UNIVERSITY OF SOUTHAMPTON

Faculty of Engineering, Science and Mathematics

School of Electronics and Computer Science

Investigating Cascades in Social Networks: Structural and Temporal Aspects

by

Nora Alrajebah

Thesis for the degree of Doctor of Philosophy in Computer Science

May 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Computer Science

Thesis for the degree of Doctor of Philosophy

INVESTIGATING CASCADES IN SOCIAL NETWORKS: STRUCTURAL AND TEMPORAL ASPECTS

By Nora Alrajebah

There has been significant interest in studying social interactions in online social networks, such as how people exchange opinions, disseminate information, and adopt certain behaviours. One phenomenon addressed is information diffusion: the way information is spread in social networks. Since their emergence, online social networks have been used by people to create and share content. They provide a set of functionalities that facilitate these and other tasks, allowing users to interact with each other. For researchers, these platforms became the basis for understanding complex human behaviours, one of which is the 'urge' to share content with others. Online social networks allow users to create and share various types of content daily. In fact, the bulk of the content displayed on these platforms is not original but shared. Thus, the ability to decipher the phenomenon of information diffusion became essential in diverse fields, such as marketeers who wish to create content that spreads, sociologists who wish to understand the underlying phenomenon, and web scientists who wish to understand the web as a socio-technical entity.

In its simplest form, the information diffusion process in online social networks consists of the content that spreads, the context that facilitates the spread, and the outcome of the process. The underlying structure on which the content spreads is the network of connections between users (the social network). Therefore, the structure of the diffusion is also a network that links users, and is based on information about who influences whom to spread the content. This network is known as the cascade. In the literature, diffusion and cascades are intersecting concepts, and they are often used interchangeably. However, this work differentiates the two. Diffusion is used to

refer to the phenomenon while cascade is used to refer to the result of the diffusion, i.e. the structural representation of the diffusion process.

This work investigates information diffusion on Tumblr, an online social network platform that provides reblogging functionality. Reblogging allows users to reblog posts, which creates a cascading behaviour that can be observed. The reblogging history is provided as a list of notes attached to each post and all of its reblogged copies. In practice, these notes have two parts: structural (who reblogged from whom) and temporal (when did the reblogging occur). These two aspects complement each other in providing an understanding of the diffusion process as it manifests in the form of a cascade. Studying such explicit cascades is important as it allows understanding the information diffusion, a phenomenon that occurs in many implicit forms on the Web.

This work's contributions include proposing an information diffusion framework that conceptualises the elements of the diffusion (namely, the content, context and cascade) and how they relate to each other. It also proposes construction models that create cascade networks minimal contextual information and missing/degraded data. In addition, this work provides a survey of the structural and temporal features of cascades, including their definitions and implications. It also investigates Tumblr as a platform for information diffusion, analyses the structural and temporal aspects of Tumblr's cascades and compares its features with cascades obtained from other platforms.

The main findings show that Tumblr's most popular content create 'large' cascades that are deep, branching into a large number of separate and long paths, having a consistent number of reblogs at each depth and at each given time. These cascades gain their popularity throughout time in various ways; some of them feature high reblogging activities followed by idleness phases, others fluctuate more slowly accumulating rebloggings. Few cascades regain their popularity after long periods of idleness, while the majority have one outstanding popularity phase that is never repeated.

Table of Contents

List of Tables

List of Figures

DECLARATION OF AUTHORSHIP

I, Nora Alrajebah 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.

Investigating Cascades in Social Networks: Structural and Temporal Aspects

I confirm that:

1. This work was done wholly or mainly 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:

    - Alrajebah, N. (2015). Investigating the Structural Characteristics of Cascades on Tumblr. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, (ASONAM 2015), ACM, pp. 910–917.
    - Alrajebah, N., Carr, L., Luczak-Roesch, M. & Tiropanis, T. (2017). Deconstructing Diffusion on Tumblr : Structural and Temporal Aspects. In: *Proceedings of the 9th ACM Conference on Web Science* (WebSci 2017), ACM.
    - Alrajebah, N., Tiropanis, T. & Carr, L. (2017). Cascades on online social networks: A chronological account. In: *International Conference on Internet Science*. (INSCI 2017), Springer, Cham, pp. 393–411.

Signed: .....

Date: .....

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors Prof. Leslie Carr and Dr. Thanassis Tiropanis, for their continued support, patience and encouragement throughout my PhD journey. I would not have been able to reach that far without their guidance. I would like to thank Prof. Carr for being kind, helpful and for providing me with endless support whenever I needed. I also would like to thank Dr. Tiropanis for all the discussions we had, I learned a lot from him on both the personal and the professional levels.

I would like to thank my examiners Dr. David Millard and Prof. Yi-Ke Guo, for all of their valuable feedback. Their feedback helped in shaping my work and massively improved my thesis.

I am thankful for Dr. Markus Luczak-Roesch, who was my mentor during the second and third year of my PhD. I am grateful for his guidance and support and the many hours he spent discussing ideas and exchanging thoughts. I also would like to thank Mr. Lester Gilbert, who generously offered some of his time to help me understand some of the statistical measures at an early stage of my PhD.

I am also thankful for the Web and Internet Sciences group for providing such a wonderful environment; I truly enjoyed the four years I spent there. I'm also thankful for my friends and lab mates: Dr. Areeb Alowisheq, Dr. Fatimah Akeel, Dr. Alaa Mashat, Dr. Nora Almuhanna, Dr. Huw Fryer, Dr. Jonathan Scott and Dr. Russell Newman, for their wonderful company, and for being there for me each time I needed to discuss ideas and solve some enigmas.

I also owe my sincere thanks to my sponsors, King Saud University, and the Saudi Arabian Cultural Bureau, who provided the financial support for my scholarship.

I am thankful for my mother Amal Alsharhan and my father Ibrahim Alrajebah, for believing in me, for being there for me and for their love, prayers and support throughout my life. I am forever at their debt. I extend my thanks to my siblings Abdulaziz, Sheikha, Shahad, Abdulrahman and Bedour, the gang, my best friends and the apple of my eyes. Thanks for your love, laughter, prayers and the good company. Last but not least, I would like to thank my friends for everything they are and everything they do.

Chapter 1: Introduction

'Ideas and products and messages and behaviors spread like viruses do.'

(Gladwell, 2000)

1.1 Social Networking on the Web

Since its emergence the Web has become a vast medium through which billions of users connect and interact with each other. Besides the ability to connect with each other, the Web provides users with the ability to create and share a huge amount of content. This has been facilitated by a large number of platforms that have evolved over the years. From the classic blogs and forums to the more complex online social network platforms such as Twitter, Facebook, Tumblr, Pinterest and Snapchat, the Web has played an important role in shaping and influencing the way humans interact with each other. Currently, there are 318.9 million blogs on Tumblr, where users are placing about 41.7 million posts each day[1]. Also, at June 2016, there were about 1.71 billion monthly active users of Facebook[2], who together share one million links every 20 minutes[3].

Online social network platforms have served as a substantial venue for research since their emergence more than a decade ago. They offer an enormous amount of data that can be analysed to cultivate insights about the way humans behave and interact with each other within the virtual borders of these networks. Such analysis is fundamental in many fields, including Marketing (Leskovec et al., 2007a), Epidemiology (Chunara et al., 2012), Natural-Disaster Management (Fraustino et al., 2012), media (Kwak et al., 2010), and Computer Science (Dow et al., 2013). Each of these disciplines perceives online social networks differently, guided by their different motivations.

In their current state, online social networks provide functionality to define and schematise users' behaviour. Users can 'follow/friend' each other; they can also 'publish' various types of user-created content. In addition, they can get involved with the published content, e.g. by 'liking' the content to show their admiration, starting a discussion by 'mentioning' or 'commenting', or by 'reblogging/retweeting/sharing' the content to spread it for any number of motives (boyd et al., 2010; Meier et al., 2014). All of these interactions are defined and bounded by the functionality offered by the platform.

[1] https://www.tumblr.com/press
[2] http://newsroom.fb.com/company-info
[3] http://www.statisticbrain.com/facebook-statistics/

The power of online social networks lies in providing their users with four features: 1) The ability to publish content, 2) The ability to follow users and to create an audience (by being followed), 3) The ability to share content using built-in functionalities such as retweet or reblog, 4) The ability to show reactions to the content, e.g. mentioning the content's author, commenting on the content or liking it. All these features contribute to the process of *content propagation*, which is one of the essential aspects of modern online social networking. To summarise, online social network platforms allow users to connect to each other to form social networks through which content can be published, shared and spread. They provide functionalities that facilitate content spread and bring users' attention to the content.

1.2 An Introduction to Information Diffusion

One aspect that has received a significant amount of research attention is the phenomenon known as information diffusion (Rogers, 2003). In the context of online social networks, studying information diffusion means studying the way that information is spread within such networks. In the previous section, this phenomenon was referred to as content propagation or spread. Research on information diffusion has varied according to the purpose of the study. The earliest research in this field utilised the blogosphere (Adar & Adamic, 2005; Leskovec et al., 2007b); and as new platforms have emerged, these have been used to analyse information diffusion dynamics (Kwak et al., 2010; Galuba & Aberer, 2010; Lerman & Ghosh, 2010; Bakshy et al., 2012).

Information diffusion has been the focus of various research including: inferring diffusion paths (Gomez Rodriguez et al., 2010), estimating user influence (Cha et al., 2010), predicting the future propagation (Yang & Counts, 2010; Cheng et al., 2014), modelling information diffusion (Gruhl et al., 2004; Liben-Nowell & Kleinberg, 2008), identifying trends, trendsetters, topics and events (Saez-Trumper et al., 2012; Romero et al., 2011), and measuring trust between users (Adali et al., 2010).

Based on the literature, the diffusion process has three components: the content that is spread, the context that facilitates the spread, and the outcome of the spread – the cascade. Of the two categories of content, the first is a platform-defined element such as a tweet in Twitter or a post in Tumblr. The second category covers any element that can be embedded within platform-defined elements such as a URL, a hashtag, a text, or a photo. Different content types require different data collection and analysis methods. Some content is more easily shared than others based on factors such as the how interesting it is and its stickiness. Thus, the content holds its value within it, it has its indigenous factors that motivate users to share that content. For example, some users will share the content if they find it humorous, regardless of the user's

agreement, or if they want to express their agreement with it, or even to associate themselves with an idea or with a community.

The context that facilitates the spread takes into account all the surrounding factors, other than the content itself. Two main factors facilitate the spread of content or diminish it. The first is related to the platform itself and its affordances. The second takes into account everything that is related to the individuals involved in the process, including their relationships (the social network), their influence, and the homophily between these individuals.

The third component, cascade, refers to the structural representation of the information diffusion process and is often perceived as the final outcome of the process (Leskovec et al., 2006b; Goel et al., 2012).

1.3 A Deeper Look into Cascades

The propagation of content is not only by the user who created that content (the author), but the result of cumulative efforts by many users who share the content with their friends and followers. Each user who participates by spreading the content therefore adds value to the overall cascade growth. So, cascades are represented as networks that show who influenced whom to propagate the content. Often such networks have the users as nodes while the edges represent the direction of information flow between them (Easley & Kleinberg, 2010). Practically, cascade networks are constructed from the series of *diffusion events* by which a specific piece of information is shared between two users. Many approaches have been used in the literature to construct cascade networks, but their baselines are similar as they all have users as nodes and edges representing the flow (Leskovec et al., 2006a). However, the construction process differs depending on the platform itself and the contextual information available to be collected from it.

Cascade networks are often perceived as the final outcome of a collective effort of many users. These networks can be temporally and structurally quantified using various measures. The importance of cascades is that they can show the negotiations that lead to their creation. Hence, cascades are vital to our understanding of how popular content becomes admired by deconstructing the process into separate diffusion events. With accurate data, cascades can be broken down to individual diffusion events, each of which consists of two users, a *source* and a *target*, with a *timestamp*. Thus, it is possible to decipher the popularity-gaining phenomenon by following the timeline of the diffusion as it manifests as a series of diffusion events.

The three aspects of studying cascades and information diffusion are temporal, structural, and social (Taxidou & Fischer, 2013). Each of these focuses on a different perspective and

complements one another. Thus they provide a better understanding of the information diffusion dynamics within social networks. Temporal analysis focuses on the speed of content diffusion and its growth over time. Structural analysis looks at the topological aspects of the diffusion and quantifies the propagation in ways that can be used to describe its structure. Together, the structural and the temporal analysis of cascades provide a better understanding of cascades, as Scott (2008) argues that the temporal aspects of social networks data strengthen the analysis beyond analysing the structural aspects of networks only. The social aspect is the one that has received the least attention in research. It is concerned with users' behaviour during the diffusion process to provide a better understating of the diffusion dynamics and the users' motives.

1.4 Motivation for this Study

Online social networks have proven, in many occasions, their vitality for a range of activities that are powered by information diffusion, such as: Mass convergence and emergency events (Hughes & Palen, 2009), Spreading information about good practices such as saving energy on earth day (Cheong & Lee, 2010), bringing people's attention to incidents that might lead to 'public shaming' behaviour (McBride, 2015). For instance, when a public relation executive lost her job over a racist tweet (Pilkington, 2013). More recently, a woman tweeted a photo of the man who harassed while on a plane (Strutner, 2016). Apart from individuals' public shaming incidents corporates might be affected, for example, Tesco's incident when a photo of a gendered sign was posted on Twitter (Gander, 2014). Therefore, studying information diffusion is vital for many purposes:

- To understand popularity gaining and virality, by deconstructing the process to single diffusion events.
- For marketeers, to understand how to make the content spread.
- To study related phenomena such as user influence.
- To predict the future spread of content.
- To Identify trends, trendsetters, events and topics.
- For modelling purposes (for epidemiologists and economists).
- As a measure of trust between users.

Investigating cascades provides many insights about the diffusion mechanism and the role of individuals involved in the process. On the one hand, understanding the temporal and structural aspects of a cascade and how it progressed can help explain why some cascades continue to spread and why others die. On the other hand, looking at cascades at the individual level provides a detailed look at the dynamics that cause such spreading. While great progress has been made in quantitative research on online information diffusion, there are some critical

reflections about this "simple epidemiological view" (Goel et al., 2015a), a critique which is amplified by studies investigating biases and inequalities in social media systems and data.

1.5 Scope of this Research

The three elements of information diffusion are: the content, the context, and the cascade. This work is concerned with the *context* and the *cascade*. For the purpose of the study, Tumblr[4] (an online social network platform) was chosen for a number of reasons.

1. There have been no thorough studies of cascades on Tumblr. Chang et al. (2014) and Xu et al. (2014) did not analyse Tumblr's cascades fully.
2. Tumblr has a unique way of presenting diffusion information as a list of notes underneath each post, minimising the chances of ambiguity during the cascade construction phase. This list is also unified, all the reblogged posts having the same list.
3. Communication style: Tumblr is a platform that is heavily dependent on content sharing. Tumblr's CEO reported that 90% of Tumblr content is actually reblogs (Karp, 2014).
4. Demographics: most Tumblr users are young (Chang et al., 2014), thus, the nature of the shared content is different and also Tumblr is a platform of fandoms (Hillman et al., 2014). This effectively changes the way users discover content as these fandoms are often identified by their corresponding tags.
5. Content-related: the most popular content on Tumblr is photos and memes, which makes it different from other social networks such as Twitter and Facebook that are more textual based (Chang et al., 2014; Xu et al., 2014).

This work studies the structural and temporal aspects of information diffusion on Tumblr in the form of a cascade of individual posts. It also investigates Tumblr's functionalities and analyses them within the context of diffusion. It looks at the implications of Tumblr's affordances such as the ability to reblog more than once and the ability to delete reblogs on the data harvesting, modelling and cascades constructions.

For the analysis, the data is collected from the acclaimed blog *Year in Review*, which is, in the words of Tumblr's staff, "a showcase of the best stuff on the Internet" each year. This blog contains a mixture of popular posts and trendy topics in different categories curated by Tumblr's staff. The detailed methodology used to create this blog is not published, but it uses the

[4] Tumblr.com

measurement of web traffic, volume of posts and tags, follower growth over time, etc. This work attempts to 'deconstruct' the most popular posts from the *Year in Review* blog structurally and temporally.

1.6 Research Questions

The main research question of this thesis is:

How does information diffusion occur on social networks?

In order to answer this question, it has been divided into four sub-questions.

RQ1: What are the factors that facilitate information diffusion in online social networks?

RQ2: How cascades networks can be constructed from minimal contextual information and missing/degraded information?

RQ3: What are the structural and temporal features of cascades?

RQ4: How is Tumblr, an online social network, used for information diffusion and what are the structural and temporal features of its cascades?

The first question aims to conceptualise the information diffusion phenomenon and draw some relations between its components. The second question aims to provide a cascade construction model that can handle situations that arise from the platform itself or the data harvesting phase. The third question surveys the features reported in the literature to describe cascades both structurally and temporally. The last question aims to quantify the structural and temporal properties of cascades on Tumblr and compare them with cascades obtained from the same or other social networks. The cascades allow the temporal and structural aspects of diffusion to be studied. Over 20 measures were used in the literature to quantify cascades. 15 of these were chosen here.

1.7 Research Contributions

This work's major quest is to provide a detailed analysis of the 'life story' of the popular content on Tumblr. Each of the research questions yields a different contribution. The contributions of this research are listed in Table 1.1.

Table 1.1 Research contributions

RQ1	An information diffusion Framework (IDF) that explains how actor factors, content factors, and platform affordances facilitate the spread of information.	Chapter 4
RQ2	A cascade construction model that yields accurate cascade networks from degraded/missing information and minimal contextual information.	Chapter 5
RQ3	A survey of the temporal and structural features of cascades and their implications.	Chapter 3
RQ4	A thorough analysis of Tumblr as a platform for content creation and sharing, including comparisons between Tumblr's main affordances.	Chapter 6 and Chapter 7
	An investigation of the popularity-gaining phenomenon from structural and temporal perspectives.	Chapter 6
	A comparison between Tumblr's top posts cascades and cascades in other OSN platforms.	Chapter 6 and Chapter 7

1.8 Terminologies Used

A number of terms are used throughout this thesis to describe information diffusion and its related phenomena. Many terms are used in the literature to describe studying the way information or content is spread, such as information diffusion, propagation, flow, and dissemination. In this thesis, the term *diffusion* refers to the phenomenon while *cascade* refers to the structural representation of the diffusion phenomenon. The cascade network is often perceived as a layer on top of the original social network, where users are linked to each other based on the direction of the flow of content between them. Table 1.2 lists other terminologies used here.

Table 1.2 Terminologies used in this thesis

Class	Term	Definition or reference
Diffusion verbs	Propagate	A term that describes the diffusion process.
	Spread	
	Diffuse	
	Share	
	Reblog	
Diffusion event	A diffusion event occurs when a piece of information is shared by one user with another. Might be: a reblog, a retweet. Generally, it is a term used to refer to cases where a user is adopting or sharing a piece of content, e.g. a hashtag, a url, a tweet, a post, etc.	
Diffusion mechanism	Sharing	The functionality that is used to spread content in online social network platforms.
	Retweeting	
	Reblogging	
Diffused item	Content	Describes the item that is diffused.
	Information	
	Message	
	Story	
	Meme	'A unit of cultural inheritance' (Leskovec et al., 2009).
	Contagion	Used to describe an item that is widely spread (Myers & Leskovec, 2012).
Contextual terms	Influence	The effect one individual has on others.
	Homophily	A property of a group of individuals.
Content creation	Content generator	The user that creates content.
	Content originator	
	Author	
	Post/publish	The act of creating content to be published.
Content sharing	Resharer	The person who spreads the content.
	Reblogger	

1.9 Outline of This Thesis

Chapter 2 sets the building blocks, and provides an overview of information diffusion and its components. Chapter 3 addresses cascades, discussing cascade definitions, how to construct cascades, and the structural and temporal features of cascades. Chapter 4 proposes the Information Diffusion Framework IDF, its components, and the relationship between them. Chapter 5 outlines the research methodology followed. It also includes information about the experimental settings and design, such as background about Tumblr, the dataset harvesting process, dataset cleansing, and cascades' networks construction. Chapter 6 presents the analyses conducted and their results, reflecting on those found on other platforms. Chapter 7 discusses some of the key concepts that emerged after analysing Tumblr cascades. The final chapter discusses the conclusions and future work directions.

Chapter 2: Information Diffusion: Background

> '… the Internet started out as nothing more than a giant Bulletin Board System (BBS) that allowed users to exchange software, data, messages, and news with each other.'
>
> (Kaplan & Haenlein, 2010)

This chapter explains the information diffusion phenomenon. The first part discusses social networks and their online versions. The advantages and affordances of online social networks are addressed highlighting those that affect diffusion. The second part focuses on information diffusion, a phenomenon that often occurs in online social networks when a piece of content spreads to many users. Information diffusion is defined, and what motivates users to spread the content. This chapter concentrates on the content and the context, while cascades will be the focus of the next chapter.

2.1 The 'Connectedness' Phenomenon

Networks, in their general sense, are structures that consist of a set of nodes and links; links associate nodes with each other, encapsulating a specific type of a relationship between the two. In mathematical terms, networks are modelled as graphs with vertices and edges (Newman, 2010). The core concept of networks is their 'connectedness', a phenomenon that has been observed in fields such as Biology, Computer Science and Sociology, and arises from the flexibility of the definition (Easley & Kleinberg, 2010). A social network can be defined as a network where the nodes represent people and the links represent the relationships and interactions between them (Kempe et al., 2003; Newman, 2010). Examples of such relationships are: acquaintance, friendship, co-authors, co-workers, affiliation, family relationships, information exchange, etc. (Grabner-Kräuter, 2009). All of these networks link people and, via these links, people interact with each other for many purposes such as: talking to each other, information sharing, and collaboration.

Since its emergence, the Internet has created a venue for human-to-human social interaction. In fact, the demand for some form of social networking was raised early on. This was facilitated by different types of computer-mediated communication (CMC), where 'humans' communicated with each other via the 'instrumentality of computers' (Herring, 1996, p1). This

comes in many forms on the Internet such as instant messaging, emails, and chat rooms. CMC was the focus of much research in the effects of such communication on social systems.

The invention of the Web in 1989 added another dimension to communication on the Internet, providing a wide range of possibilities for human interaction (Berners-Lee, 2000; Anon, 2014). The advance of the Web (Web 2.0 in particular) offered a variety of applications that fundamentally changed the way users communicate such as wikis, blogs, RSS, podcasting, and social networks (Lai & Turban, 2008). Therefore, in addition to communication and collaboration, individuals began to contribute to the Web by adding user-generated content. That is what differentiates Web 2.0 from the previous Web (O'Reilly, 2005).

Online social networks have seen a popularity surge following the proliferation of Web 2.0 applications (Heidemann et al., 2012). However, their basic concept is not new. In fact, they merely emphasise the Internet's main purpose: facilitating the exchange of information between its users (Kaplan & Haenlein, 2010).

2.2 Online Social Networks

Many terms are used to describe online social networks (OSNs) in the literature, such as online social network, social networking service, and social network site. While their definitions differ slightly, these terms have been used interchangeably.

boyd and Ellison (2007) define social network sites as web-based services that enable their users to: (i) create a profile, (ii) create a list of designated connections with whom they wish to connect, and (iii) traverse the list of connections of others. This definition focuses on viewing OSNs as 'networks' of connections. It disregarded any functionality that accounted for content creation, content sharing, or any other task that the users can perform. Similarly, Adamic and Adar (2005) describe social networking services as ones that reveal the network by showing users how they are connected to each other.

The essence of online social networks is that they are virtual communities that allow rich human interaction (Grabner-Kräuter, 2009). Schneider et al. (2009) points out that OSNs are web-based sites that allow users to form communities according to common backgrounds, interests and activities. These users are encouraged to publish content in any form of multimedia and can interact with others in different ways. Their description highlights different aspects of OSNs from boyd and Ellison's definition. There is the sense of community, the content that is being shared, and interactions using the different functions and affordances that are provided by the platform.

Kaplan and Haenlein (2010) define social networking sites as applications that allow users to create personal profiles, connect with friends via these profiles, add different forms of multimedia to their profiles, and communicate with others using some form of messaging (e-mail or instant messaging). This definition aligns with boyd and Ellison (2007) and Adamic and Adar (2005) by focusing on the ability to create a profile and allowing 'friends' to view it. This definition emphasises the ability to communicate with others using messages and creating content.

All these definitions focus on different aspects of OSNs. To summarise and provide a working definition: online social networks are platforms that allow users to create profiles of themselves, connect with others via these profiles, create a list of friends, and a form of a virtual community. These lists of friends are traversable, thus allowing others to explore the network of connections around them. Users can publish various types of content and are able to interact with each other using a handful of built-in functions offered by the platform.

2.2.1 The Advantages of OSNs

Online social network platforms have reshaped social interactions, by dramatically changing the way users connect to each other, locate interesting information, express and share ideas, and even form communities (Agrawal et al., 2011). There is a significant amount of social interaction occurring online. Consequently, this has huge implications in the way scientists (especially social scientists) conduct their research. Rather than the time-consuming data collection methods; they can gather data on a large scale almost instantly from the many sources available online (Agrawal et al., 2011).

In addition, interactions on the Web model real-world interactions. Social networks thus provide a rich resource of online behavioural data (boyd & Ellison, 2007). These behaviours vary in their nature, but include the formation of new connections (Farajtabar et al., 2015), propagating content (Myers et al., 2012), and marking content as *favourite* (Cha et al., 2012). Data gathered from OSNs has provided a way to examine the theoretical models proposed to model the propagation of information and influence within social networks (Liben-Nowell & Kleinberg, 2008; Cha et al., 2010).

Online social networks offer interaction channels for large audiences; these channels facilitate the emergence of rapidly updated reflections on current events such as elections, natural disasters, and breaking news. Both the size of the audience and the feasibility of making quick responses and reflections make it possible for information to propagate within the population. Responses occur rapidly, and they end up reaching a wide audience; this increases the

chances for trends to strengthen, and enables the rapid formation of opinions (Taxidou & Fischer, 2014).

2.2.2 OSNs Affordances

OSNs platforms provide different functionalities that allow users to interact with one another. These functionalities determine what OSNs users can and cannot do; they can be seen as different flavours of interaction between the users. Each OSN offers a set of functions that differentiate it from other platforms in addition the main purpose that motivates its usage. For example, Instagram and Pinterest are for photo-sharing, but the former is for creators and the latter is for curators, while Twitter and Facebook are social networks that are mainly used to exchange text (Mittal et al., 2013). Tumblr, on the other hand, is a platform that combines blogging with social networking (Chang et al., 2014). Bik and Goldstein (2013) categorise OSNs according to three uses: curation, community, and creation. They argue that Pinterest is mainly used for curation, while Facebook falls somewhere between curation and community. Tumblr and Twitter fall between community and creation, as well as traditional blogging. However, they also differentiate between time demands, as Twitter, Pinterest and Tumblr require moderate time, Facebook requires minimal time, while traditional blogging is extensively time demanding. Although the baseline affordances and usage of all platforms are similar, the culture that evolves around them is different (boyd & Ellison, 2007). They attract different age groups with different interests, even though they are designed in a way that makes them accessible for anyone. For example, Pinterest is dominated by women (OnlineMBA, 2012; Mittal et al., 2013) while most Tumblr users are under the age of 25 (Chang et al., 2014).

OSNs provide two basic affordances, the ability to connect with other users, and the ability to create content. Besides these, each platform provides another set of affordances that distinguish them from each other. Two sets of functionalities offer these affordances; for connectivity, it is **friend** in Facebook, Path and Snapchat, and **follow** in Tumblr, Twitter, Pinterest and Instagram. For publishing content, it is **post** in Tumblr, **tweet** in Twitter, **Create a post** in Facebook, **upload a photo** in Instagram**, add a thought** in Path, and **snap** in Snapchat, **pin it** in Pinterest.

Besides the ability to connect with others and publish content, there are a number of other affordances that affect their behaviours. These include the ability to share content (using **retweet** in Twitter, **reblog** in Tumblr, **repath** in Path, **repin** in Pinterest, and **share** in Facebook). The ability to admire content by **liking** in Facebook, Tumblr, Instagram and recently Twitter (it was

favourite previously), and **adding emoticon** in Path. Users can also express their opinions by **adding comment**s in Tumblr, Twitter, Facebook, Instagram and Path.

To summarise, OSNs provide a variation of some of the following general features. Table 2.1 presents a comparison between online social networks' functionalities

1. **Publish user-generated content**

   Users' ability to create content allows them to be part of the action; they play a role and contribute to the on-going conversation. In OSNs, users are both creators and consumers of content (Grabner-Kräuter, 2009). Some platforms (Twitter, Facebook, Tumblr, Path) allow users to publish different types of content (multimedia) including text, photos and videos, while others restrict their users to publishing special content such as photos and videos as do Snapchat, Instagram and Pinterest.

2. **Visibility of profile**

   The platforms differ in the visibility of their profiles; most of them allow users to have either a public or a private account. However, some platforms, such as Path, provide private profiles by default.

3. **Accessibility**

   Some platforms are only accessible through mobile applications, such as Path, Snapchat, and Whisper, while other platforms can be accessed using both the Web and the mobile applications.

4. **Connecting to Others**

   The ability to follow users means that a user will connect to others, and thus be exposed to stimulating ideas that will eventually impact her future interactions. Conversely, by being followed, users create an audience that consists of a set of users who are willing to listen to what the user has to say. By being exposed to a large and interested audience, the user's content will have a higher probability of being shared or it might motivate other types of user behaviour such as comments or likes. Friendship in online social networks might represent a shared interest or trust (Mislove et al., 2007). It is possible that the users have never met in person. Ahn et al. (2007) argue that social relations on OSNs are easier to form and maintain compared to offline relations and they are not affected by inactivity (Ahn et al., 2007).

   'follow' and 'friend' are two famous types of social link (social relationship) in online social networks. The difference between the two is that the follow relationship (Tumblr, Instagram

and Twitter) is unidirectional, which means that if A is following B, then B may or may not follow A; the follow relationship does not imply that the relationship is reciprocal. A friend relationship (Facebook and Path) is bidirectional, so if A is a friend of B, then B is a friend of A. Snapchat's connections are implemented as non-reciprocal but they use 'Add friend' instead of follow. Some platforms, such as Whisper, do not allow users to have connections. However, the network of connections (the social network) is not static. For instance, Twitter's social network is highly dynamic; around 9% of the connections change a month (Myers & Leskovec, 2014).

5. Ability to traverse the list of connections

Some OSNs allow users to expose their social networks and to discover those of others. This might lead to creating a connection outside of their circle that was not possible without the platform (boyd & Ellison, 2007). For example, Twitter makes this list available unless the account is private, while other platforms (Facebook and Path) make this available for friends only. In general, Tumblr does not provide these lists but it does allow its users to create their own profile layout, so that some users choose to show the list of people they follow. For example, Snapchat does not provide a list of connections even for friends.

The decision to connect or not to connect to a user determines the type of content a user will be exposed to (Myers & Leskovec, 2014). Consequently, a users' list of connections has two related processes: it is where the content is published and through it content is spread (Kwak et al., 2010; Myers & Leskovec, 2014). Myers and Leskovec (2014) state that the dynamics of link creation and deletion is hugely affected by the flow of information in the social network. A user might decide to follow another user after seeing interesting content posted by her and shared by others. Or, a user might decide to drop the connection after seeing an uninteresting or an offensive content.

6. Messaging, commenting and liking

OSNs implement different functionalities to allow users to message each other privately and to comment on each other's posts. Some platforms allow users to direct some messages to specific users publicly, usually with the mentions (@) function. They also allow users to express their admiration by using likes (**favouring** previously on Twitter). Likes or favouring convey different meanings and are used for many purposes (Meier et al., 2014).

7. Exploring content

Users of online social networks are surrounded by streams of content. They have two options for content exposure: they can either selectively follow users, or they can follow topics using the notion of **hashtags** or **tags**. The way users discover new content to share differs from one platform to another, since the platforms employ various mechanisms for attention gathering and content promotions. Consequently, this affects the way content spreads.

Tags work as a mechanism of content categorisation; they gather users from different parts of the social network so they can observe, contribute or share content (boyd et al., 2010). Tagging is used for content promotion too, such as '**trends**' on Twitter and '**Trending**' on Tumblr. Although both Twitter and Tumblr use tagging for content promotion, the actual mechanism is completely different. On Tumblr, 'trending' and 'staff picks' gathers curated posts selected by Tumblr's staff, while 'trends' on Twitter lists the top hashtags (which can be tailored to the user's location). If a user clicks on a hashtag, he will see all of the tweets that have this hashtag in their text. Facebook, on the other hand, provides a more tailored 'trends' list that is populated using information about the pages that the user follows and his location[5].

Hashtags/tags are vital for content popularity as they allow users to discover content. As the popularity of these hashtags/tags increases, users become able to locate relevant and interesting content (Romero et al., 2011). Moreover, other publicity mechanisms such as Twitter's 'Trends', which features content about the most popular topics at the moment (hashtags), attracts users' attention and consequently increase the chances that content has of spreading (Lotan, 2011). Tumblr has other forms of aggregated popular content such as staff picks and trending.

8. Spreading content

The built-in propagation functions, such as reblog or retweet, are mechanisms that have made it nearly effortless for users to share content and to make it spread virally across the platform from one user to another. However, this functionality is not implemented by all platforms. For instance, Instagram users employ a third party application to repost content, while Snapchat allows users to send specific snaps to a single user or a group, but recipients cannot repost it into their own story to be seen by all their friends.

[5] *How does Facebook determine what topics are trending?,*
https://www.facebook.com/help/737806312958641
Accessed: 18 November 2016

Functions that allow content to be spread, such as 'retweet' and 'reblog', measure the user's content pass-along value, while others that are used for more direct communication, e.g. 'mention', measures the user's degree of engagement with others and the name-value of an individual (Cha et al., 2010).

Table 2.1 Comparison between online social networks' functionalities

	Tumblr	Twitter	Facebook	Blogsphere	Flickr
Connections	Follow	Follow	Friend	Through URLs	Follow
Basic Element	Post	Tweet	Status/URL	Post	Photo
Diffusion mechanism	Reblog	Retweet	Share, posting on Walls	Manual	Favourite/fan popularity
What diffused?	Post, URL, memes	Tweet, URL, memes	Status, URL, memes	URL, content	Photo
Availability of multiple copies	Yes	Yes	Yes	Yes	Yes
Sign of admiration	Like	Favourite	Like	N/A	Like
Metrics	#Follower #Following #Posts #Likes #reblogs	#Followers #Following #Tweets #mentions #retweets	#Friends #Likes #Post #Shares #Comments	N/A genuinely via platform	#Followers #Fans/Fav. #Views
Publicity/ promotional	tags	hashtags	hashtags	tags	tags

2.3 Information Diffusion

2.3.1 Definitions and Historical Background

The 'connectedness' phenomenon is applied not only at a structural level but also at a behavioural level. At this level, a user's behaviour will have an implicit effect on everyone else's behaviour and consequently the overall outcome (Easley & Kleinberg, 2010, p. 4). One of these behaviours is information diffusion where individuals' influence each other to share the same piece of information. The theoretical roots of the research on information diffusion stem from many disciplines, including economy (Easley & Kleinberg, 2010) and communication studies. Everett Rogers, a communication studies professor, published a book "Diffusion of Innovations", where he explains how innovation spreads (2003). The term innovation in the title refers to any new idea, behaviour or technology and the theory explains how the diffusion of such innovation occurs. In the online social networks context, a behavioural change is said to spread each time an individual spreads information under the influence of others. Rogers (2003) states that "Diffusion

is the process by which an innovation is communicated through certain channels over time among the members of a social system." This definition highlights four main elements in the diffusion process, which are:

i. The Innovation: the idea, behaviour or object that diffused.
ii. Social channels: the communication channels through which an innovation is diffused.
iii. Time: the time it takes for an innovation to diffuse.
iv. Social system: the space in which an innovation diffuses.

The innovation that Rogers is referring to could be an idea, or a behaviour or an object that is considered as an alternative to current ones, where this alternative is new but might not be better than older ones (Agrawal et al., 2011). Rogers's definition implies that diffusion is a process that sheds light on the social change process which occurs regardless of the innovation type, who is adopting that innovation, and where it occurs (Agrawal et al., 2011). Rogers's definition has been adopted to describe the spread of information after the emergence of online social networks. In such contexts, information diffusion becomes more precise as to what diffuses and through what channels. Hence, within the context of online social networks information diffusion is defined as the process by which a piece of information is spread within social contexts, in environments that are either open (Myers, Zhu & Leskovec, 2012) or closed (Guille & Hacid, 2012), i.e. with or without taking external exposures into consideration.

Romero et al. (2011) state that the study of online information diffusion aims to understand the hidden tendencies and motivations of individuals to be involved in different activities within the various social network platforms, such as forwarding information, favouring and liking messages and photos, and joining communities.

Taxidou and Fischer (2014) defined the study of information diffusion as the study of the means of tracing, understanding, and predicting the way a piece of information spreads. This highlights two of the main purposes of research that have been tackled, namely: to understand and analyse existing flows, and to predict future ones.

Zafarani et al. (2014) state that information diffusion is a process in which information propagates through a population by one or more means of interaction. They identified three elements in this process: the sender(s), the receiver(s), and the medium of interaction, in which a piece of information is spread from senders to receivers.

The definition adopted here for online social network platform diffusion is:

> A **process** that consists of a number of **diffusion events**, each of which involves two users: a **sender** and a **receiver**. The receiver receives a **message** from the sender at a specific **time**. The sending and receiving process is codified by a **functionality** provided by an **online social network system**.

Zafarani et al. (2014) classify information diffusion into four types depending on the level of network observability and information availability. These are: Herd Behaviour, Information Cascades, Diffusion of Innovation, and Epidemics. They differentiate between **explicit** networks (where the network is observable, and hence the interactions between individuals are observable), and **implicit** networks (where the network is unobservable but there are other indicators that can be used to estimate the diffusion of products or diseases within the population). Thus, Diffusion of Innovation and Epidemics occur in implicit networks, while Herd Behaviour and Information Cascades occur in explicit networks. The difference between Herd Behaviour and Information Cascades is the type of information available for users to make a decision about adopting behaviour. Herd Behaviour relies on **global information** involving the whole population, while Information Cascades rely on **local information** about the immediate neighbours' decisions.

One of the earliest empirical studies was the study of diffusion of innovation by Ryan and Gross (1943), They interviewed farmers from Iowa asking them about their adoption of a new hybrid seed corn. The objective of their study was to draw some conclusions about the farmers' decisions regarding whether or not to adopt the new seed. Another early study (Coleman et al., 1957) researched the diffusion of a new drug adoption among physicians. The study concluded that peers had a significant effect on the adoption of the drug.

Researchers in the past faced the challenge of physical barriers that made research that involved detecting and analysing social interactions a difficult and expensive task (Gruhl et al., 2004). In addition, most historic studies investigated the diffusion and adoption of products, ideas and technologies over time (Goel et al., 2015b).

The history of information diffusion research began with empirical studies (Coleman et al., 1957), followed by studies that focused on mathematical modelling of the diffusion phenomenon. With online social networks, researchers are finally able to test their theories as they occur within the boundaries of these platforms.

2.3.2 Why Users Spread Information

Disseminating interesting information seems to be more of a natural habit of humankind. This behaviour is apparent in everyday life long before online social networks even existed. A few attempts have been made to investigate the motivation behind the diffusion behaviour.

Jackson (2010) argues that there are two main reasons behind diffusion in general: (i) social influence, and (ii) homophily. Interestingly, the two reasons are highly correlated with each other and can hardly be distinguished (Crandall et al., 2008; Shalizi & Thomas, 2011). Jackson explained this correlation with an example: assume that a person buys a new product soon after one of his friends did so. Is it then possible to conclude that the first one influences the other or is it because they are simply friends and both have the same interests? Building on that, it is important to differentiate between two forms of social influence: the first arises from family, peers and friends, and the second is influenced by celebrities, leaders and public figures (Huffaker et al., 2011).

Kelman (1958) proposed three different attitudes that individuals may adopt toward attempts at social influence. The first is compliance, in which an individual accepts being influenced, agrees publicly and hides his or her own opinions. The goals achieved by showing agreement are far more satisfying than disagreement or ignoring. The second attitude is identification, where an individual accepts another's influence to maintain a relationship with him or her. The third attitude is internalisation, where an individual accepts the influence of another because it is rewarding.

Anger and Kittl (2011) reflect on those three behaviours within the context of online social networks, specifically on Twitter. They stated that compliance is when a user shows public agreement, e.g. as a retweet, if the forwarded content seems valuable to establish his or her social reputation and gain popularity. Identification occurs when a user follows a liked and respected person; thus, the user will interact with that person because of his status not because of the nature of his content. Internalisation means that a user accepts a belief or behaviour due to the published content value. Anger and Kittl added two more attitudes to the framework: neglect and disagreement. The first implies that the content will be ignored, while the second implies that the user will publicly expresses his or her disagreement using mentions or even **unfollow** the person.

Easley and Kleinberg (2010) tried to explain why people imitate and follow other people's decisions. They argued that people tend to conform by behaving the same way that others do, but differentiated between informational effects and direct-benefit effects. Informational-effects occur when A (who has some private information) observes other people's decisions and assumes

that they have their own private information, which A will try to infer. In this case, information cascade will happen and, at the end, the decisions of the majority will be based on little genuine information. Direct-benefit effects occur when someone will gain a direct benefit if he aligns his decisions with those of others.

boyd et al. (2010) argue that retweets convey different messages and their meanings are not as straightforward as using a hashtag or mentioning a user. However, they discuss the different incentives for users on Twitter to retweet, some of which include:

1. To amplify content and spread it to the user's followers, potentially for followers that were not yet exposed to it.
2. To popularise users or content.
3. To grasp the attention of a specific audience.
4. As a conversation starter, by retweeting and adding comments.
5. To emphasis the user's presence.
6. To Agree with the author.
7. As a friendly gesture (sometimes a retweet is requested).
8. To gain benefits: followers or reciprocity from popular users.

2.3.3 Challenges of Studying Information Diffusion

Guille et al. (2013) cite three main assumptions in information diffusion research: (i) social networks are considered as closed-world while influence might come from external sources; (ii) social networks' structure is static; (iii) cooperating and competing diffusions are not often analysed. Although most theories in social network research derive from well-known research in sociology, most research on online social networks is significantly broad. Recent work is mainly empirical analysis and provides statistics, with very shallow projection of human behaviour and the way that people react to sociological stimuli.

Gomez Rodriguez et al. (2010) argue there are two challenges to studying information diffusion. First, tracing diffusions as they occur on the social network, and secondly, identifying the information that spreads within social networks.

2.3.4 Information Diffusion Models

Information Diffusion models are used to describe the mechanism of diffusion, either by investigating the infectious properties of the item that spreads and the properties of the social links that facilitate the spread, or the adoption timeframe. Researchers aim at modelling the

diffusion in order to reproduce them using observations drawn empirically or to facilitate prediction of the future growth of diffusion processes (Galuba & Aberer, 2010).

The next subsections address three of the most popular models. Goel et al. (2013) listed a number of other information diffusion models (generative models): Bass (1969), Watts (2002), and Dodds et al. (2003).

2.3.4.1 Probabilistic Models in Epidemiology

Probabilistic models are derived from epidemiology, which studies the propagation of diseases in populations, in which an infected person transfers the infection to others with a probability, causing the infection to propagate. This model follows the cycle of diseases in a host. The general scheme has four stages, starting from a node being susceptible (S) to an infection. It then becomes infected (I) after being exposed (E) to an infection. After that, the node becomes recovered/removed (R) (Gruhl *et al.*, 2004). Sometimes, the node is immune, and thus never gets an infection. The transitions from one stage to another are governed by two probabilities: the first is the probability that a node infects its neighbour, and the second is the probability that an infected node becomes uninfected. The epidemic threshold depends on those two probabilities (Easley & Kleinberg, 2010). This model has many variations such as SIR, SIS, SEIR and SIRS.

Terminologies drawn from epidemiological studies about the spread of infections within the population have been used to describe the diffusion process (Adar & Adamic, 2005). However, there are some differences between studying the spread of a diseases and studying the spread of information. Epidemic studies focus mostly on the speed and distance of the spread of an infection not on tracking it as it infects the population (Adar & Adamic, 2005).

2.3.4.2 Decision-Based Models/Models of Influence

i. Models of collective behaviour

a. Threshold models

This model, by Granovetter (1978), deals with situations where there are binary decisions and the person has to choose one, i.e. whether to join a riot or not. The idea is that each person has a threshold at which they will adopt a behaviour (become active) if the number of people who have adopted it exceeds that threshold. If the threshold is small, then that person is part of the early adopters, while if it is high then they will be a late adopter. This model does not take into consideration that some people are more influential than others and it treats all people as the same.

b. Cascade models

The independent cascade model introduced by Goldenberg et al. (2001b, 2001a) follows a simple process that starts with a set of nodes that are active already, i.e. have adopted the innovation. After being activated, each node tries to activate its neighbours in the subsequent step. This process is controlled by a probability chosen for the whole system independently without taking any history in consideration. Each node is given one chance to activate its neighbours and the activation process continues until no more activation is feasible.

c. A generalised model

In their attempt to study the problem of identifying the most influential users who would create a massive cascade, Kempe et al. (2003) introduced a general framework that covers the threshold and cascade models as special cases. They generalised the threshold model by defining a threshold function in which a node will become active if its activation function exceeds its threshold. They generalised the cascade model by allowing the probability that a node activates a neighbour X to depend on X's neighbours who already tried. They proved mathematically that these two models are equivalent. Their models assume that the probabilities of influence between the users are predefined. They studied the spread of influence in social networks.

ii. Game-theoretic model: Networked coordination game

This model addresses the case where nodes need to decide whether or not to adopt an idea, taking into consideration the decisions made by their neighbours. One of its simplest cases is when two neighbours have two options to adopt. If both nodes choose the same option, their payoff (benefit) will increase, while if their options are different, their payoff will be zero. In larger networks, where each node has many neighbours, the same game will be played for each neighbour and the payoff will be the sum of individual payoffs. The node will choose the option that yields the highest payoff. The complexity of this model increases with the number of players and available options (Easley & Kleinberg, 2010).

2.4 Information Diffusion Components

Guille et al. (2013) outlined a number of interesting issues in the literature about diffusion: (i) the type of information that spreads, i.e. detecting popular topics; (ii) the way in which information spreads, i.e. modelling the diffusion; and (iii) the role of people in the spread of information, i.e. identifying individuals with influence. Building on that, for each information diffusion event there are three components: content, context and cascades (Figure 2.1).

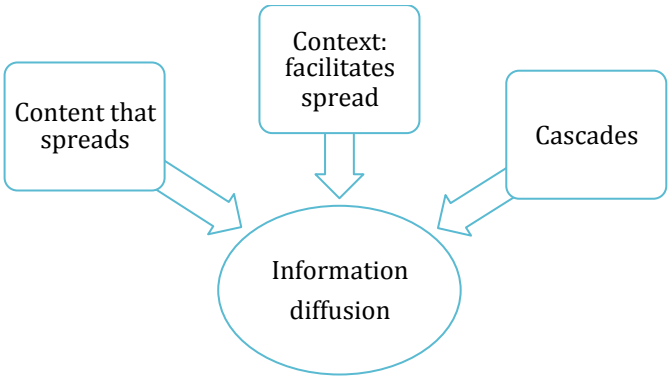


Figure 2.1 Components of the information diffusion process

2.5 Content

The literature cites many types of content as targets to be tracked and analysed, depending on the platform. Most early diffusion studies tracked URLs because they are easily tracked within blog posts, it is rare that they have multiple variations (Adar & Adamic, 2005), and they are language-independent (Galuba & Aberer, 2010). Leskovec et al. (2007, 2006) studied the diffusion in **blogosphere** as manifested by the existence of links to other posts in different blogs. No particular content was taken into consideration, just the analysis of the cascading behaviour between bloggers.

Many items can potentially diffuse as a 'contagion' in online social networks, such as URLs, pictures, text, hashtags, and memes. Two categories of content have been studied. The first category includes generic items such as URLs and hashtags. In such cases, many users are involved in the diffusion process; researchers have studied the spread of such generic items across the social network. In contrast, the second category involves content that diffuses as individual messages created by a single person and spreads via one of the information diffusion functionalities provided by the social network platform. Some items are harder to trace and analyse than others, because they can take different forms. For instance, URLs are more explicit, easier to detect, and less ambiguous than raw text or memes (Leskovec et al., 2007b). With URLs, however, it is not clear who influenced whom, even if timestamps and social network structures are used to identify diffusion. Gomez-Rodriguez et al. (2010) argued that identifying the contagion to trace it successfully is one of the fundamental challenges of studying diffusion. Memes have been targets for many diffusion-related studies, due to their nature as content that encourages users to repost them; thus memes often travel far within a social network (Adamic et al., 2012).

A small number of studies have taken content into consideration; for example, Webberley et al. (2013) inferred the 'interestingness' of tweets based on their retweet history. They argued that this interestingness has an effect on a tweet's retweetability. Bakshy et al. (2011) studied the

role of content on cascades. They used Amazon Mechanical Turk (an Internet skills marketplace that utilises crowdsourcing) to rate URLs according to six features: interestingness, perceived interestingness, positive feeling, willingness to share via email, URL type, and category. André et al. (2012) asked volunteers to rate tweets; they found that only 36 percent of tweets were actually worthy of being retweeted, and 25 percent were not worthy, while the rest were in between.

The 'retweetability' of tweets was estimated using content-related features. The fact that a tweet contains a hashtag, a mention and URLs, proved significant when predicting the retweetability (Suh et al., 2010). In fact, both Suh et al., (2010) and Petrovic et al. (2011) show that the number of followers and the number of friends (followees) are highly correlated with the retweetability of tweets. Petrovic et al. (2011) added membership in twitter lists as a successful predictor of retweetability. They state that without considering the content, the number of followers and friends are enough to predict the retweetability. They concluded that retweets are mostly about the person who retweeted, not what has been tweeted.

Other research has studied meme mutation by tracing text that users on Twitter quote in their tweets (Leskovec et al., 2009; Simmons et al., 2011). Meme mutation or the 'evolution of memes' studies fall into large-scale analysis (Adamic et al., 2016; Leskovec et al., 2009; Simmons et al., 2011; Romero et al., 2011; Cheng et al., 2016), and small-scale ones (Liben-Nowell & Kleinberg, 2008; Adar & Adamic, 2005)

Adamic et al. (2016) argue that textual mutations are easily tracked compared to photos or videos because they can be easily added to the content. They used biological analogy of genetic evolution to model meme evolution. However, studying meme mutation requires extensive access to the data that allows observation of different copies of the same content, as they are introduced by many users on the platform. For instance, Cheng et al. (2016) studied memes on Facebook taking into account multiple introduction of the same meme to the social network.

Other types of content have been studied. For example, Gruhl et al. (2004) studied the diffusion of topics in blogs. Their analysis required the identification of topics and terms as a first step towards analysing the diffusion of topics.

One of the few examples that analyse the popularity of Flickr's images using detailed features of images, such as colour and gradient, is the work by Khosla et al. (2014). In fact, they are among the few who combine content features with social context to predict image popularity.

2.6 Context

In any diffusion, the context that surrounds the users involved in the process has a huge impact on it. The word 'context' includes all the factors that are not related to the content of the message. These factors relate to the surroundings of the user: where does he reside in the social network? Who is he connected to? Do they have influence over him? Other factors include the time at which the message was published, the number of times a user was exposed to the message, its language and its geographical location. Many researchers have been looking in-depth into some of these factors. Research on influence and the impact of the social network structure have received much attention, followed by homophily. Earlier research identified three possible reasons for users to share the same piece of information. One user has *influence* so the other will share the influencer's information; or users might be interested in the same topics (*homophilous*); or there might be some *confounding factors* as they might have access to the same resources (Anagnostopoulos et al., 2008; Bakshy et al., 2012). However, it is challenging to identify the true reason of a diffusion (influence vs. homophily), as these three reasons collectively affect the possibility of sharing the same information (Aral et al., 2009). For example, Bakshy et al. (2012) studied the sharing behaviours of Facebook users experimentally by randomising the exposure to information in the users' feeds. They concluded that repeated exposure increases users' chances of sharing the same information. They also stated that weak ties (Granovetter, 1973), which they defined using the number of interactions users have, play a major role as resources for new information. Additionally, when they omit the possibility of social influence, users share the same information which might be the result of external exposure to the same resource or to homophily or both.

2.6.1 Social Network Structure

Network structure plays a major role in the overall performance of the network, especially if the purpose is to understand the way information spreads across it. Social network structure affects the way content is shared (Mislove et al., 2007). It is crucial to the propagation of information as it might amplify or weaken those flows (Adar & Adamic, 2005). Kempe et al. (2003) argued that social networks have a huge impact on the diffusion process, and determine whether a message or a product will be successfully adopted by the users or not.

It is often assumed that if one user follows another, the former will be exposed to any content that the followee publishes (Kwak et al., 2010). This might not be always true unless there are sufficient contextual data, such as clicks on feeds that can support such an assumption

(Bakshy et al., 2012). However, Galuba and Aberer (2010) state that around 33% of the retweets they encountered in their dataset were from users who do not follow the tweets' authors.

Many measures have been used to observe and analyse networks. Some of these measures quantify specific nodes' or edges' properties and others can be generalised for the whole network. In social networks studies, the term centrality is often used to refer to the measures used to quantify the structural importance of a node within a network (Borgatti & Everett, 2006), i.e. to measure to what extent a node is central to the network (Monge & Contractor, 2003). Centrality measures range from the basic measure of degree to more complex measures, such as betweenness, closeness, and eigenvector centrality.

Centrality

1. *Degree*

Degree is a node property, which is the number of links incident upon a particular node (Freeman, 1979). In online social networks, there are unidirectional links that create directed networks and bidirectional links that create undirected networks. Thus, when talking about the degree in directed networks, it is useful to specify whether it is in-degree (number of links directed into the node) or out-degree (number of links from the node). In undirected networks, the term degree is used loosely due to the lack of direction. Degree is used to demonstrate 'small-world' property in networks by utilising degree distribution curves (Newman, 2010).

The nodes' degree has a large impact in information diffusion contexts. For instance, from the information flow perspective, when novel information reaches a node with zero out-degree, it will not get any further (a sink in graph theory terminology). By contrast, in a diffusion graph, if the node has zero in-degree, it is considered a source. Additionally, nodes with high in-degrees are more prone to obtain information from various sources, while nodes with high out-degrees are capable of transferring information to many nodes. Thus out-degree measures the extent to which a node is influential.

2. *Betweenness*

As the name suggests, the betweenness of a node quantifies the number of shortest paths between nodes that pass through that node (Newman, 2010). Nodes with high betweenness values are sometimes called bridges (Hinz et al., 2011). Bridge nodes are characterised by their central location within the network that facilitates bridging different parts of the network. This means that a bridge node might be the source of novel information for the partial network speeding its propagation (Mochalova & Nanopoulos, 2013). On the other hand, the bridge node

could keep some information preventing its spread (Freeman, 1979). In addition, using the same notion, edge betweenness can be calculated, which is the number of shortest paths that pass through that edge. In such a case, high edge betweenness means that the edge acts as a weak tie that bridges different parts of the network. Based upon this intuition, Girvan and Newman (2002) proposed their famous algorithm to detect communities by repeatedly removing edges with high betweenness from the network, i.e. removing weak ties. The importance of weak ties and their impact on the overall connectivity of the network was first posited in 'The Strength of Weak Ties' (Granovetter, 1973).

3. *Closeness*

Freeman (1979) defined closeness as the summation of geodesic distance from a given node to all other nodes. Following this definition, closeness is an inverse measure of centrality and actually measures farness rather than nearness. This means that as the closeness value increases, if novel information reaches that node it will take a longer time to spread (Borgatti & Everett, 2006).

4. *Eigenvector*

Eigenvector is a variation of degree centrality introduced by Bonacich (1987), which measures the importance of a given node in a network taking into account the importance of the node's neighbours (Newman, 2010). Therefore, the eigenvector value of a node does not only depend on the node's neighbours but also upon the neighbours of neighbours (Mochalova & Nanopoulos, 2013). Google's page rank is a variant of an eigenvector measure.

Connected component

A network component is a subset of nodes that are connected (Monge & Contractor, 2003). A component is strongly connected if there is a path from each node in the subset to the others and *vice versa*, such that there is no larger set that satisfies this property (Easley & Kleinberg, 2010). A component is weakly connected if each node can be reached from any other node by following edges and ignoring their direction (Newman, 2010).

A giant component is a connected component that acquires a significant proportion of all nodes in a network and, as the network grows, the giant component preserves this property (Newman, 2010). Consequently, the giant component is crucial for information diffusion; as soon as novel information reaches any node in it, an epidemic might occur (Newman et al., 2006).

Communities

The clustering coefficient was suggested by Watts and Strogatz (1998) as a measure to identify networks that exhibit 'small-world' properties. It is calculated for each node in the network in order to examine the interconnectivity between its neighbouring nodes. In other words, it checks the existing connections between the neighbours of a given node to see how close they are to forming a clique around that node. In addition, cliques are formed if each node is connected to all other nodes in the network (Newman, 2010).

A high clustering coefficient could sometimes result in preventing novel information from travelling outside clusters. Keeling (1999) argues that when the clustering coefficient is high, only a few nodes need to be vaccinated to prevent epidemics.

Network properties

Real-world social networks are complex by nature; such complexity makes the analysis process more challenging. However, they exhibit a number of properties that enable the design of mathematical models to simulate real-world networks. These models are designed to preserve and mimic the properties encountered in real-world networks. This facilitates an efficient analysis of various phenomena that can be observed and analysed (Zafarani et al., 2014). The two central properties of social networks are that they exhibit 'small-world' features and that they are 'scale-free' networks. These properties provide an understanding of the underlying structure that online social networks follow. The structure is important to understand information diffusion as it specifies the information an individual will be exposed to and possible directions for information flow (Myers & Leskovec, 2014). However, network properties do not explain the information diffusion mechanism that takes place across them.

'Small-world' is a property that renders large social networks connected tightly in a way that makes any two nodes that seem to be far from each other actually very close. A network with the 'small-world' property has three features: (i) short paths exist between most of its nodes, (ii) it has high clustering co-efficient and cliques, (iii) the degrees of its nodes has a power law distribution (Watts & Strogatz, 1998). The term 'small-world' was established after Stanley Milligram's experiment in 1967: the 'Six Degrees of Separation' (Travers & Milgram, 1969). As the name reveals, the major finding of this famous experiment is that the average path length between any two living individuals on earth is six. Dodds et al., (2003) conducted a similar experiment but analysed the social network generated by email exchanges. They reported that the average path length in that network was also six.

'Scale-free' networks have three features: (i) the existence of hubs due to preferential attachment, (ii) its degree follows a power law distribution, (iii) it has 'small-world' properties.

Barabási and Albert (1999) analysed a portion of the World Wide Web network. They found that just a few websites (hubs) had more connections than others did. Ahn et al. (2007) argued that there are two origins for the power law property in social networks: preferential attachment (Barabási & Albert, 1999) and 'transitive linking' model (Davidsen et al., 2002), which studied 'triadic closure' in aquantenace networks. Ahn et al. (2007) suggested that in online social networks, attractive users can have many friends, and consequently it is easier for them to have more friends via transitive linking.

Social networks also tend to exhibit positive assortativity of the distribution of degree correlation (the correlation between the degree of a node and the mean degree of its neighbours). That is to say, hubs in social networks (nodes with high degree) tend to be linked to other hubs (Newman, 2004).

2.6.2 Homophily

The principle of Homophily stems from the idea that 'Similarity breeds connections' (McPherson et al., 2001). In their study of homophily in social networks, they argue that the rate of interaction is higher among pepole who are similar. Users are believed to be homophilous if they are interested in the same topics, and will thus share the same piece of information regardless of the social signals (Bakshy et al., 2012). Some research relates diffusion to homophily between friends rather than to influence (Anagnostopoulos et al., 2008).

Cha et al. (2009) reported that, most photos in Flickr get their favourite markings by users who are friends of the uploader, while users who mark the most popular photos as favourite are at most two hops away from the uploader. Jackson and López-Pintado (2013) proposed a model of diffusion that is based on epidemics model (SIS) and game theoretic models. Their model took into account biases in interaction and different tendencies towards adoption. They studied a case in which there were a small number of first adopters. They wanted to know whether the diffusion occurs in such case in heterogeneous and homophilous populations. They found that homophily facilitates diffusion, both in groups of those who have the tendency to adopt and in groups of those who do not. Also, they claimed that if there is less homophily in the population and if diffusion is facilitated by small seeds only, diffusion will not happen.

2.6.3 Influence

There has been ongoing discussion about the role of social influence on information diffusion. To some extent, it is hard to separate influence from diffusion, as diffusion is more or less an indicator of influence, and influence causes the dissemination of information (Granovetter, 1978).

Many researchers have explicitly studied influence, based on the features of individuals or communities, while others have attempted to explain diffusion without explicitly discussing influence. Leskovec et al. (2006) define information cascades as phenomena that occur when people adopt an idea or behaviour under the influence of others. Kempe et al. (2003) used the spread of influence to refer to the diffusion process, which implies that most of the diffusion processes are influence driven.

Although online social networks provide vast amounts of data, identifying influence and influencers is still a challenging task (Cha *et al.*, 2010; Bakshy *et al.*, 2012). This difficulty arises from identifying influencers, and of detecting the nature of influence, i.e. knowing that an action was taken under social influences. However, Bakshy et al. (2012) argued that it is very difficult to identify the cause of diffusion whether it is influence or homophily.

Influence is assumed to be the cause of diffusion; in Taxidou and Fischer (2014) influence paths were derived using the social graph. Hence, an influence model (Bakshy et al., 2011; Cha et al., 2010) was adapted to link users with various categories of influencers: the least recent influencer, the most recent influencer, the most followed influencer, or the most retweeted influencer.

Choobdar et al. (2015) argued that there are two roles for users in diffusion; they can either be influencers or blockages. They studied the correlation between users' structural characteristics (either theirs or those of their neighbours) and their roles as influencers or blockages. They consider the number of votes a user's story receives, as a measure of influence. They also measure blockage using the number of stories a user is exposed to but has not voted for. However, for both roles the authors assumed that a user is exposed to a story as long as he is friend with the one who posted it.

Table 2.2 summarises some of the measures used to measure influence in the literature; it shows that measures related to retweets and cascades were used as an indication of the user's influence.

Table 2.2 Measures of influence

Followers influence/structural	In-degree (Number of followers)	(Kwak et al., 2010) (Cha et al., 2010) (Ye & Wu, 2010) (Bakshy et al., 2011) (Taxidou & Fischer, 2014) (Lee et al., 2010)
	Followers to following ratio	(Anger & Kittl, 2011)
	PageRank	(Kwak et al., 2010) (Ghosh & Lerman, 2012)
	Centrality	(Ghosh & Lerman, 2012)
Retweet influence	Number of retweets	(Kwak et al., 2010) (Cha et al., 2010) (Lee et al., 2010) (Ye & Wu, 2010) (Taxidou & Fischer, 2014)
	Number of users who retweet	(Ye & Wu, 2010)
Cascade	Size of cascade	(Bakshy et al., 2011)
Replies/mentions influence	Number of replies	(Ye & Wu, 2010)
	Number of mentions	(Cha et al., 2010) (Lee et al., 2010)
	Number of users who replied	(Ye & Wu, 2010)
	Retweet to mention ratio	(Anger & Kittl, 2011)
	Interactor ratio (ratio of all the followers who interacted with a user)	(Anger & Kittl, 2011)
General	Number of tweets	(Bakshy et al., 2011)
	Joining date	
	Frequency of contact	(Bakshy et al., 2012)

2.7 Chapter Summary

This chapter presented background of online social networks and their affordances, reflecting on those that affect information diffusion. The second part reviewed the literature on *information* diffusion including its components: the content and the context. The next chapter will review cascades, the third component of the information diffusion process.

Chapter 3: Cascades

> 'The success of any kind of social epidemic is heavily dependent on the involvement of people with a particular and rare set of social gifts.'
>
> (Gladwell, 2000)

The previous chapter introduced information diffusion, a process by which information 'spreads' on social networks. This chapter will delve into cascades, the third component of information diffusion and the outcome of the process. First, cascades are defined, as well as their significance and the purpose of analysing them. Then, the different methods used for their construction are highlighted. A survey of cascade features gathered from published literature will be discussed in detail emphasising their significance as descriptors of cascades. This chapter concludes by presenting some of the most debated topics: cascades' size and virality.

3.1 What is a 'Cascade'?

3.1.1 Definitions

For economists, *information cascade* occurs when an individual decides that it is optimal to follow the behaviour of those before him after observing their behaviour, without taking into account his own information (Bikhchandani et al., 1992).

The term 'cascade' was picked by researchers to describe a similar phenomenon that has been observed in OSNs. For instance, a cascade as defined by Goel et al., (2012), comprises a seed individual who shares an item of information independently from any other individual, followed by other individuals who are influenced by the seed to share the same information. A definition by Leskovec et al., (2006) is that cascades are phenomena caused by individuals' influence in which an action or idea becomes widely adopted by others (Goldenberg et al., 2001a; Granovetter, 1978); hence, they are known as 'fads' (Bikhchandani et al., 1992). Cascades are amplified on OSNs by built-in mechanisms that allow users to share content while crediting the source or the person who shares it (boyd et al., 2010).

In the case of cascades, messages travel through the social network links from one user to another (Kwak et al., 2010). When gathered, the paths that these messages travel create a network that resides as a layer on top of the social network. These networks are the cascade

networks and the paths messages take are often called information paths in the literature (Gomez Rodriguez et al., 2013).

Leskovec et al. (2007, 2006) define a cascade as a tree that has a single root (the cascade initiator) that has links to other nodes. Further nodes can be added by linking to the existing nodes in the cascade. All of the added links follow a strict time order.

3.1.2 Significance of Cascades

To understand cascades, it is essential to understand the way information propagated on the Internet (Dow et al., 2013). Additionally, the spread of a message gives a lot of information about the users involved in the process. Obviously users have limited attention, so a successful cascade is the one that gets the most attention across the competing cascades at a particular moment (Weng et al., 2012; Myers & Leskovec, 2012). However, Dow et al. (2013) state that a user who is repeatedly exposed to a particular item by his friends increases the chances of the user sharing it further. They argued that in such a case, these users are subject to both influence and homophily (Bakshy et al., 2012); repeated exposure increases the influence factor, and being surrounded by a group of users who are susceptible to an item means that the user himself is susceptible too. Hence, the paths that information takes to reach individuals have been called influence paths in many studies, as they directly indicate that a user influenced another to spread the message.

Bild et al. (2015) refer to cascade networks as implicit networks because they are constructed using a subset of the social network, which they define as an explicit network. They argue that analysing cascade networks is important as these 'implicit' networks can serve as an accurate indication of interest or trust relationships. They conjecture that cascade networks model real-world social, interest and trust networks better than the social network. They argue that connections on the social network (follow/friend) entail that users are willing to listen to each other, but connections on the cascade network are better indicators because they are created using a forceful sharing action that pushes the content to the user's list of friends.

Analysing cascades can help detect network evolution and link creation, since users often create new links (follow/friend new users) after being exposed to novel information sources. Myers and Leskovec (2014) studied the relation between cascades and the creation of new links in the social network. They related the sudden bursts of connectivity to the dynamics of sharing on Twitter. Antoniades and Dovrolis (2015) used the number of retweets and follow reciprocity to model link formation. They studied link removal dynamics on Twitter after reading a tweet or receiving a retweet from the user. Farajtabar et al. (2015) introduced a model that takes into account both activities (sharing and link creation).

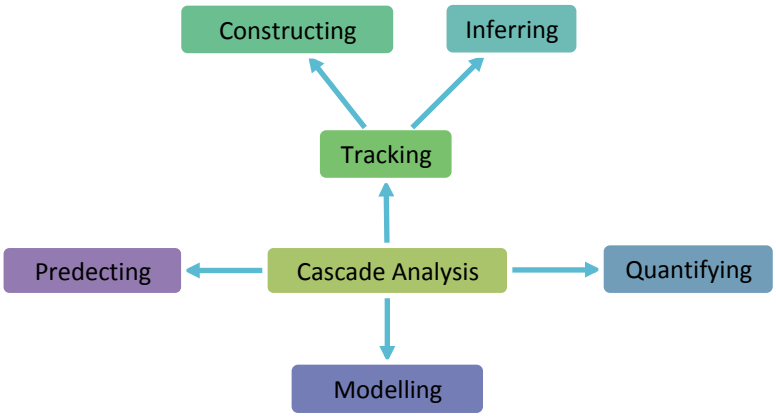


Figure 3.1 Four perspectives for cascade analysis

3.2 Purposes for Studying Cascades

Cascade studies' purposes vary depending on the objective of the study and the data available for the researchers. Throughout the years, and the different platforms that have been studied, the purposes have ranged from merely observing and quantifying cascades, to tracking them, predicting information flows, and modelling them (Gruhl et al., 2004).

Figure 3.1 illustrates the four general perspectives for studying cascades. The first, and essential aim, is tracking existing cascades and either constructing them or inferring them. The ability to construct a cascade depends solely on the data available for its construction. The second perspective focuses on quantifying cascades, structurally (Dow et al., 2013), temporally (Gruhl et al., 2004), or just numerically, combined with some platform dependent measures (Bakshy et al., 2012). Structural analysis of cascades often requires constructing the cascade first, before the analysis. Some studies focused on analysing cascades quantitatively thus did not attempt to construct cascade networks as the structure of the cascades was not their aim, e.g. Bakshy et al., (2012).

The third perspective looks at modelling the cascade, i.e. using generative algorithms to create cascade networks using the characteristics observed from the tracked networks (Gruhl et al., 2004; Liben-Nowell & Kleinberg, 2008). The fourth perspective investigates predictions such as the likelihood that a piece of information will be shared in the first place (Petrovic et al., 2011), or the possibility that a popular piece of content will continue to be popular (Ma et al., 2013), or predicting the future growth of a cascade (Cheng et al., 2014). Most of the time, a study incorporates one or more purposes in its analysis.

3.2.1.1 Modelling

A number of studies looked at modelling cascades or aspects related to it. Gruhl et al., (2004) modelled the temporal patterns of the diffusion of topics in blogspace and individual propagation. They categorise topics into chatter, spikes and mixed, based on the number of blog posts written across time. Individuals were categorised according to their position within the lifecycle of a topic: Ramp-up, ramp-down, mid-high, and spike. Their model was inspired by epidemics; users who post about a topic become infected by it, and they spread this infection to others. Gruhl et al.'s analysis was mainly focused on the temporal patterns and modelling, so cascade construction and analysis of its structure were not used.

Leskovec et al. (2007, 2006a) proposed a conceptual model that generates cascade networks. Their model is similar to the SIS model from epidemiology, in which an infected blog will infect each of its neighbours with a given probability.

Some research has modelled using actual information flow, in order to gain insights from the structure of the diffusion graph. Liben-Nowell and Kleinberg (2008) constructed trees that represent the propagation of large-scale Internet chain letters. They used a probabilistic model to produce cascade trees similar to the ones captured in their analysis of chain letters. These trees are deep and narrow, reaching individuals who are several hundreds of steps away from the root.

Lerman and Ghosh (2010) analysed the spread of news in two different social networks, Digg and Twitter. The means of spread is via votes in Digg and via retweets in Twitter. Their analysis compared temporal aspects in both sites. They found that newly-posted stories on Digg spread quickly, getting large numbers of votes (mostly from the submitter's friends). After being promoted to Digg's front page, the spread of stories on Digg tended to slow down and saturate shortly thereafter. Stories on Twitter spread more slowly than those on Digg but reached farther. Lerman and Ghosh suggested the reason for this difference in spreading patterns is that Digg's network is denser and more connected than Twitter's. They claim that, as time passes, Twitter's network will become denser as more people join the network. They also claim that story popularity (votes per story and retweets per story) follows a normal distribution rather than power-law distribution.

3.2.1.2 Inferring

Sometimes information is diffused without an easy-to-follow flow. For example, several blogs might share a URL without referring to the contents' originator. In such cases, the only way to study the flow of information is by inferring the network's structure, which would have led to this diffusion. Several attempts have been made to infer the flow of information, particularly in the

blogosphere. Gruhl et al. (2004) proposed an iterative algorithm to induce topical transmission graphs in blogs using posting time as an indicator to link two blogs. This induction algorithm is based on a closed-world assumption, in which all posts about the same topic, except the first one, are the result of social influence.

Another attempt was by Adar and Adamic (2005) who aimed to infer the propagation of URLs in blog networks. The goal of these researchers was to correctly classify and label the links between blogs in order to visualise them using what they called "infection trees". They implemented link inference, to classify links between blogs, and infection inference, to classify plausible links for infections.

Liben-Nowell and Kleinberg (2008), studying chain letters, incorporated an inference task in addition to constructing cascades from actual data. The reason is that they often encountered cases where there was either insertion, deletion or alteration to the list of users who signed the petitions. Thus, to be able to create a cascade tree, they first created a complex network then deleted some edges to retain the tree structure using the number of copies that mentioned that edge as evidence to its existence in the cascade tree.

The problem of diffusion inference was further investigated by Gomez-Rodriguez et al., (2010) who developed an approximation algorithm (NETINF) that infers near-optimal diffusion networks by tracing paths of diffusion and influence. For each different cascade, the algorithm uses data from times when nodes are infected, one after another. Thus, the algorithm finds the likelihood of one node influencing another node. It assumes that each node has one parent only, i.e. that each node is influenced by only one other node and the network is static. They validated the algorithm on synthetic and real data collected from blogs and mainstream media websites, concluding that most information tends to propagate from media to blogs, and that the links between media websites are the strongest. Links between media websites are thus detected early by NETINF.

Most inference-related research was on blogosphere, which might parallel the lack of conventional social connections similar to those existing in new online social network platforms.

3.2.1.3 Predicting

Predicting future diffusions may be based on temporal measures (which link is going to disseminate the information first?), and on structural measures (what are the features of nodes and links that would allow the information to propagate?).

Yang and Counts (2010) predicted the speed, scale and range of cascades on Twitter. The speed refers to how quickly it occurs, the scale is how many nodes are affected by the content's author, and the range is the number of hops in the cascade. They utilised features such as total number of tweets, and the number of mentions, and whether a tweet has a role as predictor of the cascades. They concluded that the rate of mentions is a better predictor for all three aspects.

Macskassy and Michelson (2011) identified four information-sharing cascades (retweeting models) that are used to compute the probability that a tweet will be retweeted:

i. General model: where users will retweet randomly.
ii. Recent Communication model: where users will retweet a tweet by a user who they've contacted recently (mention or direct message).
iii. Topic model: where users will retweet a topic of interest.
iv. Homophily model: where users will retweet a user based on his or her profile.

Yang et al. (2012) analysed three datasets collected from Twitter to predict future hashtag adoption by users. (This excludes two cases: when the user creates a hashtag or retweets it.) They selected several factors based on the assumption that there are two roles for hashtags in Twitter: as a way to tag context, and as a way to express community membership. The factors related to content tagging are relevance and preference, while prestige and influence are the factors related to community membership. They listed five non-role-specific factors for hashtags: popularity, length, age, activeness and degree of the user.

3.3 Aspects of Cascades

Cascades have two dimensions; the first is linked to the relation between the users involved in the cascade, i.e. who influenced whom to spread the content. The second is a time-series information about cascades that provides the number of diffusion events that occur at a given time. Each of these dimensions is related to a different aspect: the structural and the temporal. These aspects complement each other and provide a better understanding of cascades, as Scott (2008) argued that the temporal aspect adds value to the structural aspect.

3.3.1 Structural

Looking first at the structural (topological) properties of cascades includes studying their structure and quantifying cascade networks' properties. According to Liben-Nowell and Kleinberg (2008), a better understanding of the properties of the structure of cascades leads to better dissemination models.

Among the earliest work is that of Leskovec et al. (2006b) who studied the topology of cascade networks including their size and frequency of shapes, across different products' groups. Leskovec et al. (2007, 2006a) analysed the topology of post and blog networks taking into account the in-degree and out-degree and the network components. They also enumerate cascade networks' shapes and how frequently they are encountered.

The cascade sizes in these early studies are generally small, which explains the reason why their shapes were enumerated. With the emergence of the online social network platforms such as Twitter and Facebook, larger cascades began to be analysed, and new measures were used to quantify the structural aspect of cascades. Some of these measures are: scale and range (Yang & Counts, 2010), branching factors and subcascade sizes (Dow et al., 2013), height of cascade tree (Kwak et al., 2010), and diameter (Taxidou & Fischer, 2014).

3.3.2 Temporal

There are two approaches to the temporal aspect of cascades. The first tracks and describes existing cascades' temporal features, e.g. how fast information spreads, for how long trendy content keeps its popularity, and the overall growth of cascades over time, such as whether cascades show patterns like 'burstiness' or sparks. The other line of research uses a cascade's temporal patterns to either predict or model the cascade's future popularity. Most of these attempts do not mention the word cascade, because they are concerned about the temporal aspect of the diffusion of online content. The underpinning structure of online content diffusion is an implicit cascade network.

Gruhl et al. (2004) were among the first who studied the temporal aspects of information diffusion in blogs, tracking topic diffusion through time. They distinguished between two patterns for the diffusion of topics, chatter and spikes. Chatter refers to the steady and on-going discussions about a topic between bloggers, while spikes refers to the short periods of high-intensity volume of discussions. They categorise topics into chatter, spikes, and mixed, based on the number of blog posts written across time. Individuals, on the other hand, were categorised according to their position within the lifecycle of a topic as Ramp-up, ramp-down, mid-high, and spike.

Cheng et al. (2016) studied cascade recurrence on Facebook using large-scale data gathered over a year, that accounts for multiple introductions of the same content into Facebook. They studied the rise and fall of cascade popularity by deconstructing the growth timeline into periods of burstiness around peaks. Their analysis shows that recurrence occurs widely in large cascades,

while the temporal patterns of such cascades show periods of bursts and idleness too, due to fresh introductions into the network.

Borghol et al. (2011) modelled the popularity evolution of user-generated content on YouTube. However, their analysis and modelling was done on the assumption that popular content exhibits a simple pattern of rise and fall, meaning that their model only accounts for one peak during the popular content's lifetime.

3.4 Constructing a 'Cascade' Network

Within social networks, many sub-networks can be created using the same nodes that can be linked using edges with various meanings. As soon as information starts to spread within a population, another layer could be added on top of the original network that represents the flow of information (Gomez Rodriguez et al., 2013). This is often called a diffusion/propagation network or a cascade (Easley & Kleinberg, 2010). Using Twitter as an example, instead of creating a network of followers, we could create a network where each node represents a user and each link represents a retweet direction. Thus, if A retweeted a tweet posted by B, then there would be a link from B to A, creating what is known as a 'retweet network' (Yang et al., 2012). Or as it will be referred to a 'cascade network.'

In blogs, there are no built-in mechanisms for diffusion; thus, most of the early studies on blogs used various features to infer cascade networks. Adar and Adamic, (2005) added a link between two blogs if there is an explicit link to the other. If there is no explicit link, they infer it using a number of features related to the blog network structure, historical data about the blogs' posts, text similarity, and timestamps.

Most early studies of cascades on online social networks exploited users' typed credit attribution of content sources to construct cascade networks. Examples of credit attributions are "RT", "via", "retweet", and "reshare" (Dow et al., 2013). There were also many attempts to use the social network and timestamps to infer cascade networks (Gomez Rodriguez et al., 2010). However, with more contextual information available it is now possible to construct more accurate cascade networks. For instance, Dow et al., (2013) used information about reshares, timestamps and clicks on feed, to infer cascade networks and compare them with cascade networks constructed solely from tracked information. More recently, online social network platforms start incorporating the ability to share content with a click on a button; for example retweet on Twitter, Reblog on Tumblr and Share on Facebook. With these functionalities in place, users can share different types of content easily. As a consequence, tracking existing cascades is

now feasible with the appropriate access to data. Thus, researchers are now able to construct existing cascades directly from the platform.

3.4.1 Cascade Networks' Topology

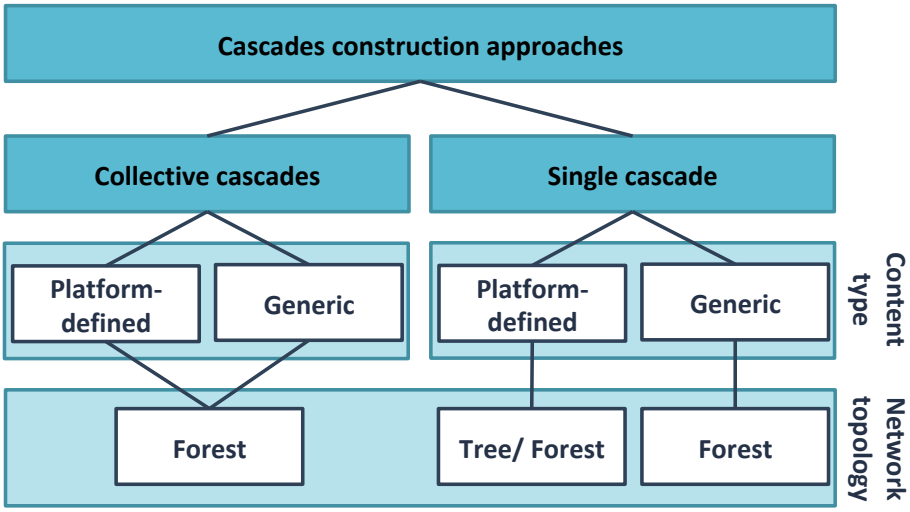


Figure 3.2 Cascade construction approaches and their outcomes

A cascade is often perceived as a tree that has a single root (the cascade initiator), which is linked to other nodes. Further nodes can be added by linking to the existing nodes in the cascade network and all of the added links follow a strict time order (Leskovec et al., 2007b). However, cascades are not always shaped as trees, in fact, their structure changes depending on the type of content these cascades networks represent. Anderson et al. (2015) classified cascade networks into: information-sharing networks in which information spread between the users and signups which mimic the adoption of a new technology. This classification does not specify the topology of the generated cascade network. Thus, the following presents a different classification of cascade networks based on their topology. The basis of this classification is the content type and the diffusion mechanism provided by the platform.

There are two main approaches to constructing cascade networks that have been used in research. Figure 3.2 illustrates them and the resulting cascades' topologies generated by each approach. The first approach is collective cascades, in which a large cascade network is constructed, linking users according to their sharing activities (retweet/reblog) collectively for a group of cascading items. The topology of this network is a forest that has several components. These large networks are useful to study the sharing activity patterns within a platform (Xu et al., 2014; Bild et al., 2015). Collective cascade networks are often weighted to represent how often a link occurs between two nodes (Leskovec et al., 2007b, 2006a).

The second approach is for single cascades in which cascade networks are constructed for each item that has been shared separately. Of the two categories of content, the first is a platform-defined elements such as a tweet in Twitter or a post in Tumblr. The second category (generic elements) covers any element that can be embedded within platform-defined elements such as a URL, a hashtag, a text, or a photo. Different content types require different data collection and analysis methods, and they create a completely different network topology.

The platform-defined elements that can be shared are for example: a post on Tumblr and Facebook, or a tweet on Twitter. This type of content spreads via explicit diffusion functionalities such as retweeting, sharing or reblogging. Their spread generates cascades that can be tracked or inferred on the platform. Cascades are constructed from the flow of information from users who might or might not be connected to each other by a relationship within the social graph (Cheng et al., 2014, 2016). These cascade networks ideally follow a tree topology; the root is the source (author) and from there content travels across the social network. However, in many cases due to the limited access to the platform, some data might be missing because it is deleted, the topology of the generated cascade network will be a forest where there will be separate components for each isolated part that can not be linked to the main tree due to missing data (Taxidou & Fischer, 2014).

Because the diffusion of generic elements, such as hashtags and URLs, does not occur via explicit diffusion functionalities in social networks, timestamps are often utilised as an indicators of diffusion between users assuming that these users have an established social relationship in the social network graph. Cascade networks of generic items are different to cascade networks of one story. These networks incorporate multiple introductions of the same item in the network, thus naturally their topology will be a forest with separate components (sub-cascades). Hence, the number of sub-cascades and their sizes can be used as structural features of these networks (Galuba & Aberer, 2010).

Collective cascades networks can be easily converted into single cascade networks by separating the different branches of the network where they are related to the same story (message). For instance, Leskovec et al. (Leskovec et al., 2006a, 2007b) generated cascade networks following the two approaches from blogosphere. They constructed a post network that links posts if they credit each other. From the post network they constructed a blog network by collapsing the links between blogs and assigning weights to them. Following this method, they constructed separate cascade trees from the post network. Sections 3.4.3 and 3.4.4 discuss cascade construction approaches used in different platforms, including the data used for their

construction, the detected diffusion mechanism, and the topology of cascades. A detail survey of the approaches is presented in Appendix B.

3.4.2 Link Directions

Edges between the nodes in a network might convey different meanings. For instance, Bild et al. (2015) identify the number of users who retweet from a user as the *popularity*; while the *prolificity* refers to the number of users a user retweet from. Hence, the direction of edges in a network can have different meanings. Consequently, all the measures that rely on the edges' direction will be affected.

Figure 3.3 illustrates two possible uses for edges' direction as used in the literature For example, suppose that we have three users, A, B and C. For simplicity, suppose that we have the following settings: user B follows user A and user C follows user B. Then, each time user A posted some content user B will be exposed to it and when user B shares that content after seeing it; user C will be exposed to the content too and can share it as well. In such a scenario, there are two possible representations:

Relationship perspective: If our concern is to represent who is linked to whom i.e., who follows whom, then the in-link from B to A means that B is linked to A, and the in-link from C to B means that C is linked to B. This is shown on the left in Figure 3.3, this representation is often referred to as the social network or the follow network.

Information flow perspective: In this case, the in-link from one user (A) to another (B) means that B is exposed to whatever information A has and when B shares that information too an edge will be drawn from A to B indicating the flow of information from A to B. This representation is often used for cascade networks. Figure 3.3 shows how this network can be constructed cumulatively at different timestamps. At timestamp t1, A posted a content, then when B was exposed to it, B decides to share it at timestamp t2, hence the edge from A to B and so on.

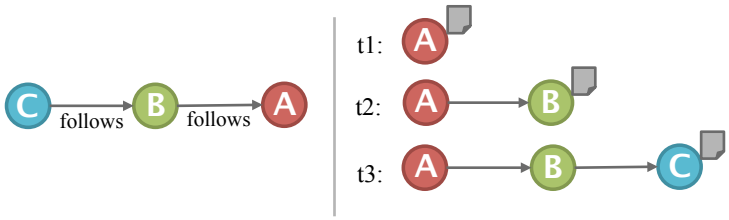


Figure 3.3 Link types in social networks

3.4.3 Cascades in Blogs, Recommendation Networks and Internet-chain letters

Data used to construct/infer cascades:

As mentioned earlier, in the early days of blogosphere there were no convenient mechanism to share content. Thus, instead of following the traces, cascades are inferred using a variety of measures such as: posts text, explicit links to other blogs, features about the blogs network, the blog and the timestamps (Adar & Adamic, 2005). In another study of cascades on blogs, the In-links/out-links between blog posts and timestamps were utilised to construct cascade networks (Leskovec et al., 2007b, 2006a).

On the other hand, on recommendation networks information about: products, time of recommendation, whether the product is purchased, and time of purchase are utilised to infer these networks (Leskovec et al., 2006b). Also, Liben-Nowell and Kleinberg (2008) used the ordered list of users who forwarded the petition to construct the cascades of chain-letters.

Diffusion mechanism:

As we can see the lack of explicit diffusion functionality means that various mechanisms of diffusion were identified such as: posting a URL in a blog (Adar & Adamic, 2005), recommending a product \cite{Leskovec2006}, linking between posts on blogs (Leskovec et al., 2007b, 2006a), and forwarding of a petition letter from one user to another (Liben-Nowell & Kleinberg, 2008).

Cascade networks topology and components:

The network topology of these cascades and their components vary based on the platform and the purpose of them. For instance, in Adar & Adamic work (2005) the cascade networks structure is trees, where the nodes are blogs and the edges between them are inferred to show the direction of diffusion of information between the blogs. While Leskovec et al. (2007b, 2006a), constructed a posts network that links posts in different blogs, and a blogs network which is a collapsed and weighted version of the posts network. Both networks are forests and they extracted separate cascade trees from the posts network.

On recommendation networks a separate group networks and a product networks are constructed, where the nodes are the customers and the edges connect customers' product recommendations (Leskovec et al., 2006b). Finally, in the work Liben-Nowell and Kleinberg (2008) the lists of users in each petition contains duplicates or missing users. Thus, the cascade networks are trees inferred by removing edges that did not appear in a sufficient number of copies. Thus, the nodes are users and the edges represent the direction of information flow between them.

3.4.4 Cascades in OSNs

Data used to construct/infer cascades:

Depending on the content type in each study and the diffusion mechanism, the data needed to construct cascade networks on OSNs. They vary from: retweets on Twitter (Kwak et al., 2010; Bhattacharya & Ram, 2012; Bild et al., 2015), reblogs on Tumblr (Chang et al., 2014; Xu et al., 2014), and share on Facebook (Dow et al., 2013; Cheng et al., 2014, 2016).

The tweet texts, timestamps and social network are used in (Galuba & Aberer, 2010) to infer cascade networks of URLs. Yang and Counts (2010) analysed tweets' texts that contain topics and mentions of other users to construct cascades. Also, text analysis (status updates that include the meme and the words 'copy', 'paste' and 'repost'), lists of users who commented on users' status and timestamps are used in (Adamic et al., 2012) to construct cascades of memes on Facebook. In another study of cascades of memes on Facebook, the social network, time, text similarity measures are used (Adamic et al., 2016). On LinkedIn signups and timestamp are used to construct cascade networks of invitations (Anderson et al., 2015).

These examples shows the diverse views of cascades on OSNs; they show how the diffused content type affects the cascade, and the varieties of data that can be used to either construct or infer cascade networks.

Diffusion mechanism:

On OSNs the main diffusion mechanism is provided by a platform's functionality (retweet, reblog, share). Other mechanisms of diffusion are: posting a URL (Galuba & Aberer, 2010), or crediting the source using `RT @' in tweet text (Galuba & Aberer, 2010; Bild et al., 2015). For memes, the diffusion mechanism is simply copy and paste of textual memes (Adamic et al., 2012, 2016).

Cascade networks topology and components:

Various cascade networks topologies are used based on the content type, as mentioned earlier platform-defined elements generate trees, while generic elements generate forests. For example, Kwak et al. (2010) created retweet trees for each tweet in their dataset and forests for each topic. Also, in Galuba and Aberer (2010) work, because the diffusion mechanism used is either posting a URL or crediting the source, the generated cascades' structure is a forest. Due to their nature, cascades of memes are forests (Adamic et al., 2012, 2016). There are also two studies that constructed large cascade networks of collective cascades (Xu et al., 2014; Bild et al., 2015).

In general, the nodes in most of the cascades on OSNs are users, and the edges always indicate the direction of information flow between them. An exception was found in (Adamic et al., 2016), where the nodes are meme variants and the edges between them link a meme variant to its parent.

3.5 Cascade Features

Due to their complex structures and features, cascade analysis relies on a set of features to use as a proxy to estimate these structures and temporal features. In general, the data available to harvest about cascades is multidimensional in its nature. It has a twofold purpose: the first is to allow cascade networks to be constructed using the detailed information about users sharing from other users; the second is to allow the creation of a time series dataset, where the number of sharing activities at a given time (day or hour) after publishing is recorded. The first is linked to the relation between the users involved in the cascade, i.e. who influenced whom to spread the content. The second (time-series) holds information about cascades and the number of diffusion events that occur at a given time. Each of these dimensions is related to a different aspect: either the structural or the temporal. These two aspects complement each other and provide a better understanding of cascades, as Scott (2008) argued that the temporal aspect adds value to the structural aspect when analysing data from social networks. The level of access researchers have to the platform's data determines the type of data they can gather. For instance, utilising a privileged access ensures that both dimensions are harvested, minimising the effect of missing or deleted data. In addition, with privileged access researchers have unlimited access to rich metadata such as clicks in News Feed (Dow et al., 2013). As a result, they can infer cascades more accurately. Figure 3.4 illustrates the two classes of cascades features; the next subsections, will explore the structural and temporal features of cascades; they provide definitions of these features and highlight their significance in relation to cascades' analysis.

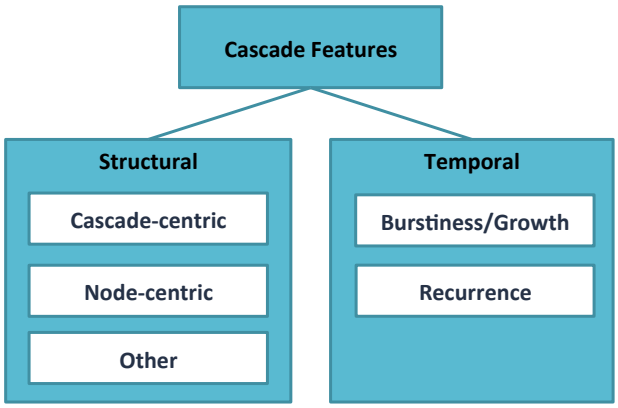


Figure 3.4 Cascade features classes

3.5.1 Structural Features

Analysing the structural features of cascades includes studying their structure and quantifying cascade networks' properties. According to Liben-Nowell and Kleinberg (2008), a better understanding of the properties of the structure of cascades leads to better dissemination models. There are three categories of the structural features of cascades: the first category is cascade-centric features; these features are computed on the cascade level as a whole. The definition and significance of each of these features is as follow:

1. **Depth, range, and distance to the root:** These measures represent the height of a cascade, they are calculated using the number of subsequent occurrences of message passing events, i.e. maximum number of hops or range of influence. Maximum depth and average depth can be measured too. Indicates the shape of a cascade, and how far it travels away from the source within the network. When all distances to the root are gathered, they can help assessing whether a cascade is shallow or deep (Liben-Nowell & Kleinberg, 2008; Yang & Counts, 2010; Kwak et al., 2010; Galuba & Aberer, 2010; Bakshy et al., 2011; Adamic et al., 2012; Goel et al., 2012; Dow et al., 2013; Chang et al., 2014).
2. **Width:** the width is computed as the maximum size of a set of nodes, which share the same depth. It indicates the extent to which a cascade is narrow or wide. It gives hints about the factors that make a message quite popular at one stage within the cascade (Liben-Nowell & Kleinberg, 2008).
3. **The fraction of nodes with exactly one child:** As the name suggests it is the number of nodes with one child only and it indicates missing or unsuccessful cascade event (Liben-Nowell & Kleinberg, 2008).
4. **Scale:** The scale is the number of nodes at depth one and it Indicates how popular/interesting a message gets soon after its first appearance (Yang & Counts, 2009).
5. **Wiener index:** It is a measure of the structural virility; it is computed as the average distance between all pairs of nodes in a cascade. The Wiener index gives an indication of the cascade shape, the higher the Wiener index, the more viral the cascade. Cascades with low Wiener index resemble a star shape, where there are few hubs that create the cascade. The Wiener index increases with the increase in cascade size (Cheng et al., 2014; Anderson et al., 2015).
6. **The percentage of adoption per depth:** Counting the percentage of adoptions within one degree of a root could indicate whether epidemic-like cascade occurs in the dataset, i.e. if the majority of adoptions recorded in the dataset are within the first few degrees from a root, then one could conclude that most cascades are shallow and small (Goel et al., 2012; Anderson et al., 2015).

7. **Number of nodes at depth = 1:** Nodes (users) at depth 1 are the ones who share directly from the author, meaning that they were exposed to the author's post directly. It might be that they arrive via external resources or direct links. Although there is a possibility that users click on the original post and share from the author rather than from user they receive the post from (Dow et al., 2013).
8. **Connectivity Rate:** It is the percentage of users who have one edge at least; hence, they were influenced by other users. It shows whether an edge exists between any two nodes in the cascade. It is useful to examine whether users get their information from the social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree (Taxidou & Fischer, 2014).
9. **Root Fragment Rate:** It is the percentage of nodes that have either direct or indirect connection to the root node. It Shows whether each node in the cascade is actually linked to the root or not. It is useful to examine whether users get their information from social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree (Taxidou & Fischer, 2014).
10. **Diameter:** The diameter of a network shows whether cascades are deep or shallow (Leskovec et al., 2007b; Taxidou & Fischer, 2014).

The second category is node-centric structural features, which are computed on nodes level. There are two features in this category: the branching factors and the subcascade size and they both measure individual's influence on the overall cascade (Gruhl et al., 2004; Dow et al., 2013; Galuba & Aberer, 2010) . However, there is a difference between the two, as the branching factor estimates the immediate influence, the subcascade size estimates the overall influence of one individual on the cascade.

The last structural feature is the frequency of distinct cascade structures. It helps to detect if there is a repeated cascade pattern, which can be investigated later. When combined with depth, it could help draw some conclusions about the shape of the cascade and how far it branches (Leskovec et al., 2006b, 2006a, 2007b; Goel et al., 2012; Chang et al., 2014).

3.5.2 Temporal Features

There are two approaches to analyse the temporal aspect of cascades. The first tracks and describes existing cascades' temporal features, e.g. how fast information spreads, for how long trendy content keeps its popularity, and the overall growth of cascades over time, such as: whether cascades show patterns like `burstiness' or sparks. The other line of research uses cascade's temporal patterns to either predict or model the cascade's future popularity. Most of

these studies do use the word `cascade', because they are concerned with the temporal aspect of the diffusion of online content. However, the underlying structure of online content diffusion is an implicit cascade network. The cascades' temporal features, their definition and significance are listed below:

1. **Time passed since message published:** it is the time since a particular message has been published. It shows the growth of cascade and the fade of interest in the message over time (Dow et al., 2013).
2. **Speed:** Calculated using the time at which the first cascade occurs, it indicates how fast users would be influenced to spread the message or generally react using other means of interaction like reply or mention (Yang & Counts, 2009).
3. **Time lag between posting and first reshare, elapsed time:** Measures the resharability of content, the larger the lag the less likely a content will be reshared (Kwak et al., 2010; Chang et al., 2014).
4. **Time lag between two sharing events:** Shows the speed at which a cascade occurs in relation to the distance between nodes, i.e. sharing events (Kwak et al., 2010).
5. **The number of spikes/peaks:** Spikes refer to high-volume of cascading activities that occur in a short period during the lifetime of a cascade. It measures the degree to which a cascade provokes high volume of cascading during its lifetime (Gruhl et al., 2004; Cheng et al., 2016).
6. **Cascading density throughout lifetime:** it is the timeline of a cascade; it shows the number of cascading activities per day. It helps assessing the temporal patterns of diffusion, whether it has spikes or maintains a steady growth. (Gruhl et al., 2004).
7. **Maximum time between reshares:** Indicates the maximum idleness period within a cascade (Cheng et al., 2016).
8. **Cascade growth/cascade popularity:** Helps to show whether a cascade size grows linearly as time passes or in different ways. This helps detect whether the growth in cascade size occurs in short intervals or whether it grows with time. It also shows the periods of idleness and spikes in the cascade timeline (Leskovec et al., 2007b, 2006a; Adamic et al., 2012; Dow et al., 2013; Anderson et al., 2015).
9. **Recurrence:** Recurrence occurs if a cascade has at least two peaks in addition to other conditions. It helps identifying cascades that regain their popularity after a period of idleness (Cheng et al., 2016).

The table in Appendix C summarises cascade features. It contains the feature definition, and briefly discusses its significance and the way it appears in the analysis.

3.6 Large and Viral Cascades

3.6.1 Large Cascades

The structure of cascades has been extensively analysed in the literature, and there is a strong debate between those who assert that cascades are deep and those who assert they are shallow.

The main finding of Liben-Nowell and Kleinberg (2008) was that the structure of the trees is rather deep and narrow, in contrast to the 'small-world' perspective in which many people are reached in a few steps. On Facebook memes reach 40 steps in depth (Adamic et al., 2012), the same conclusion reached in adoption cascades (Anderson et al., 2015). Taxidou and Fischer (2014), using the diameter measure, concluded that cascades are deep in Twitter, contradicting the work of Kwak et al. (2010). Goel et al. (2012) state that large cascades exist but are rare, one in a thousand cascades are of a medium size, while one in a million cascades are viral (2013). In another study, they state that 99% of cascades are shallow, and die in one step, and large-scale cascades are rare.

On the other hand, in blogosphere most cascades are shallow but some are relatively large (Leskovec et al., 2006b). Other studies reached the same conclusion that cascades are fragmented and shallow (Leskovec et al., 2006a, 2007b; Bakshy et al., 2011; Bhattacharya & Ram, 2012). Galuba and Aberer (2010) in their Twitter study found that cascades (sub-cascades in their case) are shallow, and the distance between any node to the root is short. They argued that the reason behind the shallowness of cascade networks is that users often follow the author of interesting content as soon as they are exposed to their messages. This makes future cascades shorter and shorter.

Anderson et al. (2015) argued that information cascades: 1) happen very quickly, 2) most of the sharing events are very close to the root, 3) there is no correlation between the size of the cascade and their structural virality (Cheng et al., 2014; Goel et al., 2012, 2015b). In adoption cascades it is the opposite; they take longer time, they occur through multiple steps and they are highly viral.

3.6.2 The Notion of Virality

In many news outlets, and in much research, terms like virality and popularity are used interchangeably to describe a content that spreads at high volumes on several online social networks. The interest in virality is caused by the bias of the commercial social media toward virality (Cebrian et al., 2016). However, there is some degree of ambiguity when it comes to the

exact meaning of these words. Dow et al., (2013) stated that the majority of cascades are non-viral, while some viral cascades do occur. They analysed two 'viral' cascades that were shared 618,015 and 150,759 times respectively. Cheng et al., (2016) used the term virality to refer to the appeal of content in its early stage, i.e. whether it will be shared in high volume in the first few days after publishing. Goel et al., (2013) suggested using the Wiener Index (WI), drawn from chemistry, which measures the degree of complexity in the structures, as a measure of virality. They differentiate between two types of cascade structure in the literature; the first is a cascade that has elements of virality, which results in creating denser and more complex structures. Viral networks branch out in multiple steps where users influence each other along the paths. The other type is viewed as a broadcast, where many individuals receive information from one source.

All of the above notions of virality take into account the structure of cascade networks and the size of a cascade in its early stages. However, in their recent work on cascades' recurrence, Cheng et al., (2016) used the early size of cascade virality to differentiate between low, moderate, and high virality. They found that moderate virality might cause cascades to recur more than low or high virality, because high virality means that a significant number of users will be exposed early on, minimising the chances of cascades' recurrence.

3.7 Chapter Summary

This chapter reviewed the literature on cascade, the third component of the information diffusion process and the outcome of the process. It presented various definitions of cascades and their significance in research. It then also outlined the different purposes behind studying cascades: tracking, modelling, prediction, and inferring. In addition, the temporal and structural aspects of cascades were discussed. Before analysing the structure of cascades, their construction phase takes place and this chapter explored the different approaches used to construct such cascades. The cascades' features were listed in a survey that included their definitions, significance, and how they were analysed. The chapter concluded by discussing two topics related to cascades: the size of cascades and cascades' virality.

Chapter 4: IDF: Information Diffusion Framework

> 'First of all, I know it's all people like you. And that's what's so scary. *Individually* you don't know what you're doing *collectively*.'
>
> Dave Eggers, The Circle

This chapter proposes an information diffusion framework IDF that comprises several aspects of the diffusion process. This chapter will explain the components of the framework and highlight their relation to each other.

4.1 A Framework for Information Diffusion?

Information diffusion is a complex phenomenon that involves several components and has many aspects. Looking at the literature about information diffusion one can see that it has been investigated from several angles. For instance, the diffusion of a particular hashtag can be used as a way to identify interesting topics (Lin et al., 2013). Moreover, in many cases the diffusion of content and influence are considered as an indicator of each other interchangeably (Bakshy et al., 2011; Taxidou & Fischer, 2014).

There have been some efforts, in previous research, towards providing an abstract overview of the information diffusion phenomenon. Examples of such efforts are present in the work of Guille et al. (2013). They published a survey of information diffusion that categorises research into three categories: (i) Detecting popular topics; (ii) Modelling the diffusion; and (iii) Identifying individuals with influence. They also proposed a taxonomy for the different approaches that have been used under each category and identified areas of improvement, which include: adding social properties, defining and using topics, studying competing and cooperating information.

The above categories as used in Guille et al. (2013) survey helped in shaping the literature review chapters. These categories were then translated into the following:

i. Detecting popular topics → The type of information that spreads → **Content**
ii. Modelling the diffusion → The way in which information spreads → **Cascade**
iii. Identifying individuals with influence → The role of people in the spread of information → **Context**

So, Chapter 2 and 3 presented a literature review about information diffusion and cascades; the discussion was split into three parts, the content that spreads, the context that facilitate the spread and the cascade which is the outcome of the diffusion. This early categorisation of information diffusion components was then extended to include more aspects of the phenomenon in the related literature such as, the social network and the affordances of the platform.

In her PhD thesis investigating information diffusion, Weng (2014) framed her research around four components: Actors, Content, Network and Diffusion. Her work studied the different aspects of the actors, including their limited attention, homophily and tie strength. It also analysed the impact of the content's topics, languages, sentiment and culture on diffusion. In addition, it analysed the effect of the network formation and model on diffusion and the possibility of generating cascades from such networks.

Subbian in his thesis studied information flow in networks (2014). His work identified three components for any piece of information that spread: content, network and time. Content-centric analysis aims to extract different flow patterns in the network, while the network-centric analysis aims to extract flow patterns efficiently using vertex-centric algorithms. Time, on the other hand, looks at analysing cascades incrementally as they arrive in the stream.

The above studies have identified the diffusion components and some of their applications and impact. They also highlight the fact that these components play a different role in the diffusion process. Although some of the studies examined some of the relations between these components, such as the interplay between the social network and diffusion (Weng, 2014), they did not provide a general overview of the diffusion process and all of the possible relations between its components in a holistic way.

4.2 The Construction of IDF

Drawn from the literature and the definitions of information diffusion and cascades, a framework of information diffusion is proposed (Figure 4.1). The framework is simple and most of the work done in the literature fits within its frame, as it not only captures the components of the diffusion but also the relations between them. The framework has three main components that conceptualise the information diffusion process: the content, the context and the cascade. The following phrase highlights the relation between the three components and summarises the framework: If there are sufficient **contextual factors** that facilitate the diffusion, **content** will spread and its spread creates a **cascade**.

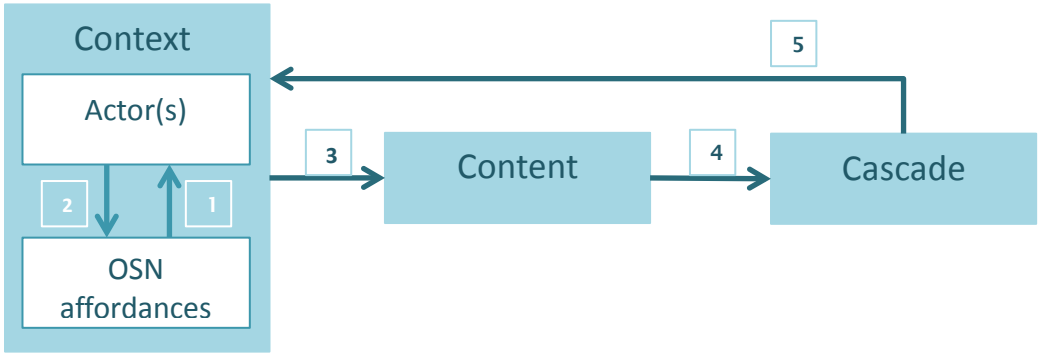


Figure 4.1 An illustration of the information diffusion framework (IDF)

4.2.1 The Context

There are two parts of the context that play a major role in the diffusion of content in online social networks. Firstly, it is either related to the Actors involved in the process i.e., the users (human or non-human) and the Online Social Network's (OSN's) affordances. The actors' factors include their relation with each other, manifested in the social network (sometimes it is referred to as the social graph), the actors' influence and their homophily. These topics have been covered in Chapter 2; however, they will be briefly explained here, focusing on their relation to the diffusion process.

The social network and the users' connections, i.e. those who are following the users, determine who gets exposed to their shared content. The probability of content spreading depends not on the quantity of the relations a user has but their quality. The more active and engaged the users' neighbours are the more their content will potentially spread. The influence of the user, which has been the focus of many studies in the literature, tries to estimate the influence of the user using measures derived from the platform itself or the user's position within the social network, while their homophily refers to the effects of the coherence among a set of users in their collective diffusion actions, i.e. whether their languages, locations, and shared interests mean that they will be interested in the same type of content. Chapter 2 discussed how these three Actor-related factors are the main motivations of diffusion, and how they relate to the emergence of cascades.

On the other hand, the OSN's affordances also play a major role in the diffusion process. The main affordances that have a direct impact on the diffusion are the ability to spread content using a built-in functionality and the ability to discover new content via content promotion and discovery mechanisms.

There is a relation between the OSNs Affordances and the Actor(s) and vice versa. To better understand arrow number one, consider the following example: if a user gets exposed to new and interesting content via the platform's ability, a user might decide to follow the content's creator, which affects both the social network and will have an impact on the content creator's influence score. Using other OSN affordances, such as the ability to send messages and to interact with non-verbal functionalities like comments and likes. The frequency of such contacts can be used as measures of the strength of the ties between users, which serves as a proxy to estimate users' influence.

Arrow number two represents the link from Actor(s) to OSN affordances: here, the OSN affordances are affected by the users. An account of an old example occurred a number of years ago, before introducing the retweet button on Twitter, users copied the entire tweet and included "RT" or similar identifiers to their tweets. Eventually, Twitter implemented that functionality allowing users to retweet with a click on a button.

4.2.2 The Content

Content that spreads can be looked at from two points of view; the first one identifies and quantifies the characteristics that makes it spread. The second one looks at the effect of the contents' characteristics on the ability to track it and analyse it, i.e. the effect of the content on the generated cascade.

In the literature, two characteristics that cause content spread are considered: its interestingness and its retweetability. Both can be quantified using either information about the sharing history or by asking humans to rate the content (Lerman & Rey, 2007; Bakshy et al., 2011; Webberley et al., 2013). However, there are other characteristics that affect content spread, namely the temporal factor, geography and language. For example, the sentence "Happy New Year" is relevant on one day of the year, for the rest of the year this sentence is not relevant and it is unlikely that it will attract users to spread it. In addition, a user might share something intriguing or funny but written in a language that most of the user's friends do not understand, or a user might write about an event that happens in a different country. The chances that content will be shared diminishes, as the content would not be appealing to most of the user's friends.

On the other hand, the content that is the subject of investigation might come in different forms. It can be a text, a hashtag, a photo, a URL or a distinct item such as a tweet or a post. All of these different types need different tracking and collecting approaches. Not only that, but they also require different approaches to analyse them. The generic items (text, hashtags, URLs) might appear anywhere in the social network, creating a cascade that is best described as a forest,

whereas items such as a tweet or a post create tree-shaped cascade networks in an ideal scenario where there is nothing missing.

The context has an effect on the content (represented by the third arrow). Firstly, OSNs Affordances control what type of content users are able to publish and what content the user will be exposed to. The former has an effect on the methods used to track and analyse cascades generated by that content, while the latter increases the likelihood that the content will spread. The social network, i.e. the types of connections a user has and the degree of homophily between the user and his friends have an impact on the type of content they will be interested in and, eventually, may spread among them. For example, a group of users might share the same URL if they are homophilous. Moreover, as explained earlier the content type affects cascade construction and analysis (arrow 4).

4.2.3 The Cascade

The definitions, features and structures (i.e. cascade construction) of cascades were explained in depth in Chapters 3 and 5. However, according to the framework there are three aspects to analysis of cascade networks: structural, temporal and social.

The fifth arrow is used to represent the impact of the cascade on the context. The major impact of cascades is their impact on the social network's evolution. As the content becomes popular, users might decide to be friends with or follow the author of the content or any user who is also involved by spreading the content.

4.3 How to Use IDF?

This IDF provides a holistic overview of the information diffusion phenomenon and its related components and aspects. It helps in organising research tasks that aim to answer questions about information diffusion and cascades.

The study presented in this thesis investigates cascades on Tumblr, a platform that offers several OSN affordances related to content spread. This study aims to investigate the relation between the OSN affordances and the increase in the likelihood that the content will spread, as well as constructing cascade networks from distinct posts and analysing cascades from structural and temporal aspects. Thus, in the context this study includes OSN affordances, namely the ability to reblog posts and content discovery mechanisms such as tags and content promotion. The content of interest in this study is that of individual posts, which affects the cascade analysis

phase: for the temporal analysis, time-series data of cascades are needed, whilst for the structural analysis, data about reblogging causality (i.e., network data) are needed.

4.4 Chapter Summary

This chapter has proposed an information diffusion framework that encapsulates the phenomenon and its components. The chapter briefly highlighted what each component means and explained their relation to each other. The importance of this framework is that it will help researchers seeking to study diffusion and cascades on social networks as it incorporates all of the different components in the diffusion process; thus, it can be tailored according to the intended research purpose.

Chapter 5: Research Methodology

> 'The greatest challenge to any thinker is stating the problem in a way that will allow a solution'
>
> Bertrand Russell

The aim of this chapter is to provide an overview of the research methodology followed to answer the questions this study is concerned about. The first section in this chapter will discuss the approaches used to analyse and interpret data gathered from online social networks. In particular, it discusses how Network, Web and Data Sciences approaches can help analysing such data. The second section presents the research rationale and questions in the light of the concepts discussed in Section 5.2, in addition to the research methodology and the deliverables from each stage. The third section explains the experimental settings of this study; it discusses Tumblr's features, data sampling and pre-processing. The last part of the chapter discusses the cascade construction models used in this study as a phase that precedes the structural analysis.

5.1 OSNs and the Co-operative Sciences

Since their emergence, online social networks (OSNs) become major channels where people share content and connect with each other. The ample amount of data generated by users on various online social network platforms has created a new strand of research, which has utilised that data for various purposes. Many fields have used data gathered from online social networks, including computer science, sociology, political science, marketing and economics (boyd & Crawford, 2011). However, this data has three characteristics that make analysing it more challenging: the size and noise of the data and the dynamism factor of the platforms (Adedoyin-Olowe et al., 2014). Thus, to provide deep insights from such data, theories from different disciplines are integrated with computational capabilities (Zafarani et al., 2014). Therefore, the methodologies followed for this strand of research require novel techniques to harvest and analyse the content of social networks and the rich context around it (Tinati et al., 2014).

What makes data harvested from social networks invaluable is that it has two facets; it represents different types of **relationships** between users and it conveys information about how the users (humans, often) **behave** on the Web. Hence, it is useful for a relatively emergent set of sciences, all of which contribute in attempts to analyse and interpret that data. These are:

Network Science, **Web Science** and **Data Science**. Each one of these "Sciences" provides approaches and perspectives that help to set the research agenda and ensure that the research phases are being followed in a methodological way. The areas of intersection between these fields and how their methodologies fit together are still being developed (Wright, 2011; Tiropanis et al., 2015; Phethean et al., 2016). They complement each other in allowing researchers to draw insightful and comprehensive overviews of the research topics in hand (Phethean et al., 2016).

Network Science is a field that studies the emergence, evolution and characteristics of networks (Tiropanis et al., 2015). It relies on a long history of network analysis in sociology and has utilised many mathematical approaches, e.g. graph theory, to model and analyse different types of networks (Watts, 2004). Networks are structures that can be found everywhere, such as transportation networks, telecommunication networks, biological networks and the Web (Newman, 2010). A special area of Network Science is the field known as Social Network Analysis (SNA). It is used to analyse social networks: the structural patterns that represent relationships between individuals, such as citations and collaboration networks (Freeman, 2011). SNA approaches take into account the characteristics of the ties between individuals rather than the individuals themselves and it also studies the implications of these structures in the individuals' behaviours (Otte & Rousseau, 2002). As mentioned earlier, social network data represent different types of **relationships** between the users themselves and the users' behaviour within the platforms. As seen in Chapter 3, the baseline of online social networks is the social network (social graph) that connects users to each other based on their follower or friend relations. Using this baseline network many networks can be added as additional layers; these include networks that represent cascade (reblogging or retweeting) relations and liking relations (Agarwal & Sureka, 2016). Thus, approaches drawn from Network Science and SNA have been utilised to analyse such data (Bródka et al., 2012; Bakshy et al., 2012; Antoniades & Dovrolis, 2015).

Web Science is an interdisciplinary field that is concerned about two paradigms: understanding the Web as a phenomenon and engineering its future growth (Berners-Lee et al., 2006). The main purpose of Web Science is to study the Web from both micro and macro perspectives, in other words, the technological aspects and the interactions of people (Hendler et al., 2008). In particular, Web Science studies the socio-technical aspect of the Web; it investigates how technology affects society and how society affects technology within the borders of the Web (Halford et al., 2010). Halford et al. state that to understand the socio-technical relations, actors' (humans and non-humans) behaviours implications on the Web must be followed. This is often achieved through mixed research methods using data collected from the Web, including data collected from social networks (Tiropanis et al., 2015). Social network platforms provide a number of functionalities and even record users' impressions on the platform. Both the functionalities and

the impressions are sources of information about users' **behaviours,** making it easy to observe, collect data about and analyse such behaviours. In the context of information diffusion, Web Science methodologies overlap with those of Network Science in two areas: they both rely on data from social networks, and they both utilise measurements, models and quantitative methods to analyse that data (Tiropanis et al., 2015).

The quantitative methods used to analyse data about information diffusion are often drawn from Data Science (Phethean et al., 2016), which is the science concerned with extracting knowledge from the data (Dhar, 2013). The aim of Data Science as described by Hayashi (1998) is "to reveal the features or the hidden structure of complicated natural, human and social phenomena with data from a different point of view". Thus it provides several methods that enable researchers to handle the data and analyse it. According to Hayashi (1998) there are three stages in Data Science research: design for data, collection of data and analysis of data. The classic data mining methodology also has three phases: data pre-processing, data analysis, and data interpretation (Adedoyin-Olowe et al., 2014).

5.1.1 Social Network Data Challenges

When analysing data collected from social networks, there are a number of challenges and it is sometimes difficult to avoid certain sources of bias that must be taken into account in any "Social Media Mining" task, as Zafarani et al. (2014) describe it.

The first one is the "**Big data Paradox**". Obviously on the macro level, the data collected is big, however, on the micro level it misses a lot of details about the users involved in the process. Thus, such data is often aggregated utilising the multidimensional aspect of social networks (Zafarani et al., 2014).

The second is "**Privileged Access**" to the data, which is not available for everyone in the research community; this creates a digital divide between those who have access and those who have not (boyd & Crawford, 2011). The immediate consequence of this situation is that it is often hard to collect sufficient samples of the data (Zafarani et al., 2014). For instance, Twitter streaming API provides 1% only of the overall tweets at any particular moment in time (Morstatter & Ave, 2014). Choudhury and Hari (2006) refer to this situation as a data acquisition bottleneck. In fact, Petrovic et al. (2011) argue that researchers who rely on the API to gather retweets might be missing a great deal. Another issue caused by limited access to the data is that the collected data might suffer from different types of biases, including: biases towards popular or viral content (Borghol et al., 2011; Cebrian et al., 2016), biases towards large cascades (Cheng et al., 2014), biases towards specific topics and time periods (Wang et al., 2013) and network

measurability bias towards easy to observe measures such as retweets and likes (Cebrian et al., 2016).

The third source of bias is the "**Noise Removal Fallacy**"; as Zafarani et al. (2014) point out, the quality of the findings mined from the data depends on the quality of the data. According to Zafarani et al., noise handling comes with its own set of challenges, as the arbitrary removal of noise might remove important information. Moreover, noise identification is based on the intended task, which complicates the noise elimination process.

The fourth issue is "**The Abstraction Pitfall**"; boyd and Crawford (2011) point out that abstraction might help in drawing some generic conclusions about the data but the context remains a vital aspect in the analysis. They explain this by referring to the "strong ties" concept proposed by Granovetter (1973): if someone appears to be spending more time with one of their colleagues, that does not necessarily mean they have a strong relationship with them as opposed to their spouses.

5.2 Research Rationale and Methodology

This thesis aims to answer the following question: **How does information diffusion occur on social networks?**

This general question has been split into four different sub-questions each of which focuses on a different aspect of the information diffusion phenomenon

RQ1: What are the factors that facilitate information diffusion in online social networks?

RQ2: How cascades networks can be constructed from minimal contextual information and missing/degraded information?

RQ3: What are the structural and temporal features of cascades?

RQ4: How is Tumblr, an online social network, used for information diffusion and what are the structural and temporal features of its cascades?

The aim of the first question is to conceptualise the information diffusion phenomenon. It investigates the different components of the diffusion process and the relations between them, highlighting their impact. The second question focuses on cascades, the outcome of the diffusion process. In particular this question aims to propose cascade construction models from minimal contextual information. Cascade construction is the fundamental stage that precedes analysis of the structural features of cascades. The third question aims to survey the structural and temporal

features of cascades and the measures used as estimates of these features. The fourth question aims to apply these measures to analyse Tumblr's top posts' cascades and compare their features to cascades on other platforms.

Guided by these four questions, Figure 5.1 demonstrates the methodology of this study, outlining its stages and the outcomes of each stage. The first stage is the **Literature Review** presented in Chapter 2 and Chapter 3. The main outcome of this stage is the **Cascades Features Survey**.

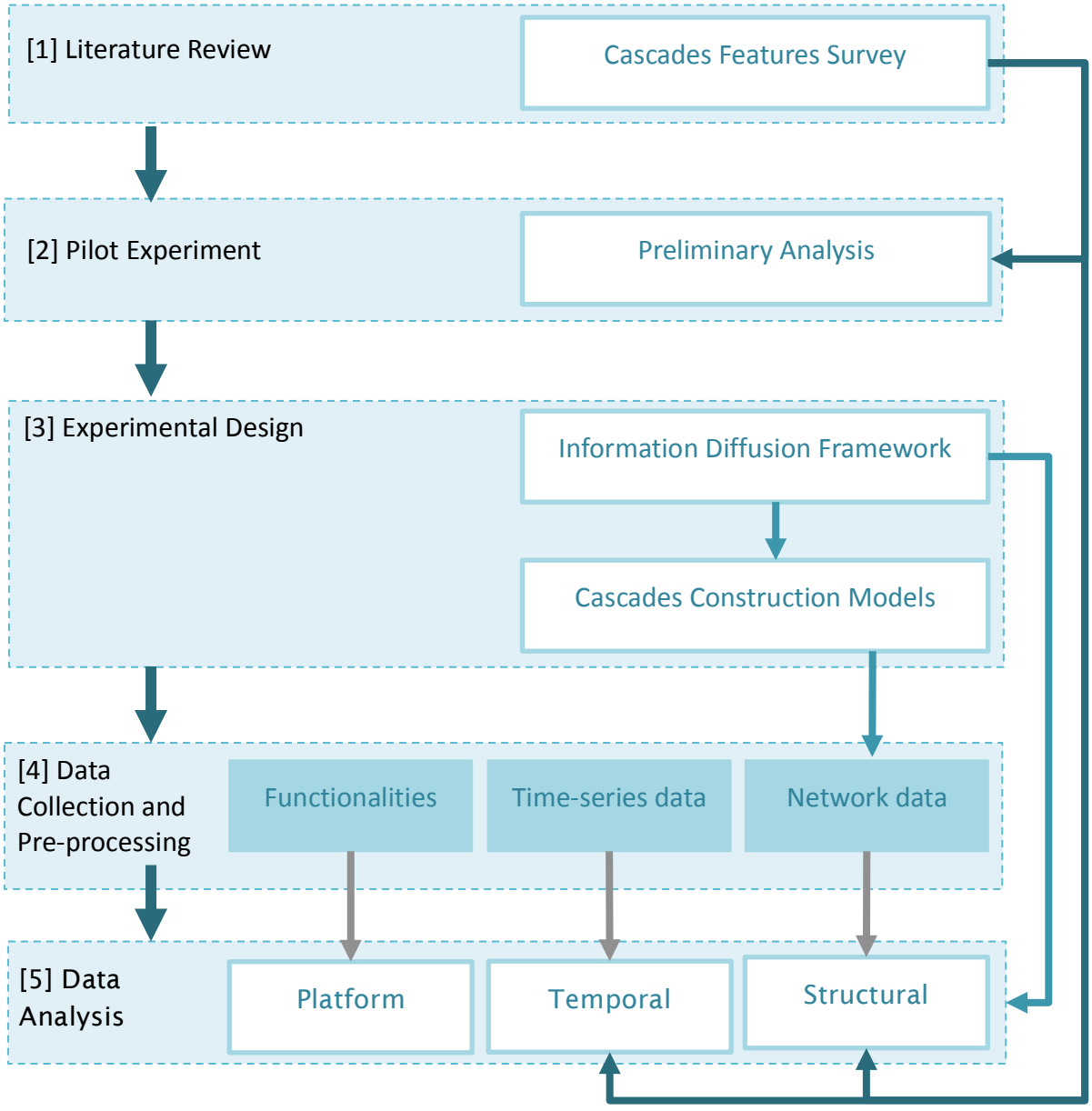


Figure 5.1 Research methodology stages and outcomes

The survey includes the structural and temporal features of cascades; it categorises these features and provides their definition, impact and how they were presented in previous research.

The second stage is the **Pilot Experiment,** presented in Appendix A. The aim of the experiment was to obtain hands-on experience of analysing cascades on Tumblr, including approaches for data collection and analysis. The **Preliminary Analysis** focuses on the structural features of cascades. The pilot experiment and the results of the preliminary analysis helped shaping the stages that follow, as its major purpose was to establish an understanding of Tumblr as a platform for content propagation, how cascades can be constructed and how to apply the structural features measures.

The stage that follows is **the Experimental Design**. There are two main outcomes from this stage, the **Information Diffusion Framework** (Chapter 4), and the **Cascades Construction Models** (Section 5.3.4). Both of these outcomes are reached based on the results of the pilot experiment. In addition to these outcomes, this stage also includes a "design for data" step, as described by Hayashi (1998), which is the step that precedes data collection in stage four. In this step, the data that will be collected is chosen, where the aim is to analyse popular content (Section 5.3.2).

Stage four is **Data Collection and Pre-processing** (discussed in Section 5.3.2 and Section 5.3.3); the outcomes of this stage comprise a multidimensional dataset with three dimensions: the users' behaviour on the platform (**Functionalities data**), the time at which each reblogging occurred (**Time-series data**), and reblogging causality between users (**Network data**). The cascade construction models from stage three are used to create cascade networks.

In stage five, **Data Analysis** takes place, where the three dimensions of the data are analysed. Here, the platform (Tumblr) is analysed using the data about users' behaviour and both the structural and temporal aspects of the cascades are also analysed. Both aspects are analysed using the measures obtained in stage one.

The research methodology used in this study and the tasks at each stage fall into the intersections between Network, Data and Web Sciences. Figure 5.2 illustrates how the different tasks at each stage in the methodology fall in the intersections between two or more disciplines. Data Collection and Pre-processing are Data Science tasks, while Cascade Construction relies on theories from Network Science and uses Data Science techniques to construct cascade networks. Hence, the Structural Analysis falls in the intersection between the three disciplines, as its interpretations reflect on the way popular content is spread on the Web; thus, it provides an understanding, an X-ray of the skeleton (i.e., the structure) of cascades on the Web. On the other hand, the Platform and Temporal Analyses are based on Data and Web Sciences. The quantitative methods performed are from Data Science while their interpretations shed light on the way users' behave on the Web and how cascades grow in relation to time.

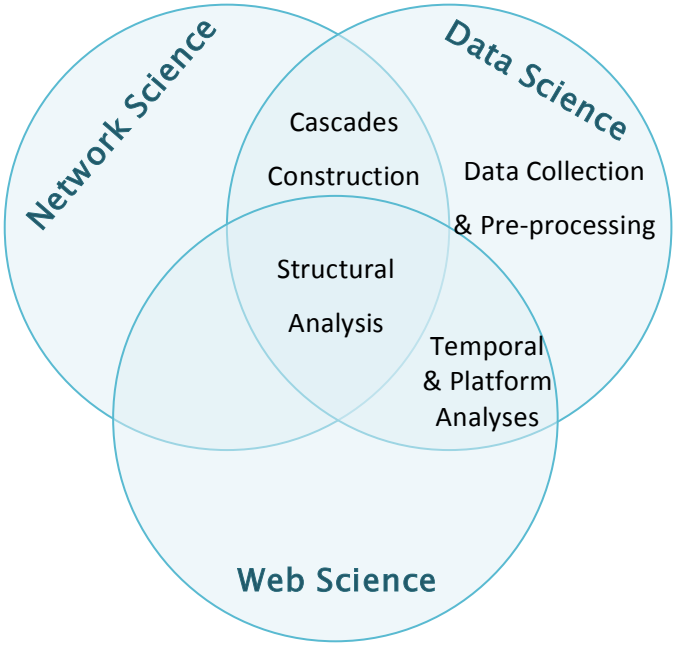


Figure 5.2. The tasks performed at different stages in this study fall in the intersections between Network, Data and Web Sciences

5.3 Experimental Design

This section aims at providing the contextual settings for this study. It will provide background information about Tumblr, the platform chosen to study information diffusion and cascades. It will then describe the dataset, how it was collected and the data pre-processing phase. Finally, cascade construction models and the approaches used for each model are explained in the last section.

5.3.1 What is Tumblr?

Tumblr is an online social network website founded in 2007, and currently owned by Yahoo! Inc. In Tumblr, each user has one or more blogs in which he or she can post any type of multimedia, such as text, photos, quotes, links, music or videos. Tumblr exhibits several characteristics from different domains, such as social media, blogosphere and social networking; this makes it a hybrid version of blogging and online social network platforms (Chang et al., 2014). Similarly to blogs, Tumblr allows its users to write longer posts in any multimedia form, yet, like any social networking platform, it provides various social interaction functionalities for users, such as following, reblogging and liking (Xu et al., 2014). Hence, the basic element of diffusion on Tumblr is the post and it diffuses/spreads using the built-in reblogging functionality. Reblogging allows users to reblog posts to their own blogs i.e., sharing it with their friends. Reblogging increases

posts' exposure rates and eventually attracts more users to reblog it. Once a post is reblogged, it will appear in the reblogger's blog with a new ID. However, it will still link to the original post and the original author. Users can add a comment with a reblog and can reblog both original and reblogged posts. Reblogs appear as notes for each post in the format:

User X reblogged this from User Y

User X reblogged this from User Y and added "a comment"

On each post, Tumblr maintains a list of notes; each item in this list gives the usernames (blog names) of all users who **reblog** or **like** this particular post. In the case of reblogging, it shows who actually reblogged from whom. This list appears on both original and reblogged posts, and it is unified across the platform. Figure 5.3 illustrates a post in Tumblr showing its list of notes and its tag components; both are separate from the content itself.

In addition to its basic role, reblogging on Tumblr is implemented in a unique way that allows users to reblog the same post more than once using the same conventional mechanism. This means that users might appear in different parts of one cascade. Chang et al. (2014) mentioned this aspect briefly in their paper; their explanation of this behaviour is that users use reblogging as a means of communication by adding comments to reblogs.

The availability of an explicit, unified and chronologically ordered list of all users who reblogged and liked a post makes the cascade construction task relatively easy (see an example in Figure 5.3). However, there are two cases that add complexity to the construction process. First, users can reblog the same post more than once; i.e., users might appear more than once in different parts of the cascade graph. Secondly, in some circumstances, some notes might be deleted; this might occur when users delete their reblogs. In such cases, the notes list will have some missing links, which creates isolated components within the cascade graph. Section 5.3.4 will discuss the cascade construction models used to handle such cases.

Tumblr allows users to follow an unlimited number of blogs. However, explicit lists of followers and followings are not necessarily shown; this is due to the separation between accounts and blogs. This makes the social network not accessible for anyone without privileged access to the data, as it is also not accessible via Tumblr's API. The impact of this is that there is no way to infer missing reblogs using information about users' connections. However, the fact that the list of notes is available and that it is ordered and explicitly represents causal relationships of reblogging allows cascade networks to be constructed.

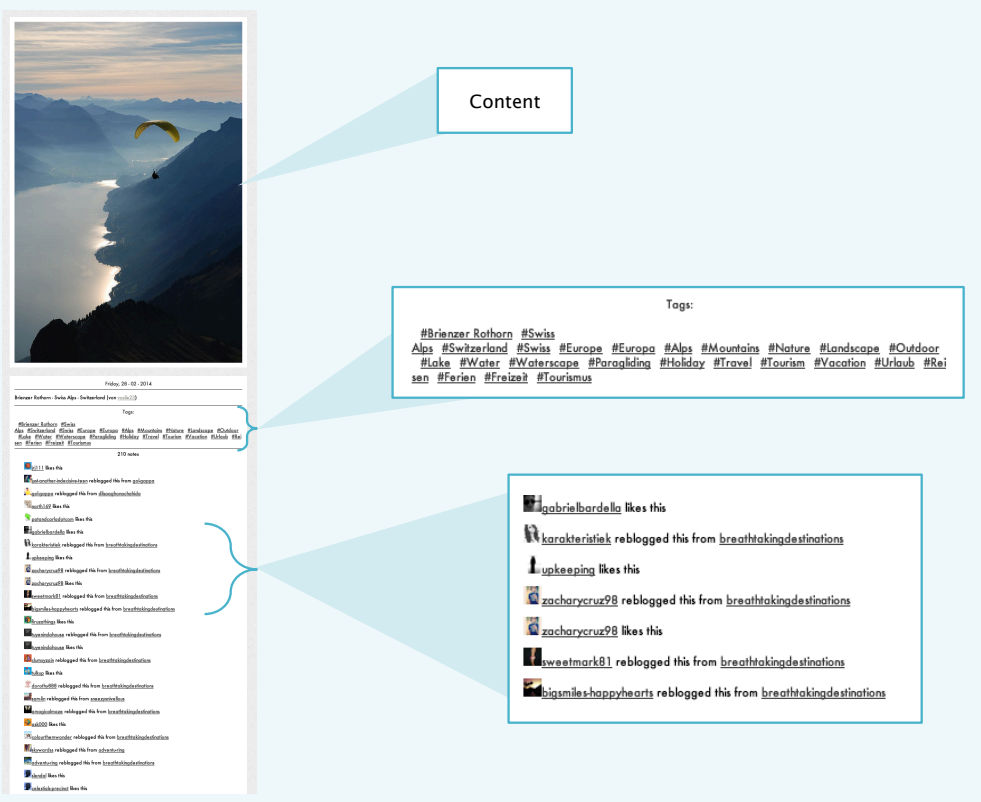


Figure 5.3 An illustration of the content, part of the notes container and the tags container in one post

5.3.2 Dataset Sampling and Collection

The dataset was harvested from Tumblr's 2014 “Year in Review” blog, which is a blog curated by Tumblr's staff at the end of each year to highlight the most popular posts and tags in that year. The details of the methodology that Tumblr follows to create this blog is not published. However, it is stated that it takes into account many factors, such as the web traffic, the growth in the number of followers, and the size of both the original and reblogged post.

This dataset was chosen because it was not possible to catalogue all of Tumblr's cascades, and select the “successful” ones that had large number of reblogs allowing the characteristics of large cascades to be analysed. Instead, the data sample was chosen by utilising Tumblr's staff effort to catalogue popular content, leveraging their privileged access.

Marres and Weltevrede (2013) make a distinction between “scraping the social” and “scraping the medium”. They differentiate between analysing media dynamics and analysing social dynamics, this leads to two types of research either to be concerned with the medium or the social life within the medium. Scraping the social simply means to capture the social life of the medium rather than to just collect data about the medium itself, which are not necessarily about the social aspect of the medium. Thus, a similar approach was used for this study, as the data

collection aimed to capture the social aspect of Tumblr. The harvesting process started by obtaining the URLs of the most reblogged posts from Tumblr's 2014 year in review blog, where each post belonged to one category that Tumblr's staff chose. For each post the following information was obtained: the URL, the author (blog-name), publishing timestamp, type and category. Initially, a web scraper was used to fetch the notes list for each post, starting from the most recent activities going down to the oldest activity. Both reblogging and liking activities were collected. And for each reblog the following information was collected: the username (blog-name) of the user who reblogged (reblogger), the username of the user from whom the post was reblogged (reblogee), the URL of the reblogged post, whether it included a comment and the comment's text.

Tumblr's API was then utilised for each reblog to get its timestamp i.e., the time at which a reblog is made. Hence, the harvested data is multidimensional and it has a twofold purpose: the first is to allow cascade networks to be constructed using the detailed information about users reblogging from other users; the second is to allow the creation of a time series dataset, where the number of reblogging activities at a given time (day or hour) after publishing is recorded. Figure 5.4 illustrates these two data representations that are used for the analysis. In addition, the information gathered about the functionalities, reblogging, commenting and liking, will be used to analyse the platform.

As a result, the harvested dataset contains the top posts in 57 different categories. The number of the harvested posts in each category ranges from 10 to 115 posts, and the total number of harvested posts is 1292. In the whole dataset there are 73,048,903 independent reblogging events in all the posts.

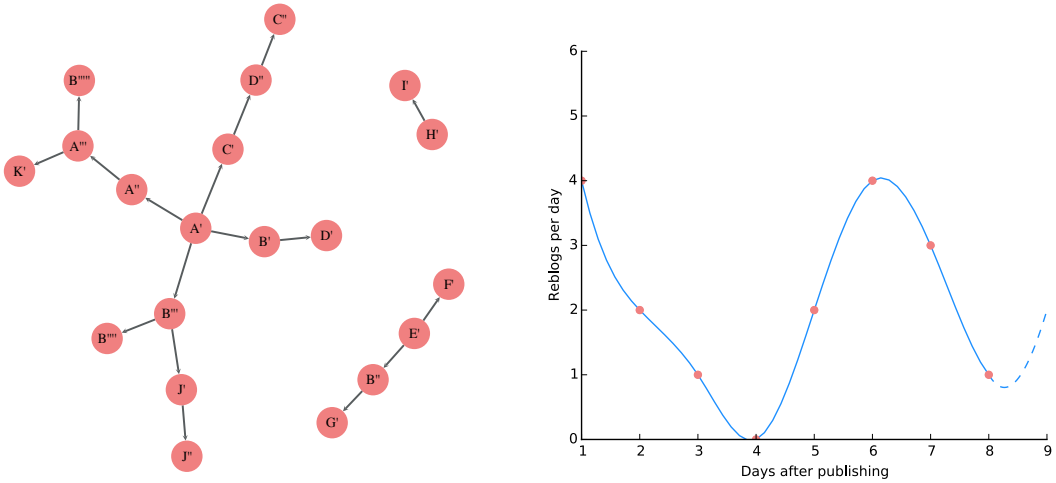


Figure 5.4 Two data representation obtained from the dataset. Left: the network data shows the relation between the users, right: The time-series data reflects the number of reblogs recorded per day

5.3.3 Data Pre-processing

Data harvesting with non-privileged access comes with its own set of challenges that must be taken into account during the harvesting process and prior to the data analysis phase. As explained above, for each post there is a list of notes that carries details about the reblogging history of each post. This list is ordered chronologically, with the most recent reblogs at the top of the list. Hence, during the harvesting process, the crawler started by getting the most recent reblogs and it kept fetching the next page of notes until it reached the very first reblog. However, because the harvesting process started long after the posts' publishing dates, a number of challenges arose, as explained below.

The first challenge is dealing with missing reblogs. Although the reasons behind their absence are not clear, they appear to be caused by reblog deletion, or due to the deactivation of the whole blog (account). The direct implication in this case is that there will be disjoint parts in the cascade network, which will be explained in Section 5.3.6.

The second challenge is related to the reblogs' timestamps. In some cases, Tumblr's API does not provide a timestamp, or when it does, it provides an invalid one. For instance, sometimes it provides an invalid timestamp from the year 2000 or even 1970, or in some cases it provides a timestamp before the posting timestamp. Hence, all of the invalid timestamps were excluded. For the missing timestamps, a simple interpolating algorithm is used. The algorithm fills in the timestamps using the time delta between two existing, consecutive and valid timestamps. In cases where there are no available timestamps to calculate the delta, the items in the list are excluded from the analysis. On average, the posts had 0.02% invalid timestamps, 15.6% interpolated timestamps and 84.38% valid timestamps.

In this thesis, two types of analyses are needed for cascades; the first is concerned with the structural aspects of cascades and the second tackles the temporal aspects. For the structural aspects, only the information about the relationship between users is needed. This is achieved by analysing the topology of the cascade network, taking into account the reblogging causality between the users. The temporal aspect, however, is concerned with analysing the relationship between the reblogging rates and the time. Thus, accurate timestamps are essential for the temporal analysis but are not needed for the structural analysis. Therefore, two conditions are proposed to accept posts for some parts of the structural analysis and all of the temporal ones. The conditions are: 1) the first reblog must be from the post's author; hence, appropriate linkage to the source is preserved, 2) the timestamps of the first reblog must be after the posting timestamp. The first constraint yields 806 posts. The second constraint excluded 10 posts, leaving 796 posts identified as "ideal" cascades. Table 5.1 summarises the dataset description.

Table 5.1 Dataset Description

No. of posts	1292
No. of posts (ideal cascades)	796
No. of reblogs/ No. of rebloggers	73,048,903
No. of reblogees	3,541,110

5.3.4 Cascade Networks Types

This section explains three types of cascade construction models utilised to construct different cascade networks from the same dataset. Each one of these networks represents a different perspective of the reblogging dynamics on Tumblr. The three network models are the **Reblog network**, **User Network** and **Event Network**. The **Reblog Network** model creates one giant network while both **User** and **Event Networks** models create four networks *User Most Recent (UM), User Least Recent (UL), Event Most Recent (EM) and Event Least Recent (EL)* networks. Thus, the overall number of generated networks is five. Table 5.2 summarises the differences between these models, which will be explained in detail below.

Table 5.2 Characteristics of cascade construction models

Model	Represents	Nodes	Edges	Mode of node connectivity
Reblog network	All posts	Users	Flow of information between users	To the parent reblogger
User network UM and UL	One post	Users	Flow of information between users	To most or least recent reblogger
Event network EM and EL	One post	Reblogging events	Reblogging event's causation	To most or least recent reblogger's event

1- Reblog Network:

The purpose of the reblog network is to provide an overview of the reblogging dynamics within the top posts in 2014 as a whole. It is a network of users that shows the reblogging relationship between users on Tumblr and links users with each other based on their reblogs, i.e., the reblogger will be linked to the user she reblogged from. This network can be loosely considered as a social network, where the edges represent follow relationship (Xu et al., 2014). This network is particularly crucial, as Tumblr's API, in its current state, does not provide information about the

social network (who follows whom); therefore, Tumblr's social network was not accessible. Nevertheless, to examine the degree to which Tumblr's reblog network resembles Tumblr's social network its structure and topology were compared with the networks of Xu et al. (2014) and Chang et al. (2014). This is based on the idea that users often reblog what their followees post (Xu et al., 2014). Thus, all of the individual cascade networks form one large network where the edges loosely represent following/follower relationships on Tumblr.

For each reblogging event, there is information about the user who reblogged the post and the user from whom the post was reblogged, i.e., the follower and the followee respectively. This information can be used to generate a social network, such that each reblogging event will add an edge between two nodes (users) in the network. Thus, the social network obtained from the reblog network is a weighted directed network, in which nodes are users and edges represent the direction of reblogging. So, if **user1** reblogs **user2** *n* times, there will be an edge from **user1** to **user2** and the weight of that edge will be *n*. Following a similar approach to that of Chang et al. (2014) and Xu et al. (2014), an edge's direction is the direction of the following relationship from the follower to the followee. Thus, a user's in-degree is the number of followers she has, while the out-degree is the number of users he is following.

The edges in this network are weighted to represent cases where a user reblogs from another user the same post more than once or different posts (reblog reoccurrences). This can also give an indication of the strength of the relationship between the users on Tumblr.

2- User Network (UM and UL):

In the second model, a separate cascade network is constructed for each post. In these networks, the nodes are the users and the edges represent the flow of information between them. This model illustrates the structure of a post's propagation as it is being reblogged by the users. It also represents the causal relationship between users, i.e., who reblogged from whom. It preserves the causal relationships between the reblogging users according to the timestamps when the reblogging occurs; hence, the flow of information and the order of reblogging is preserved.

Both users networks and the reblog network have the users as nodes in the network and the edges represent who reblogged from whom. However, there are three differences between the two. Firstly, the reblog network is weighted while the users networks are not. Secondly, the reblog network does not preserve the order of reblogging. Thirdly, the direction of edges is different on both networks, i.e., if **user1** reblogged from **user2**, there will be an edge from **user1** to **user2** in the reblog network, but in the users network there will be an edge from **user2** to **user1**, hence, the direction of the edges shows the direction of information flow. Thus, the node's

in-degree in the users network will either be one or zero because there will always be one source (parent node/rebloggee) or this information might be missing, so the in-degree will be zero. For the users networks two sub-models are used: User most-recent (UM) and User least-recent (UL) which will be explained in Section 5.3.6.

3- Event Network (EM and EL):

An events network is a network where each reblogging event is a node in the network. Each reblogging event consists of a reblogger, a rebloggee and a timestamp. Events (nodes) are linked to their parent event, i.e., the event that precedes and causes the current one. So again, the edges here represent the causal relationships between events. For example, assuming that **user2** reblogged from **user1** and **user3** reblogged from **user2**, here there are two events: **E1** (user2 reblogged from user1) and **E2** (user3 reblogged from user2) and both are nodes in the network. There will be an edge from **E1** to **E2**, because **user3** reblogged from **user2** (**E2**) who reblogged from **user1** (**E1**). For a better overall representation of the network, the first event in the cascade network (**E0**) must be the posting event. Again, two sub-models are used: Event most-recent (UM) and Event least-recent (UL) which will be explained in Section 5.3.6.

5.3.5 Construction Process: Reblog Network

A Reblog Network is represented as $R\{V, E\}$ where $\{V\}$ is the set of nodes (users or events) and $\{E\}$ is the set of edges in the network. A node $v \in V$ will be connected to its parent node $u \in V$, and an edge $(v, u) \in E$ will be drawn to indicate that user v reblogged from user u. The direction of the edge will be from v to u and it indicates the direction of information flow between u and v.

The construction process is straight forward, all the reblogging activities at all the posts will be considered in the process. Thus, the Reblog Network model construction process generates a complex and weighted network of users, connected based on their reblogging activities. To illustrate the process, assuming that there are three posts and the reblogging activities in each one of them are as listed in Figure 5.5. A 'reblog network' can be constructed for each of these posts as shown in the figure. However, as mentioned earlier a reblog network can be constructed by combining the separate reblog networks for each post in Figure 5.5 (d). Note that the edges in the reblog network are directed and weighted. As the figure shows, the thickness of the edge indicates its corresponding weight, while the direction of edges is from the reblogger to the reblogee.

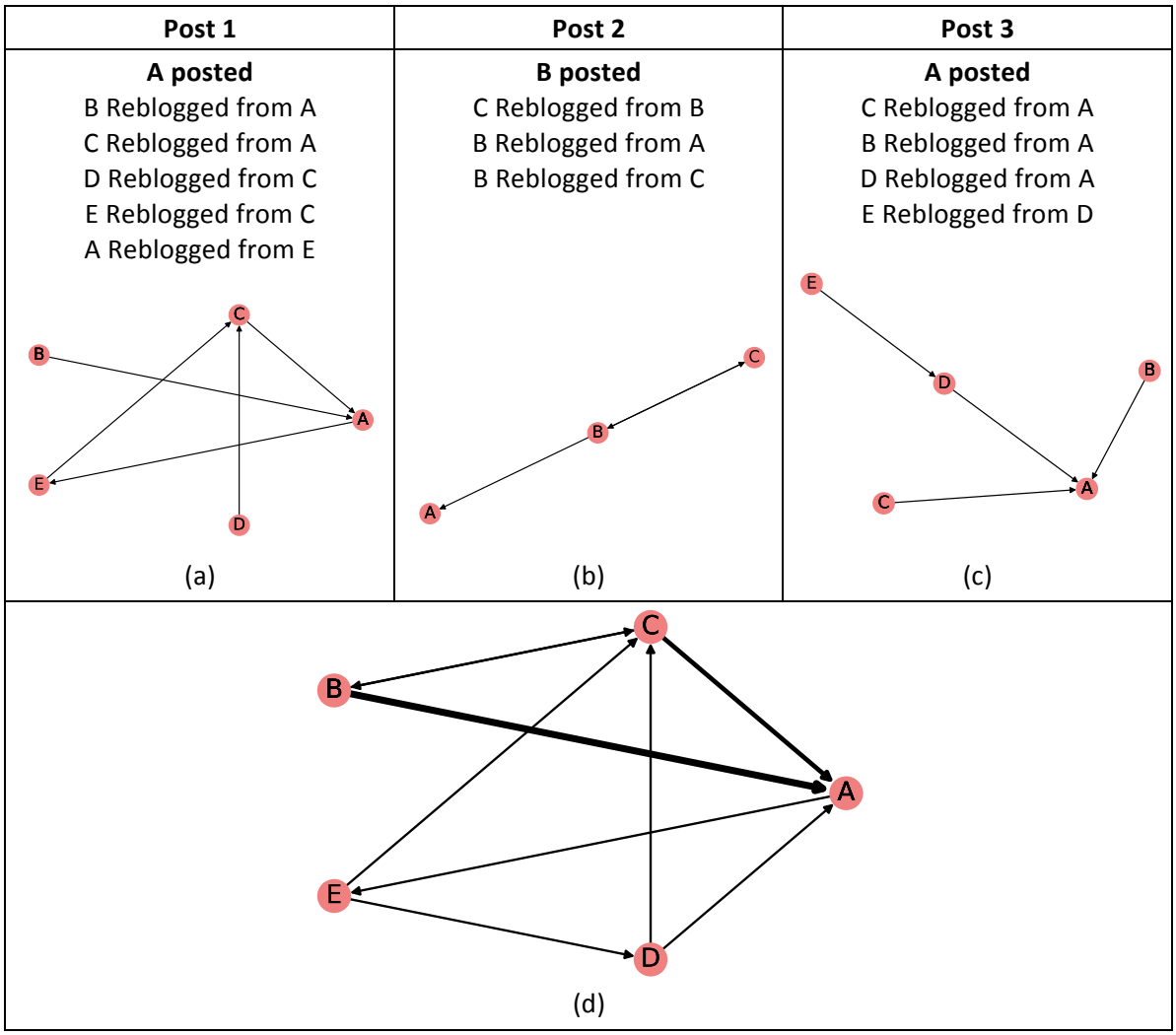


Figure 5.5 Constructing a reblog network

In this large reblog network, the set of nodes and edges are as follows:

$V = \{A, B, C, D, E\}$

$E = \{(B,A), (C,A), (D,C), (E,C), (A,E), (C,B), (B,A), (B,C), (C,A), (B,A), (D,A), (E,D)\}$.

The edge (B,A) has the highest weight that equals to three, because user B reblogged from A in the three posts. The direction of edges, from rebloggers to reblogees, indicates the flow of information following a similar approach to that of Chang et al. (2014) and Xu et al. (2014) as the resulting reblog network will be compared with the networks generated in both related work.

5.3.6 Construction Process: User and Event Network

This section explains the construction process for the four different cascade construction models explained in the previous section (UM, UL, EM and EL). A cascade network can be represented as $C\{V, E\}$ where $\{V\}$ is the set of nodes (users or events) and $\{E\}$ is the set of edges in the network. In the User Network models, a node $v \in V$ will be connected to its parent node $u \in V$, and an

edge $(v, u, t) \in E$ will be drawn to indicate that user v reblogged from user u at time t. The direction of the edge will be from u to v in the Users Network indicating the direction of information flow (in the reblog network it was the opposite from v to u). In Event Network model, an edge $(v, u) \in E$ will be drawn to indicate that event v occurs because of and after event u and the direction of this edge will be from u to v.

In User Network and Event Network models the constructed network will ideally be a tree and the root is either the post's author or the posting event. Unfortunately, not all networks are ideal; there are a number of non-ideal cases that were observed in the construction process. These cases are listed below:

Case 1: Each post must be firstly reblogged from the post author herself, but in some cases posts' authors have either never been reblogged from at all or are not the parent of the first reblog. It might be one reblog that is missing or more than one reblog.

Case 2: Sometimes reblogs are deleted and the direct consequence when this happens is that it creates cases where nodes have no parent user or event.

Case 3: Tumblr users are allowed to reblog the same post more than once, hence, it appears more than once in the notes list. This flexibility causes the following issues:

1- A parent node might appear as a child again by reblogging the same post after another user.

2- If a node reblogs the same post several times after different (or the same) parents, the node will have more than one parent; hence, it will appear as a child in different parts of the cascade network.

Case 4: Following the previous case, if a user reblogs from a user who reblogged more than once (appears in different places in the cascade network), a challenge arises in deciding which parent the user will be linked to.

To handle the cases explained above, the following approaches are used:

For **Case 1**, any cascade that does not have the author as the first parent is excluded from some of the structural analysis and all of the temporal analysis. These are cascades that are ideal but they are useful to report the reblogging dynamics in general.

In **Case 2**, to handle isolated components there are three approaches: first, ignore them and only analyse the largest component (large connected component). Second, analyse the whole network (forest), including the isolated components. The third approach is inferring the missing

links to connect the isolated fragments to the main component (Taxidou & Fischer, 2014). The first approach has a major disadvantage, because if these bits are ignored important information about the reblogging dynamics will be missing, and also the size of the cascade will be affected, as a result. To be able to reconnect the graph by inferring missing links, access to Tumblr's social network and possibly some historic reblogging information are needed to assist in a better judgement about the accuracy of the inferred links. In Tumblr's case, the social network is not accessible; thus, the second option is the most suitable one: to analyse the forest as whole taking into consideration the isolated fragments as disconnected bits of the cascade network.

To explain how **Cases 3** and **4** are tackled, assume that the following reblogging event occurs: ***A reblogged from B***. If it is a users network, one of the following scenarios will take place (Note: if a node already exists in the cascade network it can either be a previous reblogger or a previous rebloggee):

1- Neither **A** nor **B** are in the cascade network; in such case, two nodes will be added to the cascade network and an edge between them will be drawn from **B** to **A**. They will be isolated because there is no information about who is the user that **B** reblogged from.

2- **A** does not exist in the cascade network but **B** exists (either as a parent or a child). Thus, a new node will be created for **A** and will be linked to **B**.

3- Both **A** and **B** exist in the cascade network. A new node will be created for **A** and labelled differently to distinguish between the different copies of the same user. The new node will be linked to one of **B**'s.

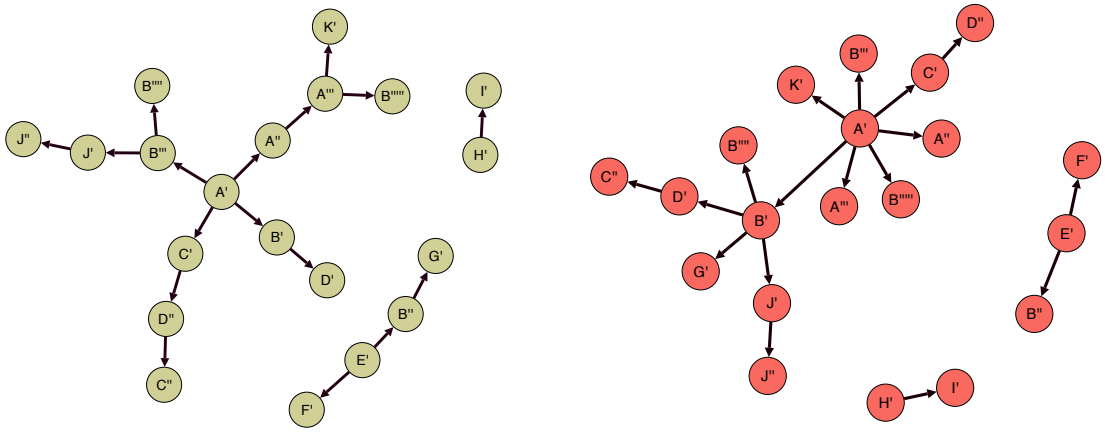
In the first and second scenarios there is always one node for user **B**. However, in the third scenario, **B** existed more than once in different threads in the cascade network. Each version is labelled differently but the problem is to decide to which version the new node will be linked.

There are many approaches in the literature that have been utilised to decide which node to choose if there are a number of possibilities to choose from (Bakshy et al., 2011; Taxidou & Fischer, 2014). These are particularly useful in the case of inferred cascades. For Tumblr's cascades two approaches can be used: the most-recent and least-recent approaches. In these approaches a node (**A** in the example) is linked to either the most-recent version of **B,** i.e. the last one that was created, or it will be linked to the least-recent version of **B,** i.e. **B**'s first appearance in the cascade network. To achieve this, a record for the most-recent and least-recent versions of each node in the network must be kept. For example, in Figure 5.6, ***E6: G reblogged from B***, in the UM network **G'** is linked to **B''**, the most-recent version of **B**, while in the UL network **G'** is linked to **B'**, the least-recent version. Similarly, in EM **E6** is linked to **E5**, while in EL **E6** is linked to **E1**.

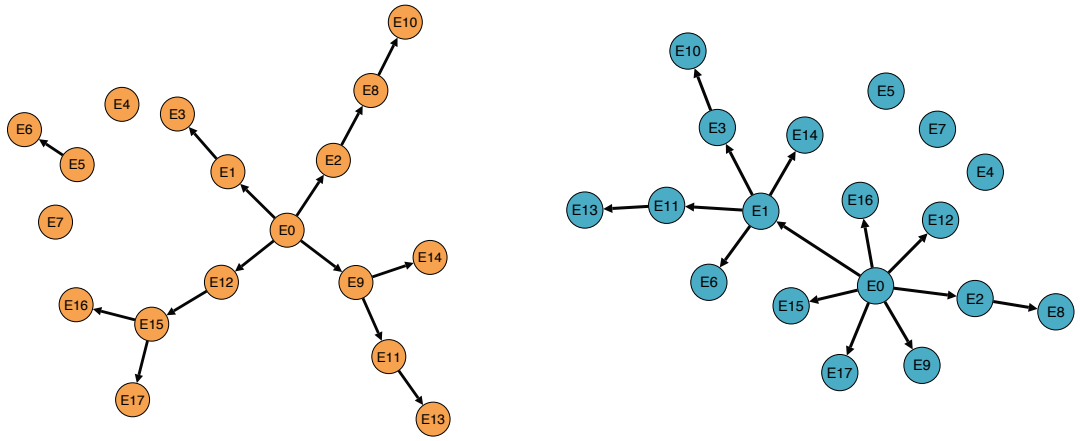
Figure 5.6 illustrates four cascade networks constructed from the reblogging information in Table 5.3: users most-recent (UM), users least-recent (UL), events most-recent (EM) and events least-recent (EL).

Table 5.3 An example of the reblogging history for one post

Notes	**Events**	**Notes**	**Events**
A posted → E0	E0	B Reblogged from A → E9	E9
B Reblogged from A → E1	E1	C Reblogged from D → E10	E10
C Reblogged from A → E2	E2	J Reblogged from B → E11	E11
D Reblogged from B → E3	E3	A Reblogged from A → E12	E12
F Reblogged from E → E4	E4	J Reblogged from J → E13	E13
B Reblogged from E → E5	E5	B Reblogged from B → E14	E14
G Reblogged from B → E6	E6	A Reblogged from A → E15	E15
I Reblogged from H → E7	E7	K Reblogged from A → E16	E16
D Reblogged from C → E8	E8	B Reblogged from A → E17	E17



(a) (b)



(c) (d)

Figure 5.6 Four constructed cascade networks (a) UM, (b) UL, (c) EM and (d) EL

The difference between the most-recent and least-recent models is that, if there are two or more options to choose from as a source for reblogging, one might assume that users will reblog from either the most-recent or the least-recent versions. The most-recent means that users will reblog from the last reblogged post, i.e. following the order of the timeline in which the new posts appear at the top. The least-recent model means that the user reblogs from the first reblog made. Any option between these two is ignored.

These two approaches have a huge impact on the structure and topology of the constructed cascade network. The fundamental difference between the two is that the most-recent model distributes the credit among the different versions of the same user in the cascade network. On the other hand, the least-recent model attributes all the credit to the first appearance of the user in the network; consequently, it creates networks that have high density around some nodes in the network, depending on the amount of the reblogging activities that was initiated by that node (user). Figure 5.7 illustrates the users most-recent network and the users least-recent network for one post in the dataset. This figure clearly shows that the least-recent models create networks where some nodes have higher out-degree, because these models link to the first appearances.

A similar approach will be followed for the events network. The differences between the events network and the users network are:

1) Events networks are slightly more condensed than the users networks, such that, rather than having two nodes for each reblogging event (the reblogger and the reblogee), there will be one node to represent the event.

2) In some cases, events networks create more disconnected networks than the users network. To explain this if ***E4 (F reblogged from E)*** and ***E5 (B reblogged from E)***, but there is no information from whom E reblogged, i.e. the event that has E as a reblogger is missing. In this case, E4 and E5 will be separate nodes not linked to each other. However, in the users networks E, B'' and F are still isolated nodes but they are still connected to each other because both B and F reblogged from E. This aspect is very important because it will amplify at scale. If there are many users reblogging from E there will be many events that are isolated and not linked.

To summarise, the following simple heuristics are used to tackle the four cases mentioned above. These heuristics are taken into consideration during the cascade network construction process for both users and events networks. The purpose of applying these heuristics is to ensure that the construction process yields accurate cascade networks that represent the flow of information and reflect the reblogging dynamics between Tumblr's users accurately.

1- A node can be a parent for an unlimited number of nodes, i.e. there is no limit on the number of children a node can have.

2- Each time a user reblogs a post it will be a child, and all the child nodes will have one parent; if a user reblog twice or more it will be added and labelled differently.

3- Whenever a new child is added to the graph, it will be linked either to the most- or least-recent parent version.

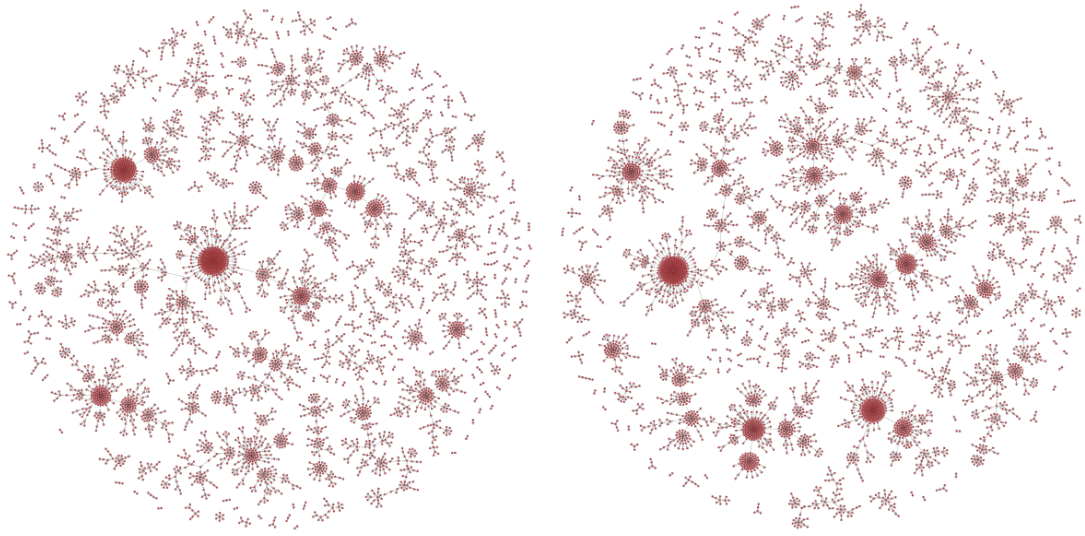


Figure 5.7 The cascade network generated from one post in the dataset, left: UM, right: UL

5.4 Chapter Summary

This chapter has discussed the research methodology followed in this thesis by dividing it into stages that have different outcomes. It has also explained how these stages fit within the Network, Web and Data Sciences perspective. The second part of this chapter presented the experimental design for this thesis, including information about Tumblr, the platform chosen to conduct the experiment. After that the Data collection and pre-processing phases were discussed, highlighting the challenges arise from the limited contextual information and the missing data. The last part of this chapter focused on the first step for the structural analysis, cascade construction models. It explained in detail how cascade networks are constructed to accurately represent the diffusion of posts on Tumblr, using three main models: the reblog network, and the users and event networks.

Chapter 6: Analysis

> 'Count what is countable. Measure what is measurable. What is not measurable, make measurable.'
>
> Galileo Galilei

This chapter will explore the results of analysing the data gathered from Tumblr's most popular posts. The previous chapters paved the path for this chapter; Chapters 2 and 3 explored the building blocks for this thesis in general: the diffusion process and cascades. Chapter 4 proposed a framework of the phenomenon of information diffusion that includes three fundamental components: the context, the content and the cascade. While Chapter 5 provided an overview of the research methodology used to answer the research questions. This chapter looks at two of these components: the context, represented by the platform's functionalities, and the result of the diffusion, the cascade. Cascade analysis receives the majority of attention in this chapter, and will be examined from two angles: structural and temporal.

This chapter is divided into four sections; the first one provides an in-depth analysis of Tumblr as a platform for content sharing, and compares its functionalities and analyses its affordances and its users' behaviour, focusing on the possible effect of each one of them on the cascades and the information available about them. The second part discusses the topology of the large reblog network generated from the most popular posts on Tumblr. It compares its structures to the structure of other networks obtained from Tumblr. The third part sheds light on the structural features of separate cascades (i.e., individual posts) as opposed to the structural features of the reblog network as a whole. The fourth part analyses the temporal aspect of cascades, i.e., it attempts to answer the following question: How do cascades grow in size in relation to the time after a post is published?

6.1 Tumblr's Functionalities

This section compares Tumblr's functionalities: reblogging, liking and commenting. It looks at some of the other affordances of Tumblr and discusses their impact such as, the ability to reblog more than once (reblogging reoccurrences), reblogging across categories and reblog deletion.

6.1.1 Cascade Size

The average cascade size of posts was 56539 (median: 36771), and 51876 (31493) respectively for the ideal cascades). The maximum cascade size was 581895 while the minimum was 3 (Figure 6.1 and Figure 6.2). The distributions in Figure 6.1 show that about 78% of the cascades were reblogged 10000 times or more but only 18% were reblogged more than 100000 times. The difference in cascade sizes is a direct impact of the fact that these posts were selected as the top ones in 57 categories curated by Tumblr's staff. Heavy-tailed distributions of cascade sizes have been widely observed in previous studies. Here, these posts were selected as the top ones, yet they exhibit similar characteristics, highly diverse cascade sizes, with larger cascades being a minority. The notion of large cascades will be discussed further in the next chapter.

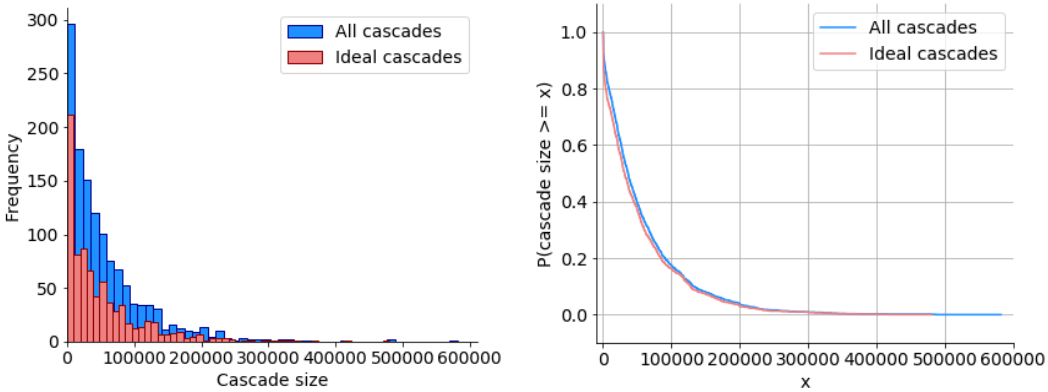


Figure 6.1 The distribution of cascade sizes, histogram and CCDF distribution for all cascades and ideal cascades

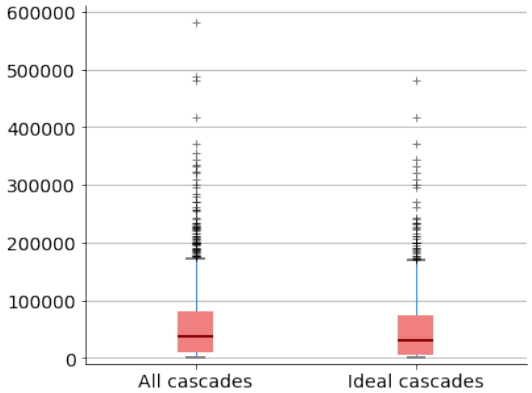


Figure 6.2. Cascade sizes boxplots for all cascades and ideal cascades, showing that the majority of cascades sizes are below 100000 reblogs

Figure 6.3 shows that the *chill* category contains some of the largest cascades in the dataset; even the average cascade size for this category is higher than the other categories (Figure 6.3 and Figure 6.6). The *tumblropenarts* category has remarkably small cascades; one possible explanation is that this category contains posts from the *tumblropenarts* blog, which is known as the official hub of art on Tumblr where users submit their work to be published on it[6]. Another category that has remarkably small cascades is *kale*, possibly because it attracts an audience with a very specific interest in vegetarian and vegan food. On the other hand, The *Tumblr gets deep* category has a number of outliers with higher number of reblogs compared to the others in the same category. Posts in the *Tumblr gets deep* category basically contain a text or a photo that becomes popular as users reblog and comment on it then the thread of comments gets popular as one piece (See an example in Figure 6.5). Categories like *animals, feminism, healthcare, lgbtq, lyrics nostalgia, plants* and *tattoos* have similar upper bound cascade sizes. However, most of the cascades in *feminism* are large compared to the others. Also, contrary to the common perception, posts that belong to categories the *3d gif* and the *gif art* are not as large as expected. Nonetheless, knowing that Tumblr's is famous as a place where users mostly share gifs, more than 84% of the posts in the dataset, in all categories, are photos (Figure 6.4).

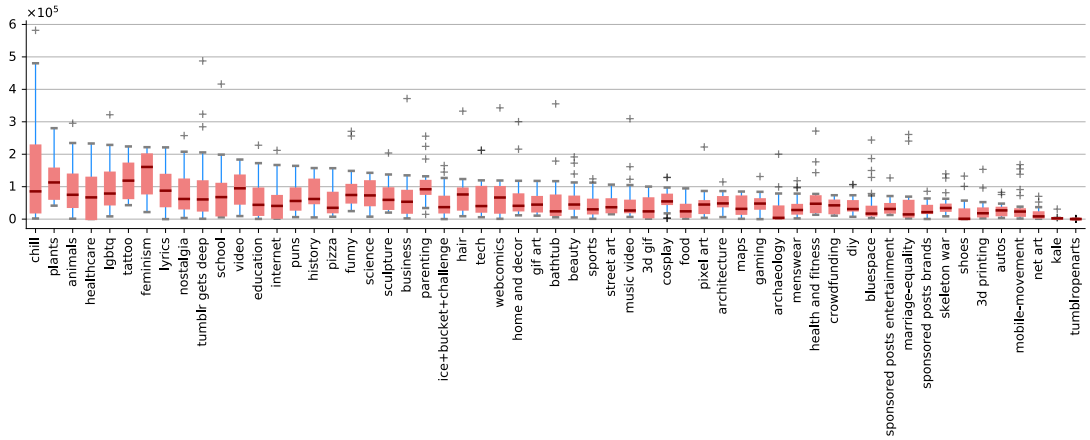


Figure 6.3 Cascade sizes boxplots by category

[6] *tumblroepnarts* was recently replaced with art.tumblr.com

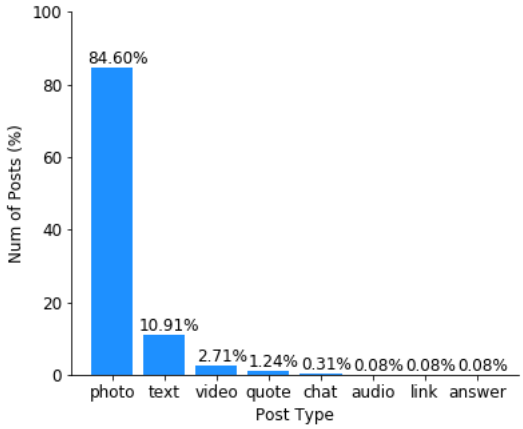


Figure 6.4 Percentages of posts by type, photo posts are the dominant type

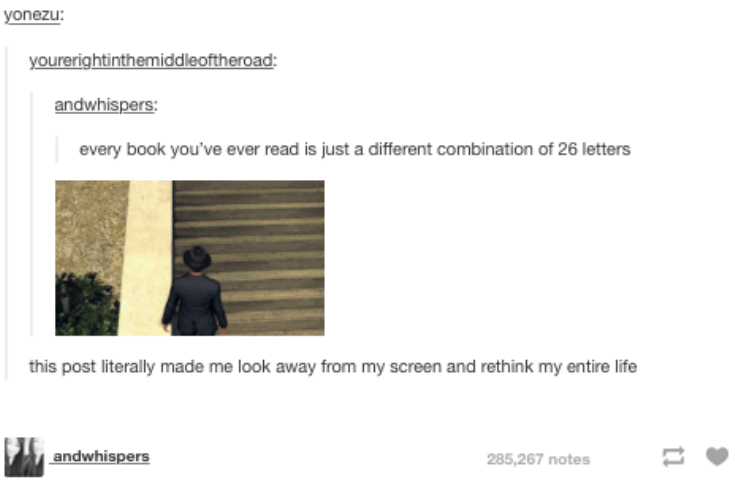


Figure 6.5 An example of a post in "Tumblr gets Deep" category, showing a thread of comments that gets reblogged as one post

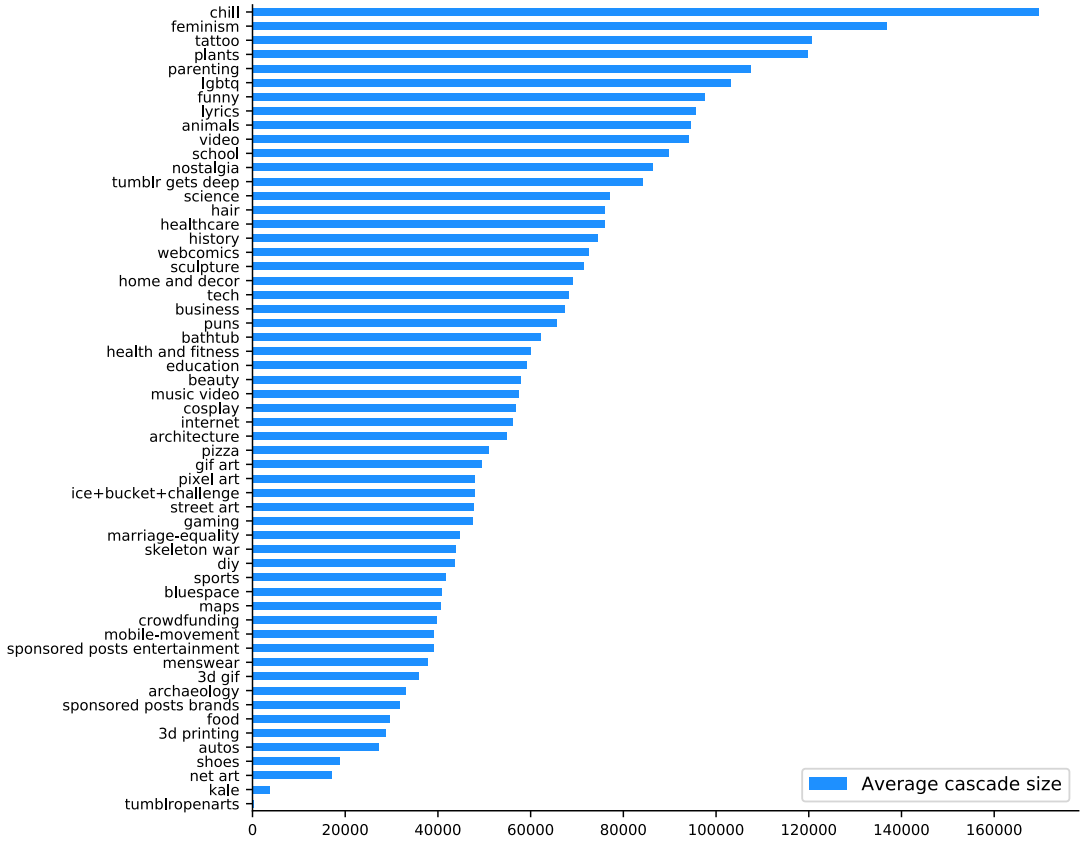


Figure 6.6 Average cascade sizes by category, the category *chill* has some posts with the largest cascade sizes

6.1.2 Reblogging Across Categories

Figure 6.7 enhances the idea of Figure 6.3, it illustrates the relation between the number of posts in each category and the corresponding average cascade size of that category. The number of posts in the majority of categories ranges between 10 and 30, while few categories stand out both in terms of the number of posts and the average cascade sizes, e.g., *chill*, *feminism*, *tattoo* and *plants* with larger average cascades sizes. *Tumblr gets deep* has a higher number of posts and relatively moderate cascade size. *tumblropenarts* and *shoes* stand out in terms of the number of posts they have, but they have smaller cascade sizes.

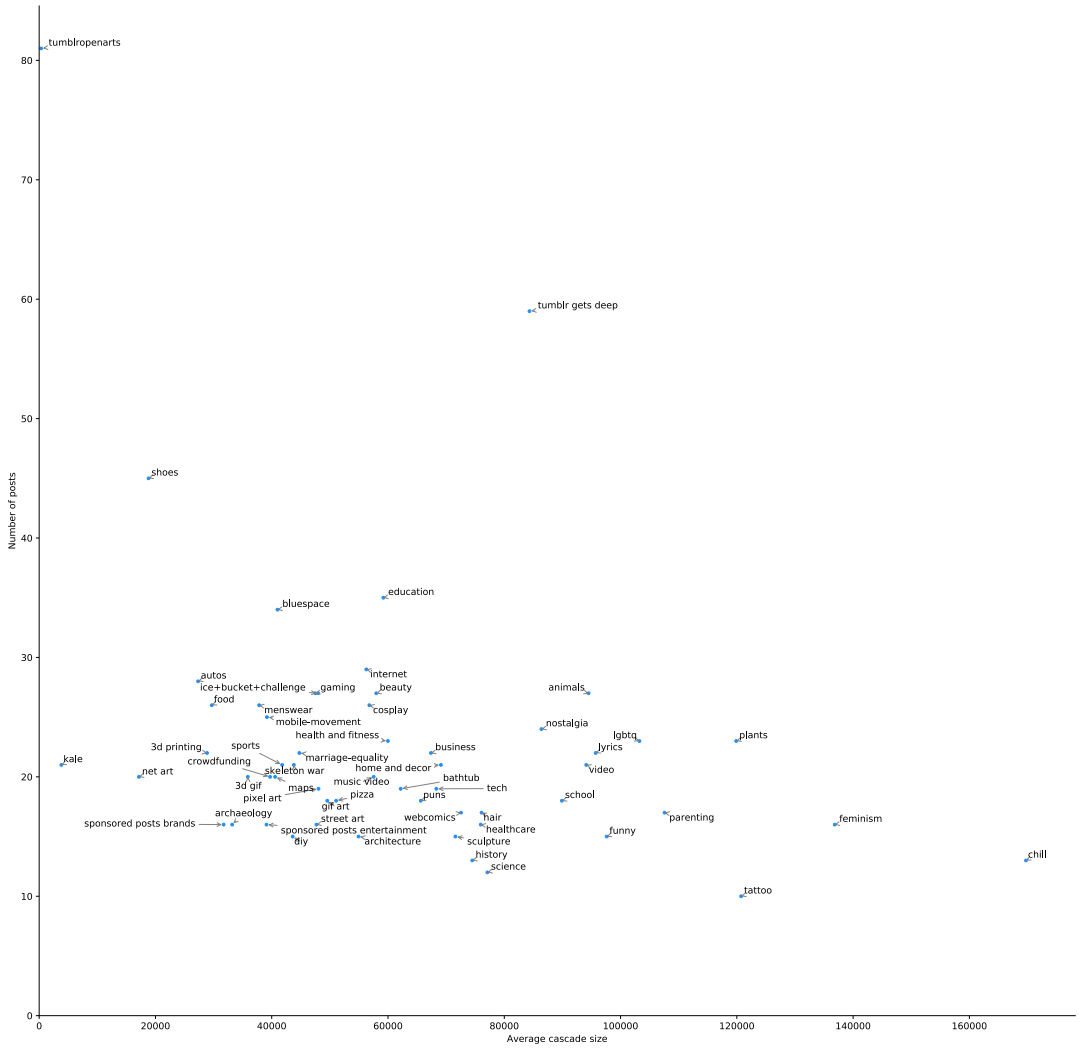


Figure 6.7 A comparison between the number of posts and the average cascade size for all categories

6.1.3 Liking

On Tumblr, Liking is the second main functionality available for the users. It allows users to express their admiration towards a post without adding it to their blogs. Figure 6.8 shows that the total number of reblogs per category is higher than the total number of likes which means that reblogging is more popular among Tumblr users in comparison to liking. However, they are correlated and relationship between the two is linear, so as the post becomes more popular by being reblogged, more people attempt to like the post (Figure 6.9). The computed Pearson correlation between the two is positive and is equals to 0.90 ($n = 1292$, $p < 0.001$, $r = 0.90$).

On average, there are 7.99 likes per 10 reblogs (median: 7.25). Having the option to like and reblog the same post, on average 22.37% of users reblogged and liked the same post (median: 22.84%).

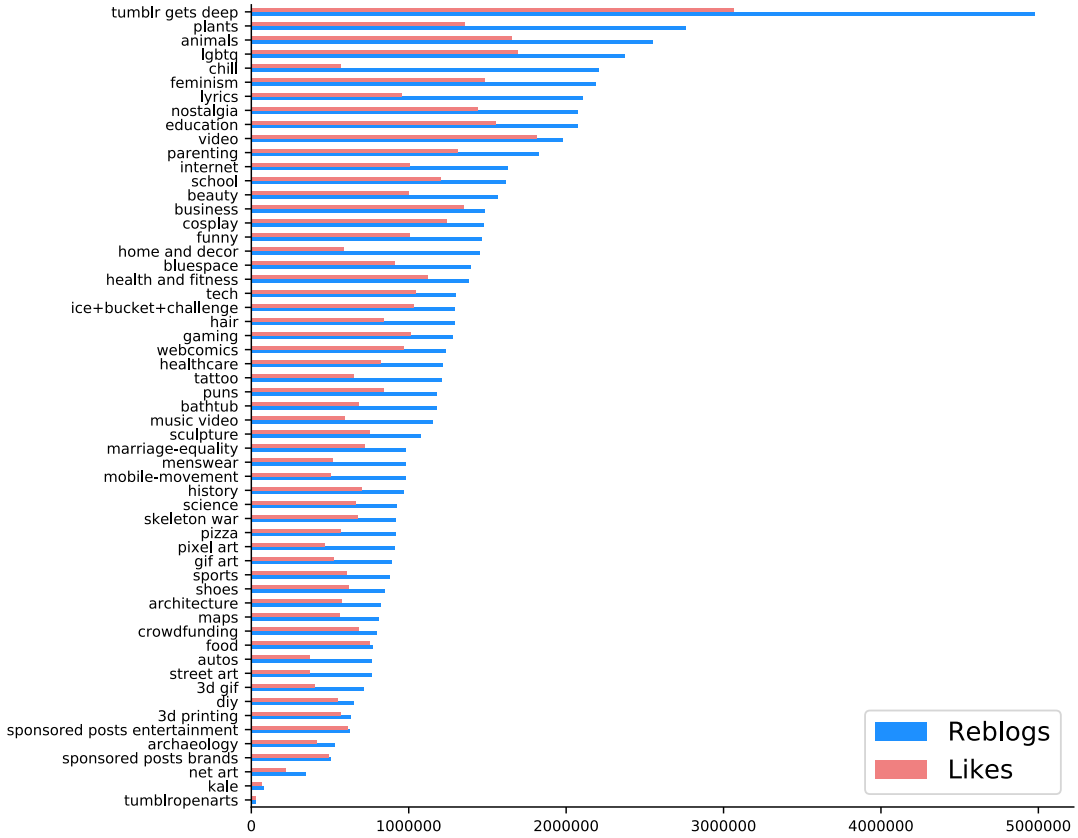


Figure 6.8 A comparison between likes and reblogs by category, showing that reblogging is more popular than liking

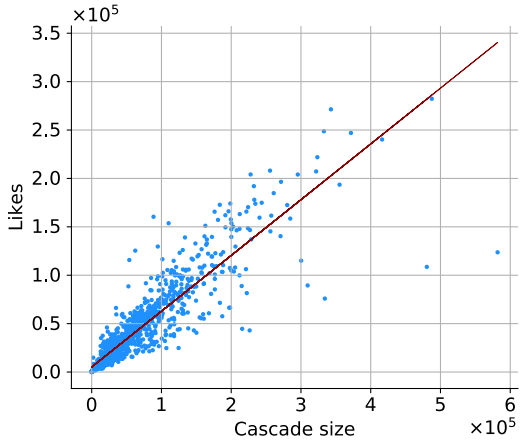


Figure 6.9 Scatter plot of cascade size (the number of reblogs) vs. the number of likes per post; the relation between the two is linear

6.1.4 Commenting

On Tumblr, users can add comments as they reblog a post. Before introducing @mention in 2014, comments added with reblogs were the only way for users to communicate publicly on Tumblr. Thus, comments were suggested as an explanation for the reoccurring reblogging behaviour

(Chang et al., 2014). However, text analysis of comments shows that only a tiny percentage of comments include @mention in their text (0.32%).

The number of reblogs with comments is exceptionally low in comparison to the total number of reblogs (Figure 6.10): only 1.55% of reblogs are with comments. The average number of comments in each cascade equals 877 (median =454.5). On average, there are 0.16 comments per 10 reblogs (median: 0.13). However, the relation between cascade size and the number of comments in a cascade is linear, using Pearson correlation R = 0.63 (n = 1292, $p < 0.001$), which means that, in most cases, as the cascade size increases the number of reblogs also increases (Figure 6.11). Regardless of the cascade size (the number of reblogs), there are three categories that have higher upper bound comment rates; these categories are *video, healthcare, and feminism* (Figure 6.12). On the other hand, *Tumblr gets deep*, a category that often invokes users to comment, has the highest number of outliers for some posts while the rest have low comment rates. Additionally, posts with videos have higher comment rates per post, which is consistent with the fact that posts that belong to the *video* category have high numbers of comments per post (Figure 6.13). Photo posts have many outliers, but the majority of these posts have low commenting rates.

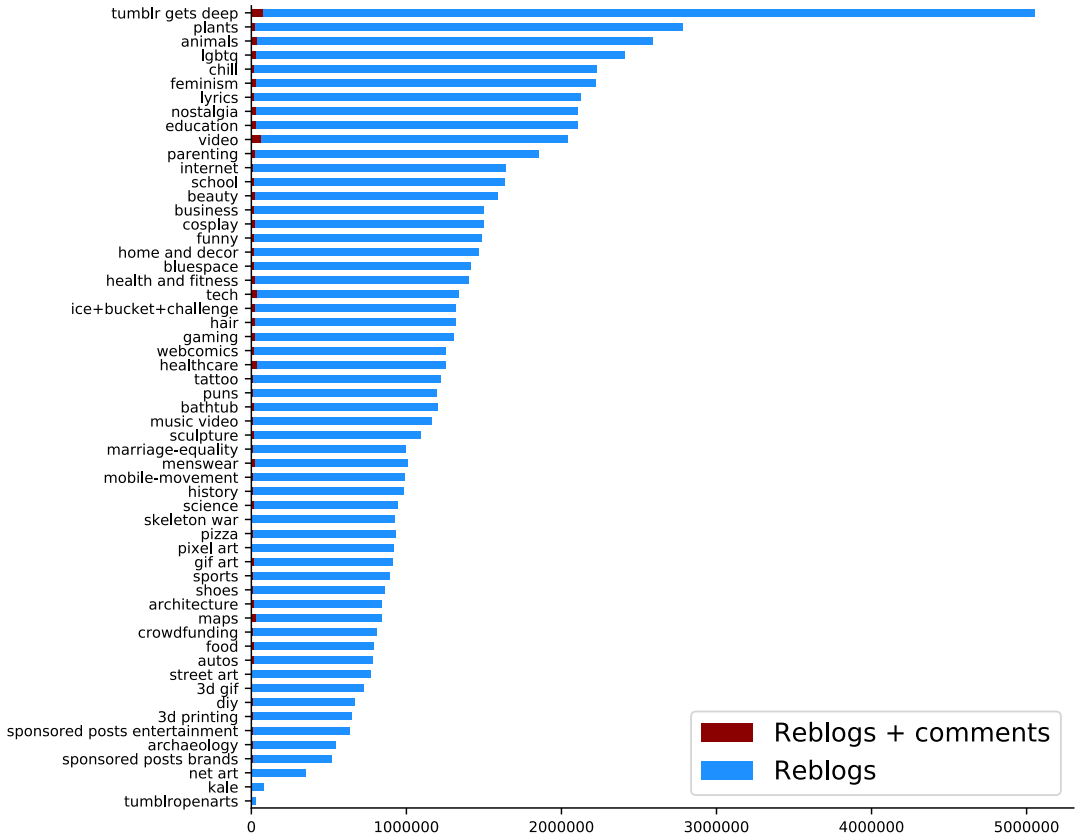


Figure 6.10 A comparison between reblogs with comments and total reblogs, by category; the number of reblogs with comments is extremely low

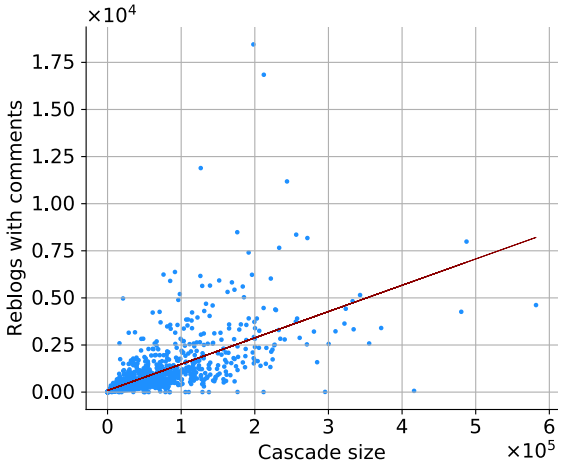


Figure 6.11 The relation between cascade size (total number of reblogs) and the number of reblogs with comments; the relation between the two is linear

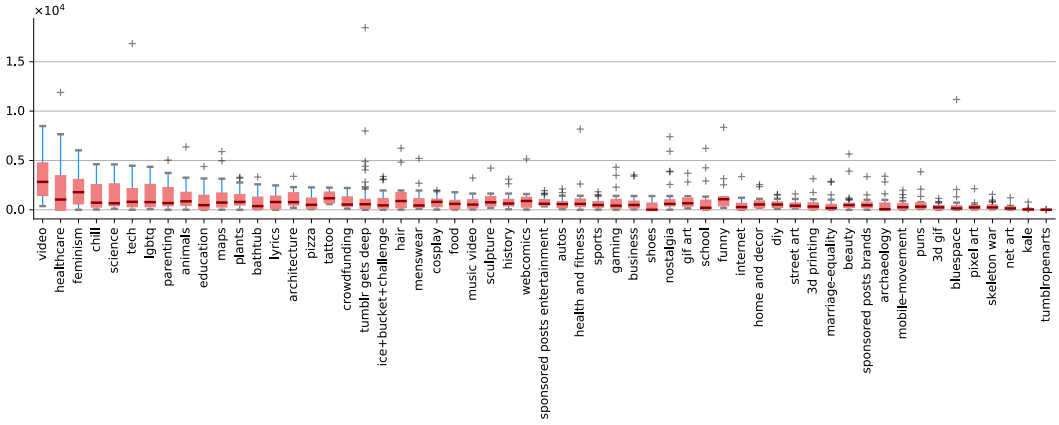


Figure 6.12 Boxplot of number of comments by category

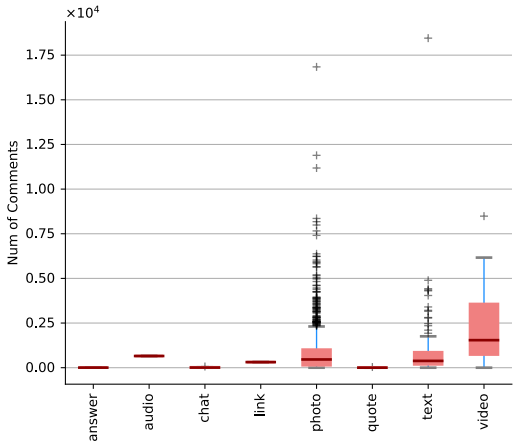


Figure 6.13 Boxplot of number of comments by post type

6.1.5 Reblogging Rate (Reblogging Reoccurrences)

Unlike most of the other platforms, Tumblr allows its users to reblog the same post more than once. This particular ability is said to be used as Tumblr's users' means of communication (Chang et al., 2014). This part will look at the rate at which this happens: i.e., if Tumblr users are allowed to reblog more than once, how often does this occur? Thus is particularly important in view of the finding that this ability is being used as a way of communication.

Figure 6.14 illustrates the distribution of reblogging reoccurrences. The median reblogging rate per user in one post equals one, which means that even though Tumblr allows reblogging more than once, most users reblog a post on one occasion only. In fact, only 7.33% of the reblogs in the dataset are reoccurrences (two and more), while the majority (92.66%) are one reblog per user in each post. The maximum number of reblogging reoccurrences in one post is 139 reblogs. Figure 6.15 plots the reblog reoccurrence counts for all categories. There are some cases where higher reblogging reoccurrence are detected, e.g. in the *chill* category, but this rarely happens. For the majority of posts, the number of reoccurrences in most categories is less than 5000.

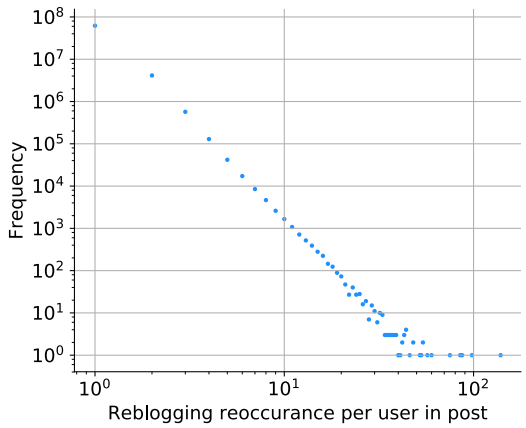


Figure 6.14 The distribution of reblogging reoccurrences per user in a post (log-log scale), showing that most of the time the user reblogs the post once only

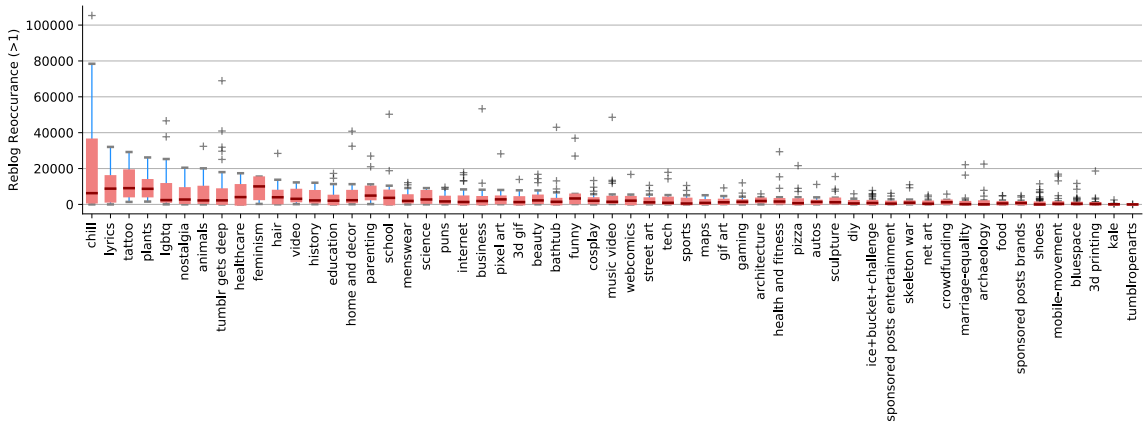


Figure 6.15 Boxplot of reblog reoccurrences by category, showing that chill has one of the highest values compared to the rest of categories

6.1.6 Reblog Deletion

In an ideal cascade, based on the cascade construction model used, the first reblog must be from the same user (post's author), i.e., the first reblog must have the post's author as a source (parent). Most of the posts had posts' authors as sources (parents) of the first reblog (806 post, 62.38%). However, that was not the case in 486 posts in the dataset (37.61%). In these posts the authors were not the source for the first reblogs. There are three posts in which the author was not the source for any reblog in the whole cascade. All of these posts are relatively large (cascade sizes: 35773, 26783, 33067) and they all belong to *Tumblr gets deep* category. This occurs if the author or a reblogger deletes their post/reblog. Hence, there will be some missing data and it is difficult to estimate the amount of the deleted posts/reblogs before started the data collection process, it might be one reblog only or many consecutive reblogs. However, there is a way to estimate deletion during the cascades' construction phase. A deletion of a reblog(s) is detected if there is a reblog event but the source (rebloggee/parent) doesn't appear as a target (reblogger/child) in all of the earlier reblogging events. On average, there are 1150 deleted reblogs (median: 747) in each post. In other words, there are 27 deleted reblogs for each 1000 reblogs. Moreover, around 60% of the posts have 1000 deleted reblogs or less, while the rest have higher deletion rates.

6.1.7 Discussion and Remarks

This section has focused on analysing the way Tumblr users utilise the available functionalities, especially within the most popular posts on Tumblr. The following are some remarks and discussion about the findings obtained from the analysis carried out on the users' activities around Tumblr's most popular content.

Reblogs vs. likes vs. comments

One of the most important findings is the fact that reblogging is slightly higher than liking on Tumblr: for each 10 reblogs there are 7.9 likes. The high reblogging rates mean that users on Tumblr are highly engaged with the shared content. I.e., reblogging entails that the posts are added to users' blogs, which indicates their level of engagement with the published content. This finding is aligned with Tumblr's CEO's remark about the platform's high reblogging rates: "Ninety percent of content on Tumblr is actually reblogged". On Twitter, which is a platform that provides similar functionalities: the ability to spread and to 'favourite' (which has been changed to *like* recently) content, 43% of tweets get at least one favourite and 36% of them get at least one retweet (ENGE, 2014). Hence, it seems that favouring is more popular on Twitter than retweeting, while it is the opposite on Tumblr, especially for the most popular content.

Another interesting finding is that the number of comments on Tumblr is remarkably low; there are only 0.16 comments per 10 reblogs and the reblogs with comments comprise only about 1.55% of the total reblogs. On Twitter, 0.7% of tweets get replies (ENGE, 2014), but, Liu, Kliman-Silver & Mislove (2014) report that about 35% of tweets are actually replies. This means that on Tumblr most of the communications between users is non-textual: they like and reblog but rarely express their opinions in textual form. This is apparent by looking at the percentage of post types in Figure 6.4: only 10.90% of posts are of type text, and even these posts show very low degrees of commenting, according to the boxplots in Figure 6.13.

On the other hand, some users (22.37% on average) liked and reblogged the same post; i.e., attempting two social interactions on the same post. This behaviour shows the degree of 'interestingness' in the post's content (Meier et al., 2014), measured by its likeability and rebloggability and shows that users are more interested in the posts' content.

Reblogging more than once: reblogging reoccurrences

The ability to reblog content more than once is not exclusive to Tumblr. Twitter allows its users to retweet a tweet again; however, the difference between Twitter and Tumblr in this regard is that when a user reblogs the same content more than once it will keep the old reblogs and it will generate a new post with a new ID for the reblogged post. However, on Twitter, when a user retweets a content it will resurface on the tweet stream in a way that brings the same old tweet up again.

However, comparing reblogging reoccurrences and commenting rates across categories, it can be noted that the number of comments is lower than reblogging reoccurrences, meaning that most re-reblogging attempts are without comments (Figure 6.16). This particular behaviour raises

the following question: why do the users re-reblog the same content if they are not using them for communication? A possible answer might be that these users might be bots, or it might be that these users are deliberately reblogging the same post at a different time of the day to get attention from a different audience.

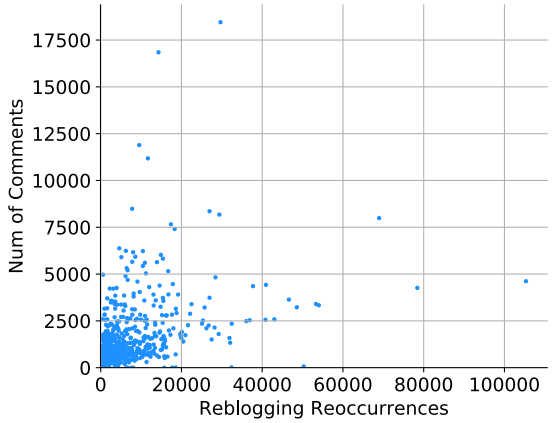


Figure 6.16 The relation between the reblogging reoccurrences and the number of comments, showing that the number of comments is significantly lower than the number of reblogging

6.2 Tumblr's Reblog Network

This section discusses the topology of Tumblr's top posts' reblog network, which is a giant network constructed from all of the reblogging information. As explained in Chapter 5, there are two purposes to construct such a network: firstly, to analyse the collective reblogging dynamics as a whole on a predetermined part of the network (the popular posts in this case) and secondly, to examine the extent to which this network resembles the social network (Xu et al., 2014). Guided by these two purposes this section analyses the topology and structure of the network obtained and compares it with two similar networks obtained by Chang et al. (2014) and Xu et al. (2014).

All of these networks were constructed from Tumblr; however, the data utilised to construct each one of them is different. The first one, in the present study, is the reblog network constructed from the cascading dynamics in the top posts in 2014. Xu et al.'s network is similar to it, as it was constructed from the cascading dynamics; however, their data harvesting method was different. Their network is constructed from all of the reblogging dynamics recorded within four months. On the other hand, Chang et al.'s network is actually Tumblr's social network. The comparison between the three is based on Xu et al.'s remark that networks constructed from cascading dynamics can be loosely considered as a social network. Table 6.1 below summarises

the results of the comparison, and the subsections that follow will discuss the comparison in detail.

Table 6.1 Reblog network's topology in comparison with other networks from Tumblr

		Reblog Network	**Chang et al.**	**Xu et al.**
Type		Reblog network	Social network	Reblog network
#nodes		6,926,665	62.8 million	18, 367, 173
#edges		51,421,042	3.1 billion	999, 548, 135
Direction		Directed	Directed	Directed
Density		0.000214%	0.000157%	0.000592%
Reoccurrences (edges)		30%	-	86%
Edge's weights	Max	296	-	-
	Min	1		
	Mean	1.42		
	Median	1		
Reciprocity		7.7%	29.03%	8.8%
In-degree	Max	251626	4.06 million	-
	Min	0	-	
	Mean	7.42		
	Median	1		
	% equal 0	~49%	41.40%	
Out-degree	Max	960	155.5k	-
	Min	0	-	
	Mean	7.42		
	Median	3		
	% equal 0	~1%	12.74%	
In-degree/out-degree	No. of node where in-degree and out-degree are non zero	3466065 (~50%)	-	11, 259, 743 (~61%)
	In-degree > out-degree	18.09%	-	22%
	In-degree > out-degree (10 or more)	6.67%	-	2.7%
	In-degree < out-degree	70.19%	-	-
	In-degree = out-degree	11.7%	-	-
Components	% of nodes in GCC *(undirected)*	99.74%	99.61%	-
	% of nodes in GCC *(directed)*	44.71%	-	-

6.2.1 Density

Network density (computed by dividing the total number of edges by the number of possible edges) determines the degree of connectedness in a given network. In general, the density of all of the three networks is significantly low, which means that Tumblr's network is sparse and the degree of connectivity on Tumblr is low. The low degree of connectivity on Tumblr aligns with that of most of the other platforms: Twitter 0.0214% (Java et al., 2007) and 0.00016% (Kwak et al., 2010); Facebook 0.000026% (Ugander et al., 2011) and Blogosphere 0.0068% (Shi et al., 2007). However, the density of the reblog network is higher than the density of Tumblr's social network (Chang et al, 2014) and lower than Tumblr's large reblog network (Xu et al. 2014), which means that Facebook's network is the 10 times less dense than Tumblr's social and reblog networks, which means that Facebook's network is less connected. Twitter's and blogosphere are the most connected networks; in fact Twitter's social network is 100 times denser than Tumblr's social and reblog networks. In general, the reblog networks are slightly denser than the social network. Which means that these networks tend to be more connected than the social network as these are constructed based on common interests.

6.2.2 Reoccurrences

In the dataset, there are 73,048,903 independent reblogging events in 1292 different posts. Each reblogging event adds an edge between two nodes (users). In the case that each user reblogs another user only once, then, the number of edges in the network must be equal to the total number of reblogging events. However, the number of edges in the obtained network is only 51,421,042, while the number of nodes equals 6,926,665. This means that almost 30% of the reblogging events are reoccurrences, in which users reblog one post or different posts from the same user. The computed percentage (30%) is higher than the one computed in Section 6.1.5 (7.33%), because here it is computed collectively across all posts, i.e., it does not differentiate between posts. Nonetheless, this percentage is high, especially given that it is observed within a concise set of posts (the most reblogged posts in 2014). The percentage of reoccurrences in the reblog network is significantly lower than the one obtained in Xu et al. (2014), in which almost 86% of all reblogging events were reoccurrences. The difference between the two percentages suggests that the top posts tend to attract a wider audience; hence, the reoccurrence rate is lower than for ordinary posts. Figure 6.17 illustrates the edges' weight distribution; it shows that for the majority (79%) of edges, their weight is equal to one. In fact, about 99% of the edges' weight is less than or equal to ten. Thus, the top posts community is sparse and the rate of reoccurrences is low compared to that found by Xu et al. (20140, since the edges' weights are an indication of the strength of the relationship between any two users.

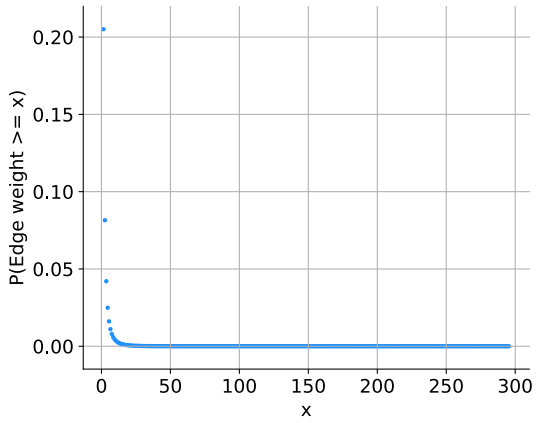


Figure 6.17 Left: the distribution of edge weight, showing that for the majority (79%) of edges, the weight equals one

6.2.3 Reciprocity

In any directed network, reciprocity is the percentage of reciprocal edges in the network, i.e., the percentage of edges in which the users reblog each other (A reblogs B and B reblogs A). The reciprocity of Tumblr's reblog network is 7.7%, meaning that only 7.7% of the edges are mutual. Our network's reciprocity percentage is similar to that in the reblog network of Xu et al. (2014) (8.8%). On the other hand, Tumblr's social network has a higher reciprocity than the reblog networks, at 29.03% (Chang et al. (2014). The low reciprocity in the reblog network is not surprising, because of the nature of the reblogging behaviour; when a user follows another user it is not necessary that he will reblog from that user. Moreover, popular accounts will always attract users to reblog from them but not the other way around.

Xu et. al. (2014) compared the reciprocity of their reblog network to the retweet and mention networks on Twitter (5.5% and 18.6% respectively). They suggest that reblogging on Tumblr indicates stronger social connections than Twitter's retweet but reblogging is weaker than Twitter's mention. On Tumblr, because the reblog networks are constructed from the reblogging activities, reciprocal connections are created either when the users simply reblog posts from each other, or when they engage in conversations via reblogs. Therefore, reblogging acts both as a content sharing functionality and a mechanism of communication, like Twitter's mention. Thus, the low reciprocity means that users rarely engage in conversations using reblogs within the top posts' reblog network.

6.2.4 Degree Distribution

The reblog network was constructed to resemble Tumblr's social network, i.e., it loosely represents follower-followee relationships, utilising the reblogging activities of users. Thus, a

user's in-degree is the number of rebloggers she has, and a user's out-degree is the number of the users she reblogs from (i.e., the followers and the followee respectively, using the social network terminologies).

The user who has the highest in-degree (the most reblogged user in the dataset) was not an author of any post; nonetheless, he was reblogged 606303 times in 199 posts (606303 is the total number of reblogs and 251626 is the total number of distinct reblogs, hence, the in-degree). In fact, only 23% of the top 100 users with high in-degree (11.9% of the top 1000 users) are authors of posts. On the other hand, the user with the highest out-degree is 'yearinreview', followed by another user (not an author of any post) who reblogged 947 posts. Around 49% of the users have zero in-degree, meaning that no one reblogged from them; this percentage is very close to that recorded on Tumblr's social network (40%) in Chang et al. (2014). In the network, these users have no impact on the spreading of content on Tumblr; hence, this suggests that they have few or no followers.

Figure 6.18 shows the degree distribution of Tumblr's reblog network, using the complementary cumulative distribution function (CCDF). The Pearson correlation between the in-degree and the out-degree in the reblog network is 0.0816 ($n = 6926665$, $p < 0.001$), which is lower than in Tumblr's social network (0.106) Chang et al. (2014). The curves show that both nodes' in-degree and out-degree drop significantly after 100. In general, the nodes' out-degree is lower than their in-degree, meaning that the majority of users in Tumblr reblog network tend to be reblogged from rather than be rebloggers themselves. The percentage of users with non-zero in-degree and out-degree is about 50%. Overall, within this subset of users there are 18% that have higher in-degree, i.e., have more rebloggers (followers) and only 6.67% of these users have in-degrees higher than out-degrees of ten times or more. These percentages are very close to the ones in Xu et al.'s (2014) reblog network, 22% and 2.79% respectively. However, most users have higher out-degree (70%) while the rest have equal in-degree and out-degree (12%). This suggests that in Tumblr's reblog network, most users reblog others more than they are being reblogged. This means two things: first, a user is more likely to reblog a post than be reblogged, which means that the community is sparse and users rarely engage in conversations using comments in the reblogs.

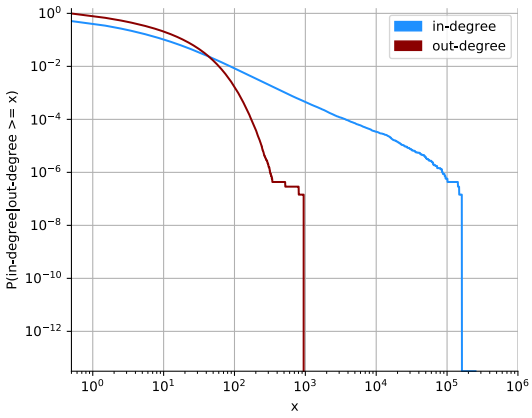


Figure 6.18 The degree distribution of Tumblr's reblog network (CCDF)

6.2.5 Components

In order to find the giant connected component (GCC), the reblog network was treated as an undirected one, to be able to compare it to Tumblr's social graph in Chang et al. (2014). About 99.74% of the nodes belong to the giant component; the percentage is very close to that in Tumblr's social graph (99.61%), which means that almost all the users who reblog the top posts are connected to all the other users, i.e., users can be reached from any user in the network, this means that the reblog network is highly connected and users reblog from each other across the different posts. When the network is treated as directed, the percentage drops to 44.71%; thus, even when the direction is taken into consideration, a significant part of the network is strongly connected and the users can reach each other via relatively short paths.

6.3 Structural Features of Cascades

Analysing the topology of the whole reblog network provides insights about the reblogging dynamics that appear within a selected part of the network (the most reblogged posts). However, analysing individual cascade networks (one per post) helps in providing further details about the reblogging dynamics on individual posts. The structural and temporal analysis make it possible to go beyond just reporting the total number of reblogs to explain how it happened, step by step, as the cascade grows.

To analyse the structure of the constructed cascade networks a number of measures were utilised, the branching factor (the out-degree), the subcascade size and depth, the number of paths and path lengths. The first two measures are related to the influence of individual users, while the rest act as estimators of the cascade's overall structure. The purpose of the structural analysis is to quantify and assess the networks' characteristics, which is very practical as the

majority of the posts in the top post dataset are large with very complex structures. These measures were applied to the individual cascades, constructed using both user and event models, including the most-recent and least-recent models, meaning that it is applied on four different cascade networks: UM, UL, EM and EL for each post (Section 5.3.6). User models show the influence of individual users while event models represent the cascade as a series of reblogging events that were influenced by a particular reblogging event.

6.3.1 Branching factor: How many users does a user influence?

Using the loose definition of influence in social networks (Cha et al., 2010; Kwak et al., 2010), the branching factor is simply the number of children a node has (the out-degree); i.e., in Tumblr's context, the number of users who were influenced by a particular user to reblog the post. This means that the branching factor works as an estimate of how much a particular node in the network (a user or an event) contributed to the cascade's overall size: the higher the branching factor of a node the higher its influence.

One way of analysing branching factors is to differentiate between three classes of users: those who have no impact (Branching factor = 0), those who at least influenced one user (Branching factor = 1), and the others who had an impact on more than one user (Branching factor > 1). Figure 6.19 illustrates the percentages of nodes with zero, one, or more branching factors, respectively, in each cascade.

Figure 6.19 shows that most of the nodes (users) in all of the four models have zero children (mean 67.44%-70.33%), meaning that in most cases the cascade stops when it reaches these users. About 17.95% - 20.21% of the nodes have one child at most, and only few nodes have contributed to the growth of the cascade by having more than one child (11.70%-12.33% on average). The average percentage for the nodes with one child is similar to the percentage found in Liben-Nowell & Kleinberg (2008), which represents 19.04%, where a 10-fold simulation was run to model information flow on Internet chain-letters. These percentages show that, in most cases, cascades die out soon after reaching a large number of users who have zero reblogging children.

Table 6.2 shows the mean and median percentage of nodes with different branching factors across the four construction models. The percentages for each model are very close to each other. The least-recent models (UL and EL) yield slightly higher rates of nodes with zero children, because they link rebloggers to their parents' first appearances (least-recent) rather than distributing the children on the other copies as well.

Table 6.2 The number of nodes with branching factors =0, 1 or more, it shows that among the four models the results are similar to each other

		UM	UL	EM	EL
BF = 0	**Mean**	67.44%	68.73%	69.10 %	70.33%
	Median	65.63%	67.13%	67.60%	68.94%
BF = 1	**Mean**	20.21%	19.20%	18.90%	17.95%
	Median	21.08%	19.98%	19.73%	18.74%
BF >1	**Mean**	12.33%	12.05%	11.98%	11.70%
	Median	13.03%	12.77%	12.63%	12.34 %

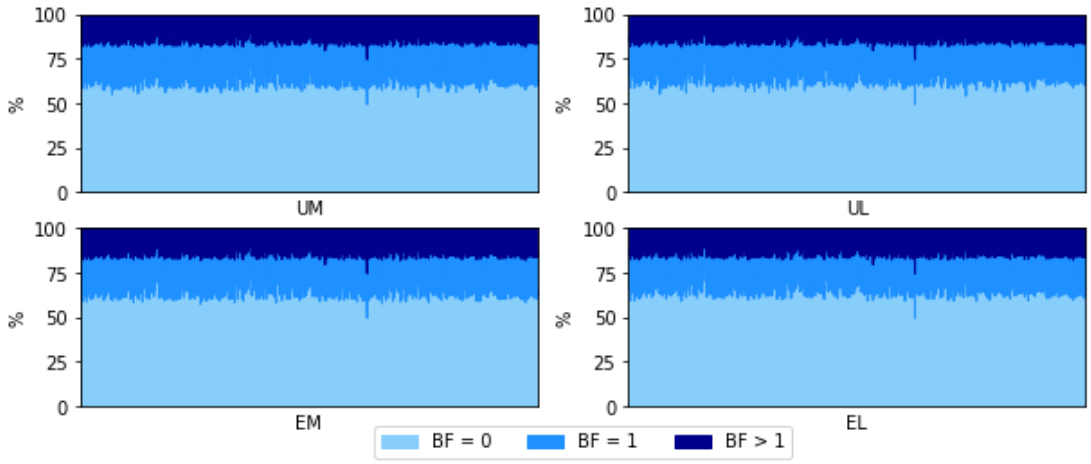


Figure 6.19 The percentages of nodes' branching factors; most nodes have zero branching factor

The rest of the analysis is carried out for the most-recent user model only, because the differences between models are small, due to the fact that most nodes have zero branching factors. The average branching factor across all the posts equals 1.02, which reflects the percentages shown in Figure 6.19. Most nodes have zero branching factors, but there are about 32.54% of nodes with branching factors of one or above. A similar average (1.02) was obtained when the branching factors are aggregated for all of the different reblog reoccurrences in the cascade.

6.3.2 Scale: The impact of the post's author

The scale is the branching factor of the author, i.e., the number of users who reblogged directly from the root. Obviously the whole cascade is initially generated by the root, but the overall cascade size cannot be attributed to the author alone, as it is the result of the cumulative contributions of everyone who reblogged it. However, the scale can be used to estimate the

influence of the author and the direct impact s/he has on the cascade. Author's contributions to the cascade are computed as a function of its branching factor in relation to the cascade size. On average, the post's author contributes about 8.94%, and the median is 0.5%. Figure 6.20 plots the CCDF (cumulative distribution function) of the percentages of author's contributions to the total cascade size. It is noted that there are some cases where the author contributed more than 20%, but for the majority of cascades, author's contributions are below 10%. These percentages must be taken in their appropriate context, as the aim is to estimate the direct impact of the author. The author's contribution to the cascade decreases for large cascades. However, it is noted that, for some cascades, up to size 40000, the author contributed up to 80% of the overall cascade size, i.e., a large number of reblogs followed from the author directly, which means that cascades with this characteristic are shallow, as most of the nodes reside one step away from the root, i.e., were reblogged directly from the author. Nevertheless, for the majority of cascades (75%), the author's contributions are lower than 2.89%. Thus, most cascades' networks are deep, branching out beyond one step away from the author.

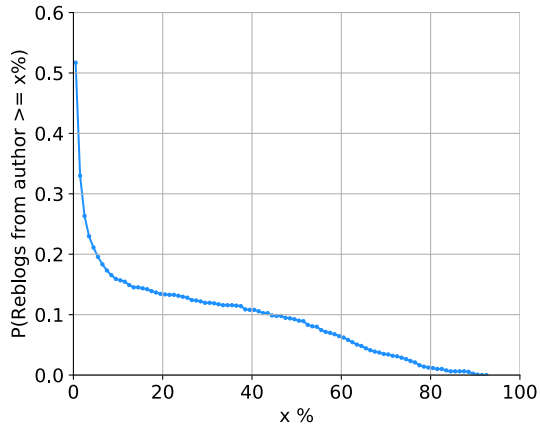


Figure 6.20 CCDF distribution of the percentages of author's contributions to the cascade (scale)

6.3.3 Sub-cascade sizes

The subcascade size extends the notion of influence measured by the branching factor to include the whole subcascade that follows from one user. Only nodes with branching factors equal to one or more are considered in the analysis in this part, as the nodes with zero branching factors will consequently have no subcascades. The size of the subcascade generated from each node, excluding the post author and all the users who have zero branching factors, is computed. The average subcascade size across all the posts equals 10.8. The maximum subcascade size is 183157, which means that for that particular post, one user's contribution to the cascade is about 98%. On average, users with high subcascade sizes contribute about 29% to the overall size of a cascade.

This percentage means that in some cases the size of the cascade might be highly dependent on one user only. Figure 6.21 shows the percentages of nodes with subcascade sizes that equal one or above, across all cascades. We can see that most nodes have subcascade sizes that are above one, meaning that they branch out beyond their direct children. On average, about 13.3% of the nodes have cascade sizes that equal one; hence, their branching factors are one, as well.

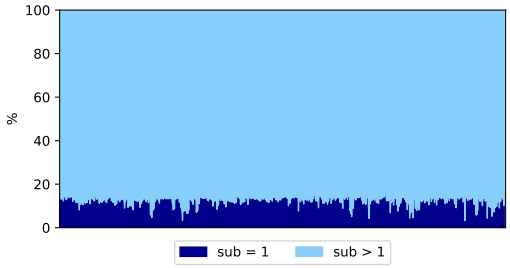


Figure 6.21 The percentages of nodes' subcascade size; the majority have more than one subcascade size

6.3.4 Number of paths, path lengths and depth

This part utilises another set of measures of the structural aspects of cascades, namely: the number of paths, paths lengths and depth. Each of these measures shed some light on a different aspect of the structure of a cascade. A node's depth in a cascade network is simply its distance from the root, i.e., how far it is from the root (the post's author). While the path lengths and the number of paths in a cascade network are measures of how far content travels away from the root until it reaches a user from whom no-one reblogs (leaf nodes, branching factor = 0). The total number of paths indicates the number of different paths a post travels along as it spreads. Because these measures are highly dependent on the distances from the root, only ideal cascades are considered, i.e., the ones that have the author as a source of the first reblog.

6.3.4.1 Adoption per depth

Figure 6.22 plots the aggregate number of nodes at each depth across all cascades, which roughly corresponds to the number of reblogs at each depth. The figure shows that the majority of nodes are at depth one. The number of nodes per depth decreases after that until depth 10, where the number of nodes starts to decrease sharply after that. This means that most of the time rebloggers are not far away from the author. Nonetheless, some cascade networks grow far away from the root, reaching 32.78 steps away from the root on average, while the maximum depth found across all cascades equals 145. In fact, the maximum depth is below 46 for about 75% of the cascades, while the rest have a maximum depth greater than 46. Overall, the depths recorded for Tumblr cascades are considered as non-trivial depth (Dow et al., 2013). Adamic, Lento and

Fiore (2012) reported a maximum depth of 40, while Liben-Nowell and Kleinberg (2008) reported a median node depth of 288. On the other hand, in their analysis of Twitter, Taxidou and Fischer (2014) noted an average diameter of only four. Thus, the depth of Tumblr's cascade' is relatively greater than that of the others.

The average proportion of reblogs per depth for all cascades is calculated (See Figure 6.23). It is clear that, on average, reblogs at depth of one comprise about 14% of the overall cascade size. The mean reblogs proportions decreases after that, remaining slightly above zero. The plot on the right in Figure 6.23 shows that the mean reblogs proportion drops below 2 from depth 5 onwards. Thus, even though posts were reblogged after depth 5, the majority of reblogging users are only few steps away from the post's author.

It is important to note that in this section, the aim is to examine the structure of cascades; thus, all of the measures are used without taking the time into consideration, meaning that nodes that appear at depth one are not necessarily caused by reblogs in the cascade's early stages.

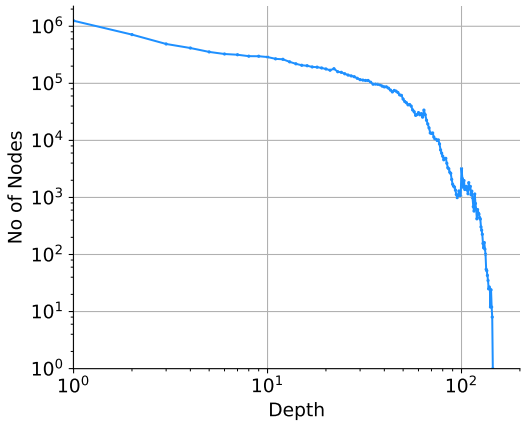


Figure 6.22 Aggregate number of reblogs at each depth, showing that the majority of nodes are at depth one

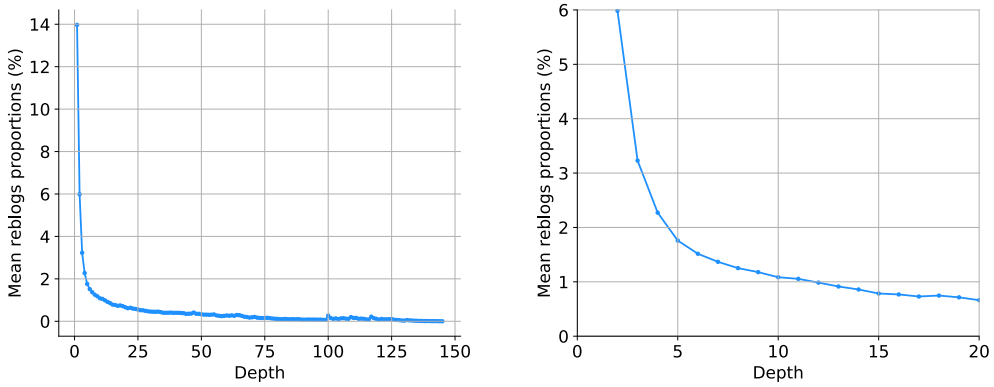


Figure 6.23 Right: The percentages of the mean reblogs proportions per depth; left: zoom-in of the same chart

6.3.4.2 Number of path and path lengths

Table 6.3 below compares the number of separate paths in the four models. The least-recent models have a slightly higher number of paths than the most-recent ones. The reason behind this is that in the least-recent models nodes are linked to the least-recent parent copy; thus, there will be more paths but they will be shorter than the ones in the most-recent models. Each path represents a separate information flow path. On average, a post has around 9196-11109 paths. The maximum number of paths was recorded for one post by Tumblr's official staff blog that was reblogged directly from the same blog most of the time; thus it has the largest number of short paths, in which almost every reblogger reblogged from the author. The chart below plots the distribution of the number of paths for the most-recent users and event models.

Table 6.3 Comparison of the number of paths in each model

	UM	UL	EM	EL
	No. of paths	**No. of paths**	**No. of paths**	**No. of paths**
Mean	9196	10877	9196	11109
Median	3768	5298	3768	5404
Minimum	2	2	2	2
Maximum	104061	124560	104061	125815

The total number of paths which the post spread through to reach different users and the length of these paths are computed. The leaf nodes (nodes that have BF = 0) are used to mark the termination of cascade paths. On average, cascades consist of 9167.01 different paths (median = 3768.5). Half of the cascades consist of 923-12531 different paths, with some outliers above and below these numbers. For instance, in one cascade there are 104061 different paths from the root (author). The Pearson correlation coefficient between the cascade size and the number of paths in the cascade network is 0.54 ($n = 796$, $p < 0.001$, $r = 0.54$), which means that it is more

likely that as the cascade grows it will have more paths, but that is not always the case, for there are some moderate sized cascades that have a large number of paths ($R^2 = 0.29$).

On average, path lengths are between 10 and 11, but there are some long information flow chains that reach 145 hops away from the post author. Table 6.3 shows the distribution of path lengths in the four construction models. We can see that most-recent models tend to produce longer paths than least-recent ones. In general, across all the four models, about 46-55% of paths are 10 in length or shorter while the rest can reach up to 104-145 in length.

6.3.4.3 Mean Branching Factor Per Depth

Looking at branching factor as a function of depth allows us to understand how cascades grow in relation to rebloggers' distances from the root (their depths). Figure 6.24 shows that the average branching factor at depth is 8.4, then it decreases to 3.6 at depth two. After that, it remains just above two and below three up to depth 79. For depths from 100 and deeper, the average branching factor fluctuates between 2 and 5; sometimes it is close to one, but overall it does not go below 1 at any given depth. A possible explanation for the fluctuating part is that there might be some users who reside away from the author. These users were not exposed to the post directly from the author or one of the author's closest users. Nonetheless, when these users reblogged that post, a surge in the post's popularity emerged, as many users reblogged after them. It is worth noting that the average branching factor on Tumblr is higher than the average branching factor on two of Facebook's large cascade memes (Dow et al., 2013), where the average branching factors per depth are between 0.5 and 1 for depths from 1 up to 20. Therefore, it appears that popular content on Tumblr exhibits higher numbers of branching factors at each depth in comparison to Facebook's popular content.

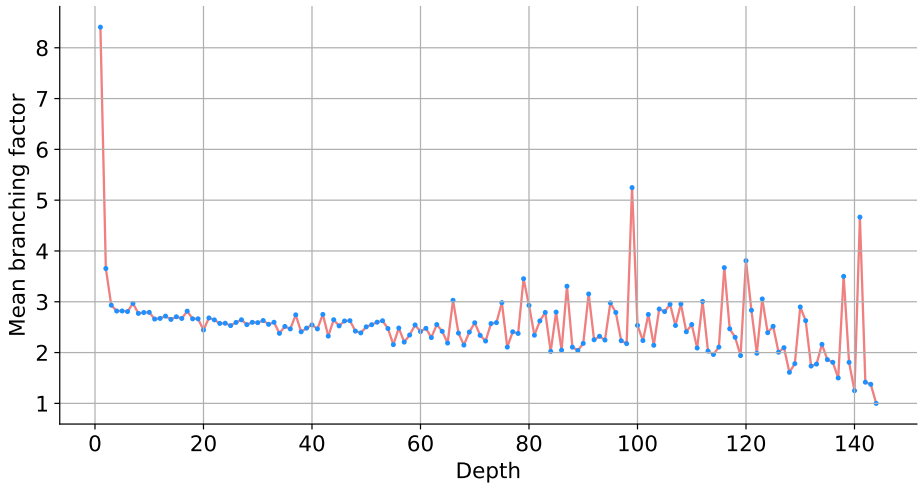


Figure 6.24 Mean branching factor per depth: at each depth the posts are being reblogged with some sparks in the mean branching factor around and after depth = 100

6.3.5 Discussion and Remarks

Estimating influence

A user with a high branching factor means that this user plays a major role in spreading the content, which might be related to the user's position in the social network (the number of engaged followers a user has). If a user with high number of followers is exposed to a particular post, the possibility that the post will continue to spread will increase. However, if a user has few followers, it is likely that it will have a small branching factor (or zero), hence, the cascade will eventually stop after reaching this particular user (a sink, in graph theory terminology). It is widely acceptable to use the number of reblogs (or retweets) that result from a user as a proxy to estimate influence (Cha et al., 2010; Kwak et al., 2010). This is reasonable because it estimates the user's direct impact, since most of the users who will reblog a post from another user have most likely done so after being exposed to the user's feed.

The analysis shows the difference between using the branching factor and the subcascade size, because in some cases branching factors can underestimate the overall impact of a user. However, branching factors are useful to assess the direct impact, while the subcascade sizes are useful to assess the overall impact, whether it is caused by the same user or possibly some other user in the network.

Small branching factor but high impact?

Considering the branching factor only as an estimator of the user's influence can sometimes lead to underestimation. In some cases, the subcascade generated from a user that has branching

factor of one is larger than the initial anticipation, especially for nodes having a branching factor that equals one. We notice that the average subcascade size for these nodes equals to 13, while the maximum subcascade size equals 182986. These numbers are computed by excluding the root, to avoid including cascades where the author's branching factor equals one.

To be able to assess the relation between the branching factors and the subcascade sizes, we divide the branching factor by the subcascade size to compute the ratio that will allow us to assess this relation. Small ratios (close to zero) means that the branching factor is small but the subcascade size is large. Which also means that this particular user has created a subcascade that goes beyond one step away from the user. As the ratio increases, it means that the user does not create a large subcascade, meaning that the subcascade is close to broadcasting, where it does not branch away from it. If the ratio equals one, then the branching factor equals the subcascade size, meaning that it does not go beyond one step away.

Figure 6.25 plots the CDF distribution for the ratios generated for nodes with one branching factor or more. It shows that for about half the nodes, their generated subcascades are larger than their branching factors, meaning that their impact is actually higher than their branching factors. Also, the figure shows that nodes with branching factors above one generate slightly larger subcascades than the ones generated from the nodes with branching factors that equal one.

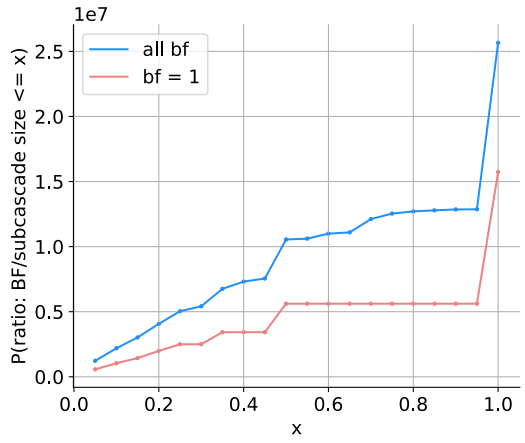


Figure 6.25 CDF distribution of the branching factor to the subcascade size ratios, showing that for more than half the nodes their subcascade sizes are larger than their branching factors, which means that these nodes (users) have a greater impact.

6.4 Temporal Features of Cascades

The previous section looked at the structural features of cascades; this section will provide an analysis of the temporal features of cascades. As explained in Chapter 5, to provide an accurate temporal analysis, only ideal cascades will be included. These have cascade networks with the author as the root and all the timestamps of the reblogging activities are valid. The first constraint yields 806 posts, which have the posts' author as the source for the first reblog. The second constraint excluded 10 more posts, leaving a total of 796 posts suitable for temporal analysis.

This section is split into two main parts: growth of cascades and cascade burstiness. Analysis of the cascades' growth looks into the cumulative increase in the number of reblogs in relation to the time after publishing the post, while cascades' burstiness is concerned with the rate at which a cascade grows.

6.4.1 Preliminaries: Cascades' active age

Before presenting the findings from the temporal analysis, there are s some points that need to be clarified. The posts included in the analysis were published sometime before Dec 2014, and Tumblr's Year in Review blog is published before New Year's Eve. Data collection took place for about a month, from 29 Dec 2014 to 2 Feb 2015. Thus, all of the reblogging activities available before 2 Feb 2015 were collected. Therefore, the post's active age was calculated, which is the difference in days between the post's publishing timestamp and the last day in which any reblogging activity was recorded.

On average, the recorded posts' active age is 237 days, which is slightly below a year. The oldest post in the dataset was active for 617 days (more than a year); it was posted on May 2013, which means that there are some posts published before 2014, which remained active during 2014. Surprisingly, this old post has a very small cascade size of 131 reblogs; nonetheless, it managed to survive for 617 days. The youngest post was posted on 9 Mar 2014 and was active for 28 days only, and it was tiny, with 68 reblogs over the period of 28 days, where about 61% of its reblogs occurred on the first day after publishing.

However, it is important to note that a cascades' age is defined as active-age, as it is not definitive, because many of these posts will remain active after the data collection. However, the aim is not to provide estimation or prediction of the cascades' life-time, the focus is to understand cascades' growth and burstiness patterns.

Figure 6.26 shows the distribution of the posts' active age: most of the posts last for up to 400 days and only few are older than 400. Looking at the correlation between the size of a cascade

and its active age, it is clear that as the cascade becomes larger its likely to have a shorter active age than smaller cascades. However, the correlation coefficient is very small (n = 796, p = 0.0017, r = -0.11082), the squared Pearson correlation coefficient $R^2 = 0.01228$. This is due to the huge variations in the combinations of the cascade sizes and their active age.

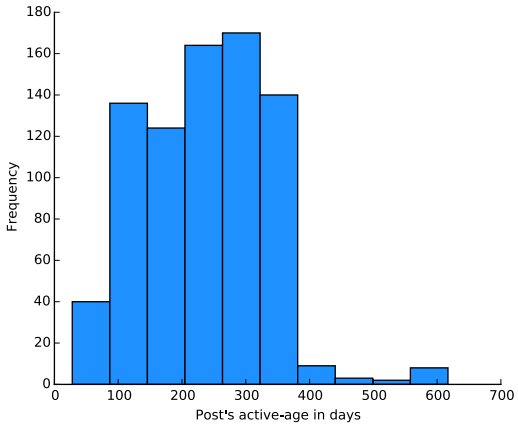


Figure 6.26 The distribution of posts' active age

6.4.2 Cascade Growth

As they attract more rebloggers, posts' cascades size increases; this section investigates cascade growth, looking at the cumulative cascade size in relation to the time after publishing from several perspectives, such as: how long it takes for a post to be reblogged and the rate at which a cascade accumulates its reblogs.

This part of the analysis will look into the patterns of the cumulative cascades' size advancements during their lifetime. The question this section attempts to answer is: *Is there a pattern of cascade growth on Tumblr that can be detected?* I.e., *does popular content follow the same growth pattern as time goes by?* To answer this question, consider the timelines of three of the largest cascades (416282, 416282, and 371600), as shown in Figure 6.27. From this figure, three different patterns can be recognised. The first one (in blue), starts off with few reblogs (27 on its first day) but picks up the pace on its 17th day with 101351 reblogs on a single day, then it maintains its popularity with a moderate number of reblogs until its last recorded reblogging activity. The second one (in dark red), in contrast, maintains a moderate number of reblogs each day, with no high spikes in its lifetime. The third one (in light red), starts off with a high volume of reblogging on its first day, then it declines, having a moderate number reblogs each day. Again, the focus here is not on how long it lasts, but rather how cascades accumulate their popularity.

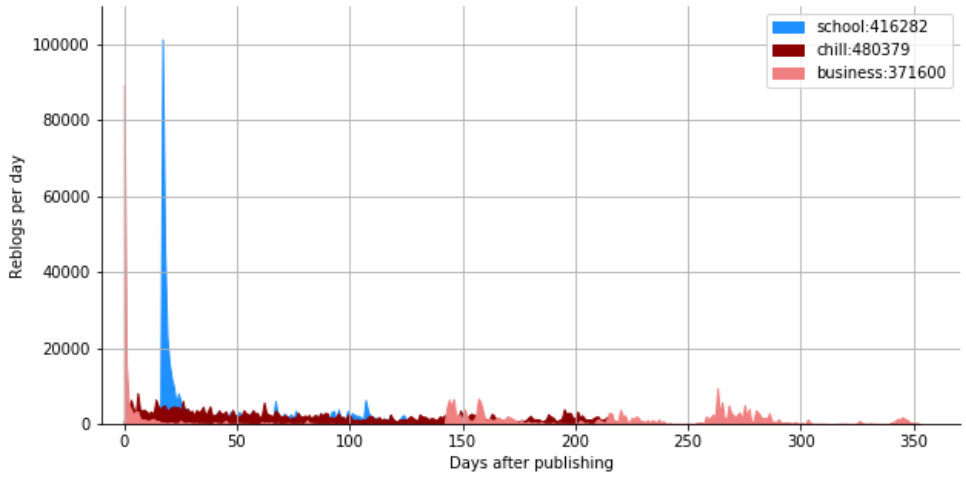


Figure 6.27. The timelines of three large cascades shows different patterns of accumulated cascade size

Considering the number of reblogs per day after publishing the post, most reblogs occur earlier in the post's lifetime. The number of reblogs per day decreases as the post gets older and after the posts' mean age, 237, the number of reblogs per day decreases and cascade growth becomes almost stable, i.e. there are no significant increases in the cascade sizes. The average number of reblogs per day is 281 reblogs (median=149).

6.4.2.1 How long does it take for a post to be reblogged?

Before delving into analysis of cascade growth, this section looks into the time lag between posting and reblogging. Around 87% of first reblogs occur in the first hour after publishing, and 97.11% of first reblogs occur within 24 hours after publishing. These figures are similar to the ones reported by Chang, Tang, Inagaki, *et al.* (2014), who also analysed cascades on Tumblr. Their results show that 75.03% of reblogs occurred in the first hour and 95.84% occurred within one day. Kwak et al.'s famous study on Twitter (Kwak et al., 2010) shows that 50% of retweets occur within an hour, while 75% occur within a day. These percentages confirm the tendency towards `recency' in Tumblr, as Chang et al. describe it. However, there is a minority of posts (2.6%) that did not get their first reblogs in the first 24 hours and 1.13% that were actually idle for more than 100 days before the first reblog occurred. Nonetheless, being late in getting their first reblog did not affect their active-age or cascade size. In fact, some of these posts managed to survive for more than 600 days. There is one post that got its first reblog on its 105th day but had 43660 reblogs cumulatively.

6.4.2.2 Growth Patterns

Motivated by the observations from the large cascades' timelines, this section examines whether cascades of similar sizes follow the same growth pattern or not. Analysing cascade growth involves categorising the patterns of cascade growth, i.e., whether it is 'bursty', alternating between high to low with idle periods of reblogging activities, or whether it grows at a steady-pace throughout its lifetime, or a third case where it gets most of the attention at its early stages then users' attention declines afterwards. The subsequent sections will discuss these topics in detail.

To be able to assess growth systematically, cascades must be categorised according to the distribution of their cascade sizes. Precisely, ideal cascades are categorised into four different categories. The first category has cascades that reside below the 25^{th} percentile, these are cascades of relatively small sizes (below 7697). The second category for cascade between roughly the 25^{th} and 50^{th} percentiles: these are moderate cascades below the median (31493). The third category is for the slightly larger cascade between the 50^{th} and the 75^{th} percentiles (above the median 31493 and below 72475) and the last category is for large cascades beyond the 75^{th} percentile. Figure 6.28 illustrates the normalised cumulative growth in cascade sizes for all cascades in the four cascade categories. Each line in the figure corresponds to one cascade, and the normalised cumulative growth is computed using the total number of reblogs a cascade has on a given day. As the figure shows, there are no particular patterns that cascades in different categories follow, which simply means that cascades reach their maximum recorded sizes in various ways. Across all categories, the increase in cascade sizes does not occur at a constant rate. In contrast, the diffusion of Linked-in invitations (Anderson et al., 2015) has a linear growth pattern for both large and medium cascades (over 4000 and 500 respectively). The lack of a uniform pattern might be related to the cascade size itself, or is probably platform specific i.e., the different mechanisms by which the users become exposed to a particular content.

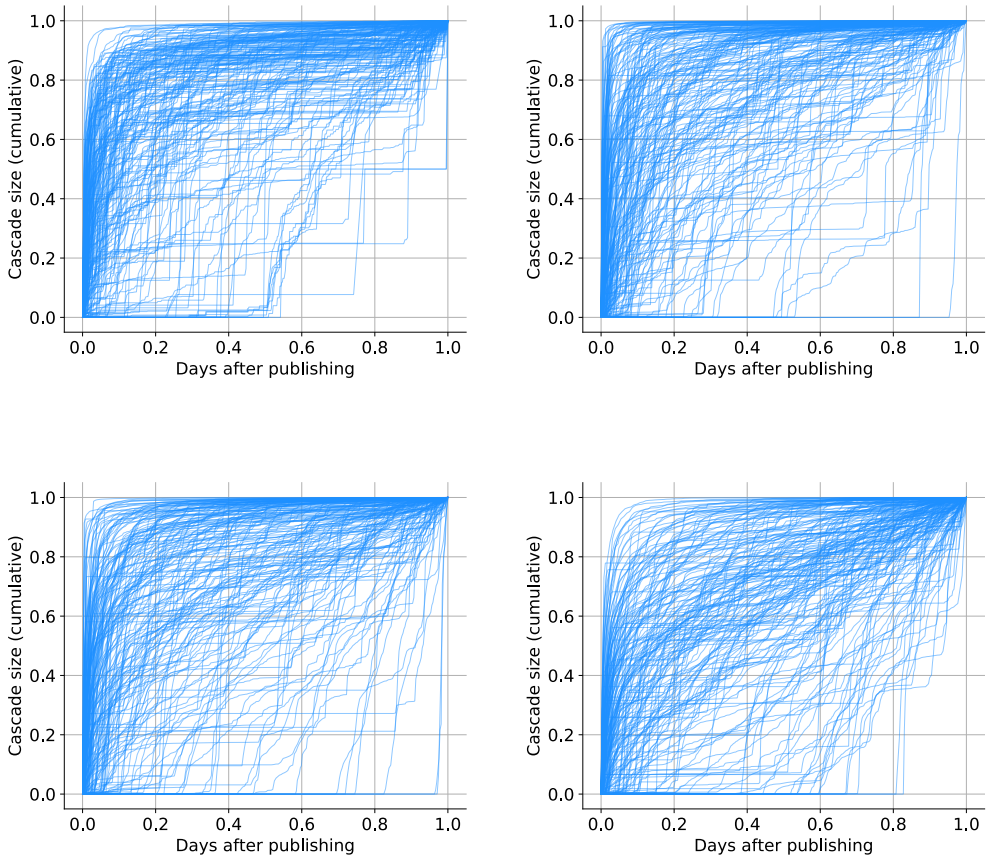


Figure 6.28 Normalised cascade cumulative growth against days after publishing for four cascade size categories, showing that there is no uniform pattern for the cascades' growth

6.4.2.3 Branching Factors Across Time

As shown in the structural analysis, branching factors per depth help in estimating how cascades grow, as a function of their average branching factors at a given time after publishing. The last temporal measure that will be used is actually a measure that combines the structure of the cascade in relation to the time. The average branching factors each day or hour after publishing will expand our understanding of the way popular content gains its popularity. Figure 6.29 and Figure 6.30 show the average branching factor per day and per hour. We can see that, during the first 100 days, the average branching factor starts below 2.8 and drops, after which it fluctuates between 2.1 and 2.4 for the rest of the period. If we consider the first 3 days (i.e., 72 hours), we can see that in this period, the branching factor starts slightly above 4, then it loses almost half of its value in the next hour as it drops to slightly above 2.5. After that it falls again, remaining above 1.5. In comparison to the average branching factors per hour in Dow et al. (2013), the average branching factors for popular content on Tumblr is higher than their counterpart son Facebook (where branching factors per hour are between 0.5 and 1). Thus, it appears that Tumblr's content attracts higher numbers of users at each point in its lifetime.

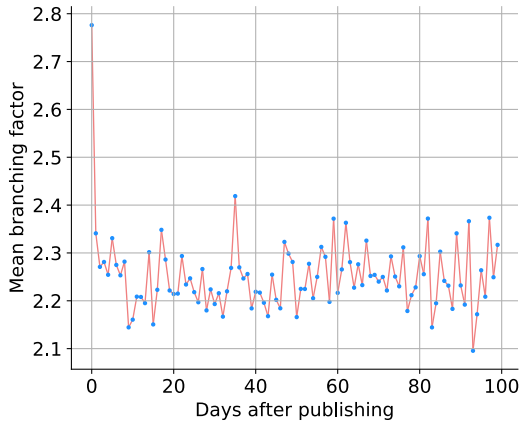


Figure 6.29 Mean branching factor per day for the first 100 days

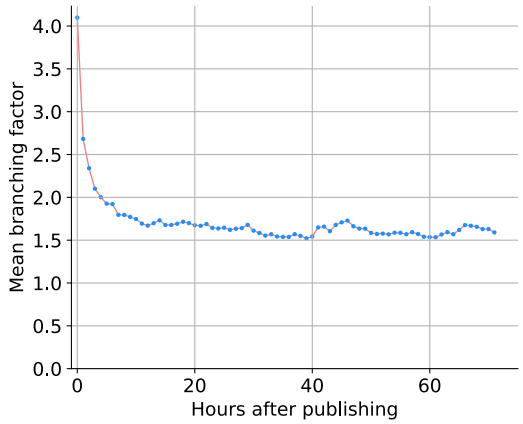


Figure 6.30 Mean branching factor per hour for 72 hours

6.4.3 Burstiness of Cascades

One of the most important temporal aspects that have been studied in the literature is a cascade's burstiness. Burstiness is often measured by analysing time series data in order to identify bursty periods (Kleinberg, 2002). These periods show a high volume of interactions and often feature a peak where the number of interactions reaches its highest value within that burst (Palshikar, 2009). Identifying peaks is an essential step towards understanding the temporal dynamics of reblogging. In fact, there are two facets of burstiness analysis that will investigated in this part of the analysis; the first one focuses on analysing peaks in the general sense, i.e., identifying peaks and estimating bursts and periods of idleness. The second facet, in contrast, digs deeply into a temporal phenomenon that has not been studied widely in cascade related studies which is cascade recurrence (Cheng et al., 2016). The following sections will investigate these two facets of burstiness in details.

6.4.3.1 Peaks Detection

The classic definition of peaks states that peaks are points in time preceded by an increase in the volume of interactions and followed by a decrease (Palshikar, 2009; Schneider, 2011). Precisely, according to Palshikar, a peak is:

1- A local maximum within a window, which means it is not necessarily a global maximum or large across the whole time series.

2- Isolated, so it stands out within a window, as not many points have similar features.

These remarks emphasise two aspects that are equally important in any peak detection tasks. The first point sets constraints on the peak itself, while the other sets constraints on the area surrounding a peak. Following from these remarks and the definitions provided by Schneider, (2011) and Cheng et al. (2016), a peak in Tumblr's context is defined as follows:

Let $f(d)$ be a function that gives the number of reblogs on a day d, $f : D \rightarrow R$ where $D = \{0,1,2,3,..n\}$ is the set of days in the time series of the cascade and R is the set that holds the number of reblogs that corresponds to D's values.

A peak p_i, that corresponds to a day $d_i \in D$, where $i \leq n$, is identified if it meets the following conditions:

1- d_i is a local maximum, i.e., the number of reblogs on d_i is higher than for its immediate neighbours.

$$f(d_{i-1}) \leq f(d_i) \geq f(d_{i+1})$$

2- The number of reblogs on d_i (the height of the peak) must be at least h and at least $m \times \bar{x}$, where $\bar{x}$ is the mean number of reblogs per day in the cascade.

3- The peak is a local maximum within a window $\pm w$, where $\pm w$ is an interval $I = (a, b), I \subset D$ and assuming that $I \cap D \neq \emptyset$:

$$f(d_i) \geq f(d), \forall\, d \in I$$

As the formal definition shows, the peak detection process varies and can be made stricter by tightening the variables related to the height of the peak namely (h and m), i.e. the number of reblogs on the peak's day and the size of the interval $\pm w$ (the observation window). These constraints can be made stricter or looser; consequently, they will affect the total number of peaks that can be detected.

6.4.3.2 Burstiness Patterns

During their lifetime, cascades often consist of days with high reblogging activities and others with very low to zero reblogging volume. Cascades' burstiness patterns are identifiable using peak detection methods. For this part of the analysis the three conditions mentioned in the previous section are used to identify peaks. So peaks are local maxima within $\pm w$, and their heights are at least h and at least m multiplied the mean number of reblogs. Practically, the values used for each of the above parameters are: $h = 10$, $m = 2$, and $w = 7$, which is similar to the values suggested by Cheng et al. in their analysis of cascades on Facebook (Cheng et al., 2016). After experimenting with these parameters, these values were considered suitable, because for about 13% of the cascades $m \times \bar{x}$ is less than 10, thus, setting h to 10 will exclude all the peaks that are less than 10. This is particularly effective for smaller cascades that have many days with almost zero reblogging, which results in a smaller mean number of reblogs per day. Additionally, multiplying the mean by two will exclude all peaks that are much less than twice the mean reblogs per day. The window size used, which is equal to 7, will ensure that these detected peaks are local maxima within 15 days (including the peak's day itself), hence, within a month there is a chance if having two peaks at most. In practice, this is important, due to the long observation period, which is about a year.

After applying the conditions above, a number of peaks were detected in each cascade, the cascades' burstiness patterns will be discussed guided by the following questions:

How many peaks are there in each cascade?

Figure 6.31 plots the distribution of the number of peaks in all cascades. As the figure shows, on average, there are about 4.47 peaks (median = 4) in each cascade and the maximum number of peaks in one cascade is 19. The boxplot in Figure 6.32 shows that the majority of cascades have less than 7 peaks. Moreover, for around 18.84% of cascades only one peak was detected during the cascade's lifetime, and for less than 1% of the cascades no peaks were detected at all.

Figure 6.32 shows the proportion of peak days as percentages of the total number of days in the cascade's active age. As the figure shows, the proportion of peak days is significantly low. On average, peak days comprise 1.9% of the total number of days in the cascade's active age. The maximum proportion recorded is 5.6%.

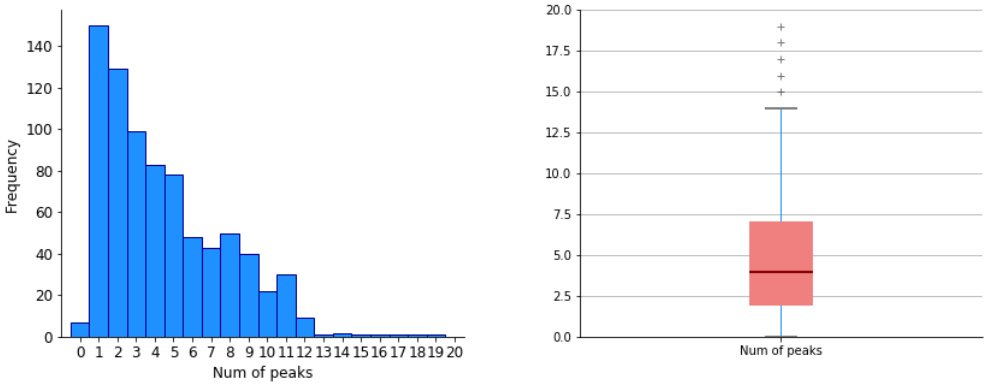


Figure 6.31 The distribution of the number of detected peaks and boxplot: there are about 4.47 peaks on average

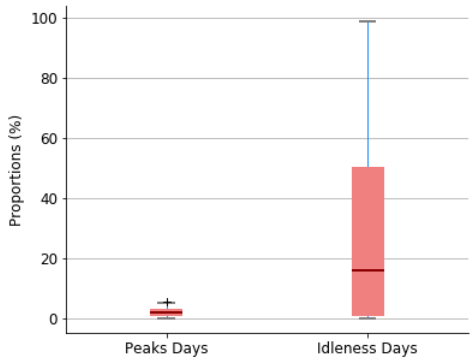


Figure 6.32. Proportions of peaks and idleness days: peaks comprise a minority of days during the cascade's lifetime, while the idleness periods comprise higher proportions.

What is the relation between cascade size and the number of detectable peaks?

The boxplot plot in Figure 6.33 shows that the number of peaks in a cascade does not follow a particular pattern: there are some large cascades with a small number of peaks and on the other hand there are some small cascades with a larger number of peaks. This is also noted in the distribution of the number of peaks in the four cascade categories in Figure 6.34: apart from the small cascades (below 25th percentile), the majority of cascades feature up to 7 peaks. This number is slightly smaller for small cascades, where the majority of cascades have less than 5 peaks.

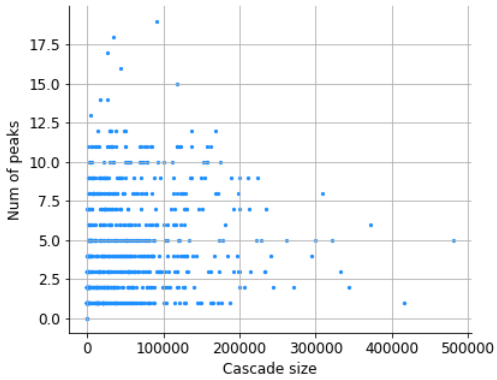


Figure 6.33. A scatter plot of the number of peaks and the number of reblogs

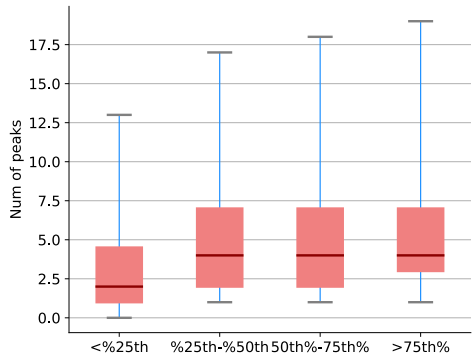


Figure 6.34. Number of peaks in four cascade categories grouped according to their size

For how many days does the cascade remain idle?

Each cascade goes through periods of high reblogging activities and other periods with very low reblogging activities. Hence, the days that have no reblogging activities are identified as idleness days. As Figure 6.34 shows, on average, the proportion of idle days is about 28% of the cascade's active age (median = 15.75%). However, for about 75% of the cascades. the proportion of idleness days is 50% or less; only a few cascades have proportions of idleness days that are higher than 50%. The maximum idleness proportion is 98.86%, for one cascade. These proportions are computed as aggregate proportions across all cascades; thus, to provide an in-depth look into the idleness periods for each cascade, the number and length of each idleness period is computed.

On average, each cascade has about 18.19 idleness periods, which last from one day to 16 days on average. Figure 6.35 illustrates the number of idleness periods in four different cascade categories, where cascades are grouped according to their sizes. Smaller cascades have relatively higher numbers of idleness periods, the number of idleness phases decreases as the cascade size increases. Large cascades (above the 75th%) have fewer idleness periods in comparison to the

others. The average number of idleness periods in large cascades is 2.6 (median = 1), while it equals 28 (median = 22) for smaller cascades.

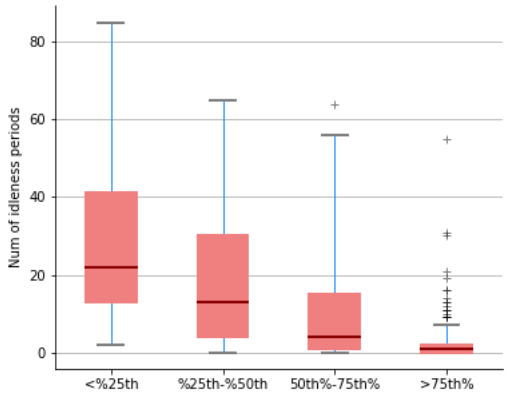


Figure 6.35 The number of idleness periods in the four cascades categories

The length of the idleness periods varies across the different cascade categories. On average, across all categories, the length of idleness periods is 16 days. However, the mean lengths of idleness periods are 9.13, 9.92, 19.69 and 33.39 days in the four cascade categories, respectively. This means that, as the cascade grow in size, it will have fewer idleness periods but its idleness periods will be longer.

Moreover, for about 148 cascades (18.6% of the ideal cascades), there are no idleness periods at all, which means that these posts were reblogged every single day during their active age, which equalled 169.45 days, on average. These cascades have moderate to large cascades; their sizes range from 7752 to 480379, while the mean size for these cascades is 103892.

When do peaks appear in the cascade's lifetime?

The distribution of peak days shows that 58.45% of the peaks occur within the first 100 days in the cascade's lifetime (See Figure 6.36) and the majority of peaks occur within the first 200 days. In fact, 11.20% of the detected peaks are on the first day after publishing the post. The rest of the peaks are distributed across the cascade's lifetime; the furthest detected peak is on the 557th day.

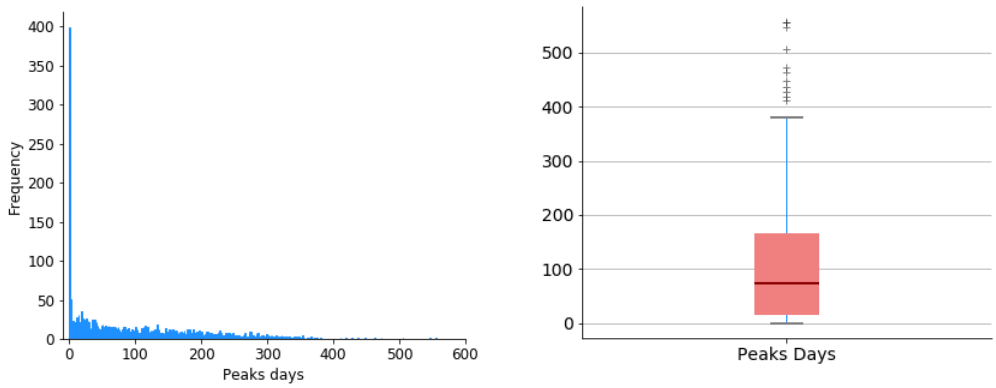


Figure 6.36 The distribution of peaks days: 58.45% of the peaks occur within the first 100 days

How many days separate any two consecutive peaks?

The number of days between peaks can be used an estimate of cascades' burstiness. The difference in number of days indicates how often a peak is detected i.e., whether the detected peaks are close to each other or not. The difference between two consecutive peaks is 36 days on average (median = 21 days). The maximum difference between two peaks is 368 days. The heavy-tailed distribution in Figure 6.37 shows that most of the time peaks are less than 50 days apart from each other, with some cases where the difference exceeds 100 days. To put these differences in perspective, the average active age for a cascade is 237 days; thus, the average difference (36 days) means that peaks are relatively far apart from each other.

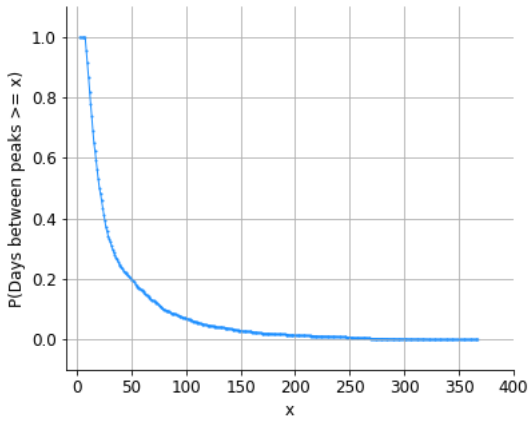


Figure 6.37 The distribution of the days between peaks

How many days does it take a cascade to reach its maximum number of reblogs in a day?

Each cascade reaches its highest number of reblogs in one day during its lifetime. This day is identified as the global maximum or the highest peak. It is defined as $d_0 \in D$ if:

$$f(d_0) \geq f(d), \forall\, d \in D$$

About 42.46 % of cascades reach their global maximum on the same day after publishing a post. This fraction drops after that to 11.05%, where cascades reach their global maximum one day after publishing a post. Collectively, about 70% of cascades reach their global maximum within a week after publishing a post. To estimate the impact of these global maxima, the proportions of the total number of reblogs are computed in relation to the total number of reblogs. Figure 6.38 plots the computed proportions in the four cascade categories. For smaller cascades the number of reblogs on global maximum days comprise about 60% or less of the total number of reblogs in one cascade. The proportion of the number of reblogs is higher for the majority of small cascades. Moderate to large cascades' reblog proportions are about 40% to 30%. Across all cascade categories, there is a minority where the number of reblogs on global maxima days comprises the majority of reblogs in the whole cascade. Figure 6.38 (left) shows the proportions of the number of reblogs taking all peaks into account. As the figure shows, the total proportion increases by similar amounts across all cascade categories. The proportions of the smaller cascades' number of reblogs remain the highest, with the majority having around 70% of total reblogs or less. The proportions for the other three categories range from about 40% to 50%.

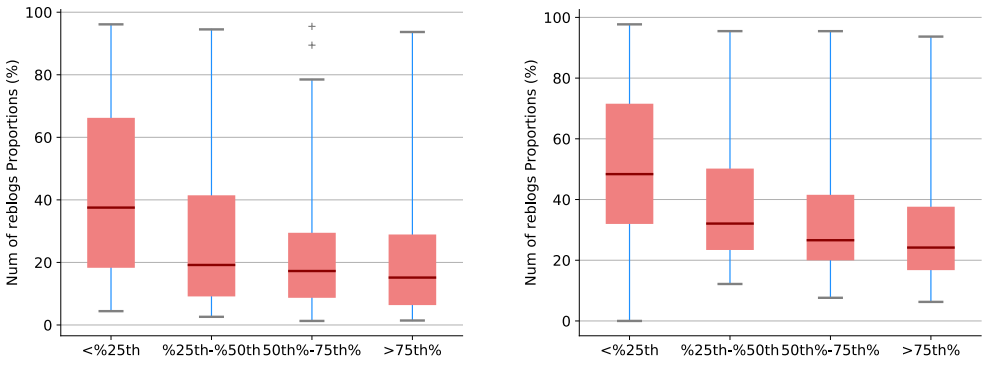


Figure 6.38 Left: proportions for global maxima only; right: proportions for all peaks

6.4.4 Recurrence

6.4.4.1 Defining Recurrence

The previous section demonstrated burstiness patterns, which included identification of peaks as an essential step. In fact, after being detected, peaks can serve as an indicator of another temporal feature of cascades, namely, recurrence. In its simplest form recurrence is said to occur if a cascade has more than one peak in its lifetime, meaning that, when recurrence happens, it is as if a cascade regains popularity after a period of low reblogging activity, where popularity is measured as the total number of reblogs.

The idea of recurrence stemmed from epidemiology, where it is simulated for infectious diseases and used to measure their periodicity. The first to apply recurrence identification to cascades were Cheng et al., who studied the recurrence of cascades on Facebook (Cheng et al., 2016). However, merely detecting peaks is not sufficient to identify recurrence. Instead, Cheng et al. imposed an additional condition on the valley between two consecutive peaks. The overall rule is that, between any two consecutive peaks, the number of reblogs must drop below a specific number. In practice, Cheng et al. state that between any two consecutive peaks p_i and p_{i+1}, that correspond to the days d_{pi} and d_{pi+1} the number of reblogs must be less than $v \times min\{f(d_{pi}), f(d_{pi+1})\}$. Formally, a valid valley satisfies the following condition:

$$f(j) \leq v \times min\{f(d_{pi}), f(d_{pi+1})\}, \forall\, j \in J \text{ and } d_{pi} \leq j \geq d_{pi+1}$$

Where $J \subset D$ and J is the set of days between the two peaks, assuming that $J \cap D \neq \emptyset$.

They set v's value to 0.5, which means that the number of reblogs between two consecutive peaks must drop below half the minimum number of reblogs of the two peaks. In supplementary materials, Cheng (2016) used a slightly different definition of the valley, precisely:

$$f(d_{pi}), f(d_{pi+1}) \geq v \times max\{f(j)\}, \forall\, j \in J \text{ and } d_{pi} \leq j \geq d_{pi+1}$$

Using the same value of v, the above condition means that number of reblogs in the two consecutive peaks must be higher than the maximum number of reblogs in the valley between the peaks.

These two conditions are applied to the peaks detected using the conditions in Section 6.4.3.1. It was not clear in Cheng et al. whether, when the valley's condition is not satisfied, it would exclude both peaks on either side or just one. For the sake of accuracy, in the following sections, each time the valley's condition is not met, both peaks, on either side of the valley, will be excluded.

In addition to the valley's condition, the size of the burst around a valid peak is also computed. It is simply the period around the peak where the number of reblogs are either increasing or decreasing while remaining above the average number of reblogs per day.

6.4.4.2 Identifying Recurrence

The two valley conditions explained above have a different effect in identifying recurrence. The second one (comparing with maximum) identified significantly higher recurrence than the first one (comparing with the minimum). The distribution in Figure 6.39 shows the difference in the detected recurrence using the two conditions. Using the first condition about 8.16% of cascades

recur, while it is 62.06% using the second condition. In comparison Cheng et al. report that the probability of recurrence is 0.4 for image memes and 0.30 for video memes (40% and 30% respectively) (Cheng et al., 2016).

Cheng et al. used v = 0.5 (Cheng et al., 2016) and do not explicitly specify its value in the supplementary materials (Cheng, 2016). Thus, the researcher experimented with the value of v, when it is changed for the second option to 2 instead of 0.5, meaning that the number of reblogs in the two peaks must be larger than twice the maximum number of reblogs per day in the valley. Changing v's value decreases the probability of recurrence using the second condition to 0.085, which means that 8.54% of cascades recur on Tumblr. The resulting distribution after changing v's value is similar to that in Figure 6.40.

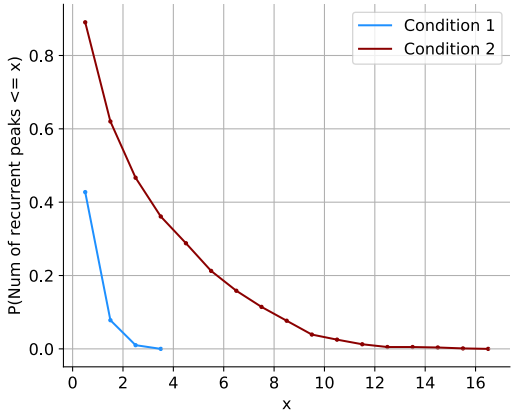


Figure 6.39 The distribution of the number of recurrent peaks using condition 2 and v = 0.5, which yields a higher number of recurrences.

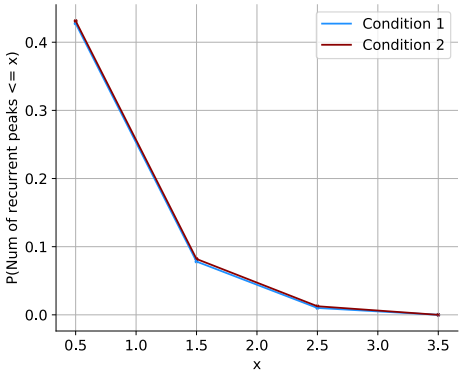


Figure 6.40 The distribution of the number of recurrent peaks using condition 2 and v = 2

On average, the number of recurrent peaks (i.e. the number of bursts) using the first condition is 2.21, and 2.23 using the second condition. In Cheng et al., the average number of peaks is 2.3 for image memes and 1.6 for video memes. Figure 6.41 shows the distribution of the

number of peaks for the two conditions, showing that the majority of cascades have no bursts at all, which means that either there are no detectable peaks or the valley's condition is not met. There are some cases where one peak was accepted and these are cases where there is only a single detectable peak in the whole timeline; thus, there are no other peaks to test the valley between the two. The rest are when there are two or more peaks, i.e., recurrent cases, where the valley's condition is satisfied for these cases. Recurrence varies in different categories, posts belonging to *Bathtub* are more likely to recur (probability = 0.33), followed by *Sponsored_posts_entertainment* (probability = 0.26). Surprisingly, none of the posts in the *Chill* and *Tumblrgetsdeep* categories recurred; especially given that these categories contain some of the largest cascades in the dataset.

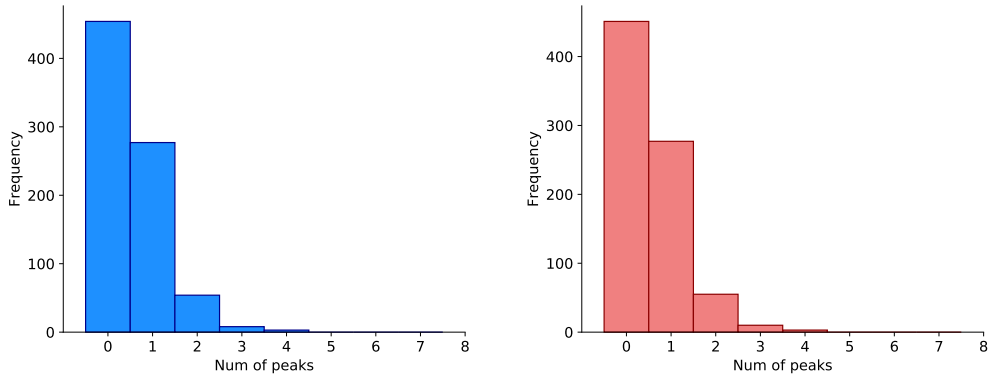


Figure 6.41 Left: peak distribution using the first condition with v=0.5. Right: peak distribution using the second condition with v=2: 8.16% and 8.54% of cascades recur using conditions 1 and 2 respectively.

6.4.4.3 Analysing Recurrence

For recurrent cascades, the difference between the first two bursts is 75.43 days on average, using the first condition, and 75.51 days on average using the second condition (medians=37.5 and 34 respectively). On Facebook, the number of days between the first and the second burst is 32 on average for photo memes and 44 for videos. Figure 6.42 plots the distribution of the differences for both conditions. As the figure shows, there is a small difference between the two conditions, and about 40% of recurring cascades regain their popularity within 50 days after the first peak.

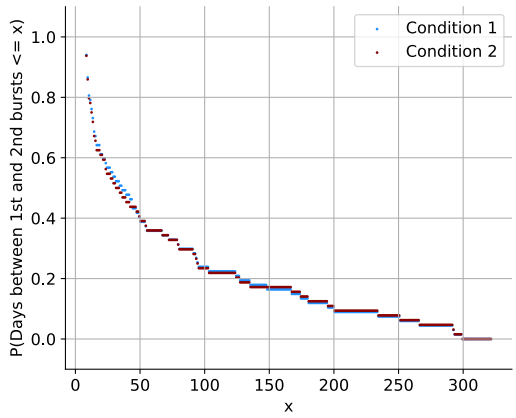


Figure 6.42 The difference between the first and second bursts in days for recurrent cascades: about 40% of cascades recur within 50 days.

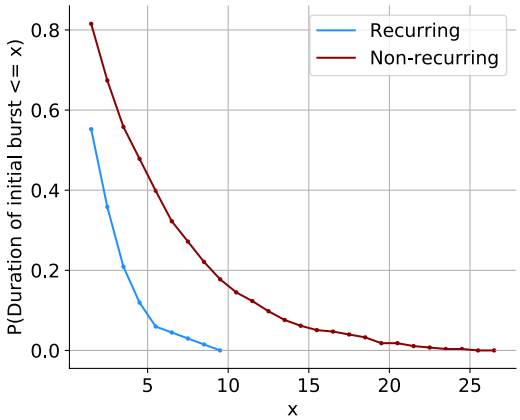


Figure 6.43 The duration of initial bursts for recurring cascades is shorter than for non-recurring cascades.

The duration of the first burst serves as an indicator of popularity and recurrence. The longer the duration of the initial burst, the longer it will remain alive, and hence will attract more users to reblog it. Figure 6.43 plots the distribution of the duration of the initial burst for both recurring and non-recurring cascades. In contrast to recurring cascades on Facebook (Cheng, 2016), the initial burst of recurring cascades on Tumblr is shorter than the initial burst of non-recurring cascades. For recurring cascades, the initial burst ranges from 2 to 9 days (mean = 2.5), while for non-recurring cascades the initial burst ranges from 2 to 27 days (mean = 5.7). On Facebook these numbers are 9.3 for recurring cascades and 6.9 for non-recurring cascades. The size of the initial burst in recurring and non-recurring cascades is as follows: for recurring cascades the size of the initial burst is 3076.57 reblogs on average, and for the non-recurring cascades there are 19609.85 reblogs, on average. This means that recurring cascades have shorter and smaller initial bursts that those on Facebook.

However, this brought up a different question: Does the number of reblogs in the first burst affect recurrence? To be able to answer this question properly the number of reblogs in the first burst must be grouped to provide a collective overview about the relation between the size of the initial burst and recurrence. Figure 6.44 plots the recurrence, measured by the number of bursts against four different first burst sizes (<=10^2, <=10^3, <=10^4, >10^5). The figures show that the average number of bursts is slightly above 2. When the size of the first burst increases, it slightly increases the cascade's chances to have up to four bursts. When the number of reblogs is above 10^5, the number of bursts decreases to two, which is the lowest possible number of bursts for recurring cascades. However, Cheng et. al. reported that when the number of reshares is between 10^4 and 10^5 the number of bursts jumps from 2.5 to slightly above 3.5 then it goes back to below 3 when the number of reshares is 10^6.

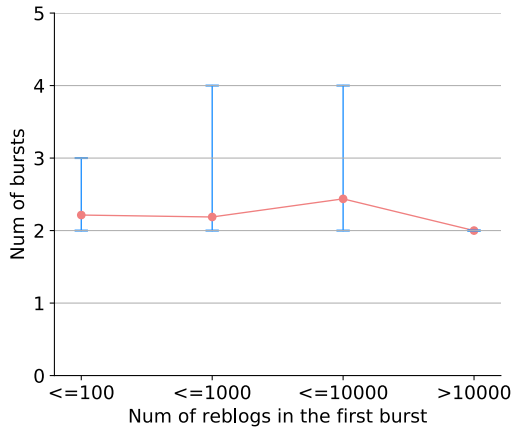


Figure 6.44 The relation between the size of the first burst and recurrence,

6.4.5 Discussion and Remarks

The number of detectable peaks

The total number of detected peaks is an indicator of how bursty the cascade is; it is particularly important because peaks are not merely days of high reblogging activities, but instead they are identified according to a number of conditions. Looking at the difference between peak days and idleness days, it is clear that the proportion of idleness days is significantly higher than peak days. In fact, the number of detectable peaks is 4.47 on average and the proportion of peak days is only 1.9% on average. As mentioned earlier, the number of detectable peaks is bound to the conditions enforced to identify them. However, small numbers of detected peaks are preferable to large numbers.

Figure 6.45 plots the distribution of the number of detected peaks in Tumblr's cascades, using three algorithms. The figure compares the algorithm used in the analysis (in blue) with two other algorithm S1 (Palshikar, 2009) (in dark red) and Findpeaks (in light red). Palshikar's algorithm relies on comparing the peak's altitude with the average and the standard deviation of the number of reblogs per day. It is also parameterised using the predefined window size $w = 7$, so for two adjacent peaks it will exclude the smaller one if they are within the same window w. Findpeaks, on the other hand, is parameterised using, minimum peak height = 10, minimum peak distance = 7, and the peak's height threshold = twice the mean number of reblogs. The total number of peaks detected for the analysis using Cheng et al.'s conditions (in blue) is roughly in the middle; it is smaller than the ones detected by Palshikar's algorithm and larger than the ones detected by Findpeaks' algorithm. In practice, an ideal algorithm to detect peaks must detect a sensible number of peaks: it must not detect too many of them or too little. However, further experimenting with algorithms and parameters is needed for a better, probably tighter, peak definition and possibly a sensible estimate of their acceptable proportions within their timeline.

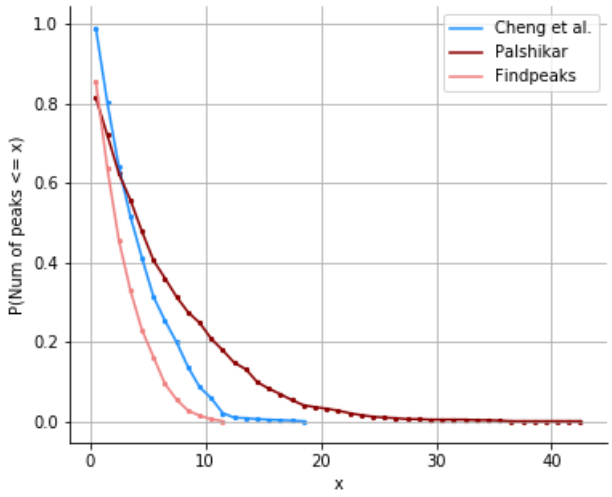


Figure 6.45 The distribution of the number of detectable peaks using three algorithms

Idleness periods

While most cascades (81.40%) suffer from idleness phases, the proportion of idleness days is less than 50% for the majority of cascades; fewer cascades have proportions of idleness days that are higher than 50%. Smaller cascades have higher numbers of idleness periods in comparison to moderate to large cascades. In contrast, the idleness periods in large cascades are longer than those in small cascades. On the other hand, about 18.59% of cascades have no idleness days at all, meaning that they were reblogged at least once every day during their life-time. The sizes cascades range from moderate to large cascades. This means that as the cascades grow in size

they will have fewer idleness periods, but their idleness periods are longer than those in small cascades.

One explanation that might decipher this pattern is that the cascade sizes and the number idleness periods are two faces of the same coin. It might be because these cascades remained active by being reblogged almost every day; that they accumulated a larger number of reblogs than smaller cascades. Further exploration of the growth and burstiness patterns of these cascades is required, especially regarding the relation between cascade sizes and idleness/activeness periods, which has not been widely explored in the literature.

Peak days

The fact that about 11.20% peaks occur on the first day means that, for a large number of cascades, they successfully attract a large number of rebloggers on their first day after publishing, which again confirms the tendency towards 'recency' on Tumblr (Chang et al., 2014). In fact, the global maximum, the day on which a cascade reaches its highest number of reblogs, is within a week after publishing a post, for about 70% of cascades, which again confirms the tendency towards 'recency' in Tumblr. In contrast, YouTube popular content peaks around the first six weeks (Borghol et al., 2011). The fact that about 42.46% of cascades reach their highest number of reblogs on the same day after publishing means that about 42.46% of Tumblr's posts start with a spark in popularity on the first day after being published.

Are Tumblr's cascades bursty?

One way to estimate cascades' burstiness is by looking at the proportions of idleness days and peak days. On average, peak days comprise 1.9% of the overall cascade lifetime, while idleness days comprise 28% of the cascade's lifetime. Apart from idleness days and peak days, the rest of the days (70.1% on average), are days where there are some reblogging activities that are not high enough to be accepted as peak days. Additionally, on average there are 4.47 peaks in each cascade, which means that the cascades successfully reach a high number of reblogging four times on average during their lifetime.

Additionally, the proportions of numbers of reblogs computed for the four cascade categories shows that the total number of reblogs per cascade on all of the peaks days comprises roughly between 70% to 30% of the total number of reblogs. This means that peak days are only partially accountable for the total number of reblogs, and that the total reblogs are accumulated gradually during the cascade's active age.

Another way to estimate burstiness is to look at the difference in days between any two consecutive peaks, which is 36 days on average. In fact, the probability that the difference between two consecutive peaks is 10 days or less is 0.13, the probability increases to 0.49 when computed for 20 or fewer days. For 50 or fewer days the probability is 0.8, which means that in most cases peaks are 50 days or less apart from each other. Does this mean that cascades are bursty? Putting the average difference in perspective with the average active age of cascades (237 days), it is noted that the detected peaks are separated from each other. Thus, the cascades are bursty, i.e. feature periods of high, medium and low to no reblogging activities. The next section will expand on the notion of burstiness, as it will reflect on recurrence as another perspective of cascades' burstiness.

Recurrence Detection

Identifying recurrence is bound to the set of conditions and parameters used in the process, whether it is the conditions to identify the peak or the valley between two identified peaks. Cheng et al. (2016) state that for a peak to be identified, the number of reshares must increase before the peak and decrease after it. However, in practice if this condition is strictly applied it will minimise the number of detectable peaks significantly, because around a peak it is rare to have a smooth increase and a smooth decline. Instead, it is more practical to ensure that the number of reblogs/reshares before and after a given peak is less than the peak's number itself, without strictly enforcing the smooth incline and decline. Moreover, there are two conditions provided: the first condition restricts the number of reblogs in the valley to be half the smallest number of reblogs between the first and the second peak, while the second condition ensures that the number of reblogs for both peaks is larger than half the maximum number of reblogs per day between the two peaks. However, it turns out that after changing the value of v both conditions yielded similar results.

Facebook's analysis carried out by Cheng et al. (2016) includes analysis of multiple introductions of the same content in the platform, which is the major difference between the analysis in this thesis which was done on individual posts that were promoted as the most popular ones in a year. For multiple introductions of the same content there are some attempts to explain the increase in the number of reshares in Facebook, for instance, as a result of the fact that these post are being shared by popular pages (Dow et al., 2013). In other cases, these increases are bound to external stimuli, especially when the item being tracked is generic, a URL, or a hashtag (Myers et al., 2012; Bakshy et al., 2012). For individual copies, Cheng et al. (2016) found that 18% of images and 30% of videos recur. The findings presented here agree with Cheng et al. (2016): popular content does recur on Tumblr, but it occurs at a smaller rate.

In contrast to Cheng et al.'s (2016) analysis, Tumblr has a short initial burst duration, which means that the cascade peaks quickly then users lose interest in the content. However, because this happens very quickly it allows the content to re-peak again after a while. The same goes for the size of the initial burst: if it lasts longer or attracts large number of users it is unlikely that it will succeed doing so after a while, i.e., it won't recur because it immunises the large number of users who are exposed to it. Thus, the analysis on Tumblr's cascades shows that the initial burst is shorter and smaller in terms of the number of reblogs than for non-recurring cascades.

6.5 Chapter Summary

This chapter has presented the analysis carried out in this thesis. The first section looked into Tumblr as a platform for content sharing, and provided an analysis regarding the platform and discussed the impact of Tumblr's affordances on the cascade analysis. Reblogging is found more popular among Tumblr's users than liking, while the rate of comments is significantly low. Meaning that most of the time users are interested in the content thus they reblog or like it, but they rarely engage in conversations about the content.

The second section covered the structural and topological aspects of Tumblr's reblog network and it compared it to two other networks, also obtained from Tumblr. The reblog network is denser than Tumblr's social network, meaning that it has more connections as the users who reblog these posts tend to reblog more than one post, which is shown by the reoccurrences percentage across all posts (30%).

The third and fourth sections were dedicated to the structural and temporal aspects of cascades, where individual cascades (posts) were analysed to understand how diffusion occurs on Tumblr and how these posts (the top posts in a year) accumulated their sizes. The analysis done in these sections has shown that Tumblr's cascades have non-trivial sizes and depth in comparison to similar cascades from the literature. These cascades branch out in separate and long paths. Temporal analysis has shown the tendency towards recency in Tumblr as the posts get reblogged within an hour after publishing. Cascades on Tumblr are bursty, they go through a series of idleness and high-activity periods (peaks), but the proportions of idleness days are higher than those of peaks. There is no particular pattern detected as the cascades grow, they reach their overall sizes in several ways. Finally, Tumblr's cascades do recur but most of them exhibit one period of high reblogging activities.

Chapter 7: Discussion

> 'The Web is a piece of computing embedded in a social setting, and its development is as much about getting the embedding right as it is doing the engineering.'
>
> (Berners-Lee et al., 2006)

This chapter builds on the findings of the analysis in the previous chapter, expands on them, and explores other related aspects. The first section sheds light on the effect of the platform on content spreading and data harvesting, while the second discusses the size of cascades and argues over the meaning of large cascades as used in the literature.

7.1 The Platform's Effect: The Case of Tumblr

The functionalities provided by any social network platform shape its users' behaviours. These functionalities schematise, and sometimes control and limit, what users are able to do within the platform, and distinguish each platform from the others.

Information diffusion, as a phenomenon, can be observed in many forms online. However, they can all be categorised into **implicit diffusions** and **explicit diffusions** (Zafarani et al., 2014). The cascades analysed here fall under explicit diffusions because their networks are observable. Zafarani et al. (2014) differentiate between two types of explicit diffusion, based on the information available to the users involved in the process. The first of these is Herd Behaviour, where individuals base their decisions on global information available across the population. The other is information cascades, which is based on local information passed from the immediate neighbours.

The following sections address how 'explicit' cascades occur on Tumblr and discuss how Tumblr's functionalities (and other platforms) affect the likelihood of cascades' emergence, how data about cascades is collected, and what the implications are of the contextual data provided by the platfrom on the cascade construction.

7.1.1 Content Exposure and Discovery

For a post to be reblogged, users must first be exposed to it. Tumblr incorporates a number of mechanisms that expose its users to new content and enable the discovery of new content. These

include: the social network, tags and searches, staff picks, trends, 'fandometrics' [7] blog, and 'Year in Review' blog.

By following other users, the user will be exposed to her friends' newly-shared content in her feed. For Tumblr and many social network platforms, following others (i.e. the social network) is the main mechanism for content exposure. However, Tumblr has other mechanisms in place to promote content and increase its exposure.

Tumblr is a platform, the majority of users of which are young (Chang et al., 2014), and most of the time they get involved in fandoms. Fandoms are communities of users with similar interests, mainly of TV shows, films, celebrities, musical groups, etc. (Renwick, 2014). Through these fandoms users express and share their devotion, passion and feelings (DeSouza, 2013). Fandoms are not explicit entities but rather they are implicit communities that are identified with a number of tags (Bourlai & Herring, 2014). Members of fandoms discover content through a set of designated tags, i.e. if they are familiar with one tag they will discover the others from the posts they become exposed to (DeSouza, 2013).

Tags are an important aspect of Tumblr, as they identify posts and make them visible, since Tumblr's search mechanism searches only tags. Posts with no tags can hardly be discovered unless the user follows the blog (Xu et al., 2014). Tags reside in a separate component alongside the content, allowing users to include as many tags as they wish in their posts, which differs from Twitter, where hashtags are part of the tweet's content. This approach increases a post's exposure, allowing many users to discover it, which might lead to an increase in the number of reblogs. The top posts analysed here have on average about 8.33 tags, while the maximum number of tags for one post was 31 (Figure 7.1).

Additionally, Tumblr has a number of content promotion tools, some present seasonally and others weekly or daily. Each day Tumblr shows what is trending at a given moment, based on users' activities, and has a staff picks page that includes a list of curated posts selected by Tumblr staff. fandometrics is an official Tumblr page that provides a weekly review of the most popular fandoms in Tumblr, based on their tags. It rates fandoms according to the number of posts posted in a tag and the number of searches, reblogs and likes on posts with that tag. At the end of each year, Tumblr publish the 'Year in Review' blog, which contains the most popular posts and the most reblogged tags during the past year.

[7] https://thefandometrics.tumblr.com

Together, these mechanisms increase the likelihood of diffusion occurring. The non-trivial sizes of cascades that have been noted in the top posts' cascades are perhaps the result of combining all of these mechanisms. There had been much discussion recently about the algorithms that decide what the users will be exposed to. Dow et al., (2013) claims that even the news feed in Facebook is a stream of a user's friends' stories that are curated automatically using a ranking algorithm.

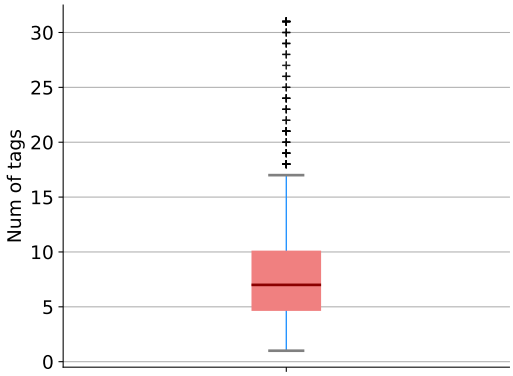


Figure 7.1 Distribution of the number of tags in each post

7.1.2 The Ability to Spread Content

As basic as it might seem, the ability to spread content is not an affordance of all platforms. Tumblr among few others (Twitter, Facebook, and Path) provides it through built-in functionality (reblog, retweet, share, repath) that allows users to spread content. For explicit diffusion, this functionality is vital as it makes the resulting cascades observable and traceable. However, the lack of this functionality in a platform does not preclude diffusion occurring. In fact, all platforms have many implicit diffusions that might occur eventually without the need of a functionality to support it. For instance, a user might decide to pick up a new tag, or she might share a URL or a photo. This, in turn, creates other forms of implicit diffusion.

Recently introduced platforms such as Snapchat[8], which is extensively used to exchange photos and short videos (snaps), has not provided a functionality that allows users to spread content. Nonetheless, other forms of diffusion might be encountered when users forward a public snap from a user's story to another user or a group, which sometimes results in these users following the original author. Sometimes users take screenshots and post them in their own stories. Instagram users, on the other hand, utilise third party applications to repost a photo while accrediting the original user.

[8] https://snapchat.com

The availability of a button to spread content has changed the way cascades are studied. Chapter 2 provided an overview of diffusion studies and showed how dramatically platform's affordance changed the methodology used to harvest cascades, construct them, and analyse them. But for a complete overview, not all cascades are explicit and in all platforms many implicit cascades occur. Explicit cascades are easier to track compared with implicit ones, as implicit ones are far more complex to observe, collect data from and to analyse.

7.1.3 Communication Style

Reblogging is more popular than liking, based on the analysis presented in chapter 6; on average, there are 7.9 likes for 10 reblogs. However, while comments were considered the main way to communicate publicly on Tumblr before the introduction of mentions (Chang et al., 2014), comment rates are extremely low; on average, there are only 0.16 comments for 10 reblogs.

These two ratios (reblogs to likes and reblogs to comments) emphasis few points of view. The majority of interactions in the top posts on Tumblr are non-verbal, i.e. via reblogging and liking. The low commenting observed on each post in the dataset means that Tumblr users rarely engage in discussions and they rarely express their reactions towards posts' content. In other words, they are interested in the content but they do not attempt to start conversations with other users. For instance, boyd et al. (2010) argued that retweets on Twitter were used as conversation starters, and they discuss different conversational practices using retweets. That is not the case on Tumblr. The main difference between the two is that Tumblr is, in most cases, a non-verbal platform, while Twitter is heavily dependent on textual interactions.

A number of reasons might explain the low commenting rate. Tumblr is mainly used to exchange photo memes and GIFs (Hillman et al., 2014). The proportion of photo posts in the dataset is 84%, Chang et al. (2014) reported 78% of their dataset was photos. These photo posts play an important role in Tumblr's culture and the multimodal communications used within the platform (Bourlai & Herring, 2014), as some of them have text within the photo itself.

Another reason for the low commenting rate might be related to the nature of personas that Tumblr users choose and the types of connections they seek. Hillman et al. (2014a) reported that Tumblr users often choose informal usernames. They also mentioned that, in contrast to Twitter and Facebook, most of their connections are not personal but based on common interests. This may explain the low rates since the users are gathered around the content they find interesting, not communication. However, Tumblr recently added a number of communication functions such as replies (@mentions), which were rare in the dataset, and messages, which are private messages between the users. Replies and messages are fairly new

features; replies was introduced in early 2014[9] while messages was introduced in November 2015[10]. Before that, and around the time when our dataset was harvested, the only way for users to communicate was via 'Ask'. Hillman et al. (2014a) mentioned that Tumblr users hesitate to communicate with each other using this 'Ask' feature because the question (message) and the response will be publicly visible. All these factors have an effect on the comment rates observed in the dataset.

In general, the high reblogging rate means that these users are actively engaged with the content, because when a post is reblogged it will be added to the reblogger's blog. Liking is more passive, as it only appears in the list of notes, though recently some blogs include a tab that lists the posts that have been liked by the user.

On some occasions, users reblog and like the same post; which raises an important question: what is the users' perception of the functionality available on social network platforms? It appears that users will always use the functionalities available in innovative ways that they were not originally designed for (Berners-Lee et al., 2006). Thus, while the intended purpose of these functionality is defined by the platform, users might exploit them for other purposes. Sometimes they can tweak their usage of a functionality to serve other needs. For instance, it is common on Snapchat to use screenshots as polls, where users take a screenshot of the snaps they choose. This will have consequences on any analysis of user interaction of a social network's functionality. boyd et al. (2010) surveyed Twitter users on their perception of the retweet functionality while Meier et al. (2014) surveyed favourite functionality likewise. Both reported that users' motivations to use the functionality were diverse, both in terms of its meaning and its possible purpose.

Kwak et al., (2010) noticed the existence of repetitive retweet, where users retweet the same tweet more than once. They did not elaborate on why this exists, or how it was dealt with in the construction step. It is not clear whether repetitive retweets were considered as new, isolated, components or as part of the same retweet tree. They constructed their cascade trees and topical forests by gathering the entire individual cascade trees for each topic.

[9] https://unwrapping.tumblr.com/post/74972171775/user-mentions-tumblr-apps

[10] https://support.tumblr.com/post/132943845192/youve-asked-us-for-real-instant-messaging-and

7.1.4 Data Harvesting Considerations

To be able to trace cascades, there are three questions that hugely affect the harvesting approach. These questions are in fact part of the components of the Information Diffusion Framework presented in Chapter 4.

Q1: What types of content will be considered?

Q2: Where can the data be found?

Q3: What information is available during the harvesting process?

These questions address three aspects of diffusion: whether it is explicit or implicit, how to fetch the appropriate data, and how accurate the cascade network can be, based on the available information.

Tracing cascades of a URL is completely different from those of a particular post or tweet. Cascades of URLs are considered as implicit diffusions. Any user can post the URL and can thus be found anywhere on the platform. Implicit cascades are, by definition, scattered, in contrast to explicit diffusions, where the causal relationships are preserved between the users. More recently, the cascades resulting from multiple introductions of memes to the network were investigated. Multiple introductions create explicit diffusion but also create disjoint cascade networks (Cheng et al., 2016).

When it comes to fetching the sources to harvest the data, there are two main types of content: *popular* (or viral or large) and *ordinary* content, which might or might not be popular. This has an effect on the data-harvesting step, i.e. if the focus is to study popular content, deliberate effort must be taken to detect and harvest such content. Gathering data from the API gets everything being shared within a given time interval, but the data collected might or might not be viral. For instance, the Twitter streaming API provides only 1% of the overall tweets at a particular moment. Researchers who rely on the API to gather retweets might miss so much (Petrovic et al., 2011). Goel et al., (2013) argue that the challenge that arises with the insufficient rate of 'small' and 'shallow' cascades is the need to harvest even more data, increasing the probability of finding a sufficient number (around one billion distinct events) of large cascades for the purposes of statistical analysis. Even with large amounts of data, however, the result is not guaranteed.

On the other hand, information available to researchers with non-privileged access to the data that comes with its own set of challenges resulting from missing data and limited access. For example, explicit cascade networks might be disjoint, with many isolated parts as a result of

missing or deleted information, which makes these branches isolated and difficult to link to the main cascade network. However, these shortcomings might be resolved with sufficient access to other information such as the social network or historic data. The missing links can be inferred using the social graph (Gomez Rodriguez et al., 2010) or by using historic information about content spreading between users. On Tumblr, the social graph (the network of who followed whom) is not accessible either via the API or via the user profiles. Some users choose to include the list of blogs they follow, but it is not mandatory as Tumblr allows its users to choose their own layout. So the only way to access the social network is with privileged access. Cascade networks can be used as a proxy to infer the possible connections between users, based on the observation that users mainly spread content they are exposed to through their social links (Kwak et al., 2010). However, this observation must be taken as an estimate because it needs sufficient contextual data such as clicks on feeds (Bakshy et al., 2012) or page impressions (Rotabi et al., 2017; Cheng et al., 2014) to support such an assumption. Again, information about clicks of feeds and page impressions needs privileged access. Also, privileged access is needed be able to study cascades of multiple introductions of the same content to increase the likelihood of detecting such content.

To summarise, the more access a researcher has, the more in-depth the analysis can be done and the more accurate the harvesting and construction can be.

7.1.5 The Value of Deletion information

Chapter 6 showed that it is common for users to revisit their reblogging decisions and they might either delete reblogs or deactivate their accounts. Deletion is an important aspect of social network behaviour. The rate of deletion indicates that the cascade network status on Tumblr (and presumably any other social network) can rapidly change. Thus, this must be considered during the data harvesting phase and analysis. Deletion has been looked at for many purposes: calculating the probability of rumour deletion (Friggeri et al., 2014), investigating the reasons behind deletion on Twitter (Almuhimedi et al., 2013), predicting deletion on Twitter (Petrovic et al., 2013), identifying regrettable tweets (Zhou et al., 2016). But it has not been addressed widely in cascades studies.

7.2 How Tumblr differs from other social networks?

This section will provide a highlight of the differences between Tumblr and the other social network platforms in light of the analysis presented in Chapter 6. The aim of this section is to provide an overview of Tumblr its functionalities and its users' behaviour and compare it with

other platforms. User behaviour includes those related to cascades and other ones that are related to other functionalities of Tumblr.

7.2.1 Tumblr's Functionalities

One of the most interesting findings is the higher reblogging rate in comparison with liking. Liking is still high and the ratio is 10 reblogs for 7.9 likes. The higher reblogging rate means that there is a higher degree of engagement with the content as reblogging means that the post will be added to the reblogger's blog. On the other hand, likes will only be shown on the same post or for some users who chose to show liked posts in the designated tab. On Twitter, which provides similar functionalities: retweet and like, 43% of tweets get at least one favourite and 36% of them get at least one retweet (ENGE, 2014). In addition, it has been reported that 25.5% of the tweets are actually retweets (using non-conventional retweet mechanisms) (Yang et al., 2010). As mentioned earlier, according to Tumblr's CEO: "Ninety percent of content on Tumblr is actually reblogged". This means that it appears that likes are more popular than retweets on Twitter but it is the opposite on Tumblr.

Another interesting finding is the remarkably low commenting rate on Tumblr; the ratio is 0.16 comments for 10 reblogs. In fact, reblogs with comments comprise only about 1.55% of the total reblogs. On Twitter, Liu, Kliman-Silver & Mislove (2014) report that about 35% of tweets are actually replies. This means that on Tumblr Users are interested in the content more than communicating with each other i.e., the majority of communication is non-verbal as oppose to Twitter.

Reblog reoccurrences happens when a user reblog a post more than one time. This affordance is not exclusive to Tumblr. Twitter employs a similar mechanism allowing users to retweet tweets several times. However, on Twitter retweeting brings an old tweet (or retweet) to the surface again. On Tumblr, each time a user reblog a post it will be considered as a new piece of content with a new ID. The analysis has shown that on Tumblr users might reblog more than once given the possibility to do so bit that is not the norm and when it happens it is not because they use reblogs for communication.

7.2.2 Tumblr's Reblog Network

On Tumblr, the social network is not accessible via the API, any attempts to analyse it require a privileged access that is not available for all researchers. In such case, the reblog network, constructed from all of the reblogging activities, acts as a proxy to estimate the social network (Xu et al., 2014). Several measures were used in Section 6.2 to analyse and compare the reblog

network with other networks in other social network platforms. The first measure is the density of the network; the analysis showed that the density of all of the three networks is significantly low. Meaning that Tumblr's networks are sparse and there is a low degree of connectivity among them. This is not significantly different from the density reported in other platforms such as Facebook, Twitter even in the blogosphere (See Section 6.2.1).

Reciprocity is another network measure that was used to measure the percentage of reciprocated edges in the reblog network. The reciprocity of the reblog network constructed from the most popular posts in 2014 is significantly low. However, it is similar to the reciprocity in another reblog network in Tumblr (Xu et al., 2014) and lower than Tumblr's social network (Chang et al., 2014). Nonetheless, the reciprocity in the reblog network is higher than the reciprocity of the retweet network according to Xu et al. (2014). This means that Tumblr's reblog network seems to be more connected than its counterpart on Twitter.

7.2.3 Cascades: Structural and Temporal Features

The analysis of Tumblr's cascades illuminated a number of Tumblr's cascade characteristics (structural and temporal) when compared to cascades on other platforms. A number of measures were used in the structural and temporal analysis (Section 6.3 and Section 6.4 respectively).

The first structural feature analysed is the branching factor (the number of children a node has). The analysis showed that the majority of nodes have zero children i.e., did not influence any user to reblog the post. The percentage of nodes with one child is similar to the percentage found in Liben-Nowell and Kleinberg famous work on Internet chain-letters (2008). While the percentage of nodes with no children is not reported, the similarity means that about 20% of the nodes have an influence on one user to share the same post. On the other hand, the mean branching factor per depth on Tumblr is higher than that on Facebook (Dow et al., 2013). In either case, the analysis was carried out on popular content: the most popular posts on Tumblr and Facebook's large cascade memes.

Another structural measure is the depth of Tumblr's cascades networks, the analysis showed that some of Tumblr cascades have non-trivial depths reaching 145 steps away from the root. Identifying these as non-trivial depths is based on the comparison with Facebook where the maximum depth reported is 40 (Adamic et al., 2012), Internet-chain letters where the maximum depth reported is 288 (Liben-Nowell & Kleinberg, 2008). On the other hand, on Twitter Taxidou and Fischer (2014) noted an average diameter of only four.

The temporal analysis aligns with the tendency towards recency on Tumblr as Chang et al. reported (2014), as 87% of first reblogs occur in the first hour after publishing, and 97.11% of first reblogs occur within 24 hours after publishing. On Twitter, only 50% of retweets occur within an hour and 75% of them occur within a day (Kwak et al., 2010). In addition, comparing the growth of cascades in size across time showed that there is no pattern that can be detected, cascades in different size categories reach their overall sizes in different ways. On the other hand, Linked-in invitations cascades has a linear growth pattern for both large and medium cascades (Anderson et al., 2015). The difference might be related to the nature of the shared content, as invitations tend to keep growing among interested users thus it gets bigger. On Tumblr, there are a number of factors that affect users' interestingness in the content such as timing factors and influencers' factors. Also, the mean branching factor across time on Tumblr is higher that that on Facebook (Dow et al., 2013) both are popular content. This means that Tumblr's content continue to attracts rebloggers at each point in its lifetime.

Cascades' burstiness analysis showed that about 42.46% of cascades reach their highest number of reblogs on the same day after publishing. In fact, the peak or global maximum when a cascade reaches its highest number of reblogs, is within a week after publishing a post. In contrast, YouTube popular videos reach its maximum views around the first six weeks after publishing (Borghol et al., 2011). This also aligns with the tendency towards recency on Tumblr. When analysing cascades' recurrence only 8.16% of cascades on Tumblr recur. Recurrence is higher on Facebook (Cheng et al., 2016), which means that the majority of cascades exhibit one period of very high reblogging activities then it either fades away or continue on attracting a moderate number of rebloggers throughout its lifetime. Also, recurring cascades have shorter and smaller initial bursts that those on Facebook, meaning that the longer a post remains alive, it will attract more users to reblog it and eventually might become very popular again.

7.3 How 'big' are large cascades?

This seems an easy question to answer, and a possible response might be, "well, large enough!" In fact, many studies in the field, especially these with privileged access to the data, use the term "large" loosely in their analysis and, in most cases, large cascades are the largest ones in the dataset. For instance, in the dataset used in this work, the mean cascade size is 56,539, while the median is 36,771, and about 18% of the cascades are larger than 100,000 (about 227 cascades). To put these numbers into perspective, in their study on Twitter, Goel et al. (2012) studied cascades of up to 10,000 nodes in their trees. Analysis by Cheng et al. (2014) of cascades on Facebook, states that the maximum cascade size is around 10,000 shares per post. Another study of Facebook analysed the diffusion of two photos that were shared 618,015 and 150,759 times

respectively (Dow et al., 2013). Thus, in light of the reported large cascade sizes on other platforms, it is noted that some of Tumblr's most reblogged posts exhibit high numbers of reblogging. However, there is much debate when it comes to the existence of large cascades. Goel et al. (2012) state that large cascades are rare while 99% of cascades are shallow and die in one step. Earlier work also claims that most cascades are shallow and fragmented (Leskovec et al., 2007b), while a few of them are relatively large (Leskovec et al., 2006a).

The first step is to achieve some level of agreement on an acceptable lower-bound at which a cascade can be considered large. Secondly, it might be true that most content posted does not end up 'large', but there is some evidence for their existence, when content spreads in high volume creating a cascade with complex structures. These need special consideration in the construction and analysis phases. Such content is a serious source for marketeers, who want to replicate its success. For a scientist wishing to reveal this phenomenon, however, the question is where to find them? In this thesis, they were deliberately harvested from posts that were promoted as popular, but it is unlikely that such content would be stumbled upon accidentally without prior planning and much speculation.

7.4 Chapter Summary

The focus of this chapter was on the implications of the platform on cascades' emergence, life, and harvesting. It compared Tumblr with the other platforms and provided discussion about its functionalities and how they are employed by its users. Besides the ability to share content with a click of a button, Tumblr makes available various content exposure and discovery mechanisms, which consequently increase the likelihood of content being shared. In addition, Tumblr is heavily dependent on GIFs, memes and non-textual communication forms. Thus, the rate at which users communicate with each other is far less than their agreement on how interesting content is by reblogging and liking it. This chapter also discussed how the type of content and the available information affects data harvesting. It concluded by discussing the meaning of large cascades and compared the sizes of cascades analysed in this work with others obtained from different platforms.

Chapter 8: Conclusions and Future Work

'We are drowning in information and starving for knowledge.'

Rutherford D. Roger

This chapter is divided into four sections; the first one provides an overview of the research conducted in this study highlighting the major outcomes at each stage. The second one discusses the research implications while the third section outlines the research contributions including the important findings and insights that have been gained from investigating cascades on Tumblr. The final section discusses future research directions that emerged after analysing the cascades of popular content.

8.1 Research Overview

The availability of rich data about human traces online has opened the door for many research opportunities that aim to unravel human behaviour on the web. Since their emergence, online social networks have been utilised to study the socio-technical aspect of the web. They create a sphere where users can create content, share it and interact with other users. This thesis investigated information diffusion, a phenomenon manifested by the spread of information on the web. The research process followed to answer the research questions is summarised in Figure 8.1. As the figure shows, the first step was to understand the information diffusion phenomenon, and what factors affect it on online social networks. The literature reviews (**Chapter 2 and Chapter 3**) provided a thorough overview about information diffusion and cascades. **Chapter 3** provided a survey of the cascade features that have been utilised in previous research. A pilot experiment then took place. The preliminary analysis focused on the structural aspect of cascades (**Appendix A**). The purpose of this phase was to grasp how Tumblr cascades can be harvested and analysed. After the literature review and the pilot experiment, an Information Diffusion Framework (IDF) was proposed in **Chapter 4** in addition to the cascade construction models in **Chapter 5**. An experiment was designed and conducted to analyse cascades on Tumblr (presented in **Chapters 6 and 7**)**.** In particular, this experiment provided analyses of Tumblr as a platform for content spread; it analysed the platform's functionalities and the structural and temporal aspects of the cascades that emerged on Tumblr, comparing their features to others in different platforms and contexts.

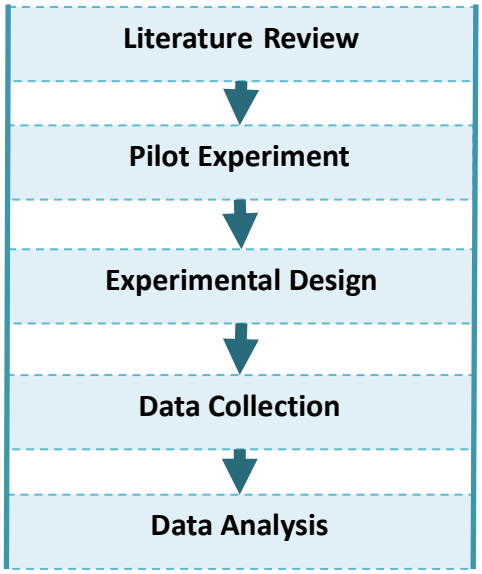


Figure 8.1 The research stages followed in this thesis

8.2 Research Questions

The main research question of this study is: **How does information diffusion occur on social networks?** This question has been divided into four sub-questions that are the focus of this thesis. This section will provide the answers for each of the four sub-questions.

RQ1: What are the factors that facilitate information diffusion in online social networks?

Information diffusion is a complex process, several factors have an effect on it including those that ignite it, facilitate it or hinder it. Studies that looked into information diffusion on online social network platforms have focused on several aspects of the process. Hence, the literature review in this thesis led to identifying these factors. These factors include factors that are related to the content itself such as the content type or degree of interestingness. Other factors are related to the context that can facilitate or hinder its spread. Contextual factors include the structure of the social network, users' influence or degree of homophily and other factors that are related to the platform under study including its affordances and the culture that emerged from using it. Then result of the diffusion process identified as the cascade can be analysed to estimate content's popularity. As a result, from the literature review and the pilot study and Information Diffusion Framework (IDF) was proposed. This framework aims to provide a holistic overview of information diffusion by identifying the components of information diffusion, namely, the content, the context and the cascade. IDF also identifies the relations between these components highlighting how they affect each other. Chapter 2 discussed information diffusion focusing on the first two components: the content and the context, while Chapter 3 discussed the third

component (the cascade) in details. IDF was introduced in Chapter 4, which provided an overview of the components and their relation to each other.

RQ2: How cascades networks can be constructed from minimal contextual information and missing/degraded information?

The answer to this question emerged after experimenting with the collected data about cascades on Tumblr. In particular, the pilot experiment presented in Appendix A helped in illuminating the challenges that arise from having minimal contextual information and degraded information. The first challenge is that Tumblr's social network is not available, also it is possible that users reblog more than once or delete their reblogs which would breaks the chain of reblogging between users. These challenges are addressed in Section 5.3.6, based on the literature (Chapter 3 and Appendix B) and the pilot experiment two cascade construction models were proposed. These two models (and their sub-models) aim at constructing cascade networks for individual posts on Tumblr. In both models, a number of heuristics are used in to handle these challenges and create accurate cascade networks that represent the flow of information between Tumblr's users accurately.

RQ3: What are the structural and temporal features of cascades?

As mentioned above, cascades are the result of the information diffusion, hence, they are the artefacts that allow researchers to analyse information diffusion in online social networks. In order to analyse cascades there must be a plausible approach to quantify cascades in a way that allows understanding them and comparing them to others on different platforms. The literature is rich with numerous ways to analyse cascades to serve various purposes. Reviewing the literature of cascades studies led to identifying two main classes of cascades features namely structural and temporal. Thus, a survey of these features was presented in Chapter 3 and Appendix C. This survey aimed at categorising these features so the structural features include those related to the cascade or to individual nodes in the cascade network. The value of this survey is that it brings those features together, defines them and highlights their significance as measure to estimate cascades.

RQ4: How is Tumblr, an online social network, used for information diffusion and what are the structural and temporal features of its cascades?

The analysis in Chapter 6 uncovered a number of facts about Tumblr as a platform for sharing content and the structural and temporal features of Tumblr's cascades. One of the most important findings in this thesis is that the most popular functionality among Tumblr users is reblogging, liking comes second but Tumblr users rarely engage in discussions. Chapter 7 looked

deeper into Tumblr's affordances and culture and how they affect the user behaviour on the platform. On the other hand, Tumblr's cascades have non-trivial have non-trivial sizes and depths in comparison to similar cascades on other platforms. Also, the analysis showed the tendency towards recency in Tumblr as reported in previous work. Tumblr cascades grown quickly in different manners, there was no particular pattern detected as the cascades grow. In addition, some of Tumblr's cascades recur meaning that they regain high popularity rates but the majority exhibit one period of high reblogging activities.

8.3 Research Contributions

This study has contributed the following:

1- An Information Diffusion Framework, IDF, that explains how actor factors, content factors, and platform affordances facilitate the spread of information.

The Information Diffusion Framework (IDF) proposed in this thesis takes into account all of the factors that facilitate the diffusion. It includes factors that help sparking the diffusion of the content and the factors that help to fuel its spread. The IDF has three main components that affect the diffusion process: the content, the context and the cascade. It encapsulates the factors that affect the diffusion under three components and it identifies the connections between each component. It makes a clear distinction between the information diffusion as a phenomenon and cascades as the result of diffusion: precisely, the structural representation of the diffusion as it manifests on online social networks. This framework can be used as a frame that can accommodate any diffusion-related research, as it conceptualises the different components of the diffusion and explains how they relate to each other.

2- A cascade construction model that yields accurate cascade networks from degraded/missing information and minimal contextual information.

Although the social graph of Tumblr (followers network) is not accessible, cascade graphs can be constructed by utilising the available information within notes, regardless of the information about the social connections. Unlike any other social network, Tumblr offers an explicit list of notes that shows who reblogged from whom. The proposed cascade construction models use both the reblogs and timestamps to construct a cascade graph for each individual post on Tumblr. The models extend the commonly used cascade construction model in a way that handles two issues often encountered on Tumblr. These are missing reblogs (e.g. deleted ones) and multiple occurrences of users in the same cascades, a case that is not investigated in the literature. These models, drawn from data mining approaches, allow structure to be found in

event sequences, even with minimal contextual information (e.g. unavailability of the social network, as on Tumblr).

3- A survey of the temporal and structural features of cascades and their implications.

Temporal and structural features of cascades are equally important. They complement each other and provide a systematic way to quantify the structural and temporal aspects of cascades, due to the complexity of these structures. The survey presented in this thesis lists the features the researcher used for diverse research purposes to quantify cascades. The survey categorises these features into: structural: cascade-centric and node-centric, in addition to the temporal features. It also defines each feature, explains its meaning and significance in relation to the cascades and presents how they were visualised in previous research. This survey of features will be useful for researchers who would like to study cascades, as it will give an overall overview of the significant measures that have been used earlier.

4- A thorough analysis of Tumblr as a platform for content creation and sharing, including comparisons between Tumblr's main affordances.

5- An investigation of the popularity-gaining phenomenon from structural and temporal perspectives.

6- A comparison between Tumblr's top posts' cascades and cascades in other OSN platforms.

This study has quantitatively analysed cascades on Tumblr in order to understand the structural and temporal aspects in Tumblr's cascades. In addition, this study has analysed Tumblr as a platform for information diffusion. The findings from this study are summarised below:

1. Reblogging is more popular among Tumblr's users than liking, while the rate of comments is significantly low. This means that most of the time, Tumblr's users communicate using non-verbal mechanisms; they are interested in the content, they reblog it or like it but they rarely talk to each other about it.

2. Although reblogging might be used for communication by adding comments, most of the time users reblog a post once. In fact, only 7.33% of all of reblogs are reoccurrences. Commenting is low in reoccurrences as well, so most of these reblogs are without comments.

3. The deletion rate of reblogs is not very high; there are about 27 deleted reblogs per 1000 reblogs, on average. However, the reblogs' deletion rate can help in estimating the dynamism of cascade network structures within the platform. On the other hand, deleted

reblogs result in creating cascade graphs that have separate components rather than the expected tree shaped cascades.

4. The reblog network is denser than Tumblr's social network. 30% of its reblogs are reoccurrences, which means that in most cases, the "Year in Review" blog attracts a wide range of users interested in different topics. Another finding that aligns with reoccurrences is that the reciprocity is low in the reblog network, which is expected, due to the fact that the blog contains the most popular content only, thus the dataset does not contain information about reblogging done by the posts' authors.

5. The branching factor, as a structural measure to quantify users' contributions to the overall cascade growth, has shown that, on average, in the four cascade models 67.44%-70.33% of the nodes have no influence and their branching factor is zero. This means that in most cases the cascade's total size is attributed to a few users who have an influence on a very large number of users.

6. A post's author's direct impact on the cascade is 8.94% on average, i.e., accounting for those who reblogged directly from the user.

7. A node's influence measured by the branching factor only can underestimate the node's actual contribution to the cascade. In some cases, nodes with a branching factor equal to one generated much larger subcascades.

8. Compared to cascades in different platforms, Tumblr's cascades have non-trivial depth, reaching a maximum depth of 32.78 on average, while the maximum depth across all cascades is 145.

9. Cascades branch out in long and separated paths; on average, a post has around 9196 or 11109 paths, depending on the construction model used.

10. The average branching factor at a depth of one is equal to 8.4; it decreases after that, remaining above two. It is higher than large cascades on Facebook.

11. The temporal analysis showed that posts get reblogged within an hour after publishing.

12. Cascades in different size categories reach their overall sizes in different ways; no particular pattern of growth was detected.

13. Cascades on Tumblr are bursty, they go through a series of idleness and high-activity periods (peaks), but the proportions of idleness days are higher than those of peaks.

14. Only 8.16% of cascades on Tumblr recur, meaning that the majority of cascades exhibit one period of very high reblogging activities and the rest are mostly moderate to low reblogging activity periods.

15. Most cascades have one high activity period at most.

16. About 40% of recurring cascades have their second peak within 50 days.

17. Cascades on Tumblr have short initial periods, which means that in most cases users lose interest in the content shortly after publishing.

8.4 Research Implications

Goel et al. (2012) state that there is one way to overcome the problem of using aggregated data in many offline diffusion studies, which is to utilise what they call "individual-level diffusion" data. This type of data conveys explicit and precise information about who influenced whom in the diffusion process and when it took place. This is, in particular, what makes Tumblr an ideal platform to study diffusion, because it provides this type of "individual-level diffusion" data explicitly. Thus, the cascades that have been analysed in this study are all explicit cascades; they are constructed from the list of notes. The availability of such an explicit list allows the cascade to be constructed as it spreads, gradually, without using aggregated or inferred data.

On the web, many implicit cascades occur: for instance, when users exchange a URL to a news article or a YouTube video using several platforms. More generally, on the Web, search engines use an established model that explains how hypertext pages become hubs. In the random surfer model, pages are visited randomly and a ranking process takes place to compute the PageRank for each page based on the number and quality of its in-links, URLs that point to that page. This model explains the overall popularity of pages as a PageRank score but it does not explain the series of events that made this page popular. Therefore, the ability to gather and analyse detailed data about the way content spreads, i.e. explicit cascades, helps in understanding these implicit cascades.

On the other hand, Bild et al. (2015) conjecture that cascade networks model real-world social, interest and trust networks better than the social network. Thus, the cascade networks can be better indicators of the shared interest and trust than the social network, especially on Tumblr, where the reblogging rates are very high and its users create their connections based on common interests.

To Summarise, the implications of this thesis are as follows:

1. IDF, the information diffusion framework conceptualises the diffusion process and provides a framework that helps the researchers to design their experiments.
2. The survey of the structural and temporal features of cascades will help the researchers to pick the most applicable features for their research purposes.
3. Harvesting and analysing cascades datasets come with their own set of challenges especially without privileged access. Such challenges include the content type (popular vs. ordinary), how is it going to be harvested and the effects on missing data and missing contextual information about the cascades.
4. Analysing explicit "individual-level diffusions" allows us to estimate the implicit diffusions that occur on the web across different platforms including online and offline diffusions.
5. Tumblr's analysis has shown the impact of the platform's affordances on the users and consequently the diffusion process, thus if someone intended to design a new platform these points must be taken into considerations.

8.5 Future Work

There are three main areas for future research directions; these areas will look into some different aspects of the information diffusion phenomenon and cascades.

The first area for the future is to investigate Tumblr's users' motives to reblog content. The analysis has shown that reblogging is very popular in Tumblr. However, reblogging is considered as one of explicit ways to show the degree to which users are engaged with the content, as it entails that the reblogged post will be added to the reblogger's blog. It has especially been shown that Tumblr is a platform with a distinctive characteristic: it revolves around fandoms, where users engage with each other based on their common interests. To investigate users' motives, qualitative research is needed to ask users about their reblogging behaviours and what it means for them in the fandom context. This area would fill the gaps between Web Science and Network Science, as shown in Figure 4.2.

The second area is related IDF, the information Diffusion Framework proposed in this thesis. The Framework takes into account all of the aspects of diffusion that are often studied separately. For instance, possible future endeavours would aim to investigate the relation between content and cascades, i.e., how different types of content create different types of cascades and what are the characteristics of these cascades.

The third area for future investigation is related to cascades' temporal features. Further experimentation with algorithms and parameters is needed for a better, probably tighter, definition of peaks and, possibly, a sensible estimate of their acceptable proportions within the cascade's timeline. In addition, further analysis is needed to examine the relationship between the cascade size and its activeness/idleness phases and the content. Do large cascades become large and accumulate more reblogs because they are being reblogged every single day? If so, what are the characteristics of this type of content?

Both the second and third future areas require a dataset of cascades, and it would be interesting to harvest a different type of dataset, as the dataset in this study is harvested from the popular content on Tumblr. A different dataset requires different harvesting methods, and it will come with its own set of challenges, especially without privileged access. For example, it would be possible to start from a seed blog and crawl its posts, then pick other blogs who have reblogged from this blog (a snowball sampling). This would allow exploration of different types of cascades that are not necessary large, and it will be interesting to compare non-popular posts' cascades to those of popular ones.

Appendices

Appendix A

The results of the preliminary analysis was Presented in The 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15).

2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

Investigating the Structural Characteristics of Cascades on Tumblr

Nora Alrajebah
University of Southampton
Southampton
United Kingdom
N.Alrajebah@soton.ac.uk

***Abstract*—Online social network platforms provide built-in functionalities that allow users to share information through their social connections. The study of information diffusion focusses on analysing how the spread of information enables us to understand the ways that users behave and interact online. Information diffusion manifests itself on social networks in the form of cascades of information, which are structural representations of diffusion events. The characteristics of information diffusion differ between online social networks partly depending on the capabilities of each. Tumblr is an online social network platform; it allows users to reblog more than once, to add a comment with a reblog, and to delete reblogs which presents certain challenges in the study of diffusion on this network. In this paper, we will identify and address those challenges, and we will examine the effects of these functionalities on users' behaviour and consequently on the structural characteristics of cascades on Tumblr.**

Keywords*—Tumblr, Social Network Analysis, Information Diffusion, Cascades.

I. Introduction

Online social network platforms provide built-in functionalities with which users can communicate and interact with others. These functionalities are categorised based on their purpose, i.e., *Following* is used to establish social links, *Liking* is used to express admiration, *Favouriting* is used as a bookmark, and functionalities such as *Retweeting* and *Reblogging* enable users to propagate content within their social circles. Users' ability to propagate content enables researchers to study the hidden tendencies and motivations for users to become involved in different activities within the various social networks, one of which is forwarding information [1]. This class of research is called information diffusion, which aims to study the way particular information spreads on social networks [2]. The information diffusion phenomenon includes a series of cascading behaviours through which users propagate information under the influence of their neighbours [3], [4]. Guille et al. [4] outlined three issues that need to be considered while studying the diffusion process: 1) the type of content that spreads, i.e., detecting popular topics; 2) the way in which information spreads, i.e., modelling and inferring the diffusion; and 3) the role of people in the spread of information, i.e., identifying individuals with influence. Building on this, for each information diffusion event, there are three fundamental components: the content that spreads, the context that facilitates the spread (influence, social network structure and homophily), and the result of the diffusion, known as a cascade. Thus, in the context of this paper, the term *cascade* is used to refer to the structural representation of a diffusion event in which a specific piece of information reaches different users [4]. Cascades are often constructed as a tree or a graph in which nodes represent users and edges represent the direction of information flow between the users [3], [5]. Cascade graphs are considered as a subset of the underalying social network that links users together through platform-related social links such as *Follow*. The direction of edges in the cascade graphs serves as an indication of a user's influence; e.g., if user B propagates a message that user A posted/propagated, then we say that A influenced B [6]. Obviously, the spread of a message is not due only to the user who created it (the author); instead, such spreading is the result of cumulative efforts by many users who participate by sharing and spreading it to their friends [7]. Each user who participates by spreading content therefore adds value to the overall cascade growth.

The structural characteristics of diffusion events (i.e., cascades) received considerable attention in the literature as many of them were utilised either to analyse cascades [8], model them [7] or predict their future growth [9], [10]. Investigating the structural characteristics of cascades provides us with many insights about the diffusion mechanism and the roles of the individuals involved in the process. On the one hand, understanding the structural aspects of a cascade and how it progresses can help us to explain why some cascades continue to spread and why others die. On the other hand, the structural characteristics of cascades can be utilised to predict the growth of cascades, which is vital for many purposes, such as determining influential users and marketing. In addition, the structural characteristics of cascades can be used to measure the degree of virality in the cascade. There are two types of cascade structures: the first one has elements of virality, which result in creating denser and more complex structures; the other one is more of a broadcast-like cascade where many individuals receive information from one source [11]. The more complex a cascade structure is, the more it exhibits contagious behaviour [11].

Tumblr, an online social network, provides a built-in reblogging functionality that allows users to add posts to their own blogs, thus contributing to the process of spreading posts by increasing their exposure rate. For each post, Tumblr maintains a list of notes that shows explicit information about who reblogged from whom, minimising the ambiguity of identifying users' influence in a cascade. This list of notes is the same in the original post and all of its copies (reblogged posts). Tumblr exhibits the unique functionality of allowing

ASONAM '15, August 25-28, 2015, Paris, France
 ISBN 978-1-4503-3854-7/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2808797.2808814

users to reblog a post more than once; i.e., users might appear more than once in different parts of the cascade graph. It is not clear why users reblog a post more than once; however, Chang et al. [12] noted that reblogging is used as a form of communication because users can add comments while reblogging. Besides communication, another possible reason could be to bring other users' attention to a post after a while in order to increase its exposure rate and, consequently, its chances of being reblogged or liked by other users. In addition, our initial observation suggests that, during the post's harvesting time, some reblogs could be missing from the list of notes, which affects the cascade graph and creates many disconnected and isolated components.

The aim of our paper is to investigate the structural characteristics of cascades on Tumblr while taking into consideration the challenges that arise from Tumblr's features. We performed an experiment to provide a preliminary analysis of cascades on Tumblr. We investigated the consequences of missing reblogs and multiple reblogging on Tumblr's cascades. To handle these cases, we propose a cascade construction model that provides a more accurate representation of Tumblr's cascades by utilising the order of reblogs to connect users to their appropriate influencers. The aims of the experiment presented in this paper are to examine:
1) The extent to which reblogging is used as a form of communication.
2) The extent to which users reblog posts more than once.
3) The significance of isolated components in relation to the overall cascade size.
4) The extent to which rebloggers contribute to the growth of cascades.

Our preliminary analysis shows that the reblogging functionality is more popular amongst Tumblr's users than the other functionalities such as liking and commenting. In addition, the percentage of reblogs with comments is notably very low. Also, most users reblog only once, and, in some cases where users did reblog more than once, the analysis shows that after three reblogs the number of reblogs per user in aggregate decreases. On average there were 98 isolated components in each cascade; this is caused by missing links in the cascade that suggest that deleting reblogs behaviour occurs quite often. Moreover, we found that 11% of nodes featured maximum branching factors belong to isolated components. Thus, isolated components play a major role in cascade growth and should be taken into account when analysing cascades on Tumblr.

This paper is organised as follows: Section 2 outlines related work on information diffusion and cascades. Section 3 presents Tumblr's characteristics, and Section 4 explains the cascade construction model implemented in this paper. Section 5 describes the experiment's design and Section 6 discusses the findings we obtained from the experiment. Finally, Section 7 concludes and outlines future work.

II. Related Work

The cascade notion (or diffusion event) studied in this paper consists of a seed individual who shares an item of information independent of any other individual, followed by other individuals who are influenced by the seed to share the same information [13]. In the literature, cascades have been studied from several perspectives; one study investigates the likelihood that a piece of information will be shared in the first place [14]. Another looks at the possibility that a popular piece of content will remain popular [15]. A third perspective focusses on the structure of cascades [3], [16], [17], and a fourth predicts the future growth of a cascade [9], [10]. This paper studies the structural characteristics of cascades on Tumblr using well-defined metrics adopted from the literature (See Section V).

In order to analyse the structural characteristics, cascades must first be constructed. According to previous research, the cascade construction task relies on two factors: i) the type of content that is propagated, and ii) the capabilities of the platform. Various types of content were tracked and analysed, and each one had a great impact on the way cascades can be constructed. For example, in the blogosphere, the type of propagating content is often URLs; thus, such cascades are constructed using timestamps, blogs' content similarities and posts' links [18], [3]. Twitter, on the other hand, has two built-in functionalities that have an impact on the cascade construction process. Users on Twitter can form an explicit social network by following each other, and they can use the 'Retweet' functionality to propagate content. Thus, depending on the content type (a URL, a hashtag, or a tweet), cascades are constructed using the social network structure [8], [7], interaction network [9], [1], timestamps [19], [9], [1], [6], [7], and retweets [8], [6], [7].

Our proposed cascade construction model uses both the reblogs and timestamps to construct a cascade graph for each individual post on Tumblr. Our model extends the commonly used cascade construction model in a way that handles two issues we encountered on Tumblr (multiple reblogging and deleted reblogs). To the best of our knowledge, the only cascade model encapsulating the effect of multiple events performed by the same individual (e.g. reblogging the same post multiple times) are the transcendental information cascades introduced in [20], [21], [22], a method which based on Kleinberg's work on bursty and hierarchical structures in streams [23]. These temporal data mining approaches allow for finding structure in event sequences even with sparse contextual information (e.g. unavailability of the social network as on Tumblr). We follow a similar approach in our cascade construction model, utilising timestamps to connect users to their most recent influencers. In addition, whenever a link is missing in a cascade, it is assumed to be an independent fragment of the cascade graph, and it will be considered a new root for a new sub-cascade graph [7].

There are two recent related studies about cascades on Tumblr; the first is the work of Chang et al. [12], who analysed Tumblr as a medium for information propagation, where reblog cascades are constructed each time a new post is added. They studied the correlation between users' reblogging frequency and their in-degree, and the correlation between the frequency of reblogging and the time since a user has registered. They also analysed cascades' depths, sizes and structures. In their paper, they noted that users might appear more than once in one cascade chain, as users tend to use reblogging with comments as a form of communication. Our paper investigates this further as we analyse multiple occurrences of users in one cascade graph, and their commenting behaviour as well.

2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

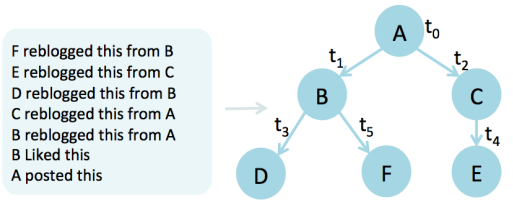


Fig. 1: A cascade graph constructed from notes list, top notes: most recent ones

Xu et al. [24] constructed a reblog network of all reblogging events (cascades) that occurred on Tumblr for four months. The reblog network is a weighted directed graph in which nodes are users and edges represent the direction of a reblog; i.e., if A reblogs B for N times, there will be an edge from A to B, and the weight of that edge is N. Our work is different from Xu et al., as our goal is to analyse individual cascades' structures rather than the dynamics of the reblogging network on Tumblr.

III. Tumblr's Characteristics

Tumblr is a hybrid social network platform [12] that allows users to create blogs and publish posts in any multimedia form, yet, like any social network platform, it provides various social interaction functionalities, such as following, reblogging and liking [24]. Reblogging is the main diffusion functionality provided by Tumblr. It allows users to reblog posts to their own blogs. Once a post is reblogged, it will appear in the reblogger's blog with a new ID. However, it will still link to the original post and the original author. Users can add a comment with a reblog and can reblog both original and reblogged posts. Reblogs appear as notes for each post in the following format:
X reblogged this from Y
X reblogged this from Y and added a comment

The availability of an explicit, unified and chronologically ordered list of all users who reblogged and liked a post makes the cascade construction task relatively easy (see an example in Figure 1). However, there are two cases that add complexity to the construction process. First, users can reblog the same post more than once; i.e., users might appear more than once in different parts of the cascade graph. Second, in some events, some notes might be deleted; this might occur when users delete their reblogs. In such cases, the notes list will have some missing links, which creates isolated components within the cascade graph. In the next section, we will discuss the cascade construction model used to handle such cases.

IV. Cascade construction model

In our experiment, the diffusion of individual Tumblr posts will be analysed; hence, for each post, one cascade graph will be constructed. Tumblr's social network graph is not accessible through the API; i.e., it is impossible to know who is following whom on Tumblr. Nevertheless, cascades can be constructed by using the reblogging information that appears in the notes section. In this graph, users are represented by nodes, and each reblogging note forms an edge between two nodes in

TABLE I: An example of the list of notes for a post on Tumblr

A posted	-
B Reblogged from A	OK
C Reblogged from A	OK
D Reblogged from B	OK
F Reblogged from E	Case 3
B Reblogged from E	Case 1
G Reblogged from B	Case 4
I Reblogged from H	Case 3
D Reblogged from C	Case 2
B Reblogged from A	Case 1 & 2
C Reblogged from D	Case 1 & 4
J Reblogged from B	Case 4

the cascade. These edges represent different information paths that the post spread amongst Tumblr's users. In addition, the explicit list of notes is ordered chronologically such that the most recent reblogs appear at the top of the list while the older reblogs are found at the bottom. Reblogs' ordering is important for making decisions about connecting nodes in certain complex situations that we will explain below.

Following the common cascade construction model, we define a cascade graph $C = \{V, E\}$ where V is the set of users and E is the set of edges in the cascade graph. A user $v \in V$ will be linked to its influencer $u \in V$, and an edge $(u, v, t) \in E$ will be added to denote that a user v reblogged from another user u at time t. The edge's direction indicates the direction of information flow between users. In an ideal scenario, the result will be a tree shape where the root is the post's author and each user who shares this information is represented as a node linked to the user whom they reblogged from. However, there are special cases that need to be taken into consideration while constructing this cascade graph. Because it is possible for one user to appear more than once in one post's notes list, many special cases arise from such flexibility. The ability to reblog a post many times and to delete reblogs/blogs cause the following cases: Case 1: A node that was once a parent might appear as a child afterwards by reblogging the same post after another user. Case 2: A node might have more than one parent as the cascade grows; i.e., a node might reblog the same post after different (or the same) parents. Case 3: A node might reblog from another node that does not exist in the list; i.e., a node might reblog from another node that might have deleted its reblog, and so it becomes isolated from the rest of the tree. Case 4: A node might reblog from another node that exists more than once in different parts of the cascade graph, which makes it difficult to choose which parent node is the correct one in such a scenario. Figure 2 demonstrates the cascade graph that is constructed following the common construction model for the list of notes in Table 1.

Taxidou et al. [7] stated that there are three approaches to handling isolated cascade components: 1) Ignore the isolated components and analyse the largest component (large connected component). 2) Analyse the forest as a whole, taking into consideration each isolated component. 3) Infer missing links between isolated fragments and the main component. The same approaches can be applied to problematic edges caused by reblogging more than once; i.e., either ignore problematic edges or consider them within a cascade. The first approach is not practical and has a major drawback, because if we ignore isolated components and problematic edges then many

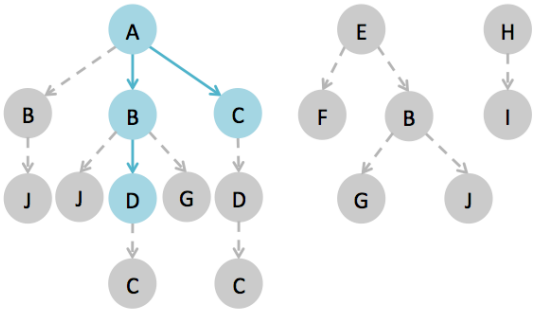


Fig. 2: The cascade graph obtained from cascade construction model without the heuristics

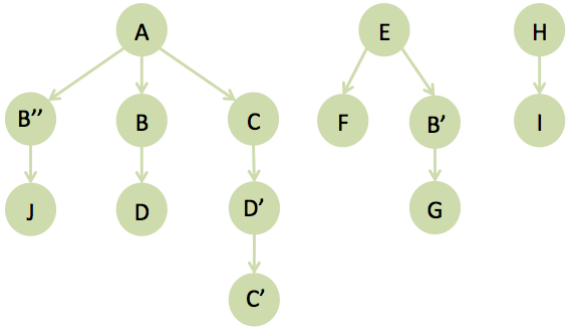


Fig. 3: The cascade graph obtained from cascade construction model with the heuristics

important aspects of the cascade will be missing. In fact, each new edge that is added to the cascade graph enhances our understanding of a cascade's dynamics and growth over time. Therefore, to overcome these challenges, the reblogs' order is used as an indication for linking nodes to their parents in such a way that no significant information about a cascade is missing [20], [21], [22]. Therefore, a cascade-graph construction model is used to construct a cascade graph that follows these simple heuristics:

1) Each node can be a parent for an unlimited number of nodes.
2) Each node can only be a child once, and if it becomes a child again then it will be added to the graph and labelled differently (keeping a list of the most recent copy of each node).
3) Whenever a new child is added to the graph, it will be linked to the most recent parent copy, adopting the most recent influencer model used in [25], [7]. It is also possible to link to either the least recent or the most recent influencer; however, linking to the most recent influencer is arguably more accurate given the fact that Tumblr's users encounter the most recent posts or reblogs from the people they follow on their dashboard before older ones.

Applying these heuristics to cascades is only possible by utilising the order of reblogs, which helps to keep track of the most recent copies of nodes in the cascade. Figure 3 illustrates the same cascade from Figure 2, constructed following the three heuristics in the cascade-graph construction model. The purpose of adding the same user with different labels is to distinguish cases where users appear in different threads within the same cascade. For example, when the user G reblogs a post from B, linking G to B (least recent) has a completely different meaning than linking G to B' (most recent) as each one is located in a different part of the cascade. Using labels is important for representing the cascade structure accurately; hence, the cascade will be represented as a graph that consists of chains of nodes that are linked in such a way that they show the paths a post follows as it reaches different users within Tumblr's social network graph. Those paths are used as an indication of influence in this context: if a user X reblogs another user, Y, then it is said that user Y influenced user X to reblog that post. Figure 4 illustrates four examples of cascades constructed from Tumblr; some has many isolated components, while the others are well-connected and their shape resembles a tree.

V. Experiment Design

The goals of our experiment are to explore cascades on Tumblr, focussing on the structural features of such cascades in order to understand their dynamics and to gain some insights. Using DataSift, 5027 different interactions were collected in one day. The query used to harvest interactions was very simple: we asked DataSift to retrieve any interaction with a 'Game of Thrones' tag. From the interactions that were active during that timeframe, there were 2475 distinct interactions with such a tag. For each of these distinct interactions, we retrieved the original posts and we harvested all of the reblogging and liking activities for up to three months later. Only 1575 posts were chosen for the analysis; the rejected posts either had no identified author or had not been reblogged at least once.

A. Cascade Measures

For each cascade, a number of measures were calculated for the purpose of the analysis:
1. Number of reblogs: cascade size
2. Number of reblogs with comments
3. Number of likes
4. Number of users who reblog and like the same post
5. Number of reblogs per user in each post
6. Number of comments per user in each post
7. Number of components in the cascade graph
8. Number of nodes in the main component (the one that has the author as a root)
9. Number of nodes in isolated components
10. Branching factor (number of children per node) [17], which is used to calculate the fraction of nodes with no children, the fraction of nodes with exactly one child [16], and the fraction of nodes with more than one child. In addition, the branching factor of the root node is considered to be the scale, which measures the influence of the author on the cascade's growth [9].
11. Depth: Number of hops from the root node (author) [17].
12. Wiener Index (structural virality measure), it is computed

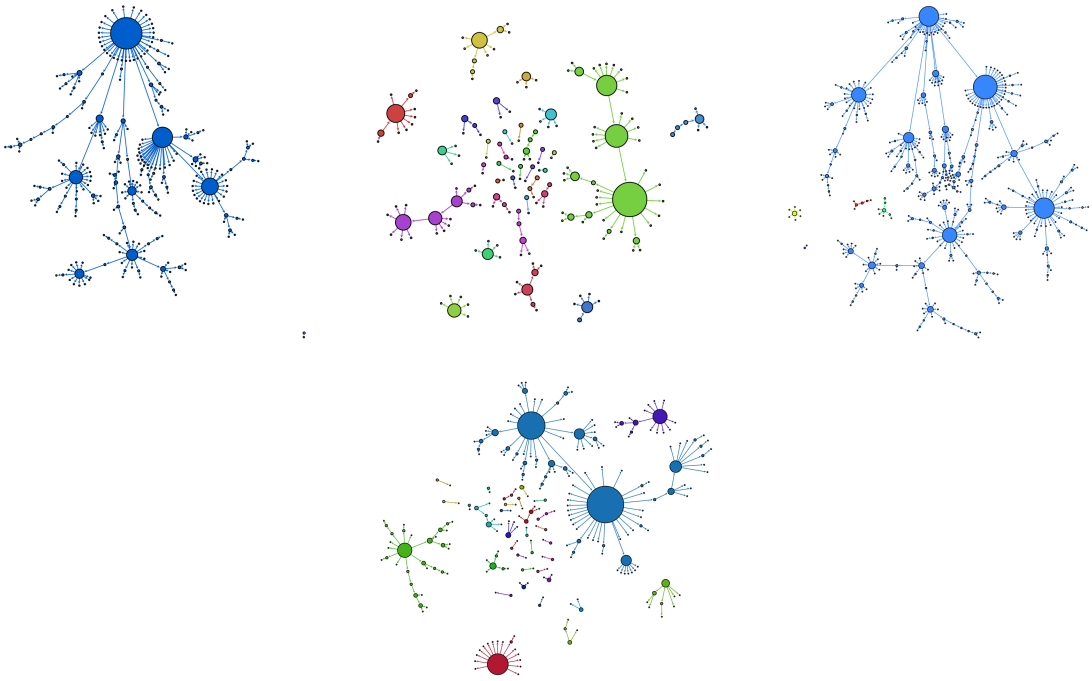


Fig. 4: Examples of Tumblr's cascade graphs, different colours denote different components while the node size denotes the out-degree (branching factor)

as the average distance between all pairs of nodes in a cascade [11], it indicates the degree of virality a cascade has, the higher the wiener index is the more viral the cascade is (the more branching it has at each level). Cascades with low wiener index do not exhibit virality features, the post will reach many people then stops (resembles broadcasting).

VI. Results and Discussions

This section discusses cascades' structural characteristics; it is divided into four subsections. The first looks at reblogging, liking and commenting functionalities, while the second focusses on cases where users reblog more than once. The third part discusses the effects of missing links and the significance of isolated components in cascade growth. Finally, the fourth part discusses rebloggers' contributions to the cascade. The cascade construction model allows the same user to appear more than once in the cascade graph using different labels. Therefore, instead of handling every single occurrence of a user separately, in some cases a user's features were taken in aggregate if the purpose was to find the cumulative contribution of a certain user to the cascade.

Figure 5 illustrates the distribution of cascade size, which shows that the majority of cascades in our dataset are small. The cascade size ranges from tiny ones (minimum cascade size = 2) to large ones (maximum cascade size = 170060), while the mean cascade size is 6190. We then computed the wiener index for all cascades and plot it in Figure 6 that shows the relationship between posts' popularity (measured as cascade size [11]) and their structural virality. The wiener index value increases as the cascade size increases. Hence, posts on Tumblr seem to have a high degree of branching at each depth, which increases as the post attracts more users to reblog it (becomes more popular).

A. *Reblogging*

On average, a Tumblr post is liked 4266 times, which is less than the average number of reblogs per post (the mean cascade size = 6190). However, knowing that Tumblr users can reblog and like the same post, the analysis showed that 76% of users only reblog (Figure 7a). The percentage of users who attempt reblog and like the same post is small (24%) in comparison with the percentage of those who reblog only. This suggests that Tumblr users tend to engage with content explicitly by reblogging it and adding it to their own blogs. Based on [12] observations, Tumblr users use reblogging as a form of communication by utilising comments that are permitted if a user reblogs a post. However, the analysis shows that, in general, the percentage of reblogs with comments is very low (1.18%) (Figure 7b). This makes the commenting functionality highly insignificant, since Tumblr users rarely use them. On average, posts only have about 72 comments, and the maximum number of comments one post has is 4199. Looking at individual users' reblogs with comments, the maximum number of comments was 21 per user per post in aggregate, and the mean was 0.1, which confirms that comments are not often used on Tumblr.

2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining

B. Re-Reblogging

As stated by [12], Tumblr users sometimes appear more than once in one list. The analysis shows that, on average, the number of reblogs per user per post is 1, which means that most users (94.3%) only reblog once. The distribution of reblogging per user per post shows that the reblogging rate decreases drastically after the third reblog by the same user (Figure 8).

We then looked at the maximum reblogs per user in each post separately. Amongst all posts, 24.8% were reblogged once by all users (maximum reblogs per user equals one), followed by 22.6% of posts with a maximum of three reblogs per user. Cumulatively, 98.4% of posts had a maximum number of reblogs per user less than or equal to 10, which suggests that only a few users reblog a post more than ten times.

C. Isolated Components

Normally, the shape of a cascade graph for one post would look like a tree that has one root (post's author), and the nodes represent users who reblogged this post, connected to their influencers by edges. However, this was not the case on Tumblr; in most cases the cascade structure is not a tree but a forest where there are many isolated parts that are not connected to the root (author) because of missing links. On average, there were 98 isolated components in one cascade, and the maximum number of components in one cascade was 4525. It is obvious that, as the cascade grows in size, the possibility that it will have a greater amount of isolated components also increases. In addition, those isolated components make up a significant portion of the cascade. On average, the percentage of nodes that belong to isolated components is 21% of the overall cascade's size (out-Tree). To evaluate the contributions of nodes that belong to isolated components in relation to the overall size of a cascade, we use the branching factor measure, which indicates the degree to which individual users contribute to the spread of posts. Looking at all nodes that have the maximum branching factor and using the number of hops away from the author, 57% of these nodes were actually root nodes, 32% were nodes that belonged to the main cascade component and had a link to the root, and the remaining 11% were nodes that belonged to isolated components. In most cases, the author has the highest impact on cascade growth in comparison with other users within the same cascade. Although the contributions of nodes that belong to isolated components are small in comparison with the other two types, they should not be ignored.

D. Rebloggers' contributions to the cascade

We considered the overall contribution of authors (scale) and users with the maximum branching factor to the cascade size. On average, posts' authors contribute around 22% to the overall cascade size, i.e., on average, 22% of individual cascade events follow directly from the authors. However, users with the highest branching factor contribute an average of 25% to the overall cascade size. Thus, posts' authors are actually playing a major role in the growth of cascades, along with other users who score high in terms of the values of their branching factor.

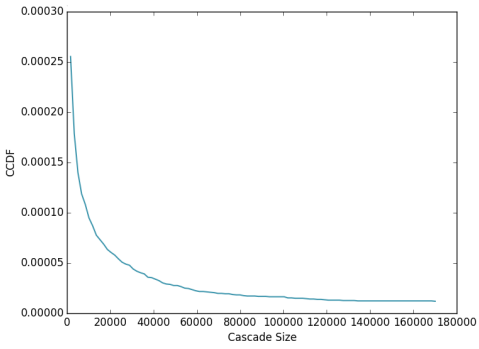


Fig. 5: The distribution of cascade sizes on a log-log scale

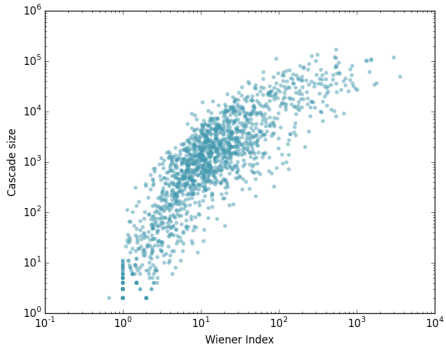


Fig. 6: The relationship between the structural virality and popularity (cascade size) on a log-log scale

Another aspect of a cascade structure is the number of children a node has; i.e., whether a cascade will continue to spread after reaching some users or if it will stop. On average, 71% of individual cascade events fail to continue (no children), while 18% of individual cascade events continue to spread for one hop away (exactly one child), and the rest (11%) spread either in depth or breadth (Figure 7c). This indicates that, in most cases, posts fail to spread after reaching certain number of users.

In addition, almost half the nodes with a maximum branching factor are within one hop away from the root node (54%), and, cumulatively, 92% of nodes are located within ten hops away from the root node. This means that major spikes in cascade growth occur within ten hops from the root.

VII. Conclusion and Future Work

In this paper, we analysed Tumblr as a platform for information diffusion. Our goal was to investigate the structural characteristics of cascades on Tumblr. We examined users' reblogging behaviour on Tumblr, such as reblogging more than once, using reblogging to communicate, and causing missing reblogs due to deleted reblogs or other unclear reasons. These cases were taken into consideration in our cascade construction model. Our model provides an accurate representation of the

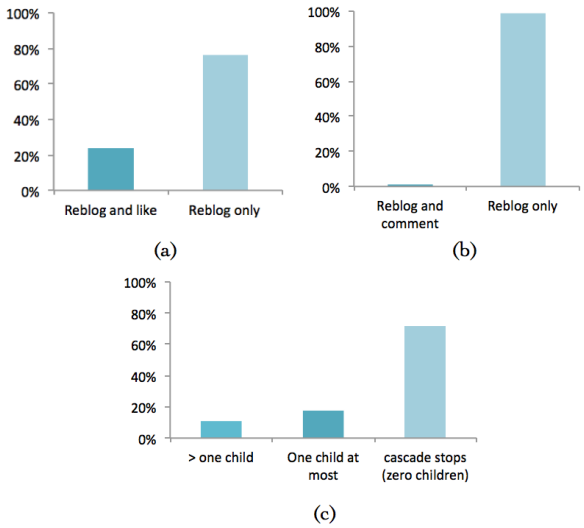


Fig. 7: (a) the percentage of users who reblog and like and only reblog, (b) the percentage of reblogs with/without comments, (c) the percentage of nodes that have a maximum branching factor in three categories: root, main component, and isolated component

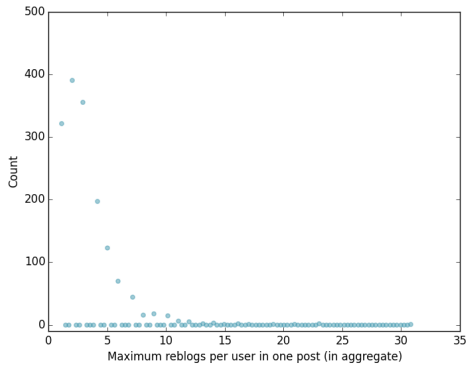


Fig. 8: The distribution of maximum reblogs per user in each post in aggregate

dynamics of cascades because it utilises reblogs in order to connect nodes to their most recent parent copy (i.e., most recent influencer). The term 'influence' is used in this context to refer to some sort of a local influence that is observed in a single diffusion event (i.e., a cascade). The findings from our experiment suggest that most users only reblog once, and, in some cases where users did reblog more than once, we found that after three reblogs the number of reblogs per user decreases. In addition, in a few cases, users reblog the same post more than 10 times. Although repeated reblogging is not significantly high, knowing that the rate of comments is very low raises the following questions: if users rarely comment, then why do they reblog the same content more than once? What is the probability that the users who reblog more than one time are actually bots? And if they're not bots what is their motivation to promote a post by reblogging it again? These questions open an interesting area for future work. On the other hand, cascade graphs on Tumblr are not tree-shaped ones but rather graphs with many isolated components due to missing reblogs. These isolated components represent a considerable portion of the overall cascade; around 11% of nodes that feature maximum branching factors belong to these components.

The dataset used for the analysis was small and collected over a short period of time, and it has the same tag; thus, our analysis is limited. For future work, we are interested in applying the same analysis to a more representative dataset extracted from Tumblr (e.g., posts with other tags and a larger sample size). Another area for future work is to investigate cascade reconstruction, to link isolated components to the main cascade component. The underlying social network is not accessible through the API, so reconstructing cascades might rely on historic information about previously reblogged posts. In addition, we plan to apply cascades' structural characteristics to predict future growth in cascades' sizes.

ACKNOWLEDGMENT

We thank Leslie Carr and Thanassis Tiropanis for their guidance and valuable insights, we also thank Markus Luczak-Rösch, for valuable discussions and comments that greatly improved this work. This work is supported by a scholarship from King Saud University.

REFERENCES

[1] D. M. Romero, B. Meeder, and J. 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*, ser. WWW '11. ACM, 2011, pp. 695–704.

[2] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," in *Proceedings of the 13th International Conference on World Wide Web*, ser. WWW '04. ACM, 2004, pp. 491–501.

[3] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst, "Cascading behavior in large blog graphs," in *7th SIAM International Conference on Data Mining (SDM)*, 2007, pp. 1–21.

[4] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Record*, vol. 42, no. 2, pp. 17–28, 2013.

[5] I. Taxidou and P. Fischer, "Realtime analysis of information diffusion in social media," *Proc. VLDB Endow.*, vol. 6, no. 12, pp. 1416–1421, Aug. 2013.

[6] L. Yang, T. Sun, M. Zhang, and Q. Mei, "We know what @you #tag: Does the dual role affect hashtag adoption?" in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. ACM, 2012, pp. 261–270.

[7] I. Taxidou and P. M. Fischer, "Online analysis of information diffusion in twitter," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, ser. WWW Companion '14. ACM, 2014, pp. 1313–1318.

[8] K. Lerman and R. Ghosh, "Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*. AAAI, 2010, pp. 90–97.

[9] J. Yang and S. Counts, "Predicting the Speed, Scale, and Range of Information Diffusion in Twitter," in *Fourth International AAAI Conference on Weblogs and Social Media ICWSM*, 2010.

[10] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. ACM, 2014, pp. 925–936.

[11] S. Goel, A. Anderson, J. Hofman, and D. Watts, "The structural virality of online diffusion," *Preprint*, vol. 22, p. 26, 2013.

[12] Y. Chang, L. Tang, Y. Inagaki, and Y. Liu, "What is tumblr: A statistical overview and comparison," *ACM SIGKDD Explorations Newsletter*, vol. 16, no. 1, pp. 21–29, 2014.

[13] S. Goel, D. J. Watts, and D. G. Goldstein, "The structure of online diffusion networks," in *Proceedings of the 13th ACM Conference on Electronic Commerce*, ser. EC '12. ACM, 2012, pp. 623–638.

[14] S. Petrovic, M. Osborne, and V. Lavrenko, "RT to win! predicting message propagation in twitter," in *Proceedings of the Fifth International Conference on Weblogs and Social Media*, 2011, pp. 586–589.

[15] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in twitter," *JASIST*, vol. 64, no. 7, pp. 1399–1410, 2013.

[16] D. Liben-Nowell and J. Kleinberg, "Tracing information flow on a global scale using internet chain-letter data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 12, pp. 4633–4638, 2008.

[17] P. A. Dow, L. A. Adamic, and A. Friggeri, "The anatomy of large facebook cascades," in *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM*, 2013.

[18] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," in *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '05. Washington, DC, USA: IEEE, 2005, pp. 207–214.

[19] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010, pp. 1019–1028.

[20] M. Luczak-Rösch, R. Tinati, K. O'Hara, and N. Shadbolt, "Socio-technical computation," in *the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, ser. CSCW'15 Companion. New York, NY, USA: ACM, 2015, pp. 139–142.

[21] M. Luczak-Rösch, R. Tinati, and N. Shadbolt, "When resources collide: Towards a theory of coincidence in information spaces," in *Proceedings of the 24th International Conference on World Wide Web Companion*, ser. WWW '15 Companion. Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 2015, pp. 1137–1142.

[22] M. Luczak-Rösch, R. Tinati, M. Van Kleek, and N. Shadbolt, "From coincidence to purposeful flow? properties of transcendental information cascades," in *International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2015*. IEEE, to appear.

[23] J. Kleinberg, "Bursty and hierarchical structure in streams," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '02. ACM, 2002, pp. 91–101.

[24] J. Xu, R. Compton, T.-C. Lu, and D. Allen, "Rolling through tumblr: Characterizing behavioral patterns of the microblogging platform," in *Proceedings of the 2014 ACM Conference on Web Science*, ser. WebSci '14. ACM, 2014, pp. 13–22.

[25] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: Quantifying influence on twitter," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. ACM, 2011, pp. 65–74.

Appendix B

The table in this appendix presents the different cascade construction approaches used in different platforms, it differentiates between the different items that diffuse and the diffusion mechanism either if it is provided by the platform or adopted by the users. Moreover, it differentiates between the structure of the resulted cascade network and consequently the types of nodes and edges.

Citation	Platform	Data used for construction	Diffusion mechanism detected	Structure of cascade network	Nodes	Edges
(Adar & Adamic, 2005)	Blogs	Cascades are inferred using: Posts text, explicit links to other blogs, and other features about the blogs network, the blog and the timestamps.	Posting a URL.	Cascade networks (Trees)	Blogs	Inferred edges show the direction of diffusion of information between them.
(Leskovec et al., 2006b)	Recommendation networks	Information about: Product, time of recommendation, was the product purchased, time of purchase.	Recommending a product.	Separate group recommendation networks. And separate product recommendation networks.	Customers	Customers' product recommendations.
(Leskovec et al., 2007b) (Leskovec et al., 2006a)	Blogs	In-links and out-links between blog posts and timestamps.	Linking between posts: posts have links to other posts written in the past.	Post network: links posts in different blogs. And, blog network: collapsed and weighted version of post network. Both are forests. From post network they extracted separate cascade trees.	Posts and blogs	Links between posts and blogs.
(Liben-Nowell & Kleinberg, 2008)	Internet chain letters	Each copy has an ordered list of users who forwarded the petition.	Forwarding of a petition letter from one user to another.	Inferred cascade trees by removing edges that did not appear in a sufficient number of copies.	Users	Direction of information flow.
(Kwak et al., 2010)	Twitter	Details are not mentioned.	Retweet	Retweet trees for each tweet in the dataset. And forests for each topic.	Users	Direction of retweeting activities.

Citation	Platform	Data used for construction	Diffusion mechanism detected	Structure of cascade network	Nodes	Edges
(Galuba & Aberer, 2010)	Twitter	Tweet texts, timestamps, social network.	Posting a URL independently, or crediting the source using 'RT @'	For each unique URL: 1 F-cascade network constructed using the follow (social) network and timestamps. 2 RT-cascade network using retweet credits (Forest due to the nature of crediting).	Users	Direction of retweeting activities.
(Yang & Counts, 2010)	Twitter	Tweets' texts contain topics and mentions other users.	Tweeting	Diffusion networks linking users if they mention each other and include a topic with timestamps.	Users	Direction of topics' diffusion.
(Adamic et al., 2012)	Facebook	Status updates that include the meme (or its variations) and 'copy', 'paste' and 'repost'. Lists of users who commented on users' status. Timestamps	Copy and paste of memes	Large cascade network (Forest) where there are nodes with no parents.	Users	Denotes that a user commented on a status and posted it to his friends.
(Bhattacharya & Ram, 2012)	Twitter	News agencies' tweet streams. Users' tweets that contain URLs from the chosen news agencies.	Retweeting from news agencies and users. Posting of news URLs.	Separate cascade networks for each news agency (Ego networks).	Users	Weighted edges between users denote retweeting or replying relationships.
(Dow et al., 2013)	Facebook	Tracked: reshare Inferred: reshare, timestamp, clicks in News Feed	Reshare	Separate cascades networks (Trees). Radial representation of cascades where nodes' distance from the centre is a function of time.	Users	Direction of reblogging activities.

Appendix B

Citation	Platform	Data used for construction	Diffusion mechanism detected	Structure of cascade network	Nodes	Edges
(Chang et al., 2014)	Tumblr	Reblogs	Reblog	Separate cascades networks (Trees)	Users	Direction of reblogging activities.
(Xu et al., 2014)	Tumblr	Reblogs	Reblog	Large reblog network	Users	Direction of reblogging activities.
(Anderson et al., 2015)	LinkedIn	Signups, timestamp	Invitations	Forests Roots: uninfluenced signups	Users	Direction of invitations between users.
(Bild et al., 2015)	Twitter	Retweets	Retweet + “RT@” in tweet’s text	Large retweet network	Users	Direction of retweeting activities to source retweeted destination.
(Adamic et al., 2016)	Facebook, memes	Social network, time, text similarity measures	Copy and paste of textual status update	Forests	Meme variants	Links to variant parent.
(Cheng et al., 2016)	Facebook	Resharers: users and pages.	Reshare	Forests as it accounts for multiple introductions.	Resharers	Not mentioned.

Appendix C

The table in this appendix presents a survey of the structural and temporal features of cascades. It categorises them, defines them and provides information about their significance in cascade analysis studies.

Type	Feature	Definition	Significance	How it was analysed	References
Structural: cascade-centric	Depth, Range, distance to the root	Represents the height of a cascade, it is calculated using the number of subsequent occurrences of message passing events, i.e. maximum number of hops or range of influence. Maximum depth and average depth can be measured too.	Indicates the shape of a cascade, and how far it travels away from the source within the network. When all distances to the root are gathered, they can help assessing whether a cascade is shallow or deep.	Distribution of nodes depths	(Galuba & Aberer, 2010) (Bakshy, *et al.*, 2011) (Adamic et al., 2012) (Goel et al., 2012) (Dow et al., 2013) (Chang *et al.*, 2014)
				Regression model to quantify the predictability of range using different features.	(Yang & Counts, 2009)
	Median depth	Median distance from root	Similar to depth	As a quantity for each cascade which then was averaged for all the cascades together.	(Liben-Nowell & Kleinberg, 2008)
	Maximum depth, maximum hop count, cascade height	Maximum distance from the root	Similar to depth	Distribution of cascade heights	(Kwak et al., 2010)
	Width	The maximum size of a set of nodes which share the same depth.	Indicates the extent to which a cascade is narrow or wide. It gives hints about the factors that make a message quite popular at one stage within the cascade.	As a quantity for each cascade which then was averaged.	(Liben-Nowell & Kleinberg, 2008)
	The fraction of nodes with exactly one child	How many nodes in a cascade with exactly one child.	Indicates missing or unsuccessful cascade event.	As a quantity for each cascade which then was averaged.	(Liben-Nowell & Kleinberg, 2008)

Type	Feature	Definition	Significance	How it was analysed	References
Structural: cascade-centric	Scale	The number of nodes influenced by a message at depth equals one.	Indicates how popular/interesting a message gets soon after its first appearance.	Regression model to quantify the predictability of scale using different features.	(Yang & Counts, 2009)
	Wiener index	It is used to measure the structural virality of a cascade. It is computed as the average distance between all pairs of nodes in a cascade.	Gives an indication of the cascade shape, the higher the Wiener index, the more viral the cascade. Cascades with low Wiener index resemble a star shape, where there are few hubs that create the cascade. Wiener index increases with the increase in cascade size.	Distribution of Wiener index.	(Cheng, Adamic, Dow, *et al.*, 2014) (Anderson et al., 2015)
	The percentage of adoption per depth	The percentage of adoption events that occur at each depth from the root.	Counting the percentage of adoptions within one degree of a root could indicate whether epidemic-like cascade occurs in the dataset, i.e. if the majority of adoptions recorded in the dataset are within the first few degrees from a root, then one could conclude that most cascades are shallow and small.	The distribution of adoptions by depth.	(Goel et al., 2012) (Anderson et al., 2015)
	Number of nodes at depth = 1	The number of nodes that are one step away from the author.	Nodes (users) at depth 1 are the ones who share directly from the author, meaning that they assumingly were exposed to the author's post directly. It might be that they arrive via external resources or direct links. Although there is a possibility that users click on the original post and share from the author rather than from user they receive the post from.	Fraction of reshares by depth.	(Dow et al., 2013)

Type	Feature	Definition	Significance	How it was analysed	References
Structural: cascade-centric	Connectivity Rate	The percentage of users who have one edge at least, hence they were influenced by other users.	Shows whether an edge exists between any two nodes in the cascade. It is useful to examine whether users get their information from the social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree.	Distribution of connectivity rate	(Taxidou & Fischer, 2014)
	Root Fragment Rate	The percentage of nodes that have either direct or indirect connection to the root node.	Shows whether each node in the cascade is actually linked to the root or not. It is useful to examine whether users get their information from social links (i.e. explicit links via following) if this information was taken into account while constructing the cascade tree.	Distribution of root fragment rate	(Taxidou & Fischer, 2014)
	Diameter	The diameter of a network.	Shows whether cascades are deep or shallow.	Distribution of diameter	(Leskovec et al., 2007b) (Taxidou & Fischer, 2014)
Structural: Node-centric	Fanout/ Branching factor	The number of subsequent cascades that follow directly from a particular node (user).	Measures individual's influence on the overall cascade.	-Distribution of fanout per user. -As a function of depth and time (mean branching factor per hour & mean branching factor by depth). -Branching factor by audience size.	(Gruhl et al., 2004) (Dow et al., 2013)
	Size of Sub-cascade	The size of the sub-cascade under a particular node.	Measures individual's influence.	Sub-cascades size by audience size.	(Galuba & Aberer, 2010) (Dow et al., 2013)

Type	Feature	Definition	Significance	How it was analysed	References
Structural: Other	Frequency of distinct cascade structures	After constructing all of the cascade trees in a dataset, it is possible to compute the frequency of cascade structures. This process is computationally expensive as it aggregates all the generated cascade trees to identify similar structures, e.g. trees with root only, or trees with a root and two child nodes.	It helps to detect if there is a repeated cascade pattern, which can be investigated later. When combined with depth, it could help draw some conclusions about the shape of the cascade and how far it branches.	Distribution of cascade structures.	(Leskovec et al., 2006b) (Leskovec et al., 2006a) (Leskovec et al., 2007b) (Leskovec et al., 2007b) (Goel et al., 2012) (Leskovec et al., 2006) (Chang *et al.*, 2014)
Temporal factors	Time passed since message published	It is the time since a particular message has been published.	Shows the growth of cascade and the fade of interest in the message over time.	Distribution of reshares to hours after upload.	(Dow et al., 2013)
	Speed	Detecting whether and when the first cascade will occur (depth = 1).	Indicates how fast users would be influenced to spread the message or generally react using other means of interaction like reply or mention. Time lag also.	Regression model to quantify the predictability of speed using different features.	(Yang & Counts, 2009)
	Time lag between posting and first reshare, elapsed time.	The difference between posting time and the first reshare.	Measures the resharability of content: the larger the lag the less likely a content will be reshared.	Distribution of time lag.	(Kwak et al., 2010) (Chang et al., 2014)
	Time lag between two sharing events.	The difference between two nodes in a cascade.	Shows the speed at which a cascade occurs in relation to the distance between nodes, i.e. sharing events.	Distribution of elapsed time of two sharing events in relation to the number of hops between them.	(Kwak et al., 2010)

Type	Feature	Definition	Significance	How it was analysed	References
Temporal factors	The number of spikes.	Spikes refer to high-volume of cascading activities that occur in a short period during the lifetime of a cascade.	Measures the degree to which a cascade provokes high volume of cascading during its lifetime.	Distribution of average daily volume of spikes.	(Gruhl et al., 2004) (Cheng et al., 2016)
	Cascading density throughout lifetime.	The timeline of a cascade, it shows the number of cascading activities per day.	Helps assessing the temporal patterns of diffusion, whether it has spikes or maintains a steady growth.	Plots of density of diffusion throughout time.	(Gruhl et al., 2004)
	Maximum time between reshares.	The maximum time difference between reshares.	Indicates the maximum idleness period within a cascade.		(Cheng et al., 2016)
	Recurrence	Recurrence occurs if a cascade has at least two peaks in addition to other conditions.	Helps identifying cascades that regain their popularity after a period of idleness.	The probability of recurrence. Distribution of days between bursts and the duration of the first burst.	(Cheng et al., 2016)
	Cascade growth/cascade popularity	The relation between the growth in cascade size through time. The rate at which cascades gain their size (i.e. popularity).	Helps to show whether a cascade size grows linearly as time passes or in different ways. This helps detect whether the growth in cascade size occurs in short intervals or whether it grows with time. It also shows the periods of idleness and spikes in the cascade timeline.	Plotting the growth in size for cascades through time.	(Leskovec et al., 2007b) (Leskovec et al., 2006a) (Adamic et al., 2012) (Dow et al., 2013) (Anderson et al., 2015)

Type	Feature	Definition	Significance	How it was analysed	References
Other	Size	How many times a particular message has been passed.	Indicates the overall message's popularity.	Distribution of cascade sizes and cumulative average cascade sizes	(Leskovec et al., 2006a) (Leskovec et al., 2006b) (Leskovec et al., 2007b) (Goel et al., 2012) (Bakshy et al., 2011) (Dow et al., 2013) (Cheng et al., 2014) (Chang et al., 2014) (Taxidou & Fischer, 2014) (Cheng et al., 2016)

Bibliography

Adali, S., Escriva, R., Goldberg, M.K., Hayvanovych, M., Magdon-ismail, M., Szymanski, B.K., Wallace, W.A. & Williams, G. (2010). Measuring Behavioral Trust in Social Networks. In: *Proceedings of the IEEE International Conference on Intelligence and Security Informatics (ISI'10)*. 2010, IEEE Computer Society, pp. 150–152.

Adamic, L.A., Lento, T.M. & Fiore, A.T. (2012). How You Met Me. In: *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2012.

Adamic, L. & Adar, E. (2005). How to search a social network. *Social Networks*. 27 (3). p.pp. 187–203.

Adamic, L., Lento, T., Adar, E. & Ng, P. (2016). Information Evolution in Social Networks. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining (WDSM'16)*. 2016, New York, NY, USA: ACM, pp. 473–482.

Adar, E. & Adamic, L. (2005). Tracking Information Epidemics in Blogspace. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI '05)*. 2005, IEEE Computer Society, pp. 207–214.

Adedoyin-Olowe, M., Gaber, M.M. & Stahl, F. (2014). A Survey of Data Mining Techniques for Social Media Analysis. *Journal of Data Mining & Digital Humanities*.

Agarwal, S. & Sureka, A. (2016). Spider and the Flies : Focused Crawling on Tumblr to Detect Hate Promoting Communities. *arXiv preprint arXiv:1603.09164*.

Agrawal, D., Budak, C. & Abbadi, A. El (2011). Information Diffusion In Social Networks : Observing and Influencing Societal Interests. In: *Proceedings of International Conference on Very Large Data Bases (VLDB'11)*. 2011, pp. 1512–1513.

Ahn, Y.Y., Han, S., Kwak, H., Moon, S. & Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In: *Proceedings of the 16th international conference on World Wide Web (WWW'07)*. 2007, New York, New York, USA: ACM, pp. 835–844.

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N. & Acquisti, A. (2013). Tweets are forever: a large-scale quantitative analysis of deleted tweets. In: *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*. 2013, ACM, pp. 897–908.

Anagnostopoulos, A., Kumar, R. & Mahdian, M. (2008). Influence and Correlation in Social Networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. 2008, Las Vegas, Nevada, USA: ACM, pp. 7–15.

Anderson, A., Huttenlocher, D., Kleinberg, J., Leskovec, J. & Tiwari, M. (2015). Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily. In: *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. 2015, ACM, pp. 66–76.

André, P., Bernstein, M.S. & Luther, K. (2012). Who Gives A Tweet? Evaluating Microblog Content Value. In: *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. 2012, Seattle, Washington, USA: ACM, pp. 471–474.

Anger, I. & Kittl, C. (2011). Measuring influence on Twitter. In: *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW '11)*. 2011, Graz, Austria: ACM.

Bibliography

Anon (2014). *World Wide Web Timeline*. [Online]. 2014. Pew Research Center. Available from: http://www.pewinternet.org/2014/03/11/world-wide-web-timeline. [Accessed: 5 December 2016].

Antoniades, D. & Dovrolis, C. (2015). Co-evolutionary dynamics in social networks: A case study of Twitter. *Computational Social Networks*. 2 (14).

Aral, S., Muchnik, L. & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences of the United States of America*. 106 (51). p.pp. 21544–21549.

Bakshy, E., Hofman, J., Mason, W.A. & Watts, D.J. (2011). Everyone's an Influencer: Quantifying Influence on Twitter. In: *Proceedings of the fourth ACM international conference on Web search and data mining (WSDM '11)*. 2011, ACM, pp. 65–74.

Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. (2012). The Role of Social Networks in Information Diffusion. In: *Proceedings of the 21st international conference on World Wide Web (WWW '12)*. 2012, Lyon, France: ACM, pp. 519–528.

Barabási, A.-L. & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*. 286 (5439). p.pp. 509–512.

Bass, F.M. (1969). A New Product Growth for Model Consumer Durables. *Management Science*. 15 (2). p.pp. 215–227.

Berners-Lee, T. (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. HarperInformation.

Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N. & Weitzner, D.J. (2006). A Framework for Web Science. *Foundations and Trends in Web Science*. 1 (1). p.pp. 1–130.

Bhattacharya, D. & Ram, S. (2012). Sharing news articles using 140 characters: A diffusion analysis on twitter. In: *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, (ASONAM '12)*. 2012, IEEE Computer Society, pp. 966–971.

Bik, H.M. & Goldstein, M.C. (2013). An introduction to social media for scientists. *PLoS biology*. 11 (4).

Bikhchandani, S., Hirshleifer, D. & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*. 5. p.pp. 992–1026.

Bild, D.R., Liu, Y., Dick, R.P., Mao, Z.M. & Wallach, D.S. (2015). Aggregate Characterization of User Behavior in Twitter and Analysis of the Retweet Graph. *ACM Transactions on Internet Technology*. 15 (1). p.p. 24.

Bonacich, P. (1987). Power and Centrality : A Family of Measures. *American Journal of Sociology*. 92 (5). p.pp. 1170–1182.

Borgatti, S.P. & Everett, M.G. (2006). A Graph-theoretic perspective on centrality. *Social Networks*. 28 (4). p.pp. 466–484.

Borghol, Y., Mitra, S., Ardon, S., Carlsson, N., Eager, D. & Mahanti, A. (2011). Characterizing and modelling popularity of user-generated videos. *Performance Evaluation*. 68 (11). p.pp. 1037–1055.

Bourlai, E. & Herring, S.C. (2014). Multimodal Communication on Tumblr: ' I Have So Many Feels !' In: *Proceedings of the 2014 ACM Conference on Web Science (WebSci'14)*. 2014, ACM Press,

pp. 171–175.

boyd, danah & Crawford, K. (2011). Six Provocations for Big Data. In: *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*. 2011.

boyd, danah, Golder, S. & Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: *2010 43rd Hawaii International Conference on System Sciences (HICSS)*. January 2010, IEEE Computer Society, pp. 1–10.

boyd, danah m. & Ellison, N.B. (2007). Social Network Sites: Definition, History, and Scholarship. *Journal of Computer-Mediated Communication*. 13 (1). p.pp. 210–230.

Bródka, P., Kazienko, P. & Kołoszczyk, B. (2012). Predicting Group Evolution in the Social Network. In: *International Conference on Social Informatics (SocInfo 2012)*. 2012, Springer Berlin Heidelberg, pp. 54–67.

Cebrian, M., Rahwan, I. & Pentland, A. 'Sandy' (2016). Beyond Viral. *Communications of the ACM*. 59 (4). p.pp. 36–39.

Cha, M., Benevenuto, F., Ahn, Y.-Y. & Gummadi, K.P. (2012). Delayed information cascades in Flickr: Measurement, analysis, and modeling. *Computer Networks*. 56 (3). p.pp. 1066–1076.

Cha, M., Haddadi, H., Benevenuto, F. & Gummadi, P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In: *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM'10)*. 2010, AAAI, pp. 10–17.

Cha, M., Mislove, A. & Gummadi, K.P. (2009). A Measurement-driven Analysis of Information Propagation in the Flickr Social Network. In: *Proceedings of the 18th international conference on World Wide Web (WWW'09)*. 2009, ACM, pp. 721–730.

Chang, Y., Tang, L., Inagaki, Y. & Liu, Y. (2014). What is Tumblr: A Statistical Overview and Comparison. *SIGKDD Explorations*. 16 (1). p.pp. 21–29.

Cheng, J. (2016). *Do Cascades Recur?* [Online]. 2016. Available from: https://speakerdeck.com/jcccf/do-cascades-recur.

Cheng, J., Adamic, L.A., Dow, P.A., Kleinberg, J. & Leskovec, J. (2014). Can cascades be predicted? In: *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. 2014, Seoul, Korea: ACM, pp. 925–935.

Cheng, J., Adamic, L.A., Kleinberg, J. & Leskovec, J. (2016). Do Cascades Recur? In: *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. 2016, ACM, pp. 671–681.

Cheong, M. & Lee, V. (2010). Twittering for earth: A study on the impact of microblogging activism on earth hour 2009 in Australia. In: *Intelligent Information and Database Systems (ACIIDS 2010)*. 2010, Springer Berlin Heidelberg, pp. 114–123.

Choobdar, S., Ribeiro, P., Parthasarathy, S. & Silva, F. (2015). Dynamic inference of social roles in information cascades. *Data Mining and Knowledge Discovery*. 29 (5). p.pp. 1152–1177.

Choudhury, M. De, Lin, Y., Sundaram, H., Candan, K.S., Xie, L. & Kelliher, A. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? In: *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2010, AAAI, pp. 34–41.

Chunara, R., Andrews, J.R. & Brownstein, J.S. (2012). Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak. *The American Journal of Tropical Medicine and Hygiene*. 86 (1). p.pp. 39–45.

Coleman, J., Katz, E. & Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry*. 20 (4). p.pp. 253–270.

Crandall, D., Cosley, D., Huttenlocher, D., Kleinberg, J. & Suri, S. (2008). Feedback Effects between Similarity and Social Influence in Online Communities. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. 2008, ACM, pp. 160–168.

Davidsen, J., Ebel, H. & Bornholdt, S. (2002). Emergence of a Small World from Local Interactions: Modeling Acquaintance Networks. *Physical Review Letters*. 88 (12).

DeSouza, M.E. (2013). A Case of the Red Pants Mondays: The Connection Between Fandom, Tumblr, and Consumption. *Major Papers by Master of Science Students*. [Online]. Available from: http://digitalcommons.uri.edu/tmd_major_papers/3/.

Dhar, V. (2013). Data Science and Prediction. *Communications of the ACM*. 56 (12). p.pp. 64–73.

Dodds, P.S., Muhamad, R. & Watts, D.J. (2003). An Experimental Study of Search in Global Social Networks. *Science*. 301 (5634). p.pp. 827–829.

Dow, P., Adamic, L. & Friggeri, A. (2013). The Anatomy of Large Facebook Cascades. In: *Proceedings of the Seventh International Conference on Weblogs and Social Media, (ICWSM)*. 2013, Cambridge, Massachusetts, USA: AAAI, pp. 145–154.

Easley, D. & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

ENGE, E. (2014). *Twitter Engagement Unmasked: A Study of More than 4M Tweets*. [Online]. 2014. stonetemple. Available from: https://www.stonetemple.com/twitter-engagement-umasked/. [Accessed: 7 April 2017].

Farajtabar, M., Gomez-Rodriguez, M., Wang, Y., Li, S., Zha, H. & Song, L. (2015). Co-evolutionary Dynamics of Information Diffusion and Network Structure. In: *Proceedings of the 24th International Conference on World Wide Web (WWW'15)*. 2015, ACM, pp. 619–620.

Fraustino, J.D., Liu, B. & Jin, Y. (2012). *Social Media Use during Disasters: A Review of the Knowledge Base and Gaps*.

Freeman, L.C. (1979). Centrality in Social Networks Conceptual Clarification. *Social Networks*. 1 (3). p.pp. 215–239.

Freeman, L.C. (2011). The Development of Social Network Analysis – with an Emphasis on Recent Events. *The sage handbook of social network analysis*. p.pp. 29–39.

Friggeri, A., Adamic, L., Eckles, D. & Cheng, J. (2014). Rumor Cascades. In: *International AAAI Conference on Web and Social Media Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2014, AAAI, pp. 101–110.

Galuba, W. & Aberer, K. (2010). Outtweeting the Twitterers - Predicting Information Cascades in Microblogs. In: *Proceedings of the 3rd Wonference on Online Social Networks (WOSN'10)*. 2010, Boston, MA: USENIX Association, pp. 1–9.

Gander, K. (2014). *Girl, 7, gets Tesco to remove 'stupid' sign suggesting superheroes are 'for boys'*. [Online]. 2014. The Independent. Available from: http://www.independent.co.uk/life-style/health-and-families/girl-7-gets-tesco-to-remove-stupid-sign-suggesting-superheroes-are-for-boys-9882725.html. [Accessed: 1 November 2016].

Ghosh, R. & Lerman, K. (2012). Rethinking Centrality: The Role of Dynamical Processes in Social

Network Analysis. *arXiv preprint arXiv:1209.4616*.

Girvan, M. & Newman, M.E.J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences (PNAS)*. 99 (12). p.pp. 7821–7826.

Gladwell, M. (2000). *The Tipping Point: How Little Things Can Make a Big Difference*. Boston: Little, Brown.

Goel, A., Munagala, K., Sharma, A. & Zhang, H. (2015a). A Note on Modeling Retweet Cascades on Twitter. In: *International Workshop on Algorithms and Models for the Web-Graph*. 2015, Springer International Publishing, pp. 119–131.

Goel, S., Anderson, A., Hofman, J. & Watts, D.J. (2015b). The structural virality of online diffusion. *Management Science*. 62 (1). p.pp. 180–196.

Goel, S., Watts, D. & Goldstein, D. (2012). The structure of online diffusion networks. In: *Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012)*. 2012, ACM, pp. 623–638.

Goldenberg, J., Libai, B. & Muller, E. (2001a). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*. 12 (3). p.pp. 211–223.

Goldenberg, J., Libai, B. & Muller, E. (2001b). Using Complex Systems Analysis to Advance Marketing Theory Development: Modeling Heterogeneity Effects on New Product Growth through Stochastic Cellular Automata. *Academy of Marketing Science Review*. 2001 (9). p.pp. 1–19.

Gomez Rodriguez, M., Leskovec, J. & Krause, A. (2010). Inferring Networks of Diffusion and Influence. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining - (KDD '10)*. 2010, ACM, pp. 1019–1028.

Gomez Rodriguez, M., Leskovec, J. & Schölkopf, B. (2013). Structure and dynamics of information pathways in online media. In: *Proceedings of the sixth ACM international conference on Web search and data mining (WSDM '13)*. 2013, ACM, p. 23.

Grabner-Kräuter, S. (2009). Web 2.0 Social Networks: The Role of Trust. *Journal of Business Ethics*. 90 (SUPPL. 4). p.pp. 505–522.

Granovetter, M.S. (1973). The Strength of Weak Ties. *American Journal of Sociology*. 78 (6). p.pp. 1360–1380.

Granovetter, M.S. (1978). Threshold Models of Collective Behavior. *American Journal of Sociology*. 83 (6). p.pp. 1420–1443.

Gruhl, D., Guha, R., Liben-Nowell, D. & Tomkins, A. (2004). Information Diffusion Through Blogspace. In: *Proceedings of the 13th international conference on World Wide Web (WWW '04)*. 2004, ACM, pp. 491–501.

Guille, A. & Hacid, H. (2012). A predictive model for the temporal dynamics of information diffusion in online social networks. In: *Proceedings of the 21st International Conference Companion on World Wide Web (WWW '12 Companion)*. 2012, ACM, pp. 1145–1152.

Guille, A., Hacid, H., Favre, C. & Zighed, D. (2013). Information Diffusion in Online Social Networks: A Survey. *SIGMOD Record*. 42 (2). p.pp. 17–28.

Halford, S., Pope, C. & Carr, L. (2010). A Manifesto for Web Science? *Proceedings of the WebSci10: Extending the Frontiers of Society*.

Hayashi, C. (1998a). What is Data Science?: Fundamental Concepts and a Heuristic Example. *Data Science, Classification, and Related Methods*. p.pp. 40–51.

Hayashi, C. (1998b). What is Data Science?: Fundamental Concepts and a Heuristic Example. *Data Analysis, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27-30, 1996*. p.pp. 40–51.

Heidemann, J., Klier, M. & Probst, F. (2012). Online social networks: A survey of a global phenomenon. *Computer Networks*. 56 (18). p.pp. 3866–3878.

Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T. & Weitzner, D. (2008). Web Science: An Interdisciplinary Approach to Understanding the Web. *Communications of the ACM*. 51 (7). p.pp. 60–69.

Herring, S.C.S. (1996). *Computer-mediated communication: linguistic, social, and cross-cultural perspectives*. John Benjamins Publishing.

Hillman, S., Procyk, J. & Neustaedter, C. (2014). Tumblr fandoms, community & culture. In: *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW Companion '14)*. 2014, pp. 285–288.

Hinz, O., Skiera, B., Barrot, C. & Becker, J. (2011). Seeding Strategies for Viral Marketing: An Empirical Comparison. *Journal of Marketing*. 75 (6). p.pp. 55–71.

Huffaker, D., Teng, C., Simmons, M.P., Gong, L. & Adamic, L.A. (2011). Group Membership and Diffusion in Virtual Worlds. In: *IEEE 3rd international conference on social computing (socialcom)*. 2011, IEEE Computer Society, pp. 331–338.

Hughes, A.L. & Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*. 6 (May). p.p. 248.

Jackson, M.O. (2010). An Overview of Social Networks and Economic Applications. In: J. Benhabib, A. Bisin, & M. O. Jackson (eds.). *The Handbook of Social Economics*. North-Holland, pp. 511–585.

Jackson, M.O. & López-Pintado, D. (2013). Diffusion and contagion in networks with heterogeneous agents and homophily. *Network Science*. 1 (1). p.pp. 49–67.

Java, A., Song, X., Finin, T. & Tseng, B. (2007). Why We Twitter: Understanding Microblogging Usage and Communities. In: *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007, ACM, pp. 56–65.

Kaplan, A.M. & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*. 53 (1). p.pp. 59–68.

Karp, D. (2014). *"Ninety percent of content on Tumblr is actually...: Unwrapping Tumblr*. [Online]. 2014. Available from: https://unwrapping.tumblr.com/post/99263588212/90-percent-content-reblogged. [Accessed: 14 October 2016].

Keeling, M.J. (1999). The effects of local spatial structure on epidemiological invasions. *Proceedings. Biological sciences / The Royal Society*. 266 (1421). p.pp. 859–67.

Kelman, H.C. (1958). Compliance, identification, and internalization: Three processes of attitude change. *Journal of Conflict Resolution*. 2 (1). p.pp. 51–60.

Kempe, D., Kleinberg, J. & Tardos, E. (2003). Maximizing the Spread of Influence through a Social Network. In: *Proceedings of the ninth ACM international conference on Knowledge discovery*

and data mining (KDD'03). 2003, ACM, pp. 137–146.

Khosla, A., Das Sarma, A. & Hamid, R. (2014). What Makes an Image Popular ? In: *Proceedings of the 23rd International Conference on World Wide Web (WWW' 14)*. 2014, ACM, pp. 867--876.

Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02)*. 2002, ACM, p. 91.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). What is Twitter, a social network or a news media? In: *Proceedings of the 19th international conference on World wide web (WWW '10)*. 2010, ACM, pp. 591–600.

Lai, L.S.L. & Turban, E. (2008). Groups formation and operations in the web 2.0 environment and social networks. *Group Decision and Negotiation*. 17 (5). p.pp. 387–402.

Lee, C., Kwak, H., Park, H. & Moon, S. (2010). Finding Influentials Based on the Temporal Order of Information Adoption in Twitter. *Proceedings of the 19th international conference on World wide web (WWW '10)*. 9 (1). p.pp. 1137–1138.

Lerman, K. & Ghosh, R. (2010). Information Contagion: An Empirical Study of the Spread of News on Digg and Twitter Social Networks. In: *Fourth International AAAI Conference on Weblogs and Social Media*. 2010, AAAI, pp. 90–97.

Lerman, K. & Rey, M. (2007). Social Information Processing in Social News Aggregation Anatomy of Digg. *Information Sciences*. (2). p.pp. 1–17.

Leskovec, J., Adamic, L.A. & Huberman, B.A. (2007a). The Dynamics of Viral Marketing. *ACM Transactions on the Web*. 1 (1). p.pp. 1–39.

Leskovec, J., Backstrom, L. & Kleinberg, J. (2009). Meme-tracking and the Dynamics of the News Cycle. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '09)*. 2009, ACM, pp. 497–506.

Leskovec, J., Mcglohon, M., Faloutsos, C., Glance, N. & Hurst, M. (2006a). Cascading Behavior in Large Blog Graphs Patterns and a model. *Science*. (October).

Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. & Hurst, M. (2007b). Cascading Behavior in Large Blog Graphs. In: *7th SIAM International Conference on Data Mining (SDM)*. 2007, pp. 1–21.

Leskovec, J., Singh, A. & Kleinberg, J. (2006b). Patterns of influence in a recommendation network. In: *Proceedings of the 10th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining (PAKDD'06)*. 2006, Singapore: Springer-Verlag, pp. 380–389.

Liben-Nowell, D. & Kleinberg, J. (2008). Tracing information flow on a global scale using Internet chain-letter data. *Proceedings of the National Academy of Sciences*. 105 (12). p.pp. 4633–4638.

Lin, Y.-R., Margolin, D., Keegan, B., Baronchelli, A. & Lazer, D. (2013). #Bigbirds Never Die: Understanding Social Dynamics of Emergent Hashtags. In: *7th International AAAI Conference on Weblogs and Social Media (ICWSM)*. 2013, pp. 370–379.

Liu, Y., Kliman-Silver, C. & Mislove, A. (2014). The Tweets They are a-Changin': Evolution of Twitter Users and Behavior. In: *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*. 2014, AAAI, pp. 305–314.

Lotan, G. (2011). Mapping Information Flows on Twitter. In: *International AAAI Conference on Web and Social Media Fifth International AAAI Conference on Weblogs and Social Media (ICWSM'11)*. 2011, AAAI, pp. 23–27.

Ma, Z., Sun, A. & Cong, G. (2013). On predicting the popularity of newly emerging hashtags in Twitter. *Journal of the American Society for Information Science and Technology*. 64 (7). p.pp. 1399–1410.

Macskassy, S. & Michelson, M. (2011). Why do people retweet? anti-homophily wins the day! In: *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM'11)*. 2011, AAAI, pp. 209–216.

Marres, N. & Weltevrede, E. (2013). Scraping the Social? *Journal of Cultural Economy*. 6 (3). p.pp. 313–335.

McBride, K. (2015). *Journalism and public shaming: Some guidelines*. [Online]. 2015. Poynter. Available from: http://www.poynter.org/2015/journalism-and-public-shaming-some-guidelines/326097/. [Accessed: 1 November 2016].

McPherson, M., Smith-Lovin, L. & Cook, J. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*. 27 (1). p.pp. 415–444.

Meier, F., Elsweiler, D. & Wilson, M.L. (2014). More than Liking and Bookmarking ? Towards Understanding Twitter Favouriting Behaviour. In: *Proceeding of the 8th International AAAI Conference on Weblogs and Social Media*. 2014, AAAI, pp. 346–355.

Mislove, A., Marcon, M., Gummadi, K.P., Druschel, P. & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In: *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*. 2007, ACM, p. 29.

Mittal, S., Gupta, N., Dewan, P. & Kumaraguru, P. (2013). The Pin-Bang Theory : Discovering The Pinterest World. *arXiv preprint arXiv:1307.4952*.

Mochalova, A. & Nanopoulos, A. (2013). On The Role Of Centrality In Information Diffusion In Social Networks. In: *Proceedings of the 21st European Conference on Information Systems (ECIS)*. 2013, pp. 1–12.

Monge, P.R. & Contractor, N.S. (2003). *Theories of communication networks*. Oxford University Press.

Morstatter, F. & Ave, S.M. (2014). When is it Biased? Assessing the Representativeness of Twitter's Streaming API. In: *Proceedings of the 23rd International Conference on World Wide Web (WWW'14)*. 2014, ACM, pp. 555–556.

Myers, S. a. & Leskovec, J. (2012). Clash of the Contagions: Cooperation and Competition in Information Diffusion. In: *Proceeding of IEEE 12th International Conference on Data Mining*. December 2012, IEEE, pp. 539–548.

Myers, S. a., Zhu, C. & Leskovec, J. (2012). Information Diffusion and External Influence in Networks. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '12)*. 2012, ACM, pp. 33–41.

Myers, S. & Leskovec, J. (2014). The Bursty Dynamics of the Twitter Information Network. In: *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. 2014, ACM, pp. 913–923.

Newman, M., Barabási, A.-L. & Watts, D.J. (2006). *The Structure and Dynamics of Networks*.

Princeton University Press.

Newman, M.E.J. (2004). Fast algorithm for detecting community structure in networks. *Physical Review E*. 69 (6). p.p. 66133.

Newman, M.E.J. (2010). *Networks: an introduction*. Oxford University Press.

O'Reilly, T. (2005). *What Is Web 2.0*. [Online]. 2005. O'Reilly Media, Inc. Available from: http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html. [Accessed: 5 December 2016].

OnlineMBA (2012). *A Case Study in Social Media Demographics*. [Online]. 2012. Available from: http://www.onlinemba.com/blog/social-media-demographics/. [Accessed: 4 December 2016].

Otte, E. & Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information. *Journal of Information Science*. 28 (6). p.pp. 441–453.

Palshikar, G.K. (2009). Simple Algorithms for Peak Detection in Time-Series. In: *Proceedings 1st Int. Conf. Advanced Data Analysis, Business Analytics and Intelligence*. 2009.

Petrovic, S., Osborne, M. & Lavrenko, V. (2013). I Wish I Didn't Say That! Analyzing and Predicting Deleted Messages in Twitter. *arXiv preprint arXiv:1305.3107*. 1.

Petrovic, S., Osborne, M. & Lavrenko, V. (2011). RT to Win! Predicting Message Propagation in Twitter. In: *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM)*. 2011, AAAI, pp. 586–589.

Phethean, C., Simperl, E., Tiropanis, T., Tinati, R. & Hall, W. (2016). The Role of Data Science in Web Science. *IEEE Intelligent Systems*. 31 (3). p.pp. 102–107.

Pilkington, E. (2013). *Justine Sacco, PR executive fired over racist tweet, 'ashamed'*. [Online]. 2013. Available from: https://www.theguardian.com/world/2013/dec/22/pr-exec-fired-racist-tweet-aids-africa-apology. [Accessed: 1 November 2016].

Renwick, L. (2014). Audience research project: Tumblr study group research "How do 'Fandoms' on Tumblr react to new media content? *Enquiry-The ACES Journal of Undergraduate Research*. 4. p.pp. 1–24.

Rogers, E.M. (2003). *Diffusion of Innovations*. 5th Ed. Simon and Schuster.

Romero, D.M., Meeder, B. & Kleinberg, J. (2011). 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 (WWW '11)*. 2011, ACM, pp. 695–704.

Rotabi, R., Kamath, K., Kleinberg, J. & Sharma, A. (2017). Cascades: A view from Audience. In: *Proceedings of the 26th International Conference on World Wide Web (WWW'17)*. 2017, ACM, pp. 587–596.

Ryan, B. & Gross, N. (1943). The diffusion of hybrid seed corn in two Iowa communities. *Rural Sociology*. 8 (1). p.pp. 15–24.

Saez-Trumper, D., Comarela, G., Almeida, V., Baeza-Yates, R. & Benevenuto, F. (2012). Finding trendsetters in information networks. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '12*. 2012, ACM, p. 1014.

Schneider, F., Feldmann, A., Krishnamurthy, B. & Willinger, W. (2009). Understanding online social

network usage from a network perspective. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference IMC 09*. 2009, ACM, p. 35.

Schneider, R. (2011). *Survey of Peaks / Valleys identification in Time Series*.

Scott, J. (2008). Network analysis. In: W. A. Darity (ed.). *International Encyclopaedia of the Social Sciences*. Macmillan Reference USA.

Shalizi, C.R. & Thomas, A.C. (2011). Homophily and Contagion Are Generically Confounded in Observational Social Network Studies. *Sociological Methods & Research*. 40 (2). p.pp. 211–239.

Shi, X., Tseng, B. & Adamic, L. a (2007). Looking at the blogosphere topology through different lenses. In: *Proceedings of the International Conference on Weblogs and Social Media (ICWSM' 07)*. 2007, AAAI.

Simmons, M., Adamic, L. & Adar, E. (2011). Memes Online: Extracted, Subtracted, Injected, and Recollected. In: *Proceedings of 5th International Conference on Weblogs and Social Media (ICWSM'11)*. 2011, AAAI, pp. 353–360.

Strutner, S. (2016). *Here's How Airlines Really Handle Sexual Assault*. [Online]. 2016. Huffington Post. Available from: http://www.huffingtonpost.com/entry/sexual-assault-on-planes_us_5808f758e4b000d0b15552f8. [Accessed: 1 November 2016].

Subbian, K. (2014). *Scalable Analysis of Information Flows in Networks*. University Of Minnesota.

Suh, B., Hong, L., Pirolli, P. & Chi, E.H. (2010). Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. In: *Proceedings of IEEE Second International Conference on Social Computing*. 2010, IEEE, pp. 177–184.

Taxidou, I. & Fischer, P. (2013). Realtime analysis of information diffusion in social media. *Proceedings of the VLDB Endowment*. 6 (12). p.pp. 1416–1421.

Taxidou, I. & Fischer, P.M. (2014). Online Analysis of Information Diffusion in Twitter. In: *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion (WWW Companion '14)*. 2014, ACM, pp. 1313–1318.

Tinati, R., Halford, S., Carr, L. & Pope, C. (2014). Big Data: Methodological Challenges and Approaches for Sociological Analysis. *Sociology*. 48 (4). p.pp. 663–681.

Tiropanis, T., Hall, W., Crowcroft, J., Contractor, N. & Tassiulas, L. (2015). Network Science, Web Science, and Internet Science. *Communications of the ACM*. 58 (8). p.pp. 76–82.

Travers, J. & Milgram, S. (1969). An Experimental Study of the Small World Problem. *Sociometry*. 32 (4). p.pp. 425–443.

Ugander, J., Karrer, B., Backstrom, L., Marlow, C. & Alto, P. (2011). The Anatomy of the Facebook Social Graph. *arXiv preprint arXiv:1111.4503*.

Wang, G., Gill, K. & Mohanlal, M. (2013). Wisdom in the Social Crowd: an Analysis of Quora. In: *Proceedings of the 22nd international conference on World Wide Web (WWW '13)*. 2013, ACM, pp. 1341–1351.

Watts, D.J. (2002). A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*. 99 (9). p.pp. 5766–5771.

Watts, D.J. (2004). The 'New' Science of Networks. *Annual Review of Sociology*. 30 (1). p.pp. 243–270.

Watts, D.J. & Strogatz, S.H. (1998). Collective dynamics of 'small-world' networks. *Nature*. 393 (6684). p.pp. 440–442.

Webberley, W., Allen, S.M. & Whitaker, R.M. (2013). Inferring the Interesting Tweets in Your Network. In: *Proceeding of IEEE 3rd International Conference on Cloud and Green Computing*. September 2013, IEEE, pp. 575–580.

Weng, L. (2014). *Information Diffusion On Online Social Networks*. Indiana University.

Weng, L., Flammini, A., Vespignani, A. & Menczer, F. (2012). Competition among memes in a world with limited attention. *Scientific Reports*. 2. p.pp. 1–9.

Wright, A. (2011). Web Science Meets Network Science. *Communications of the ACM*. 54 (5). p.p. 23.

Xu, J., Compton, R., Lu, T.-C. & Allen, D. (2014). Rolling through Tumblr : Characterizing Behavioral Patterns of the Microblogging Platform. In: *Proceedings of the 2014 ACM Conference on Web Science (WebSci '14)*. 2014, ACM, pp. 13–22.

Yang, J. & Counts, S. (2010). Predicting the Speed, Scale, and Range of Information Diffusion in Twitter. In: *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM '10)*. 2010, AAAI, pp. 355–358.

Yang, L., Sun, T., Zhang, M. & Mei, Q. (2012). We know what@ you# tag: does the dual role affect hashtag adoption? In: *Proceedings of the 21st international conference on World Wide Web (WWW '12)*. 2012, ACM, pp. 261–270.

Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L. & Su, Z. (2010). Understanding Retweeting Behaviors in Social Networks. *Idea*. [Online]. p.pp. 1633–1636. Available from: http://keg.cs.tsinghua.edu.cn/persons/tj/publications/CIKM10-Yang-et-al-Understanding-Retweeting.pdf.

Ye, S. & Wu, S. (2010). Measuring Message Propagation and Social Influence on Twitter.com. In: *Proceedings of the Second international conference on Social informatics (SocInfo'10)*. Lecture Notes in Computer Science. 2010, Springer Berlin Heidelberg, pp. 216–231.

Zafarani, R., Abbasi, M.A. & Liu, H. (2014). *Social Media Mining: An Introduction*. Cambridge University Press.

Zhou, L., Wang, W. & Chen, K. (2016). Tweet Properly: Analyzing Deleted Tweets to Understand and Identify Regrettable Ones. In: *Proceedings of the 25th International Conference on World Wide Web (WWW'16)*. 2016, ACM, pp. 603–612.