University of Southampton

Faculty of Social Sciences
Social Statistics and Demography

Topics of statistical analysis with social media data

Martina Patone

Thesis for the degree of Doctor of Philosophy
January, 2020

Abstract

This thesis investigates the use of social media data in social research from a statistical perspective. A broad review is given of how these data has been used by researchers from different disciplines and the extent and means of the investigations carried out with these data is assessed. Special attention has been given to the common obstacles faced by using social media data for statistical analysis and to the graph representation of these data that is generally available to the researcher and to its use for statistical inference.

Most of the literature about the use of social media data for statistical analysis is concerned with the fact that these data represent a non-random sample from the population of interest. We have instead highlighted another fundamental challenge presented by these data, which is, however, rarely taken explicitly into consideration. The problem is that the object of sampling and the unit of interest might be distinct. To tackle this problem, we have shown how two different approaches of statistical inference can be distinguished in the literature. Under each approach, we have provided a discussion about the target of inference and make explicit their limitations in relation with the statistical methods used. Our exposition offers a framework for dealing with unruly data sources.

However, the problems of non-random sample and various unavoidable non-sampling errors do not admit a universally valid statistical approach. One can cope with them if needed to, but one cannot really hope to solve these problems. Meanwhile the graph structure inherent of social media data (and other forms of big data) seems to us a more rewarding area of research.

We have investigated how to use the structure of the graph for estimation. The Horvitz-Thompson (HT) estimator operates by weighting each sample motif by the inverse of its inclusion probability. Generalising the work of Birnbaum and Sirken (1965), we demonstrated that infinite types of *incidence weights* can be constructed for unbiased estimation. We define the Incidence Weighting Estimator (IWE) as a large class of linear design-based unbiased estimators based on the edges of the a Bipartite Incidence Graph (BIG), of which the HT estimator is a special case. This class of estimator has no equivalence in traditional list sampling.

More ways of using the incidence structure of the BIG for estimation has been explored and in doing so we enter in a completely new territory. We have investigated how to use the incidence structure of the BIG to estimate a total based on the sampling units, and, once we have obtained such estimator we have discussed if and how it can be used together with the IWE to improve the inference. We have also seen that it is possible to use the reverse incidence weights in combination with the incidence weights. The weights obtained in such ways, can be used to construct an unbiased estimator in both directions, although the idea seems somewhat impractical at the moment. The final chapter wants to offer a flavour of what can be done under the BIG framework and inspire future research in this direction.

The thesis is organised in four papers: the first paper discusses the current statistical analysis made using social media data, while the other three papers deal with the topic of graph sampling and estimation.

Key words: social media data, finite-graph sampling, quality, non probability samples, network.

A nonna Lena.

Contents

List of papers **xviii**

Authors' contribution to the publications **xx**

1 Introduction **1**
1.1 Big data and Social media data 2
1.1.1 Definitions and characteristics of big data 2
1.1.2 Social Media Data 6
1.2 Use of social media data for social research 8
1.2.1 Quantitative social research 9
1.2.2 The statistical characteristics of social media data 9
1.3 Types of inference with big data 19
1.3.1 Challenges with the model-based approaches of inference . 21
1.4 The network structure 22
1.4.1 Structure of social media 23
1.4.2 Use of the graph structure for design-based inference . . . 26
1.5 A structure of the thesis 27

2 On two existing approaches to statistical analysis of social media data **30**
2.1 Introduction 31
2.2 General issues of representation and measurement 33
2.2.1 Representation 33
2.2.2 Measurement 36
2.3 One-phase approach 38
2.3.1 Case: Social Media Index (SMI) 39
2.3.2 Formal interpretation 40
2.3.3 Statistical validation 41
2.3.4 Discussion 44

2.4 Two-phase approach . . . 46
2.4.1 Case: Residence location from tweets . . . 47
2.4.2 Quality assessment . . . 48
2.4.3 Discussion: Statistical analysis . . . 51
2.5 Concluding remarks . . . 52

3 Graph sampling **55**
3.1 Introduction . . . 55
3.2 Sampling on a graph . . . 57
3.2.1 Terms and notations . . . 57
3.2.2 Definition of sample graph . . . 58
3.2.3 Some graph sampling methods . . . 59
3.3 Graph parameter and HT-estimation . . . 61
3.3.1 Graph totals of a given order . . . 63
3.3.2 Graph totals of unspecified order . . . 66
3.3.3 HT-estimation . . . 67
3.4 T-stage snowball sampling . . . 69
3.4.1 Inclusion probabilities of nodes and edges in G_s . . . 70
3.4.2 Arbitrary M_k with $k \geq 2$ and $s_2 = s_1 \times U \cup U \times s_1$ . . . 72
3.4.3 Arbitrary M_k with $k \geq 2$ and $s_2^* = s_1 \times s_1$ . . . 74
3.4.4 Proportional representative sampling in graphs . . . 75
3.5 Network sampling methods . . . 78
3.5.1 Sampling patients via hospitals . . . 80
3.5.2 Sampling children via parents . . . 81
3.5.3 Sampling siblings via households . . . 82
3.5.4 Adaptive cluster sampling of rare species . . . 83
3.6 Concluding remarks . . . 84

4 Incidence weighting estimation under sampling from a bipartite incidence graph **86**
4.1 Introduction . . . 87
4.2 Basics of BIG sampling and estimation . . . 88
4.2.1 BIG sampling . . . 89
4.2.2 Three existing estimators under BIG sampling . . . 91
4.2.3 More on Phat $\hat{\theta}_p$ . . . 93
4.3 Incidence weighting estimator . . . 96
4.3.1 Definition . . . 96

4.3.2 Theory 97
4.4 Unbiased IWE 99
4.4.1 Zhat 100
4.4.2 HT weights 101
4.4.3 Priority weights 102
4.4.4 Discussion on the efficiency of the different unbiased IWE 103
4.5 Simulations 104
4.6 Concluding remarks 107

5 Reverse incidence weighting under BIG sampling 115
5.1 Introduction 115
5.2 The reverse incidence weighting estimator 117
5.2.1 Examples of reverse incidence estimators 120
5.2.2 Simulations 122
5.2.3 A discussion 125
5.3 Ratio-type estimators on a BIG 127
5.3.1 Simulation study 130
5.3.2 A particular class of ratio-type and Hajek-type estimators 133
5.4 Conclusions 134

Conclusions 136

References 138

List of Figures

1.1 The social media data pipeline (Halford et al., 2017). 13
1.2 Two examples of Twitter networks from the blog 'Digital Humanities Specialist' 23
1.3 Conceptual model of Twitter activities. 24

2.1 Comparison of Dutch CCI and SMI on a monthly basis. A correlation coefficient of 0.88 is found for the two series (Daas et al., 2015). 39
2.2 The CCI series with 95% confidence interval, 2000-2014. 43
2.3 P-values of test H_0 vs. H_1 for varying CVs, level 0.05 mark by horizontal line 44
2.4 Phase-two life-cycle model of Zhang (2012) 50

3.1 Population graph (top) and four sample graphs (i) - (iv) based on $s_1 = \{3, 6, 10\}$. 62
3.2 Population graph G with 10 nodes and 11 edges (left), a sample graph G_s by 2-stage snowball sampling starting from $s_{1,0} = \{3, 4\}$ by simple random sampling (right). 72
3.3 Inclusion probability $\pi_{(i)}$: true vs. (3.3), left; $\pi_{(ij)}$: true vs. (3.4), right. 72
3.4 Population graph G with 13 nodes and 19 edges (top); sample graphs G_s (bottom left) and G_s^* (bottom right) by 2-stage snowball sampling with initial $s_{1,0} = \{4, 5, 10\}$. 76

4.1 Top, population bipartite incidence graph $G = (F, U; A)$. Sample graph G_s given $s = \{1, 2\}$: bottom-left, by incident reciprocal observation; bottom-right, by incident ancestral observation, with additional information marked by dotted edges. 90
4.2 The two observed degree distributions for the units set F in G_1 and G_2. 105

4.3 The average of the estimates with associated Monte Carlo error for the IWE plotted against the increasing sample sizes for G_1 and G_2 considering the three ordering of the frame. Three types of multiplicity weighting are used: Equal-Share, Inverse-Degree and Power of Inverse-Degree weighting 108
4.4 The variances estimator and the true variances for the IWE with fixed weights plotted against the increasing sample sizes for both graph G_1 and G_2. Three types of multiplicity weighting are used: Equal-Share, Inverse-Degree and Power of Inverse-Degree weights. 109
4.5 The average of the estimates with associated Monte Carlo error for the IWE plotted against the increasing sample sizes for G_1 and G_2 considering the priority weighting. Three ordering of the frame are considered. 110
4.6 The true variances and its two estimators with associated Monte Carlo error for the priority IWE plotted against the increasing sample sizes considering different ordering of the frame for grap G_1. 111
4.7 The true variances and its two estimators with associated Monte Carlo error for the priority IWE plotted against the increasing sample sizes considering different ordering of the frame for grap G_2. 112
4.8 The average of the estimates with associated Monte Carlo error for the IWE corresponding to the Yhat, plotted against the increasing sample sizes for G_1 and G_2 113
4.9 The variances of the six IWE plotted against the increasing sample sizes for both graph G_1 and G_2. 114

5.1 The degree distributions for the sampling units F and the motifs set U for the three BIG. 122
5.2 The sampling distribution of the estimators of M, under SRS of size $m = 2$ for the three BIGs. 124
5.3 The sampling distribution of the three IWE $Y = 308.54$, under SRS of size $m = 2$ for the BIG G_3. 127
5.4 Comparison of the accuracy of the 10 estimators of N for BIG 1. . 131
5.5 Comparison of the accuracy of the 10 estimators of N for BIG 2. . 132
5.6 Comparison of the accuracy of the 10 estimators of N for BIG 3. . 132

List of Tables

2.1 Many-one relations a from post to account, and b from account to user 34

3.1 Inclusion probability $\pi_{(M_3)}$ of selected $M_3 = \{i_1, i_2, i_3\}$ 77
3.2 Third-order graph total estimate, expectation and standard error 77

4.1 Probability $p_{(ki)}$ for population BIG in Figure 4.1. 95
4.2 Weights, measures, variances for Fig. 4.1 using three choices of multiplicity weighting: ES = equal-share; ID = inverse-degree; ID2 = power of inverse-degree weights with $\alpha = 2$ and ID3 with $\alpha = 3$. 101
4.3 HT incidence weights and corresponding measures for the BIG in Fig. 4.1. 102
4.4 Priority weights and measures for the graph in Fig. 4.1 under the unorder and descending ordering of the sampling frame. 103
4.5 The true variances of the IWE for $\theta = N$ by different choices of weights under SRS with $m = 2$ of F in the graph in Figure 4.1. The ordering of the sampling frame for the priority estimator is given by: (I) - random; (II) - descending and (III) - ascending. . . 104

5.1 The true variances of the estimators for M showed in Figure 5.2. . 123
5.2 The sampling distribution of the $\hat{Y}$, $\tilde{M}$ and $\tilde{L}$, using equal-share incidence and reverse incidence weights for the BIG G, with $F = \{1, 2, 3\}$, $U = \{4, 5, 6\}$ and $A = \{(1, 4), (1, 5), (2, 6), (3, 4)\}$, under simple random sampling from F of size 2. 130
5.3 The estimators considered in the simulations in this section. . . . 131

Declaration of authorship

Print name: Martina Patone

Title of thesis: Topics of statistical analysis with social media data

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research. 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:
   (a) Patone, M. and Zhang, L. C. (2019), On two existing approaches to statistical analysis of social media data, `arXiv:1905.00635`;
   (b) Zhang, L. C. and Patone, M. (2017), Graph Sampling, *Metron*, 75:277.

Date: July 14, 2020

Acknowledgments

I could not have written this thesis without the help of many.

First of all, my supervisors. My first supervisor, Li-Chun Zhang, who not only has played a great part in the creation of this work, but also in my life during the past four years. He has left me with indelible moments and valuable lessons and has probably changed deeply me as a researcher and as a person. My second supervisors, Agnese Vitali, who has been my anchor and my safe place, as well as a good friend; and, my third supervisors, Les Carr, whose presence has added a different perspective to my work. I also need to thank my sponsor, the DSTL, and especially, Keith Hermiston. Meeting Keith every three months has been the only certainty during this PhD journey. Although I have hated it sometimes, I admit that it is the thing that has saved me. Undoubtably Keith's kindness and enthusiasm has always been rewarding, but more importantly, no matter how lost I could have been, every three months I knew I needed to find a way to get back on track.

Because it is so easy to get lost during a PhD. It is so easy to lose confidence, to be intimidated and to finally get emotionally drained. A lot of students deal with that. It hurts most of us and it hurts especially those who are most vulnerable. Of course, during a PhD we should develop those skills that will enable us to go on in our discipline. It can be tough, hard to cope with the pressure sometimes and it is a difficult process we need to face. But, even when struggling during this process, we should never feel or let anyone to make us feel that we are 'out of place'. When that happens, our research becomes disconnected and irrelevant and we stop caring about our discipline, because our discipline does not elevate ourselves anymore. To cite Professor Francis Su: our discipline 'should help us flourish as human being, so why should we care about our discipline if it does not connect deeply to some human desire: to play, seek truth, pursue beauty,

fight for justice?' (Prof. Su refers to mathematics, but I believe that could be extended here as well. His speech can be found here `shorturl.at/tFMQ3`).

I thank myself that I still have a little desire to flourish.

I also want to thank my family and my friends. It would have been impossible without their constant support and help. My mom Maria, my father Arcangelo and my brothers Lucio and Jacopo, which have always tried to support me at their best, even when I was not able to show that I needed help. My friends Alessandro, Sara, Allegra, Anna Clara, Anna Chiara, Carla ed Eleonora who have been near me even when not even me could have stand myself. All the friends that have shared this journey here with me: Gabi (I don't know how, but we managed to live together for more than 4 years!), Ingi, Carina, Angela, Mia, Rubi, Shoaib, Kibuchi, Francesco, Viktor, Chloe, Dan, Jamie, Fabio, Giannis, Alex, Gaya, Paolo, Stephanie and all the many others which I cannot mentioned for a limit of space. The people from the Sunday Hop and my squash friends. And the PGR office: thanks Claire for having sorted it out everything. A special thanks to all the people who I have met in these years and have agreed to hear about my joys and sorrows, depending on the time they have met me; in most cases both. Finally, I also want to thank Angela Luna. No one else could have helped me as she did, both professionally and personally.

List of papers

The work of this thesis is based on the following papers:

1. Patone, M. and Zhang, L. C. (2019), On two existing approaches to statistical analysis of social media data, `arXiv:1905.00635`;
2. Zhang, L. C. and Patone, M. (2017), Graph Sampling, *Metron*, 75:277;
3. Patone, M. and Zhang, L. C., Incidence weighting estimation under sampling from a bipartite incidence graph;
4. Patone, M. and Zhang, L. C., Reverse weighting estimation under BIG sampling.

Authors' contribution to the publications

Authors' contribution statements to the paper *Graph Sampling*:

L.-C. Zhang and M. Patone conceived the ideas presented in sec.1-4. L.-C. Zhang developed the theoretical formalism, devised the main conceptual ideas in sec.5 and took the lead in writing the manuscript. M. Patone performed the analytic calculations and performed the numerical simulations. All authors discussed the results and commented on the manuscript.

Chapter 1

Introduction

In recent times, the use of social media data as a source of social science data has considerably increased. This type of data is easily available, cheap and in real time; they can provide information about behaviour and opinion, making possible to observe directly what people 'do' or 'think', rather than what people 'claim to do' or 'claim to think'. Nevertheless, the application of statistical inference to obtain valid insight from them is still under debate. In fact, these data present several challenges and limitations that need to be addressed for their statistical analysis and a coherent statistical framework for analysing social media data is currently lacking.

This thesis investigates the challenging of conducting statistical analysis this particular new form of data. The thesis focuses on three specific aspects involved in the analysis of social media data, namely: 1. the problem of the statistical validity of the conclusions drawn; 2. the problem of the sampling and 3. the problem of estimation. These three problems are investigated separately in four stand-alone papers.

The first chapter in this thesis presents the context of the research and is divided into three sections. The first section provides an overview of big data and social media data, discussing what defines them and how they emerged. The second section examines how social media data has been used in social research; the population that social media data represents and the measures that are observable in social media. Also an examination of the limitations involved with applying statistical methods to this data type is made. In addition, the process of data collection is described, distinguishing three modes: the API streaming,

the data purchasing from social media data aggregation services, such as GNIP and Web scraping. A third section is devoted to the graph structure of the data. In particular, it is seen how the available relationship amongst the elements of social media data, can be used to obtain new sampling methods and improve the efficiency of the estimators. Finally, the four papers are briefly presented.

1.1 Big data and Social media data

Data can be generated by organizations, such as transactions, emails, databases, etc.; by Internet users themselves through their surfing habits, online discussion forum, or sensors and other devices that exchange data.

The term 'big data' describes a significant volume of heterogenous data from different sources and which are often unavailable in standard database formats.

1.1.1 Definitions and characteristics of big data

There is no rigorous definition of big data. The Oxford English Dictionary defines big data as "data of a very large size, typically to the extent that their manipulation and management present significant logistical challenges".

This definition emphasises the scale and complexity of big data and the methodological and structural challenges that they pose. The definition also alludes to the fact that challenges arise because big data are unlike traditional data; they are not only large in size, but their nature is intrinsically different from the data that has been known so far.

One of the most common definition of big data was proposed by the 2001 Gartner report (Laney, 2001) and updated in 2012 in the Gartner IT Glossary, which states "big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation".

Laney's definition of big data considers the difficulties which are often encountered in the process of extracting knowledge from the data: the technological capabilities to store large and unstructured data, how to link different types of data and how to perform comprehensive analysis. According to the definition of Laney (2001), there are three characteristics that distinguishes big data from other data types.

The first of such characteristics is *volume*. Volume refers to the amount of data available. Two main factors contribute to the big volume of data: the increasing number of data collection tools, such as social media, mobile phones, sensors, cameras and scanners, among others; and the improvements in data storing. Facebook, for instance, has 2 billion monthly active users uploading 350 million new photos every day [1]; it is expected that connected cars, i.e. equipped with Internet access, will upload every hour twenty-five gigabytes of data regarding the routes, speeds or road conditions among others[2]; 1.6 million packages are shipped every day by Amazon[3]. The term 'volume' also indicates that big data are not generated as a random sample of a given population, but are often the result of observations of real time occurrences, which sometimes refers to the whole population, sometimes it refers to a non-representative sample of it.

The second characteristics of big data in Laney (2001) definition is *velocity*. Velocity refers to the speed at which the data is generated. Data is streamed at real-time; social media are a classical example, also sensor data are becoming increasingly popular, transmitting bits of data at a constant rate. The flow of data is significant as well as continuous. The velocity of big data also makes them appealing for evidence-based decisions and real-time analytics. Social media data, for instance, facilitate the analysis of marketing campaigns which provide information about customers, such as their location, demographics, and their engagement with the product.

Finally, *variety* refers to the many types and formats of these data: text, images, audio, video etc... These are all examples of types of unstructured data, and they are often all collected simultaneously. As a consequence of this large variety of data types, the process of cleaning the data requires greater effort. New data management technologies and analytics are emerging, such as facial recognition technologies or methods which collect and analyse clickstream data.

Over the years, the definition of big data given by Laney (2001) has evolved to accommodate other characteristics of big data. Japec et al. (2015) includes *Variability*, i.e. the inconsistency of the data over time, *Veracity*, i.e. the trustworthiness of the data, *Complexity*, i.e. the need for a edge with multiple data sources.

Groves (2011a) suggests that big data are characterised by four aspects which distinguish it from traditional data. He proposes that big data:

1. tend to measure behaviours, not internalized states like attitudes or beliefs;
2. tend to offer near-real-time records of phenomena, and they are highly granulated temporally;
3. tend to observe a significant number of variables, many merely having some sort of identifier;
4. rarely offer well-defined coverage of a large population.

Also these characteristics are used to focus on the difficulties that big data present in extracting meaning from them, in comparisons with traditional data.

Taylor (2013) distinguishes between 'found vs made data', where 'found' refers to the non-research purpose of the data. He argues that to conduct scientific analysis, data should be 'made', i.e. constructed by the researcher to answer the specific research question. The statistical analysis of big data is of secondary use, since they were intended for different primary use, in contrast with other form of data, such as survey data, which are collected in such a way to permit statistical analysis in valid and reliable ways.

Groves (2011a,b) makes a similar distinction between *organic* and *designed* data. Groves suggests that designed data are created with a specific idea in mind and organic as 'a self-measure in increasingly broad scope'. He considers the difference in the amount of knowledge that can be obtained from the two types of data. The ratio of knowledge to data is higher for designed data than for organic data for the use of interest, primarily because the data was created in order to extract the maximum level of information, with minimal noise.

Another definition proposed by Forbes[4] advocates that big data represents "the belief that the more data you have the more insights and answers will rise automatically from the pool of ones and zeros". Another characteristic of big data is that it represent a disruptive innovation which enables new approaches to science (Kitchin, 2014). Some believe that we are heading towards a data-intensive science; while others suggest that we are in a new era of empiricism, where the data speaks for itself and where theory is not necessary.

The new empiricism approach advocates that insightful and meaningful knowledge can now be produced directly from the mere observation of the data without any theoretical method to extract such knowledge: "with enough data, the numbers speak for themselves" (Chris Anderson, ex editor-in-chief WIRED[5]).

This is a completely different approach to the traditional type of analysis, which begins with relevant questions from which the appropriate methods are selected based on theory, and then the data is collected accordingly to answer these initial questions. On the other hand, advocates of this new approach may insist that hundreds of different algorithms can be applied to a dataset to find hidden knowledge, without having to justify the use of them. The weaknesses of this empiricism argument are several. First, the data is not exhaustive, but instead it is a sample, it is determined by the technology used and subject to regulatory environment and sampling bias (Crawford, 2013). Second, the data is not generated free from theory. The algorithms and analytic methods used to capture certain types of data are based on scientific methods and were tested for scientific validity. Any attempt to identify patterns is not free from scientific theory. Third, the data is not free from human bias and framing. In order for data analysis to make sense, it needs to be contextualized within a particular scientific approach. Finally, any analysis should be interpreted within a context or domain-specific knowledge; the data cannot just speak for itself (Kitchin, 2014).

Those believing that science is becoming data-driven suggest that hypotheses and questions are found in the data rather than in the theory. However, theory is used in developing knowledge discovery techniques which identifies questions that are worthy of further examination. Therefore, data is not generated by every possible means; data is generated and analysed under assumptions which guarantees that the techniques used will produce valid insights.

The technological advances which propelled the process of data generation has increased over the previous decades and are unlikely to decelerate. Therefore, the era of 'big data' is likely to continue into the future as well, so it is necessary to develop statistical methods and techniques that have the ability to utilize such new forms of data and exploit their potential.

The term 'big data' is quite broad and comprises different types of data and industries. We will focus our discussion on social media data. In the next section we present an overview of social media data, outlining their characteristics and emphasising their differences compared to traditional sources of data.

1.1.2 Social Media Data

The UNECE (United Nations Economic Commission for Europe) in 2013 set up by the High Level Group for the Modernisation of Official Statistics, which developed a classification of big data containing three main types of big data sources (Beręsewicz et al., 2018). Firstly, we have machine-generated data, usually captured by sensors. Secondly, the classification includes data generated as a by-product of IT system. These data is generated by people as they interact with IT systems. And thirdly, human generated data which is stored in digitalised form. Note that, while the first two classes include data with high level of detail, data belonging to the last class is unstructured and of poor resolution.

Also, all data sources, with exception of human-volunteered data, produce designed data, maybe not for statistical purpose, but for some purposed of data processing.

Social media data belong to the third class. Social media consists of conversational platforms. People, who join them, utilise this technology to share content and communicate their ideas and opinions with others. The communication may take place through actively participating in already existing debates or creating new ones, or it may be passive observation without interaction. Users can act as their true self or they can create new identities; they can decide when to enter and when to leave, what to share publicly and what to keep private.

Social media can be seen as collections of anecdotes: short stories that are significant to describe a topic or a population. They may be real or fictional, and perhaps involve subtle exaggeration and drama in order to entertain the audience. Social media data is generated in real-time and often as an answer to external, real-life events. People may upload their personal stories or interact over phenomena that are occurring in the offline sphere.

The data generated from surveys and experiments follows specific criteria and theory which allows valid conclusions to be drawn. Unlike survey or experimental data, social media data is generated directly from the users without any specification or questions to be answered. In other words, social media data is constructed free of a theory that is motivated by external purpose; the data coming from social media is instead generated as a consequence of socio-technological factors.

Following Citro (2014), social media data belongs to the class of data obtained

by the interactions of individuals with the World Wide Web: the individuals that are providing information in their posts, are not asked to respond to a questionnaire or required to supply administrative records; they choose to share their information autonomously.

Every social media platform has a particular topic or theme of interest. The underlying idea behind every social media site is to create communities of users. In certain cases the community is broad and encompasses different topics. Facebook, for instance, is an agglomerate of many small communities with variable interests; in other cases, the intent is to form a niche in order to host a more specific theme, such as LinkedIn, where users are connected by professional relationships or ResearchGate for example, which connects mainly academics.

A last consideration. Some of the data obtained from social media data can be structured and highly detailed, for example the geographical information that the individual voluntarily decides to share.

An example: Twitter Twitter is an example of an online news and social networking site where people communicate. The method of communication occurs through short messages, called *tweets*; the act of sending tweets is called *tweeting*. Some people refer to the act of tweeting as microblogging, since people often share tweets to their followers with the hope of being useful or interesting to them, as well as increase their audience.

A peculiarity of this social site is that each tweet has a limit of 280 characters (changed in 2017 from the original 140 characters). This limitation, on a side, promotes the use of clever and direct language which makes the text easier to scan; although, it can also incite the use of abbreviations which might require more effort to interpret from a textual analysis perspective, if the researcher is interested in the tweets' contents.

To be able to tweet, an account needs to be created. To register the user has to provide an email address, a username and a password. Optional fields are a profile picture, a bio and a location. Once an account is registered, it can start sending tweets. Tweets are by default publicly available, although the user may change the privacy setting of his or her profile in order to make it private.

1.2 Use of social media data for social research

Social media are arenas where issues are debated and opinion formed. This has caused an increased interest from social researchers to use them to understand societal phenomena and characteristics. Three main applications of social media data can be distinguished for social research (Japec et al., 2015):

1. to capture what people are thinking or talking about;
2. to analyse public sentiments and opinions on specific topics;
3. to understand demographics about a specific population.

Twitter is often considered a source of real-time news from its users. The content generated by the user in tweets covers daily stories, local news or world-wide news which are reported as they happen. It is therefore plausible to consider the use of tweets to gain an understanding of events occurring in real time. Petrovic et al. (2013) have shown how traditional newswire and Twitter equally cover the same major events. They also found that Twitter has better coverage of sport events, unpredictable high impact phenomena and small or local events. This type of analysis is known as event detection and a survey of its techniques is presented in Hasan et al. (2017).

In addition, users do not only report news or stories, but they share their opinions about them, as a result, social media also includes a substantial display of sentiments. This provides the opportunity to analyse tweets' content to gain insights about what people's opinions are. For instance, a manufacturing company may be interested in understanding what people think about their product and how positive (or negative) their opinions are. Political parties and social organizations may be interested in people's opinions about current debates. This type of interest is known under the term 'sentiment analysis'. Sentiment analysis offers a series of techniques involving the analysis of a text, identification of key words and the classification of opinions. See Pang and Lee (2008) for a review of the current state on opinion mining and sentiment analysis.

In these first two applications, analysis and interpretation occur at the tweet level, the text which constitutes the tweets being the object of analysis. Research can, however, be also conducted at the Twitter account level. This type of research involves an understanding of the characteristics of the accounts, for instance demographic characteristics of the user behind each account, as user profiling.

Daas et al. (2016) demonstrated how some demographic characteristics can be extracted from some Twitter-specific characteristics of the users. This type of research is also aimed at obtaining auxiliary information that can be used to link social media units to traditional survey units.

1.2.1 Quantitative social research

Social researchers study social phenomena though quantifiable evidence and rely on the empirical investigation of observable phenomena via statistical and computational techniques. Social research is centred around the collection of data, which is based on a given hypothesis.

The researcher starts her investigation with an hypothesis, formulated in terms of a research question, and collect the appropriate data to extend, revise and test the hypothesis through the data analysis.

In many cases, it might be impractical to collect data directly, and the researcher has to rely on secondary data, i.e. data which has already been collected by someone other than the user, such as social media data.

When secondary data are used for social research, the quality of the data becomes of primary interest. From the statistical inference point of view, what really matters is the way these data are generated, in particular to assess whether the methodological assumptions behind the use of a statistical method are met and the theory drawn from the evidence is consequently valid.

1.2.2 The statistical characteristics of social media data

In the following, we will discuss the statistical characteristics of social media data, in particular their non-probabilistic character, and their organic and unstructured nature.

1.2.2.1 Imperfect coverage

The representation problem of social media is a well-known problem which undermines the generalization of the results to the broader population. In this section, we explore the population of social media users and how it differs from the general population.

Demographics on social media The Internet usage is continuously spreading over the year. Approximately 89% of adults in the UK used the Internet in 2017 and in 2016 63% of adults in the UK had reported using the Internet for social networking (ONS, 2016, 2017). Social media platform undoubtedly accelerated Internet usage: they are easy to use, cover different and generic topics of interest and are used by a broad spectrum of the population. This last characteristic in particular contributes to increasing penetration of social media in the population, due to the network effect which incentives people to join the networking sites.

Social media users are not spread evenly throughout the population. Approximately 90% of adults resident in the UK, aged $16-34$ are active on social networks while there are only 23% of those aged 65 and over using social media (ONS, 2016, 2017). Greenwood et al. (2016) describe the demographic characteristics associated to some social media platforms in the USA; for example, younger Americans are more likely to be on Twitter, while the number of older adults (> 40 years old) joining Facebook has been increasing. LinkedIn continues to be popular amongst college graduates and high income earners.

Different usages of social media It is important to note the different behaviours that users display on social media. Who creates the most content and what type of content they create depends on different factors and diverging patterns have been found (Bright et al., 2014). However, some common ways of using the sites can be described.

Power users: The majority of content is created by a small group of users. The general term to describe these types of users is *influencers*, since their purpose is often to influence other people's opinions about current salient issues or for commercial reasons. There are no clear conclusions concerning why some users may share more content than others. Some results show that age and skills could have an effect on the differences in content creation (Hargittai and Walejko, 2008). Others show that different types of contents are affected by different user's characteristics. For instance, those with higher academic qualifications are more likely to engage in political discussions and less likely to create entertainment kind of content (Blank, 2013). Content from power users will be over-represented in any social media platform.

Residents vs visitors: we can distinguish two types of users. There are those

who actively spend most of their life online (residents), i.e. those who regularly (every week or every day) sign into a social media platform and those who come online only to satisfy specific needs, following which they leave (visitors). For the visitors in particular, the usage pattern is not uniform across time. Visitors tend to participate in the platform dialogue in correspondence of an important event, usually happening offline, which could have a public nature, as e.g. Twitter, or a private one, as e.g. in Facebook.

Not human users: social media platform are not only populated by users who are individuals. Institutions, governments and brands have their own profile that they use to communicate and promote their content within their communities. There are also accounts which are automatically controlled and can produce content. Those accounts are called *bots* and the most refined would appear as human as possible. Another example of not human users are the parody accounts, such as Elizabeth Windsor (@Queen_UK), which have a high number of followers and retweets.

Duplicated users: it is also possible for the same individual to have multiple accounts in a networking site. This is generally the case for celebrities or public professional figures. A journalist may have a professional profile, where the information shared is related to the newspaper for which she is working and a private one, where they can share freely their opinions.

1.2.2.2 Measures of interest not observed

The unstructured nature of social media data means that often they do not consist of direct measurements of the variables of interest but only proxies. What is observed in social media data cannot always be assumed to be the measure under investigation; in most of cases, the measure of interest requires extrapolation from metadata and interpretation.

In some cases the socio-demographic characteristics of the user are of interest. Sometimes, this information is given by the social site, while in the majority of cases, they need to be inferred by the researcher. Techniques for the automatic detecting of these characteristics are still under development. For example, in the case of Twitter, the users' age and sex can be identified from the users' profile picture and their writing style (Daas et al., 2016; Yildiz et al., 2017), their political views from their follow relationship (Golbeck and Hansen, 2014), their residence

from their geolocalized tweets (Swier et al., 2015).

In addition, opinions also needs to be identified. *Sentimental analysis* incorporates a family of techniques which address the problem of automatic detection of opinions and sentiments (Feldman, 2013; Pang and Lee, 2008). For example, a basic algorithm assigns to specific words a score which represent a sentiment, and the message is classified based on the frequency of the words with a given score. The field is still new and the challenges are numerous, such as detecting sarcasm[6].

Finally, even when the characteristics of a user or their opinions are clearly identified, it should be taken into consideration that they do not always represent what the researcher wants to observe. For instance, during the Iranian protests in 2009/2010, tweets posted in Iran appeared to be written outside the country, since people were afraid of repercussion from the government, while Iranians tweeting outside the country decided to geolocalise their tweets in Iran as a form of support (Halford et al., 2017).

On social media platforms, people interact with each other, in contrast to answering questionnaires, and they make their profile public, so that it is likely that they can be influenced by other people's opinions, or they want to publicise an image of themselves which does not necessarily represent the truth.

1.2.2.3 Secondary data

Halford et al. (2017) describe how the construction and circulation of social media data consists of a set of processes. To represent these processes, they use the data pipeline model as in Figure 1.1.

The pipeline model provides evidence of the way in which social media data are generated and processed beyond the researcher's control; in contrast, they are a consequence of technical, sociological and political factors. They are subject to the social media companies' technological infrastructures, the individuals' perception, legal regulations and ethical implications.

At the bottom of Figure 1.1 the process of the construction of social media data is conceptualised. The subject represents the user that generates the content on a laptop, mobile phone or a tablet. The information shared by the subject is controlled by the company's APIs, which determine what passes through the company's servers.

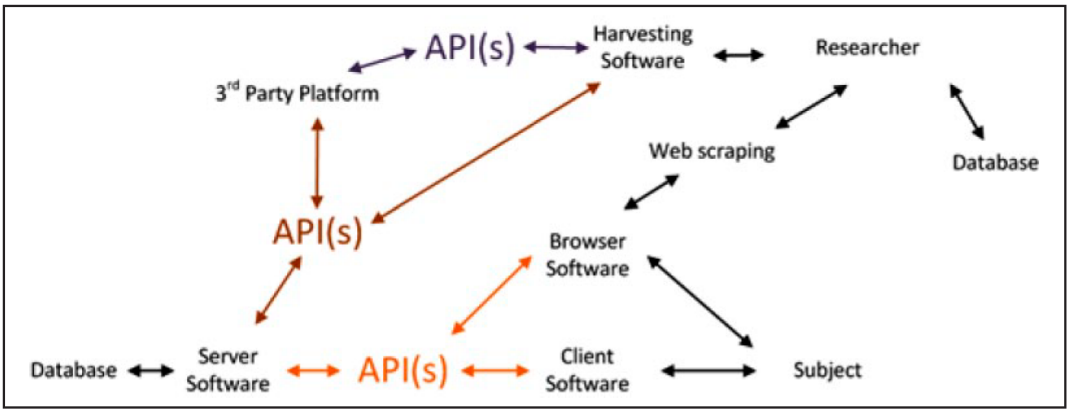


Figure 1.1: The social media data pipeline (Halford et al., 2017).

API Acronym for 'Application Programming Interfaces', the APIs are a set of routines, protocols and tools used for building software applications. They describe which functionalities are available and how they must be used. The use of a public API is strictly regulated by the policy of the API provider (Janetzko, 2017; Lomborg and Bechman, 2014).

Finally the information is organised in particular formats and structures to form databases. The information is shaped according to which client server is used; for instance, geo-referenced content will be more likely created on mobile phones, rather than laptops. The APIs take a core role in the process, given that they provide whether certain information is taken or not (see also section 1.2.2.4). When a database is created, the process also can be read inversely. The information sent back to the subject is also part of a process that involves the same actors: servers, APIs, client or browser software; for example, changes in the API will change what users can do.

The upper part of Figure 1.1 describes the processes which allow researchers to collect social media data. These processes move in different lines, all having a database at the endpoints: one database is generated by the content produced by the users and shaped by the company's API, server and interfaces; the other database is the one that the researcher constructed with the data collected after cleaning and processing. The data can be collected by the researcher from the company or through Web scraping. Web scraping involves downloading data directly from the browser; the data available on the browser is constrained by the APIs and by the browser as well (browsers, in general, offer personalised Web surfing experiences to their users). Data can also be collected from the company

directly or via a third party using their APIs.

In particular, different colours are used to distinguish the APIs used between different actors. In orange are the APIs which regulate what, and how, is passed through to the company's server software and the software that store the data. The APIs in blue and red are those that provide the set of rules determining which data can be harvested and their limitations from a third party or the server software respectively.

These networks of heterogeneous actors, that can be read in both directions, offer an illustration of the socio-technical factors that shape the process of data production and collection.

1.2.2.4 Selectivity issues

All the activities made on a social media platform are collected and stored. Access to all of them is, in some cases, impossible; in other cases, really expensive. However it is possible to collect a part of the data stored by the media site from its API.

There are two types of APIs in terms of accessability: the *restrict* APIs, when the access is granted only under special conditions, and *public* APIs, which give universal access. Public APIs are normally used by researchers to get access to a social media to collect data.

Via API Every social network has its own APIs, which works in different ways, however some common filters used to obtain the sample can be distinguished. Most of the APIs have limitations set by the owner of the sites. These limitations include the amount of data that can be retrieved and the time of data collection; for instance Twitter does not allow the collection of tweets older than seven days.

The data collection starts with the identification of the objects used for the selection, which depends on the social media of interest. Facebook involves users and post, Twitter involves hashtag and tweets, Instagram involves pictures. These objects have different attributes assigned to them, for instance a tweet contains a time stamp, the username of the user who posted it and in some cases, the location where it was posted. Finally, the different objects are able to interact with each other within the network.

The most common criteria used to collect data for social media are illustrated below (Mayr and Weller, 2016).

Based on topics and keywords: social media contents, i.e. tweets, Facebook posts, blog posts etc., can be obtained by searching for a specific topic, for example a specific event or a general topic. Note that there are many limitations to achieve completeness, such as the use of different vocabulary and language or use of different hashtag to indicate the same event. For example during the EU referendum in the UK many hashtags where used to express opinions on the topic: those in favour to leave: `#beleave`, `#brexit` (even though this hashtag was quite generic and used to describe the political event), and `#voteout`; those pro-EU: `#bremain`, `#strongerin` and `#hugabrit`; and the neutral most commonly used: `#EU`, `#UK` and `#EUreferendum`. These final hashtags are quite generic and, if not used together with some of the previous ones, could be misleading and not related to the topic of investigation. Tweets without hashtags also sometimes occur. This criteria of selecting data can be used when the interest is on a particular topic or to define a subpopulation.

Based on structural metadata: filter metadata, for example geolocation, timeframe, language or format (for instance only retweets or only status containing a URL) can be used to select social media data. These methods can be used when the investigation concerns particular characteristics of the target population (for instance the residency) or to define a subpopulation (for instance, those who speak a specific language).

Random Sample: if the interest is not a specific topic or characteristic, the API can provide a random sample of objects, in the case of Twitter a sample of tweets. The algorithm used to obtain the sample is unknown and property of the site itself.

Based on user accounts: a sample from a given population of users can also be collected. This approach of selecting a sample can only be feasible if the usernames are known in advance. Consider the case where the interest is on the political candidates during an election. If a complete list of their usernames are available, all the data that they produce on the social media platform can be retrieved. This approach is only attainable when the group is made of 'elite' users (a small group of known people), rather than 'ordinary' users; for instance it is not always possible to identify all those

who are eligible to vote, hence a random sample from the population of potential voters cannot be extracted. Rebecq (2018) uses the user ID number to randomly select a set of users from Twitter. It has to be noted, however, that some ID numbers are missing as well as the maximum number between them, indicating that the number of existing users is not known.

The Twitter APIs The collection of all the tweets is called the *firehose*. Using the free available Twitter's APIs a complete access to the firehose is not possible, however there are other ways to obtain it (see below). Twitter's free APIs are organized into two categories, each of which provides a set of rules and criteria used to collect the data: the REST API and the streaming API. Both APIs are constantly changing; terms of usage, data access limits, technical features (like geo-tagging) can be updated in times. These changes of regulations are not only due to technological advance, but they are also related to specific strategies of the company.

Streaming API: it offers access to the global stream of tweet data. The streaming data has two different endpoints, which specify where the data that can be accessed is situated:

Filter endpoint. Used to obtain a stream of tweets which match one or more keywords;

Sample endpoint. It offers a random sample of 1% of the firehose.

REST API: it is used to retrieve past tweets. It is characterized by some limitations: 1. Only tweets between the last seven days to 24 hours can be collected; 2. It is focused on 'relevance and not completeness'[7], so that not all tweets will be indexed and made available; 3. The Rest API does not provide the same results as the Twitter Web search, i.e. if the same keyword is used to search on the Twitter search (in the website), the list of the resulting tweets will not necessarily be the same as the one obtained via the Rest API.

There is only one endpoint:

Search endpoint. Via a variety of query operators, such as hashtags, text, usernames, etc., a specific search is made possible among tweets.

Via GNIP or similar Data collection can also be made using social media API aggregation services. They are paid service that offer completely access to the Firehose and retrieval of historical tweets.

The companies offering these services "gather data from the APIs of over 40 different publishers, normalize the content into one format (Activity Streams), enrich the stream with relevant metadata, and send those streams on to our customers through one pipe"[8] (Rob Johnson, Gnip).

The most established companies in this sector are DataSift and Gnip (which was acquired by Twitter in April 2014). Note that DataSift lost access to the complete full data stream of Twitter and to historical data in 2015.

Web scraping Web scraping is another form of data collection which involves the use of programs to process Web pages and extrapolate the required information, such as social media content. Using this method, the researcher does not have to go through the company's APIs, and she is able to collect data which may not be given by the company's terms. Although, it has to be noted that when the data is scraped from the web, the content available is often personalised for the registered users from the company. Furthermore, even when the session is anonymous, the content can still vary according to the geographic location or the browser used, amongst others, during the request.

1.2.2.5 The unit problem

Research conducted with social media data may vary from one form of social media to another, however a generic scheme illustrating how the research is conducted can be described in three steps. The first two steps are related to the identification of the population of interest via the unit of data collection. The third step concerns the measurement of the variable of interest, which has to be constructed from the content that the user has posted or provided on her profile. If the measure has to be taken from the posts related to each user, they need to be aggregated to produce a single value for the variable.

1. Identify a time frame and a geographical place. The construction of a frame for the target population begins here.

2. Identify a set of relevant keywords or metadata. For example, if the filter involves the use of keywords, then the selection of those filters refines the

frame of the target population. The choice of words to include in the set is quite delicate, since it can produce under or over coverage errors. For instance, if a term is too generic, it will likely include units which are irrelevant to the research, resulting in overcoverage. On the contrary, if terms are too specific, the risk is to exclude units which are relevant. Note that units could include both those which are linked to people (users) or linked to the content produced by people (post, tweets, etc.), according to the data collection from the social media network.

3. Finally in this step the focus is on extracting meaning from each unit according to the format of the data, which can be text, URL, images etc. and constructing a method for measuring the variable(s) of interest. This process can be made both by a human or machine learning algorithm. For instance, Yildiz et al. (2017), compare the results of human vs. machine learning algorithms for identifying sex and age and find the accuracy to be higher when humans are asked to perform the task.

These three steps are quite generic and might accommodate data from each different social media platform, according to the permitted procedures used for selecting the data.

An important feature of the statistical analysis of social media data, which transpire from the above discussion, is that the definition of the frame and the observable measure is consequent to the choice of the units of data collection. Because, in most of the cases, the direct observation of the elements of interest, i.e. the user, is unpractical, the sample is taken indirectly from the posts or the hashtags, as above described.

The problem of the unit is not isolated to the representation dimension. As seen above, the measure of interest might also be computed as a function of the measures observed on the sampling units relevant to each element of interest.

In the first chapter of this thesis (Patone and Zhang, 2019), the current state of analysis of social media data will be discussed and classified into two approaches on the basis of how they relate to the problem of the units.

1.3 Types of inference with big data

In the previous section we have discussed the quality of social media data for making statistical inference. It appears clear that one of the main obstacle to achieve valid conclusion is that rarely a social media dataset represents a random sample of the population of interest, due to missing data, imperfect coverage and non random selection.

If the sample selection is not random, then no valid statistical inference can be made using a design-based approach. In sampling theory, randomisation plays a dominant role. It is employed to determine which units should be observed and randomisation distribution provides the basis for statistical inference.

Model-based approach to inference does not make explicit use of randomisation or probability sampling, and it offers a formal way to made statistical valid inference from non-random samples.

A review of the inference methods which use non-random samples is given by Buelens et al. (2018). Three broad types of model-based approaches to inference are distinguished: pseudo-randomisation or pseudo-design-based inference, model-based inference and machine learning or algorithmic methods. In their paper, Buelens et al. (2018) compare different methods in each class to derive which class is more able to remove selection bias in non-random samples.

A comprehensive overview of non-probability sampling and the their methodological issues for official statistics is provided by Beręsewicz et al. (2018).

Pseudo-randomisation or pseudo-design-based inference Pseudo randomisation includes all methods where the probability of being in the sample, which is unknown, can instead be modelled: pseudoinclusion probabilities are estimated and used to correct for selection bias and used in Horvitz–Thompson type estimators to account for unequal selection probabilities (Valliant and Dever, 2011). See Elliott and Valliant (2017) for a review of the pseudo-randomisation approach. Examples of pseudo-design-based inference are: propensity score methods (Rosenbaum and Rubin, 1983); linear weighting methods to non-probability sample (Baker et al., 2013); combine a non-probability sample with a reference sample to construct propensity models for the non-probability sample (Elliott, 2009); Sample matching (Rivers, 2007). Matching and propensity score adjustments are based on strong ignorability assumptions and can lead to serious bias

if these assumptions are not met (see, e.g. Young and Karr (2017)).

Model-based inference In the modelling approach, a model is assume to have generated the distribution of the variable of interest. The model is fitted using a sample. Smith (1983) has discussed how the sampling mechanism affect the inference drawn under a model-based approach. Sverchkov and Pfeffermann (2004) and Pfeffermann and Sverchkov (2003) discuss model-based approaches for informative sampling. Small area estimation has also been used to combine high-quality small-sized probability samples with large non-probability samples (Marchetti et al., 2015; Blumenstock et al., 2015; Brakel et al., 2017; Pfeffermann and Sverchkov, 2007). Finally, Pfeffermann et al. (2015) gives an overview of problems and issues with the use of big data in official statistics. See e.g. Smith (1983), Elliott and Valliant (2017) and Zhang (2019) for inference approaches assuming non-informative selection of the observed sample; see e.g. Rubin (1976) and Pfeffermann et al. (1998) for examples of approaches that explicitly adjust for the informative selection mechanism. Statistical models are also used to predict the units not in the sample (Royall, 1970).

Machine learning or algorithmic methods Machine learning methods are algorithms that are able to predict unseen values based on a given data set with known values. Unlike model-based prediction, these algorithmic models cannot be formulated as relatively simple analytic expressions (Buelens et al., 2018). Hastie et al. (2016) give a recent overview of common algorithmic machine learning methods. Examples are k-nearest neighbours, regression trees, artificial neural networks and support vector machines.

In their study, machine learning methods are the more powerful. They conclude that pseudo-design-based methods are too restrictive and will often be insufficient to remove selection bias from non-random samples. Also, a set of auxiliary variables explaining the missing-data mechanism is an essential ingredient for successfully employing non-probability samples in producing accurate valid statistical inference.

1.3.1 Challenges with the model-based approaches of inference

Two important characteristics of big data are the high-dimensionality and the large sample size. These characteristics are suggested to understand aspects of the data which would be difficult to understand with small data. The biggest promise that big data propose, thanks to its features, is related to the development of methods which can describe the relationship between outcome and predictor variables and efficiently predict future observations. Moreover, large sample sizes are essential to be able to identify and study subpopulations, especially when those consist of rare individuals, who may be difficult to capture with a small sample size (where, if they are captured, they are likely to be considered as outliers).

However, as Fan et al. (2014) argue, high dimensionality and large sample size are also the features which cause most of the challenges which invalidate traditional statistical methods and require a new set of tools to analyse big data. Below we list those challenges.

Heterogeneity. Being able to observe large enough sample sizes for different subpopulations, big data are useful to understand heterogeneity and investigate problems such as the association of certain covariates and rare features (which are now observable) or the effects of a certain treatment on a specific subpopulation. However, inferring models which represent a subpopulation in a large dimension can be problematic with standard models, since it can lead to overfitting or a further issue of noise accumulation.

Noise accumulation. Given the large number of variables available, it may be tempting to use them in the statistical analysis, increasing the number of parameters that needs to be tested or estimated. When a decision or a prediction rule is conceptualised based on such parameters their estimation errors will increase, causing noise accumulation. Furthermore, when a large number of variables are used, included variables which have a low signal-to-noise ratio, for classification or regression prediction, the models will provide low performances.

Spurious correlation. Again spurious correlation is due to the high dimensionality of the data and it refers to the fact that some variables, which are scientifically unrelated, might erroneously have high sample correlation. This may

lead to false conclusions, which challenge variables selection and incorrect statistical inference.

Incidental endogeneity. The term 'endogeneity' implies that some predictors may be correlated to the residual terms. Since for many statistical methods, the assumption of independence between residual terms and predictors is essential, incidental endogeneity may invalidate the statistical analysis. High dimensionality is a possible cause of endogeneity.

All four of the issues highlighted above are motivated by examples provided in Fan et al. (2014).

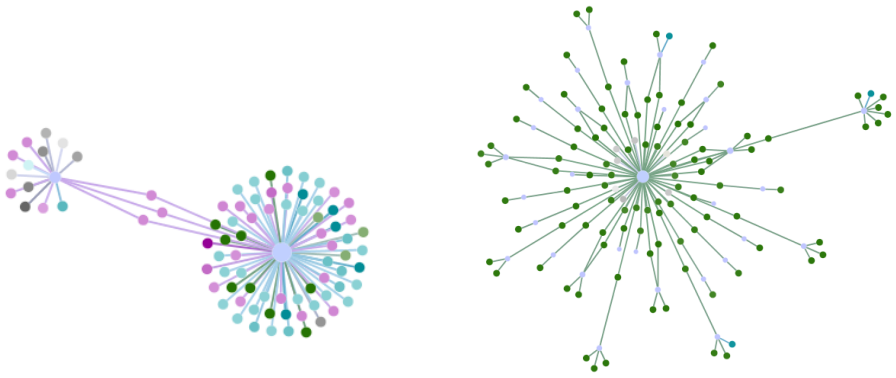
An example of a big data study that did not go as intended Take for instance the Google Flu Trend, which is the most recognized example of failure in the use of big data for scientific research. The idea was to use Google searches on flu symptoms, remedies, et similar to estimates the flu activities in the United States and 24 countries worldwide. At first, Google Flu Trend provided estimates which were more accurate than the Centers for Disease Control and Prevention (CDC), but after a while, it started to predict more than double the proportion of doctor visits than the CDC.

Lazer et al. (2014) attributes this failure in prediction to two causes: *algorithmic dynamics* and *big data hubris*. The algorithm dynamics could lead to the creation of bias in the data. In this specific example, the algorithm was such that, when someone searched for flu related terms, the algorithm was suggesting the search of flu symptoms and treatments, which were the terms used to predict flu. It is comparable to the case in survey sampling, when the interviewer suggests to the respondent that are coughing that they are coughing and could therefore have flu and asks the respondent if he has flu. The big data hubris implies that the large amount of data does not imply the data does not suffer from any other error, such those mentioned above.

1.4 The network structure

Social media can be reduced to an abstract structure which captures the main objects and the connections between them; this structure is known as a *network* (see Figure 1.2 for two examples of a network). More than one network can be constructed, according to the objects of interest (i.e. the users and the content

they create) and the relationships between them. A network, in its simplest form, consists of a set of *nodes*, i.e. the objects; a set of *edges*, i.e. the connections between them; and a set of attributes, i.e. different measures associated to both the nodes and the edges.



(a) (b)

Figure 1.2: Two examples of Twitter networks from the blog 'Digital Humanities Specialist'(`https://dhs.stanford.edu/gephi-workshop/twitter-network-gallery/`. (a): A conversation between Twitter users focused on whether terror and Islam can be separated; (b): a user actively writing different tweets at a variety of users during a short period.

1.4.1 Structure of social media

Social media data does not only consist of conversations between users, but also acts as a series of complex platforms involving different actors. Often, two main actors are distinguishable: the user and the content produced, which varies according to the specific platform. The two types of actors interact within their similar as well as between them.

The types of connection, between the objects of social media vary, according to the specific platform, changing the dynamics and structure of the platform and the way in which it is accessed. For instance, Facebook does not allow access to content unless the connection between users is mutual, while Twitter, which also allows for one-way relationships (user A can follow user B, without necessarily being followed by them), allows that there is open access to content to all users, depending on the user's privacy.

Consider Twitter. Each account can follow and be followed from other accounts. On the news feed, the tweets posted from the followed users will appears.

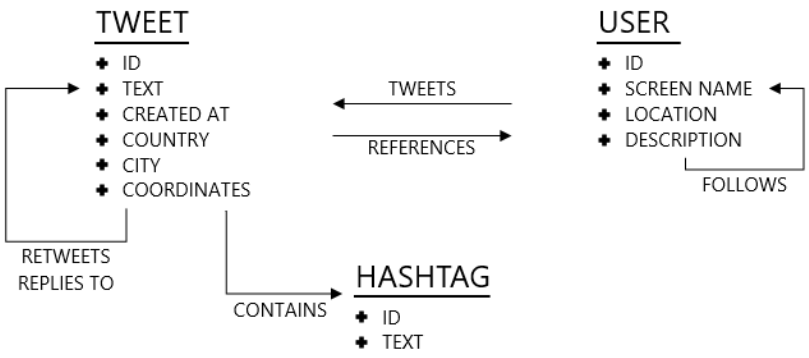


Figure 1.3: Conceptual model of Twitter activities.

Each tweet can be original, a reply to another tweet or a copy of a different tweet, known as a retweet (RT); it can mention a username account (@), to address a specific user, and it can contain hashtag (#), to declare the topic of the tweet. Hashtags offer a way to categorize tweets into specific topics (e.g. a tv show, a sport event, a news); for instance football matches, film festivals or conferences which may have an official hashtag under which the content generated by the users watching/attending the event is classified. Hashtags can also be user-specific and may not understandable for the general public.

Together with the hashtag, the two other main objects of Twitter are the account and the tweet and Twitter records both with their corresponding metadata.

Figure 1.3 represents a conceptual model of the Twitter platform as defined in Brown and Soto-Corominas (2017). The authors describe the Twitter platform with three objects, which are the users, the tweets and the hashtags, and the relationships that those objects share between them. For instance, a user tweets a tweet or a tweet replies to a tweet, among others. Further, to each one of these objects a series of attributes can be associated. For instance, for each user we have an ID, a screen-name, a location and a picture.

1.4.1.1 Measures of complexity

Representing the social media systems as a network can be a powerful tool to discover and understand patterns of connections or interactions between the elements of the systems or between some of their components. Different metrics or properties, borrowed from graph theory, are used to represent the form and the function of the system represented by the network, i.e. to analyse the structure

of the network. In the following we present a brief introduction to some of those properties.

Let A and B be two nodes in a network N.

Measures of connections: density of N, the ratio between the number of edges in the graph and the number of all the edges that could be present; *Out-degree of* A, the sum of the edges connecting A to the others nodes; *In-degree*, the sum of the edges connecting the nodes in the network to B.

Measures of distance: walk between A *and* B, the sequence of nodes and edges from a node A to a node B; *Geodesic distance between* A *and* B, the shortest walk between all the possible walks from A and B; *Diameter of* N, the longest between all the geodesic distances.

Measures of power: degree of A, the sum of edges from or to A; *Closeness centrality of* A, the sum of the geodesic distances between A and the other nodes.

For instance, an interesting characteristic of social networks is that there is a small but significant number of nodes with an extremely high degree. Those nodes are called 'hubs' and they usually play an important role inside the network, changing the performance and behaviour of the other nodes in the system and act as propagators of information.

Another related characteristic of social network is known as the 'small-world effect' which says that the mean geodesic distance between two nodes is usually short, which increases as the logarithm of the number of nodes in the network. This implies that information spreads rapidly around the network (six degree of separation).

A final example of a distinctive property of the network is the formation of clusters or communities. In Facebook, for instance, everyone is extremely connected with their close friends, which are likely to be connected to each other (high transitivity).

In the last decades, the study of the patterns of connections in social media networks has rapidly increased. On one side, the structure of such networks can clearly have an important effect on the behavior of the whole system; it seems therefore necessary to include it when we aim to understand how the social media network works. For example, the connections in a social network affect how

people debate, share opinions, and gather news, as well as how information is spread inside the network. On another side, when dealing with social media data, these patterns of connection can be easily observed and stored, when it was otherwise unpracticable, if not impossible, with more traditional type of data. The technological progresses has clearly motivated the interest on understanding these system.

1.4.2 Use of the graph structure for design-based inference

Stephan (1969) discusses modern sampling theory and suggests several further development of it. In particular, he calls *nexus sampling* the statistical inference in graphs. Stephan recognized that the conventional way of looking at a population disregard completely the interactions that might exists between the units and focuses only on the measurements attached to each individual. However, these interactions might be of value during the construction of a sampling design and for the estimation of the characteristic of interest.

For example, there are situations when some individuals can be observed only if the individuals related to them are observed in the sample. Stephan point out that a general theory of graph sampling is missing. Several attempts have been made: Goodman (2010) proposes *snowball sampling* as a sampling technique which allows to enlarge the initial sample by recruiting more elements, which are related to the initially sampled ones. *Adaptive sampling* (Thompson, 1990) is also a way to expand a sample, by adding only elements related to elements of interest. Birnbaum et al. (1965) and Lavallée (2007) consider instead the situation where the sampling frame and the population of interest consist of different elements, and the observation of an element of interest is subject to the observation of the sampling unit related to it.

All these attempts have aimed to solve a relevant problem, but a general theory which can incorporate all of them, by recognising the graph structure of the population, has not been provided. The first attempt to establish a general framework for finite graph sampling and estimation belongs to the pioneer work of Frank (1971, 1977a, 1979, 1980b, 1981, 2011). Frank does not focus on specific populations, but he investigates how different sampling methods can be applied to any population graph.

The second chapter of this thesis Zhang and Patone (2017) reviews the work made by Frank and includes some more extension. Also, as suggested by Stephan (1969), the connections available in the population graph can be used as a way of accessing into the population, with the purpose of obtaining more data: the third and four chapter of this thesis investigate how these extra information can be used in the estimation of totals of a population.

1.5 A structure of the thesis

The purpose of the first paper "On two existing approaches of statistical analysis with social media data" is to identify a range of theoretical and methodological challenges for a valid descriptive statistical inference. Two existing approaches of statistical analysis, aimed to overcome the basic challenges associated with these data, are delineated. In the first approach, the analysis is applied to the social media data that are organised around the objects directly observed in the data; in the second one, a pseudo survey dataset, aimed to transform the observed social media data to a set of units from the target population, is constructed and analysis applied to it. From the review of these two approaches, we conclude that the main difficulty in the one-phase approach is to identify an analytic connection to the target parameter, whereas the two-phase approach, besides facing the same challenges of non-probability sampling and measurement errors, introduces a new type of error that involves the transformation of the data into the units and measures of interest. The paper is currently under review in *International Statistical Review*.

The bigger part of the thesis concerns graph sampling and estimation. In the paper 'Graph sampling', published in *Metron*, we synthesize the existing theory of graph sampling and develop a general approach of HT-estimation based on T-stage snowball sampling, under the requirement of ancestral observation procedure. While the ancestral requirement might be hard to fulfil in many applications, it can be possible with social media data, by technological means. A key message of this paper is that the parameters which can be studied under a network representation differ from the conventional target parameters. It seems, in fact, feasible to investigate the interactions between the elements, their structural positions, etc. which are instead hard to be defined in a list representation of the population.

The other two papers deal with unbiased estimation methods from the sample graph. Both papers focus on the use of a Bipartite Incidence Graph (BIG), which offers an useful representation of many unconventional situations of sampling. In a BIG, the two distinct sets of nodes are represented by the sampling frame and the population of motifs; and a edge exists from a sampling unit to a motif, if its observation leads to the observation of the motif. In the paper 'Incidence weighting estimation under sampling from a bipartite incidence graph' we exploit the use of the observed edges in the BIG for estimation: each sampling method induces an incidence structure on the BIG, that, together with a relevant observation procedure, allows the estimation of characteristics of the population of motifs to be carried out on the sampled units rather than on the motifs directly. This use of the BIG is advantageous in situations when the probabilities of inclusion of the motifs cannot easily be computed, e.g. if a frame is not available, or in situations where the incidence structure can be used to improve the estimation. The use of the incidence structure of the BIG is also investigated in the last paper, 'Reverse weighting estimation under BIG sampling', but in the reverse direction. In this paper, we turn around to the opposite direction the incidence estimation described in the previous paper and make use of observed edges to carry out the estimation of cardinality of the frame on the sampled motifs, rather than on the sample of units. Without using any known additional variable, but only the size of the frame, we showed how the structure alone of the BIG can be used to improve the reduce the variance of the estimator, in the form of a Hajek-type and ratio-type estimators.

Notes

[1] http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9?IR=T, (accessed January 14, 2017).

[2] https://qz.com/344466/connected-cars-will-send-25-gigabytes-of-data-to-the-cloud-every-hour/, (accessed January 14, 2017).

[3] https://www.inc.com/business-insider/facts-about-amazon-jeff-bezos-seattle-2017.html, (accessed January 14, 2017).

[4] https://www.forbes.com/sites/gilpress/2014/09/03/12-big-data-definitions-whats-yours/2/#66ad3ea85b3f, (accessed March 4, 2017).

[5] https://www.wired.com/2008/06/pb-theory/, (accessed February 21, 2017).

[6] https://qz.com/598648/researchers-have-developed-an-extremely-effective-sarcasm-detector/, (accessed March 11, 2017).

[7] https://developer.twitter.com/en/docs/tweets/search/overview/standard.html, (accessed December 3, 2017).

[8] https://www.quora.com/What-is-the-difference-between-Gnip-and-DataSift, (accessed March 4, 2017).

Chapter 2

On two existing approaches to statistical analysis of social media data

Using social media data for statistical analysis of general population faces commonly two basic obstacles: firstly, social media data are collected for different objects than the population units of interest; secondly, the relevant measures are typically not available directly but need to be extracted by algorithms or machine learning techniques. In this paper we examine and summarise two existing approaches to statistical analysis based on social media data, which can be discerned in the literature. In the first approach, analysis is applied to the social media data that are organised around the objects directly observed in the data; in the second one, a different analysis is applied to a constructed pseudo survey dataset, aimed to transform the observed social media data to a set of units from the target population. We elaborate systematically the relevant data quality frameworks, exemplify their applications, and highlight some typical challenges associated with social media data.

Key words: quality, representation, measurement, test, non-probability sample.

2.1 Introduction

There has been a notable increase of interest from researchers, companies and governments to conduct statistical analysis based on social media data collected from platforms such as Twitter or Facebook. At the same time, there is also a growing concern about various issues associated with these new types of data. For instance, Boyd and Crawford (2012) ask whether such data may alter what 'research' means, and call for the need to interrogate relevant assumptions and biases. Bright et al. (2014) argue that caution is needed when interpreting social media data, and major questions remain on how to employ such data properly. Hsieh and Murphy (2017) highlight what they call coverage error, query error and interpretation error in relation to Twitter data. Halford et al. (2017) urge to develop better understanding of the construction and circulation of social media data, to evaluate their appropriate uses and the claims that might be made from them.

The aim of this paper is to examine and summarise two existing approaches to statistical analysis based on social media data, when the analysis otherwise would have been possible based on the traditional approach of survey sampling. To fix the scope, let $U = \{1, 2, ..., N\}$ be a target population of *persons*. Let y_i be an associated value for each $i \in U$. Let the parameter of interest be a function of $y_U = \{y_1, ..., y_N\}$, denoted by

$$\theta = \theta(y_U)$$

For instance, θ can be the population total or mean of the y-values. The quality of sample survey data can generally be examined with respect to two dimensions: representation and measurement (Groves et al., 2004). The representation dimension concerns the relationship between U and the *observed* set of persons, denoted by s. For example, s suffers from under-coverage if there are persons in U who have no chance of being included in s. The measurement dimension concerns the potential discrepancy between y_i and the *obtained* measures, denoted by y_i^* for $i \in s$. For instance, y_i^* may be subjected to various causes of measurement error, such that $y_i^* \neq y_i$ for some persons in s.

Thus, when social media data are employed, one needs to address two basic obstacles with respect to each quality dimension. Firstly, social media data are initially organised around different units than persons; secondly, the relevant measures typically cannot be directly observed but need to be processed using

algorithms or machine learning techniques. For example, one may like to make use of the relevant tweets to estimate the mean of a value associated with the resident population of a country. The directly observed unit (or data object) is then the tweets, whereas the statistical unit of interest is the residents. Next, instead of using designed survey instruments to measure the value of interest as one could in survey sampling, one will need to process a proxy to the target value from the Twitter texts by means of text mining.

Two existing approaches can be discerned in the literature. In what we refer to as the *one-phase approach*, statistical analysis is directly applied to the observed social media data that are organised around data objects other than persons. An example is Yan et al. (2019), who document statistical association between available drug-related tweets (processed by text mining techniques) in May - December 2012 and US county crimes rates (calculated against population size adjusted for non-residents) over 2012 - 2013. Next, in the *two-phase approach*, a different analysis is applied to a constructed *pseudo survey dataset*, after transforming the observed social media data to a set of persons from the target population. An example is Rampazzo et al. (2018), who document correlation between fertility rate published by the UN and that can be calculated for Facebook users. The pseudo survey dataset is collected directly from the Facebook Advertising Platform, which is assumed to be cleared of bots or other non-human accounts. The variable 'number of children' is also prepared by Facebook based on the information the company has about the users.

In this paper we shall delineate these two approaches more generally and systematically than they have hitherto been treated in the literature, where the Social Media Index for Dutch Consumer Confidence (Daas and Puts, 2014) serves as a typical case of the one-phase approach, and the ONS study on residency and mobility data constructed from geolocalised tweets (Swier et al., 2015) is used to illustrate the construction of pseudo survey dataset under the two-phase approach. We shall elaborate the relevant data quality frameworks and methodologies, and highlight some typical challenges to statistical analysis.

The rest of the paper is organised as follows. In Section 2.2, we systematise and describe in greater details the general issues of representation and measurement of social media data. In Section 2.3 and 2.4, we delineate and examine the one-phase and two-phase approaches, respectively. Finally, some concluding remarks are provided in Section 2.5.

2.2 General issues of representation and measurement

2.2.1 Representation

A major concern about the use of social media for research is the non-representativeness of data, when the population of interest does not coincide with the social media population (Boyd and Crawford, 2012; Bright et al., 2014; Halford et al., 2017; Hsieh and Murphy, 2017). Meanwhile, when investigating the representativeness of a social media population, one often compares it to the resident population of a country, about which one has high-quality statistics. For instance, Pew Research Centre publishes every year a report on the use and participation in social media of the US population. It is shown that US users of Twitter and Facebook tend to be younger and more educated than the US resident population (Greenwood et al., 2016). In the UK, Blank and Lutz (2017) find that Facebook users are more likely to be younger and female, while LinkedIn users are more likely to have an higher income than non-users. Mellon and Prosser (2016) examine how Twitter and Facebook users differ from the UK resident population in terms of demographics, political attitudes and political behaviour.

Twitter provides a typical example of online news and social networking site. Communication occurs through short messages, called *tweets*; the act of sending tweets is called *tweeting*. To be able to tweet, an account needs to be created. To register, a user has to provide an email address, a username and a password. A user can be a person, a business, a public institution, or even softwares (bots), etc. In case of person, the user is not obliged to create an account reflecting her physical persona. Optional fields include a profile picture, a bio and a location, which are neither verified nor expected to accurately characterise the user. By default tweets are publicly available, although the user may change the privacy setting to make it private. Each tweet can be original, a reply to another tweet or a copy of a different tweet, known as a retweet. It can mention a username account (@) to address a specific user, and it can contain hashtag (#) to declare the topic of the tweet. Hashtags offer a way to categorise tweets into specific topics (e.g. a tv show, a sport event, a news story). Some events such as football matches, film festivals or conferences may have an official hashtag under which the relevant tweets about the event are classified. Hashtags can also be user-specific and not intelligible to the general public.

As in the Twitter example, one can identify two directly observable units of data on most social media platforms, which we will refer to as the *post* and the *account*:

Post We use the generic term post to refer to the immediate packaging of social media content, which otherwise has a platform-specific name: Facebook has posts, Twitter has tweets and Instagram uses picture, etc.

Account An account is the ostensible generator of a post. As in Twitter, the user(s) operating a social media account can be different entities including but not limited to persons. Moreover, the same user can have multiple accounts, but the connections between these accounts and the user are not publicly accessible.

Denote by P and A, respectively, the totality of all the posts and accounts on a given social media platform. There is a many-one relationship from posts to the active accounts, denoted by $A_P = a(P)$, and the inactive accounts $A \setminus A_P$ is non-empty in general. Next, there is a many-one relationship from accounts to the users, denoted by $b(A)$. The *observable* persons are given by the joint set of the target population U and $u_{AP} = b(A_P) = b\big(a(P)\big)$, i.e. via the active accounts. Moreover, $U \setminus u_{AP}$ is non-empty as long as there are persons not engaged with the given social media platform, and $u_{AP} \setminus U$ is non-empty as long as they are other users than persons. These relationships are summarised in Table 2.1.

Table 2.1: Many-one relations a from post to account, and b from account to user

	Post	Account	Person
Totality	P	A	U
Observable	P	$A_P = a(P)$ $A \setminus A_P \neq \emptyset$	$U \cap u_{AP},\ u_{AP} = b(A_P) = b\big(a(P)\big)$ $U \setminus u_{AP} \neq \emptyset,\ u_{AP} \setminus U \neq \emptyset$
Sample	i. $s_P \subset P$ ii. $s_P \subset a^{-1}(s_A)$	i. $s_A = a(s_P)$ ii. $s_A \subset A$	$U \cap s_{AP},\ U \setminus s_{AP} \neq \emptyset,\ s_{AP} \setminus U \neq \emptyset$ i. $s_{AP} = b\big(a(s_P)\big)$, ii. $s_{AP} = b(s_A)$

Next, a common way of collecting data from a given social platform is via the public APIs, either directly or indirectly through third-party data brokers; Web scraping provides another option, albeit with unclear legal implications at this moment. Via the APIs, a sample of posts or, less commonly, accounts is harvested directly from the social media company and the obtainable sample depends on the company's terms and conditions. Depending on the API, the obtained datasets may differ in terms of being real-time or historical, or the amount of data that is allowed for.

Gaffney and Puschmann (2013) provides an overview of the tools available to extract Twitter data. For example, the `Streaming API*` returns two possible samples: a 1% sample of the total firehose (the firehose is the totality of tweets ever tweeted), without specifying any filter; or a sample of posts on specific keywords or other metadata associated to the post. However, if the number of posts matching these filters is greater than 1% of the firehose, the Twitter API returns at most 1% of the firehose. In addition, historical tweets can be retrieved using the `Search API`, which provides tweets published in the previous 7 days, with a selection based on "relevance and not completeness" (Twitter Inc.). For both APIs, Twitter does not provide the details of the process involved, nor guarantees that the sampling is completely random. See e.g. studies that have been conducted to understand and describe how the data generation process works with Twitter (Morstatter et al., 2013; González-Bailón et al., 2014; Wang et al., 2015).

Sampling of accounts is less common, which is only feasible if the usernames are known in advance. Consider the case where the interest is on the political candidates during an election. If a complete list of their usernames are available, sampling can be performed by the analyst; all the posts generated by the sample accounts on the social media platform can possibly be retrieved. The approach is only applicable when the group is made of 'elite' users (of known people), rather than 'ordinary' users; for instance it is not always possible to identify all the eligible or potential voters. Rebecq (2018) and Berzofsky et al. (2018) use the user ID number to randomly select a set of users from Twitter. Both the authors use also the available connections between users to propogate the initial sample.

Thus, the actually observed units are generally either a subset of P or A to start with. An initial observed *sample* of posts, denoted by $s_P \subset P$, can lead one to a corresponding sample of accounts $s_A = a(s_P)$ and then, in principle, a sample of users $s_{AP} = b\big(a(s_P)\big)$. Given a sample s_A directly selected from A, we can possibly acquire a sample of users $s_{AP} = b(s_A)$ and a sample of associated posts, denoted by $s_P = a^{-1}(s_A)$. The observed sample of persons are given by the joint set of U and s_{AP}. Again, both $U \setminus s_{AP}$ and $s_{AP} \setminus U$ are non-empty in general. The relationships are summarised in Table 2.1 as well.

2.2.2 Measurement

Unlike in sample surveys, social media data are not generated for the purpose of analysis. They have been referred to as "organic data" (Groves, 2011b) to emphasise their non-designed origin. One can only decide what is best to do with the data given the state in which they are 'found'. In light of the discussion of representation above, the obtained measures from social media data are either associated with the sample of posts or accounts. These may be based on the content of a post such as a text or an image, or the metadata of a post or account, such as the geo-location of a post or the profile of an account. According to Bright et al. (2014) and Japec et al. (2015), social media data are seen to provide the opportunity to study the following social aspects: 1. to capture what people are thinking, 2. to analyse public sentiment and opinion, and 3. to understand demographics of a population. To this one may add that social media data can obviously provide data about certain network relationships between posts, accounts or users.

Take Twitter for examples of all the possibilities mentioned above. While Twitter does not provide the information whether a user is a parent or not, it may sometimes be possible to infer that the user behind a tweet is a parent based on its content. Similarly, while Twitter does not provide the location of a user, it is sometimes possible to infer this from the location (or content) of the relevant tweets. When opinions about a particular topic are of interest, sentiment analysis can be performed on each tweet. By analysing the frequency of different hashtags, it could be possible to investigate the major topics that capture people's attention at a given moment. Finally, retweeting or the inclusion of certain hashtags can reveal particular network connections between the different users.

Generally we shall distinguish three types of data extraction from the sample posts and accounts, while at the same time noting the associated challenges in each respect:

Content Thought, opinion and sentiment provide typical examples of content extraction, which are the direct interest of study. Sentiment analysis is a common technique for extracting opinion-oriented information in a text. However, social media posts present some distinct challenges, because the expressions may be exaggerated or too subtle (Pang and Lee, 2008). Moreover, the posts on social media are public by nature, such that a user may easily be influenced by other opinions, or she may want to project an image

of herself which does not necessarily represent the truth.

Feature Demographics, location and socio-economic standing are common examples of feature extraction, when these are not the direct interest of study but may be useful or necessary for disaggregation and weighting of the results. Various techniques of 'profiling' have been used for feature extraction. For instance, Daas et al. (2016) and Yildiz et al. (2017) consider the problem of estimating age and gender of Twitter users based on the user's first name, bio, writing style and profile pictures. Or, Swier et al. (2015) derive the likely place of residence of a user, from all the geo-located tweets that the user has posted. Completely accurate feature extraction is generally impossible regardless of the techniques.

Network Directional posting, reposting, sharing, following and referencing all provide the possibility of observing network relationships among the posts, accounts or users. Common interests regarding the pattern and interaction among social network actors include identifying the most influential actor, discovering network communities, etc. Tabassum et al. (2018) provide an overview for social network analysis. However, it should be noted that the possibility and ease of network extraction is to a large extent limited by the APIs provided for a given social media platform.

In light of the above, whether by content, feature or network extraction from available social media data, one should generally consider the obtained measures as proxy values to the ideal target values. Of course, measurement errors are equally omnipresent in sample surveys. For instance, survey responses to questions of opinion may be subjected to mode effects, social desirability effects and various other causes of measurement error (e.g. Biemer et al. (2004)). So there is certainly scope for exploring social media data for relevant studies.

There is a noteworthy distinction between measurement errors in survey and social media data. In sample surveys, a measurement error does not affect the representation of the observed sample. The matter differs with social media data. For instance, when relevant accounts to a study are selected based on the metadata of an account, such as place of residence, errors can arise if the information recorded at the time of registration is not updated despite there has actually been a change of the situation. Such an error can then directly affect which accounts are selected for the study, i.e. the representation dimension of data quality. An initial measurement error in the description of the account can

thus result in a coverage error with respect to the study population. Similarly, one may fail to include a post in a study if it is classified as not containing the relevant opinion of interest. In sample surveys, the sampling frame is chosen to best fit the target population and it is obtained from external sources, such as registers of addresses or persons. Any error arising from an incomplete or erroneous frame is classified as coverage error and only affects the representation dimension.

It may be envisaged that combining multiple platforms, such as Twitter and Linkedln, can be useful for enhancing the accuracy of data extraction, although we have not been able to find any documented examples. This could be due to ethical reasons or the limitations imposed by the terms of conditions of the social media companies. An additional concern could be the 'interaction' between representation and measurement just mentioned above, where e.g. the accounts for which data combination is possible are subjected to an extra step of selection from the initially observed sample of accounts.

2.3 One-phase approach

In the one-phase approach, one needs to estimate the target parameter $\theta = \theta(y_U)$ directly from the obtained measures, denote by z_j, associated with a different observed set of units s_P or s_A, despite the differences to y_i and U.

To see why this may be possible at all, consider the following example. Suppose one is interested in the totality of goods (θ) that have been purchased in a shop over a given time period. One could survey all the people who have been in the given shop during the period of interest and ask what they have purchased. The population U then consists of all the relevant persons and y_i is the number of goods they have purchased (possibly over multiple visits to the shop). Alternatively, θ can be defined based on the transactions registered over the counter. The population P consists then of all the relevant transactions, and z_j is the number of goods associated with each transaction $j \in P$. Clearly, despite the differences in (y_i, U) and (z_j, P), either approach validly aims at the same target parameter θ.

Below we reexamine the Social Media Index (Daas and Puts, 2014) as an application, to formalise this approach and the relevant quality issues and methodological challenges.

2.3.1 Case: Social Media Index (SMI)

Every month, Statistic Netherlands conducts a sample survey to compute the Consumer Confidence Index (CCI). It is based on a questionnaire of people's assessment of the country economy and their financial situation. As part of the research on the use of social media data in official statistics (Daas and Puts, 2014; Daas et al., 2015), the authors collected posts from different social media platforms and constructed the Social Media Index (SMI) from these posts. They observed and compared the CCI and SMI over time and concluded that the two series are highly correlated (see Figure 2.1).

The SMI is constructed as an index that measures the overall sentiment of social media posts. The posts were purchased, in the time period between June 2010 and November 2013, from the Dutch company Coosto, which gather social media posts written in the Dutch language on the most popular social media of the country (Facebook, Twitter, LinkedIn, Google+ and Hyves). Coosto also assigns a sentiment classification, positive, neutral or negative to each post based on sentiment analysis (Pang and Lee, 2008), which determines the overall sentiment of the combination of words included in the text of the post. A neutral label is assigned when the text does not show apparent sentiment.

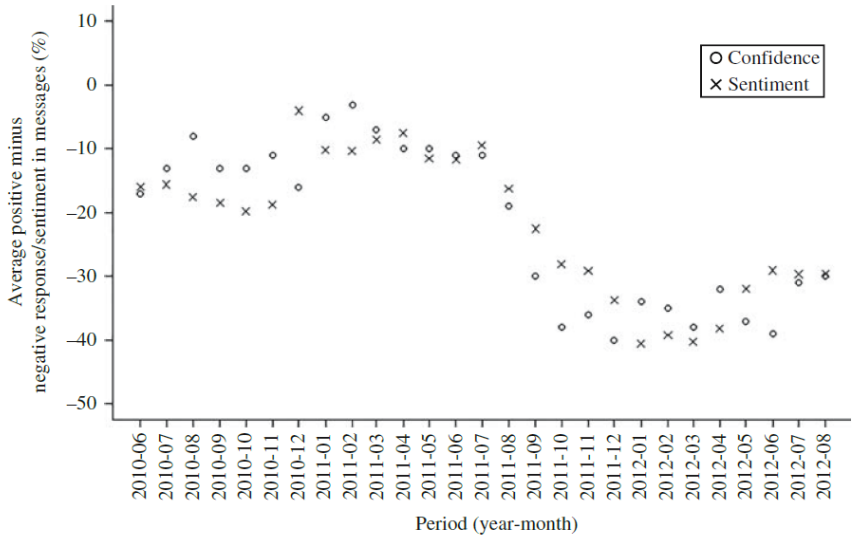


Figure 2.1: Comparison of Dutch CCI and SMI on a monthly basis. A correlation coefficient of 0.88 is found for the two series (Daas et al., 2015).

Let P_t be the totality of all the observed posts in month t. Let $s_{P,t}$ be a subset of posts that are selected from P_t. Let m_t be the size of $s_{P,t}$. The posts included in

$s_{P,t}$ can have positive, neutral or negative sentiment value, respectively denoted by $z_j = 1, 0, -1$, for $j \in s_{P,t}$. The SMI is calculated as the percentage difference between the positive and negative posts in $s_{P,t}$, i.e. a function of $z_{s_{P,t}} = \{z_j; j \in s_{P,t}\}$:

$$\text{SMI}_t = \text{SMI}(z_{s_{P,t}}) = \frac{100}{m_t} \sum_{j \in s_{P,t}} z_j \ .$$

Daas and Puts (2014) experimented with different ways of selecting the sample $s_{P,t}$. The choices involve a decision about which social media platforms to include, and whether to accept all the posts from an included platform or only certain groups. The groups can be filtered using a set of keywords, such as posts containing personal pronouns like 'I', 'me', 'you' and 'us', or words related to the consumer confidence or the economy, or words that are used with high frequency in the Dutch language. The idea is that selecting only certain groups of posts could affect the association between the SMI and the CCI. For instance, from a previous study (Daas et al., 2012) the same authors found that nearly 50% of the tweets produced in the Netherlands can be considered a 'pointless bubble'. In the end $s_{P,t}$ is chosen to include all the Facebook posts and filtered Twitter posts, for which the resulting SMI achieved the highest correlation coefficient with the CCI (Figure 2.1).

Finally, considering the SMI as an estimator with its own expectation and variance, let

$$\text{SMI}_t = \xi_t + d_t \ , \tag{2.1}$$

where ξ_t is the expectation of the SMI, and d_t has mean 0 and variance τ_t^2.

2.3.2 Formal interpretation

To assess the SMI as a potential replacement of the CCI, let us now formalise the CCI and its target parameter. Let U_t be the Dutch *household* population in month t, which is of the size N_t. Let y_i, for $i \in U_t$, be a consumer confidence score for household i based on positive, neutral or negative responses to five survey questions. The target parameter of the CCI is given by

$$\theta_t = \theta(y_{U_t}) = \frac{100}{N_t} \sum_{i \in U_t} y_i \ .$$

The CCI based on the sample survey is an estimator of θ_t, which can be given by

$$\text{CCI}_t = \theta_t + e_t \ , \tag{2.2}$$

where e_t is the sample survey error of the CCI. For our purpose here, we shall assume that $e_t \sim N(0, \sigma_t^2)$, i.e. normally distributed with mean 0 and variance σ_t^2.

Now that there is a many-one relationship between persons and households, the generic relationships from posts to persons apply equally from posts to households. The households corresponding to the SMI sample $s_{P,t}$ can thus formally be given as

$$s_t = U_t \cap a\big(b(s_{P,t})\big) \ .$$

Let s_t be of the size n_t. Let the target parameter defined for s_t be given by

$$\theta_{s,t} = \theta(y_{s_t}) = \frac{100}{n_t} \sum_{i \in s_t} y_i \ .$$

In order to replace the CCI by the SMI, it is clear that one would like to have $\theta_t = \xi_t$. However, given the underlying relationship between the social media data posts and the target population, one can only establish an analytic connection between ξ_t and $\theta_{s,t}$, based on the relationship between $(z_j, s_{P,t})$ and (y_i, s_t). It is therefore clear that the principal difficulty for the one-phase approach in this case is the lack of an explicit connection between ξ_t and $\theta_t = \theta(y_{U_t})$, or between $\text{SMI}(z_{s_{P,t}})$ and $\theta(y_{U_t})$. Moreover, it seems that in such situations external validation will be necessary in order to establish the validity of the analysis results based on social media data, which we consider next.

2.3.3 Statistical validation

In the case of the SMI, one does have the possibility of validating its statistical relationship to the CCI, despite the lack of an analytic connection between the two. As can be seen in Figure 2.1, the two indices display a high correlation with each other over time: the empirical correlation coefficient is 0.88 over the 27 months displayed. However, a high correlation between the two indices alone is not enough. Below we formulate a test to exemplify a possible venue for statistical validation in similar situations.

As a conceivable scenario in which the SMI can replace the CCI, we set up the null and alternative hypotheses below:

$$H_0 : \theta_t - \xi_t = \mu \quad vs. \quad H_1 : \theta_t - \xi_t \neq \mu \ ,$$

i.e. whether or not the target parameters of the SMI and CCI differ by a constant over time. Or, one can apply the procedure below on the log-scale to test if θ_t/ξ_t is a constant.

For our purpose here, we shall make a simplifying assumption that $\tau_t^2 = 0$, and thereby remove the conceptual distinction between SMI as an estimator and its theoretical target ξ_t. In light of the large amount of posts in $s_{P,t}$, the assumption seems plausible. It follows then from (2.1) and (2.2) that, under H_0, we have

$$X_t = \text{CCI}_t - \text{SMI}_t = \mu + e_t \ ,$$

where $e_t \sim N(0, \sigma_t^2)$. Thus, one may compare the total deviation of X_t from its mean $\bar{X} = \sum_{t=1}^{T} X_t$, over the available T time points, to the variances of the CCI: the larger the total deviation exceeds that which is allowed for by the CCI variances, the stronger is the evidence against H_0 compared to H_1.

Formally, let $P = I - \mathbf{1}\mathbf{1}^\top/T$, where I is the $T \times T$ identity matrix and $\mathbf{1}$ is the $T \times 1$ unity vector, and the matrix P is idempotent such that $PP^\top = PP = P$. We have

$$\begin{aligned} E(PX) &= \mathbf{0} \qquad \text{for} \quad X = (X_1, ..., X_T)^\top \ , \\ V(PX) &= P\Sigma P \qquad \text{for} \quad \Sigma = \text{Diag}(\sigma_1^2, ..., \sigma_T^2) \ . \end{aligned}$$

The diagonal matrix Σ corresponds to the assumption that the CCI's are uncorrelated over time. If this is not the case, one may specify the true covariance matrix appropriately, without this affecting the generality of the following development. Now that $\mathbf{1}^\top PX \equiv 0$, one of the component is redundant. Let $X' = (PX)_{(-t)}$ on deleting the t-th component of PX, for any $1 \leq t \leq T$. Let Q be the correspond $(T-1) \times (T-1)$ sub-matrix of $P\Sigma P$, such that X' has the $T-1$-variate normal distribution

$$X' \sim N(\mathbf{0}, Q) \ .$$

Let $LL^\top = Q$ be the Cholesky decomposition with lower-triangular L, such

that

$$L^{-1}Q(L^{-1})^\top = L^{-1}LL^\top(L^{-1})^\top = I_{(T-1)\times(T-1)}$$

and

$$R = L^{-1}X' \sim N(\mathbf{0}, I) \ .$$

A test statistic for H_0 can thus given as

$$D = R^\top R \sim \chi^2_{T-1} \ .$$

Under the alternative hypothesis, the test statistic D follows a noncentral chi-squared distribution with same degree of freedom $T-1$ and noncentrality parameter $\lambda = \sum_{\substack{i=0 \\ i\neq t}}^{T} \gamma_i^2$, where $\gamma = (\gamma_1, \ldots, \gamma_{t-1}, \gamma_{t+1}, \ldots, \gamma_T)$ is expected value of X' under the alternative hypothesis, that is

$$\gamma_i = \mu_i - \frac{1}{T}\mu_1 - \cdots - \frac{1}{T}\mu_T.$$

The smallest the value of λ, the bigger the overlapping between the null and the alternative hypothesis ($\lambda = 0$ iff $\gamma_i = 0$ for all $i = 1, \ldots, T$, therefore $\mu_1 = \mu_2 = \cdots = \mu_T$). On the other hand, the bigger λ, the smaller is the overlap and the higher is the statistical power of the test.

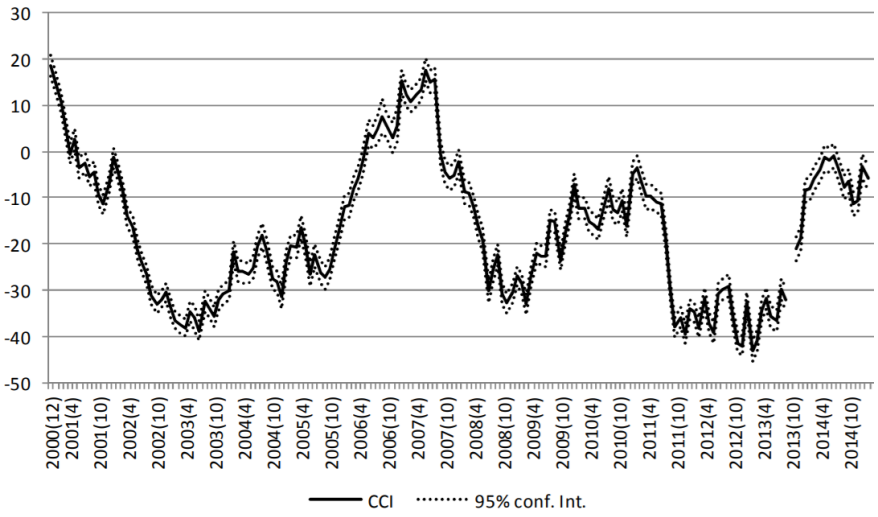


Figure 2.2: The CCI series with 95% confidence interval, 2000-2014.

Due to confidentiality restrictions, we can only obtain the CCI (from the homepage of Statistics Netherlands), but not the actual values of the SMI, nor the variances of the CCI. The calculations below serve only for the purpose of illus-

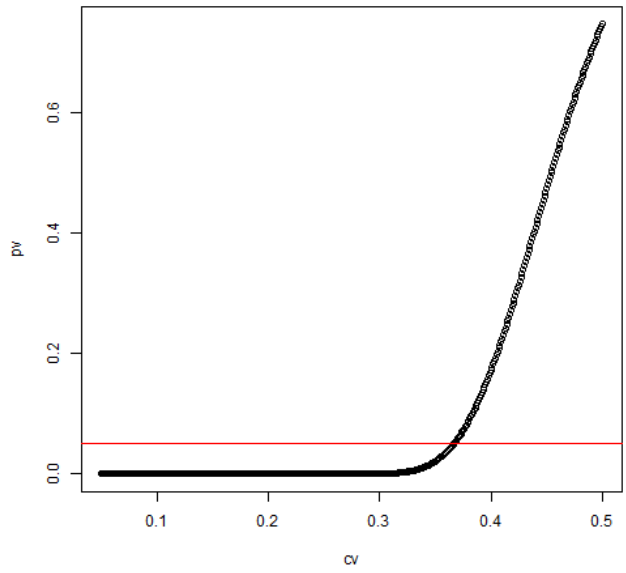


Figure 2.3: P-values of test H_0 vs. H_1 for varying CVs, level 0.05 mark by horizontal line

tration. Firstly, we eyeball Figure 2.1 to obtain the approximate values of the SMI, where the empirical correlation coefficient between two series is 0.88 over the 27 months. Next, Figure 2.2 reproduced from Brakel et al. (2017) plots the 95% confidence interval of CCI over 2000 - 2014, where the coefficient of variation (CV), denoted by $\eta_t = \sigma_t/\text{CCI}_t$, varies approximately between 0.01 to 0.34 over the period relevant to Figure 2.1. Based on these approximate σ_t^2's, the p-value of the test above is virtually zero, such that H_0 is rejected at the level of 0.05 or much lower. Moreover, for the illustration purpose here, we stipulate the values of σ_t^2 in relation to the CCI via a constant coefficient of variation over time, denoted by η, such that $\sigma_t = \eta\text{CCI}_t$. Figure 2.3 shows the p-value of the test as η varies from 0.05 to 0.5, where the p-value exceeds 0.05 for $\eta > 0.367$. In other words, unless the CV of the CCI is larger than 36.7% for all the 27 months of concern here, the null hypothesis is rejected at the level of 0.05.

2.3.4 Discussion

Firstly, in the above we have considered the validity of the SMI, assuming the aim is to replace the CCI with it. Of course, even if the SMI cannot do this directly, there is still the possibility to use it to improve the CCI. Brakel et al.

(2017) study the two indices over time using a bivariate time series model:

$$\begin{pmatrix} Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} L_t^Y \\ L_t^Z \end{pmatrix} + \begin{pmatrix} S_t^Y \\ 0 \end{pmatrix} + \begin{pmatrix} \beta^{11}\delta_t^{11} \\ 0 \end{pmatrix} + \begin{pmatrix} v_t^Y \\ v_t^Z \end{pmatrix} \ ,$$

where Z_t is the SMI that is decomposed into trend L_t^Z and an error term v_t^Z, and Y_t is the CCI that is decomposed into trend L_t^Y, seasonal component S_t^Y, an error term v_t^Y, and $\beta^{11}\delta_t^{11}$ that is an outlier term introduced to accommodate the economic downturn at the corresponding time point. The authors find that using the SMI series as an auxiliary series slightly improves the precision of the model based estimates for the CCI, at a time when the SMI for the current month is available but not the CCI – due to the longer production lag required for the latter. Notice that such uses of social media data as the auxiliary information for survey sampling does not pose any new theoretical challenges.

Next, disregarding the distinction between $\theta_{s,t} = \theta(y_{s_t})$ and the CCI-target $\theta_t = (y_{U_t})$, where one faces a difficulty of representation between s_t and U_t, there is a question whether the SMI (2.1) appropriately targets the 'intermediary' parameter $\theta_{s,t}$. As remarked by Brakel et al. (2017), the CCI survey questions involve the amount of purchases of expensive goods during the last 12 months and the tendency of households to buy expensive goods. It seems relevant to utilise internet search data and actual purchase data of such expensive goods. The implication is that one needs not to rely exclusively on social media data for content extraction, but could seek to combine them with other non-survey data. On the one hand, combining data to improve content extraction seems desirable regarding the quality of measurement. On the other hand, doing so is likely to affect the representation dimension of data quality, as previously noticed in Section 2.2.2. But the quality of representation is worth examining in any case. In the current definition of SMI (2.1), each post is given the same weight. It is unclear whether this is the most appropriate treatment, because the number of posts per account or user is likely to vary in different subsets of s_t. Indeed, provided a method of differential weighting of the posts in $s_{P,t}$ can be justified with respect to $\theta(y_{s_t})$, targeting $\theta(y_{U_t})$ may no longer be as elusive as it is currently.

Finally, despite our focus in this paper on target parameter θ defined for (y_i, U), it is conceivable that one may be interested in target parameter ξ defined for (z_j, P) directly. In such situations, the quality considerations are analogous to those in the case of targeting θ based on a sample s, for $s \subset U$, and the associ-

ated measures $y_s^* = \{y_i^*; i \in s\}$. A basic issue regarding representation is the fact that the sample s_P is not selected from the totality P according to a probability sampling design. Inference from non-probability samples have received much attention. See e.g. Smith (1983), Elliott and Valliant (2017) and Zhang (2019) for inference approaches assuming non-informative selection of the observed sample; see e.g. Rubin (1976) and Pfeffermann et al. (1998) for examples of approaches that explicitly adjust for the informative selection mechanism. When it comes to the measurement dimension of data quality, the traditional treatment of measurement errors in surveys (e.g. Biemer et al., 2004) may be less relevant because, as discussed in Section 2.2.2, content, feature or network extraction from social media data faces quite different challenges and uses quite different techniques than data collection via survey instruments.

2.4 Two-phase approach

In the two-phase approach, one aims to estimate the target parameter $\theta = \theta(y_U)$ based on a pseudo survey dataset constructed from the sample of social media data to resemble a survey dataset from the target population. Denote by s_{AP} the sample of statistical units in the pseudo survey dataset, and by y_i^* the constructed proxy to y_i for $i \in s_{AP}$.

The quality of the pseudo survey dataset (y_i^*, s_{AP}) with respect to the ideal census data (y_i, U) can be assessed with respect to representation and measurement, under the quality framework of Groves et al. (2004) for traditional sample survey data. The key extra concern is the necessary transformation from the initial social media data, which is a process that does not exist for sample survey data. Zhang (2012) outlines a two-phase life-cycle model of statistical data before and during integration, respectively, which includes the transformation from multiple first-phase input datasets to the ones to be integrated at the second phase. The total-error framework of Zhang (2012) is applicable as well to the two-phase approach to statistical analysis based on social media data.

Below we examine the study of Swier et al. (2015), which aims to construct pseudo survey datasets of residence and mobility from geolocated tweets. In particular, this illustrates the generic transformation process under the two-phase approach: from the first-phase data objects (posts) to the second-phase statistical units (persons) in terms of representation, and from values obtained at the first-phase

(e.g. the geolocation of a post) to the second-phase statistical variable (e.g. location of residence) in terms of measurement. Moreover, we analyse the quality of the resulting pseudo survey dataset according to the total-error framework of Zhang (2012), and highlight some relevant methodological challenges.

2.4.1 Case: Residence location from tweets

Swier et al. (2015) conducted a pilot study at the Office for National Statistics, on the potential of Twitter to provide residence and mobility data for official statistics. The main efforts concerned the construction of relevant pseudo survey datasets, which we summarise below. In addition, some simple analyses were performed, giving indications of the possible target parameters envisaged. We do not explicitly discuss these analyses here.

There are two first-phase input datasets. The first one is collected via the Twitter `Streaming API`, covering the period 11th of April to 14th of August in 2014. The search criteria involve a set of bounding rectangles covering the British Isles, for which a tailor made application is developed and deployed. The second dataset is purchased from GNIP (a reseller of data, now owned by Twitter), covering the period 1st to 10th of April and 15th August to 31st of October in 2014. Unlike the API data, the GNIP data is filtered by tweets with a "GB" country code. The tweets from the same period, which cannot be geo-located in either way, are excluded.

Next, the two datasets are merged to create a single dataset, during which a number of tweets are removed. These include e.g. the ones that are detected to be generated by bots, or without exact GPS location, or non-GB tweets in the first dataset (mainly those from the Republic of Ireland). In particular, for privacy protection reasons, any tweet from the first dataset is removed, unless it is associated with an account in the purchased GNIP data. All the retained tweets have latitude and longitude (GPS) coordinates.

The process of merging can therefore equally be represented as in the life-cycle model of integrated data (Zhang, 2012), where linkage of separate datasets are carried out via the second-phase units associated each input datasets. In other words, one may first identify the associated Account IDs (second-phase units here) in the API and GNIP datasets, respectively; and then merge the data for the same Account ID, provided it is present in the GNIP dataset. In this case one

could merge the datasets before transforming the data organised around Tweet ID to Account ID, because the two first-phase datasets share the same identifiable objects (i.e. tweets with Tweet ID)

In this way, at the beginning of the second-phase processing, one obtains a single set of GB-located tweets (81.4 million over 7 months) and the associated accounts. No further second-phase data processing takes place in the representation dimension. For instance, one does not attempt to identify and classify the users behind the observed accounts. Second-phase processing in the measurement is primarily concerned with content extraction of residential location and its classification. This is carried out in the following steps.

- The tweets associated with a given account are *clustered*, using the density-based spatial clustering algorithm with noise (DBSCAN). It groups together points that are closer to each other in terms of spatial density; the cluster formed is regarded valid only if it contains a specified minimum number of points. The points in clusters below the minimum threshold are considered as noise. Of the 81.4 million tweets, 67.4 million are included in one or another cluster that contains three or more tweets. The rest clusters with only one or two tweets are classified as 'invalid'.
- Next, each valid cluster is classified as 'residential', 'commercial' or 'others' in terms of address type, using the AddressBase that is the definitive source of address information for Great Britain. To this end, one calculates a weighted centroid of the cluster and finds the closest property to it in the AddressBase. The cluster address type is then classified according to this 'nearest neighbour' property.
- Then, for each account with one or several residential clusters, the one of them with the most tweets is classified as the 'dominant' residential cluster.
- Finally, additional classification may be attached to each cluster, such as the administrative geography it belongs to, the number of tweets it contains, the time span of these tweets (short-term if less than 31 days vs. long-term otherwise).

2.4.2 Quality assessment

Before we assess the quality of the pseudo survey dataset (y_i^*, s_A) obtained under the two-phase approach when targeting θ defined for (y_i, U), it is helpful to reca-

pitulate some of the relevant technical issues, even if they do not account for all the sources of errors.

Firstly, some additional API data are actually collected on the 10th of April and 15th of August, which overlaps with the GNIP data on these two days. A small number of API tweets are found not be included in the GNIP set, all of which are associated with protected accounts – users may opt to protect their accounts so that their tweets can only be viewed by approved followers. More generally, retrospective changes made by a user to its account or specific tweets may prevent them from being included in the historic point-in-time data available from GNIP, despite these accounts or tweets are accessible via the real-time `Streaming` API. This exemplifies a general cause for discrepancy between Twitter data collected in different ways. Two other examples of general causes are as below.

Filter criteria The filter criteria may not be fully compatible between the APIs and the data brokers. As explained above, in the case here, the geographic filter works differently with the `Streaming API` and GNIP.

Missing data Data from APIs may be missing due to technical problems, such as moving of IT equipment or broadband router failure.

Next, once the data from the first phase have been merged and transformed, there are generally technical issues with data extraction and processing that are necessary at the second phase. In this case, the DBSCAN clustering of tweets is an unsupervised machine learning technique, for which it is generally difficult to verify the truthfulness of the results. The address type classification is in principle a supervised learning technique. However, it may be resource demanding to obtain a training-validation dataset, by which the classification method can be improved and its accuracy evaluated. Similarly for the classification of the dominant residual cluster.

The quality of the dataset (y_i^*, s_A) can be assessed according to the second-phase life-cycle model (Figure 2.4), along the two dimensions of representation and measurement. The exact nature of the potential errors needs to be related to the envisaged analysis. Below we consider first representation and then measurement.

In terms of representation, the "Linked Sets" in Figure 2.4 is given by $b(s_A)$, which is subjected to coverage errors. Over-coverage is the case if $b(s_A) \setminus U \neq \emptyset$. This is unavoidable here because some of the accounts in $b(s_A)$ are not persons at

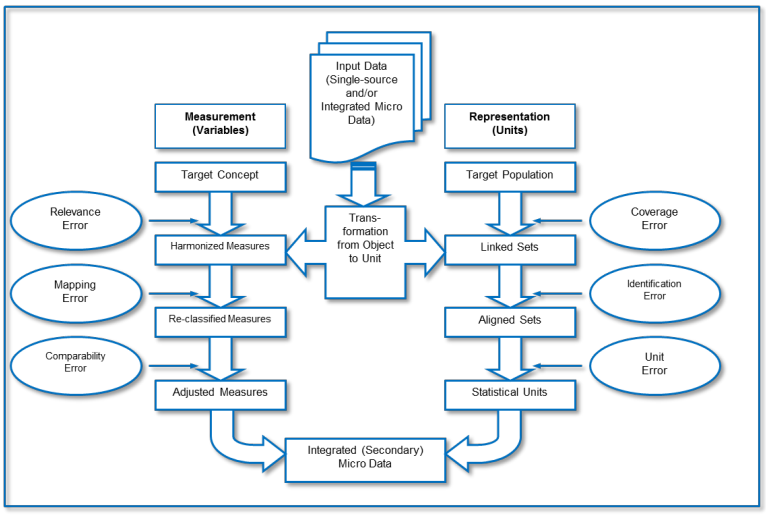


Figure 2.4: Phase-two life-cycle model of Zhang (2012)

all and all the bots are not completely removed. Moreover, there may be multiple accounts in s_A that correspond to the same person; such duplicates are another form of over-coverage error. Whether s_A entails under-coverage depends on the assumption. For instance, let the target population U be the adult residents of England. If one assumes that in principle there is an unknown but non-zero probability for everyone in U to have a Twitter account and to have tweeted at least three times from the same location during the 7 months in 2014, then there would be no under-coverage error of $b(s_A)$ for U, but only a non-probability selection issue. However, insofar as these assumptions are untenable, then there would be an under-coverage error in addition.

Next, the identification error may be an issue if domain classification of the target population needs to be based on feature extraction, which is prone to errors; whereas unit error is potentially troublesome if additional statistical units (e.g. household) need to constructed. Neither seems relevant to any of the analyses of Swier et al. (2015).

In terms of measurement, an example of "Harmonized Measures" in Figure 2.4 is the dominant residential cluster here. Suppose the "Target Concept" is the de facto place of residence of a person. Relevance error is mostly like the case, unless everyone sends most tweets from her de facto place of residence. Or, suppose the "Target Concept" is whether a person is a tourist, and short-term vs. long-term classification of the dominant residential cluster is used as a proxy measure of the corresponding person. Again, relevance error is mostly like the case, unless no

tourist stays longer than a month and no usual resident stops tweeting after less than a month.

Next, the mapping error is e.g. the case when someone does tweet from her de facto place of residence but the clustering-classification algorithm fails to identify it as the dominant residential cluster. This can happen e.g. if the person tweets more when at her friend's place, or if the person more often than not switches off GPS location when tweeting at home, or if the person's home is in a dense area and the chosen nearest neighbour property in the AddressBase happens to be a commercial address. Finally, the comparability error could arise if e.g. the classified dominant residential cluster is further adjusted in light of other available measures, although this is not the case in the study of Swier et al. (2015).

In summary, the main errors of the pseudo survey dataset (y_i^*, s_A) here are coverage errors in terms of representation, and relevance and mapping errors in terms of measurement.

2.4.3 Discussion: Statistical analysis

In the above we outlined the data processing required under the two-phase approach to social media data, using the study of Swier et al. (2015) as the case-in-point. It is shown that the life-cycle model of (Zhang, 2012) can be applied as a total-error framework for evaluating the quality of the resulting pseudo survey dataset (y_i^*, s_A), where $s_A = a(s_P)$. The study of Swier et al. (2015) does not specify any definitive target of analysis. For a discussion of possible statistical analysis of the target parameter θ defined for (y_j, U), let us consider two situations, depending on whether it involves additional datasets or not.

Consider the situation where only the pseudo survey dataset (y_i^*, s_A) is to be used for an analysis targeted at $\theta(y_U)$. The first key issue regarding representation is over-coverage adjustment, from $s' = b(s_A)$ to $s = U \cap b(s_A)$, due to the fact that $s' \setminus U \neq \emptyset$. This could be either based on the mapping from s' to s or, provided it can be specified, from $t(y_{s'}^*)$ to $t(y_s^*)$, where $t(\cdot)$ denotes the sufficient statistics for θ. Given the over-coverage adjustment, the remaining issues are non-probability representation of s for U, and measurement discrepancy between y_i^* and y_i caused by lack of relevance and imperfect data extraction, similarly to what has been discussed earlier in Section 2.3.4.

A potentially more promising scenario is to utilise additional datasets, in order

to overcome or reduce the deficiency of each dataset on its own. Integration with other Sign-of-Life data can possibly improve the quality of the pseudo survey dataset constructed from social media data. For example, in the case of data for residence and mobility, other Sign-of-Life data on employment, education, utility services, etc. can probably improve the classification of the dominant residential cluster, provided these data are available and can be combined with the tweets data. However, it is also possible that one cannot always overcome the inherent deficiencies of social media data in this way. Making statistics based on multiple sources is a broad challenging topic. It is currently an area of active research and development. See e.g. De Waal et al. (2017); Di Zio et al. (2017) for overviews of related situations and methodological issues. See Zhang (2018) for an overview of estimation methods in the presence of multiple proxy variables.

2.5 Concluding remarks

In the above we systematically delineated two existing approaches to statistical analysis based on social media data. The fundamental challenge with the one-phase approach in some situations is a lack of analytic connection to the target parameter, which is defined for a different set of units and another associated measure. Nevertheless, external data can in principle be used to verify the statistical validity of this approach. Compared to observational studies based on data subjected to non-probability selection and survey measurement errors, the key extra issues with the two-phase approach revolve around the transformation process from the initial data objects to the statistical units of interest and the algorithmic data extraction required for measurement. In addition, an explicit adjustment for the over-coverage error will be needed in many situations.

For assessment of data quality, we have demonstrated that it is possible to apply relevant total-error frameworks formulated in terms of representation and measurement of generic statistical data. In particular, for both approaches, it seems more promising if one does not simply restrict oneself to the available social media data, but seeks to combine them with additional relevant datasets, in order to overcome or reduce the deficiency of each source, despite data integration is by no means a straightforward undertaking in general.

We would like to close with a few remarks. Firstly, in the paper we have focused on target parameters that are finite-population functions. Such a parameter is

often referred to as a descriptive target, in contrast to analytic target parameters that can never be directly observed, regardless how large the observed number of units and how perfect the obtained measurement may be. For example, the ordinary least squares fit of some specified linear regression coefficients based on a perfect census of the current population is a descriptive target parameter; at the same time it is an estimate of the theoretical (or super-population) values of these coefficients of the postulated regression model, i.e. the analytic target parameter in this case. Our focus on descriptive target parameters helps to simplify the exposition, since the differences between descriptive and analytic inference can be subtle and many, but are nevertheless not critical to our aim in this paper. See e.g. Skinner et al. (1989), Chambers and Skinner (2003), and Skinner and Wakefield (2017) for introductions to analytic vs. descriptive inference based on sample surveys.

Next, there are certainly many similarities to statistical analysis based on administrative data. As we have demonstrated, the total-error framework (Zhang, 2012) for statistical data integration involving administrative sources is applicable as well to the two-phase approach based on social media data. It is worth reiterating the two extra difficulties in comparison. The first one relates to the transformation from the original data objects P to the statistical units U. The same requirement exists equally for administrative data in general. For instance, exams are part of the initial education data objects. However, while the transformation from exams (say, P) to students (say, U) can be carried out unproblematically by the school administration, such straightforward processing is often impossible from social media data objects to the target population of interest. The second extra difficulty concerns data extraction. The available measures in the administrative sources do often suffer from relevance error. Nevertheless, the actual mapping to the "Re-classified Measures" (Figure 2.4) seldom requires content or feature extraction that are necessary for social media data which, as has been discussed, is generally an additional cause of discrepancy between y_i^* and y_i or between z_j and y_i.

Finally, there seems to be currently an under-explored potential regarding the rich network relationships that can be extracted from social media data. Such network relationships may be difficult to obtain via traditional survey methods, both due to the limitations of the usual survey instruments and the relatively high cognitive and memorial requirements for correct information retrieval by the

respondents. In contrast, for network relationships that are directly observable on the social media platform, no subjective information processing will be needed and the errors associated with such processing are thereby avoided. Making greater use of the network relationships in social media data and developing suitable sampling and analysis methods appear fruitful venues forward, in order to harness the opportunities that have emerged with such big data sources.

Chapter 3

Graph sampling

We synthesise the existing theory of graph sampling. We propose a formal definition of sampling in finite graphs, and provide a classification of potential graph parameters. We develop a general approach of Horvitz-Thompson estimation to T-stage snowball sampling, and present various reformulations of some common network sampling methods in the literature in terms of the outlined graph sampling theory.

Key words: network, finite-graph sampling, multiplicity sampling, indirect sampling, adaptive cluster sampling.

3.1 Introduction

Many technological, social and biological phenomena exhibit a network structure that may be the interest of study; see e.g. Newman (2010). As an example of technological networks, consider the Internet as consisting of routers that are connected to each other via cables. There are two types of objects, namely routers and cables. A router must be connected to a cable to be included in the Internet, and a cable must have two routers at both ends. As another example, consider the social network of kinships. Again, there are two types of objects, namely persons and kinships. Each person must have two or more kinships, and each kinship must represent a connection between two persons. However, while it is obvious that any two routers must be connected by cables to each other either

directly or via other routers in the Internet, it is not sure that any two persons can be connected to each other in the network of kinships. The difference can be articulated in terms of some appropriate characterisation of their respective network structures.

Following Frank (1980, 2011), we refer to *network* as a valued graph, and *graph* as the formal structure of a network. The structure of a network, i.e. a graph, is defined as a collection of nodes and edges (between the nodes); measures may be attached to the nodes or the edges or both to form a valued graph, i.e. a network. For a statistical approach to networks one may choose to model the entire *population network* as a random realisation, or to exploit the variation over possible *sample networks* taken from a given fixed population network. Graph sampling theory deals with the structure of a network under the latter perspective. In comparison, finite-population sampling (Neyman, 1934; Cochran, 1977) can mostly be envisaged as sampling in a 'graph' with no edges at all. We shall refer to such a setting as *list* sampling.

Ove Frank has undoubtedly made the most contributions to the existing graph sampling theory. See e.g. Frank (1977c, 1979, 1980b, 1981, 2011) for his own summary. However, the numerous works of Frank scatter over several decades, and are not easily appreciable as a whole. For instance, Frank derives results for *different* samples of nodes (Frank, 1971; 1977c; 1994), dyads (Frank, 1971; 1977a; 1977b; 1979) or triads (Frank, 1971; 1979). But he never proposes a formal definition of the "sample graph" which unifies the different samples. Or, Frank studies various characteristics of a graph, such as order (Frank, 1971; 1977c; 1994), size (Frank, 1971; 1977a; 1977b; 1979), degree distribution (Frank, 1971; 1980a), connectedness (Frank, 1971; 1978), etc. But he never provides a structure of possible graph parameters which allows one to classify and contrast the different interests of study. Finally, Frank does not appear to have articulated the role of graph sampling theory in relation to some common "network sampling methods" (e.g. Birnbaum and Sirken, 1965; Thompson, 1990; Lavallée, 2007), which "are not explicitly stated as graph problems but which can be given such formulations" (Frank, 1977c).

The aim of this paper is to synthesise and extend the existing graph sampling theory, many elements of which are only implicit in Frank's works. In particular, we propose a definition of sample graph taken from a given population graph, together with the relevant observation procedures that enable sampling

in a graph (Section 3.2). In Section 3.3, we provide a structure of graph totals and graph parameters, which reflects the extended scope of investigation that can be difficult or impossible using only a list representation. Next, we develop a general approach to HT-estimation under arbitrary T-stage snowball sampling (Section 3.4). In Section 3.5, we present various graph sampling reformulations of multiplicity sampling (Birnbaum and Sirken, 1965), indirect sampling (Lavallée, 2007) and adaptive cluster sampling (Thompson, 1990), all of which are referred to as unconventional sampling methods in contrast to the more familiar finite-population sampling methods, such as stratified multi-stage sampling. Finally, some concluding remarks are given in Section 3.6, together with a couple of topics of current research.

An interactive illustration of the graph notation, as used in this paper, and of the graph sampling methods defined in section 3.2.3 can be found at the following link `http://tiny.cc/to8wpz`. To have access to the notebook, a Google account is required. Once you have clicked on it, you will be asked to switch to the playground mode to run the R code. If a warning message appears ("this notebook was not been authorized by Google"), continue by clicking on 'run anyway'.

3.2 Sampling on a graph

3.2.1 Terms and notations

A graph $G = (U, A)$ consists of a set of nodes U and a set of edges A. Define $|U| = N$ and $|A| = R$ as the *order* and *size* of G, respectively. Let $A_{ij} \subset A$ be the set of all edges from i to j; let $a_{ij} = |A_{ij}|$ be its size. If $a_{ij} > 1$ for some $i, j \in U$, the graph is called a multigraph; it is a simple graph if $a_{ij} = 0, 1$. The edges in $A_{i+} = \bigcup_{j \in U} A_{ij}$ and $A_{+i} = \bigcup_{j \in U} A_{ji}$ are called the outedges and inedges at i, respectively. Let $a_{i+} = |A_{i+}| = \sum_{j \in U} a_{ij}$ and $a_{+i} = |A_{+i}| = \sum_{j \in U} a_{ji}$. The node i is *incident* to each outedge or inedge at i. The number of edges *incident* at a node i is called the degree of i, denoted by $d_i = a_{i+} + a_{+i}$. Two nodes i and j are *adjacent* if there exists at least one edge between them, i.e. $a_{ij} + a_{ji} > 1$. For any edge in A_{ij}, i is called its initial node and j its terminal node. Let α_i be the *successors* of i, which are the terminal nodes of outedges at i; let β_i be the *predecessors* of i, which are the initial nodes of inedges at i. For a simple graph, we have $a_{i+} = |\alpha_i|$ and $a_{+i} = |\beta_i|$. A graph is said to be directed (i.e. a *digraph*) if $A_{i+} \neq A_{+i}$; it is undirected if $A_{i+} = A_{+i}$, in which case there is no distinction

between outedge and inedge, so that $d_i = a_{i+} = a_{+i}$, and $\alpha_i = \beta_i$. Finally, an edge a_{ii} connecting the same node i is called a *loop*, which can sometimes be a useful means of representation. Whether or not loops are included in the definitions of the terms and notations above is purely a matter of convention.

Remark Adjacency refers to relationship between nodes, as objects of the same kind; incidence refers to relationship between nodes and edges, i.e. objects of different kinds.

Remark Let the $N \times N$ *adjacency matrix* $\mathbf{A}$ have elements $a_{ij} = |A_{ij}|$. It is defined to be symmetric for undirected graphs. Put the diagonal degree matrix $\mathbf{D} = \text{diag}(\mathbf{A}1_{N\times 1})$. The Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{A}$ sums to 0 by row and column, which is of central interest in Spectral Graph Theory (e.g. Chung, 1997).

3.2.2 Definition of sample graph

Denote by s_1 an *initial sample* of nodes, for $s_1 \subseteq U$. Under a probability design, let π_i and π_{ij} (or $\bar{\pi}_i$ and $\bar{\pi}_{ij}$) be the probabilities of inclusion (or exclusion) of respectively a node and a pair of nodes in s_1. (The exclusion probability of i is the probability of $i \notin s_1$, and the exclusion probability of a pair (i, j) is the probability that neither i nor j is in s_1.) A defining feature of sampling on graphs is that one makes use of the edges to select the *sample graph*, denoted by G_s. Given s_1, the relevant nodes are either in $\alpha(s_1) = \bigcup_{i\in s_1} \alpha_i$ or $\beta(s_1) = \bigcup_{i\in s_1} \beta_i$, where $\alpha(s_1) = \beta(s_1)$ for undirected graphs. An *observation procedure* of the edges needs to be specified, and the observed edges can be given in terms of a *reference set* of node pairs, denoted by s_2 where $s_2 \subseteq U \times U$, under the convention that the set of edges A_{ij} are observed whenever $(ij) \in s_2$. Notice that generally speaking (ij) and (ji) are considered as two distinct elements in $U \times U$. Denote by $\pi_{(ij)}$ (or $\bar{\pi}_{(ij)}$) the corresponding inclusion (or exclusion) probability of $(ij) \in s_2$, and by $\pi_{(ij)(kl)}$ (or $\bar{\pi}_{(ij)(kl)}$) the inclusion (or exclusion) probability of these two pairs in s_2. Denote by $A_s = A(s_2)$ the edge set inherent of s_2, and $U_s = s_1 \cup \text{Inc}(A_s)$ the union of s_1 and those nodes that are incident to A_s. The sample graph is $G_s = \big(U_s, A_s\big)$.

Example 1 Let $U = \{1, 2, 3\}$, and $a_{12} = 1$. Suppose $s_1 = \{1\}$. Provided $s_2 = s_1 \times \alpha(s_1)$, where $\alpha(s_1) = \{2\}$ in this case, the sample graph G_s has $A_s =$

$A(s_2) = A_{12}$ and $U_s = \{1, 2\}$. The same sample graph can equally be given by $s_2' = s_1 \times U$, since $A(s_2') = A_{12} = A(s_2)$.

Observation procedure Frank (1977c) considers several observation procedures, which can be formalised as follows. First, given s_1, a procedure is *induced* if A_{ij} is observed iff both $i \in s_1$ and $j \in s_1$, or *incident reciprocal* if A_{ij} and A_{ji} are both observed provided either $i \in s_1$ or $j \in s_1$. Second, for digraphs, an *incident non-reciprocal* procedure is *forward* if A_{ij} is observed provided $i \in s_1$, or *backward* if A_{ij} is observed provided $j \in s_1$. For example, provided $i \in s_1$ and $j \notin s_1$ and $a_{ij} > 0$ and $a_{ji} > 0$, we would observe both A_{ij} and A_{ji} given an incident reciprocal procedure; only A_{ij} if it is incident forward; only A_{ji} if it is incident backward; neither A_{ij} nor A_{ji} given an induced procedure from s_1.

Initial sampling of edges Sample graph initiated by a sample of edges can be defined analogously. Bernoulli or Poisson sampling can be useful, because it is not required to know all the edges in advance. Notice that when one is interested in the totals or other functions of the edges of a graph, initial Bernoulli or Poisson sampling of edges is meaningful – see e.g. Frank (1977c, Section 12), whereas initial simple random sampling (of edges) would have been a trivial set-up, because one would need to know all the edges to start with.

3.2.3 Some graph sampling methods

We describe some sampling methods based on the aforementioned observation procedures. Frank (1977c) elicited several sampling methods based on the aforementioned observation procedures. We include several alternative specifications which are marked by †. By way of introduction, the first- and second-order inclusion probabilities of (ij) in s_2 are given in terms of the relevant inclusion probabilities in s_1, which facilitates Horvitz-Thompson (HT) estimation of any totals defined on $U \times U$. As will be illustrated, given s_1 and the observation procedure, the sample graph can be specified using different reference sets s_2, but the inclusion probabilities are more readily obtained for some choices of s_2.

(i) $s_2 = s_1 \times s_1$ [Induced]: Both $(ij) \in s_2$ and $(ji) \in s_2$ iff $i \in s_1$ and $j \in s_1$. Then, $\pi_{(ij)} = \pi_{ij}$ and $\pi_{(ij)(kl)} = \pi_{ijkl}$.

(ii.1) $s_2 = s_1 \times s_a$, $s_a = \alpha(s_1) \cup s_1$ [Incident forward]: $(ij) \in s_2$ iff $i \in s_1$ and $j \in s_a$. Let $B_j = \{j\} \cup \beta_j$, i.e. itself and its predecessors, then $j \in s_a$ iff $B_j \cap s_1 \neq \emptyset$. Thus,

$$\bar{\pi}_{(ij)} = \bar{\pi}_i + \bar{\pi}_{B_j} - \bar{\pi}_{B_j \cup \{i\}}.$$

Similarly, $(ij), (kl) \in s_2$ iff $i, k \in s_1$ and $B_j \cap s_1 \neq \emptyset$ and $B_l \cap s_1 \neq \emptyset$, so that

$$\begin{aligned}\bar{\pi}_{(ij)(kl)} = \bar{\pi}_{ik} + \bar{\pi}_{B_j \cup \{k\}} + \bar{\pi}_{B_l \cup \{i\}} + \bar{\pi}_{B_j \cup B_l} \\ - \bar{\pi}_{B_j \cup \{i,k\}} - \bar{\pi}_{B_l \cup \{i,k\}} - \bar{\pi}_{B_j \cup B_l \cup \{i\}} - \bar{\pi}_{B_j \cup B_l \cup \{k\}} + \bar{\pi}_{B_j \cup B_l \cup \{i,k\}}.\end{aligned}$$

(ii.2) $s_2 = s_1 \times U$ [Incident forward]: $(ij) \in s_2$ iff $i \in s_1$. Then, $\pi_{(ij)} = \pi_i$ and $\pi_{(ij)(kl)} = \pi_{ik}$.

Remark The sample edge set $A(s_2)$ is the *same* in (ii.2) and (ii.1), because the observation procedure is the same given s_1. For the estimation of any total over A, the two reference sets would yield the same HT-estimate: any $(ij) \in s_2$ with $a_{ij} = 0$ does not contribute to the estimate, regardless of its $\pi_{(ij)}$; whereas for any $(ij) \in s_2$ with $a_{ij} > 0$, we have $\pi_{(ij)} = \pi_i$ given s_2 in (ii.2), just as one would have obtained in (ii.1) since $B_j = B_j \cup \{i\}$ provided $a_{ij} > 0$. But it appears easier to arrive at $\pi_{(ij)}$ and the HT-estimator in (ii.2) than (ii.1).

(ii.3)† $s_2 = s_b \times \alpha(s_1)$, $s_b = s_1 \cap \beta\big(\alpha(s_1)\big)$ [Incident forward]: This is the smallest Cartesian product that contains the same sample edge set as in (ii.1) and (ii.2).

(ii.4)† $s_2 = \bigcup_{i \in s_1} i \times \alpha_i$, where $i \times \alpha_i = \emptyset$ if $\alpha_i = \emptyset$ [Incident, forward]: Only (ij) with $a_{ij} > 0$ is included in s_2. This is the smallest reference set for the same G_s in (ii.1) - (ii.4).

(iii) $s_2 = s_a \times s_a$, $s_a = \alpha(s_1) \cup s_1$ [Induced from s_a]: $(ij) \in s_2$ even if $i \in s_a \setminus s_1$ and $j \in s_a \setminus s_1$. Similarly to (ii.1), $(ij) \in s_2$ iff $B_i \cap s_1 \neq \emptyset$ and $B_j \cap s_1 \neq \emptyset$, and so on. Then,

$$\begin{aligned}\bar{\pi}_{(ij)} &= \bar{\pi}_{B_i} + \bar{\pi}_{B_j} - \bar{\pi}_{B_i \cup B_j}, \\ \bar{\pi}_{(ij)(kl)} &= \bar{\pi}_{B_i \cup B_k} + \bar{\pi}_{B_i \cup B_l} + \bar{\pi}_{B_j \cup B_k} + \bar{\pi}_{B_j \cup B_l} \\ &\quad - \bar{\pi}_{B_i \cup B_k \cup B_l} - \bar{\pi}_{B_j \cup B_k \cup B_l} - \bar{\pi}_{B_i \cup B_j \cup B_k} - \bar{\pi}_{B_i \cup B_j \cup B_l} + \bar{\pi}_{B_i \cup B_j \cup B_k \cup B_l}.\end{aligned}$$

Remark Observation of the edges between $i \in s_a \setminus s_1$ and $j \in s_a \setminus s_1$ may be demanding in practice, even when the observation procedure is reciprocal. For example, let the node be email account. Then, by surveying $i \in s_1$ only, it is possible to observe all the email accounts that have exchanges with i due to reciprocality. But one would have to survey the accounts in $\alpha_i \setminus s_1$ additionally, in order to satisfy the requirement of (iii).

(iv.1) $s_2 = s_1 \times U \cup U \times s_1$ [Incident reciprocal]: $(ij) \notin s_2$ iff $i \notin s_1$ and $j \notin s_1$. Then, $\pi_{(ij)} = 1 - \bar{\pi}_{ij}$ and $\pi_{(ij)(kl)} = 1 - \bar{\pi}_{ij} - \bar{\pi}_{kl} + \bar{\pi}_{ijkl}$.

(iv.2)† $s_2 = s_1 \times s_a \cup s_a \times s_1$, $s_a = \alpha(s_1) \cup s_1$ [Incident reciprocal]: We have $s_a \times s_a = s_2 \cup (s_a \setminus s_1) \times (s_a \setminus s_1)$, where the two sets on the right-hand side are disjoint. The inclusion probabilities can thus be derived from those in (iii) and those of $(s_a \setminus s_1) \times (s_a \setminus s_1)$. However, the sample edge set $A(s_2)$ is the same as in (iv.1), and it is straightforward to derive the HT-estimator of any total over A based on the reference set s_2 specified in (iv.1).

(iv.3)† $s_2 = \big(\bigcup_{i \in s_1} i \times \alpha_i \big) \cup \big(\bigcup_{i \in s_1} \beta_i \times i \big)$ [Incident reciprocal]: This is the smallest reference set of the sample edge set in (iv.1) - (iv.3).

Example 2 Figure 3.1 illustrates the four sampling methods (i) - (iv) described above, all of which are based on the same initial sample $s_1 = \{3, 6, 10\}$.

3.3 Graph parameter and HT-estimation

Frank (1980b) reviews some statistical problems based on population graphs. In a list representation, the target population U is a collection of elements, which are associated with certain values of interest. In a graph representation $G = (U, A)$, the elements in U can be the nodes that have relations to each other, which are presented by the edges in A. It becomes feasible to investigate the interactions between the elements, their structural positions, etc. which are difficult or unnatural using a list representation. The extended scope of investigation is above all reflected in the formulation of the target parameter. In this Section, we provide our own classification of the potential target parameters based on a graph in terms of graph totals and graph parameters.

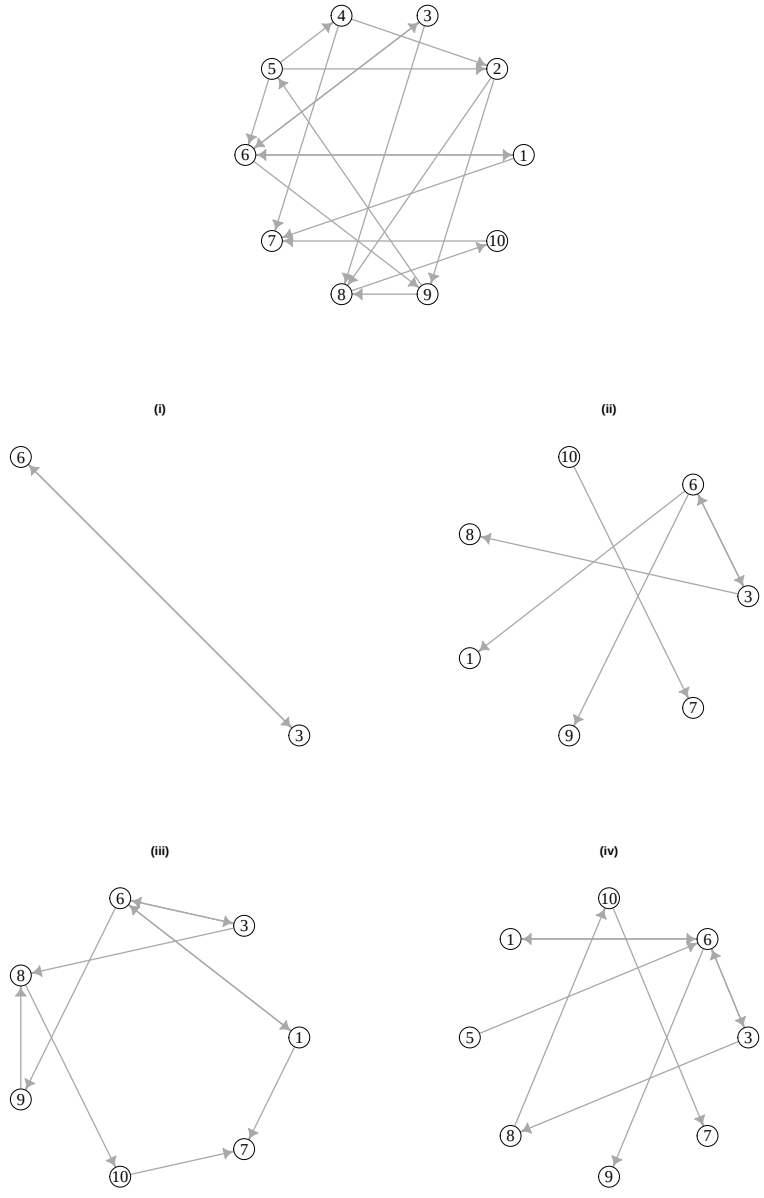


Figure 3.1: Population graph (top) and four sample graphs (i) - (iv) based on $s_1 = \{3, 6, 10\}$.

Graph total and graph parameter Let M_k be a subset of U, where $|M_k| = k$. Let $\mathcal{C}_k$ be the set of all possible M_k's, where $|\mathcal{C}_k| = N![k!(N-k)!]^{-1}$. Let $G(M_k)$ be the subgraph induced by M_k. Let $y\big(G(M_k)\big)$, or simply $y(M_k)$, be a function of integer or real value. The corresponding *k-th order graph total* is given by

$$\theta = \sum_{M_k \in \mathcal{C}_k} y(M_k). \tag{3.1}$$

We refer to functions of graph totals as *graph parameters*.

Remark *Network totals* can as well be defined by (3.1), where $y(\cdot)$ can incorporate the values associated with the nodes and edges of the induced subgraph $G(M_k)$.

Motif A subset $M \subset U$ with specific characteristics is said to be a *motif*, denoted by $[M]$. For example, denote by $[i : d_i = 3]$ a 1st-order motif, i.e. a node with degree 3. Or, denote by $[i, j : a_{ij} = a_{ji} = 1]$ the motif of a pair of nodes with mutual simple relationship, or by $[i, j : a_{ij} = a_{ji} = 0]$ the motif of a pair of non-adjacent nodes. A motif may or may not have a specific order, giving rise to graph totals with or without given orders.

3.3.1 Graph totals of a given order

3.3.1.1 First-order graph total: $M_1 = \{i\}$

Each M_1 corresponds to a node. In principle any first-order graph total can be dealt with by a list sampling method that does not make use of the edges, against which one can evaluate the efficiency of any graph sampling method. For the two parameters given below, estimation of the order by snowball sampling is considered by Frank (1971, 1977c, 1994), and estimation of the degree distribution is considered by Frank (1971, 1980a).

Order (of G) Let $y(i) \equiv 1$, for $i \in U$. Then, $\theta = |U| = N$.

Number of degree-d nodes Let $y(i) = \delta(d_i = d)$ indicate whether or not d_i equals d, for $i \in U$. Then, θ is the number of nodes with degree d.

3.3.1.2 Second-order graph total: $M_2 = \{i, j\}$

An M_2 of a pair of nodes is called a dyad, for $M_2 \subset U$ and $|M_2| = 2$. Some dyad totals are considered by Frank (1971, 1979).

Size (of G) Let $y(M_2) = a_{ij} + a_{ji}$ be the adjacency count between i and j in a digraph, or $y(M_2) = a_{ij}$ for an undirected graph. Then, $\theta = \sum_{M_2 \in \mathcal{C}_2} y(M_2) = R$ is the size (of G).

Remark If there are loops, one can let $y(M_1) = a_{ii}$ for $M_1 = \{i\}$, and $\theta' = \sum_{M_1 \in \mathcal{C}_1} y(M_1)$. Then, $R = \theta + \theta'$ is a graph parameter based on a 1st- and a 2nd-order graph totals.

Remark Let N_d be the no. degree-d nodes, which is a 1st-order graph total. Then,

$$2R = \sum_{i \in U} d_i = \sum_{d=1}^{D} dN_d, \qquad \text{where} \quad D = \max_{i \in U} d_i.$$

This is an example where a higher-order graph total (R) can be 'reduced' to lower-order graph parameters (N_d). Such reduction can potentially be helpful in practice, e.g. when it is possible to observe the degree of a sample node without identifying its successors.

Number of adjacent pairs Let $y(M_2) = \delta(a_{ij} + a_{ji} > 0)$ indicate whether i and j are adjacent. Then, θ is the total number of adjacent pairs in G. Its ratio to $|\mathcal{C}_2|$ provides a graph parameter, i.e. an index of *immediacy* in the graph. Minimum immediacy is the case when a graph consists of only isolated nodes, and maximum immediacy if the graph is a *clique*, where every pair of distinct nodes are adjacent with each other.

Number of mutual relationships Let $y(M_2) = \delta(a_{ij}a_{ji} > 0)$ indicate whether i and j have reciprocal edges between them, in which case their relationship is *mutual*. Then, θ is the number of mutual relationships in the graph. Goodman (1961) studies the estimation of the number of mutual relationships in a special digraph, where $a_{i+} = 1$ for all $i \in U$.

3.3.1.3 Third-order graph total: $M_3 = \{i, j, h\}$

An M_3 of three nodes is called a triad, for $M_3 \subset U$ and $|M_3| = 3$. Some triad totals are considered by Frank (1971, 1977a, 1977b, 1979).

Number of triads Let $y(M_3) = \delta(a_{ij}a_{jh}a_{ih} > 0)$ indicate whether the three nodes form a triangle in an undirected graph. Then, θ^* by (4.1) is the total number of triangles. Triangles on undirected graphs are intrinsically related to equivalence relationships: for a relationship (represented by an edge) to be transitive, every pair of connected nodes must be adjacent; hence, any three connected nodes must form a triangle. For a simple undirected graph, transitivity is the case iff $\theta' = 0$, when θ' is given by (4.1), where

$$y(M_3) = a_{ij}a_{jh}(1 - a_{hi}) + a_{ih}a_{jh}(1 - a_{ij}) + a_{ij}a_{ih}(1 - a_{jh}).$$

Provided this is not the case, one can e.g. still measure the extent of transitivity by

$$\tau = \theta^*/(\theta^* + \theta'),$$

i.e. a graph parameter. Next, for digraphs and ordered (jih), let $z(jih) = a_{ji}a_{ih}a_{hj}$ be the count of *strongly connected* triangles from j via i and h back to j. Let $\widetilde{M}_3$ contain all the possible orderings of M_3, i.e. (ijh), (ihj), (jih), (jhi), (hij) and (hji). Then, the number of strongly connected triangles in a digraph is given by (4.1), where

$$y(M_3) = \sum_{(ijh)\in\widetilde{M}_3} z(ijh).$$

Remark For undirected simple graphs, Frank (1981) shows that there exists an explicit relationship between the mean and variance of the degree distribution and the triads of the graph. Let the numbers of triads of respective size 3, 2 and 1 be given by

$$\theta_{3,3} = \sum_{M_3\in\mathcal{C}_3} a_{ij}a_{jh}a_{ih},$$
$$\theta_{3,2} = \sum_{M_3\in\mathcal{C}_3} a_{ij}a_{ih}(1 - a_{jh}) + a_{ij}a_{jh}(1 - a_{ih}) + a_{ih}a_{jh}(1 - a_{ij}),$$
$$\theta_{3,1} = \sum_{M_3\in\mathcal{C}_3} a_{ij}(1 - a_{jh})(1 - a_{ih}) + a_{ih}(1 - a_{ij})(1 - a_{jh}) + a_{jh}(1 - a_{ij})(1 - a_{ih}).$$

Let $\mu = \sum_{d=1}^{N} dN_d/N = 2R/N$ and $\sigma^2 = Q/N - \mu^2$, where $Q = \sum_{d=1}^{N} d^2 N_d$. We have

$$R = \frac{1}{N-2}\big(\theta_{3,1} + 2\theta_{3,2} + 3\theta_{3,3}\big), \qquad Q = \frac{2}{N-1}\big(\theta_{3,1} + N\theta_{3,2} + 3(N-1)\theta_{3,3}\big).$$

3.3.2 Graph totals of unspecified order

A motif is sometimes defined in an order-free manner. Insofar as the corresponding total can be given as a function of graph totals of specific orders, it can be considered a graph parameter. Below are some examples that are related to the connectedness of a graph. The number of connected components is considered by Frank (1971, 1978).

Number of connected components The subgraph induced from M_k is a connected component of order k, provided there exists a path for any $i \neq j \in M_k$ and $a_{ij} = a_{ji} = 0$ for any $i \in M_k$ and $j \notin M_k$, in which case let $y(M_k) = 1$ but let $y(M_k) = 0$ otherwise. Then, θ_k given by (4.1) is the number of connected components of order k. The number of connected components (i.e. as a motif of unspecified order) is the graph parameter given by $\theta = \sum_{k=1}^{N} \theta_k$. At one end, where $A = \emptyset$, i.e. there are no edges at all in the graph, we have $\theta = N = \theta_1$ and $\theta_k = 0$ for $k > 1$. At the other end, where there exists a path between any two nodes, we have $\theta = \theta_N = 1$ and $\theta_k = 0$ for $k < N$.

Number of trees in a forest In a simple graph, a motif $[M_k]$ is a *tree* if the number of edges in $G(M_k)$ is $k - 1$. As an example where θ can be reduced to a specific graph total, suppose the undirected graph is a forest, where every connected component is a tree. We have then $\theta = N - R$, where R is the size of the graph, which is a 2nd-order parameter.

Number of cliques A clique is a connected component, where there exists an edge between any two nodes of the component. It is a motif of unspecified order. The subgraph induced by a clique is said to be complete. A clustered population can be represented by a graph, where each cluster of population elements (i.e. nodes) form a clique, and two nodes i and j are adjacent iff the two belong to the same cluster.

Index of demographic mobility Given the population of a region (U), let there be an undirected edge between two persons i and j if their family trees intersect, say, within the last century, i.e. they are relatives of each other within a 'distance' of 100 years. Each connected component in this graph G is a clique. The ratio between the number of connected components θ and N, where N is the maximum possible θ, provides an index of demographic mobility that varies between $1/N$ and 1. Alternatively, an index can be given by the ratio between the number of edges R and $|\mathcal{C}_2|$, which varies between 0 and 1, and is a function of a 2nd-order graph total. This is an example where the target parameter can be specified in terms of a lower-order graph total than higher-order totals.

Remark In the context of estimating the number of connected components, Frank (1971) discusses the situation where observation is obtained about whether a pair of sample nodes are connected in the graph, without necessarily including the paths between them in the sample graph. The observation feature is embedded in the definition of the graph here.

Geodesics in a graph Let an undirected graph G be connected, i.e. $U = M_N$ is a connected component. The geodesic between nodes i and j is the shortest path between them, denoted by $[M_k]$, where M_k contains the nodes on the geodesic, including i and j. A geodesic $[M_k]$ is a motif of order k, whereas geodesic is generally a motif of unspecified order. Let θ be the harmonic mean of the length of the geodesics in G, which is a closeness centrality measure (Newman, 2010). For instance, it is at its minimum value 1 if G is complete. Alternatively, let $y(M_2) = 1/(k-1)$, provided $[M_k]$ is the geodesic between i and j, so that θ can equally be given as a 2nd-order graph parameter. Again, this is an example where a lower-order graph parameter can be used as the target parameter instead of alternatives involving higher-order graph totals, provided the required observation.

3.3.3 HT-estimation

A basic estimation approach in graph sampling is the HT-estimator of a graph total (4.1). Provided the inclusion probability $\pi_{(M_k)}$ for $M_k \in \mathcal{C}_k$, the HT-estimator is given by

$$\hat{\theta} = \sum_{M_k \in \mathcal{C}_k} \delta_{[M_k]} y(M_k)/\pi_{(M_k)}, \tag{3.2}$$

where $\delta_{[M_k]} = 1$ if $[M_k]$ is observed and $\pi_{(M_k)}$ is its inclusion probability. The observation of $[M_k]$ means not only $M_k \subseteq U_s$, but also it is possible to identify whether M_k is a particular motif in order to compute $y(M_k)$. The probability $\pi_{(M_k)}$ is defined with respect to a chosen reference set s_2 and the corresponding sample graph G_s. It follows that a motif $[M_k]$ is observed in G_s if $M_k \subseteq U_s$ and $M_k \times M_k \subseteq s_2$. More detailed explanation of $\pi_{(M_k)}$ will be given in Section 3.4. The example below illustrates the idea.

Example 3 Consider an undirected simple graph. Let 3-node star be the motif of interest, and $y(M_3) = a_{ij}a_{ih}(1 - a_{jh}) + a_{ij}a_{jh}(1 - a_{ih}) + a_{ih}a_{jh}(1 - a_{ij})$ the corresponding indicator. Suppose incident observation and $s_2 = s_1 \times U$. Consider $M_3 = \{i, j, h\} \subset U_s$. To be able to identify whether it is the motif of interest, all the three pairs (ij), (ih) and (jh) need to be in s_2; accordingly, $\pi_{(M_3)} = \Pr\big((ij) \in s_2, (ih) \in s_2, (jh) \in s_2\big)$. An example where this is not the case is $i \in s_1$ and $j, h \in \alpha(s_1) \setminus s_1$, so that the observed part of this triad is a star, but one cannot be sure if $a_{jh} = 0$ in the population graph, because $(jh) \notin s_2$.

Symmetric designs The inclusion probability $\pi_{(M_k)}$ depends on the sampling design of initial s_1. At various places, Frank consider simple random sampling (SRS) without replacement, Bernoulli sampling and Poisson sampling for selecting the initial sample. In particular, a sampling design is *symmetric* (Frank, 1977a) if the inclusion probability $\pi_{M_k} = \Pr(M_k \in s_1)$ only depends on k but is a constant of M_k, for all $1 \leq k \leq N$. SRS with or without replacement and Bernoulli sampling are all symmetric designs. SRS without replacement is the only symmetric design with fixed sample size of distinct elements.

Approximate approach The initial inclusion probability π_{M_k} has a simpler expression under Bernoulli sampling than under an SRS design. Provided negligible sampling fraction of s_1, many authors use Bernoulli sampling with probability $p = |s_1|/N$ to approximate any symmetric designs. Similarly, initial unequal probability sampling may be approximated by Poisson sampling with the same π_i, for $i \in U$, provided negligible sampling fraction $|s_1|/N$. Finally, Monte Carlo simulation (Fattorini, 2006) may be used to approximate the relevant π_{M_k} under sampling without replacement.

3.4 T-stage snowball sampling

An incident observation procedure (Section 3.2.3) provides the means to enlarge a set of sample nodes by their out-of-sample adjacent nodes. It yields a method of 1-stage snowball sampling, which can be extended successively to yield the T-stage snowball sampling. Below we assume that all the successors are included in the sample. But it is possible to take only some of the successors at each stage (e.g. Snijders, 1992). In particular, taking one successor each time yields a T-stage walk (e.g. Klovdahl, 1989). Two different observation procedures will be considered, i.e. incident forward in digraphs and incident reciprocal in directed or undirected graphs. We develop general formulae for inclusion probabilities under T-stage snowball sampling. It is shown that additional observation features are necessary for the HT-estimator based on T-stage snowball sampling, which will be referred to as *incident ancestral*. Previously, Goodman (1961) has studied the estimation of mutual relationships between i and j, where $a_{ij}a_{ji} > 0$ for $i \neq j \in U$, based on T-stage snowball sampling in a special digraph with fixed $a_{i+} \equiv 1$; Frank (1977c) and Frank and Snijders (1994) considered explicitly HT-estimation based on 1-stage snowball sampling.

Sample graph $G_s = (U_s, A_s)$ Let $s_{1,0}$ be the initial sample of *seeds*, and $\alpha(s_{1,0})$ its successors. Let $U_0 \subseteq U$ be the set of possible initial sample nodes. The additional nodes $s_{1,1} = \alpha(s_{1,0}) \setminus s_{1,0}$ are called the first-wave snowball sample, which are the seeds of the second-wave snowball sample, and so on. At the t-th stage, let $s_{1,t} = \alpha(s_{1,t-1}) \setminus \big(\bigcup_{h=0}^{t-1} s_{1,h} \big)$ be the t-th stage seeds, for $t = 1, 2, ..., T$. If $s_{1,t} = \emptyset$, set $s_{1,t+1} = \cdots = s_{1,T} = \emptyset$ and terminate, otherwise move to stage $t+1$. Let $s_1 = \bigcup_{t=0}^{T-1} s_{1,t}$ be the *sample of seeds*. This may result in two different sample graphs.

I. Let $s_2 = s_1 \times U$ provided incident forward observation in digraphs, such that the sample graph G_s has edge set $A_s = \bigcup_{i \in s_1} \bigcup_{j \in \alpha_i} A_{ij}$ and node set $U_s = s_1 \cup \alpha(s_1)$.

II. Let $s_2 = s_1 \times U \cup U \times s_1$ provided incident reciprocal observation, digraphs or not, such that G_s has edge set $A_s = \bigcup_{i \in s_1} \bigcup_{j \in \alpha_i} (A_{ij} \cup A_{ji})$ and node set $U_s = s_1 \cup \alpha(s_1)$.

Remark One may disregard any loops in snowball sampling, because they do not affect the propagation of the waves of nodes, but only cause complications to their definition.

3.4.1 Inclusion probabilities of nodes and edges in G_s

Below we develop the inclusion probabilities $\pi_{(i)}$ and $\pi_{(i)(j)}$ of nodes in U_s, and $\pi_{(ij)}$ and $\pi_{(ij)(hl)}$ of edges in A_s, under T-stage snowball sampling with s_2 as specified above.

Forward observation in digraphs The stage-specific seed samples $s_{1,0}, ..., s_{1,T-1}$ are disjoint, so that each observed edge, denoted by $\langle ij \rangle \in A_s$, can only be included at a particular stage. For $i \in U$, let $\beta_i^{[0]} = U_0 \cap \{i\}$; let $\beta_i^{[t]} = U_0 \cap \Big(\beta(\beta_i^{[t-1]}) \setminus \big(\bigcup_{h=0}^{t-1} \beta_i^{[h]}\big)\Big)$ be its *t-th generation* predecessors, for $t > 0$, which consists of the nodes that would lead to i in t-stages from $s_{1,0}$ but not sooner. Notice that $\beta_i^{[0]}, \beta_i^{[1]}, \beta_i^{[2]}, ...$ are disjoint. We have

$$\pi_{(i)} = 1 - \bar{\pi}_{B_i} \qquad \text{for} \quad B_i = \bigcup_{t=0}^{T} \beta_i^{[t]},$$

$$\pi_{(ij)} = 1 - \bar{\pi}_{B_{ij}} \qquad \text{for} \quad B_{ij} = \bigcup_{t=0}^{T-1} \beta_i^{[t]}.$$

The respective joint inclusion probabilities follow as $\pi_{(i)(j)} = 1 - \bar{\pi}_{B_i} - \bar{\pi}_{B_j} + \bar{\pi}_{B_i \cup B_j}$ and $\pi_{(ij)(hl)} = 1 - \bar{\pi}_{B_{ij}} - \bar{\pi}_{B_{hl}} + \bar{\pi}_{B_{ij} \cup B_{hl}}$.

Incident reciprocal observation Each $\langle ij \rangle \in A_s$ can only be included at a particular stage, where either i or j is in the seed sample, regardless if the graph is directed or not. For $i \in U$, let $\eta_i = \{j \in U; a_{ij} + a_{ji} > 0\}$ be the set of its adjacent nodes. Let $\eta_i^{[0]} = U_0 \cap \{i\}$; let $\eta_i^{[t]} = U_0 \cap \Big(\eta(\eta_i^{[t-1]}) \setminus \big(\bigcup_{h=0}^{t-1} \eta_i^{[h]}\big)\Big)$ be its *t-th step* neighbours, for $t > 0$, which are the nodes that would lead to i in t-stages from $s_{1,0}$ but not sooner. We have

$$\pi_{(i)} = 1 - \bar{\pi}_{R_i} \qquad \text{for} \quad R_i = \bigcup_{t=0}^{T} \eta_i^{[t]}, \tag{3.3}$$

$$\pi_{(ij)} = 1 - \bar{\pi}_{R_{ij}} \qquad \text{for} \quad R_{ij} = \bigcup_{t=0}^{T-1} \eta_i^{[t]} \cup \eta_j^{[t]}. \tag{3.4}$$

The respective joint inclusion probabilities follow as $\pi_{(i)(j)} = 1 - \bar{\pi}_{R_i} - \bar{\pi}_{R_j} + \bar{\pi}_{R_i \cup R_j}$ and $\pi_{(ij)(hl)} = 1 - \bar{\pi}_{R_{ij}} - \bar{\pi}_{R_{hl}} + \bar{\pi}_{R_{ij} \cup R_{hl}}$.

Incident ancestral observation procedure It is thus clear that additional features of the observation procedure is required in order to calculate $\pi_{(i)}$ and $\pi_{(i)(j)}$ given any $T \geq 1$, or $\pi_{(ij)}$ and $\pi_{(ij)(hl)}$ given any $T \geq 2$. Reciprocal or not, an incident procedure is said to be *ancestral* in addition, if one is able to observe all the nodes that would lead to the inclusion of a node $i \in U_s$, which will be referred to as its *ancestors*. These are the predecessors of various generations for forward observation in digraphs, or the neighbours of various steps for reciprocal observation in directed or undirected graphs. Notice that the edges connecting the sample nodes in U_s and their out-of-sample ancestors are *not* included in the sample graph G_s. More comments regarding the connections between snowball sampling and some well-known network sampling methods will be given in Section 3.5.

Remark Frank (1971) defines the *reach* at i as the order of the connected component containing node i. The requirement of observing the reach, without including the whole connected component in the sample graph, is similar to that of an ancestral observation procedure, even though the two are clearly different.

Example 4 To illustrate the inclusion probabilities (3.3) and (3.4), consider the population graph $G = (U, A)$, and a sample graph $G_s = (U_s, A_s)$ by 2-stage snowball sampling, with the initial sample $s_{1,0} = \{3, 4\}$ by SRS with sample size 2. The 1st- and 2nd-wave snowball samples are $s_{1,1} = \{8, 9, 10\}$ and $s_{1,2} = \{1, 5, 7\}$, respectively. The sample of seeds is $s_1 = \{3, 4, 8, 9, 10\}$. Both G and G_s are given in Figure 3.2. To the left of Figure 3.3, the true node inclusion probabilities $\pi_{(i)}$ are plotted against those given by (3.3), where there are 5 distinct values; to the right, the true edge inclusion probabilities $\pi_{(ij)}$ are plotted against those given by (3.4), where there are 4 distinct values. In both cases, the true inclusion probabilities are calculated directly over the 45 possible initial samples of size 2.

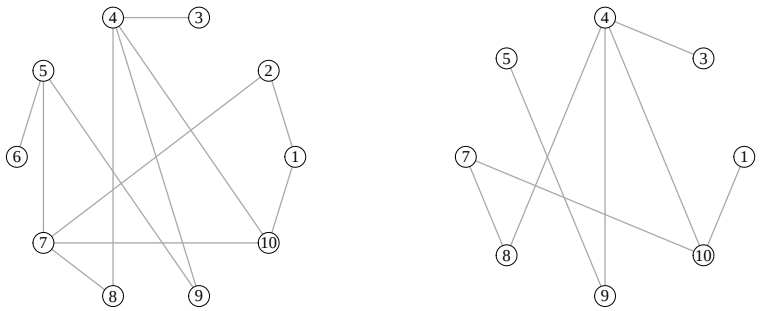


Figure 3.2: Population graph G with 10 nodes and 11 edges (left), a sample graph G_s by 2-stage snowball sampling starting from $s_{1,0} = \{3, 4\}$ by simple random sampling (right).

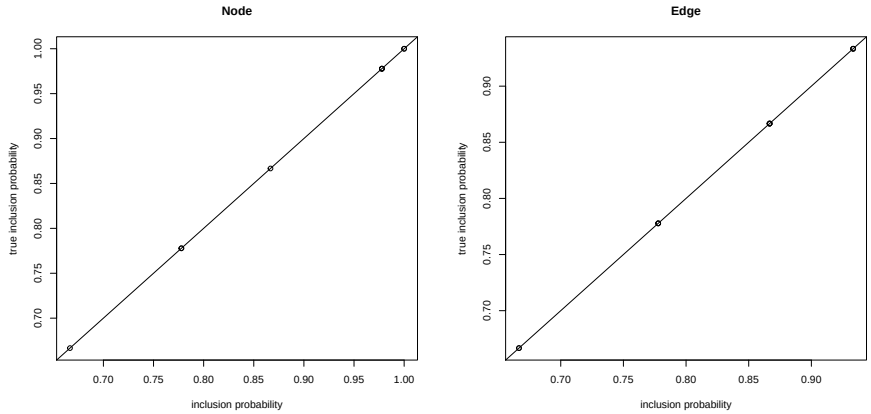


Figure 3.3: Inclusion probability $\pi_{(i)}$: true vs. (3.3), left; $\pi_{(ij)}$: true vs. (3.4), right.

3.4.2 Arbitrary M_k with $k \geq 2$ and $s_2 = s_1 \times U \cup U \times s_1$

To fix the idea, consider incident reciprocal observation in directed or undirected graphs. Notice that one can as well let $s_2 = s_1 \times U$ in the case of undirected graphs.

Definition of $\pi_{(M_k)}$ for $M_k \subset U$ To be clear, write $\{i_1, i_2, ..., i_k\}$ for $M_k \subset U$. Let $M_k^{(h)} = M_k \setminus \{i_h\}$ be the subset obtained by dropping i_h from M_k, for $h = 1, ..., k$. As explained in Section 4.6, to be able to identify the motif $[M_k]$, there can be at most one node in M_k that belongs to the last wave of snowball sample $(s_{1,T})$. In other words, at least one of the k subsets $M_k^{(h)}$ must be in the sample

of seeds s_1. We have

$$\pi_{(M_k)} = \Pr\Big(M_k^{(1)} \subseteq s_1 \text{ or } M_k^{(2)} \subseteq s_1 \text{ or } \cdots \text{ or } M_k^{(k)} \subseteq s_1 \text{ or } M_k \subseteq s_1\Big)$$
$$= \sum_{h=1}^{k} \Pr\Big(M_k^{(h)} \subseteq s_1\Big) - (k-1)\Pr\Big(M_k \subseteq s_1\Big), \tag{3.5}$$

where $\Pr\big(M_k \subseteq s_1\big) = \pi_{(i_1)(i_2)\cdots(i_k)}$ is joint inclusion probability of the relevant nodes in s_1, similarly for $\Pr\big(M_k^{(h)} \subseteq s_1\big)$, where $h = 1, ..., k$. The expression (3.5) follows from noting $\{M_k^{(h)} \subseteq s_1\} \cap \{M_k \subseteq s_1\} = \{M_k \subseteq s_1\}$, and $\{M_k^{(h)} \subseteq s_1\} \cap \{M_k^{(l)} \subseteq s_1\} = \{M_k \subseteq s_1\}$, and $\Big(\{M_k^{(h)} \subseteq s_1\} \setminus \{M_k \subseteq s_1\}\Big) \cap \Big(\{M_k^{(l)} \subseteq s_1\} \setminus \{M_k \subseteq s_1\}\Big) = \emptyset$.

Joint inclusion probability $\pi_{(M_k)(M_k')}$ For $M_k \subset U$ and $M_k' \subset U$, the joint observation of $[M_k]$ and $[M_k']$ requires that (i) at most one node i in $s_{1,T}$, provided $i \in M_k \cap M_k'$, or (ii) at most two nodes i, j in $s_{1,T}$, provided $i \in M_k \setminus M_k'$ and $j \in M_k' \setminus M_k$. Put $M = M_k \cup M_k'$. The relevant subsets are $M^{(i)}$ for all $i \in M_k \cap M_k'$, and $M^{(ij)}$ for all $i \in M_k \setminus M_k'$ and $j \in M_k' \setminus M_k$. The joint inclusion probability $\pi_{(M_k)(M_k')}$ follows, similarly as above for $\pi_{(M_k)}$, as the probability that at least one of these subsets is in the sample of seeds s_1.

Probability $\pi_{(i_1)(i_2)\cdots(i_k)}$ In the case of $k = 2$, $\pi_{(i)(j)}$ is as given earlier in Section 3.4.1. To express $\pi_{(i_1)(i_2)\cdots(i_k)}$ in terms of the probabilities for the initial seed sample $s_{1,0}$, we have

$$\pi_{(i_1)(i_2)\cdots(i_k)} = \sum_{L \subseteq M_k} (-1)^{|L|} \bar{\pi}(L), \tag{3.6}$$

where L includes $\emptyset$, and $|L|$ is its cardinality, and $\bar{\pi}(L)$ is the exclusion probability

$$\bar{\pi}(L) = \Pr(L \cap s_1 = \emptyset) = \Pr(R_L \cap s_{1,0} = \emptyset) = \bar{\pi}_{R_L} = \sum_{D \subseteq R_L} (-1)^{|D|} \pi_D, \tag{3.7}$$

where $R_L = \bigcup_{i \in L} R_i$ and $R_i = \bigcup_{t=0}^{T-1} \eta_i^{[t]}$ is the ancestors of i up to the $T-1$ steps, and π_D is joint inclusion probability of the nodes in D in the initial sample of seeds $s_{1,0}$.

3.4.3 Arbitrary M_k with $k \geq 2$ and $s_2^* = s_1 \times s_1$

By dropping the nodes $s_{1,T}$ of the last wave of T-stage snowball sampling, we ensure that the motif of any subset $M_k \in s_1$ is observable. The idea is developed below.

Definition of $\pi_{(M_k)}$ for $M_k \subseteq s_1$ Let $G_s = (U_s, A_s)$ be the sample graph of T-stage snowball sampling, with reference set $s_2 = s_1 \times U \cup U \times s_1$. Let $G_s^* = (U_s^*, A_s^*)$ be the reduced sample graph obtained from dropping $s_{1,T}$, with reference set $s_2^* = s_1 \times s_1$, where $A_s^* = A_s \setminus \{\langle ij \rangle; i \in s_1, j \in s_{1,T}\}$ and $U_s^* = U_s \setminus s_{1,T} = s_1$. Notice that A_s^* contains all the edges between any $i, j \in s_1$ in the population graph G, and G_s^* is the *same* sample graph that is obtained from s_1 by induced observation directly. It follows that one observes the motif for any $M_k \in s_1$, so that the inclusion probability $\pi_{(M_k)}$ is given by

$$\pi_{(M_k)} = \Pr\big(M_k \subseteq s_1\big) = \pi_{(i_1)(i_2)\cdots(i_k)}, \tag{3.8}$$

where $\pi_{(i_1)(i_2)\cdots(i_k)}$ is given by (3.6) and (3.7) as before.

Joint inclusion probability $\pi_{(M_k)(M_k')}$ For $M_k \subset s_1$ and $M_k' \subset s_1$, the joint observation of $[M_k]$ and $[M_k']$ requires simply $M = M_k \cup M_k' \subseteq s_1$. The joint inclusion probability $\pi_{(M_k)(M_k')}$ is therefore given by $\pi_{(M)}$ on replacing M_k by M in (3.8), (3.6) and (3.7).

Other reduced graphs The sample graph G_s^* is obtained from dropping the T-th wave nodes $s_{1,T}$. Rewrite G_s^* as $G_s^{(T-1)}$; it can be reduced to $G_s^{(T-2)}$ by dropping $s_{1,T-1}$ as well. This yields $G_s^{(T-2)}$ as the induced graph among $s_1 \setminus s_{1,T-1}$. The inclusion probability $\pi_{(M_k)}$ for $M_k \subset A_s^{(T-2)}$ can be derived similarly as (3.8). Carrying on like this, one obtains in the end the reduced graph $G_s^{(0)}$, with reference set $s_2 = s_{1,0} \times s_{1,0}$, which is just the induced graph among $s_{1,0}$. The inclusion probability $\pi_{(M_k)}$ for $M_k \in s_{1,0}$ is $\pi_{M_k} = \Pr\big(M_k \subseteq s_{1,0}\big)$ directly. Notice that the sample graph $G_s^{(0)}$ under T-stage snowball sampling can equally be obtained as $G_s^{(0)}$ under 1-stage snowball sampling. It follows that the additional $T-1$ wave-samples would simply have been wasted, had one only used $G_s^{(0)}$ for estimation. For the same reason it is equally implausible to use $G_s^{(1)}, ..., G_s^{(T-2)}$. However, $G_s^{(T-1)} = G_s^*$ is different because the last wave serves to establish G_s^* as an induced sub-population graph, i.e. with no potentially missing edges among the relevant

nodes.

Comparisons between G_s^* and G_s On the one hand, whichever motif of interest, G_s always has a larger or equal number of observations than G_s^*. Hence, one may expect a loss of efficiency with G_s^*. On the other hand, estimation based on G_s requires more computation than G_s^*. Firstly, for any $M_k \subseteq s_1$, it requires about k times extra computation for $\pi_{(M_k)}$ by (3.5) than by (3.8). This is due to the need to compute the probability of possibly observing M_k as $M_k^{(h)} \subset s_1$ and $h \in s_{1,T}$, even if M_k is observed as $M_k \subset s_1$, which is unnecessary with respect to s_2^*, where the observations are restricted to those among the nodes in s_1 without involving $s_{1,T}$. Secondly, corresponding to each $M_k \subseteq s_1$, there are additional observations with respect to s_2, which are all the possible $M_k' = \{M_k^{(h)}, j; h \in M_k, j \notin s_1\}$, because the motif of such an M_k' can be identified. The motif of any M_k' is unknown, if it differs from any $M_k \subseteq s_1$ by at least two nodes.

Example 5 To illustrate the inclusion probabilities (3.5) and (3.8), consider the population graph $G = (U, A)$ in Figure 3.4, where $|U| = 13$ and $|A| = 19$, together with the two 2-stage snowball sample graphs G_s and G_s^*, both with $s_{1,0} = \{4, 5, 10\}$ by SRS of sample size 3. We have $s_{1,1} = \{1, 2, 8, 9\}$, $s_{1,2} = \{3, 6, 12, 13\}$ and $s_1 = \{1, 2, 4, 5, 8, 9, 10\}$. Table 3.1 lists 6 selected triad (M_3) inclusion probabilities given by (3.5) and (3.8), respectively, with respect to $s_2 = s_1 \times U$ and $s_2^* = s_1 \times s_1$. These are seen to be equal to the true probabilities calculated directly over all possible initial samples $s_{1,0}$, under SRS of sample size 3. Table 3.2 shows the estimates of the four 3rd-order graph totals $\hat{\theta}_{3,h}$, for $h = 0, 1, 2, 3$, which are as defined in Section 3.3.1.3, based on these two sample graphs G_s and G_s^*. The expectation and standard error of each estimators are also given in Table 3.2, which are evaluated directly over all the possible initial sample $s_{1,0}$. The true totals in the population graph G are $(\theta_{3,0}, \theta_{3,1}, \theta_{3,2}, \theta_{3,3}) = (121, 123, 40, 2)$. Clearly, both HT-estimators are unbiased, and using G_s^* entails a loss of efficiency against G_s, as commented earlier.

3.4.4 Proportional representative sampling in graphs

A traditional sampling method is sometimes said to be (proportional) representative if the sample distribution of the survey values of interest is an unbiased estimator of the population distribution directly. This is the case provided equal

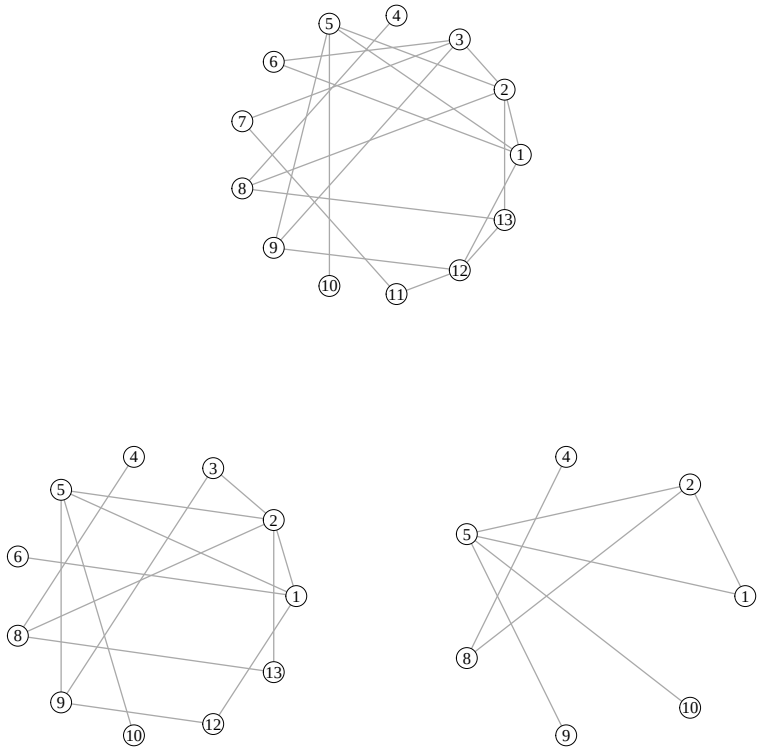


Figure 3.4: Population graph G with 13 nodes and 19 edges (top); sample graphs G_s (bottom left) and G_s^* (bottom right) by 2-stage snowball sampling with initial $s_{1,0} = \{4, 5, 10\}$.

probability selection. Equipped with the general formulae for $\pi_{(M_k)}$ under T-stage snowball sampling, below we propose and examine a proportional representativeness concept for graph sampling.

Graph proportional representativeness Let $m_k \neq m'_k$ be two distinct motifs of the order k. A graph sampling method is *k-th order proportionally representative* (PR_k) if

$$E[\theta_s]/\theta = E[\theta'_s]/\theta', \tag{3.9}$$

where θ is the number of m_k in the population graph G, and θ_s that of the observed m_k in the sample graph G_s with reference set s_2, and similarly with θ' and θ'_s for m'_k. Let $y(M_k) = 1$ if $[M_k] = m_k$ and 0 otherwise. Let $\delta_{[M_k]}$ be the observation indicator with respect to s_2. We have $\theta = \sum_{M_k \in \mathcal{C}_k} y(M_k)$ and $\theta_s = \sum_{M_k \in \mathcal{C}_k} \delta_{[M_k]} y(M_k)$. Clearly, a graph sampling method will be PR_k if $\pi_{(M_k)}$

Table 3.1: Inclusion probability $\pi_{(M_3)}$ of selected $M_3 = \{i_1, i_2, i_3\}$

			With $s_2 = s_1 \times U$		With $s_2^* = s_1 \times s_1$	
i_1	i_2	i_3	By (3.5)	True	By (3.8)	True
1	2	3	0.9230769	0.9230769	0.5664336	0.5664336
1	2	4	0.8531469	0.8531469	0.2657343	0.2657343
1	3	4	0.8321678	0.8321678	0.2027972	0.2027972
2	3	4	0.8531469	0.8531469	0.2552448	0.2552448
1	2	5	0.8671329	0.8671329	0.6223776	0.6223776
1	3	5	0.8881119	0.8881119	0.5384615	0.5384615

Table 3.2: Third-order graph total estimate, expectation and standard error

Based on sample graph G_s	$\hat{\theta}_{3,0}$	$\hat{\theta}_{3,1}$	$\hat{\theta}_{3,2}$	$\hat{\theta}_{3,3}$
Estimate	96.251	89.260	26.289	2.515
Expectation	121	123	40	2
Standard error	22.977	18.591	7.025	0.768
Based on sample graph G_s^*	$\hat{\theta}_{3,0}$	$\hat{\theta}_{3,1}$	$\hat{\theta}_{3,2}$	$\hat{\theta}_{3,3}$
Estimate	59.128	63.209	19.211	1.607
Expectation	121	123	40	2
Standard error	78.694	49.929	15.038	1.195

is a constant for different motifs of order k. Under any PR_k design, one may estimate the relative frequency between m_k and m'_k by θ_s/θ'_s.

Result 1. Induced observation from s_1 is PR_k for $k \geq 1$, provided $s_2 = s_1 \times s_1$ and symmetric design $p(s_1)$. The result follows since, for any $M_k \subset A_s = s_1$, we have $\pi_{(M_k)} = \pi_{M_k}$, which is a constant of $[M_k]$ by virtue of symmetric design $p(s_1)$.

Result 2. One-stage snowball sampling is PR_k for $k \geq 2$, provided $s_2 = s_1 \times U \cup U \times s_1$ and symmetric design $p(s_1)$. Suppose first reciprocal observation. We have $R_i = \{i\} \cup \eta_i^{[1]}$, whose cardinality varies for different nodes in G. It follows that $\pi_{(M_1)} = \pi_{(i)}$ by (3.3) is not a constant over U, i.e. the design is not PR_1. Next, for M_k with $k \geq 2$, $\pi_{(M_k)}$ by (3.5) depends on $k+1$ probabilities given by (3.6) and (3.7). Each relevant probability $\bar{\pi}(L)$ is only a function of $|R_L|$ provided symmetric design $p(s_1)$, where $R_L = \bigcup_{i \in L} R_i = L$ since $R_i = \{i\}$ given $T = 1$. It follows that $|R_L| = |L|$ regardless of the nodes in M_k, such that $\pi_{(M_k)}$ is a constant of M_k, i.e. PR_k. Similarly for forward observation in digraphs.

Remark Setting $s_2^* = s_1 \times s_1$ yields induced sample graph from s_1 and Result 1.

Result 3. T-stage snowball sampling is generally not PR_k for $k \geq 1$ and $T \geq 2$, despite symmetric design $p(s_1)$. As under 1-stage snowball sampling, the design is not PR_1. Whether by (3.5) or (3.8) for $k \geq 2$, $\pi_{(M_k)}$ depends on $\bar{\pi}(L)$ in (3.6), which is only a function of $|R_L|$ provided symmetric design $p(s_1)$. However, given $T \geq 2$ and $|L|$, $R_L = \bigcup_{i \in L} R_i$ generally varies for different L, so that neither R_L nor $|R_L|$ is a constant of the nodes in M_k, i.e. the design is not PR_k. Similarly for forward observation in digraphs.

3.5 Network sampling methods

As prominent examples from the network sampling literature we consider here multiplicity sampling (Birnbaum and Sirken, 1965), indirect sampling (Lavallée, 2007) and adaptive cluster sampling (Thompson, 1990). Below we first summarise broadly their characteristics in terms of target parameter, sampling and estimator, and then discuss four salient applications of these methods using the snowball sampling theory developed in Section 3.4.

Target parameter In all the network sampling methods mentioned above, the target parameter is the total of a value associated with each node of the graph, denoted by y_i for $i \in U$, which can be referred to as a 1st-order network total $\theta = \sum_{i \in U} y_i$ in light of (4.1). This does not differ from that when "conventional" sampling methods are applied for the same purpose, where Sirken (2005) uses the term conventional in contrast to network. In other words, these network sampling methods have so far only been applied to overcome either certain deficiency of frame or lack of efficiency of the traditional sampling methods, as discussed below in terms of sampling and estimator, but not in order to study genuine network totals or parameters, which are of orders higher than one.

Sampling Like in the definition of sample graph, these network sampling methods start with an initial sample s_1. The sampling frame of s_1 can be *direct* or *indirect*. In the latter case, the sampling units are not the population elements. This may be necessary because a frame of the population elements is unavailable, such as when siblings are identified by following up kins to the household mem-

bers of an initial sample of households (Sirken, 2005). Or, a frame of the elements may be available but is unethical to use, such as when children are accessed via a sample of parents (Lavallée, 2007). In cases a direct frame of elements is used, the initial sample s_1 may be inefficient due to the low prevalence of in-scope target population elements, so that an observation procedure depending on the network relationship (between the elements) is used to increase the effective sample size. This is the case with adaptive cluster sampling (Thompson, 1989).

Estimator For 1-st order network parameters (4.1), where the population elements are represented as nodes in the population graph $G = (U, A)$, the HT-estimator (3.2) is defined for the observed nodes in the sample graph $G_s = (U_s, A_s)$. Another approach in the aforementioned methods is the HT-estimator defined for the selected sampling units. Let F be the frame of sampling units, where $l \in F$ has inclusion probability π_l. We have

$$\sum_{l \in F} z_l = \sum_{l \in F} \Big(\sum_{i \in U} w_{li} y_i \Big) = \sum_{i \in U} y_i \sum_{l \in F} w_{li} = \sum_{i \in U} y_i = \theta,$$

where $z_l = \sum_{i \in U} w_{li} y_i$ is a value constructed for the sampling units, based on *any* chosen weights, provided $\sum_{k \in F} w_{ki} = 1$, as noted by Birnbaum and Sirken (1965). The corresponding HT-estimator that is unbiased for θ can be given by

$$\tilde{\theta}_{HT} = \sum_{l \in s_1} z_l / \pi_l = \sum_{l \in F} z_l \delta_l / \pi_l, \tag{3.10}$$

where $\delta_l = 1$ if $l \in s_1$ and 0 otherwise. To ensure that z_l can be calculated no matter which actual sample s_1, the weights w_{li} must not depend on s_1. A common approach is to set $w_{li} = 1/m_i$, where l a sampling unit in s_1 which gives rise to i, and m_i is the number of *all* sampling units in F that could lead to the observation of i, for $i \in U$. The number m_i is referred to as the *multiplicity* of the element (Birnbaum and Sirken, 1965). The observation of m_i for each sample element is the same kind of requirement as the observation of the ancestors of a node in U_s under snowball sampling. The literature is inconclusive on the relative efficiency between the two estimators (3.2) and (3.10).

3.5.1 Sampling patients via hospitals

Birnbaum and Sirken (1965) consider this situation, without using graph representation. To fix the idea, suppose a sample of hospitals is selected according to a probability design. From each sample hospital, one observes a number of patients of a given type, who are treated at this hospital. Let the target parameter θ be the population size of such patients. The complication arises from the fact that a patient may receive treatment at more than one hospital. Sirken (2005) refers to conventional sampling where every population element is linked to one and only one sampling unit, whereas in the case of network sampling a population element (i.e. patient of a certain type) can be linked to a varying number of sampling units (i.e. hospitals). Sirken (2005) refers to 'cluster' as the group of population elements which are linked to the same sampling unit, and 'network' the group of sampling units which are linked to the same population element. The distinction between cluster and network here needs to be accounted for in estimation.

(P) Projection graph The HT-estimator (3.2) can be obtained using the following graph sampling set-up. Denote by H the known set of hospitals and P the unknown set of in-scope patients, where $\theta = |P|$. Let $G = (U, A)$ have $U = H \cup P$. For any $i \in H$ and $j \in P$, $a_{ij} \in A$ iff patient j receives treatment at hospital i. Let the simple graph be undirected. Notice that (H, P) form a bipartition of U, where there are no other edges at all except those that *project* H onto P. Given $s_1 \subset H = U_0$, let $s_2 = s_1 \times P$ for 1-stage snowball sampling. The observation procedure must be incident ancestral, so that m_i is observed for $i \in \alpha(s_1)$, without including in the sample graph G_s all the edges that are incident at i but outside of s_2. The inclusion probability $\pi_{(i)}$ is given by (3.3), where we have $\eta_i^{[0]} = \emptyset$ since $U_0 \cap P = \emptyset$, and $\eta_i^{[1]} = \beta_i$, so that $R_i = \beta_i$ and $|R_i| = m_i$. Let $y_i = 1$ for all $i \in P$.

Remark The HT-estimator (3.2) and (3.10) correspond to the first two estimators proposed by Birnbaum and Sirken (1965). Their third estimator is defined for the edges in the projection graph, which however lacks a formulation that allows it to be applied generally.

Two-stage snowball sampling Consider 2-stage snowball sampling in the same graph, under which the observation procedure is incident but needs not be ancestral in addition. Given $s_{1,0} \subset H$, let $s_{1,1} = \alpha(s_{1,0}) \subseteq P$ and $s_{1,2} = \alpha(s_{1,1}) \subseteq$

H, i.e. reverse projection. The HT-estimator (3.2) makes only use of the nodes (i.e. motif of interest) in $s_{1,1}$, where $y_i \equiv 1$, and $\pi_{(i)}$ is given by (3.3), for which $R_i = \beta_i$ is fully observed due to the addition of $s_{1,2}$.

3.5.2 Sampling children via parents

Lavallée (2007) considers this situation. Children are the population elements. Suppose a sample of parents is selected according to a probability design. One obtains all the children of each sample parent. Without losing generality, let the target parameter θ be the number of children who are not orphans. The same complication arises from the fact that a child may be accessed via two parents if they are both in the sampling frame. Clearly, the situation is conceptually the same as sampling patients via hospitals above.

Remark Lavallée (2007) represents the situation using the same graph (P) above, where $U = P \cup C$, and P consists of the parents and C the children. The HT-estimator (3.2) based on either 1- or 2-stage snowball sampling formulation is the same as above, with $y_i \equiv 1$ for $i \in C$. Lavallée (2007) considers only the HT-estimator (3.10).

(M) Multigraph Put $G = (U, A)$ where $U = P$ and $A = C$, i.e. with parents as the nodes and children as the edges. Let A_{ij} represent the a_{ij} children of parents i and j. Let loops A_{ii} at node i represent the a_{ii} children of single-parent i. Given $s_1 = s_{1,0} \subset P = U_0$, let $s_{1,1} = \alpha(s_{1,0}) \setminus s_{1,0}$, i.e. 1-stage snowball sampling. The incident observation procedure is ancestral by construction here. Let $s_2 = s_1 \times U$. The inclusion probability $\pi_{(ij)}$ of a child $\langle ij \rangle \in A$ is given by (3.4), where $\eta_i^{[0]} = \{i\}$ and $\eta_j^{[0]} = \{j\}$ under 1-stage snowball sampling; whereas $\pi_{(ii)}$ of a child $\langle ii \rangle$ of a single parent is also given by (3.4), where $\eta_i^{[0]} = \{i\}$.

Remark Making population elements the edges of the graph is not convenient for the hospital-patient application, because while each child corresponds to only one edge, each patient may appear as multiple edges incident to different nodes (i.e. hospitals).

3.5.3 Sampling siblings via households

Sirken (2005) discusses this situation, without resorting to explicit graph representation. To fix the idea, suppose a sample of households is selected according to a probability design. For each member of the household, one obtains all the siblings who may or may not live in the same household. The observation elements are siblings, denoted by S, which excludes anyone who has no siblings. Without losing generality, let θ be the number of siblings.

(2P) Twice projection graph Denote by H the households, P the persons, and S the siblings, where $i \in S$ is considered a different element to $j \in P$, even if i and j refer to the same person in real life. Let $G = (U, A)$, where $U = H \cup P \cup S$ and $A = A^{HP} \cup A^{PS}$. Each $A_{hj} \subset A^{HP}$ is such that $h \in H$ and $j \in P$, i.e. A^{HP} projects H onto P; each $A_{ij} \in A^{PS}$ is such that $i \in P$ and $j \in S$ are siblings, including when the two refer to the same person, i.e. A^{PS} projects P onto S. Let the twice projection graph from H to P to S be undirected. Consider 2-stage snowball sampling starting from $s_{1,0} \subset H = U_0$. Let $s_2 = s_1 \times U$, where $s_1 = s_{1,0} \cup s_{1,1}$ is the sample of seeds. The observation procedure must be incident ancestral, provided which the HT-estimator (3.2) is only based on $s_{1,2}$. For $i \in S$, we have $y_i = 1$ and $\pi_{(i)}$ given by (3.3), where $\eta_i^{[0]} = \eta_i^{[1]} = 0$ because it can only be reached from $s_{1,0}$ in exactly two waves, and $\eta_i = \eta_i^{[2]}$ where $|\eta_i| = m_i$ is the number of households that can lead to i from $s_{1,0}$, i.e. its multiplicity according to Birnbaum and Sirken (1965).

(PR) Projection relation graph Put $G = (U, A)$, where $U = H \cup P$. Let $a_{ij} \in A$ if (i) person j belongs to household i, or (ii) persons i and j are siblings of each other. The edges of type (i) project H on to P, whereas those of type (ii) are relations within P. Notice that each group of siblings form a clique; a person without siblings is a single-node clique. To ensure ancestral observation, consider 3-stage snowball sampling. Given $s_{1,0} \subset H = U_0$, $s_{1,1}$ consists of the members of the households in $s_{1,0}$, and $s_{1,2}$ the siblings of $s_{1,1}$ which are outside of the initial sample households, and $s_{1,3} \subseteq H$ consists of the households to $s_{1,2}$. Let $s_2 = s_1 \times U$, where $s_1 = s_{1,0} \cup s_{1,1} \cup s_{1,2}$. The HT-estimator (3.2) makes use of $i \in s_1 \cap S$, with $y_i \equiv 1$. The corresponding $\pi_{(i)}$ is given by (3.3), where $\eta_i^{[0]} = 0$, and $\eta_i^{[1]}$ is the household of i, and $\eta_i^{[2]}$ contains the households of its out-of-household siblings. In other words, η_i contains all the households that can lead to i, where $|\eta_i| = m_i$.

Remark Sampling in the graphs (2P) and (PR) makes use of relationships among the population elements, unlike sampling of patients or children in the projection graph (P).

(HP) Hypernode projection graph Let each clique in the graph (PR) above be a *hypernode* — all the nodes of a hypernode are always observed together or not at all. Let $G = (U, A)$, where $U = H \cup \mathcal{P}$, and $\mathcal{P}$ consists of all the hypernodes of P. Let $a_{ij} = 1$ iff at least one node in the hypernode j belongs to household i. This yields an undirected simple graph as the hypernode projection graph. Consider 2-stage snowball sampling with $U_0 = H$ as in the projection graph, such that observation is ancestral by construction. Both HT-estimators (3.2) and (3.10) follow directly, where y_i is the number of nodes in $i \in \mathcal{P}$.

3.5.4 Adaptive cluster sampling of rare species

In contrast to conventional sampling, Thompson (1990) characterises adaptive sampling designs as those in which the procedure to include units in the sample depends on the values of interest observed during the survey. To fix the idea, suppose an area is divided into (spatial) grids as the units of sampling and observation. Each grid in an initial sample of grids is surveyed for a given species of interest. If it is not found there, one would move on to another grid in the initial sample. However, whenever the species is found in grid i, one would survey each of its neighbour grids in four directions, beyond the initial sample, provided not all of them have been surveyed before. This observation procedure can help to increase the number of in-scope grids, compared to random sampling of the same amount of grids, provided the species is more likely to be found given that it is found in a neighbour grid than otherwise. Once in a new grid, the procedure is repeated and the survey may or may not continue to the neighbour grids, depending on the finding in the current grid. The sampling is finished if no new grids can be added to the sample, or if one has reached a predetermined limit in terms of the number of surveyed grids, time, resource, etc. The observed in-scope grids form sampling as well as observation clusters, in the sense that all the member grids of a cluster are sampled and observed if any one of them is.

(T) Transitive graph Adaptive cluster sampling (ACS) can be represented as 2-stage snowball sampling in a transitive graph as follows. Let $G = (U, A)$, where U contains all the grids in ACS. Let U_A contain all the grids where the

rare species is present. Let $U_A^c = U \setminus U_A$. Let $a_{ij} = 1$ iff $i, j \in U_A$ and i and j belong to the same observation cluster under the ACS. This yields an undirected simple *transitive* graph, where each $i \in U_A^c$ is an isolated node, and each group of connected nodes in U_A form a clique. Without losing generality, let $\theta = |U_A|$. The snowball sampling starts with $s_{1,0} \subset U = U_0$, i.e. any grid can be selected initially. Let $s_{1,1} = \alpha(s_{1,0})$. Notice that the isolated nodes in $s_{1,0}$ do not lead to any nodes in $s_{1,1}$, while a connected node in $s_{1,0}$ leads to all the nodes in ies observation cluster but none in U_A^c, since edges exist only among the nodes in U_A. In reality, a neighbour grid of $i \in U_A \cap s_{1,0}$ which belongs to U_A^c is also surveyed, but it will not lead to any additional nodes in the next wave, nor will it be the motif of interest in estimation. It is therefore convenient to represent this adaptive nature of the ACS by not including in $s_{1,1}$ any node from U_A^c at all. The 2nd-wave snowball sample will be empty, i.e. $s_{1,2} = \emptyset$, because all the connected nodes in a clique will already be observed either in $s_{1,0}$ or $s_{1,1}$. But the 2nd-stage is needed to ensure that the observation is ancestral by construction. The HT-estimator (3.2) uses every node $i \in s_1 = s_{1,0} \cup s_{1,1}$, with $y_i = 1$, and $\pi_{(i)}$ is given by (3.3), where $\eta_i^{[0]} = \{i\}$, and $\eta_i^{[1]}$ contains all its adjacent nodes.

Remark The graph (T) is the same as the relation part of the graph (PR) in the case of sampling siblings via households. The projection part is not necessary here because the initial sampling uses a direct frame, unlike the other applications above.

Remark The ACS can as well be represented by the graph (HP), with the cliques in the graph (T) above as the hypernodes. Both HT-estimators (3.2) and (3.10) follow directly.

3.6 Concluding remarks

In this paper we synthesised the existing graph sampling theory, and made several extensions of our own. We proposed a definition of sample graph, to replace the different samples of nodes, dyads, triads, etc. This provides formally an analogy between sample graph as a sub-population graph and sample as a sub-population. Next, we developed a general approach of HT-estimation based on arbitrary T-stage snowball sampling. It is clarified that design-based estimation based on snowball sampling requires the observation procedure to be ancestral,

which can be hard to fulfil in many practical applications of snowball or snowball-like sampling, including the estimation of a clandestine target population size. Without satisfying the ancestral requirement, the estimation will have to be based on an appropriate statistical model instead.

We presented various graph sampling formulations of the existing design-based network sampling methods. It is seen that different graph representations reveal the different estimators more or less readily, so the choice matters in applications. The graph sampling theory provides a more general and flexible framework to study and compare these unconventional methods, and to develop possible alternatives and modifications.

Moreover, it transpires that these existing network sampling methods do not really differ from conventional sampling with respect to the target parameter. We believe that the scope of investigation can be greatly extended if one starts to consider other genuine network parameters, which can only be studied using a graph representation. Two research directions can be identified in this respect. First, we are currently examining the scope of problems that can be studied using the (hypernode) projection graph, and the properties of the design-based estimation methods. Second, it seems intuitive that a lower-order network parameter can be estimated using a 'smaller' or more fragmented sample graph than a higher-order parameter. It is therefore interesting to understand better the conditions, by which a high-order network parameter can be expressed as a function of lower-order parameters. Although this is perhaps more of a mathematical than statistical problem, such transformations can potentially be very useful for the applications of the graph sampling theory. Developing a comprehensive finite-graph sampling theory, beyond the established finite-population sampling theory, seems an exciting area for future research.

Chapter 4

Incidence weighting estimation under sampling from a bipartite incidence graph

We consider design-unbiased estimation on a bipartite incidence graph. The bipartite incidence graph can be used to represent many graph sampling situations for the purpose of estimation, including also the so-called unconventional sampling methods in the literature, such as indirect sampling, network sampling and adaptive cluster sampling. We propose a class of linear estimators based on the edges of the sample bipartite incidence graph, subjected to a general condition of design unbiasedness. The proposed class of estimators contains as special cases the classic Horvitz-Thompson estimator, as well as the other existing unbiased estimators under unconventional sampling, which can be traced back to Birnbaum and Sirken (1965). The generalisation allows one to devise new unbiased estimators, and thereby greatly increase the scope of efficiency improvement in applications. Numerical illustrations are provided for a number of incidence weighting estimators.

Key words: graph sampling, incidence weight, multiplicity weighting

4.1 Introduction

Birnbaum and Sirken (1965) consider the situation in which patients are sampled indirectly via the hospitals where they receive treatment. Since a patient may receive treatment from more than one hospital, the patients are not nested in the hospitals like elements in clustered sampling. Birnbaum and Sirken propose three estimators for such *indirect sampling*. The first one is the standard Horvitz-Thompson (HT) estimator (Horvitz and Thompson, 1952) in finite-population sampling, where each sampled patient is weighted by the inverse of the probability of being observed. The other two estimators are unusual: one is based on the sampled hospitals and a constructed measure for each of them, the other based on a sub-sample of hospitals determined by a *priority rule* and a constructed measure. Later, the first of these two estimators was recast as a generalised *weight share* method for indirect sampling (Lavallée, 2007); it was reused for *network sampling* (Sirken, 2005) and adaptive cluster sampling (Thompson, 2012, Ch. 24). However, the other priority-rule estimator appears to have vanished from the literature.

Zhang and Patone (2017) synthesise the existing graph sampling theory, extending previous works on this topic by Frank (1980a, 1980b, 2011). A formal definition is given for sampling from finite graphs and the HT-estimator is developed for general T-stage snowball sampling. In particular, they show that all the aforementioned unconventional sampling techniques can be given as various instances of graph sampling. In this paper, we shall use a *bipartite incidence graph (BIG)* to represent all these situations of sampling. For instance, the nodes can be the hospitals and the patients and an edge exists between a hospital and any patient that has received treatment at the hospital. This is a bipartite graph since the nodes of this graph are naturally divided into two disjoint sets.

The unified BIG representation allows us to reconsider and to extend the three estimators of Birnbaum and Sirken (1965), under a much more general setting that is immediately applicable to all these and other situations of sampling that can be represented by the BIG. We will show how the three estimators of Birnbaum and Sirken (1965) are particular cases of a general class of estimators which we call *incidence weighting estimators (IWEs)*. Not only can their two unusual estimators be given a unified treatment, which is hitherto unknown in the literature, they can both be extended in various ways, which increases the possibility for gains of estimation efficiency in applications.

In Section 2, we introduce the BIG formally and explain how it can be used to represent all the aforementioned situations of sampling. We recast the three estimators of Birnbaum and Sirken (1965) as estimators on the BIG, and provide the variance of the priority-rule estimator explicitly, which was lacking in Birnbaum and Sirken (1965). In Section 3, we develop the IWE, which is based on the sample incidence relationships (edges of the sample BIG). We develop the general condition for design unbiased IWE, derive its theoretical sampling variance and associated variance estimation. Some examples of unbiased incidence weights are presented in Section 4, where we recast the three estimators of Birnbaum and Sirken (1965) as the IWEs, as well as proposing new estimators. Numerical illustrations will be given for several of them in Section 5. Section 6 contains some brief concluding remarks and some topics for future research.

4.2 Basics of BIG sampling and estimation

Denote by $G = (F, U; A)$ a bipartite simple directed graph, where (F, U) forms a bipartition of the node set $F \cup U$, and each edge in A points from one node in F to another in U. The graph is directed, i.e. the edge that goes from i to j is different from the edge that goes from j to i. The graph is bipartite since there does not exist any edge among the nodes in F, nor so in U, but only between F and U. Let $F = \{1, \ldots, M\}$ and $U = \{1, \ldots, N\}$. In using G to represent BIG sampling, we assume that F is the frame containing the set of initial *sampling units*, and U is the population containing the set of *motifs* of interest, and an edge (ki) that is incident to $k \in F$ and $i \in U$ exists, if and only if the selection of k in a sample s from F leads to the observation of motif i in U. The incidence relationships corresponding to the edges in A can thus also be interpreted as *incidence of sampling*. In particular, provided $k \in s$, a sampling unit k in F will lead to the observation of all the motifs in U that are adjacent to k in G, denoted by $\alpha_k = \{i; i \in U, (ki) \in A\}$.

Henceforth we shall refer to G as BIG. Some examples of BIG sampling are as follows.

- Indirect sampling (Birnbaum and Sirken 1965; Lavallée, 2007): F consists of the hospitals, U the patients, and an edge exists between a hospital in F and a patient in U if and only if the patient receives treatment at the hospital.
- 'Network sampling' (Sirken, 2005): F consists of the households, U the cliques

of siblings, and an edge exists between a household in F and a sibling-clique in U if and only if at least one of the siblings belong to the household.

- Adaptive cluster sampling (Thompson, 1990): F consists of the spatial grids over a given area of habitat for a rare species, U the clusters of neighbouring grids where one can find the species of interest, and an edge exists between a grid in F and a grid-cluster in U if and only if the grid belongs to the cluster.

4.2.1 BIG sampling

Insofar as $\beta_i = \{k; k \in F, (ki) \in A\}$, for $i \in U$, may contain more than one unit in F, one needs to know β_i, so as to able to calculate the probability of observing i under BIG sampling. This requires the observation procedure of BIG sampling to be *ancestral* (Zhang and Patone, 2017). Ancestral observation procedure is also needed to implement the other two unusual estimators of Birnbaum and Sirken (1965) under indirect sampling.

By way of introduction, consider BIG sampling on the population graph G given at the top of Figure 4.1, where $F = \{1, 2, 3, 4\}$ and $U = \{5, 6, 7, 8, 9, 10, 11\}$. The edge set A and the set U are unknown. But the set F is known, and serves as the sampling frame of the initial sample, where $M = |F| = 4$ is its size. Given an initial sample s of size m, for $s \subset F$, the sample graph (Zhang and Patone, 2017) is given by $G_s = (s \cup \alpha(s), A_s)$, where

$$A_s = \text{Inc}(s) = \{(ki); k \in s, i \in U, a_{(ki)} = 1\} = \big(s \times \alpha(s)\big) \cap A \ ,$$

which consists of all the edges in A that are incident to the units in sample s. Thus, given $s = \{1, 2\}$, we have $G_s = \{\{1, 2\} \cup \alpha\big(\{1, 2\}\big); A_s\}$, as shown to the bottom-left in Figure 4.1, where

$$\begin{aligned} &\alpha\big(\{1, 2\}\big) = \{10\} \cup \{5, 7, 9\} \\ &A_s = \{(1, 10), (10, 1), (2, 5), (5, 2), (2, 7), (7, 2), (2, 9), (9, 2)\} \end{aligned}$$

Notice that the observation procedure given s is by default incident. More importantly, to facilitate estimation based on the sampling design, the observation procedure needs to be ancestral in addition, such that all the units that would lead to the inclusion for each sample motif $i \in \alpha(s)$ are observed. Thus, given $s = \{1, 2\}$, the motifs in $\beta\big(\alpha(s)\big) \setminus s = \{3, 4\}$ are observed due to the ancestral

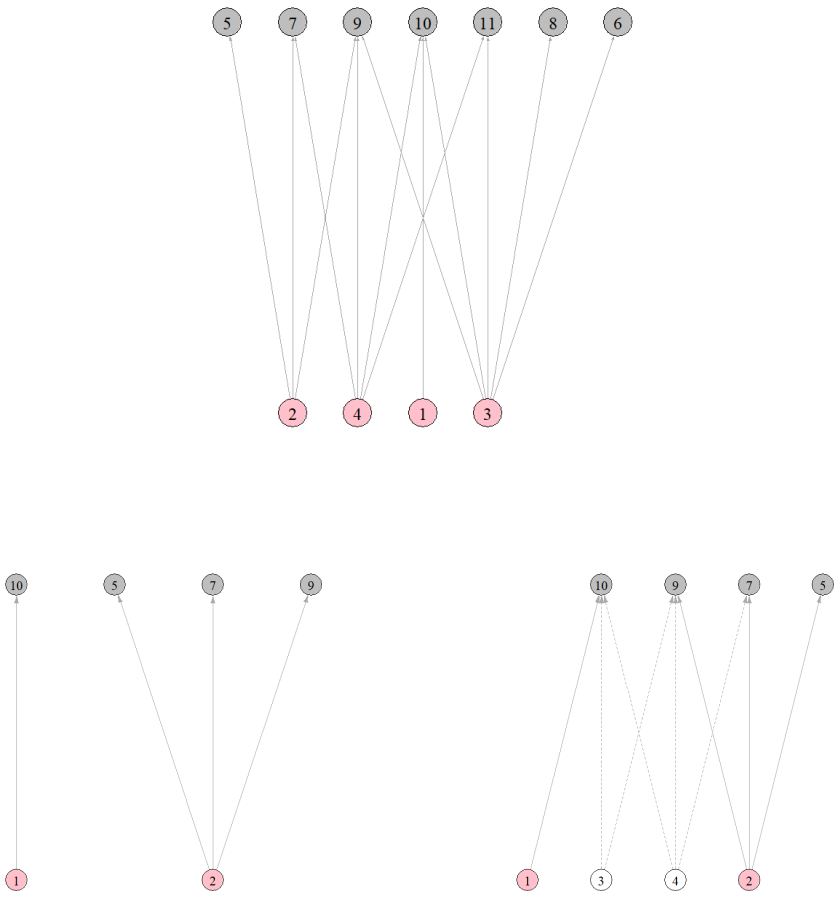


Figure 4.1: Top, population bipartite incidence graph $G = (F, U; A)$. Sample graph G_s given $s = \{1, 2\}$: bottom-left, by incident reciprocal observation; bottom-right, by incident ancestral observation, with additional information marked by dotted edges.

nature of the observation procedure, even though they are not part of the sample graph; nor are the dotted edges to the bottom-right of Figure 4.1 included in the sample graph, which exist between $\beta\big(\alpha(s)\big) \setminus s \subset F$ and $\alpha(s) \subset U$. However, the knowledge of the *existence* of these dotted edges is necessary in order to be able to calculate the HT and the other estimators to be described later.

Finally, as explained by Zhang and Patone (2017), incident ancestral observation in graph sampling can generally be achieved by T-stage snowball sampling, but retaining only the edges observed in the first $T - 1$ stages. For BIG sampling from G in Figure 4.1 and given $s = \{1, 2\}$, 2-stage snowball sampling would lead

to the observation of the dotted edges at the second stage, retaining only the edges from the 1st-stage allows one to retain the knowledge of their existence while removing them from the sample graph.

4.2.2 Three existing estimators under BIG sampling

Let y_i be the value of interest associated with each motif $i \in U$. Let the target parameter for estimation be

$$\theta = \sum_{i \in U} y_i = \sum_{k \in F} z_k = \sum_{(ki) \in A} w_{ki} y_i, \tag{4.1}$$

where z_k is a constructed measure for each unit in F, which is given by

$$z_k = \sum_{i \in \alpha_k} w_{ki} y_i \qquad \text{and} \qquad \sum_{k \in \beta_i} w_{ki} = 1 \tag{4.2}$$

(Birnbaum and Sirken, 1965). The weight w_{ki} is a fixed constant of sampling, for $(ki) \in A$, and it takes value 0 if $a_{(ki)} = 0$. Below we present the three estimators in Birnbaum and Sirken (1965) under the BIG framework, denoted by Yhat, Zhat and Phat. None of them dominates another in term of efficiency generally speaking.

Yhat Given the sample graph G_s, where $s \subset F$ is selected according to a probability sampling design. Let π_k and π_{kl} be, respectively, the first and second-order inclusion probabilities of $k, l \in F$. Let the HT-estimator based on $\{y_i; i \in \alpha(s)\}$ be given by

$$\hat{\theta}_y = \sum_{i \in \alpha(s)} \frac{y_i}{\pi_{(i)}} = \sum_{i \in U} \frac{\delta_{(i)}}{\pi_{(i)}} y_i \ ,$$

where $\delta_{(i)} = 1$ if $i \in \alpha(s)$ and 0 otherwise. The probability $\pi_{(i)} = \Pr[i \in \alpha(s)]$, for $i \in U$, is notationally distinguished from π_k for $k \in F$. It can be derived from the sampling distribution $p(s)$, since we have

$$\pi_{(i)} = 1 - \Pr[\beta_i \cap s = \emptyset] = 1 - \Pr[\text{none of } \beta_i \text{ is included in } s]$$

(Birnbaum and Sirken, 1965; Frank, 1971). The variance of $\hat{\theta}_y$ follows the standard variance formula for HT-estimator, which requires the second-order inclusion probabilities $\pi_{(i)(j)}$ for $i, j \in U$; see Zhang and Patone (2017) for more details.

For simplicity in discussion of alternative estimators later on, we shall refer to the HT-estimator as the *Yhat*.

Zhat Let w_{ki} be a value (i.e., weight) associated with the edge (ki) connecting the motif i with the sampling unit k. A measure $z_k = \sum_{i\in\alpha_k} w_{ki}y_i$ is defined for the sampling unit k in (4.2). Let $\delta_k = 1$ if $k \in s$ and 0 otherwise. An inverse probability weighted estimator of θ based on $\{z_k; k \in s\}$ can now be given as

$$\hat{\theta}_z = \sum_{k\in s} \frac{z_k}{\pi_k} = \sum_{k\in F} \frac{\delta_k}{\pi_k} z_k \qquad \text{and} \qquad \sum_{k\in\beta_i} w_{ki} = 1 \ . \tag{4.3}$$

We shall refer to this estimator as the *Zhat*. It is unbiased since $E(\hat{\theta}_z) = \sum_{k\in F} z_k = \theta$ by construction (4.1) and (4.2). For the so-called *multiplicity* estimator, which was first proposed by Birnbaum and Sirken (1965) and later developed by Sirken and Levy (1974), Sirken (2004) and by Lavallée (2007) for his generalised weight share methods, the default choice for w_{ki} is the *equal-share* weight:

$$w_{ki} = \frac{1}{d_i} \qquad \text{where} \quad d_i = |\beta_i| \ .$$

Birnbaum and Sirken(1965) actually pointed out that the w_{ki}'s for the same motif i can be unequal, as long as they sum to one for each motif and do not vary according to which other sampling units are selected in the initial sample s. Lavallée (2007) explores optimal weight-sharing which minimises $V(\hat{\theta}_z)$, and finds the result to be inconclusive. Although an optimal choice might be hard to find, there still can be many different choices of weights subjected to (4.2), which are all unbiased but have different variances. We will present a new and often more efficient choice of w_{ki} in Section 4.

Phat The third expression of θ in (4.1) suggests the possibility of estimation based on $\{w_{ki}; (ki) \in A_s\}$. However, under BIG sampling, where all the edges incident to k are observed together, whenever $k \in s$, we have $\pi_{(ki)} = \Pr[(ki) \in A_s] = \pi_k$, such that

$$\sum_{(ki)\in A_s} \frac{w_{ki}y_i}{\pi_{(ki)}} = \sum_{(ki)\in A_s} \frac{w_{ki}y_i}{\pi_k} = \sum_{k\in s} \frac{1}{\pi_k} \sum_{i\in\alpha(k)} w_{ki}y_i = \hat{\theta}_z.$$

Instead, Birnbaum and Sirken (1965) base the priority-rule estimator on a *prioritised subset* of A_s, denote by A_{sp}. Let $I_{(ki)} = 1$ if the edge (ki) is in A_{sp} and 0

otherwise. Birnbaum and Sirken (1965) let $I_{(ki)} = 1$, for $i \in \alpha(s) \subset U$, if

$$k = \min\left(s \cap \beta_i\right)$$

i.e. if k happens to be enumerated first in the frame, among all the sample units that can lead to i. Clearly, other priority rules are possible, though it was not explicitly mentioned. In any case, let the *Phat* based on $\{w_{ki}; (ki) \in A_{sp}\}$ be given by

$$\hat{\theta}_p = \sum_{(ki)\in A_{sp}} \frac{w_{ki} y_i}{\pi_{(ki)} p_{(ki)}} = \sum_{(ki)\in A_s} \frac{I_{(ki)}}{p_{(ki)}} \cdot \frac{w_{ki} y_i}{\pi_{(ki)}} \tag{4.4}$$

where $p_{(ki)}$ is the conditional probability that (ki) is prioritised given $(ki) \in A_s$, i.e.

$$p_{(ki)} = \Pr\left[I_{(ki)} = 1 | (ki) \in A_s\right]$$

Since the unconditional probability of $(ki) \in A_{sp}$ is $\Pr\left[(ki) \in A_{sp}\right] = \pi_k p_{(ki)}$, we have $E(\hat{\theta}_p) = \sum_{(ki)\in F} w_{ki} y_i = \theta$ by construction (4.1) and (4.2), provided $p_{(ki)} > 0$ for all $(ki) \in A_s$. Under BIG sampling, we have

$$\hat{\theta}_p = \sum_{(ki)\in A_s} \frac{I_{(ki)}}{p_{(ki)}} \cdot \frac{w_{ki} y_i}{\pi_k} = \sum_{k\in s} \frac{Z_k}{\pi_k} \qquad \text{and} \qquad Z_k = \sum_{i\in\alpha(k)} \frac{I_{(ki)}}{p_{(ki)}} w_{ki} y_i \ .$$

Although this looks like the Zhat $\hat{\theta}_z$, with the constructed measure Z_k instead of z_k, there is a key difference: unlike z_k that is a constant of sampling, Z_k is a variable. Birnbaum and Sirken (1965) did not provide an expression of the variance of their priority-rule estimator, but indicated that it is unwieldy.

4.2.3 More on Phat $\hat{\theta}_p$

Below we derive $V(\hat{\theta}_p)$ via the general expression (4.4). We will show that the Phat can be biased as the sample size increases, and provide a condition for unbiasedness. The problem of the bias of the Phat was not mentioned by Birnbaum and Sirken (1965).

Proposition 4.2.1. For the variance of $\hat{\theta}_p$ by (4.4), we have

$$V(\hat{\theta}_p) = \sum_{(ki)\in A} \sum_{(lj)\in A} \left(\frac{\pi_{kl} p_{(ki)(lj)}}{\pi_k \pi_l p_{(ki)} p_{(lj)}} - 1 \right) w_{ki} w_{lj} y_i y_j \tag{4.5}$$

where $p_{(ki)(lj)} = \Pr\big[I_{ki}I_{lj} = 1 | \delta_k\delta_l = 1\big]$ is the conditional probability that both (ki) and (lj) are prioritised given that both k and l are in the sample s.

Proof.

$$\begin{aligned}
V(\hat{\theta}_p) &= \sum_{(ki)\in A}\sum_{(lj)\in A} E\left(\frac{\delta_k I_{ki}\delta_l I_{lj}}{\pi_k p_{(ki)}\pi_l p_{(lj)}}\right) w_{ki}w_{lj}y_iy_j - \theta^2 \\
&= \sum_{(ki)\in A}\sum_{(lj)\in A} \pi_{kl}\ E\left(\frac{I_{ki}I_{lj}}{\pi_k\pi_l p_{(ki)}p_{(lj)}}\middle| \delta_k\delta_l = 1\right) w_{ki}w_{lj}y_iy_j - \theta^2 \\
&= \sum_{(ki)\in A}\sum_{(lj)\in A} \frac{\pi_{kl}p_{(ki)(lj)}}{\pi_k\pi_l p_{(ki)}p_{(lj)}} w_{ki}w_{lj}y_iy_j - \theta^2 \\
&= \sum_{(ki)\in A}\sum_{(lj)\in A} \left(\frac{\pi_{kl}p_{(ki)(lj)}}{\pi_k\pi_l p_{(ki)}p_{(lj)}} - 1\right) w_{ki}w_{lj}y_iy_j\ .
\end{aligned}$$

□

The difference to the variance of Zhat can be given by

$$\begin{aligned}
V(\hat{\theta}_p) - V(\hat{\theta}_z) &= \sum_{(ki)\in A}\sum_{(lj)\in A} \left(\frac{p_{(ki)(lj)}}{p_{(ki)}p_{(lj)}} - 1\right)\frac{\pi_{kl}}{\pi_k\pi_l} w_{ki}w_{lj}y_iy_j \\
&= \sum_{(ki)\in A}\sum_{(lj)\in A} Cov\Big(\frac{I_{ki}}{p_{(ki)}},\ \frac{I_{lj}}{p_{(lj)}}\Big|\delta_k\delta_l = 1\Big)\frac{\pi_{kl}}{\pi_k\pi_l} w_{ki}w_{lj}y_iy_j\ .
\end{aligned}$$

Thus, as long as the covariances are not all positive or negative, neither will the Zhat dominate the Phat in terms of efficiency, nor the other way around.

An unbiased variance estimator can be given by:

$$\hat{V}(\hat{\theta}_p) = \sum_{(ki)\in A_s}\sum_{(lj)\in A_s} \Big(\frac{\pi_{kl}p_{(ki)(lj)}}{\pi_k\pi_l p_{(ki)}p_{(lj)}} - 1\Big)\frac{w_{ki}w_{lj}y_iy_j}{\pi_{kl}}\ . \tag{4.6}$$

Illustration Let us make an illustration of BIG sampling on the population graph in Figure 4.1. Suppose simple random sampling (SRS) without replacement of s from F. To compute the probability of prioritising an edge (ki) requires the knowledge of the number of units in F with higher priority. For each $i \in \alpha(s)$, let $d_{k(i)} = \sum_{k'\in F; k'<k} a_{k'i}$, for which ancestral observation is required. We have,

then,

$$\pi_{(ki)} = \Pr\big[(ki) \in A_s\big] = \Pr(\delta_k = 1) = \pi_k \equiv m/M \ ,$$
$$p_{(ki)} = \Pr(I_{ki} = 1 | \delta_k = 1) = \binom{M-1-d_{k(i)}}{m-1} \Big/ \binom{M-1}{m-1} \ .$$

Table 4.1: Probability $p_{(ki)}$ for population BIG in Figure 4.1.

	5	6	7	8	9	10	11
1	-	-	-	-	-	1	-
2	1	-	1	-	1	-	-
3	-	1	-	1	0.67	0.67	1
4	-	-	0.67	-	0.33	0.33	0.67

The inclusion probability of $(ki) \in A_{sp}$ is given by $\pi_k p_{(ki)}$. For the population graph in Figure 4.1, the conditional probabilities $p_{(ki)}$ of being prioritised are given in Table 4.1. Birnbaum and Sirken (1965) did not provide expressions of the second-order probabilities of being included in A_{sp}. These are given by $\pi_{kl} p_{(ki)(lj)}$, where

$$p_{(ki)(lj)} = \begin{cases} p_{(ki)} & \text{if } i = j, k = l \\ 0 & \text{if } i = j, k \neq l \\ \binom{M-1-d_{k(i,j)}}{m-1} \Big/ \binom{M-1}{m-1} & \text{if } i \neq j, k = l \\ \binom{M-2-d_{k(i),l(j)}}{m-2} \Big/ \binom{M-2}{m-2} & \text{if } i \neq j, k \neq l \text{ with } |\beta_i^k \cap l| + |\beta_j^l \cap k| = 0 \\ 0 & \text{if } i \neq j, k \neq l \text{ with } |\beta_i^k \cap l| + |\beta_j^l \cap k| > 0 \end{cases} \tag{4.7}$$

where β_i^k is the set of the neighbours of i which have higher priority than k, and $d_{k(i,j)} = |\beta_i^k \cup \beta_j^k|$ is the number of units in $\beta_i \cup \beta_j$ which have higher priority than k, and $d_{k(i),l(j)} = |\beta_i^k \cup \beta_j^l|$. For instance, with $m = 2$, the variances of the three basic estimators of $\theta = |U|$, for $y_i \equiv 1$, are $V(\hat{\theta}_y) = 3.986$, $V(\hat{\theta}_z) = 5.370$ with equal-share weighting, and $V(\hat{\theta}_p) = 3.064$ by the priority-rule of Birnbaum and Sirken (1965).

Bound for unbiasedness There are circumstances where $p_{(ki)} = 0$, i.e. a edge has zero probability of being prioritised, such that the Phat is biased. Take for example the case when a motif i in U is adjacent to all the sampling units in F; the edge between i and its ancestor enumerated as the last one in F will never be prioritised, if the sample size is greater than 1. The next proposition provides

a general condition: essentially, the Phat will be biased, if there exists an ancestor of some motif i, which has zero probability of being the only one among β_i in the sample s. Generally, for given a BIG, the likelihood of this happening increases with the size of s.

Proposition 4.2.2. The Phat estimator is biased if there exists at least a motif $i \in U$, such that:

$$|\beta_i| > 1 \qquad \text{and} \qquad \Pr\Big(\sum_{k\in\beta_i} \delta_k \leq 1\Big) = 0 \ . \tag{4.8}$$

Proof. Let i be a motif with $|\beta_i| > 1$. Let $h = \max(\beta_i)$ and $p_{(hi)} = \Pr(I_{(hi)} = 1|\delta_h = 1)$. Assume that $h \in s$. Because $\Pr\Big(\sum_{k\in\beta_i} \delta_k \leq 1\Big) = 0$, then it must exist at least another ancestor of i, say h', where $h' \in s$ and $h' < h$ by definition of h. It follows that $h \neq \min(\beta_i \cap s)$ for all possible s containing h and consequently $p_{(hi)} = 0$, i.e. k is sampled but never prioritised. In this case, the Phat is biased. □

Remark Under SRS of the initial sample from F, the probability in Equation 4.8 can be easily calculated from:

$$\Pr\Big(\sum_{k\in\beta_i} \delta_k = 1\Big) = \binom{M - |\beta_i|}{m-1} \Big/ \binom{M}{m} \ .$$

Therefore, under SRS, the Phat is biased for any m such as $m > M - |\beta_i| + 1$.

4.3 Incidence weighting estimator

The proposed class of linear estimators under BIG sampling, called the incidence weighting estimator (IWE) is presented in this section, which encompasses all the three estimators described in the previous section.

4.3.1 Definition

Given the sample BIG, $G_s = (s \cup \alpha(s); A_s)$, let $\{W_{ki}; (ki) \in A_s\}$ be the *incidence weights*, where the capital letter W is used to emphasise that the incidence weights are not necessarily constants of sampling. The IWE based on $W = \{W_{ki}; (ki) \in$

$A_s\}$ is given by

$$\hat{\theta} = \sum_{(ki)\in A_s} \frac{W_{ki} y_i}{\pi_{(ki)}} = \sum_{k\in s} \frac{Z_k}{\pi_k} = \sum_{i\in\alpha(s)} \gamma_{(i)} y_i \ , \tag{4.9}$$

where

$$Z_k = \sum_{i\in\alpha_k} W_{ki} y_i \qquad \text{and} \qquad \gamma_{(i)} = \sum_{k\in\beta_i\cap s} \frac{W_{ki}}{\pi_k} \ . \tag{4.10}$$

4.3.2 Theory

We denote by t any quantity apart from the sample graph G_s, which may be used for the construction of the incidence weights. The properties of the IWE $\hat{\theta}$ will be assessed with respect to the joint distribution of (s,t), denoted by $p(s,t)$. In this paper we consider only t, which is such that $p(s,t) = p(s)$, i.e., the sampling distribution of s. For instance, $t = d_i$, the degree of motif i in the population graph, which is a constant associated with i and is observed given ancestral observation procedure for any $i \in \alpha(s)$; or $t = d_{i,s}$, the degree of motif i in the sample graph, which is a function of the sample graph G_s.

Remark It is in principle possible to allow t to be random given s, with conditional distribution $p(t|s)$, such that the properties of the IWE are evaluated with respect to $p(s,t) = p(s)p(t|s)$. However, any such estimator can be subjected to the Rao-Blackwell method, conditional on the sample graph G_s which depends only on $p(s)$, and we have not been able to devise an estimator which leads to efficiency gains that can justify the extension. We therefore do not pursue this line of development here.

Proposition 4.3.1. The IWE by (4.9) is unbiased for θ by (4.1) provided, for each $i \in U$,

$$\sum_{k\in\beta_i} E(W_{ki}|\delta_k = 1) = 1 \ . \tag{4.11}$$

The condition (4.11) implies

$$\sum_{k\in\beta_i} \frac{1}{\pi_k} \sum_{s;\ k\in s} W_{ki}(s) p(s) = \sum_{s;\ i\in\alpha(s)} p(s) \left(\sum_{k\in s\cap\beta_i} \frac{W_{ki}(s)}{\pi_k} \right) = 1 \ . \tag{4.12}$$

Remark Because the second term of (4.12) can be written as

$$\sum_{s;\ i\in\alpha(s)} p(s)\left(\sum_{k\in s\cap\beta_i}\frac{W_{ki}(s)}{\pi_k}\right) = \sum_{s;\ i\in\alpha(s)} p(s)\gamma_{(i)} = \pi_{(i)}E(\gamma_{(i)}|\delta_{(i)}=1) = 1\ ,$$

we have, in terms of the quantities in the definition of IWE (4.9):

$$\sum_{k\in\beta_i} E(W_{ki}|\delta_k=1) = 1 \qquad \text{or} \qquad E(\gamma_{(i)}|\delta_{(i)}=1) = \frac{1}{\pi_{(i)}}\ .$$

Proof. The expectation of $\hat{\theta}$ with respect to $p(s)$ is given by

$$\begin{aligned} E(\hat{\theta}) &= \sum_{k\in F}\frac{1}{\pi_k}E(\delta_k Z_k) = \sum_{k\in F}\frac{E(\delta_k)}{\pi_k}E(Z_k|\delta_k=1) = \sum_{k\in F}E(Z_k|\delta_k=1) \\ &= \sum_{k\in F}\sum_{i\in\alpha_k}E(W_{ki}|\delta_k=1)y_i = \sum_{i\in U}y_i\sum_{k\in\beta_i}E(W_{ki}|\delta_k=1) = \theta\ , \end{aligned}$$

where the first equality in the last line above follows from (4.10), and the third equality follows from the stipulation of this proposition. □

Proposition 4.3.2. The variance of an unbiased IWE can be given by:

$$V(\hat{\theta}) = \sum_{k\in F}\sum_{h\in F}\left(\frac{\pi_{kl}}{\pi_k\pi_l}\sum_{i\in\alpha(k)}\sum_{j\in\alpha(l)}E(W_{(ki)}W_{(lj)}|\delta_k\delta_l=1) - 1\right)y_i y_j\ . \qquad (4.13)$$

Proof. By definition we have:

$$\begin{aligned} V(\hat{\theta}) &= \sum_{ki\in F}\sum_{l\in F}\mathrm{Cov}\left(\frac{\delta_k Z_k}{\pi_k},\frac{\delta_l Z_l}{\pi_l}\right) \\ &= \sum_{k\in F}\sum_{l\in F}\left(\frac{E(\delta_k\delta_l Z_k Z_l)}{\pi_{kl}} - E\left(\frac{\delta_k Z_k}{\pi_k}\right)E\left(\frac{\delta_l Z_l}{\pi_l}\right)\right) \\ &= \sum_{k\in F}\sum_{l\in F}\left(\frac{\pi_{kl}}{\pi_k\pi_l}E(Z_k Z_l|\delta_k\delta_l=1) - E(Z_k|\delta_k=1)E(Z_l|\delta_l=1)\right) \\ &= \sum_{k\in F}\sum_{l\in F}\left(\sum_{i\in U}\sum_{j\in U}E(W_{(ki)}W_{(lj)}|\delta_k\delta_l=1) - \sum_{i\in U}E(W_{(ki)}|\delta_k=1)\sum_{j\in U}E(W_{(lj)}|\delta_l=1)\right)y_i y_j. \end{aligned}$$

Equation (4.13) follows from the unbiasedness condition. □

Proposition 4.3.3. An unbiased estimator of $V(\hat{\theta})$ is given by

$$\hat{V}(\hat{\theta}) = \sum_{k\in s}\sum_{h\in s}\left(\frac{\pi_{kl}}{\pi_k\pi_l}\sum_{i\in\alpha(k)}\sum_{j\in\alpha(l)} W_{(ki)}W_{(lj)} - 1\right)\frac{y_i y_j}{\pi_{kl}} \ . \tag{4.14}$$

Proof. By definition $W_{ki}W_{lj}$ is an unbiased estimator of $E(W_{ki}W_{lj}|\delta_k\delta_l = 1)$, for any $k, l \in s$. □

4.4 Unbiased IWE

Below we first show the three estimators defined in section 4.2.2 can be casted as unbiased IWEs. We will then discuss some variations of them.

Zhat Let w_{ki} be constant for $(ki) \in A$ such that $\sum_{k\in\beta_i} w_{ki} = 1$. When $w_{ki} = 1/d_i$, the IWE is the multiplicity estimator of Birnbaum and Sirken (1965).

Yhat The HT estimator (as a Yhat) is obtained by using any W_{ki} satisfying

$$\sum_{k\in s\cap\beta_i}\frac{W_{ki}(s)}{\pi_k} = \frac{1}{\pi_{(i)}} \ . \tag{4.15}$$

Notice that (4.15) is satisfied by any

$$W_{ki}(s) = \frac{c_s\pi_k}{\pi_{(i)}} \qquad \text{where} \qquad \sum_{k\in s\cap\beta_i} c_s = 1 \ ,$$

A possible choice is $c_s = 1/d_{i,s}$; but one obtains the same HT-estimator in any case.

Phat Given any fixed w_{ki} such that $\sum_{k\in\beta_i} w_{ki} = 1$, let $W_{ki} = w_{ki}H_{ki}$. Then (4.12) holds for the first term if, for each $k \in \beta_i$:

$$\sum_{s;\ k\in s} H_{ki}(s)\frac{p(s)}{\pi_k} = E(H_{ki}|\delta_k = 1) = 1 \ .$$

The Zhat can be considered as an unbiased IWE with $H_{ki} = 1$. The weights used for the Phat belongs to this type, given $|\beta_i \cap s| > 1$. Instead of attaching weights to all the sample edges incident to motif i, one could assign a non-zero weight only to one of them, depending on the observed sample. Let an indicator

variable be defined as $I_{ki} = 1$ if $W_{ki} \neq 0$ and $I_{ki} = 0$ if $W_{ki} = 0$. Then, H_{ki} is given as

$$H_{ki} = \frac{I_{ki}(s)}{p_{(ki)}} \qquad \text{where} \qquad p_{(ki)} = \Pr(I_{ki} = 1 | \delta_k = 1) = \frac{\sum_{s;\ I_k(s)=1} p(s)}{\sum_{s;\ k \in s} p(s)} \ . \quad (4.16)$$

We call $W_{ki} = w_{ki} H_{ki}$ with H_{ki} given by (4.16) the *priority weights*.

4.4.1 Zhat

Below are some choices of fixed weights $W_{ki} = w_{ki}$ that yield different Zhats.

Equal-share weights The equal-share weights are given by

$$w_{ki} = \frac{1}{d_i} \qquad \text{with} \qquad t = d_i \ ,$$

where d_i is the degree of the motif i. The equal-share weights have been commonly used in the literature, and are known as *multiplicity weights*.

Inverse-degree weights We define the inverse-degree weights as:

$$w_{ki} = \frac{1}{d_k} \Big/ \sum_{l \in \beta_i} \frac{1}{d_l} \qquad \text{with} \qquad t = \{d_k\}_{k \in \beta_i} \ .$$

Under simple random sampling (SRS) without replacement of s from F, they provide a choice of weighting which could potentially reduce the variance of the estimator, by making the constructed z_k as similar as possible. On the one hand, the weight w_{ki} is increased compared to $1/d_i$ under equal-share weighting, provided k has relatively lower degree compared to the other units in β_i. On the other hand, the measure z_l of another unit $l \in s$ will receive 'shares' from more motifs in U than z_k, provided $d_l > d_k$ and $l, k \in \beta_i$. Thus, these weights can possibly reduce the population variance of $z_k = \sum_{i \in \beta_i} w_{ki} y_{(i)}$.

Power of inverse-degree weights The inverse-degree weights above defined can be generalised as follow:

$$w_{ki} = \left(\frac{1}{d_k}\right)^{\alpha} \Big/ \sum_{l \in \beta_i} \left(\frac{1}{d_l}\right)^{\alpha} \qquad \text{with} \qquad t = \{d_k, \alpha\}_{k \in \beta_i} \ .$$

Notice that the feasibility of a particular choice of any fixed weights depends on the information available from sampling, and therefore on the observation procedure employed. The multiplicity weights are those requiring the minimum information, since for their computation only the number of the ancestor of each sample motif is needed. Whereas, information about the non-sampled ancestors of each sample motif is required to construct inverse-degree weights and the power of inverse-degree weights.

Illustration Let us compare the different choices above for estimating $N = |U|$ for the graph in Figure 4.1. Suppose SRS of s from F of size $m = 2$. The equal-share, inverse-degree and power of inverse-degree weights with $\alpha = 2$ or3, together with their corresponding constructed measures z_k are given in Table 4.2.

Table 4.2: Weights, measures, variances for Fig. 4.1 using three choices of multiplicity weighting: ES = equal-share; ID = inverse-degree; ID2 = power of inverse-degree weights with $\alpha = 2$ and ID3 with $\alpha = 3$.

	$w_{1,10}$	$w_{2,5}$	$w_{2,7}$	$w_{2,9}$	$w_{3,10}$	$w_{3,8}$	$w_{3,11}$	$w_{3,9}$	$w_{3,6}$	$w_{4,7}$	$w_{4,10}$	$w_{4,11}$	$w_{4,9}$
ES	0.33	1	0.5	0.33	0.33	1	0.5	0.33	1	0.5	0.33	0.5	0.33
ID	0.69	1	0.57	0.43	0.14	1	0.44	0.26	1	0.43	0.17	0.56	0.32
ID2	0.90	1	0.64	0.52	0.04	1	0.39	0.19	1	0.36	0.06	0.61	0.29
ID3	0.98	1	0.70	0.61	0.007	1	0.34	0.14	1	0.30	0.013	0.66	0.25

	z_1	z_2	z_3	z_4	S_w^2	$V(\hat{\theta}_z)$
ES	0.33	1.83	3.17	1.67	1.34	5.37
ID	0.69	2	2.83	1.48	0.81	3.26
ID2	0.91	2.16	2.61	1.32	0.60	2.41
ID3	0.98	2.31	2.48	1.23	0.57	2.28

The variances of the IWEs with equal-share, inverse-degree and power inverse-degree weights with $\alpha = 2$ and 3 are respectively 5.37, 3.26, 2.41 and 2.28. The power inverse-degree weights can possibly reduce the variance of the IWE $\hat{\theta}_z$, according to the choice of α.

4.4.2 HT weights

Here we consider the HT estimator given by:

$$W_{ki} = \frac{\pi_k}{d_{i,s}\pi_{(i)}} \ ,$$

where $d_{i,s} = \sum_{k \in \beta_i \cap s} a_{ki}$ is degree of i in sample graph.

Illustration The variance of the Yhat is 3.98. Table 4.3 shows the HT incidence weights W_{ki} and their corresponding Z_k measures for the graph in Fig. 4.1.

Table 4.3: HT incidence weights and corresponding measures for the BIG in Fig. 4.1.

s	$W_{1,10}$	$W_{2,5}$	$W_{2,7}$	$W_{2,9}$	$W_{3,10}$	$W_{3,8}$	$W_{3,11}$	$W_{3,9}$	$W_{3,6}$	$W_{4,7}$	$W_{4,10}$	$W_{4,11}$	$W_{4,9}$
$\{1,2\}$	0.5	1	0.6	0.5	-	-	-	-	-	-	-	-	-
$\{1,3\}$	0.25	-	-	-	0.25	1	0.6	0.5	1	-	-	-	-
$\{1,4\}$	0.25	-	-	-	-	-	-	-	-	0.6	0.25	0.6	0.5
$\{2,3\}$	-	1	0.6	0.25	0.5	1	0.6	0.25	1	-	-	-	-
$\{2,4\}$	-	1	0.3	0.25	-	-	-	-	-	0.3	0.5	0.6	0.25
$\{3,4\}$	-	-	-	-	0.25	1	0.3	0.25	1	0.6	0.25	0.30	0.25

s	Z_1	Z_2	Z_3	Z_4
$\{1,2\}$	0.5	2.1	-	-
$\{1,3\}$	0.25	-	3.35	-
$\{1,4\}$	0.25	-	-	1.95
$\{2,3\}$	-	1.85	3.35	-
$\{2,4\}$	-	1.55	-	1.65
$\{3,4\}$	-	-	2.8	1.40

4.4.3 Priority weights

The priority weights proposed by Birnbaum and Sirken (1965) for the Phat estimator, as described in Section 4.2.2 belongs to this class of incidence weights, by setting:

$$H_{(ki)} = \frac{I_{(ki)}}{p_{(ki)}} \ ,$$

where $I_{(ki)}$ is the prioritization indicator and $p_{(ki)} = \Pr(I_{(ki)} = 1 | \delta_k = 1)$.

Illustration Using the Birnbaum and Sirken (1965) rule we have that $V(\hat{\theta}_r) = 3.064$. Furthermore, the variance of the same estimator after the sampling units are arranged in descending order of their degree is 2.555, whereas it becomes 6.315 when using the units are re-arranged in the 'opposite' order. The priority weights and their corresponding Z_k measures in Table 4.4 under both random and descending order of the sampling frame.

Table 4.4: Priority weights and measures for the graph in Fig. 4.1 under the unorder and descending ordering of the sampling frame.

Random order of sampling frame													
s	$W_{1,10}$	$W_{2,5}$	$W_{2,7}$	$W_{2,9}$	$W_{3,10}$	$W_{3,8}$	$W_{3,11}$	$W_{3,9}$	$W_{3,6}$	$W_{4,7}$	$W_{4,10}$	$W_{4,11}$	$W_{4,9}$
$\{1,2\}$	0.33	1	0.50	0.33	-	-	-	-	-	-	-	-	-
$\{1,3\}$	0.33	-	-	-	0	1	0.50	0.50	1	-	-	-	-
$\{1,4\}$	0.33	-	-	-	-	-	-	-	-	0.75	0	0.75	1
$\{2,3\}$	-	1	0.50	0.33	0.50	1	0.50	0	1	-	-	-	-
$\{2,4\}$	-	1	0.50	0.33	-	-	-	-	-	0	1	0.75	0
$\{3,4\}$	-	-	-	-	0.50	1	0.50	0.50	1	0.75	0	0	0
Descending order of the sampling frame													
s	$W_{4,10}$	$W_{3,5}$	$W_{3,7}$	$W_{3,9}$	$W_{1,10}$	$W_{1,8}$	$W_{1,11}$	$W_{1,9}$	$W_{1,6}$	$W_{2,7}$	$W_{2,10}$	$W_{2,11}$	$W_{2,9}$
$\{1,2\}$	-	-	-	-	0.33	1	0.5	0.33	1	0.5	0	0	0
$\{1,3\}$	-	1	0.75	0	0.33	1	0.50	0.33	1	-	-	-	-
$\{1,4\}$	0	-	-	-	0.33	1	0.5	0.33	1	-	-	-	-
$\{2,3\}$	-	1	0	0	-	-	-	-	-	0.5	0.5	0.75	0.5
$\{2,4\}$	0	-	-	-	-	-	-	-	-	0.5	0.5	0.75	0.5
$\{3,4\}$	1	1	0.75	1	-	-	-	-	-	-	-	-	-

	Random order of F				Descending order of F			
s	Z_1	Z_2	Z_3	Z_4	Z_1	Z_2	Z_3	Z_4
$\{1,2\}$	0.33	1.83	-	-	3.17	0	-	-
$\{1,3\}$	0.33	-	3.00	-	3.17	-	1.75	-
$\{1,4\}$	0.33	-	-	2.50	3.17	-	-	0
$\{2,3\}$	-	1.83	3.00	-	-	2.25	1	-
$\{2,4\}$	-	1.83	-	1.75	-	2.25	-	0
$\{3,4\}$	-	-	3.50	0.75	-	-	2.75	1

4.4.4 Discussion on the efficiency of the different unbiased IWE

Finally, Table 4.5 provides a summary of the variances of the IWEs for $N = |U|$ in Figure 4.1. The IWEs by inverse-degree and power of inverse-degree weighting perform best compared to the others. Under SRS of s, any choice of fixed weights which reduces the population variance of z_k's will result in a gain of efficiency, compared to the standard multiplicity weights. The power of inverse-degree weights provide a general means for reducing the variability amongst the constructed z-measures.

The priority weights under the Birnbaum and Sirken (1965)'s rule yields a more efficient estimator when the sampling frame is organized by descending order of the degree. Since the efficiency of a given ordering depends on the population graph, further investigation is required to understand this relationship in general

Table 4.5: The true variances of the IWE for $\theta = N$ by different choices of weights under SRS with $m = 2$ of F in the graph in Figure 4.1. The ordering of the sampling frame for the priority estimator is given by: (I) - random; (II) - descending and (III) - ascending.

	$\hat{\theta}_z$ (ES)	$\hat{\theta}_z$ (ID)	$\hat{\theta}_z$ (ID2)	$\hat{\theta}_z$ (ID3)	$\hat{\theta}_r$ (I)	$\hat{\theta}_r$ (II)	$\hat{\theta}_w$ (III)	$\hat{\theta}_y$
$V(\hat{\theta})$	5.37	3.25	2.41	2.28	3.06	2.55	6.32	3.98

terms.

The HT estimator is not as good as inverse-degree-based weights, but also not as bad as the equal-share weights or the priority weights under the ascending ordering of the sampling units by their degrees. It seems to remain a default benchmark under BIG sampling, against which the other unbiased estimators can be assessed. The insight that the HT estimator is a special case of unbiased IWE is potentially important for future research on this topic.

4.5 Simulations

To further illustrate and explore the IWEs by simulations, we construct two graphs, denoted by $G_1 = (F, U; A_1)$ and $G_2 = (F, U; A_2)$, respectively, where $|F| = 54$ and $N = |U| = 310$. The two graphs are set to have the same number of edges, $|A_1| = |A_2| = 1200$, but different incidence relationships. In A_1, the distribution of d_k, for $k \in F$, is relatively uniform over a small range of values; in A_2, the distribution of d_k is constructed to be more skewed and asymmetrical. The two distributions of d_k are shown in Figure 4.2.

Suppose we are interested in estimating the total number of motifs $\theta = N$. For these simulations we assume SRS from F with various sample sizes $m = 2, 5, 8 \ldots, 53$, and the incident ancestral observation procedure.

We consider the following choices of incidence weights:

- the three types of fixed weights: equal-share (ES), inverse-degree (ID) and square of inverse-degree (ID2);
- BS-priority weights as given in Birnbaum and Sirken (1965);
- the HT weights that yields the HT estimator.

Moreover, for the priority estimator, we consider three different orderings of the frame: one is given by the frame as initially constructed, where the units can

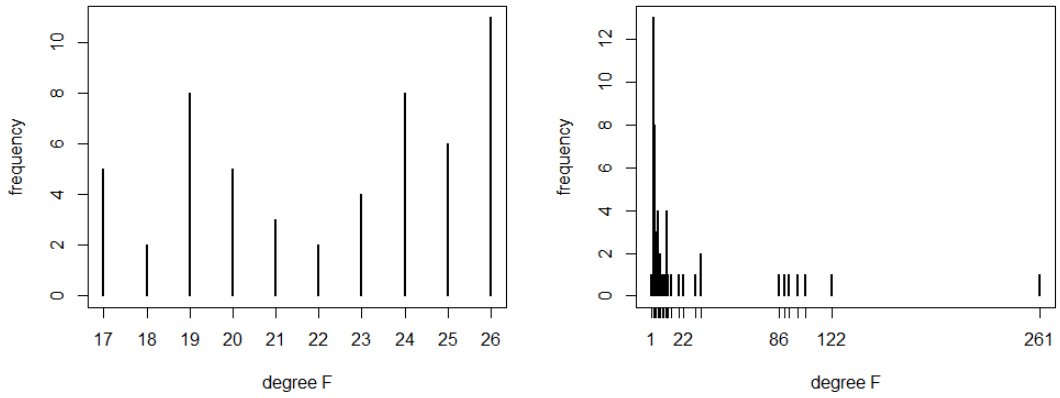


Figure 4.2: The two observed degree distributions for the units set F in G_1 and G_2.

be considered to be arranged in a random order, and the other two are obtained from rearranging the units by descending and ascending order of the degree d_k, for $k \in F$, respectively.

The plots in Figure 4.3 show the results of 10000 simulations for the IWE with fixed weights for both graph G_1 and G_2. The average of the estimates for the different choices of IWE are plotted against the increasing sample sizes with associated Monte Carlo error. It can be seen that the IWE which uses the fixed weights is unbiased and, as the sample size increases, the variance reduces to zero.

Next, Figure 4.4 shows the true variances for these estimators, again plotted against the sample sizes and with associated 95% confidence interval for the Monte Carlo error. It is visible that the unequal-share weights are more efficient, as can be expected under SRS, where the ID2-weights appear to have the smallest variance.

A more peculiar situation is presented in the case of the BS-priority weights, as shown in Figure 4.5. As previously explained, when the priority weights are used, the IWE can become biased beyond a certain threshold of sample size. For the graph with uniform degree distribution of the sampling units, where the maximum degree of the motifs in U is 10, this occurs at $m = 45$; when the degree distribution of the sampling units is skewed, where the maximum degree of the motifs is 9, this occurs at $m = 46$.

Next, the variance of the BS-priority IWE increases as the sample size increases, so that this particular estimator seems to perform well only for small sample sizes. This aspect of the Phat did not emerge in the illustration using the graph in Figure 4.1, due to the small frame size. Moreover, the ordering of the sampling units matters. When the sampling units are arranged in descending ordering of d_k, for $k \in F$, there seems to be an improvement in efficiency (see Figure 4.9), as seen in the previous illustrations; whereas ascending order entails loss of efficiency of the BS-priority estimator.

In Section 4.3.2, a general variance estimator is given by (4.14), which uses the observed values of the incidence weights as the estimates of their conditional expectation. However, for a specific IWE, it may be possible to analytically derive an expression for the corresponding $E(W_{ki}W_{lj}|\delta_k\delta_l = 1)$, which is the case with the BS-priority estimator, where the variance estimator with the exact expression of $E(W_{ki}W_{lj}|\delta_k\delta_l = 1)$ is given by (4.6). In Figure 4.6 we have plotted the variances estimators with associated Monte Carlo error together with the true value of the variances for the graph G_1 and under the three ordering of the frame. Clearly, as one would expect, the variance estimator by (4.6) is more precise.

Figure 4.8 shows the average of the HT estimates and their variances with associated 95% confidence interval.

Finally, in Figure 4.9 the variances of the six IWEs are plotted together against the increasing sample size for both G_1 and G_2. Immediately we notice that the variance is much larger for G_2 than for G_1. Moreover, a similar pattern emerges for both graphs G_1 and G_2, but more pronounced for G_2. The inverse-degree weights seems to perform better, together with the power inverse-degree weights. For larger sample sizes, however, the HT estimator is more efficient than the inverse-degree and power inverse-degree weights. The BS-priority estimator perform well only for small sample sizes, and especially if the frame is rearranged in descending ordering of the degrees of sampling units, when its variance can be lower than the IWE making use of the unequal-share weights, for the graph G_2. The estimators with higher variance are the one with equal-share weights and the BS-priority weights when the frame is organised in ascending order the degrees of sampling units.

4.6 Concluding remarks

In the above, we proposed a general linear class of IWEs for any situation that can be represented as sampling on a BIG, based on the incidence relationships underlying the sampling. The estimators presented by Birnbaum and Sirken (1965) for indirect sampling are special cases of the proposed class, and their underlying ideas generalised and synthesised into a unified condition of design unbiasedness. The BIG representation of unconventional sampling methods has proven to be extremely useful in order to simplify the problem; and the definition of IWE unifies the existing estimators under a broader theory of estimation on BIGs. In so doing we reveal the potentials of sampling strategy consisting of BIG and IWE for future research.

The performance of the IWE depends on the definition of the corresponding incidence weights. In principle, many more incidence weights can be proposed which satisfy the unbiasedness of the IWE. The general definition of IWE includes also those which are based on sample-dependent weights, such as the priority estimator and the HT estimator. Importantly, we have shown that the HT estimator is an example of IWE, which has not been discovered previously in the literature. It can be noted that the gain of efficiency is often associated with incidence weights that require the observation of a greater portion of the graph. This is not surprising, since more information is utilised in such situations.

Further investigation is needed to obtain a better theoretical understanding of the potentials of using sample-dependent weights or additional characteristics of the graph. For instance, a general variance estimator has been proposed. But it is not precise for sample-dependent weights, as seen in the numerical results for the priority estimator. Although when the conditional expectation involved in the variance can be analytically derived for a given IWE, a more accurate variance estimator can be obtained. The simulation results indicate that the variance of the priority estimator may decrease with sample size that is relatively small, but it quickly increases with the sample size beyond some threshold value. This is another example where more theoretical understanding is desirable by future research.

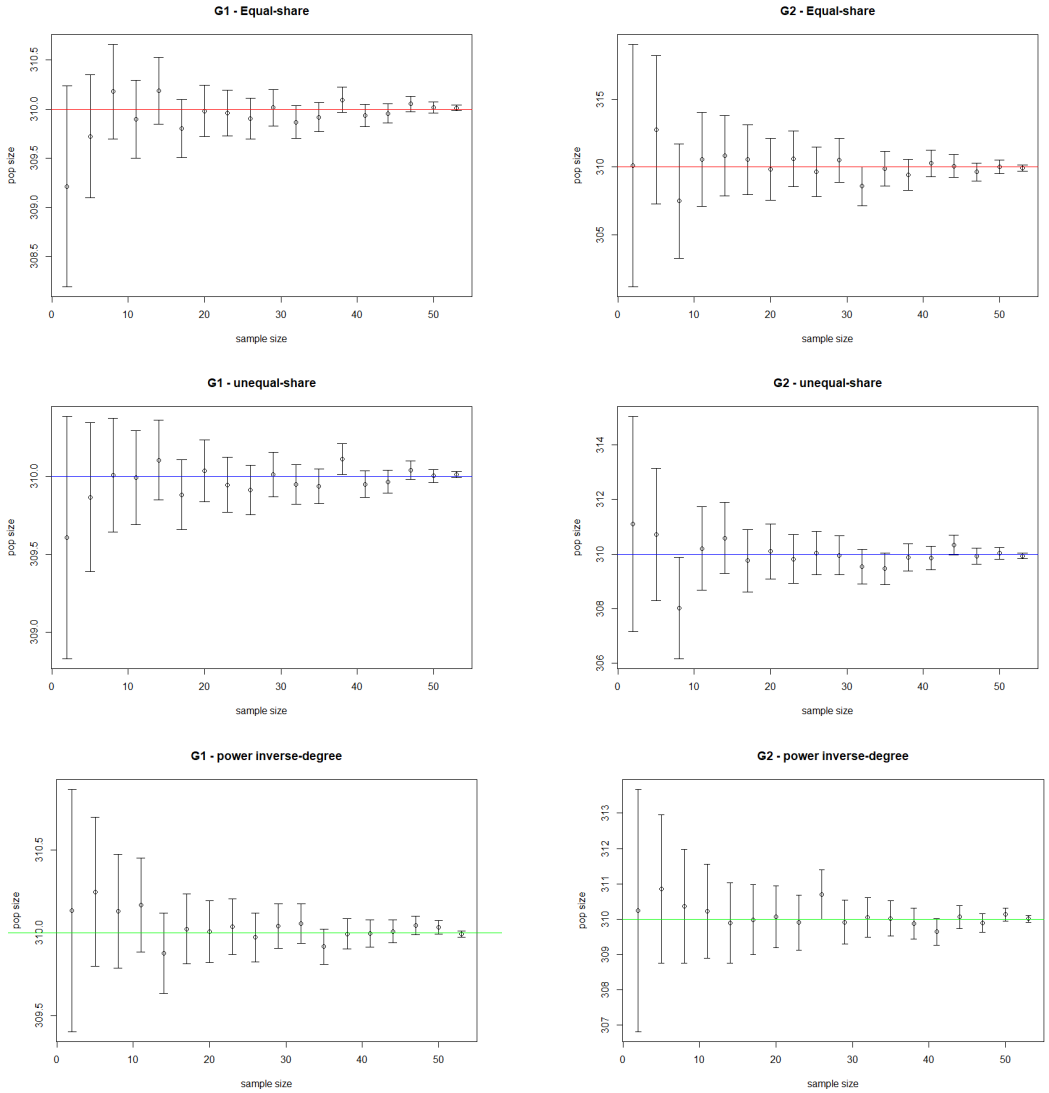


Figure 4.3: The average of the estimates with associated Monte Carlo error for the IWE plotted against the increasing sample sizes for G_1 and G_2 considering the three ordering of the frame. Three types of multiplicity weighting are used: Equal-Share, Inverse-Degree and Power of Inverse-Degree weighting

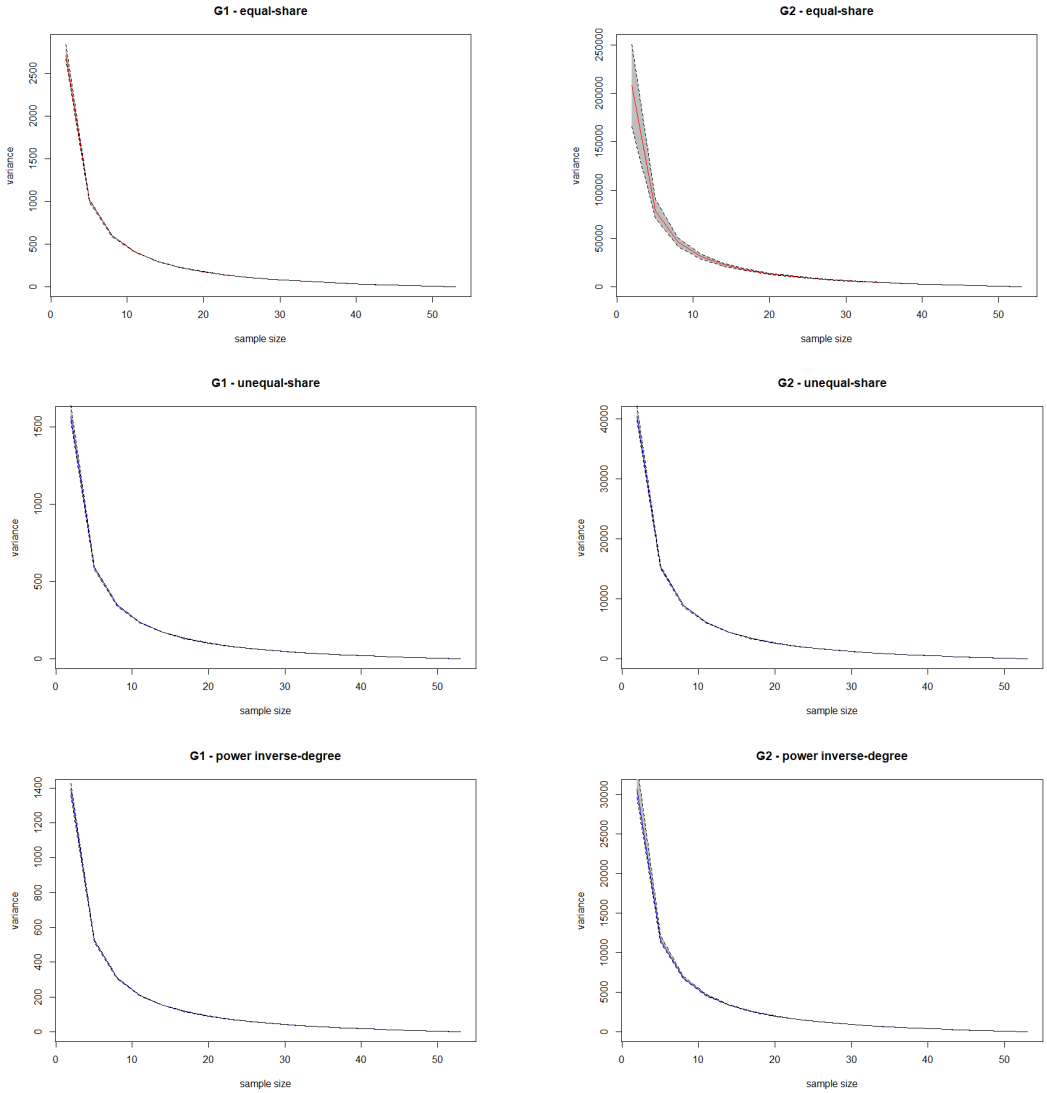


Figure 4.4: The variances estimator and the true variances for the IWE with fixed weights plotted against the increasing sample sizes for both graph G_1 and G_2. Three types of multiplicity weighting are used: Equal-Share, Inverse-Degree and Power of Inverse-Degree weights.

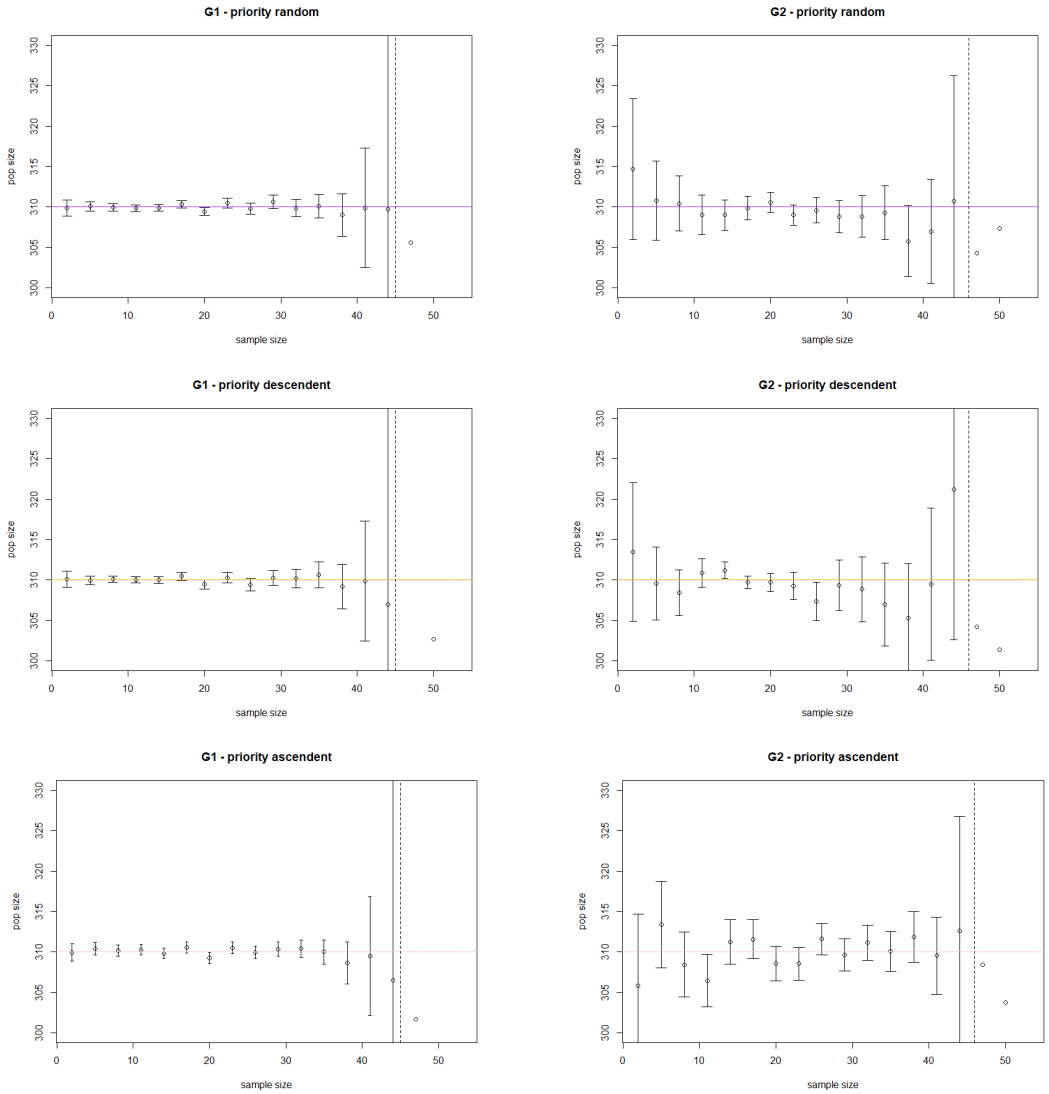


Figure 4.5: The average of the estimates with associated Monte Carlo error for the IWE plotted against the increasing sample sizes for G_1 and G_2 considering the priority weighting. Three ordering of the frame are considered.

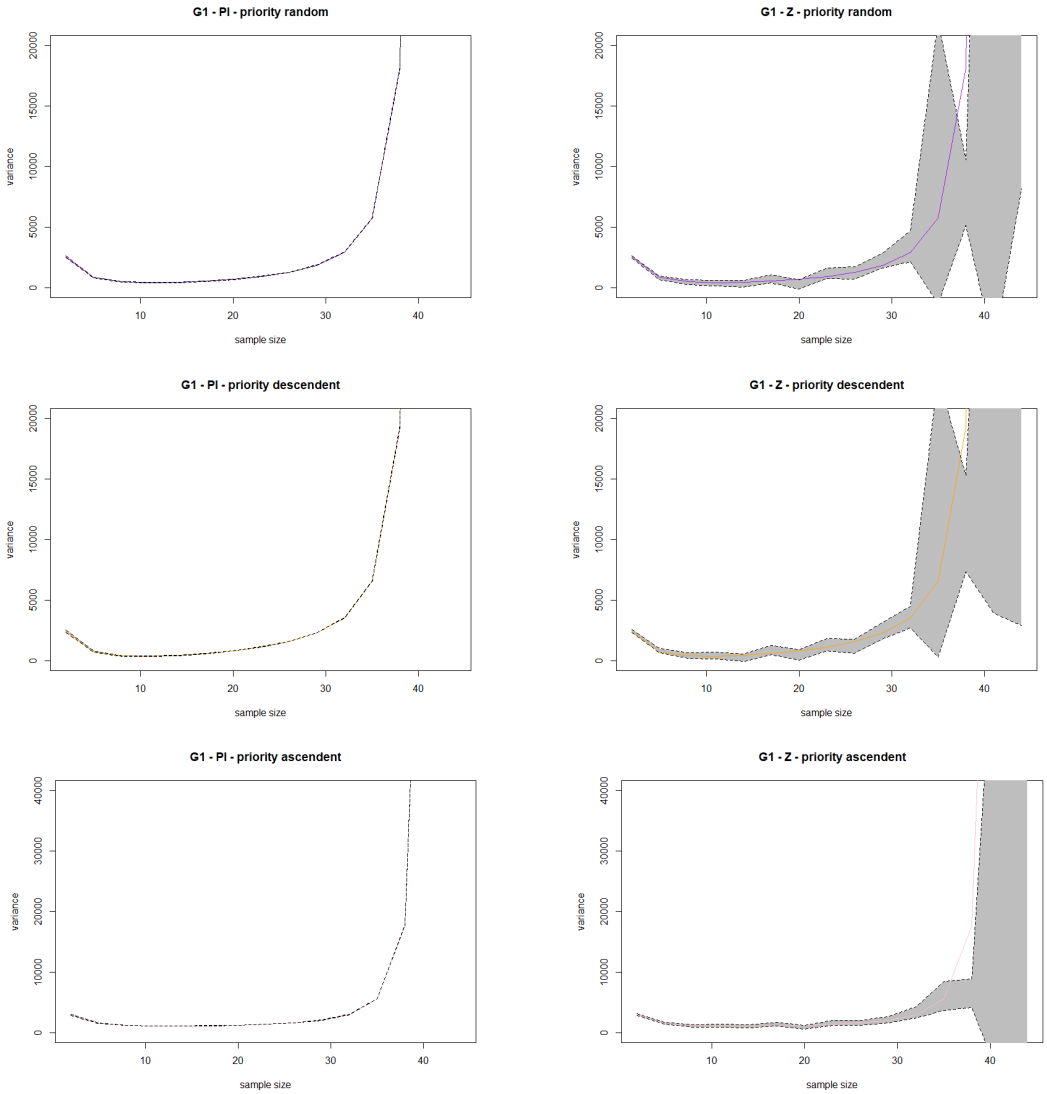


Figure 4.6: The true variances and its two estimators with associated Monte Carlo error for the priority IWE plotted against the increasing sample sizes considering different ordering of the frame for grap G_1.

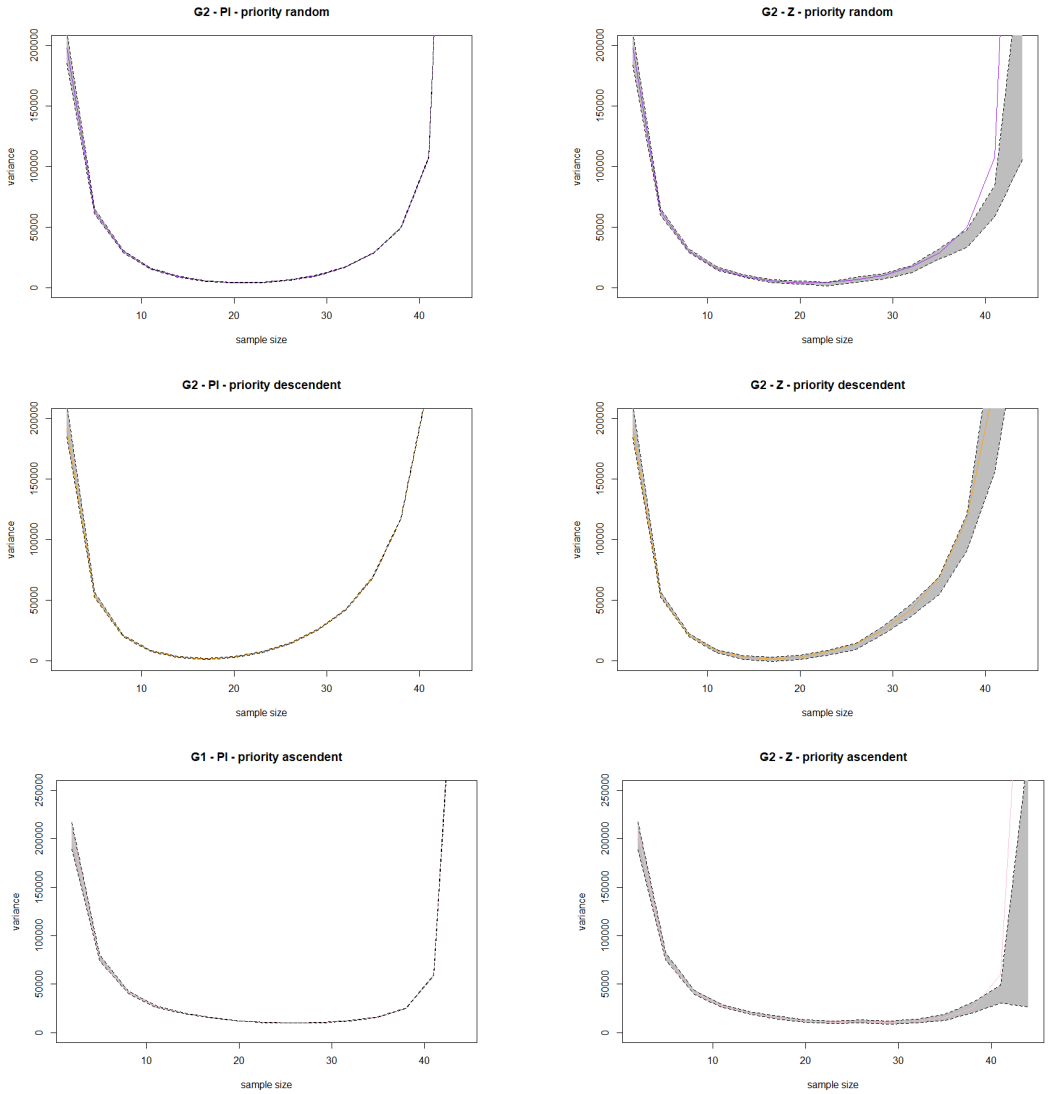


Figure 4.7: The true variances and its two estimators with associated Monte Carlo error for the priority IWE plotted against the increasing sample sizes considering different ordering of the frame for grap G_2.

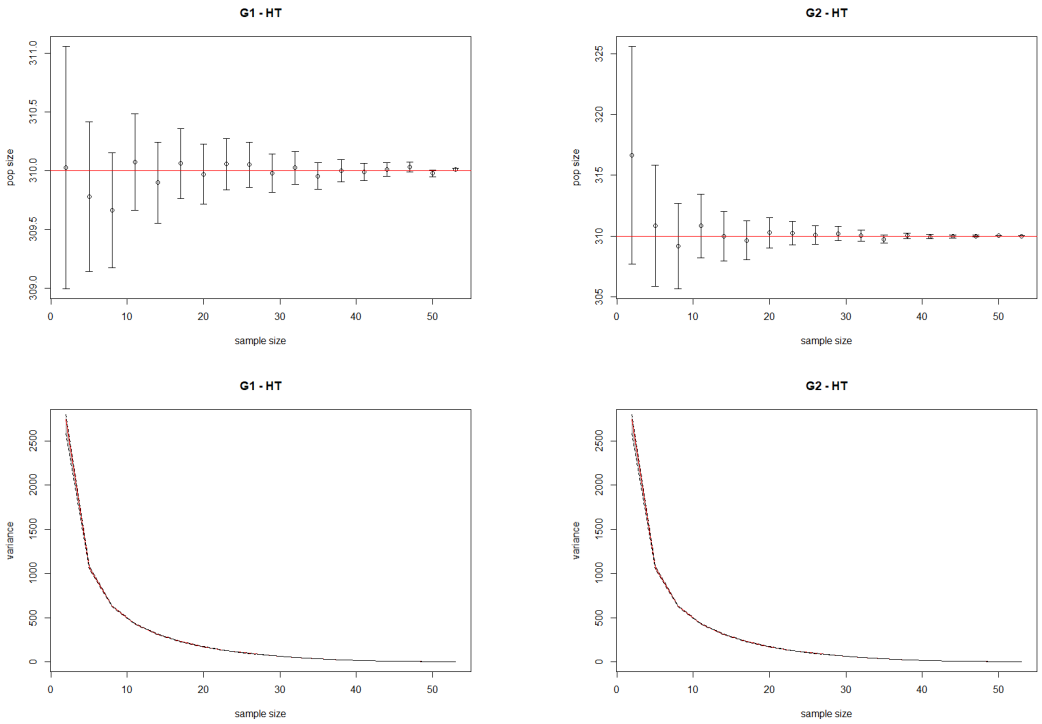


Figure 4.8: The average of the estimates with associated Monte Carlo error for the IWE corresponding to the Yhat, plotted against the increasing sample sizes for G_1 and G_2

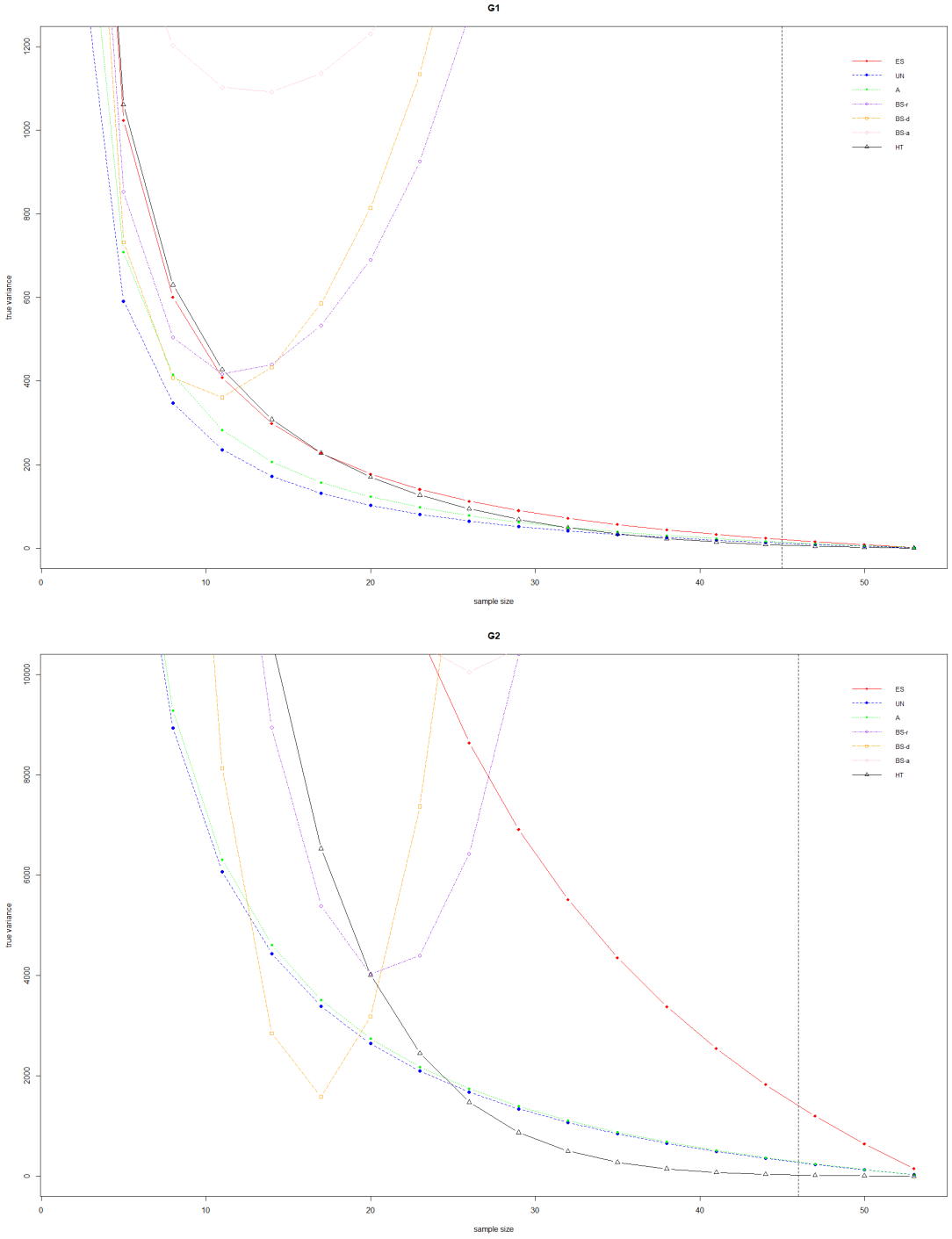


Figure 4.9: The variances of the six IWE plotted against the increasing sample sizes for both graph G_1 and G_2.

Chapter 5

Reverse incidence weighting under BIG sampling

In the BIG, a form of information, which does not exist in traditional list sampling, is given by the incidence structure of the graph. This has been used, in the previous chapter, to directly estimate a total Y of a variables measured on the motifs by means of incidence weighting. In this chapter, we see how this can be used for the estimation of a total measured on the sampling frame instead. In particular, an estimator of the size of the sampling frame, when it makes use of the incidence structure, can be considered important for improving the precision of the estimator of Y. In this chapter we show how such estimators can be formulated and how to use them to make better estimates for Y.

Key words: graph sampling, reverse incidence weighting, auxiliary information, ratio estimation, Hajek estimation.

5.1 Introduction

In the previous chapter, we have discussed how the structure of the BIG can be used for the estimation of totals of the finite population of motifs. For unconventional sampling methods, several authors (Birnbaum and Sirken, 1965; Thompson, 1990; Lavallée, 2007) have implicitly done the same, each author proposing

an estimator relevant for its problem, but without really exploring the potential of the structure of the graph in general terms. We have shown how the existing estimators can be formulated as a particular case of what we have called the *Incidence Weighting Estimator* (IWE), a general unbiased linear estimator that can be defined on the entities of the BIG. We have envisaged multiple choices of incidence weighting, showing how the performance depends on the relevant BIG structure.

If the IWE offers a way to incorporate the available graph structure in the estimation, one might suspect that this does not have to be the only way. We want to explore more deeply the possibilities that the structure of the BIG can offer during the estimation phase. This can be motivated by at least two reasons. Firstly, this is a completely unexplored territory, which does not have a corresponding version in list sampling. We are therefore driven by pure intellectual curiosity to investigate what are the differences between BIG sampling and conventional list sampling, and what is the potential of the use of the BIG compared to the traditional list. Secondly, we realize that the structure of the BIG is a form of auxiliary information, which we can use to define the incidence weights, in accordance with the observation procedure. Given the same sample of units, the sample graph can be different, depending on the observation procedure employed. In fact, some observation procedures will return a larger sample graph than other. For instance, in many situations, it will be possible to add further steps to the incident ancestral observational procedure and collect more edges so as to observe the elements they connect; in this situation, the incidence weights are computed by using this extra observed structure of the graph. It is reasonable to ask, how this information, to the extent that is available to us, can be use to make a more efficient estimator than the IWE. Obviously, when more of the graph structure is observed, more is the freedom to construct incidence weights; therefore, a natural way to use the extra information is to improve the efficiency of the IWE by means of incidence weighting.

Also, as done for the IWE, we should be able to use the graph structure to estimate the total of a variable attached to the sampling units. In this chapter, we focus on this topic. What we aim for is to make use of the incidence structure of the BIG in the estimation. We considered two ways of doing this: by means of sample-dependent incidence weighting; or by *reverse incidence weighting*, which are essentially the incidence weights, but constructed in the reverse sense. If, in

the IWE, the edges incident to each sampled unit are used for the estimation of a characteristics of the population of motifs by means of incidence weights, we can revert this argument and use the edges incident to the sampled motifs to estimate characteristics of the sampling frame, while preserving unbiasedness. In both scenarios, it will appear clear that an extra-step in the observation procedure is required. The extra step necessary for estimation can be obtained by *two-step incident ancestral* observation procedure. By incident ancestral observation procedure, we are able to observe which are the ancestors of the sample motifs; by adding an ulterior step, we are now able to also observe the successors of these ancestors. The questions we raise in this chapter are, *can the extra observed graph structure be used for improving the estimation of Y?* and if yes, *how?*

The rest of the chapter is organised as follows. In Section 2, we define the Reverse Incidence Estimator (RIWE), as a linear unbiased estimator of a characteristic of the sampling frame, which employs the additional structure of the graph provided by a two-stage incident observation procedure. A condition for the unbiasedness of the estimator is given. The variance and an estimator of it are provided. Also some numerical illustration are offered. In Section 3, we consider ratio-type estimators formulated by making use of the RIWE. The accuracy of the proposed ratio-type estimators are tested with simulated datasets. Finally, some concluding remarks are provided in Section 5.4.

A final note. The notation and setting up used in this chapter, when not otherwise specified, are the same as the ones introduced in the previous chapter.

5.2 The reverse incidence weighting estimator

Let x_k be a known value attached to a sampling unit k. Assume that the target of estimation is the total $X = \sum_{k \in F} x_k$. An estimator for X, can be given by the IWE

$$\hat{X} = \sum_{(ki) \in A_s} \frac{W_{ki} x_k}{\pi_{ki}} = \sum_{k \in s} \frac{x_k}{\pi_k} \sum_{i \in \alpha_k} W_{ki} \ . \tag{5.1}$$

where, to guarantee unbiasedness, the weights W_{ki} need to satisfy that for all $k \in F$,

$$\sum_{i \in \alpha_k} E(W_{ki} | \delta_{(i)}) = 1 \ . \tag{5.2}$$

In particular, when only fixed weights are considered, i.e. $W_{ki} = w_{ki}$, given the unbiasedness constraint, we have that:

$$\sum_{i \in \alpha_k} w_{ki} = 1 \ ,$$

and the IWE in Equation (5.1) is equal to the HT estimator $\hat{X}_{HT}$.

What appears clear is that when the IWE, based on the sample edges A_s, is used to estimate X, the incidence structure of the BIG is not taken into account. Following this observation, we can imagine at least two ways of making use of the incidence structure. One way involves sample-dependent weights. For example, the same priority rule used in the previous chapter can be formulated in this context. Once the sample BIG is observed, for each sample unit, only the edge which is incident to its 'smaller' sampled neighbour is considered in the estimation. An example of this priority estimator is presented in the next Illustration. To be able to derive this estimator, we need to know all the successors of the ancestors of the sampled motifs. A second way is obtained by enlarging the observed sample, so that the estimator is based on the set of edges incident to each sampled motif, $\tilde{A}_s = \beta(\alpha(s)) \times \alpha(s)$, where $\Pr((ki) \in \tilde{A}_s) = \pi_{(i)}$, which would result in a different estimator, as defined below.

Definition 5.2.1. Let $\tilde{s} = \beta(\alpha(s))$ and $\tilde{A}_s = \beta(\alpha(s)) \times \alpha(s)$. The *reserve incidence weighting estimator (RIWE)* for X is given by:

$$\tilde{X} = \sum_{(ki) \in \tilde{A}_s} \frac{W_{ki} x_k}{\pi_{(i)}} = \begin{cases} \sum_{k \in \tilde{s}} \gamma_k, & \text{where } \gamma_k = \sum_{i \in \alpha_k \cap \alpha(s)} \frac{W_{ki} x_k}{\pi_{(i)}} \\ \sum_{i \in \alpha(s)} \frac{z_{(i)}}{\pi_{(i)}}, & \text{where } z_{(i)} = \sum_{k \in \beta_i} W_{ki} x_k \ . \end{cases} \tag{5.3}$$

Essentially, this is the reverse problem of what we have seen in the previous chapter. The IWE and RIWE look apparently symmetrical: one is obtained by the other exchanging π_k with $\pi_{(i)}$ and $y_{(i)}$ with x_k. However, the two estimators are defined on two different sets, A_s and $\tilde{A}_s$ respectively, where $A_s = s \times \alpha(s)$.

We call the weights w_{ki} used in the RIWE, the *reverse incidence weights*.

The estimator is unbiased if Equation (5.2) holds. In fact,

$$\begin{aligned} E(\tilde{X}) &= E\left(\sum_{(ki)\in A} \frac{W_{ki}\delta_{(i)}x_k}{\pi_{(i)}}\right) \\ &= \sum_{(ki)\in A} \frac{E(W_{ki}|\delta_{(i)}=1)E(\delta_{(i)})x_k}{\pi_{(i)}} \\ &= \sum_{k\in F}\sum_{i\in\alpha_k} E(W_{ki}|\delta_{(i)}=1)x_k = \sum_{k\in F} x_k = X \ . \end{aligned}$$

Note that the conditions for unbiasedness for the IWE of Y and the RIWE of X are exactly symmetrical.

The variance of an unbiased RIWE can be given by

$$V(\tilde{X}) = \sum_{(ki)\in A}\sum_{(lj)\in A} \frac{\pi_{(i)(j)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} E(W_{ki}W_{lj}|\delta_{(i)}\delta_{(j)})x_k x_l \ .$$

In fact, we have that

$$\begin{aligned} V(\tilde{X}) &= \sum_{i\in U}\sum_{j\in U} \frac{E(\delta_{(i)}\delta_{(j)})}{\pi_{(i)}\pi_{(j)}} E(Z_{(i)}Z_{(j)}|\delta_{(i)}\delta_{(j)}=1)x_k x_l - X^2 \\ &= \sum_{i\in U}\sum_{j\in U} \frac{\pi_{(i)(j)}}{\pi_{(i)}\pi_{(j)}} \Big(\sum_{k\in\beta_i}\sum_{l\in\beta_j} E(W_{ki}W_{lj}|\delta_{(i)}\delta_{(j)}=1)x_k x_l\Big) - X^2 \\ &= \sum_{(ki)\in A}\sum_{(lj)\in A} \frac{\pi_{(i)(j)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} E(W_{ki}W_{lj}|\delta_{(i)}\delta_{(j)}=1)x_k x_l \ . \end{aligned}$$

An unbiased estimator of $V(\tilde{X})$ is given by

$$\hat{V}(\tilde{X}) = \sum_{(ki)\in A_s}\sum_{(lj)\in A_s} \frac{\pi_{(i)(j)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} \frac{E(W_{ki}W_{lj}|\delta_{(i)}\delta_{(j)})x_k x_l}{\pi_{(i)(j)}} \ .$$

Clearly, to be able to observe $\tilde{A}_s$, an extra step in the observation procedure is required; in the first step $\alpha(s)$ is observed, whereas the second step allows the observation of $\beta(\alpha(s))$. Therefore, we have available more knowledge about the structure of the graph. The question is whether the effort made to observe the extra structure can be useful to improve the estimation of Y, which is ultimately

the target of inference.

5.2.1 Examples of reverse incidence estimators

Similarly to the IWE, there are many possible choices of reverse incidence weights. We are going to consider the reverse versions of the Zhat, Yhat and Phat as described in the previous chapter, which are based on $\tilde{A}_s$. It will be obvious how they are symmetrical to their corresponding IWE, but with the necessary extra step for the observation procedure.

Ztilde In analogy with the Zhat, but looking reversely, the Ztilde makes use of the reverse incidence weights given by

$$w_{ki} = \frac{1}{d_k} \ .$$

Consequently, the $Z_{(i)}$ and γ_k are given by:

$$Z_{(i)} = \sum_{k \in \beta_i} \frac{x_k}{d_k} \qquad \text{and} \qquad \gamma_k = \frac{x_k}{d_k} \sum_{i \in \alpha_k \cap \alpha(s)} \frac{1}{\pi_{(i)}} \ .$$

For unbiased estimation of the Zhat the knowledge of d_i is essential, that can be provided under an incident or incident ancestral observation procedure. As expected, none of them allows the estimation of the Ztilde, for which knowledge of d_k is necessary for all $k \in \beta(\alpha(s))$. Two-stage incident ancestral observation procedure is needed, where the relevant measures to be observed are x_k and d_k, for all the $k \in \beta(\alpha(s))$.

Xtilde In order to provide analogy with the Yhat, we define the Xtilde as the RIWE corresponding to the IWE

$$\text{Xtilde} = \sum_{k \in \tilde{s}} \frac{x_k}{\tilde{\pi}_k} \ ,$$

where $\tilde{\pi}_k = P(k \in \tilde{s})$. Note that $\tilde{\pi}_k = 1 - Pr(\alpha_k \notin \alpha(s)) = 1 - Pr(\beta(\alpha_k) \notin s)$.

However, the strict analogy would be given by the HT-estimator of X, which is instead defined on the initial sample s and with the initial inclusion probabilities

π_k:

$$\hat{X}_{HT} = \sum_{k \in s} \frac{x_k}{\tilde{\pi}_k} \ .$$

The incidence weights corresponding to this estimator are given by

$$W_{ki} = \frac{\pi_{(i)}}{d_{k,\alpha(s)} \tilde{\pi}_k} \ ,$$

where $d_{k,\alpha(s)} = \sum_{i \in \alpha_k \cap \alpha(s)} a_{ki}$ is degree of k observed in the sample BIG and $\frac{\pi_{(i)}}{\tilde{\pi}_k} = \Pr(i \in \alpha(s) | k \in \tilde{s})$. The $Z_{(i)}$ and γ_k follows as:

$$Z_{(i)} = \sum_{k \in \beta_i} \frac{x_k \pi_{(i)}}{d_{k,s} \tilde{\pi}_k} \qquad \text{and} \qquad \gamma_k = \frac{x_k}{\tilde{\pi}_k} \ .$$

The observation procedure needed to estimate Yhat is incident ancestral, since to compute the $\pi_{(i)}$ we need to known its ancestors. As regarding the Xtilde, to know the $\tilde{\pi}_k$, we need instead to known the $\pi_{(i)}$ for all its successors $\alpha(\tilde{s})$. Two-step incident ancestral observation procedure is also required to determine the sample $\tilde{s}$.

Ptilde The same way that was used to define the Phat, can be reversed to define a RIWE using a priority rule. Let $\tilde{A}_{sp}$ be the set consisting of the prioritised edges, which are defined by the indicator variable I_{ki}, such that $I_{ki} = 1$, if for $k \in \tilde{s} \subset F$,

$$i = \min \left(\alpha(s) \cap \alpha_k \right) \ .$$

We have that:

$$Z_{(i)} = \sum_{k \in \tilde{s}} \frac{I_{ki} x_k}{d_k \tilde{p}_{(ki)}} \ ,$$

where $\tilde{p}_{(ki)} = \Pr(I_{ki} = 1 | (ki) \in \tilde{A}_s)$. We are basically assuming that the sample $\tilde{s}$ is observed, but only some of the units of $\tilde{s}$ contribute to the computation of $Z_{(i)}$, namely those that are connected with the smallest amongst their successors.

The corresponding γ_i is given by

$$\gamma_i = \frac{1}{d_k} \sum_{k \in \beta_i \cap \tilde{s}} \frac{I_{ki} x_k}{\tilde{p}_{ki} \pi_{(i)}} \ .$$

Regarding the observation procedure needed, the situation is analogous to the

Xtilde, but for computing the prioritisation probabilities in this case.

5.2.2 Simulations

Here, we illustrate the aforementioned RIWEs, providing an appreciation of their properties.

For these simulations, three BIG graphs are considered with different number of sampling units and motifs. The first BIG, G_1 has $|F_1| = 50$, $|U_1| = 20$ and $|A_1| = 100$; the second BIG G_2 has instead $|F_2| = 20$, $|U_2| = 50$ and $|A_2| = 100$, and finally the last BIG, G_3 has $|F_3| = |U_3| = 25$ and $|A_3| = 100$. The degree distributions of the both the set of units and motifs for the three BIG are shown in Figure 5.1

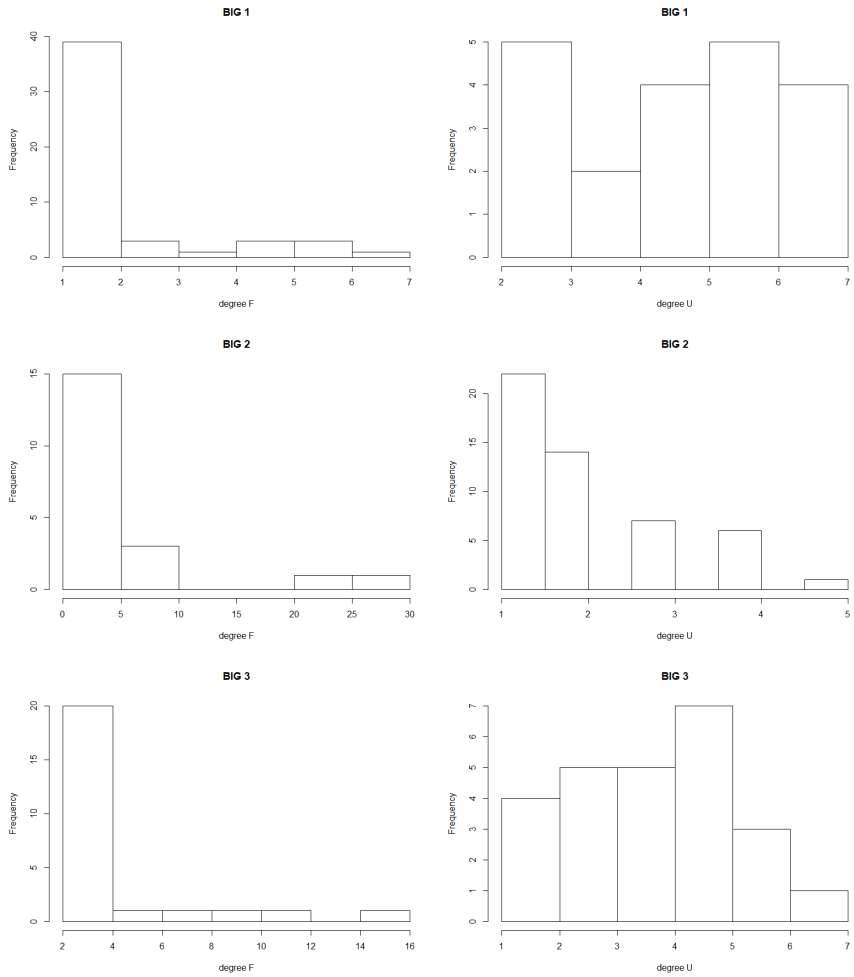


Figure 5.1: The degree distributions for the sampling units F and the motifs set U for the three BIG.

We assume simple random sampling on the sampling frame of size $m = 2$ and two-step incidence observation procedure. The objective is to estimate the total number of sampling units M for each of the three graphs. What will appear clear is that by using the structure of the BIG, we are able to compute several estimators of the sampling frame size, none of them having constant value M, as we would expect when the population is represented as a list, under SRS.

Five estimators are considered: Ztilde, Xtilde, Ptilde and the Phat of M as given in Equation (5.1). In particular, for the Ztilde, two choices of reverse incidence weighting are proposed:

$$w_{ki} = \frac{1}{d_k} \qquad \text{and} \qquad w_{ki} = \frac{1}{d_i} \Big/ \sum_{j \in \alpha_k} \frac{1}{d_j} \ .$$

The priority rule used for the Phat of M is defined by the indicator I_{ki} where,

$$I_{ki} = \begin{cases} 1, & \text{if } i = \min(\alpha(s) \cap \alpha_k) \ , \\ 0, & \text{otherwise.} \end{cases}$$

so the Phat of M is written as

$$Phat = \sum_{(ki) \in A_s} \frac{I_{ki}}{d_i p_{ki} \pi_k} \ ,$$

where $p_{ki} = \Pr(I_{ki} = 1 | i \in \alpha(s))$. The sampling distribution of each estimators is shown in 5.2, where the red points and the solid lines represent respectively the expectations and the medians of each distribution.

Table 5.1 shows the true variances of the estimators.

Table 5.1: The true variances of the estimators for M showed in Figure 5.2.

	Ztilde (ES)	Ztilde (ID)	Xtilde	Ptilde	Phat
BIG 1	623.43	653.31	468.66	459.52	55.09
BIG 2	130.34	105.64	63.46	103.24	7.72
BIG 3	149.01	158.81	37.77	54.21	9.73

Two main observations can be deduced from Figure 5.2. The Ptilde estimator is biased when the sampling units are less than the motifs (G_2) and slightly biased when both sets have the same cardinality (G_3). Similarly to the Phat, the Ptilde

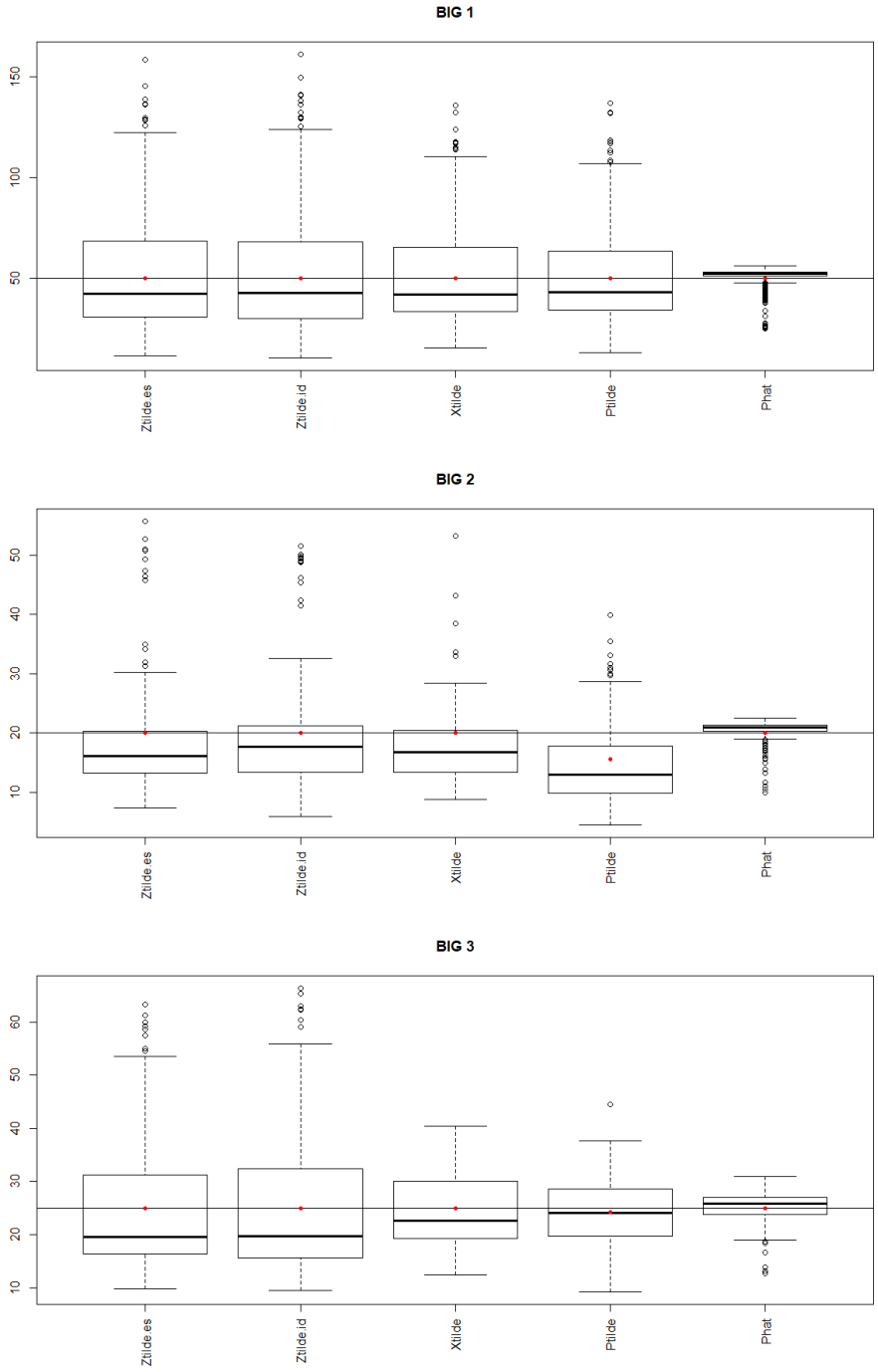


Figure 5.2: The sampling distribution of the estimators of M, under SRS of size $m = 2$ for the three BIGs.

is biased, if exists at least a $k \in F$ such that:

$$P\left(\sum_{i \in \alpha_k} \delta_{(i)} \leq 1\right) = 0 \ ,$$

where $\delta_{(i)} = 1$ if $i \in \alpha(s)$; 0 otherwise. Also, the Phat of M seems to be the more efficient amongst the estimators. Although, also this estimator suffers of bias when the above equation is the case.

5.2.3 A discussion

In the previous section, we have considered several possible estimators for the total X of any auxiliary variable measured on the sampling frame. Similarly to the IWE, our concern is to make use of the incidence structure of the BIG in the estimation. We have seen however, how this is not possible, when X is estimated using the IWE with fixed weights; in fact, in that particular case, the IWE is only based on the sampling units, and the incidence structure does not need to be employed.

Two other possibilities are explored: the use of sample dependent weights and the RIWE. The key characteristic of both estimators is that they require two-step incident ancestral observation procedures, in the first case to compute the probabilities of prioritization of each edge and in the second case to observe the set of edges which are used to define the estimator. In this way we have used the incidence structure of the BIG to estimate X. The next step is to understand how can we use these estimators to improve the estimation of Y.

A natural way, as commonly done in list sampling, is to use the estimate of X for the observed sample and the known true value of X to produce ratio-type estimators, as we will discuss in the next section. Alternatively, because the structure of the graph does not exists in list sampling, we must entertain the idea that more ways, which do not have a counterpart in list sampling, might exists on how to use it for improving the estimation of the parameter of interest. Also, it would be natural to ask ourself if there is any gain in constructing incidence weights which are unbiased for both the RIWE and the IWE, although, as it will be show, these weights present great limitations.

For unbiased estimation, the constraint for both RIWE and the IWE under fixed incidence weights, are given by

$$Y = \sum_{(ki)\in A} w_{ki} y_{(i)} = \sum_{k\in F}\sum_{i\in\alpha_k} w_{ki} y_{(i)} = \sum_{k\in F} z_k \qquad \text{with} \qquad \sum_{k\in\beta_i} w_{ki} = 1 \ ;$$

$$X = \sum_{(ki)\in A} w_{ki} x_k = \sum_{i\in U}\sum_{k\in\beta_i} w_{ki} x_k = \sum_{i\in U} z_{(i)} \qquad \text{with} \qquad \sum_{i\in\alpha_k} w_{ki} = 1 \ .$$

We use the small letter z_k to highlight that the constructed measures are fixed. Therefore, fixed weights w_{ki} can exist which simultaneously satisfy

$$\sum_{i \in \alpha_k} w_{ki} = 1 \qquad \text{and} \qquad \sum_{k \in \beta_i} w_{ki} = 1 \ . \tag{5.4}$$

An advantageous characteristics of these weights is that, under simple random sampling of the sampling units, they often improve the IWE. Intuitively, the idea is explained by the following. Under SRS, improving the efficiency of the IWE equals reducing the finite population variance of the z_k. Notice that, for each k, the measure z_k is a linear combination of the $y_{(i)}$, with $i \in \alpha(s)$, where the coefficients of the combination are the weights w_{ki}. Insofar, we had made no assumptions on the set of values that the w_{ki} should take, just that $w_{ki} > 0$. Instead, when using the weights discussed here, we are restricting the choices of weights to those respecting that $\sum_{k \in \beta_i} w_{ki} = 1$. As an immediate consequence, the range of values that z_k can take goes between the minimum and the maximum of the $y_{(i)}$, with $i \in \alpha(s)$. This suggests that compared to other weights, the weights here presented can reduce the finite population variance of z_k and consequently the variance of the IWE.

Special case: $y_{(i)} \equiv 1$**.** When $y_{(i)} \equiv 1$, using the weights satisfying both Equations (5.5) returns $z_k \equiv 1$. In fact, we have:

$$z_k = \sum_{i \in \alpha_k} w_{ki} y_{(i)} = \sum_{i \in \alpha_k} w_{ki} = 1 \qquad \text{for all} \qquad k \in U \ .$$

A necessarily condition for these weights to exist is that the size F is equal to the size U, i.e. $N = M$. In fact,

$$|U| = \sum_{i \in U} \sum_{k \in \beta_i} w_{ki} = \sum_{k \in F} \sum_{i \in \alpha_k} w_{ki} = |F| \ .$$

Notice that if $|F| = |U|$, the these weights w_{ki} are obtained by solving the system of linear equations given by:

$$\Lambda W = 1_L \ , \tag{5.5}$$

where L is the number of edges, W is the vector of weights and Λ is the incidence

$(2N) \times L$ matrix with $2N$ rows, representing the nodes of the graph; L columns, representing the edges in the graph and whose entries are 1 if the node is incident to the edge and 0 otherwise. The incidence matrix Λ is not always singular, in which case there are infinite solutions for the system in Equation (5.5). In any case, it can always be found an pseudoinverse inverse Λ^{+}, such that the weights given by $\Lambda^{+}1_L$, return approximately unbiased IWE.

Illustration Consider the BIG G_3, which has $M = N = 25$. We compute three IWE of Y, where Y is the total of variable $y_{(i)}$, where:

$$y_i = 3d_i + e_i \qquad \text{where} \qquad e_i = \mathcal{N}(0, 2) \ .$$

Three choices of weights are used: equal-share weights, inverse-degree weights and 'twice-unbiased' weights as discussed above. A sample s of size $m = 2$ is taken with simple random sampling from F. The sampling distribution of the three estimators is given in Figure 5.3. The variances are respectively: 57230.61, 14725.55 and 3199.16.

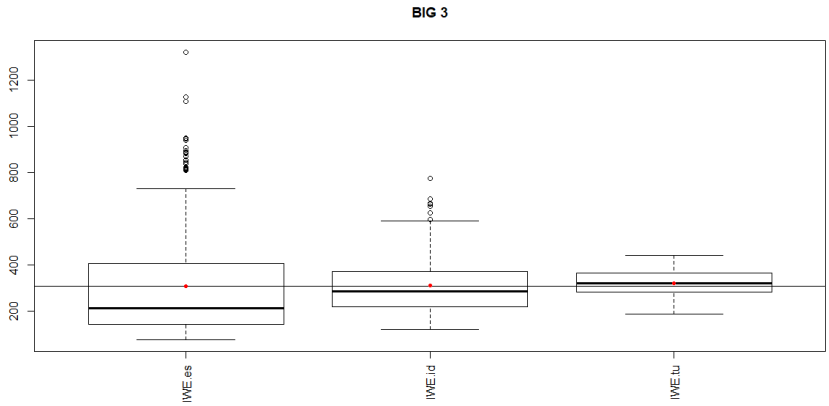


Figure 5.3: The sampling distribution of the three IWE $Y = 308.54$, under SRS of size $m = 2$ for the BIG G_3.

5.3 Ratio-type estimators on a BIG

Let $\tilde{M}$ be the RIWE of M, the size of the sampling frame, which is known and let $\hat{Y}$ be the IWE of Y, the total of the variable of interest measured on each

motif in U. The two estimators are given by

$$\hat{Y} = \sum_{(ki)\in A_s} \frac{W_{ki} y_{(i)}}{\pi_k} \qquad \text{and} \qquad \tilde{M} = \sum_{(ki)\in \tilde{A}_s} \frac{\tilde{W}_{ki}}{\pi_{(i)}} \ ,$$

where W_{ki} and $\tilde{W}_{ki}$ are respectively the incidence and reverse incidence weights and π_k and $\pi_{(i)}$ are the inclusion probability of respectively a sampling unit and a motif. We assume a simple random sample is taken from F of size m.

Two estimators of the ratio $R = Y/M$ are given by

$$\hat{R} = \frac{\hat{Y}}{\hat{M}} = \frac{\sum_{(ki)\in A_s} \frac{W_{ki} y_{(i)}}{\pi_k}}{\sum_{(ki)\in A_s} \frac{\tilde{W}_{ki}}{\pi_k}} \qquad \text{and} \qquad \tilde{R} = \frac{\hat{Y}}{\tilde{M}} = \frac{\sum_{(ki)\in A_s} \frac{W_{ki} y_{(i)}}{\pi_k}}{\sum_{(ki)\in \tilde{A}_s} \frac{\tilde{W}_{ki}}{\pi_{(i)}}} \ ,$$

where $\hat{M}$ is the IWE estimator of M.

It follows that two ratio-type estimators of the total Y on a BIG can be given as

$$\hat{Y}_R = \frac{\hat{Y}}{\hat{M}} M \qquad \text{and} \qquad \tilde{Y}_R = \frac{\hat{Y}}{\tilde{M}} M \ . \tag{5.6}$$

Some observations about the two ratio-type estimators can be made. Firstly, the two estimators in both ratio estimators $\hat{R}$ and $\tilde{R}$ are not necessarily defined over the same sample graph. Consider $\tilde{R}$. While $\hat{Y}$ is defined over A_s, $\tilde{M}$ is defined over $\tilde{A}_s$. Also, even if apparently it seems that the $\hat{Y}$ and $\hat{M}$ in $\hat{R}$ are both defined over A_s, if the priority weights are used, $\hat{Y}$ will be defined over the prioritised set A_{sp}, which is different from A_s, used to defined $\hat{M}$. Secondly, even when the auxiliary variable x_k is equal to 1 for each $k \in F$, the IWE for X with sample-dependent weights and the RIWE for X are constructed by using the observed structure of the graph, i.e. by constructing a sample dependent measure $Z_k \neq 1$ in the IWE and a measure $Z_{(i)} \neq 1$ in the RIWE. Then, even when a simple random sample is taken from F, the two estimators $\hat{M}$ and $\tilde{M}$ do not result equal to M in each sample. This suggests that even in this case of minimal information, when the additional variable $x_k = 1$ for all $k \in F$, the ratio-type estimators still exists in a non-trivial way. Finally, note that the two ratio-type estimators in Equation (5.6) are defined by assigning a value to each edge via the incidence weights. It is also certainly possible to assign just the value 1 to each edge, in this case we

obtain two Hajek-type estimators. Let L be the known total number of edges in the graph and let $\hat{L}$ and $\tilde{L}$ be the two estimators of it defined as

$$\hat{L} = \sum_{(ki)\in A_s} \frac{1}{\pi_{(ki)}} = \sum_{(ki)\in A_s} \frac{1}{\pi_k} \qquad \text{and} \qquad \tilde{L} = \sum_{(ki)\in \tilde{A}_s} \frac{1}{\tilde{\pi}_{(ki)}} = \sum_{(ki)\in \tilde{A}_s} \frac{1}{\pi_{(i)}} \ .$$

The two Hajek-type estimators are then given by

$$\hat{Y}_W = \frac{\hat{Y}}{\hat{L}} L \qquad \text{and} \qquad \tilde{Y}_W = \frac{\hat{Y}}{\tilde{L}} L \ . \tag{5.7}$$

It is remarkable that, under SRS and using only the knowledge of M (or L), it is possible to obtain at least four classes of ratio-type or Hajek-type estimators, which can be modified under different choices of incidence weights. However, it is not clear how these estimators can be motivated to improve the estimation of Y. Take for instance $\hat{Y}_R$ (note that the following discussion holds also for the other three estimators $\tilde{Y}_R$, $\hat{Y}_W$, $\tilde{Y}_W$). By its Taylor approximation around the point $(E(\hat{Y}), E(\hat{M}))$, we have:

$$\hat{Y}_R \approx Y + \left(\hat{Y} - \frac{Y}{M} \hat{M} \right) \ . \tag{5.8}$$

In fact:

$$\begin{aligned} \frac{\hat{Y}}{\hat{M}} &\approx \frac{E(\hat{Y})}{E(\hat{M})} + \frac{1}{E(\hat{M})}\big(\hat{Y} - E(\hat{Y})\big) - \frac{E(\hat{Y})}{E(\hat{M})^2}\big(\hat{M} - E(\hat{M})\big) \\ &= \frac{Y}{M} + \frac{1}{M}\big(\hat{Y} - Y\big) - \frac{Y}{M^2}\big(\hat{M} - M\big) \\ &= \frac{Y}{M} + \frac{1}{M}\big(\hat{Y} - \frac{Y}{M}\hat{M}\big) \ . \end{aligned}$$

Multiplying the last equation by M, we obtain the linearization for $\hat{Y}_R$.

From Equation (5.8), it follows that:

$$E(\hat{Y}_R) \approx Y \qquad \text{and} \qquad V(\hat{Y}_R) \approx V(\hat{Y}) + \left(\frac{Y}{M}\right)^2 V(\hat{M}) - 2\frac{Y}{M} Cov(\hat{Y}, \hat{M}) \ .$$

The ratio-type estimator $\hat{Y}_R$ on a BIG is approximately unbiased and it will be

more efficient compared to the IWE $\hat{Y}$ if

$$\left(\frac{Y}{M}\right)^2 V(\hat{M}) - 2\frac{Y}{M}Cov(\hat{Y},\hat{M}) < 0 \ . \tag{5.9}$$

This means, that there in an improvement of efficiency if the covariance term is enough big.

Consider the following example. Let G be a BIG with $F = \{1,2,3\}$, $U = \{4,5,6\}$ and $A = \{(1,4),(1,5),(2,6),(3,4)\}$. Under equal-share incidence and reverse incidence weights, we have that $z_k = (1.5, 1, 0.5)$ and $z_{(i)} = (1.5, 0.5, 1)$. Assume simple random sampling of size 2, then $\pi_k = 2/3$ for all $k \in F$ and $\pi_{(i)} = (1, 2/3, 2/3)$, for $i = 4,5,6$. Table 5.2 shows the sampling distribution of the three estimators $\hat{Y}$, $\tilde{M}$ and $\tilde{L}$.

Table 5.2: The sampling distribution of the $\hat{Y}$, $\tilde{M}$ and $\tilde{L}$, using equal-share incidence and reverse incidence weights for the BIG G, with $F = \{1,2,3\}$, $U = \{4,5,6\}$ and $A = \{(1,4),(1,5),(2,6),(3,4)\}$, under simple random sampling from F of size 2.

s	A_s	$\tilde{A}_s$	$\hat{Y}$	$\tilde{M}$	$\tilde{L}$
$\{1,2\}$	$\{(1,4),(1,5),(2,6)\}$	$\{(1,4),(1,5),(2,6),(3,4)\}$	3.75	3.75	5
$\{1,3\}$	$\{(1,4),(1,5),(3,4)\}$	$\{(1,4),(1,5),(3,4)\}$	3	2.25	3.5
$\{2,3\}$	$\{(2,6),(3,4)\}$	$\{(1,4),(2,6),(3,4)\}$	2.25	3	3.5

Clearly, there is no positive covariance between the two estimators. First of all, there is not a clear way to define a correlation between the values z_k and $z_{(i)}$, for k and i which are connected; also, the values z_k and $z_{(i)}$ are weighted differently, respectively by π_k and $\pi_{(i)}$.

5.3.1 Simulation study

We considered the same BIG used in the illustration in Section 3. Here, the objective is to estimate the population total when $y_{(i)} \equiv 1$, i.e. $Y = N$. We assume SRS of size 7 from F. The computed estimators are given in Table 5.3

where the weights are defined as following: HT = HT weights; ES = equal-share weights and ID = for inverse-degree weights.

Monte Carlo simulations are used to study the accuracy of the 11 estimators of the total number of motifs. We have run 1000 simulations. The resulting estimators

Table 5.3: The estimators considered in the simulations in this section.

	Estimator	Formula	w	$\tilde{w}$
1	HT	$\hat{Y}_{HT} = \sum_{i \in \alpha(s)} \frac{y_{(i)}}{\pi_{(i)}}$	-	-
2	IWE	$\hat{Y} = \sum_{k \in s} \frac{z_k}{\pi_k}$	ES	-
3			ID	-
4	Hajek-type	$\tilde{Y}_W = L \frac{\hat{Y}_{HT}}{\tilde{L}}$	-	-
5	Ratio-type	$\hat{Y}_R = M \frac{\hat{Y}^w}{\tilde{M}^{\tilde{w}}}$	HT	ES
6			HT	ID
7			ES	ES
8			ID	ID
9			ES	ID
10			ID	ES

for N in BIG 1, BIG 2 and BIG 3 are given respectively in Figures 5.4, 5.5 and 5.6, numbered as in Table 5.3.

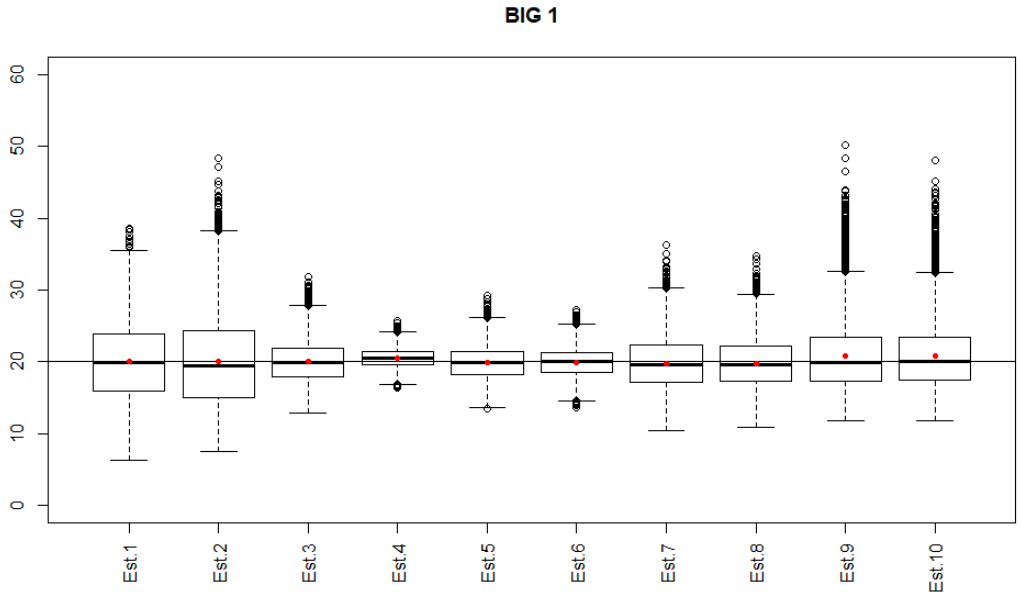


Figure 5.4: Comparison of the accuracy of the 10 estimators of N for BIG 1.

Most of the sampling units in BIG 2 have smaller degrees than the units in BIG 1 and BIG 3, so the sample of motifs in BIG 2 is on average smaller than the samples of motifs in BIG 1 and BIG 3, given the same sample $s \subset F$. It follows that the estimators in BIG 1 and BIG 3 are more efficient than those in BIG 2.

In particular, in all the three cases, we see that the IWE with inverse-degree

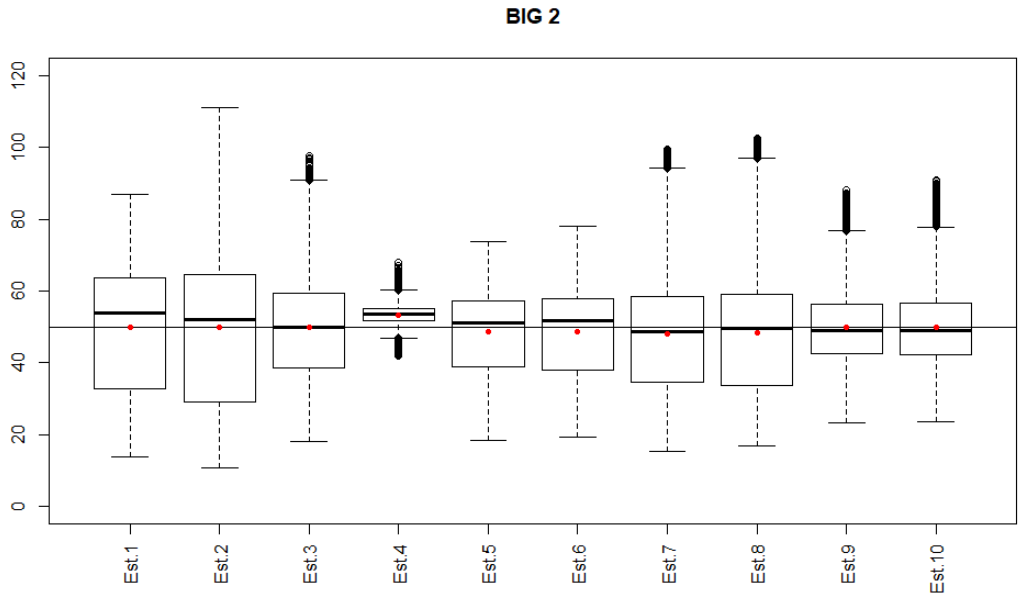


Figure 5.5: Comparison of the accuracy of the 10 estimators of N for BIG 2.

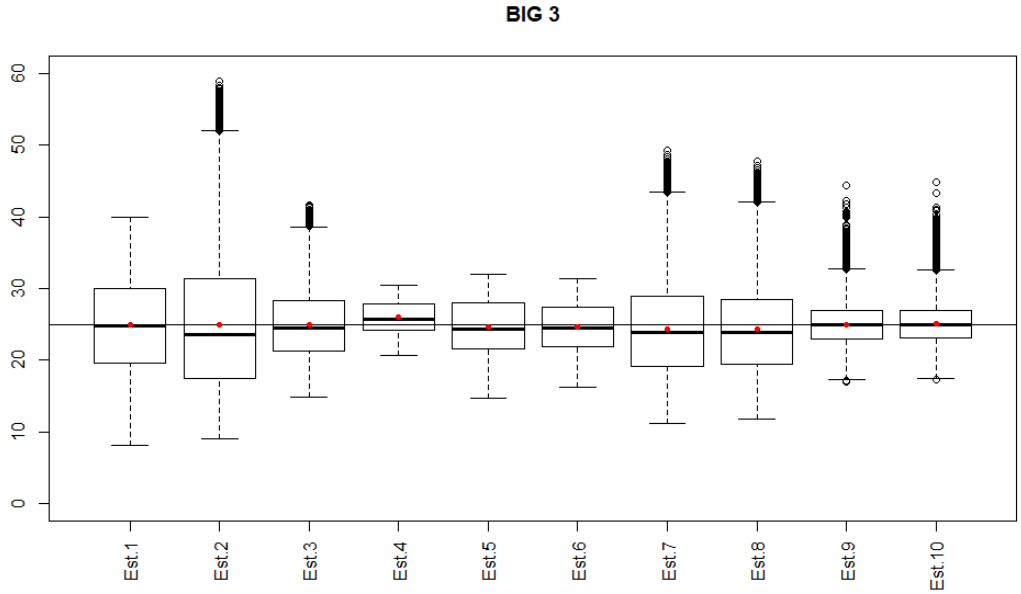


Figure 5.6: Comparison of the accuracy of the 10 estimators of N for BIG 3.

weights reduce the variance compared to the one which uses the equal-share weights. All the ratio-type and the Hajek-type estimators are approximately unbiased, and we noted that the bias is larger for the Hajek-type estimator in all the three graphs. Also, the two ratio-type estimators defined on the sampled

motifs reduce the variance compared to the HT and IWE. Also the estimator H_3 and H_4 tend to be more efficient, particularly in BIG 2 and BIG 3, when they are the most efficient estimators. More investigation is needed on these ratio-type estimators.

Finally, the Hajek-type estimator seems to be the most efficient estimators amongst the 10 estimators; however, as already noted, it suffers from bigger bias.

5.3.2 A particular class of ratio-type and Hajek-type estimators

To simplify the situation, we consider only the ratios between estimators defined over the same set and with same inclusion probability for each sampled element. We exclude from the discussion both $\hat{Y}_R$ and $\hat{Y}_W$, since under simple random sampling they will not provide an estimator different from $\hat{Y}$. We also exclude the estimators which make use of the priority rules, since it will unnecessarily complicate the discussion at the moment. We are left with the following two estimators:

$$\tilde{Y}_R = \frac{\hat{Y}_{HT}}{\tilde{M}} M \qquad \text{and} \qquad \tilde{Y}_W = \frac{\hat{Y}_{HT}}{\tilde{L}} L \ ,$$

where $\tilde{M}$ is defined by using fixed reverse incidence weights.

In this cases, Equation (5.8) becomes:

$$\begin{aligned} \tilde{Y}_R &\approx Y + \tilde{\epsilon}_{HT} \qquad \text{with} \qquad \epsilon_{(i)} = y_{(i)} - \frac{Y}{M} z_{(i)} \ ; \\ \tilde{Y}_W &\approx Y + \tilde{e}_{HT} \qquad \text{with} \qquad e_{(i)} = y_{(i)} - \frac{Y}{L} \ . \end{aligned}$$

And the variances are given by

$$\begin{aligned} V(\tilde{Y}_R) &\approx \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{(ij)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} \right) \epsilon_{(i)} \epsilon_{(j)} \ , \\ V(\tilde{Y}_W) &\approx \sum_{i \in U} \sum_{j \in U} \left(\frac{\pi_{(ij)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} \right) e_{(i)} e_{(j)} \ . \end{aligned}$$

The covariance $Cov(\delta_i, \delta_j)$ is weighted by $\frac{e_{(i)}e_{(j)}}{\pi_{(i)}\pi_{(j)}}$ or $\frac{\epsilon_{(i)}\epsilon_{(j)}}{\pi_{(i)}\pi_{(j)}}$ instead of $\frac{y_{(i)}y_{(j)}}{\pi_{(i)}\pi_{(j)}}$.

The estimators of the variances are given by

$$\hat{V}(\tilde{Y}_R) \approx \sum_{i\in\alpha(s)} \sum_{j\in\alpha(s)} \left(\frac{\pi_{(ij)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} \right) \frac{\hat{\epsilon}_{(i)}\hat{\epsilon}_{(j)}}{\pi_{(i)(j)}} ,$$

$$\hat{V}(\tilde{Y}_W) \approx \sum_{i\in\alpha(s)} \sum_{j\in\alpha(s)} \left(\frac{\pi_{(ij)} - \pi_{(i)}\pi_{(j)}}{\pi_{(i)}\pi_{(j)}} \right) \frac{\hat{e}_{(i)}\hat{e}_{(j)}}{\pi_{(i)(j)}} .$$

where $\hat{\epsilon}_{(i)} = y_{(i)} - \frac{\hat{Y}_{HT}}{\tilde{M}} z_{(i)}$ and $\hat{e}_{(i)} = y_{(i)} - \frac{\hat{Y}_{HT}}{\tilde{L}}$.

5.4 Conclusions

In this chapter we have shown how to construct the RIWE estimator of the total X of a variable measured on the sampling units in a BIG. However, because the final target of estimation is a function of the motifs, the RIWE is used to improve the efficiency of the IWE. Its properties are described and some of the peculiarities of the estimation on a BIG, which do not exists in the list case are highlighted. Particular emphasis on the construction of ratio-type estimators is given.

First of all, we have recognized that in a BIG, the structure of the graph is a type of auxiliary information itself, which does not exist in list sampling. We have utilised the graph structure by the incidence weights and by the reverse incidence weights. In particular, it can be used to improve the efficiency of the IWE. It is seen that the reverse incidence weights carry more information about the structure of the BIG, and, in fact, the observational procedure necessary to compute them returns a larger sample graph.

With the help of the RIWE, we have proposed several ratio-types estimators and shown some of their properties, analytically and by means of numerical illustrations. These ratio-type estimators improve the estimation of Y. Especially, in the simulation studies, we have seen that using a ratio between estimators which are not defined over the same sets, might nevertheless improve the estimation, and further exploration on this type of estimators is required.

More importantly, under the BIG framework, several ratio-estimators for a function of the motifs can be constructed, which are all approximately unbiased. The number of them is potentially infinite, since it depends on the possible choices of incidence and reverse incidence weights. Adding this to the fact the IWE offers

infinite unbiased linear estimator motivates our idea that the BIG framework is quite a promising area of research for sampling and estimation and future works is necessary to understand its potential.

Finally, we have two last observations. Firstly, we have seen that also the IWE can be used to estimate the total X. We have shown that by using fixed weights, the structure of the graph is not used, whereas priority weights make uses of it. We did not explore the use of sample-dependent weights any further in this chapter, but this can be a valuable topic of future research. Secondly, we have discussed the use of weights which return unbiased IWE and RIWE simultaneously and explore their advantages and limitations for the estimation of Y. We believe this should not be the only possible way, and we think more investigation is required. In particular, this exploration can be aimed to the construction of other types of estimators, that are also unbiased.

Conclusions

In this thesis we have discussed how the increasing use of social media data, as well as other type of big data has brought both exciting opportunities and considerable challenges for researcher in many disciplines and in particular in Statistics.

We have distinguished four basic obstacles to making valid statistics analysis using traditional methods. Firstly, there is a contrast between the population of interest and the users of a relevant platform; secondly, there is no control over how social media data can be selected from the relevant platform; thirdly, in general the objects of data collection are different from those of interest and finally, the measures of interest need to be extrapolated by algorithms and machine learning techniques. We have argued how the identification of such errors is crucial for understanding the quality of the data and systematically delineated two existing approaches to statistical analysis based on social media data.

Having examined the difficulties that social media data carry for conducting statistical analysis, we have focused on the possibilities that such data offers, in particular, their graph structure. A peculiar characteristic of social media data is that they offer the potential to observe several network relationships, which are seldom available via traditional surveys. We have provided a review of the existing methods of graph sampling and delineate a general framework of sampling and estimation for graph data that makes use of the BIG, a bipartite incidence graph.

To conclude, the research undertaken in this thesis offers a more rigorous approach of the use of these types of data, by delineating and discussing the validity of the existing methods for making appropriate quantitative estimates from the use of social media data. Moreover, it introduces the new topics of graph sampling and estimation, which is particularly relevant nowadays, given the increasing

availability of the graph representation of the data. Currently, sampling and analysis methods of network data are under-explored and we hope to have opened the doors for more research in the future.

References

Baker R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., and Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology,*, 1:90–143.

Beręsewicz, M., Lehtonen, R., Fernando, R., Di Consiglio, L. and Karleber, M. (2018). *An overview of methods for treating selectivity in big data sources*. Technical report, Eurostat.

Berzofsky, M., Mckay, T., Hsieh, Y., and Smith, A. (2018). Probability-based samples on Twitter: methodology and application. *Survey Practice*, 11:1–12. doi: 10.29115/SP-2018-0033.

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A., and Sudman, S. (2004). *Measurement Errors in Surveys*. John Wiley & Sons. doi: 10.1002/9781118150382.

Birnbaum, Z. W. and Sirken, M. G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. *Vital and Health Statistics*, Series 2, No.11. Washington:Government Printing Office.

Blank, G. (2013). Who creates content? Stratification and content creation on the internet. *Information, Communication & Society*, 16(4):590–612.

Blank, G. and Lutz, C. (2017). Representativeness of social media in Great Britain: investigating Facebook, Linkedin, Twitter, Pinterest, Google+, and Instagram. *American Behavioral Scientist*, 61:741–756. doi: 10.1177/0002764217717559.

Blumenstock, J., Cadamuro, G. and On, R. (2015). Predicting Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264), 1073–1076.

boyd, d. and Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5):662–679. doi: 10.1080/1369118X.2012.678878.

Brakel, J., Söhler, E., Daas, P., and Buelens, B. (2017). Social media as a data source for official statistics; the Dutch consumer confidence index. *Survey Methodology*, 43:183–210. doi: 10.13140/RG.2.2.19294.64326.

Bright, J., Margetts, H., Hale, S., and Yasseri, T. (2014). *The Use of Social Media for Research and Analysis: A Feasibility Study*. Technical report, Department for Work and Pensions, United Kingdom.

Brown, D. M. and Soto-Corominas, A. (2017). Overview of the social media data processing pipeline. *The SAGE Handbook of Social Media Research Methods,*, 125.

Buelens, B., Burger, J., and van den Brakel, J. A. (2018). Comparing Inference Methods for Non-probability Samples. *International Statistical Review*, 86(2), 322-343.

Chambers, R. L. and Skinner, C. J. (2003). *Analysis of Survey Data*. John Wiley & Sons. doi: 10.1002/0470867205.

Chung, F. R. K. (1997). Spectral Graph Theory. Providence, RI: American Mathematical Society.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, 40(2):137-161.

Cochran, W. G. (1977). *Sampling Techniques (3rd ed.)*. New York: Wiley.

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review Blog*.

Daas, P., Roos, M., Van de Ven, M., and Neroni, J. (2012). *Twitter as A Potential Data Source for Statistics*. Technical report, 201221, Statistics Netherlands, The Hague/Heerlen.

Daas, P. J. and Puts, M. J. (2014). *Social Media Sentiment and Consumer Confidence*. Series No. 5; European Central Bank Statistics Paper, Frankfurt, Germany.

Daas, P. J., Puts, M. J., Buelens, B., and van den Hurk, P. A. (2015). Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262. doi: 10.1515/jos-2015-0016.

Daas, P. J., Burger, J., Le, Q., ten Bosch, O., and Puts, M. J. (2016). *Profiling of Twitter Users: A Big Data Selectivity Study.* Technical report, 201606, Statistics Netherlands, The Hague/Heerlen.

De Waal, T., van Delden, A., and Scholtus, S. (2017). *Multi-source Statistics: Basic Situations and Methods.* Technical report, 201712; Statistics Netherlands, The Hague/Heerlen.

Di Zio, M., Zhang, L.-C., and de Waal, T. (2017). Statistical methods for combining multiple sources of administrative and survey data. *The Survey Statistician*, 76:17–26.

Elliott, M. R. (2009). Combining data from probability and non? probability samples using pseudo-weights. *Survey Practive.*

Elliott, M. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32:249–264. doi: 10.1214/16-STS598.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2):293–314.

Fattorini, L. (2006). Applying the Horvitz–Thompson criterion in complex designs: A computer-intensive perspective for estimating inclusion. *Biometrika*, 93:269–278.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, 56(4):82–89.

Frank, O. (1971). *Statistical inference in graphs.* Stockholm: Försvarets forskningsanstalt.

Frank, O. (1977a). Estimation of graph totals. *Scandinavian Journal of Statistics*, 4:81–89.

Frank, O. (1977b). A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scandinavian Journal of Statistics*, 4:178–180.

Frank, O. (1977c) Survey sampling in graphs. *Journal of Statistical Planning and Inference*, 1(3):235–264.

Frank, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scandinavian Journal of Statistics*, 5:177–188.

Frank, O. (1979). Sampling and estimation in large social networks. *Social networks*, 1(1):91–101.

Frank, O. (1980a). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference*, 4(1):45–50.

Frank, O. (1980b). Sampling and inference in a population graph. *International Statistical Review/Revue Internationale de Statistique*, 48:33–41.

Frank, O. (1981). A survey of statistical methods for graph analysis. *Sociological methodology*, 12:110–155.

Frank O. and Snijders T. (1994). Estimating the size of hidden populations using snowball sampling. *Journal of Official Statistics*, 10:53–53.

Frank, O. (2011). Survey sampling in networks. *The SAGE Handbook of Social Network Analysis*, pages 389–403.

Gaffney, D. and Puschmann, C. (2013). *Data collection on Twitter*, In *Twitter and Society*, pp. 55–67. Weller, K., Bruns, A., Burgess, J., Mahrt M., Puschmann C. (eds.), Peter Lang Publishing, Inc. doi: 10.3726/978-1-4539-1170-9.

Golbeck, J. and Hansen, D. (2014). A method for computing political preference among Twitter followers. *Social Networks*, 36:177–184.

Goldenberg, A., Zheng, A.X., Fienberg, S.E. and Airoldi, E.M. (2010). A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 2:129–233.

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., and Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks*, 38:16–27. doi: 10.1016/j.socnet.2014.01.004.

Goodman, L. A. (1961). Snowball sampling. *Annals of Mathematical Statistics*, 32:148–170.

Greenwood, S., Perrin, A., and Duggan, M. (2016). *Social Media Update, November 2016.* Technical report, Pew Research Centre.

Groves, M. R., Fowler Jr., J. F., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology.* John Wiley & Sons.

Groves, M. R. (2011a). 'Designed data' and 'organic data'. *Directors Blog.*

Groves, M. R. (2011b). Three eras of survey research. *Public Opinion Quarterly*, 75:861–871. doi: 10.1093/poq/nfr057.

Halford, S., Weal, M., Tinati, R., Pope, C., and Carr, L. (2017). Understanding the production and circulation of social media data: towards methodological principles and praxis. *New Media & Society*, 20(9):3341–3358. doi: 10.1177/1461444817748953.

Hargittai, E. and Walejko, G. (2008). The participation divide: content creation and sharing in the digital age. *Information, Community and Society*, 11(2):239–256.

Hasan, M., Orgun, M. A., and Schwitter, R. (2017). A survey on real-time event detection from the twitter data stream. *Journal of Information Science.*

Hastie, T., Tibshirani, R., and Friedman, J. (2016). *The Elements of Statistical Learning. 2nd edition.* New York, NY, USA: Springer New York Inc.

Horvitz, D. G. and Thompson, D. J. (1952) A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association*, 47:260, 663-685.

Hsieh, Y. P. and Murphy, J. (2017). Total Twitter error: decomposing public opinion measurement on Twitter from a total survey error perspective. In *Total Survey Error in Practice: Improving Quality in The Era of Big Data*, pp. 23–46. Biemer, P. P., de Leeuw, E. D., Eckman, S, Edwards, B., Kreuter, F., Lyberg, L. E., Tucker, N. E., West, B. T (eds.), Wiley. doi: 10.1002/9781119041702.ch2.

Janetzko, D. (2017). The role of apis in data sampling from social media. *The SAGE Handbook of Social Media Research Methods.*, p.146.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C., and Usher, A. (2015). Big data in survey research: AAPOR task force report. *Public Opinion Quarterly*, 79(4):839–880. doi: 10.1093/poq/nfv039.

Kitchin, R. (2014). Big data, new epistemologies and paradigm shifts. *Big Data & Society*.

Klovdahl, A. S. (1989). Urban social networks: Some methodological problems and possibilities. In M. Kochen (ed.) *The Small World*. Norwood, NJ: Ablex Publishing, pp. 176–210.

Laney, D. (2001). 3D data management: controlling data volume, velocity, and variety. *META Group Research Note*, 6:70.

Lavallée, P. (2007). *Indirect Sampling*. Springer.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google flu: traps in big data analysis. *Science*, 343(6176):1203–1205.

Lomborg, S. and Bechmann, A. (2014). Using APIs for data collection on social media. *Science*, 343(6176):1203–1205.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L., and Gabrielli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, 31(2):263–281.

Mayr, P. and Weller, K. (2016). Think before you collect: setting up a data collection approach for social media studies. arXiv:1601.06296.

Mellon, J. and Prosser, C. (2016). Twitter and Facebook are not representative of the general population: political attitudes and demographics of social media users. *SSRN Electronic Journal*. doi: 10.2139/ssrn.2791625.

Morstatter, F., Pfeffer, J., Liu, H., and Carley, K. M. (2013). Is the sample good enough? Comparing data from Twitter's streaming api with Twitter's firehose. arXiv:1306.5204.

Newman, M. E. J. (2010). *Networks: An Introduction*. Oxford University Press.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

Office for National Statistics (2016). *Internet access in Great Britain: 2016*. Technical report, Office for National Statistics, UK.

Office for National Statistics (2017). *Internet access in the UK: 2017*. Technical report, Office for National Statistics, UK.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135. doi: `10.1561/1500000011`.

Patone, M., Zhang, L. C. (2019). On two existing approaches to statistical analysis of social media data. `arXiv:1905.00635`.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., and Shrimpton, L. (2013). Can Twitter replace newswire for breaking news? In *International Conference on Web and Social Media*.

Pfeffermann, D., Krieger, A., and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8:1087–1114.

Pfeffermann, D., and Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. *Analysis of survey Data*, 175–195. Wiley, New York, USA.

Pfeffermann, D., and Sverchkov, M. (2007). Small-Area Estimation under Informative Probability Sampling of Areas and Within the Selected Areas. *Journal of the American Statistical Association*, 102(480):1427–1439.

Pfeffermann, D., Eltinge, J. L., Brown, L. D. (2015). Methodological issues and challenges in the production of official statistics: 24th annual Morris Hansen lecture. *Journal of Survey Statistics and Methodology*, 3(4):425-483.

Rampazzo, F., Zagheni, E., Weber, I., Testa, M., and Billari, F. (2018). Mater certa est, pater numquam: what can Facebook advertising data tell us about male fertility rates? `arXiv:1804.04632`.

Rebecq, A. (2018). Extension sampling designs for big networks: application to Twitter. In *Springer Proceedings in Mathematics & Statistics*, pp. 251–270. doi: 10.1007/978-3-319-96941-1_17.

Rivers, D. (2007). Sampling for Web Surveys. In *Joint Statistical Meetings. "Internet 2010 in numbers."*.

Rosenbaum, P. R., and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, 57(2):377–387.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592. doi: 10.2307/2335739.

Sirken, M. G. and Levy, P. S. (1974). Multiplicity estimation of proportion based on ratios of random variable. *Journal of the American Statistical Association*, 69:68–73.

Sirken, M. G. (2004). Network sample survey of rare and elusive populations: a historical review. In *Preceedings of Statistics Canada Symposium: Innovative Methods for Surveying Difficult-to Reach Population*.

Sirken, M. G. (2005). Network Sampling. In *Encyclopedia of Biostatistics*, John Wiley & Sons, Ltd.

Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. Wiley.

Skinner, C.J. and Wakefield, J. (2017). Introduction to the design and analysis of complex survey data. *Statistical Science*, 32:165–175. doi: 10.1214/17-STS614.

Smith, T. M. F. (1983). On the validity of inferences from non-random sample. *Journal of the Royal Statistical Society. Series A*, 146(4):394–403. doi: 10.2307/2981454.

Snijders, T. A. B. (1992). Estimation on the basis of snowball samples: How to weight. *Bulletin de Methodologie Sociologique*, 36:59–70.

Stephan F. F., (1969). Three extensions of sample survey tecniques. In *New developments in survey samplings.*, pp. 81–104. Wiley, New York.

Sverchkov, M. and Pfeffermann, D. (2004). Prediction offinite population totals based on the sample distribution. *Surv. Method.*, 30:79–92.

Swier, N., Komarniczky, B., and Clapperton, B. (2015). *Using Geolocated Twitter Traces to Infer Residence and Mobility.* Technical report, GSS Methodology Series.

Tabassum, S., Pereira, F. S. F., Fernandes, S., and Gama, J. (2018). Social network analysis: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):e1256. doi: 10.1002/widm.1256.

Taylor, Sean J (2013). Real Scientists Make Their Own Data. *Sean J. Taylor Blog.*

Thompson, S. K. (1990) Adaptive cluster sampling. *Journal of the American Statistical Association*, 85:1050–1059.

Thompson, S. K. (2012). *Sampling.* John Wiley & Sons, Inc.

Valliant, R., and Dever J. A. (2011). Estimating propensity adjustments for volunteer web surveys. *Sociological Methods & Research*, 40(1):105-137.

Wang, Y., Callan, J., and Zheng, B. (2015). Should we use the sample? analyzing datasets sampled from Twitter's stream API. *ACM Transactions on the Web (TWEB)*, 9(3):13. doi: 10.1145/2746366.

Yan, W., Wenchao, Y., Sam, L., and Sean, D. Y. (2019). The relationship between social media data and crime rates in the united states. *Social Media + Society*, 5(1):1–9. doi: 10.1177/2056305119834585.

Yildiz, D., Munson, J., Vitali, A., Tinati, R., and Holland, J. A. (2017). Using Twitter data for demographic research. *Demographic Research*, 37(46):1477–1514. doi: 10.4054/DemRes.2017.37.46.

Young, S. S. and Karr, A. (2017). Deming, data and observational studies. *Significance,*, 8:116–120.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, 66(1):41–63. doi: 10.1111/j.1467-9574.2011.00508.x.

Zhang, L.-C. and Patone, M. (2017). Graph Sampling *Metron*, 75:277–299.

Zhang, L.-C. (2018). On the use of proxy variables in combining register and survey data. In *Administrative Records for Survey Methodology*. John Wiley and Sons Ltd, to appear. Chun, A. Y., Larsen, M. D. (eds.), Wiley Series in Survey Methodology.

Zhang, L.-C. (2019). On valid descriptive inference from non-probability sample. *Statistical Theory and Related Fields*, doi: 10.1080/24754269.2019.1666241.