects of the previous draft (I) on the next (F): due to these new terms where I plays no active role, there is slightly more confounding from P and S to the effect of I on F.

In conclusion, although there are a few small differences (of the order of a few single-digit percentage points) for some documents in terms of the extent of bias in the original corpus and in the retrospective corpus, overall both corpora result in broadly the same overall amount of bias, and the same characteristics and patterns (over time and across documents) of causal and selection effects. Overall, the same conclusions are reached about the extent of bias, and about the characteristics and patterns of causal and selection effects, for both corpora.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.1: Effects of Participation in Email Communications on Current Draft, Using the Retrospective Corpus: Similarly to Figure 7.1, the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; causal and selection effects are not always positive, they are sometimes negative; effects are generally relative small, with causal effects never exceeding the 30% mark; causal effects tend to get smaller over time, which is not the case for selection effects which instead often peak at later intervals; the magnitude of confounding bias and the temporal patterns show some variation across documents.

E.2 Expanded real-time corpus analysis

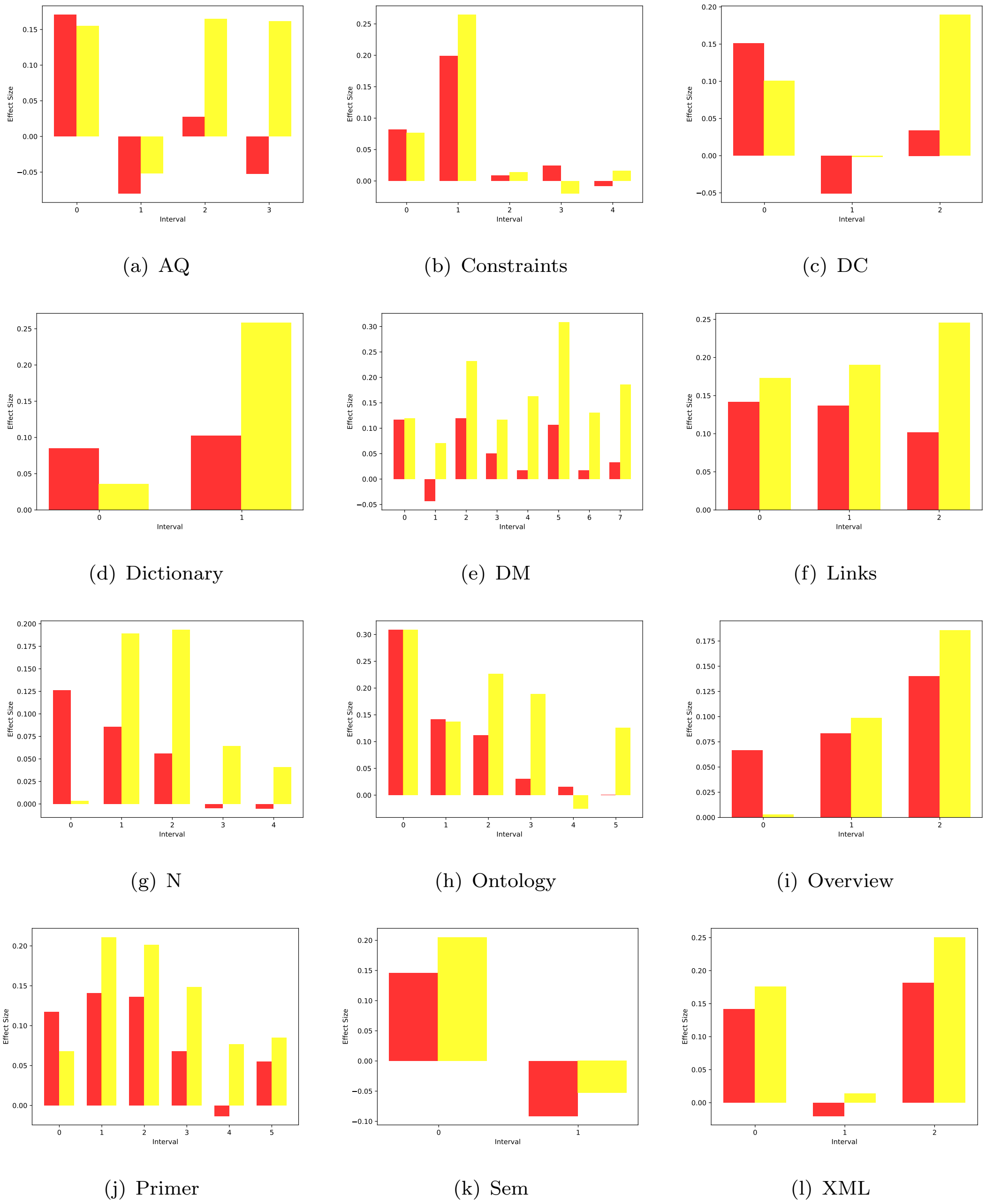
This subsection considers a corpus containing terms from the same document drafts as the original, real time corpus of Section 7.1, only this time we allow in *more* of the top concepts of each draft. Where originally the corpus obtained the top terms for each draft using only the top 100 features, the corpus in the present subsection will be expanded, as it will instead use the top 200 feature to select the top frequent terms from each draft. The details of this have been presented in Chapter 6.4.

This subsection begins with the analyses of the impact of the characteristics of email conversations (first Participation, then Sentiment) on the contents of each draft, and proceeds with the analysis of the causal and selection effects of the previous outcome (draft) on the current one.

E.2.1 Social Influence of email communications

We first consider the effects of participation (P) and of sentiment (S) in email communications on the next draft (F).

Participation Figure E.4 shows the causal and selection effects of participation (P) in email communications on the next draft (F), across all documents. Here, there are differences in effect sizes for several documents and time intervals, compared to Figure 7.1 for all documents. This is to be expected, as the concept corpora are very different, the current one containing many more terms than the original one. However, the overall patterns remain the same: effects are generally small, sometimes positive and sometimes negative, often being smaller at later time stages than at earlier, and patterns vary across documents, as in Figure 7.1.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.4: Effects of Participation Volume in Emails on Current Draft, Using the Larger Corpus: Similarly to Figure 7.1, the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; effects are generally relative small, with causal effects rarely exceeding the 25% mark; causal and selection effects are sometimes positive and sometimes negative; causal effects often (but not always) are smaller in later intervals than in earlier ones, which is not the case for selection effects which instead often peak at later intervals; the magnitude of confounding bias and the temporal patterns show some variation across documents.

As the extent of confounding bias is of particular interest, Table E.10 quantifies the extent of confounding bias in more detail, using the ARD measure, for all documents.

In addition, Table E.11 compares bias levels here versus for the original real-time corpus, for all documents, while Table E.12 compares summary descriptives for the two corpus implementation.

Table E.10: Participation: Descriptives of confounding bias (ARD) across documents, with the larger corpus

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	9%	503%	240%	223%	484%	220%
Constraints	7%	300%	118%	65%	120%	109%
DC	33%	456%	195%	97%	345%	186%
Dictionary	58%	152%	105%	105%	22%	47%
DM	3%	859%	334%	227%	797%	282%
Links	22%	141%	67%	39%	28%	52%
N	97%	1548%	584%	244%	3204%	566%
Ontology	0%	530%	180%	103%	398%	200%
Overview	19%	95%	49%	33%	11%	33%
Primer	42%	680%	165%	52%	536%	232%
Sem	40%	42%	41%	41%	0%	1%
XML	24%	169%	77%	38%	42%	65%
All documents	0%	1548%	214%	97%	878%	296%
Descriptives of descriptives						
Mean	30%	456%	180%	106%	499%	166%
Median	23%	378%	141%	81%	233%	148%

Table E.11: Participation: Comparison of median confounding bias (ARD) across documents, lager corpus versus original corpus

	Larger corpus	Original Corpus
Document	Median	Median
AQ	223%	83%
Constraints	65%	183%
DC	97%	33%
Dictionary	105%	78%
DM	227%	208%
Links	39%	119%
N	244%	183%
Ontology	103%	81%
Overview	33%	37%
Primer	52%	50%
Sem	41%	54%
XML	38%	79%
All documents	97%	97%
Descriptives of descriptives		
Median	81%	80%

Table E.12: Participation: Comparison of median confounding bias (ARD) across descriptives, original corpus versus lager corpus

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Original	5%	11460%	482%	102%	27745%	1666%
Larger	0%	1548%	214%	97%	878%	296%
Medians of descriptives						
Original	41%	545%	224%	87%	549%	229%
Larger	23%	378%	141%	81%	233%	148%

Together, these plots and tables lead to the following conclusions on the pattern of causal and selection effects, and the amount of bias between them:

Confounding bias Table E.10 shows that confounding bias is not negligible, but rather quite large: the median of medians is at 81%, with the median across all documents at 97%. For several documents (AQ, DM, N) the median is higher than 200%, and for several others it is higher than 100%. The maximum bias value present in any document and any interval is 1,548% for the N (Notation) document. High bias values such as this affect the mean for 'all documents' and the mean of means, which are hence higher than the median, at 214% and 141% respectively.

Effect sign Not all effects are positive here; rather, there are several cases of negative effects, both for the causal and the selection effects. Therefore, high participation volume for a given term in the emails does not guarantee a high chance of that term being prominent in the next draft, neither causally nor in terms of association, as was also the case in Section 7.1.

Effect magnitude Effects here are relatively small, and follow similar patterns as in Section 7.1, with causal effects in Figure E.4 rarely being higher than 25%.

Evolution over time Similarly to Section 7.1, Figure E.4 over time, the causal effect of participation often gets smaller (e.g. Constraints, DC, DM, Links, N, Ontology; although for Overview it increases), which however does not hold for the selection effect. For example, here, for Links, the selection effect actually increases over time, while the causal effect falls; For Ontology and N, the causal effect decreases, while the selection effect increases and then falls only a little bit, staying at much higher levels than the causal effect . Hence, if one were to only calculate the (biased) selection effect, they would in many cases reach quite different conclusions about the temporal evolution of the effects of participation than if they had calculated the causal effect; that is, confounding bias leads to different temporal patterns. Moreover, within each document, there is large variability over time, as indicated by the variance and standard deviation figures in Table E.10, where the median variance is very high at 233% with the median standard deviation at 148%.

Patterns across documents From Table E.10 one can see that the magnitude of bias varies across documents. For instance, if one takes the median bias across documents, one sees that N has the highest median bias at 244%, followed by DM at 227% and AQ at 223%, while the lowest median bias is 33% for Overview. Overall, given that the highest and lowest median bias figures (244% versus 33%) differ by a factor of almost 7.4, and that the median values across documents take a range of values, the overall level of bias can differ quite broadly across documents (contexts). In addition, temporal patterns vary across documents, in some cases exhibiting opposite tendencies - for instance, the causal effect gets smaller over time for Constraints, Links, N, and Ontology (Figure E.4), but it increases for Overview and for Dictionary, while for AQ, DC, and DM, monotonicity varies more wildly (several increases and decreases, less smooth development over time). This shows that temporal patterns were not the same for all documents, but rather there were some documents where the impact of participation in emails on their content did not decrease over time.

Table E.11 shows that median bias is for some documents higher for the larger corpus (e.g. AQ), and sometimes larger for the original corpus (e.g Constraints). So the retrospective corpus does not consistently have lower or larger bias for all documents than the original corpus. The overall median biases across documents are very close, within 5-7 percentage points: the medians of medians are within 1 percentage points (87% and 80%) while the medians when all documents' datapoints are used together are identical, at 97%. So, on the aggregate, very similar levels of bias are present in the results of both implementations. However, at the individual document level, the median biases tend to be further apart: there are some documents where bias values differ by only a few percentage points (e.g. DM, Primer, Sem), there also exist documents for which values vary much more widely, e.g. by a factor of around two (DC, Dictionary), by a factor of 2.8 (Constraints), and more extreme cases, where median bias varies by a factor of 3.5 (Links) and 4.1 (Overview). Still, in all cases. Still, even in those extreme cases, there is never a document that has negligible median bias (e.g. single digit) in one implementation and very large bias (e.g. 3 digits) in the other implementation. So, even bias levels are very similar, and even though at the individual document level they may differ, the bias is not negligible in any of these cases (always at least 25%).

Table E.12 compares the medians of all descriptives across both implementations. In the 'All documents' descriptives, we see that minima are close (0% versus 5%), however the maxima are far apart, with the 'all documents' maximum of the original corpus (corresponding to Ontology) being 7.4 larger than the maximum for the larger corpus (corresponding to N). Indeed, the maximum values of median bias correspond to different documents, as in the plot for Overview when using the original corpus (Figure 7.1(i)) both causal and selection effects are different, and have a larger discrepancy between causal and selection effects (particularly interval 0 and 2), than in the plot for Overview using the larger corpus (Figure E.4(i)). Moreover, in the plot for N under the original implementation (Figure 7.1(g)), the dicrepancies between causal and selection effects are somewhat smaller than in the plot for N under the larger corpus (Figure E.4(g)), hence the median bias differs under the two implementations.

In Table E.12, the medians of maximum bias values are not very close, but are in the same order of magnitude, for both the original and the large corpus (545% and 378%). The means are close (noting that the means are sensitive to extreme values), especially the medians of means (224% and 141%). The 'all documents' medians are very close,being only 5 percentage points apart, and so are the medians of medians (8% and 81%)). Variances and standard deviations are all high overall, although they are higher for the original corpus, similarly to what was observed for the maximum values in this table. Importantly, the medians are close, as discussed above, at 87% and 81%, so bias levels are comparable and very similar overall.

Overall, the levels of confounding for this implementation are very similar to those in the original implementation of Section 7.1 (81% median bias here versus 87% in the original implementation). The overall characteristics of causal and selection effects, and the patterns of how they vary over time and across documents, are also the same: effects are generally small (largely lower than 25% in magnitude), with causal effects often being smaller in later stages than earlier stages of documents' lifetimes, which does not hold for selection effects, and with variation in temporal patterns across documents.

In summary, although there are some differences in bias levels for some documents, the bias levels are non-negligible, and generally quite high for all documents, for both implementation designs. On the aggregate across documents, confounding bias levels are very close, and high. Patterns of causal and selection effects are also generally very similar, in terms of sign, effect magnitude, and variation over time and across documents. Overall, it can be said the the effects of email participation on document contents, as well as the level of confounding bias present in them, are generally robust to whether the original corpus or the larger corpus is used, and the same overall conclusions are reached in terms of their patterns.

Sentiment Figure E.5 shows the causal and selection effects of participation (S) in email communications on the next draft (F), across all documents, over time. Here, there are some differences in effect sizes for several documents and time intervals, compared to Figure 7.2 for all documents. This is to be expected, as the concept corpora are very different, the current one containing many more terms than the original one. However, the overall patterns remain the same: in Figure E.5, it can be seen that counfounding bias exists (causal effects heights differ from selection effect heights), effects are generally small, sometimes positive and sometimes negative, often being smaller at later time stages than at earlier, and patterns also vary across documents, as in Figure 7.2.

As the extent of confounding bias is of interest, Table E.13 quantifies the extent of confounding bias in more detail, using the ARD measure, for all documents. To determine whereas confounding results are significantly different in this implementation compared to the original one for Section 7.1, Table E.14 compares bias levels here versus for the original real-time corpus, for all documents, while Table E.15 compares summary descriptives for the two corpus implementation.

TABLE E.13: Sentiment: Descriptives of confounding bias (ARD) across documents, with larger corpus

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	23%	411%	159%	102%	252%	159%
Constraints	33%	300%	124%	65%	107%	103%
DC	76%	456%	209%	97%	304%	174%
Dictionary	105%	152%	128%	128%	6%	24%
DM	13%	859%	332%	227%	806%	284%
Links	25%	141%	68%	39%	27%	52%
N	120%	1548%	593%	244%	3118%	558%
Ontology	1%	530%	182%	103%	393%	198%
Overview	19%	56%	36%	33%	2%	15%
Primer	32%	680%	162%	51%	545%	234%
Sem	40%	42%	41%	41%	0%	1%
XML	12%	169%	73%	38%	47%	68%
All documents	1%	1548%	209%	99%	855%	292%
Descriptives of descriptives						
Mean	42%	445%	176%	97%	467%	156%
Median	28%	356%	144%	81%	180%	131%

TABLE E.14: Sentiment: Comparison of mean and median confounding bias (ARD) across documents, lager corpus versus original corpus

	Larger Corpus	Original Corpus
Document	Median	Median
AQ	102%	68%
Constraints	65%	183%
DC	97%	210%
Dictionary	128%	67%
DM	227%	202%
Links	39%	138%
N	244%	179%
Ontology	103%	76%
Overview	33%	136%
Primer	51%	70%
Sem	41%	40%
XML	38%	25%
All documents	99%	101%
Descriptives of descriptives		
Median	81%	106%

TABLE E.15: Sentiment: Comparison of median confounding bias (ARD) across descriptives, original corpus versus lager corpus

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Original	3%	11460%	465%	101%	27662%	1663%
Larger	1%	1548%	209%	99%	855%	292%
Medians of descriptives						
Original	22%	424%	217%	106%	217%	146%
Larger	28%	356%	144%	81%	180%	131%

Together, the plots of Figure E.5 Tables E.13, E.14 and E.15 lead to the following conclusions on the pattern of causal and selection effects, and the amount of bias between them.

Confounding bias Table E.13 shows that confounding bias is not negligible, but rather quite large: the median of medians is at 81%, with the median across all documents at 99%. For some documents (DM, N) the median is higher than 200%, with the lowest median bias being at 33% for Overview, which is not negligible. The maximum bias value present in any document and any interval is 1,548% for the N (Notation) document. High bias values such as this affect the mean for ‘all documents’ and the mean of means, which are hence higher than the median, at 209% and 176% respectively.

Effect sign Not all effects are positive in the plots of Figure E.5; rather, there are several cases of negative effects, both for the causal and the selection effects. Therefore, high sentiment for a given term in the emails does not guarantee a high chance of that term being prominent in the next draft, neither causally nor in terms of association (as causal and selection effects are not guaranteed to be positive), as was also the case in Section 7.1.

Effect magnitude Effects here are relatively small, and follow similar patterns as in Section 7.1, with causal effects in Figure E.5 rarely being higher than 25% (in fact, only once, for interval 0 of Ontology).

Evolution over time Similarly to Section 7.1, Figure E.5 over time, the causal effect of the Sentiment variable often is smaller in the last few intervals than it is at earlier intervals (e.g. Constraints, DC, N, Ontology, Primer), which however does not hold for the selection effect. For example, here, for Links, the selection effect actually increases over time, while the causal effect first slightly rises and then slightly falls; for N, the causal effect decreases, while the selection effect increases and then falls only a little bit, staying at much higher levels than the causal effect . Hence, if one were to only calculate the (biased) selection effect, they would in many cases reach quite different conclusions about the temporal evolution of the effects of participation than if they had calculated the causal effect; that is, confounding bias leads to different temporal patterns. Moreover, within each document, there is large variability over time, as indicated by the variance and standard deviation figures in Table E.13, where the median variance is very high at 180% with the median standard deviation at 131%.

Patterns across documents From Table E.13 one can see that the magnitude of bias varies across documents. For instance, if one takes the median bias across documents, one sees that N has the highest median bias at 244%, followed by DM at 227%, while the lowest median bias is 33% for Overview. Overall, given that the highest and lowest median bias figures (244% versus 33%) differ by a factor of almost 7.4, and that the median values across documents take a range of values, the overall level of bias can differ quite broadly across documents (contexts). In addition, temporal patterns vary across documents, in some cases exhibiting opposite tendencies - for instance, the causal effect are smaller in later stages than at earlier stages for Constraints, N, Ontology and Primer (Figure E.5), but it increases for Dictionary; for Links, the causal effect falls slightly then rises, whereas for XML the positive happens, as there the causal effect first falls and then increases; while for AQ, and DM, monotonicity varies more (several increases and decreases, less smooth development over time). This shows that temporal patterns were not the same for all documents, but rather there were some documents where the impact of email sentiment on their content did not decrease over time.

In confounding bias subsection: Table E.14 shows how the median confounding bias values, across all documents, are very close for both corpus choices (at 99% and 101%). The median of medians values are also close, even if not as close, at 81% and 106%. Similarly to what was observed for the Participation variable, here too median bias values for each document often differ widely across the two corpus choices: for instance, for Links, median bias values differ by 3.5 times (39% versus 138%), and for Overview by a factor of 4.12 (33% versus 136%), while for other documents median bias values are closer, for example for Sem (41% and 40%) and for Primer (51% versus 70%). But even for cases where median values of bias are far apart, the confounding bias is never negligible (minimum of median values at 33% for the larger corpus, and at 25% for the original corpus). In addition, bias in one corpus is not consistently larger than bias in the other.

Table E.6 compares the medians of all descriptives across both implementations. In the ‘All documents’ descriptives, we see that minima are very close (3% and 1%), but maxima are quite far apart (11,460% versus 1,548%), as was also the case for the Participation variable, discussed in the previous section. The means are relatively close and in the same order of magnitude (465% and 209%, noting that means are sensitive to extreme values), and medians are very close (at 101% and 99%). Variances and standard deviations are quite different, similarly to what was observed for the Participation variable, in the previous section. In the medians of descriptives, the minima are only six percentage points apart, while the maxima are not very far apart (424% and 356%). The means, and importantly, the medians are relatively close (medians at 106% and 81%), while variances and standard deviations are also close. So, bias levels are comparable and very similar overall.

Overall, the levels of confounding present in the results for this implementation are very similar to those in the original implementation of Section 7.1: as Table E.15 shows the median of medians bias is 81% here versus 106% in the original implementation, while the medians across all documents taken together are very close, at 99% and 101% respectively. The overall characteristics of causal and selection effects, and the patterns of how they vary over time and across documents, are also the same: effects are generally small (mostly lower than 25% in magnitude), with causal effects often being smaller in later stages than earlier stages of documents' lifetimes, which does not hold for selection effects, and with variation in temporal patterns across documents.

In summary, although there are some differences in bias levels for some documents, for both implementation designs one may conclude that the bias levels are non-negligible, and generally quite high for all documents. On the aggregate across documents, confounding bias levels are close, and quite high. Patterns of causal and selection effects are also very similar, in terms of sign, effect magnitude, and variation over time and across documents. Overall, it can be said the the effects of email sentiment on document contents, as well as the level of confounding bias present in them, are generally robust to whether the original or the expanded corpus is used, and the same conclusions are reached in terms of their patterns.

E.2.2 Effect of previous outcome

For the effects of the previous outcome, that is, the previous draft (variable I) on the next one (variable F), the causal effects and selection effects for each inter-draft interval of each documents are presented in Figure E.6.

The plots of Figure E.6 show that the causal effect is generally not exactly equal, but is still relatively close, to the selection effect, across documents. Hence, one can observe the presence of some confounding bias due to the contents of the email conversations (Sentiment and Participation variables) which the selection effect ignores. In addition, in many cases, the causal effect is smaller than the selection effect, but in some cases the opposite holds.

To summarize the magnitude of confounding bias present for each document and across documents, the descriptives of confounding bias (measured using the ARD formula, of Equation 6.1) are presented in Table E.16. Moreover, Table E.17 compares bias levels here versus for the original real-time corpus, for all documents, while Table E.18

compares summary descriptives for the two corpus implementation. Together, these plots and tables lead to the following conclusions on confounding bias and effect patterns.

Table E.16: Previous draft: Descriptives of confounding bias (ARD) across documents, with larger corpus

Document	Minimum	Maximum	Mean	Median	Variance	Std Dev
AQ	5%	11%	8%	8%	0%	2%
Constraints	0%	38%	10%	1%	3%	16%
DC	3%	3%	3%	3%	0%	0%
Dictionary	2%	2%	2%	2%	0%	0%
DM	2%	23%	9%	5%	1%	7%
Links	2%	5%	3%	3%	0%	1%
N	1%	22%	7%	4%	1%	8%
Ontology	1%	110%	25%	4%	18%	43%
Overview	15%	25%	20%	20%	0%	5%
Primer	0%	13%	6%	6%	0%	4%
Sem	9%	9%	9%	9%	0%	0%
XML	4%	10%	7%	7%	0%	3%
All documents	0%	110%	10%	5%	3%	18%
Descriptives of descriptives						
Mean	4%	23%	9%	6%	2%	8%
Median	2%	12%	8%	5%	0%	4%

Table E.17: Previous draft: Comparison of median confounding bias (ARD) across documents, lager corpus versus original corpus

	Larger corpus	Original corpus
Document	Median	Median
AQ	8%	7%
Constraints	1%	2%
DC	3%	9%
Dictionary	2%	5%
DM	5%	9%
Links	3%	2%
N	4%	3%
Ontology	4%	7%
Overview	20%	7%
Primer	6%	3%
Sem	9%	13%
XML	7%	13%
All documents	5%	7%
Descriptives of descriptives		
Median	5%	7%

Confounding bias Figure E.6, similarly to Figure 7.3, shows that there is some confounding bias, but not to the extent that the (biased) selection effect would yield different conclusions for the evolution of the effect over time compared to the (unbiased) causal effect, as both effects generally increase or decrease together. Table

Table E.18: Previous draft: Comparison of median confounding bias (ARD) across descriptives, original corpus versus lager corpus

Corpus	Minimum	Maximum	Mean	Median	Variance	Std Dev
All documents						
Original	0%	99%	11%	7%	2%	16%
Larger	0%	110%	10%	5%	3%	18%
Medians of descriptives						
Original	4%	17%	9%	7%	0%	5%
Larger	2%	12%	8%	5%	0%	4%

E.16 indeed shows the magnitude of confounding bias is generally small: the median of per-document medians is at 5%, with the median of ‘all documents’ being also 5%. These values are very close to the respective for the real-time corpus, which both were at 7%. Similarly, the mean of means is at 9% with the mean of ‘all documents’ at 10%. These mean values are also low, but higher than the median values, as they are affected by a few high bias values for some documents - in particular, Ontology has a very high mean at 25%, while Overview has a a very high mean bias (and the highest median bias), at 20%, so these values brings up the mean over all documents. Apart from Overview, all other documents have single-digit median bias values, and all documents apart from Constraints, Ontology, and Overview have singe-digit mean bias values.

Effect magnitude As in Figure 7.3, the causal and selection effects are quite large in Figure E.6. They are particularly high at later stages of the documents’ lifetimes, often taking values larger than 60% (AQ, DC, Dictionary, Links, Primer, Overview, Sem, XML), and in several cases larger than 80% for the four Recommendation documents (Constraints, DM, N, Ontology). And similarly to Figure 7.3, there is a drop in the effects of DM at interval 3, corresponding to the time when DM was split into DM, N and Constraints. Drops shown in Figure E.3 in other documents (Primer interval 3 and 5, Overview interval 2) also mirror those in Figure 7.3.

Effect sign Effect signs are universally positive in Figure E.6, except for the selection effects at interval 0 which are negative, as in Figure 7.3. Here, for most documents, the interval 0 selection effects are just as negative as those in Figure 7.3, except for the cases of DM, Links, and Sem, that have slightly less negative selection effects at interval 0 here than for the original corpus. This shows that, even if we consider more of the contents of the documents, we still generally do not find more similarities between the contents of the charter and the first draft of each document.

Patterns over time Similarly to Section 7.1, over time, the causal and selection effects tend to both get larger, in Figure E.6. In terms of bias, the two effects tend to increase in pace with each other, and their heights are not very different. Hence, if one were to only calculate the (biased) selection effect, one would not reach particularly different conclusions about the temporal evolution of the effects than if they had calculated the causal effect. So, if one is willing to sacrifice some accuracy for the ease of not adjusting for confounders, the conclusions reached would not be dramatically different. Moreover, within each document, there is not much variability in bias over time, as per the variance and standard deviation figures in Table E.16, where most documents' variance is at 0% or 1%, with the exception of Ontology (18%) and Constraints (3%), while standard deviations are not extremely large either, ranging from 0% to 43% (the latter corresponding to Ontology), with a median of standard deviations at 4%.

Patterns across documents Table E.16 shows how bias varies across documents: median bias ranges from only 1%, for Constraints, to a maximum of 20 times that, at 20%, for Overview. Except for Overview, for all other documents median bias is very low, and one might say negligible, at single-digit values. The median when the datapoints of all documents are taken together is low, at only 5%, with the median of medians also at 5%. So, on the aggregate, bias is certainty low, and might be considered negligible. In terms of the temporal patterns of effects, these are mostly consistent, with effects monotonically increasing in most documents, with some cases showing occasional drops (DM, Primer, Overview), as was also the case in Section 7.1. So conclusions about temporal patterns of effects are the same for both implementations designs.

As Table E.17 shows, the medians over all documents and the medians of medians are very close under both implementations (5% for the lager corpus, and 7% for the original corpus). For most documents, those values are the same or very close, within a few percentage points. However, for some documents, the median bias values can be different; this is particularly the case for Overview, where the median bias with the original corpus is only 7% but with the larger corpus it is much higher at 20%, while DC and XML have median bias values that are 6 percentage points lower with the original corpus. In the case of Overview, comparing Figures E.6(i) and Figure 7.3(i) shows that this large difference in bias is due to intervals 2 in particular, where there is very little discrepancy in causal and selection effect with the original corpus (Figure 7.3(i)) while this discrepancy is much bigger with the larger corpus (Figure E.6(i)). So, while aggregate statistics across documents are very similar across implementations, at the individual document level there are cases where bias levels are quite different.

Table E.18 compares the medians of all descriptives across both implementations. In the 'All documents' descriptives, the minima are identical (0%), and maxima are quite close (99% for the original corpus and 110% for the larger corpus - these both correspond to the maximum bias observed in the Ontology document, as per Tables E.8 and 7.3). The means are very close (11% and 10%, noting that means are sensitive to extreme values), and medians are only two percentage points apart (at 7% and 5%). Variances differ by only one percentage point and are both quite low (2% and 3%), and similarly standard deviations are both quite low and very close, differing by only 2 percentage points (16% and 18%). In the medians of descriptives, the minima are only two points apart (4% and 2%), while the maxima are five points apart (17% and 12%). The medians of means, and importantly, of medians are very close across both implementations (only one and two percentage points apart, respectively), while median variances are identical at 0% (as mentioned, numbers are rounded to the nearest integer), and median standard deviations are also low and very close (5% and 4%). So, bias levels are comparable and very similar overall.

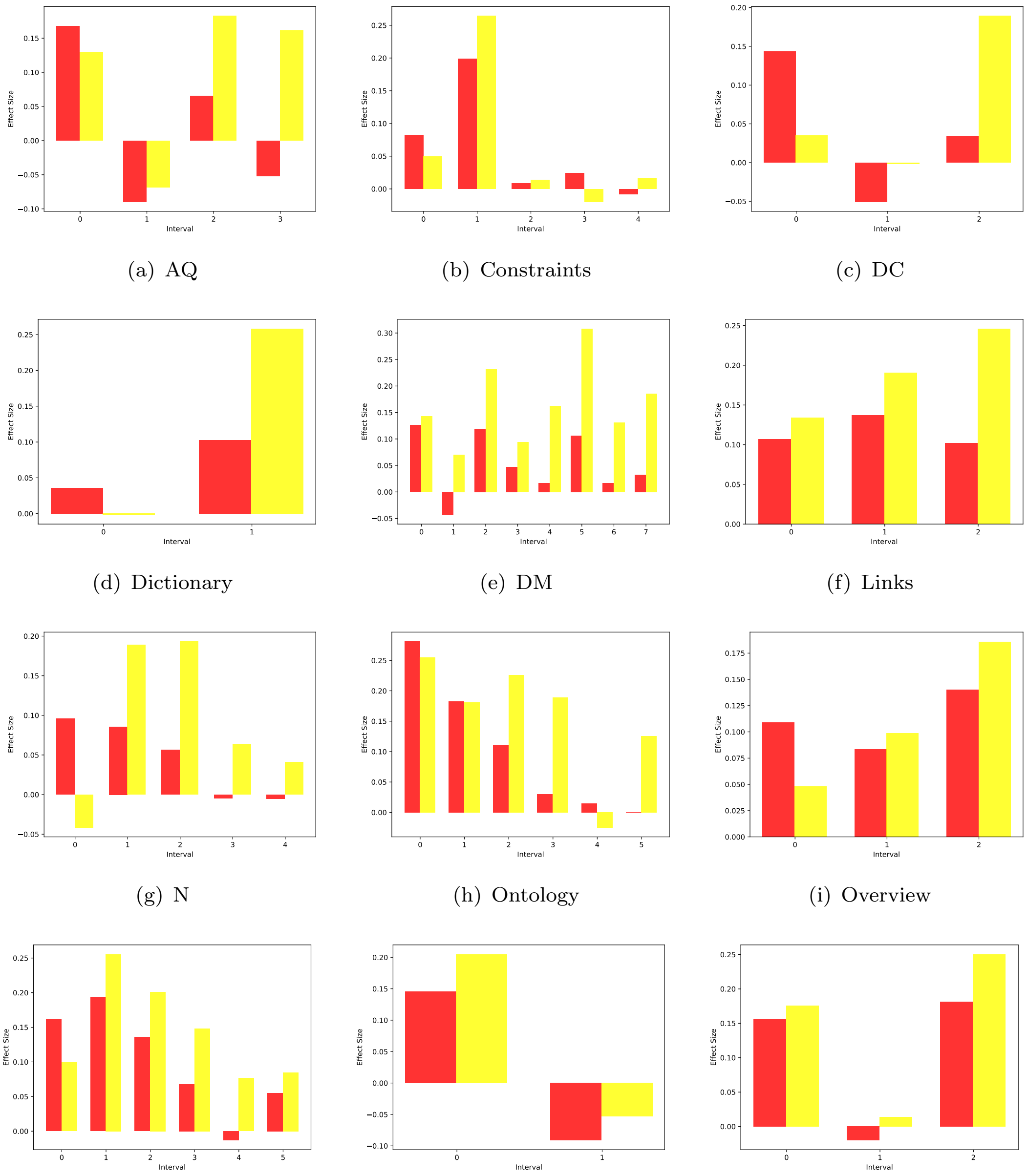
In summary, although there are some small individual-level differences in bias levels for some documents, for both implementation designs one may conclude that the bias levels are generally low across documents, and do not particularly affect the conclusions reached about the properties and patterns of causal versus selection effects.

Therefore one might consider the overall level of bias negligible. Hence, for both corpus sizes, one reaches the same conclusions in terms of bias and effect properties. So, overall, it can be said the the effects of the previous draft on the next, as well as the level of confounding bias present in them, are generally robust to whether the original or the larger corpus is used, and the same conclusions are reached in terms of their patterns.

E.2.3 Summary

Overall, the results obtained under a larger corpus presented here are overall the same

In conclusion, the results obtained under a larger corpus presented here lead to the same overall conclusions as the main implementation whose results were presented in Section 7.1. The amount of confounding bias present, for the effects of each variable on the outcome, is very similar to the bias present in Section 7.1; interestingly, bias here is slightly lower, for all variables, than in the original corpus. In addition, the characteristics and patterns (over time and across contexts) of the effects of the emails on outcomes, and of the effects of previous outcomes on next outcomes, are overall the same. Hence, the findings of Section 7.1 are robust to the implementation variations presented here.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

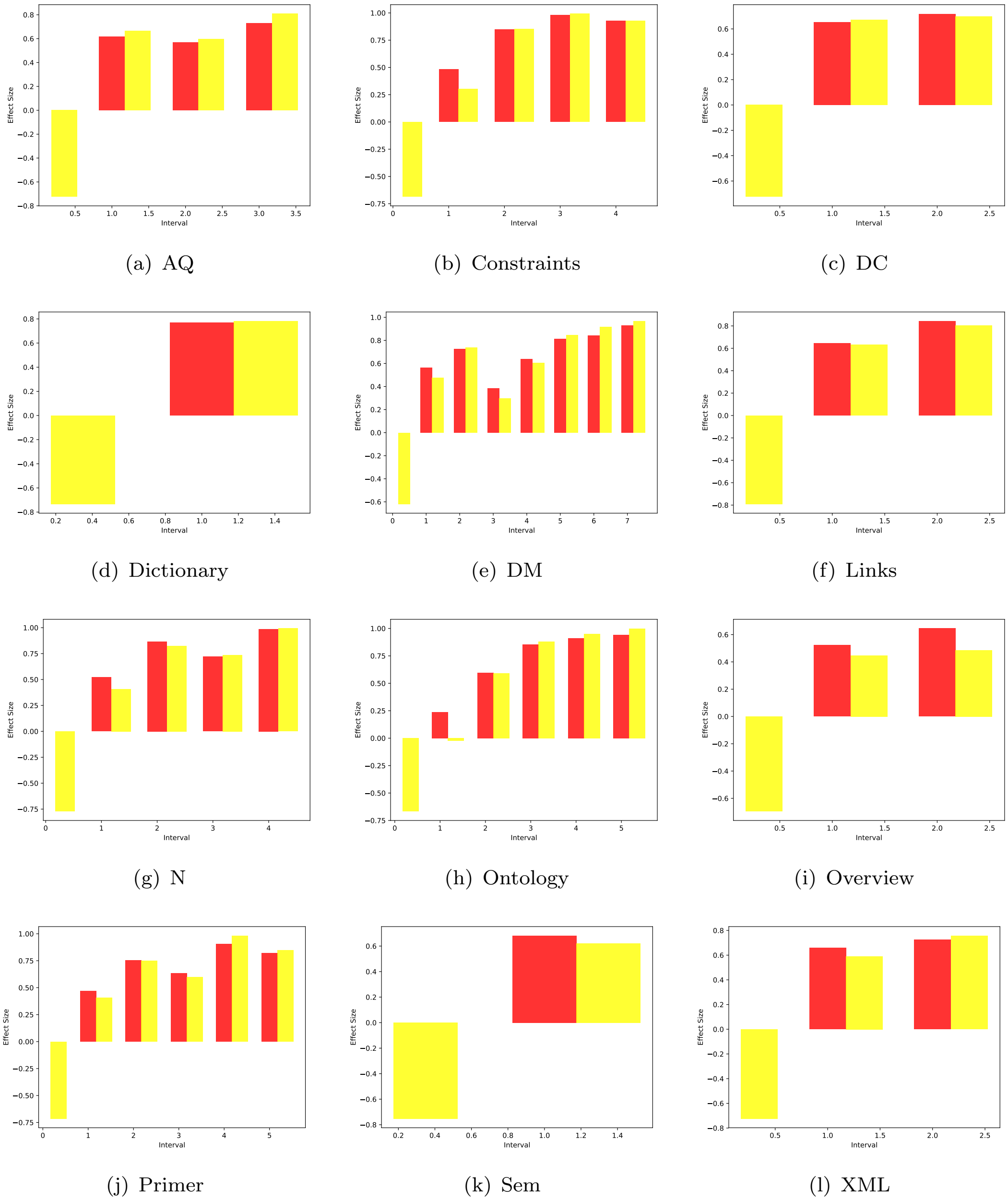
(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.5: Effects of Sentiment in Online Communications on Current Draft, Using the Larger Corpus: Similarly to Figure 7.2(for Sentiment using the original corpus) and Figure E.4 (for Participation using the larger corpus), the causal effect (red) is generally not very close to the selection effect (yellow), meaning that confounding bias is often large; causal and selection effects are not always positive, they are sometimes negative; effects are generally relative small, with causal effects never exceeding the 25% mark; causal effects tend to get smaller over time, however selection effects often remain large until later intervals; the magnitude of confounding bias and the temporal patterns show some variation across documents.



(a) AQ

(b) Constraints

(c) DC

(d) Dictionary

(e) DM

(f) Links

(g) N

(h) Ontology

(i) Overview

(j) Primer

(k) Sem

(l) XML

Figure E.6: Effects of Previous Draft on Current Draft, Using the Larger Corpus: Similarly to Figure 7.3 (for the effects of the previous draft using the original corpus), the causal effect (red) is generally not exactly equal, but still relatively close, to the selection effect (red), across documents, meaning that the presence of some confounding bias can be observed, but that bias is not as large as in Figures E.4 or E.5; the causal effect is often, but not always, smaller than the selection effect; both causal and selection effects are always positive, except at Interval 0 (which will be discussed), for all documents; effects are generally large, compared to Figures E.4 or E.5, with values that can exceed 60% or even 80%, particularly in later intervals; causal and selection effects generally grow, in sync, over time (with the exception of some temporary drops in some cases); across documents, the magnitude of confounding bias can vary, but temporal patterns are often very similar.

E.3 Summary

In conclusion, both alternative implementation choices presented here lead to the same overall conclusions as the main implementation whose results were presented in Section 7.1: the amount of confounding bias present when estimating the effects of email conversations on outcomes is not negligible, but is rather quite large, while the characteristics and patterns (over time and across contexts) of the effects of the emails on outcomes, and of the effects of previous outcomes on next outcomes, are overall the same. Hence, the findings of Section 7.1 are robust to the implementation variations presented here.

Appendix F

Causal model evaluation for all intervals of all PROV documents

As a supplement to Chapter 7.2, which presented evaluation results only for some documents and some intervals, this appendix presents the full results, for all three tests. It is structured as follows: for each document (ordered alphabetically), for each interval of that document, results are presented for Test 1, then Test 2, and finally Test 3. As per Chapter 7.2, for all documents and intervals, Tests 1, 2, and 3 do not hold, meaning that the contagion-based paradigm's causal model violated the dependencies and independencies in the empirical data and should be rejected, in favour of the causal model proposed in this thesis.

First, all tests are presented, for all intervals, for all values of all variables, for AQ. This takes up three pages, so if this done for all 12 documents, it would take up 36 pages. Recalling that, to determine that a test is not obeyed (equality violated), it is sufficient for one combination of values to violate the test (equality between the two expressions), for the remaining documents, we only present some (not all) combinations of values, which violate the equalities, for economy of space. So, for the remaining documents, we only show some of the tests, for cases where the outcome variable has the value 1 (the results for the corresponding values of the independent variables when the dependent, i.e. the outcome, is 0, can be inferred from these, as they are complementary - they sum to 1).

However, even if we shorten results tables in this way, each document would still take up two (or three) pages. Therefore, in addition to the AQ document, for economy of space we present here only results for the four core Recommendations documents: Constraints, DM, N, Ontology.

Test results for the remaining documents, can be found at the URL https://www.dropbox.com/sh/xohqt9n84iviwm3/AAD2leCctQtiIXK2rpbyupsca?dl=0, which contains the complete tests for all value combinations of all documents and all intervals.

F.1 AQ

For the AO document, the results of Tests 1, 2, and 3 are presented in Tables F.1, F.2, and F.3, respectively.

Table F.1: AQ Test 1

interval 0	interval 1	interval 1	interval 2
2010-12-14-2012-01-10	2012-01-10-2012-06-19	2012-06-19-2013-03-12	2013-03-12-2013-04-30
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (1 1) 0.6512	(s,f) = (1 1) 0.4688	(s,f) = (1 1) 0.6279	(s,f) = (1 1) 0.6552
$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$
(i,f,s) = (1 1 1) 0.4828	(i,f,s) = (1 1 1) 0.8333	(i,f,s) = (1 1 1) 0.8462	(i,f,s) = (1 1 1) 1.0
(i,f,s) = (0 1 1) 1.0	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0.2941	(i,f,s) = (0 1 1) 0.2308
. . .	. . .	. . .	. . .
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (0 1) 0.5	(s,f) = (0 1) 0.4938	(s,f) = (0 1) 0.3375	(s,f) = (0 1) 0.3402
$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$
(i,f,s) = (1 1 0) 0.1944	(i,f,s) = (1 1 0) 0.6923	(i,f,s) = (1 1 0) 0.6207	(i,f,s) = (1 1 0) 0.8421
(i,f,s) = (0 1 0) 1.0	(i,f,s) = (0 1 0) 0.3095	(i,f,s) = (0 1 0) 0.1765	(i,f,s) = (0 1 0) 0.0169
. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (1 1) 0.617	(p,f) = (1 1) 0.5	(p,f) = (1 1) 0.5769	(p,f) = (1 1) 0.6552
P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)
(i,f,p) = (1 1 1) 0.4545	(i,f,p) = (1 1 1) 0.85	(i,f,p) = (1 1 1) 0.8333	(i,f,p) = (1 1 1) 1.0
(i,f,p) = (0 1 1) 1.0	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0.2273	(i,f,p) = (0 1 1) 0.2308
. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (0 1) 0.5185	(p,f) = (0 1) 0.481	(p,f) = (0 1) 0.338	(p,f) = (0 1) 0.3402
P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)
(i,f,p) = (1 1 0) 0.1875	(i,f,p) = (1 1 0) 0.6757	(i,f,p) = (1 1 0) 0.6	(i,f,p) = (1 1 0) 0.8421
(i,f,p) = (0 1 0) 1.0	(i,f,p) = (0 1 0) 0.3095	(i,f,p) = (0 1 0) 0.1957	(i,f,p) = (0 1 0) 0.0169
. . .	. . .	. . .	. . .
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (1 0) 0.3488	(s,f) = (1 0) 0.5312	(s,f) = (1 0) 0.3721	(s,f) = (1 0) 0.3448
$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$
(i,f,s) = (1 0 1) 0.5172	(i,f,s) = (1 0 1) 0.1667	(i,f,s) = (1 0 1) 0.1538	(i,f,s) = (1 0 1) 0
(i,f,s) = (0 0 1) 0	(i,f,s) = (0 0 1) 1.0	(i,f,s) = (0 0 1) 0.7059	(i,f,s) = (0 0 1) 0.7692
. . .	. . .	. . .	. . .
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (0 0) 0.5	(s,f) = (0 0) 0.5062	(s,f) = (0 0) 0.6625	(s,f) = (0 0) 0.6598
$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$	$P(F = f \mid S = s, Initial = i)$
(i,f,s) = (1 0 0) 0.8056	(i,f,s) = (1 0 0) 0.3077	(i,f,s) = (1 0 0) 0.3793	(i,f,s) = (1 0 0) 0.1579
(i,f,s) = (0 0 0) 0	(i,f,s) = (0 0 0) 0.6905	(i,f,s) = (0 0 0) 0.8235	(i,f,s) = (0 0 0) 0.9831
. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (1 0) 0.383	(p,f) = (1 0) 0.5	(p,f) = (1 0) 0.4231	(p,f) = (1 0) 0.3448
P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)
(i,f,p) = (1 0 1) 0.5455	(i,f,p) = (1 0 1) 0.15	(i,f,p) = (1 0 1) 0.1667	(i,f,p) = (1 0 1) 0
(i,f,p) = (0 0 1) 0	(i,f,p) = (0 0 1) 1.0	(i,f,p) = (0 0 1) 0.7727	(i,f,p) = (0 0 1) 0.7692
. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (0 0) 0.4815	(p,f) = (0 0) 0.519	(p,f) = (0 0) 0.662	(p,f) = (0 0) 0.6598
P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)	P(F=f — P=p, Initial=i)
(i,f,p) = (1 0 0) 0.8125	(i,f,p) = (1 0 0) 0.3243	(i,f,p) = (1 0 0) 0.4	(i,f,p) = (1 0 0) 0.1579
(i,f,p) = (0 0 0) 0	(i,f,p) = (0 0 0) 0.6905	(i,f,p) = (0 0 0) 0.8043	(i,f,p) = (0 0 0) 0.9831

Table F.2: AQ Test 2

interval 0	interval 1	interval 2	interval 3
2010-12-14-2012-01-10	2012-01-10-2012-06-19	2012-06-19-2013-03-12	2013-03-12-2013-04-30
$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$
(i,s) = (1 1) 0.4462	(i,s) = (1 1) 0.3158	(i,s) = (1 1) 0.4727	(i,s) = (1 1) 0.2963
(i,s) = (0 1) 0.3889	(i,s) = (0 1) 0.25	(i,s) = (0 1) 0.25	(i,s) = (0 1) 0.1806
(i,s) = (1 0) 0.5538	(i,s) = (1 0) 0.6842	(i,s) = (1 0) 0.5273	(i,s) = (1 0) 0.7037
(i,s) = (0 0) 0.6111	(i,s) = (0 0) 0.75	(i,s) = (0 0) 0.75	(i,s) = (0 0) 0.8194
P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.5743	s = 0 : 0.7168	s = 0 : 0.6504	s = 0 : 0.7698
s = 1 : 0.4257	s = 1 : 0.2832	s = 1 : 0.3496	s = 1 : 0.2302
- - -	- - -	- - -	- - -
$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$
(i,p) = (1 1) 0.5077	(i,p) = (1 1) 0.3509	(i,p) = (1 1) 0.5455	(i,p) = (1 1) 0.2963
(i,p) = (0 1) 0.3889	(i,p) = (0 1) 0.25	(i,p) = (0 1) 0.3235	(i,p) = (0 1) 0.1806
(i,p) = (1 0) 0.4923	(i,p) = (1 0) 0.6491	(i,p) = (1 0) 0.4545	(i,p) = (1 0) 0.7037
(i,p) = (0 0) 0.6111	(i,p) = (0 0) 0.75	(i,p) = (0 0) 0.6765	(i,p) = (0 0) 0.8194
P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.5347	p = 0 : 0.6991	p = 0 : 0.5772	p = 0 : 0.7698
p = 1 : 0.4653	p = 1 : 0.3009	p = 1 : 0.4228	p = 1 : 0.2302

Table F.3: AQ Test 3

interval 1	interval 2	interval 3
2012-01-10-2012-06-19	2012-06-19-2013-03-12	2013-03-12-2013-04-30
$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$
(s, s_{{-}1}) = (1 1) 0.5581	(s, s_{{-}1}) = (1 1) 0.6562	(s, s_{{-}1}) = (1 1) 0.4651
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (1 1 1) 0.6552	(s,s_{{-}1},i) = (1 1 1) 0.8333	(s,s_{{-}1},i) = (1 1 1) 0.5
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (1 1 0) 0.3571	(s,s_{{-}1},i) = (1 1 0) 0.4286	(s,s_{{-}1},i) = (1 1 0) 0.4118
. . .	. . .	. . .
$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$
(s, s_{{-}1}) = (1 0) 0.1143	(s, s_{{-}1}) = (1 0) 0.2418	(s, s_{{-}1}) = (1 0) 0.1084
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (1 0 1) 0.1944	(s,s_{{-}1},i) = (1 0 1) 0.2564	(s,s_{{-}1},i) = (1 0 1) 0.0345
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (1 0 0) 0.0294	(s,s_{{-}1},i) = (1 0 0) 0.2308	(s,s_{{-}1},i) = (1 0 0) 0.1481
. . .	. . .	. . .
$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$
(s, s_{{-}1}) = (0 1) 0.4419	(s, s_{{-}1}) = (0 1) 0.3437	(s, s_{{-}1}) = (0 1) 0.5349
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (0 1 1) 0.3448	(s,s_{{-}1},i) = (0 1 1) 0.1667	(s,s_{{-}1},i) = (0 1 1) 0.5
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (0 1 0) 0.6429	(s,s_{{-}1},i) = (0 1 0) 0.5714	(s,s_{{-}1},i) = (0 1 0) 0.5882
. . .	. . .	. . .
$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$	$P(S = s \mid S_{-1} = s_{-1})$
(s, s_{{-}1}) = (0 0) 0.8857	(s, s_{{-}1}) = (0 0) 0.7582	(s, s_{{-}1}) = (0 0) 0.8916
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (0 0 1) 0.8056	(s,s_{{-}1},i) = (0 0 1) 0.7436	(s,s_{{-}1},i) = (0 0 1) 0.9655
$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$	$P(S = s \mid S_{-1} = s_{-1}, I = i)$
(s,s_{{-}1},i) = (0 0 0) 0.9706	(s,s_{{-}1},i) = (0 0 0) 0.7692	(s,s_{{-}1},i) = (0 0 0) 0.8519
. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$
(p, p_{{-}1}) = (1 1) 0.5745	(p, p_{{-}1}) = (1 1) 0.6471	(p, p_{{-}1}) = (1 1) 0.4231
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (1 1 1) 0.6364	(p, p_{{-}1}, i) = (1 1 1) 0.8	(p, p_{{-}1}, i) = (1 1 1) 0.4333
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (1 1 0) 0.4286	(p, p_{{-}1}, i) = (1 1 0) 0.4286	(p, p_{{-}1}, i) = (1 1 0) 0.4091
. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$
(p, p_{{-}1}) = (1 0) 0.1061	(p, p_{{-}1}) = (1 0) 0.3371	(p, p_{{-}1}) = (1 0) 0.0946
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (1 0 1) 0.1875	(p, p_{{-}1}, i) = (1 0 1) 0.4054	(p, p_{{-}1}, i) = (1 0 1) 0.04
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (1 0 0) 0.0294	(p, p_{{-}1}, i) = (1 0 0) 0.2885	(p, p_{{-}1}, i) = (1 0 0) 0.1224
. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$
(p, p_{{-}1}) = (0 1) 0.4255	(p, p_{{-}1}) = (0 1) 0.3529	(p, p_{{-}1}) = (0 1) 0.5769
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (0 1 1) 0.3636	(p, p_{{-}1}, i) = (0 1 1) 0.2	(p, p_{{-}1}, i) = (0 1 1) 0.5667
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (0 1 0) 0.5714	(p, p_{{-}1}, i) = (0 1 0) 0.5714	(p, p_{{-}1}, i) = (0 1 0) 0.5909
. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$	$P(P = p \mid P_{-1} = p_{-1})$
(p, p_{{-}1}) = (0 0) 0.8939	(p, p_{{-}1}) = (0 0) 0.6629	(p, p_{{-}1}) = (0 0) 0.9054
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (0 0 1) 0.8125	(p, p_{{-}1}, i) = (0 0 1) 0.5946	(p, p_{{-}1}, i) = (0 0 1) 0.96
$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$	$P(P = p \mid P_{-1} = p_{-1}, I = i)$
(p, p_{{-}1}, i) = (0 0 0) 0.9706	(p, p_{{-}1}, i) = (0 0 0) 0.7115	(p, p_{{-}1}, i) = (0 0 0) 0.8776

F.2 Constraints

For the Constraints document, the results of Tests 1, 2, and 3 are presented in Tables F.4, F.5, and F.6, respectively.

Table F.4: Constraints Test 1

interval 0	interval 1	interval 2	interval 3	interval 4
2010-12-14-2012-05-03	2012-05-03-2012-09-11	2012-09-11-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (1 1) 0.551	(s,f) = (1 1) 0.5714	(s,f) = (1 1) 0.4444	(s,f) = (1 1) 0.3571	(s,f) = (1 1) 0.5
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 1) 0.3333	(i,f,s) = (1 1 1) 0.6923	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 1.0
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	P(F=f — S=s, Initial=i)	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 1) 1.0	(i,f,s) = (0 1 1) 0.2941	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0
. . .	. . .	. . .	. . .	. . .
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (0 1) 0.5517	(s,f) = (0 1) 0.2581	(s,f) = (0 1) 0.4167	(s,f) = (0 1) 0.413	(s,f) = (0 1) 0.4118
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	P(F=f — S=s, Initial=i)	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 0) 0.1875	(i,f,s) = (1 1 0) 0.45	(i,f,s) = (1 1 0) 0.9688	(i,f,s) = (1 1 0) 0.9268	(i,f,s) = (1 1 0) 1.0
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 0) 1.0	(i,f,s) = (0 1 0) 0.1667	(i,f,s) = (0 1 0) 0.0769	(i,f,s) = (0 1 0) 0	(i,f,s) = (0 1 0) 0.0476
. . .	. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (1 1) 0.5714	(p,f) = (1 1) 0.5714	(p,f) = (1 1) 0.4444	(p,f) = (1 1) 0.3571	(p,f) = (1 1) 0.5
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 1) 0.3438	(i,f,p) = (1 1 1) 0.6923	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 1.0
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 1) 1.0	(i,f,p) = (0 1 1) 0.2941	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0
. . .	. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (0 1) 0.5345	(p,f) = (0 1) 0.2581	(p,f) = (0 1) 0.4167	(p,f) = (0 1) 0.413	(p,f) = (0 1) 0.4118
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 0) 0.1818	(i,f,p) = (1 1 0) 0.45	(i,f,p) = (1 1 0) 0.9688	(i,f,p) = (1 1 0) 0.9268	(i,f,p) = (1 1 0) 1.0
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 0) 1.0	(i,f,p) = (0 1 0) 0.1667	(i,f,p) = (0 1 0) 0.0769	(i,f,p) = (0 1 0) 0	(i,f,p) = (0 1 0) 0.0476

Table F.5: Constraints Test 2

interval 0	interval 1	interval 2	interval 3	interval 4
2010-12-14-2012-05-03	2012-05-03-2012-09-11	2012-09-11-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$
(i,s) = (1 1) 0.5077	(i,s) = (1 1) 0.661	(i,s) = (1 1) 0.3333	(i,s) = (1 1) 0.1961	(i,s) = (1 1) 0.1875
(i,s) = (0 1) 0.381	(i,s) = (0 1) 0.2881	(i,s) = (0 1) 0.2778	(i,s) = (0 1) 0.2609	(i,s) = (0 1) 0.125
(i,s) = (1 0) 0.4923	(i,s) = (1 0) 0.339	(i,s) = (1 0) 0.6667	(i,s) = (1 0) 0.8039	(i,s) = (1 0) 0.8125
(i,s) = (0 0) 0.619	(i,s) = (0 0) 0.7119	(i,s) = (0 0) 0.7222	(i,s) = (0 0) 0.7391	(i,s) = (0 0) 0.875
P(S=s):	P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.5421	s = 0 : 0.5254	s = 0 : 0.7	s = 0 : 0.7667	s = 0 : 0.85
s = 1 : 0.4579	s = 1 : 0.4746	s = 1 : 0.3	s = 1 : 0.2333	s = 1 : 0.15
- - -	- - -	- - -	- - -	- - -
$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$
(i,p) = (1 1) 0.4923	(i,p) = (1 1) 0.661	(i,p) = (1 1) 0.3333	(i,p) = (1 1) 0.1961	(i,p) = (1 1) 0.1875
(i,p) = (0 1) 0.4048	(i,p) = (0 1) 0.2881	(i,p) = (0 1) 0.2778	(i,p) = (0 1) 0.2609	(i,p) = (0 1) 0.125
(i,p) = (1 0) 0.5077	(i,p) = (1 0) 0.339	(i,p) = (1 0) 0.6667	(i,p) = (1 0) 0.8039	(i,p) = (1 0) 0.8125
(i,p) = (0 0) 0.5952	(i,p) = (0 0) 0.7119	(i,p) = (0 0) 0.7222	(i,p) = (0 0) 0.7391	(i,p) = (0 0) 0.875
P(P=p):	P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.5421	p = 0 : 0.5254	p = 0 : 0.7	p = 0 : 0.7667	p = 0 : 0.85
p = 1 : 0.4579	p = 1 : 0.4746	p = 1 : 0.3	p = 1 : 0.2333	p = 1 : 0.15

Table F.6: Constraints Test 3

interval 1	interval 2	interval 3	interval 4
2012-05-03-2012-09-11	2012-09-11-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$
(s, s_prev) = (1 1) 0.7143	(s, s_prev) = (1 1) 0.4643	(s, s_prev) = (1 1) 0.5556	(s, s_prev) = (1 1) 0.5357
$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 1 1) 0.6364	(s,s_prev,i) = (1 1 1) 0.3846	(s,s_prev,i) = (1 1 1) 0.375	(s,s_prev,i) = (1 1 1) 0.7
$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 1 0) 0.875	(s,s_prev,i) = (1 1 0) 0.6471	(s,s_prev,i) = (1 1 0) 0.7	(s,s_prev,i) = (1 1 0) 0.4444
. . .	. . .	. . .	. . .
$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$	$P(S = s\|s_{-1} = s_{-1})$
(s, s_prev) = (1 0) 0.3043	(s, s_prev) = (1 0) 0.1562	(s, s_prev) = (1 0) 0.0952	(s, s_prev) = (1 0) 0.0326
$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 0 1) 0.125	(s,s_prev,i) = (1 0 1) 0.15	(s,s_prev,i) = (1 0 1) 0.0938	(s,s_prev,i) = (1 0 1) 0.0488
$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$	$P(S = s\|s_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 0 0) 0.4595	(s,s_prev,i) = (1 0 0) 0.1591	(s,s_prev,i) = (1 0 0) 0.0962	(s,s_prev,i) = (1 0 0) 0.0196
. . .	. . .	. . .	. . .
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$
(p, p_prev) = (1 1) 0.7347	(p, p_prev) = (1 1) 0.4643	(p, p_prev) = (1 1) 0.5556	(p, p_prev) = (1 1) 0.5357
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 1 1) 0.6562	(p, p_prev, i) = (1 1 1) 0.3846	(p, p_prev, i) = (1 1 1) 0.375	(p, p_prev, i) = (1 1 1) 0.7
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 1 0) 0.8824	(p, p_prev, i) = (1 1 0) 0.6471	(p, p_prev, i) = (1 1 0) 0.7	(p, p_prev, i) = (1 1 0) 0.4444
. . .	. . .	. . .	. . .
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$
(p, p_prev) = (1 0) 0.2899	(p, p_prev) = (1 0) 0.1562	(p, p_prev) = (1 0) 0.0952	(p, p_prev) = (1 0) 0.0326
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 0 1) 0.1212	(p, p_prev, i) = (1 0 1) 0.15	(p, p_prev, i) = (1 0 1) 0.0938	(p, p_prev, i) = (1 0 1) 0.0488
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 0 0) 0.4444	(p, p_prev, i) = (1 0 0) 0.1591	(p, p_prev, i) = (1 0 0) 0.0962	(p, p_prev, i) = (1 0 0) 0.0196

F.3 DM

For the DM document, the results of Tests 1, 2, and 3 are presented in Tables F.7, F.8, and F.9, respectively.

Table F.7: DM Test 1

interval 0	interval 1	interval 2	interval 3
2010-12-14-2011-10-18	2011-10-18-2011-12-15	2011-12-15-2012-02-02	2012-02-02-2012-05-03
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (1 1) 0.6078	(s,f) = (1 1) 0.5385	(s,f) = (1 1) 0.6591	(s,f) = (1 1) 0.6364
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 1) 0.4118	(i,f,s) = (1 1 1) 0.875	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 0.875
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 1) 1.0	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0.1667	(i,f,s) = (0 1 1) 0
. . .	. . .	. . .	. . .
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (0 1) 0.5357	(s,f) = (0 1) 0.4944	(s,f) = (0 1) 0.4026	(s,f) = (0 1) 0.453
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 0) 0.1613	(i,f,s) = (1 1 0) 0.7556	(i,f,s) = (1 1 0) 0.8438	(i,f,s) = (1 1 0) 0.6136
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 0) 1.0	(i,f,s) = (0 1 0) 0.2273	(i,f,s) = (0 1 0) 0.0889	(i,f,s) = (0 1 0) 0.3562
. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (1 1) 0.5882	(p,f) = (1 1) 0.5862	(p,f) = (1 1) 0.6591	(p,f) = (1 1) 0.6
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 1) 0.4	(i,f,p) = (1 1 1) 0.8421	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 0.7895
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 1) 1.0	(i,f,p) = (0 1 1) 0.1	(i,f,p) = (0 1 1) 0.1667	(i,f,p) = (0 1 1) 0
. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (0 1) 0.5536	(p,f) = (0 1) 0.4767	(p,f) = (0 1) 0.4026	(p,f) = (0 1) 0.4561
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 0) 0.1667	(i,f,p) = (1 1 0) 0.7619	(i,f,p) = (1 1 0) 0.8438	(i,f,p) = (1 1 0) 0.6341
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 0) 1.0	(i,f,p) = (0 1 0) 0.2045	(i,f,p) = (0 1 0) 0.0889	(i,f,p) = (0 1 0) 0.3562
interval 4	interval 5	interval 6	interval 7
2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (1 1) 0.7059	(s,f) = (1 1) 0.6087	(s,f) = (1 1) 0.4565	(s,f) = (1 1) 0.45
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 1) 0.9091	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 0.9545	(i,f,s) = (1 1 1) 1.0
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 1) 0.3333	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0
. . .	. . .	. . .	. . .
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (0 1) 0.4265	(s,f) = (0 1) 0.4104	(s,f) = (0 1) 0.3874	(s,f) = (0 1) 0.3932
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (1 1 0) 0.75	(i,f,s) = (1 1 0) 0.9107	(i,f,s) = (1 1 0) 0.9149	(i,f,s) = (1 1 0) 1.0
$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$	$P(F = f\|S = s, Initial = i)$
(i,f,s) = (0 1 0) 0.2	(i,f,s) = (0 1 0) 0.0513	(i,f,s) = (0 1 0) 0	(i,f,s) = (0 1 0) 0
. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (1 1) 0.7143	(p,f) = (1 1) 0.6286	(p,f) = (1 1) 0.4565	(p,f) = (1 1) 0.45
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 1) 0.9286	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 0.9545	(i,f,p) = (1 1 1) 1.0
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 1) 0.2857	(i,f,p) = (0 1 1) 0.0714	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0
. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (0 1) 0.4167	(p,f) = (0 1) 0.3852	(p,f) = (0 1) 0.3874	(p,f) = (0 1) 0.3932
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (1 1 0) 0.7358	(i,f,p) = (1 1 0) 0.898	(i,f,p) = (1 1 0) 0.9149	(i,f,p) = (1 1 0) 1.0
$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$	$P(F = f\|P = p, Initial = i)$
(i,f,p) = (0 1 0) 0.2025	(i,f,p) = (0 1 0) 0.0411	(i,f,p) = (0 1 0) 0	(i,f,p) = (0 1 0) 0

Table F.8: DM Test 2

interval 0	interval 1	interval 2	interval 3
2010-12-14-2011-10-18	2011-10-18-2011-12-15	2011-12-15-2012-02-02	2012-02-02-2012-05-03
$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$
(i,s) = (1 1) 0.5231	(i,s) = (1 1) 0.2623	(i,s) = (1 1) 0.4483	(i,s) = (1 1) 0.2667
(i,s) = (0 1) 0.4048	(i,s) = (0 1) 0.1852	(i,s) = (0 1) 0.2857	(i,s) = (0 1) 0.0759
(i,s) = (1 0) 0.4769	(i,s) = (1 0) 0.7377	(i,s) = (1 0) 0.5517	(i,s) = (1 0) 0.7333
(i,s) = (0 0) 0.5952	(i,s) = (0 0) 0.8148	(i,s) = (0 0) 0.7143	(i,s) = (0 0) 0.9241
P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.5234	s = 0 : 0.7739	s = 0 : 0.6364	s = 0 : 0.8417
s = 1 : 0.4766	s = 1 : 0.2261	s = 1 : 0.3636	s = 1 : 0.1583
- - -	- - -	- - -	- - -
$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$
(i,p) = (1 1) 0.5385	(i,p) = (1 1) 0.3115	(i,p) = (1 1) 0.4483	(i,p) = (1 1) 0.3167
(i,p) = (0 1) 0.381	(i,p) = (0 1) 0.1852	(i,p) = (0 1) 0.2857	(i,p) = (0 1) 0.0759
(i,p) = (1 0) 0.4615	(i,p) = (1 0) 0.6885	(i,p) = (1 0) 0.5517	(i,p) = (1 0) 0.6833
(i,p) = (0 0) 0.619	(i,p) = (0 0) 0.8148	(i,p) = (0 0) 0.7143	(i,p) = (0 0) 0.9241
P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.5234	p = 0 : 0.7478	p = 0 : 0.6364	p = 0 : 0.8201
p = 1 : 0.4766	p = 1 : 0.2522	p = 1 : 0.3636	p = 1 : 0.1799
interval 4	interval 5	interval 6	interval 7
2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$
(i,s) = (1 1) 0.1642	(i,s) = (1 1) 0.2	(i,s) = (1 1) 0.3188	(i,s) = (1 1) 0.2812
(i,s) = (0 1) 0.0698	(i,s) = (0 1) 0.1034	(i,s) = (0 1) 0.2727	(i,s) = (0 1) 0.2366
(i,s) = (1 0) 0.8358	(i,s) = (1 0) 0.8	(i,s) = (1 0) 0.6812	(i,s) = (1 0) 0.7187
(i,s) = (0 0) 0.9302	(i,s) = (0 0) 0.8966	(i,s) = (0 0) 0.7273	(i,s) = (0 0) 0.7634
P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.8889	s = 0 : 0.8535	s = 0 : 0.707	s = 0 : 0.7452
s = 1 : 0.1111	s = 1 : 0.1465	s = 1 : 0.293	s = 1 : 0.2548
- - -	- - -	- - -	- - -
$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$
(i,p) = (1 1) 0.209	(i,p) = (1 1) 0.3	(i,p) = (1 1) 0.3188	(i,p) = (1 1) 0.2812
(i,p) = (0 1) 0.0814	(i,p) = (0 1) 0.1609	(i,p) = (0 1) 0.2727	(i,p) = (0 1) 0.2366
(i,p) = (1 0) 0.791	(i,p) = (1 0) 0.7	(i,p) = (1 0) 0.6812	(i,p) = (1 0) 0.7187
(i,p) = (0 0) 0.9186	(i,p) = (0 0) 0.8391	(i,p) = (0 0) 0.7273	(i,p) = (0 0) 0.7634
P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.8627	p = 0 : 0.7771	p = 0 : 0.707	p = 0 : 0.7452
p = 1 : 0.1373	p = 1 : 0.2229	p = 1 : 0.293	p = 1 : 0.2548

Table F.9: DM Test 3

interval 1	interval 2	interval 3	interval 4
2011-10-18-2011-12-15	2011-12-15-2012-02-02	2012-02-02-2012-05-03	2012-05-03-2012-07-24
$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$
(s, s_prev) = (1 1) 0.3922	(s, s_prev) = (1 1) 0.5862	(s, s_prev) = (1 1) 0.2955	(s, s_prev) = (1 1) 0.4091
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 1 1) 0.4194	(s,s_prev,i) = (1 1 1) 0.6562	(s,s_prev,i) = (1 1 1) 0.3793	(s,s_prev,i) = (1 1 1) 0.5
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 1 0) 0.35	(s,s_prev,i) = (1 1 0) 0.5	(s,s_prev,i) = (1 1 0) 0.1333	(s,s_prev,i) = (1 1 0) 0.25
. . .	. . .	. . .	. . .
$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$
(s, s_prev) = (1 0) 0.0938	(s, s_prev) = (1 0) 0.1587	(s, s_prev) = (1 0) 0.0947	(s, s_prev) = (1 0) 0.0611
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 0 1) 0.1	(s,s_prev,i) = (1 0 1) 0.1923	(s,s_prev,i) = (1 0 1) 0.1613	(s,s_prev,i) = (1 0 1) 0.0755
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$
(s,s_prev,i) = (1 0 0) 0.0882	(s,s_prev,i) = (1 0 0) 0.1351	(s,s_prev,i) = (1 0 0) 0.0625	(s,s_prev,i) = (1 0 0) 0.0513
. . .	. . .	. . .	. . .
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$
(p, p_prev) = (1 1) 0.4118	(p, p_prev) = (1 1) 0.5862	(p, p_prev) = (1 1) 0.2955	(p, p_prev) = (1 1) 0.44
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 1 1) 0.5	(p, p_prev, i) = (1 1 1) 0.6562	(p, p_prev, i) = (1 1 1) 0.3793	(p, p_prev, i) = (1 1 1) 0.5333
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 1 0) 0.2857	(p, p_prev, i) = (1 1 0) 0.5	(p, p_prev, i) = (1 1 0) 0.1333	(p, p_prev, i) = (1 1 0) 0.3
. . .	. . .	. . .	. . .
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$
(p, p_prev) = (1 0) 0.125	(p, p_prev) = (1 0) 0.1587	(p, p_prev) = (1 0) 0.1263	(p, p_prev) = (1 0) 0.0781
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 0 1) 0.129	(p, p_prev, i) = (1 0 1) 0.1923	(p, p_prev, i) = (1 0 1) 0.2581	(p, p_prev, i) = (1 0 1) 0.1154
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$
(p, p_prev, i) = (1 0 0) 0.1212	(p, p_prev, i) = (1 0 0) 0.1351	(p, p_prev, i) = (1 0 0) 0.0625	(p, p_prev, i) = (1 0 0) 0.0526
interval 5	interval 6	interval 7	
2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30	
$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	
(s, s_prev) = (1 1) 0.2468	(s, s_prev) = (1 1) 0.7391	(s, s_prev) = (1 1) 0.587	
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	
(s,s_prev,i) = (1 1 1) 0.2667	(s,s_prev,i) = (1 1 1) 0.6429	(s,s_prev,i) = (1 1 1) 0.619	
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	
(s,s_prev,i) = (1 1 0) 0.2187	(s,s_prev,i) = (1 1 0) 0.8889	(s,s_prev,i) = (1 1 0) 0.56	
. . .	. . .	. . .	
$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	$P(S = s\|S_{-1} = s_{-1})$	
(s, s_prev) = (1 0) 0.05	(s, s_prev) = (1 0) 0.2164	(s, s_prev) = (1 0) 0.1171	
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	
(s,s_prev,i) = (1 0 1) 0.08	(s,s_prev,i) = (1 0 1) 0.2364	(s,s_prev,i) = (1 0 1) 0.1163	
$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	$P(S = s\|S_{-1} = s_{-1}, I = i)$	
(s,s_prev,i) = (1 0 0) 0.0364	(s,s_prev,i) = (1 0 0) 0.2025	(s,s_prev,i) = (1 0 0) 0.1176	
. . .	. . .	. . .	
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	
(p, p_prev) = (1 1) 0.3766	(p, p_prev) = (1 1) 0.6286	(p, p_prev) = (1 1) 0.587	
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	
(p, p_prev, i) = (1 1 1) 0.4222	(p, p_prev, i) = (1 1 1) 0.5	(p, p_prev, i) = (1 1 1) 0.619	
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	
(p, p_prev, i) = (1 1 0) 0.3125	(p, p_prev, i) = (1 1 0) 0.8462	(p, p_prev, i) = (1 1 0) 0.56	
. . .	. . .	. . .	
$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	$P(P = p\|P_{-1} = p_{-1})$	
(p, p_prev) = (1 0) 0.075	(p, p_prev) = (1 0) 0.1967	(p, p_prev) = (1 0) 0.1171	
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	
(p, p_prev, i) = (1 0 1) 0.08	(p, p_prev, i) = (1 0 1) 0.234	(p, p_prev, i) = (1 0 1) 0.1163	
$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	$P(P = p\|P_{-1} = p_{-1}, I = i)$	
(p, p_prev, i) = (1 0 0) 0.0727	(p, p_prev, i) = (1 0 0) 0.1733	(p, p_prev, i) = (1 0 0) 0.1176	

F.4 N

For the N (Notation) document, the results of Tests 1, 2, and 3 are presented in Tables F.10, F.11, and F.12, respectively.

TABLE F.10: N Test 1

interval 0	interval 1	interval 2	interval 3	interval 4
2010-12-14-2012-05-03	2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (1 1) 0.4792	(s,f) = (1 1) 0.6364	(s,f) = (1 1) 0.6038	(s,f) = (1 1) 0.4062	(s,f) = (1 1) 0.6111
$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$
(i,f,s) = (1 1 1) 0.3243	(i,f,s) = (1 1 1) 0.7931	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 1.0
$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$
(i,f,s) = (0 1 1) 1.0	(i,f,s) = (0 1 1) 0.3333	(i,f,s) = (0 1 1) 0.125	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0
. . .	. . .	. . .	. . .	. . .
$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$	$P(F = f\|S = s)$
(s,f) = (0 1) 0.5	(s,f) = (0 1) 0.3188	(s,f) = (0 1) 0.3333	(s,f) = (0 1) 0.4762	(s,f) = (0 1) 0.4343
$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$
(i,f,s) = (1 1 0) 0.0714	(i,f,s) = (1 1 0) 0.6	(i,f,s) = (1 1 0) 0.9524	(i,f,s) = (1 1 0) 1.0	(i,f,s) = (1 1 0) 1.0
$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$	$P(F = f\|S = s, I = i)$
(i,f,s) = (0 1 0) 1.0	(i,f,s) = (0 1 0) 0.2041	(i,f,s) = (0 1 0) 0.0238	(i,f,s) = (0 1 0) 0	(i,f,s) = (0 1 0) 0.0175
. . .	. . .	. . .	. . .	. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (1 1) 0.5217	(p,f) = (1 1) 0.6364	(p,f) = (1 1) 0.6038	(p,f) = (1 1) 0.4062	(p,f) = (1 1) 0.6111
$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$
(i,f,p) = (1 1 1) 0.3529	(i,f,p) = (1 1 1) 0.7931	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 1.0
$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$
(i,f,p) = (0 1 1) 1.0	(i,f,p) = (0 1 1) 0.3333	(i,f,p) = (0 1 1) 0.125	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0
. . .	. . .	. . .		. . .
$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$	$P(F = f\|P = p)$
(p,f) = (0 1) 0.463	(p,f) = (0 1) 0.3188	(p,f) = (0 1) 0.3333	(p,f) = (0 1) 0.4762	(p,f) = (0 1) 0.4343
$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$
(i,f,p) = (1 1 0) 0.0645	(i,f,p) = (1 1 0) 0.6	(i,f,p) = (1 1 0) 0.9524	(i,f,p) = (1 1 0) 1.0	(i,f,p) = (1 1 0) 1.0
$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$	$P(F = f\|P = p, I = i)$
(i,f,p) = (0 1 0) 1.0	(i,f,p) = (0 1 0) 0.2041	(i,f,p) = (0 1 0) 0.0238	(i,f,p) = (0 1 0) 0	(i,f,p) = (0 1 0) 0.0175

TABLE F.11: N Test 2

interval 0	interval 1	interval 2	interval 3	interval 4
2010-12-14-2012-05-03	2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$	$P(S = s\|I = i)$
(i,s) = (1 1) 0.5692	(i,s) = (1 1) 0.5918	(i,s) = (1 1) 0.58	(i,s) = (1 1) 0.2453	(i,s) = (1 1) 0.2075
(i,s) = (0 1) 0.3143	(i,s) = (0 1) 0.2344	(i,s) = (0 1) 0.3636	(i,s) = (0 1) 0.3016	(i,s) = (0 1) 0.1094
(i,s) = (1 0) 0.4308	(i,s) = (1 0) 0.4082	(i,s) = (1 0) 0.42	(i,s) = (1 0) 0.7547	(i,s) = (1 0) 0.7925
(i,s) = (0 0) 0.6857	(i,s) = (0 0) 0.7656	(i,s) = (0 0) 0.6364	(i,s) = (0 0) 0.6984	(i,s) = (0 0) 0.8906
P(S=s):	P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.52	s = 0 : 0.6106	s = 0 : 0.5431	s = 0 : 0.7241	s = 0 : 0.8462
s = 1 : 0.48	s = 1 : 0.3894	s = 1 : 0.4569	s = 1 : 0.2759	s = 1 : 0.1538
- - -	- - -	- - -	- - -	- - -
$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$	$P(P = p\|I = i)$
(i,p) = (1 1) 0.5231	(i,p) = (1 1) 0.5918	(i,p) = (1 1) 0.58	(i,p) = (1 1) 0.2453	(i,p) = (1 1) 0.2075
(i,p) = (0 1) 0.3429	(i,p) = (0 1) 0.2344	(i,p) = (0 1) 0.3636	(i,p) = (0 1) 0.3016	(i,p) = (0 1) 0.1094
(i,p) = (1 0) 0.4769	(i,p) = (1 0) 0.4082	(i,p) = (1 0) 0.42	(i,p) = (1 0) 0.7547	(i,p) = (1 0) 0.7925
(i,p) = (0 0) 0.6571	(i,p) = (0 0) 0.7656	(i,p) = (0 0) 0.6364	(i,p) = (0 0) 0.6984	(i,p) = (0 0) 0.8906
P(P=p):	P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.54	p = 0 : 0.6106	p = 0 : 0.5431	p = 0 : 0.7241	p = 0 : 0.8462
p = 1 : 0.46	p = 1 : 0.3894	p = 1 : 0.4569	p = 1 : 0.2759	p = 1 : 0.1538

TABLE F.12: N Test 3

interval 1	interval 2	interval 3	interval 4
2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$
(s, s_prev) = (1 1) 0.5417	(s, s_prev) = (1 1) 0.7045	(s, s_prev) = (1 1) 0.4906	(s, s_prev) = (1 1) 0.4062
$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$
(s,s_prev,i) = (1 1 1) 0.4865	(s,s_prev,i) = (1 1 1) 0.7241	(s,s_prev,i) = (1 1 1) 0.3793	(s,s_prev,i) = (1 1 1) 0.5385
$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$
(s,s_prev,i) = (1 1 0) 0.7273	(s,s_prev,i) = (1 1 0) 0.6667	(s,s_prev,i) = (1 1 0) 0.625	(s,s_prev,i) = (1 1 0) 0.3158
. . .	. . .	. . .	. . .
$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$	$P(S = s \mid S_{-1} = s_{prev})$
(s, s_prev) = (1 0) 0.2769	(s, s_prev) = (1 0) 0.3056	(s, s_prev) = (1 0) 0.0952	(s, s_prev) = (1 0) 0.0588
$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$
(s,s_prev,i) = (1 0 1) 0.1071	(s,s_prev,i) = (1 0 1) 0.3	(s,s_prev,i) = (1 0 1) 0.0952	(s,s_prev,i) = (1 0 1) 0.1
$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$	$P(S = s \mid S_{-1} = s_{prev}, I = i)$
(s,s_prev,i) = (1 0 0) 0.4054	(s,s_prev,i) = (1 0 0) 0.3077	(s,s_prev,i) = (1 0 0) 0.0952	(s,s_prev,i) = (1 0 0) 0.0222
. . .	. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$
(p, p_prev) = (1 1) 0.5652	(p, p_prev) = (1 1) 0.7045	(p, p_prev) = (1 1) 0.4906	(p, p_prev) = (1 1) 0.4062
$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$
(p, p_prev, i) = (1 1 1) 0.5294	(p, p_prev, i) = (1 1 1) 0.7241	(p, p_prev, i) = (1 1 1) 0.3793	(p, p_prev, i) = (1 1 1) 0.5385
$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$
(p, p_prev, i) = (1 1 0) 0.6667	(p, p_prev, i) = (1 1 0) 0.6667	(p, p_prev, i) = (1 1 0) 0.625	(p, p_prev, i) = (1 1 0) 0.3158
. . .	. . .	. . .	. . .
$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$	$P(P = p \mid P_{-1} = p_{prev})$
(p, p_prev) = (1 0) 0.2687	(p, p_prev) = (1 0) 0.3056	(p, p_prev) = (1 0) 0.0952	(p, p_prev) = (1 0) 0.0588
$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$
(p, p_prev, i) = (1 0 1) 0.0968	(p, p_prev, i) = (1 0 1) 0.3	(p, p_prev, i) = (1 0 1) 0.0952	(p, p_prev, i) = (1 0 1) 0.1
$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$	$P(P = p \mid P_{-1} = p_{prev}, I = i)$
(p, p_prev, i) = (1 0 0) 0.4167	(p, p_prev, i) = (1 0 0) 0.3077	(p, p_prev, i) = (1 0 0) 0.0952	(p, p_prev, i) = (1 0 0) 0.0222

F.5 Ontology

For the Ontology, the results of Tests 1, 2, and 3 are presented in Tables F.13, F.14, and F.15, respectively.

TABLE F.13: Ontology Test 1

interval 0	interval 1	interval 2	interval 3	interval 4	interval 5
2010-12-14-2011-12-13	2011-12-13-2012-05-03	2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (1 1) 0.6061	(s,f) = (1 1) 0.5	(s,f) = (1 1) 0.5323	(s,f) = (1 1) 0.4848	(s,f) = (1 1) 0.3235	(s,f) = (1 1) 0.4828
$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$
(i,f,s) = (1 1 1) 0.4583	(i,f,s) = (1 1 1) 0.5556	(i,f,s) = (1 1 1) 0.8065	(i,f,s) = (1 1 1) 0.9655	(i,f,s) = (1 1 1) 1.0	(i,f,s) = (1 1 1) 1.0
$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$
(i,f,s) = (0 1 1) 1.0	(i,f,s) = (0 1 1) 0.4211	(i,f,s) = (0 1 1) 0.2581	(i,f,s) = (0 1 1) 0.1081	(i,f,s) = (0 1 1) 0	(i,f,s) = (0 1 1) 0
. . .	. . .	. . .	. . .	. . .	. . .
$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$	$P(F = f \mid S = s)$
(s,f) = (0 1) 0.5286	(s,f) = (0 1) 0.3765	(s,f) = (0 1) 0.2716	(s,f) = (0 1) 0.3038	(s,f) = (0 1) 0.4018	(s,f) = (0 1) 0.359
$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$
(i,f,s) = (1 1 0) 0.1951	(i,f,s) = (1 1 0) 0.3	(i,f,s) = (1 1 0) 0.6667	(i,f,s) = (1 1 0) 0.8846	(i,f,s) = (1 1 0) 0.9778	(i,f,s) = (1 1 0) 1.0
$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$	$P(F = f \mid S = s, I = i)$
(i,f,s) = (0 1 0) 1.0	(i,f,s) = (0 1 0) 0.4182	(i,f,s) = (0 1 0) 0.1053	(i,f,s) = (0 1 0) 0.0189	(i,f,s) = (0 1 0) 0.0149	(i,f,s) = (0 1 0) 0
. . .	. . .	. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (1 1) 0.6579	(p,f) = (1 1) 0.4524	(p,f) = (1 1) 0.5323	(p,f) = (1 1) 0.4848	(p,f) = (1 1) 0.3235	(p,f) = (1 1) 0.4828
$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$
(i,f,p) = (1 1 1) 0.48	(i,f,p) = (1 1 1) 0.52	(i,f,p) = (1 1 1) 0.8065	(i,f,p) = (1 1 1) 0.9655	(i,f,p) = (1 1 1) 1.0	(i,f,p) = (1 1 1) 1.0
$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$
(i,f,p) = (0 1 1) 1.0	(i,f,p) = (0 1 1) 0.3529	(i,f,p) = (0 1 1) 0.2581	(i,f,p) = (0 1 1) 0.1081	(i,f,p) = (0 1 1) 0	(i,f,p) = (0 1 1) 0
. . .	. . .	. . .	. . .	. . .	. . .
$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$	$P(F = f \mid P = p)$
(p,f) = (0 1) 0.4923	(p,f) = (0 1) 0.4045	(p,f) = (0 1) 0.2716	(p,f) = (0 1) 0.3038	(p,f) = (0 1) 0.4018	(p,f) = (0 1) 0.359
$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$
(i,f,p) = (1 1 0) 0.175	(i,f,p) = (1 1 0) 0.3438	(i,f,p) = (1 1 0) 0.6667	(i,f,p) = (1 1 0) 0.8846	(i,f,p) = (1 1 0) 0.9778	(i,f,p) = (1 1 0) 1.0
$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$	$P(F = f \mid P = p, I = i)$
(i,f,p) = (0 1 0) 1.0	(i,f,p) = (0 1 0) 0.4386	(i,f,p) = (0 1 0) 0.1053	(i,f,p) = (0 1 0) 0.0189	(i,f,p) = (0 1 0) 0.0149	(i,f,p) = (0 1 0) 0

TABLE F.14: Ontology Test 2

interval 0	interval 1	interval 2	interval 3	interval 4	interval 5
2010-12-14-2011-12-13	2011-12-13-2012-05-03	2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s \mid I = i)$	$P(S = s \mid I = i)$	$P(S = s \mid I = i)$	$P(S = s \mid I = i)$	$P(S = s \mid I = i)$	$P(S = s \mid I = i)$
(i,s) = (1 1) 0.3692	(i,s) = (1 1) 0.4737	(i,s) = (1 1) 0.5636	(i,s) = (1 1) 0.5273	(i,s) = (1 1) 0.1964	(i,s) = (1 1) 0.25
(i,s) = (0 1) 0.2368	(i,s) = (0 1) 0.2568	(i,s) = (0 1) 0.3523	(i,s) = (0 1) 0.4111	(i,s) = (0 1) 0.2556	(i,s) = (0 1) 0.1667
(i,s) = (1 0) 0.6308	(i,s) = (1 0) 0.5263	(i,s) = (1 0) 0.4364	(i,s) = (1 0) 0.4727	(i,s) = (1 0) 0.8036	(i,s) = (1 0) 0.75
(i,s) = (0 0) 0.7632	(i,s) = (0 0) 0.7432	(i,s) = (0 0) 0.6477	(i,s) = (0 0) 0.5889	(i,s) = (0 0) 0.7444	(i,s) = (0 0) 0.8333
P(S=s):	P(S=s):	P(S=s):	P(S=s):	P(S=s):	P(S=s):
s = 0 : 0.6796	s = 0 : 0.6489	s = 0 : 0.5664	s = 0 : 0.5448	s = 0 : 0.7671	s = 0 : 0.8014
s = 1 : 0.3204	s = 1 : 0.3511	s = 1 : 0.4336	s = 1 : 0.4552	s = 1 : 0.2329	s = 1 : 0.1986
- - -	- - -	- - -	- - -	- - -	- - -
$P(P = p \mid I = i)$	$P(P = p \mid I = i)$	$P(P = p \mid I = i)$	$P(P = p \mid I = i)$	$P(P = p \mid I = i)$	$P(P = p \mid I = i)$
(i,p) = (1 1) 0.3846	(i,p) = (1 1) 0.4386	(i,p) = (1 1) 0.5636	(i,p) = (1 1) 0.5273	(i,p) = (1 1) 0.1964	(i,p) = (1 1) 0.25
(i,p) = (0 1) 0.3421	(i,p) = (0 1) 0.2297	(i,p) = (0 1) 0.3523	(i,p) = (0 1) 0.4111	(i,p) = (0 1) 0.2556	(i,p) = (0 1) 0.1667
(i,p) = (1 0) 0.6154	(i,p) = (1 0) 0.5614	(i,p) = (1 0) 0.4364	(i,p) = (1 0) 0.4727	(i,p) = (1 0) 0.8036	(i,p) = (1 0) 0.75
(i,p) = (0 0) 0.6579	(i,p) = (0 0) 0.7703	(i,p) = (0 0) 0.6477	(i,p) = (0 0) 0.5889	(i,p) = (0 0) 0.7444	(i,p) = (0 0) 0.8333
P(P=p):	P(P=p):	P(P=p):	P(P=p):	P(P=p):	P(P=p):
p = 0 : 0.6311	p = 0 : 0.6794	p = 0 : 0.5664	p = 0 : 0.5448	p = 0 : 0.7671	p = 0 : 0.8014
p = 1 : 0.3689	p = 1 : 0.3206	p = 1 : 0.4336	p = 1 : 0.4552	p = 1 : 0.2329	p = 1 : 0.1986

TABLE F.15: Ontology Test 3

interval 1	interval 2	interval 3	interval 4	interval 5
2011-12-13-2012-05-03	2012-05-03-2012-07-24	2012-07-24-2012-12-11	2012-12-11-2013-03-12	2013-03-12-2013-04-30
$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$
(s, s_prev) = (1 1) 0.5323	(s, s_prev) = (1 1) 0.7174	(s, s_prev) = (1 1) 0.6613	(s, s_prev) = (1 1) 0.4545	(s, s_prev) = (1 1) 0.4706
$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$
(s,s_prev,i) = (1 1 1) 0.6316	(s,s_prev,i) = (1 1 1) 0.8261	(s,s_prev,i) = (1 1 1) 0.6667	(s,s_prev,i) = (1 1 1) 0.3125	(s,s_prev,i) = (1 1 1) 0.6364
$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$
(s,s_prev,i) = (1 1 0) 0.375	(s,s_prev,i) = (1 1 0) 0.6087	(s,s_prev,i) = (1 1 0) 0.6552	(s,s_prev,i) = (1 1 0) 0.5882	(s,s_prev,i) = (1 1 0) 0.3913
. . .	. . .	. . .	. . .	. . .
$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$	$P(S = s\|S_{prev} = s_{prev})$
(s, s_prev) = (1 0) 0.1884	(s, s_prev) = (1 0) 0.299	(s, s_prev) = (1 0) 0.3012	(s, s_prev) = (1 0) 0.05	(s, s_prev) = (1 0) 0.1161
$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$
(s,s_prev,i) = (1 0 1) 0.1579	(s,s_prev,i) = (1 0 1) 0.375	(s,s_prev,i) = (1 0 1) 0.3182	(s,s_prev,i) = (1 0 1) 0.0417	(s,s_prev,i) = (1 0 1) 0.1556
$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$	$P(S = s\|S_{prev} = s_{prev}, I = i)$
(s,s_prev,i) = (1 0 0) 0.2	(s,s_prev,i) = (1 0 0) 0.2615	(s,s_prev,i) = (1 0 0) 0.2951	(s,s_prev,i) = (1 0 0) 0.0536	(s,s_prev,i) = (1 0 0) 0.0896
. . .	. . .	. . .	. . .	. . .
$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$
(p, p_prev) = (1 1) 0.4677	(p, p_prev) = (1 1) 0.7143	(p, p_prev) = (1 1) 0.6613	(p, p_prev) = (1 1) 0.4545	(p, p_prev) = (1 1) 0.4706
$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$
(p, p_prev, i) = (1 1 1) 0.5789	(p, p_prev, i) = (1 1 1) 0.8421	(p, p_prev, i) = (1 1 1) 0.6667	(p, p_prev, i) = (1 1 1) 0.3125	(p, p_prev, i) = (1 1 1) 0.6364
$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$
(p, p_prev, i) = (1 1 0) 0.2917	(p, p_prev, i) = (1 1 0) 0.6087	(p, p_prev, i) = (1 1 0) 0.6552	(p, p_prev, i) = (1 1 0) 0.5882	(p, p_prev, i) = (1 1 0) 0.3913
. . .	. . .	. . .	. . .	. . .
$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$	$P(P = p\|P_{prev} = p_{prev})$
(p, p_prev) = (1 0) 0.1884	(p, p_prev) = (1 0) 0.3168	(p, p_prev) = (1 0) 0.3012	(p, p_prev) = (1 0) 0.05	(p, p_prev) = (1 0) 0.1161
$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$
(p, p_prev, i) = (1 0 1) 0.1579	(p, p_prev, i) = (1 0 1) 0.4167	(p, p_prev, i) = (1 0 1) 0.3182	(p, p_prev, i) = (1 0 1) 0.0417	(p, p_prev, i) = (1 0 1) 0.1556
$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$	$P(P = p\|P_{prev} = p_{prev}, I = i)$
(p, p_prev, i) = (1 0 0) 0.2	(p, p_prev, i) = (1 0 0) 0.2615	(p, p_prev, i) = (1 0 0) 0.2951	(p, p_prev, i) = (1 0 0) 0.0536	(p, p_prev, i) = (1 0 0) 0.0896

Bibliography

Robert Ackland. *Web social science: Concepts, data and tools for social scientists in the digital age.* Sage, 2013.

Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013.

Carlos Alós-Ferrer, Sabine Hügelschäfer, and Jiahui Li. Inertia and decision making. *Frontiers in psychology*, 7:169, 2016.

Aamena Alshamsi, Fabio Pianesi, Bruno Lepri, Alex Pentland, and Iyad Rahwan. Beyond contagion: Reality mining reveals complex patterns of social influence. *PloS one*, 10(8):e0135740, 2015.

Aris Anagnostopoulos, Ravi Kumar, and Mohammad Mahdian. Influence and correlation in social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 7–15. ACM, 2008.

Roy M Anderson, Robert M May, and B Anderson. *Infectious diseases of humans: dynamics and control*, volume 28. Wiley Online Library, 1992.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

Kevin Arceneaux, Alan S Gerber, and Donald P Green. A cautionary note on the use of matching to estimate causal effects: An empirical example comparing matching estimates to an experimental benchmark. *Sociological methods & research*, 39(2): 256–282, 2010.

Adi Avnit. The million followers fallacy. http://blog.pravdam.com/the-million-followers-fallacy-guest-post-by-adi-avnit/, 2009.

Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 65–74. ACM, 2011.

Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM, 2012.

Nicola Barbieri, Francesco Bonchi, and Giuseppe Manco. Influence-based network-oblivious community detection. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 955–960. IEEE, 2013.

Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.

William P Barnett and Glenn R Carroll. Modeling internal organizational change. *Annual review of sociology*, 21(1):217–236, 1995.

Pamela S Barr, John L Stimpert, and Anne S Huff. Cognitive change, strategic action, and organizational renewal. *Strategic management journal*, 13(S1):15–36, 1992.

Samy Bengio and Yoshua Bengio. Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks*, 11(3):550–557, 2000.

Jonah Berger. *Contagious: Why things catch on.* Simon and Schuster, 2013.

Jonah Berger, Marc Meredith, and S Christian Wheeler. Contextual priming: Where people vote affects how they vote. *Proceedings of the National Academy of Sciences*, 105(26):8846–8849, 2008.

Grant Blank. The digital divide among Twitter users and its implications for social research. *Social Science Computer Review*, 35(6):679–697.

Grant Blank and Christoph Lutz. Representativeness of social media in Great Britain: Investigating Facebook, Linkedin, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61(7):741–756, 2017.

Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8, 2011.

Phillip Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1):113–120, 1972.

Francesco Bonchi. Influence propagation in social networks: A data mining perspective. *IEEE Intelligent Informatics Bulletin*, 12(1):8–16, 2011.

Rick Bonney, Caren B Cooper, Janis Dickinson, Steve Kelling, Tina Phillips, Kenneth V Rosenberg, and Jennifer Shirk. Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11):977–984, 2009.

Javier Borge-Holthoefer, Nicola Perra, Bruno Gonçalves, Sandra González-Bailón, Alex Arenas, Yamir Moreno, and Alessandro Vespignani. The dynamics of information-driven coordination phenomena: A transfer entropy analysis. *Science advances*, 2(4): e1501158, 2016.

Jonathan Bright, Scott A Hale, Bharath Ganesh, Andrew Bulovsky, Helen Margetts, and Phil Howard. Does campaigning on social media make a difference? evidence from candidate use of Twitter during the 2015 and 2017 uk elections. *arXiv preprint arXiv:1710.07087v2*, 2017.

John T Cacioppo, James H Fowler, and Nicholas A Christakis. Alone in the crowd: the structure and spread of loneliness in a large social network. *Journal of personality and social psychology*, 97(6):977, 2009.

Rafael Cappelletti and Nishanth Sastry. IARank: Ranking users on Twitter in near real-time, based on their information amplification potential. In *Social Informatics (SocialInformatics), 2012 International Conference on*, pages 70–77. IEEE, 2012.

Paul R Carlile and Ann Langley. *How matter matters: Objects, artifacts, and materiality in organization studies*, volume 3. Oxford University Press, 2013.

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

Claudio Castellano and Romualdo Pastor-Satorras. Competing activation mechanisms in epidemics on networks. *Scientific reports*, 2:371, 2012.

Manuel Cebrian, Iyad Rahwan, and Alex Sandy Pentland. Beyond viral. *Communications of the ACM*, 59(4):36–39, 2016.

Damon Centola. Social media and the science of health behavior. *Circulation*, 127(21): 2135–2144, 2013.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113(3):702–734, 2007.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

Chaomei Chen. Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology*, 57(3):359–377, 2006.

Wei Chen, Laks VS Lakshmanan, and Carlos Castillo. Information and influence propagation in social networks. *Synthesis Lectures on Data Management*, 5(4):1–177, 2013.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. International World Wide Web Conferences Steering Committee, 2014.

Belkacem Chikhaoui, Mauricio Chiazzaro, and Shengrui Wang. A new granger causal model for influence evolution in dynamic social networks: The case of dblp. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

François Chollet et al. Keras. https://github.com/fchollet/keras, 2015.

Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *New England journal of medicine*, 357(4):370–379, 2007.

Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *New England journal of medicine*, 358(21):2249–2258, 2008.

Nicholas A Christakis and James H Fowler. *Connected: The surprising power of our social networks and how they shape our lives.* Little, Brown, 2009.

Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32(4):556–577, 2013.

Scott Counts, Munmun De Choudhury, Jana Diesner, Eric Gilbert, Marta Gonzalez, Brian Keegan, Mor Naaman, and Hanna Wallach. Computational social science: Cscw in the social media era. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 105–108. ACM, 2014.

Kate Crawford, Mary L Gray, and Kate Miltner. Big data— critiquing big data: Politics, ethics, epistemology— special section introduction. *International Journal of Communication*, 8:10, 2014.

Nick Crossley and Joseph Ibrahim. Critical mass, social networks and collective action: Exploring student political worlds. *Sociology*, page 0038038511425560, 2012.

Robert A Dahl. The concept of power. *Behavioral science*, 2(3):201–215, 1957.

Karen Dale. Building a social materiality: Spatial and embodied politics in organizational control. *Organization*, 12(5):649–678, 2005.

Adnan Darwiche. Bayesian networks. *Communications of the ACM*, 53(12):80–90, 2010.

Thomas H Davenport and Laurence Prusak. *Information ecology: Mastering the information and knowledge environment.* Oxford University Press on Demand, 1997.

Morton Deutsch and Harold B Gerard. A study of normative and informational social influences upon individual judgment. *The journal of abnormal and social psychology*, 51(3):629, 1955.

Francis X Diebold. *Elements of forecasting*. Citeseer, 1998.

Francis X Diebold. *Forecasting*. Department of Economics, University of Pennsylvania, 2015.

Pedro Domingos and Matt Richardson. Mining the network value of customers. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66. ACM, 2001.

Elizabeth Dubois and Devin Gaffney. The multiple facets of influence: identifying political influentials and opinion leaders on twitter. *American Behavioral Scientist*, 58 (10):1260–1277, 2014.

David Easley and Jon Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

Dean Eckles and Eytan Bakshy. Bias and high-dimensional adjustment in observational studies of peer effects. *arXiv preprint arXiv:1706.04692*, 2017.

Michael Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 153(1-2):233–268, 2012.

Michael Eichler. Causal inference with multiple time series: principles and problems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(1997):20110613, 2013.

ESRC. Big data network phase 3: New and emerging forms of data policy demonstrator projects, 2017.

Enrique Estellés-Arolas and Fernando González-Ladrón-de Guevara. Towards an integrated crowdsourcing definition. *Journal of Information science*, 38(2):189–200, 2012.

Daniel Faggella. Your feed is all you: The nuanced art of personalization at Facebook. *Motherboard*, 2016.

Emily Falk and Christin Scholz. Persuasion, influence, and value: Perspectives from communication and social neuroscience. *Annual review of psychology*, 69:329–356, 2018.

Brian J Fogg. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):5, 2002.

Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, page 40. ACM, 2009.

James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *Bmj*, 337:a2338, 2008.

Deen Freelon. On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, 58(1):59–75, 2014.

Linton C Freeman. A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41, 1977.

Noah E Friedkin. *A structural theory of social influence*, volume 13. Cambridge University Press, 2006.

Yang Gao, Yan Chen, and KJ Ray Liu. On cost-effective incentive mechanisms in microtask crowdsourcing. *IEEE Transactions on Computational Intelligence and AI in Games*, 7(1):3–15, 2015.

Rumi Ghosh and Kristina Lerman. Community detection using a measure of global influence. In *Advances in Social Network Mining and Analysis*, pages 20–35. Springer, 2010a.

Rumi Ghosh and Kristina Lerman. Predicting influential users in online social networks. In *Proceedings of KDD workshop on Social Network Analysis (SNA-KDD)*, July 2010b.

Rumi Ghosh and Kristina Lerman. Rethinking centrality: The role of dynamical processes in social network analysis. *Discrete and Continuous Dynamical Systems Series B*, 19(5):1355 – 1372, July 2014.

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Gautam Korlam, Fabricio Benevenuto, Niloy Ganguly, and Krishna Phani Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*, pages 61–70. ACM, 2012.

Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *ICWSM*, pages 59–65, 2010.

Malcolm Gladwell. *The tipping point: How little things can make a big difference*. Back Bay Books, 2002.

Sharad Goel, Duncan J Watts, and Daniel G Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. ACM, 2012.

Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028. ACM, 2010.

Sandra González-Bailón, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. The dynamics of protest recruitment through an online network. *Scientific reports*, 1, 2011.

Amit Goyal, Francesco Bonchi, and Laks VS Lakshmanan. Learning influence probabilities in social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 241–250. ACM, 2010.

Mark Granovetter. Threshold models of collective behavior. *American journal of sociology*, 83(6):1420–1443, 1978.

Mark S Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

Alex Graves et al. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.

Julia Greenberg. Advertisers don't like Facebook's reactions. They love them. *WIRED*, 2016.

Aiden P Gregg, Nikhila Mahadevan, and Constantine Sedikides. The spot effect: People spontaneously prefer their own theories. *The Quarterly Journal of Experimental Psychology*, 70(6):996–1010, 2017.

Michael T Hannan and John Freeman. Structural inertia and organizational change. *American sociological review*, pages 149–164, 1984.

Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

James O Huff, Anne S Huff, and Howard Thomas. Strategic renewal and the interaction of cumulative stress and inertia. *Strategic Management Journal*, 13(S1):55–75, 1992.

Carlos Jensen, Shelly D. Farnham, Steven M. Drucker, and Peter Kollock. The effect of communication modality on cooperation in online environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '00, pages 470–477, New York, NY, USA, 2000. ACM. ISBN 1-58113-216-6.

Ece Kamar and Eric Horvitz. Incentives for truthful reporting in crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 3*, AAMAS '12, pages 1329–1330, Richland, SC, 2012. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 0-9817381-3-3, 978-0-9817381-3-0.

Elihu Katz and Paul Felix Lazarsfeld. *Personal Influence, The part played by people in the flow of mass communications*. New York: The Free Press, 1955.

Edward Keller and Jonathan Berry. *The influentials: One American in ten tells the other nine how to vote, where to eat, and what to buy.* Simon and Schuster, 2003.

Herbert C Kelman. Processes of opinion change. *Public opinion quarterly*, 25(1):57–78, 1961.

David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

David Kempe, Jon Kleinberg, and Éva Tardos. Influential nodes in a diffusion model for social networks. In *Automata, languages and programming*, pages 1127–1138. Springer, 2005.

William O Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences*, volume 115, pages 700–721. The Royal Society, 1927.

Martin Kilduff, Dan S Chiaburu, and Jochen I Menges. Strategic use of emotional intelligence in organizational settings: Exploring the dark side. *Research in organizational behavior*, 30:129–152, 2010.

Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Copy at http://j. mp/1sexgVw Download Citation BibTex Tagged XML Download Paper*, 378, 2016.

Maksim Kitsak, Lazaros K Gallos, Shlomo Havlin, Fredrik Liljeros, Lev Muchnik, H Eugene Stanley, and Hernán A Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, 2010.

Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. *Algorithmic game theory*, 24:613–632, 2007.

Jon Kleinberg. The convergence of social and technological networks. *Communications of the ACM*, 51(11):66–72, 2008.

Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

Konstantin Klemm, M Ángeles Serrano, Víctor M Eguíluz, and Maxi San Miguel. A measure of individual role in collective dynamics. *Scientific reports*, 2, 2012.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

Christopher CM Kyba, Janna M Wagner, Helga U Kuechly, Constance E Walker, Christopher D Elvidge, Fabio Falchi, Thomas Ruhtz, Jürgen Fischer, and Franz Hölker. Citizen science provides valuable data for monitoring global night sky luminance. *Scientific reports*, 3, 2013.

George Lakoff. Dont think of an elephant! know your values and frame the debate. the essential guide for progressives, including post-election updates, 2004.

George Lakoff. Why it matters how we frame the environment. *Environmental Communication*, 4(1):70–81, 2010.

Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. Finding effectors in social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1059–1068. ACM, 2010.

Bruno Latour. Reassembling the social. *London: Oxford*, 2005.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, and Marshall Van Alstyne. Computational social science. *Science*, 323(5915):721–723, 2009.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

Elissa Lee and Laura Leets. Persuasive storytelling by hate groups online: Examining its effects on adolescents. *American Behavioral Scientist*, 45(6):927–957, 2002.

RTAJ Leenders. Longitudinal behavior of network structure and actor attributes: modeling interdependence of contagion and selection. *Evolution of social networks*, 1, 1997.

Kristina Lerman. Information is not a virus, and other consequences of human cognitive limits. *Future Internet*, 8(2), 2016.

Kristina Lerman and Rumi Ghosh. Information contagion: An empirical study of the spread of news on digg and twitter social networks. *Icwsm*, 10:90–97, 2010.

Jure Leskovec, Lada A Adamic, and Bernardo A Huberman. The dynamics of viral marketing. *ACM Transactions on the Web (TWEB)*, 1(1):5, 2007.

Jure Leskovec, Ajit Singh, and Jon Kleinberg. Patterns of influence in a recommendation network. In *Advances in Knowledge Discovery and Data Mining*, pages 380–389. Springer, 2006.

Sam Levin. Facebook to give congress thousands of ads bought by Russians during election. *The Guardian*, 2017.

David Liben-Nowell and Jon Kleinberg. Tracing information flow on a global scale using internet chain-letter data. *Proceedings of the national academy of sciences*, 105(12): 4633–4638, 2008.

Dimitra Liotsiou, Luc Moreau, and Susan Halford. Social influence: From contagion to a richer causal understanding. In *International Conference on Social Informatics*, volume 10047, pages 116–132. Springer, 2016.

Jiaying Liu, Siman Zhao, Xi Chen, Emily Falk, and Dolores Albarracín. The influence of peer behavior as a function of social and cultural closeness: A meta-analysis of normative influence on adolescent smoking initiation and continuation. *Psychological bulletin*, 143(10):1082, 2017.

Russell Lyons. The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, and Policy*, 2(1), 2011.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Farhad Manjoo. Slack, the office messaging app that may finally sink email. *The New York Times*, 2015.

Helen Margetts, Peter John, Scott Hale, and Taha Yasseri. *Political turbulence: How social media shape collective action*. Princeton University Press, 2015.

John Markioff. Government aims to build a data eye in the sky. *The New York Times*, 2011.

Noortje Marres. Digital sociology: The reinvention of social research, 2017.

Winter Mason, Jennifer Wortman Vaughan, and Hanna Wallach. Computational social science and social computing. *Machine Learning*, 95(3):257, 2014.

Winter A Mason, Frederica R Conrey, and Eliot R Smith. Situating social influence processes: Dynamic, multidirectional flows of influence within social networks. *Personality and social psychology review*, 11(3):279–300, 2007.

Rose McDermott, James H Fowler, and Nicholas A Christakis. Breaking up is hard to do, unless everyone else is doing it too: Social network effects on divorce in a longitudinal sample. *Social Forces*, 92(2):491–519, 2013.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

Florian Meier, David Elsweiler, and Max L Wilson. More than liking and bookmarking? towards understanding Twitter favouriting behaviour. In *ICWSM*, 2014.

Panagiotis T Metaxas, Eni Mustafaraj, and Dani Gayo-Avello. How (not) to predict elections. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 165–171. IEEE, 2011.

Gilad Mishne, Natalie S Glance, et al. Predicting movie sales from blogger sentiment. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 155–158, 2006.

Brent Mittelstadt. Automation, algorithms, and politics— auditing for transparency in content personalization systems. *International Journal of Communication*, 10:12, 2016.

Luc Moreau, Paul Groth, James Cheney, Timothy Lebo, and Simon Miles. The rationale of prov. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35: 235–257, 2015.

Gareth Morgan, Fred Gregory, and Cameron Roach. Images of organization. 1997.

Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference.* Cambridge University Press, 2014.

Peter Morriss. *Power: a philosophical analysis.* Manchester University Press, 1987.

Seth A Myers, Chenguang Zhu, and Jure Leskovec. Information diffusion and external influence in networks. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 33–41. ACM, 2012.

Mark EJ Newman. Networks: an introduction. 2010.

Mark EJ Newman. Prediction of highly cited papers. *EPL (Europhysics Letters)*, 105 (2):28002, 2014.

JS Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9.(tlanslated and edited by dm dabrowska and tp speed, statistical science (1990), 5, 465-480). *Annals of Agricultural Sciences*, 10:1–51, 1923.

David W Nickerson. Is voting contagious? evidence from two field experiments. *American Political Science Review*, 102(01):49–57, 2008.

Andrzej Nowak, Jacek Szamrej, and Bibb Latané. From private attitude to public opinion: A dynamic theory of social impact. *Psychological Review*, 97(3):362, 1990.

Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas Wood. Taking corrections literally but not seriously? the effects of information on factual beliefs and candidate favorability. 2017.

Brendan Nyhan and Jason Reifler. When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330, 2010.

Elizabeth L Ogburn. Challenges to estimating contagion effects from observational data. *arXiv preprint arXiv:1706.08440*, 2017. To appear in Spreading Dynamics in Social Systems; Eds. Sune Lehmann and Yong-Yeol Ahn, Springer Nature.

Daniel Olguín Olguín and Alex Pentland. Assessing group performance from collective behavior. In *Proc. of the CSCW*, volume 10, 2010.

Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. Social data: Biases, methodological pitfalls, and ethical boundaries. 2016.

Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2017.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Talcott Parsons. On the concept of influence. *Public opinion quarterly*, 27(1):37–62, 1963.

Paul B Paulus and Bernard A Nijstad. *Group creativity: Innovation through collaboration*. Oxford University Press, 2003.

Judea Pearl. Bayesian networks, causal inference and knowledge discovery. 2001.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009a.

Judea Pearl. *Causality*. Cambridge University Press, 2009b.

Judea Pearl. The foundations of causal inference. *Sociological Methodology*, 40(1):75–149, 2010.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: a primer*. John Wiley & Sons, 2016. Chapter previews can be found at http://bayes.cs.ucla.edu/PRIMER/.

Alex Pentland. *Social physics: How good ideas spread-the lessons from a new science*. Penguin, 2014.

Garry Robins, Philippa Pattison, and Peter Elliott. Network models for social influence processes. *Psychometrika*, 66(2):161–189, 2001.

Everett M Rogers. *Diffusion of Innovations*. Simon and Schuster, 2003.

Daniel M Romero, Brendan Meeder, and Jon Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *Proceedings of the 20th international conference on World wide web*, pages 695–704. ACM, 2011.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Jakob Runge. Quantifying information transfer and mediation along causal pathways in complex systems. *Physical Review E*, 92(6):062829, 2015.

Matthew J Salganik, Peter Sheridan Dodds, and Duncan J Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762): 854–856, 2006.

Elliot Schrage. Hard questions: Russian ads delivered to Congress, 2017.

Cosma Shalizi. *Advanced data analysis from an elementary point of view*. Cambridge University Press, 2013.

Cosma Rohilla Shalizi. Return of "homophily, contagion, confounding: Pick any three", or, the adventures of irene and joey along the back-door paths. `http://bactra.org/weblog/656.html`, 2010. Accessed: 2017-12-02.

Cosma Rohilla Shalizi and Andrew C Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research*, 40(2):211–239, 2011.

Amit Sharma and Dan Cosley. Distinguishing between personal preferences and social influence in online activity feeds. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 1091–1103, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3592-8.

Amit Sharma, Jake M Hofman, and Duncan J Watts. Estimating the causal impact of recommendation systems from observational data. In *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, pages 453–470. ACM, 2015.

Kijung Shin, Tina Eliassi-Rad, and Christos Faloutsos. Corescope: Graph mining using k-core analysispatterns, anomalies and algorithms. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 469–478. IEEE, 2016.

Justin Shriber. How b2b sellers are offering personalization at scale. *Harvard Business Review*, 2017.

Jonathan Silvertown. A new dawn for citizen science. *Trends in Ecology Evolution*, 24 (9):467 – 471, 2009. ISSN 0169-5347.

Sergio Sismondo. An introduction to science and technology studies. 2009.

Olivia Solon and Sabrina Siddiqui. Russia-backed Facebook posts 'reached 126m americans' during US election. *The Guardian*, 2017.

Dan Sperber. *Explaining culture: A naturalistic approach*. Cambridge University Press, 1996.

Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11(May):1643–1662, 2010.

Vasumathi Sridharan, Vaibhav Shankar, and Minaxi Gupta. Twitter games: how successful spammers pick targets. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 389–398. ACM, 2012.

David Strang and Sarah A Soule. Diffusion in organizations and social movements: From hybrid corn to poison pills. *Annual review of sociology*, 24(1):265–290, 1998.

Gianluca Stringhini, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. Poultry markets: On the underground economy of twitter followers. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN '12, pages 1–6, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1480-0.

Eric Sun, Itamar Rosenn, Cameron Marlow, and Thomas M Lento. Gesundheit! modeling contagion through Facebook news feed. In *ICWSM*, 2009.

Araz Taeihagh. Crowdsourcing, sharing economies and development. *Journal of Developing Societies*, 33(2):191–222, 2017.

Richard H Thaler and Cass R Sunstein. Nudge: Improving decisions about health, wealth, and happiness., 2009.

AC Thomas. The social contagion hypothesis: comment on social contagion theory: examining dynamic social networks and human behavior. *Statistics in medicine*, 32 (4):581–590, 2013.

Ramine Tinati, Markus Luczak-Roesch, Elena Simperl, and Wendy Hall. Because science is awesome: Studying participation in a citizen science game. In *Proceedings of the 8th ACM Conference on Web Science*, WebSci '16, pages 45–54, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4208-7.

Ramine Tinati, Max Van Kleek, Elena Simperl, Markus Luczak-Rösch, Robert Simpson, and Nigel Shadbolt. Designing for citizen data analysis: A cross-sectional case study of a multi-domain citizen science platform. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 4069–4078, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6.

Fani Tsapeli and Mirco Musolesi. Investigating causality in human behavior from smartphone sensor data: a quasi-experimental approach. *EPJ Data Science*, 4(1):24, 2015.

Zeynep Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM*, 14:505–514, 2014.

Johan Ugander. Truth, lies, and an ethics of personalization. *Medium*, 2017.

Greg Ver Steeg, Rumi Ghosh, and Kristina Lerman. What stops social epidemics? In *ICWSM*, 2011.

Robin Wagner-Pacifici, John W Mohr, and Ronald L Breiger. Ontologies, methodologies, and new uses of big data in the social and cultural sciences, 2015.

Hanna Wallach. Computational social science: Toward a collaborative future. *Computational Social Science: Discovery and Prediction*, 2016.

Duncan Watts. Challenging the influentials hypothesis. *WOMMA Measuring Word of Mouth*, 3(4):201–211, 2007.

Duncan J Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

Duncan J Watts. *Everything is obvious:* Once you know the answer.* Crown Business, 2011.

Duncan J Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

Dennis G Wilson. The ethics of automated behavioral microtargeting. *AI Matters*, 3 (3):56–64, October 2017. ISSN 2372-3483.

Fang Wu, Bernardo A Huberman, Lada A Adamic, and Joshua R Tyler. Information flow in social groups. *Physica A: Statistical Mechanics and its Applications*, 337(1-2): 327–335, 2004.