SAMPLING DESIGNS AND ROBUSTNESS FOR THE ANALYSIS OF NETWORK DATA

by
Marios Papamichalis

A dissertation submitted in partial
fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Supervisor: Patrick J. Wolfe
Second Supervisor: Simon Lunagomez
Department of Statistical Science
University College London
February 11, 2019

I, Marios Papamichalis, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

February 11, 2016

(Marios Papamichalis)

Impact Statement

This thesis is consisted from three projects. In the first project, our objective is to develop a new method for comparing sampling design on network data by using information theory. We make the case that different designs are more suitable for different random network models. At the end of the day, a practitioner can use our framework to associate sampling designs which are more informative for different random networks. Regarding the theory of this section, we associate the problem with the rationality behind the statistical interpretation of the reference priors. As a consequence, we show that following the same assumptions of the reference priors, a framework for comparing sampling designs on network data can be derived.

Subsequently, in the second part of this thesis we use the previous Bayesian algorithm to provide valid and generic ways to translate statements for partially observed networks to fully observed networks and via versa. The most interesting thing we want to investigate is to theoretically and practically understand what information is required to enable us to use statements at the level of partially observed networks and turn them into statements for fully observed networks that they produced them. Our goal is to investigate to what extend a statement that we can make for a partially observed network can be translated to a statement to a fully observed network. We prove that in the general case the answer is that the joint distribution of their features and the sampling designs is required in order to create relevant statements.

In the third project, we adopt the statistical framework on robustness to provide tools to the modeler to evaluate how the quality of inference for a specific feature of a random network model is degraded when the approximating model is misspecified. We try to answer how sensitive could be the quality of an inference be when the data is not coming exactly from the true exchangeable model. More specifically, we provide methodology to examine whether and how much

an approximating random network model is suitable for describing a true random network model in terms of a specific feature. In terms of methodology, our main challenge is to combine stochastic optimization and graph limits tools to explore the model space.

Explicitly, the benefits inside academia, as described above is the development of methodology which solves practical problems for practitioners regarding sampling designs and robustness on random networks. To the best of our knowledge the research around the first topic has not received much attention and around the second topic is the first attempt when we deal with networks. The benefits outside academia could occur to firms which are active with social, computer, biological and information networks. Our frameworks can be used either to perform inference when a sampling design is applied to the generative mechanism that produced the network date or to check and criticize how robust an attribute of a generative model that produces network data is.

Abstract

This manuscript addresses three new practical methodologies for topics on Bayesian analysis regarding sampling designs and robustness on network data:

- In the first part of this thesis we propose a general approach for comparing sampling designs. The approach is based on the concept of data compression from information theory. The criterion for comparing sampling designs is formulated so that the results prove to be robust with respect to some of the most widely used loss functions for point estimation and prediction. The rationale behind the proposed approach is to find sampling designs such that preserve the largest amount of information possible from the original data generating mechanism. The approach is inspired by the same principle as the reference prior, with the difference that, for the proposed approach, the argument of the optimization is the sampling design rather than the prior. The information contained in the data generating mechanism can be encoded in a distribution defined either in parameter's space (posterior distribution) or in the space of observables (predictive distribution).

  The results obtained in this part enable us to relate statements about a feature of an observed subgraph and a feature of a full graph. It is proven that such statements can not be connected by invoking conditional statements only; it is necessary to specify a joint distribution for the random graph model and the sampling design for all values of fully and partially observed random network features. We use this rationale to formulate statements at the level of the sampling graph that help to make non-trivial statements about the full network. The joint distribution of the underlying network and the sampling mechanism enable the statistician to relate both type of conditional statements. Thus, for random network partially and fully observed features joint distribution is considered and useful statements for practitioners are provided.

- The second general theme of this thesis is robustness on networks. A method for robustness on exchangeable random networks is developed. The approach is inspired by the concept of graphon approximation through a stochastic block model. An exchangeable model is assumed to infer a feature of a random networks with the objective to see how the quality of that inference gets degraded if the model is slightly modified. Decision theory methods are considered under model misspecification by quantifying stability of optimal actions to perturbations to the approximating model within a well defined neighborhood of model space. The approach is inspired by all recent developments across the context of robustness in recent research in the robust control, macroeconomics and financial mathematics literature.

In all topics, simulation analysis is complemented with comprehensive experimental studies, which show the benefits of our modeling and estimation methods.

Acknowledgements

I would like to express my gratitude to my first supervisor, Prof. Patrick Wolfe, whose encouragement and generous support during my PhD has been invaluable to me. His insightful research guidance is beyond my ability to describe in words. I remembered the first time we talked during my visit at UCL, I felt so inspired that I knew I wanted him to be my advisor.

I am also deeply grateful to Dr. Simon Lunagomez, who has been to me a great friend and advisor rather than simply a co-author. I was very fortunate to meet him the years of my PhD. His enthusiasm and advises, both in professional and personal level, were catalytic for me. Without him, this thesis would not be possible. I will always be indebted to him. Thank you.

In professional level I would like to thank the Stochastic Processes Group at UCL and the Design, Bayes, and Causal at Purdue, for their curiosity and the great seminars with many enlightening questions. A special mention is needed for Prof. Arman Sabbaghi, Prof. Karthik Kannan, Prof. Bilionis, Prof. Alex Beskos, Prof. Dellaportas, Dr. Sam Livingstone, Dr. Arthur Barthe, Dr. Beate Franke, Tran Viet Long and Prof. Sofia Olhede for their great advices, creativity, energy, sharing their experience, and always readily lending an ear.

I would like to thank Fyodor Dostoevsky, Anton Chekhov, Albert Camus, Martin Heidegger, Ernest Hemingway, Julio Cortazar, Gabriel Garca Mrquez,Mario Vargas Llosa, Carlos Fuentes, Constantine P. Cavafy, Epicurus, Ingmar Bergman, Federico Fellini, Theodoros Angelopoulos, Akira Kurosawa, Pier Paolo Pasolini, Bernardo Bertolucci, Luchino Visconti di Modrone etc for being my resorts in the lonely times of my life and bringing peace to my heart.

In personal level, I am really grateful to my friend Maria-Lida Kounadi for

making me grow my personality. Without her, my life would be completely different. I am eternally indebted to her and I will never forget the things she taught me. Special thanks to my cozy and warm papamix family. Last but not least, I would like to thank my friends, George Rovis, Sotiris Sotiriou, Panos Maroudas, Alfredos Theodorakopoulos, Orestis Polychronakis, Dr. Panos Vekris, Giannis Vlachoulis, Eleanna Mixa, Nick Angelopoulos, Dr. Kyriakos Ispoglou, Giannis Gemenis, Giannis Orfanidis, Katerina and Angelos Zymvragoudakis, Christina Golfi etc for their understanding and support. Mountains of love to Katerina and Thanasis Katopodis, Valia Emmanouilidi, all the guys from UK Gate 13 and to my labradorable child Ben. Thank you guys and I will never forget what you did for me in my ups and downs.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The purpose of the chapter is to provide enough context regarding network data so that we can state what the contributions of the thesis are. The topics of this thesis involve the study of models on network data. Network data is currently receiving considerable attention from the mathematics, computer science and engineering communities because of its relevance to real-world networks. In principle, network models enable us to draw inference from data, but for the results to be defensible we must be able to quantify them and judge them. Many systems of critical importance are commonly modeled as networks. Understanding networks and anticipating weaknesses of different modeling approaches are vital task to numerous applications. Here, we continue to explore three emerging topics at the interface between random graph theory and network science and we aim at facilitating the transfer of ideas, insights and interdisciplinary approaches to tackle three exciting problems in random graphs and real networks.

The thesis is structured as follows: In chapter 1, we first motivate the problems addressed in this thesis and 1) describe why network data and models are important and statistician care about them 2) provide a taxonomy how people use Bayesian inference on network data and why 3) present a detailed discussion of the prominent challenges for performing Bayesian inference on networks and finally 4) briefly mention our contribution. We describe the current state-of-the-art for networks in this regard. Having mentioned the complexity of the data above, all the three problems we consider in this thesis are practical and are tackled with computational Bayesian methods which provide practitioners and statisticians with tools flexible enough to overcome problems which are hard (or even impossible) to be solved through analytical frameworks. Our review of the

current challenges in network modeling is not exhaustive but we rather focus our discussion on the three topics motivated above. In chapter 2 we present all the random network model and their features we will use in this thesis. Having established a framework and given the right context, we then proceed in chapters 3-5 which turn to our original contributions. Lastly, in chapter 6 we recap this thesis by stating what we have learnt from this work, enumerating its limitations and discussing potential extensions for the future.

1.1 Motivation

1.1.1 Networks

In many sciences there has been a conceptual shift away from the study of individual entities and towards the analysis of entire systems-not least because of the technological advances that enable us to collect the corresponding data. In every system, these entities interact either directly or induced as a summary of their dependencies. Networks give us a means to describe and analyze these interactions between entities. In contrast to classical statistics, networks allow us to model complex dependencies while assuming very little structure. For instance, there is no natural ordering and thus no geometry inherited in a network as it is in time series or spatial statistics.

More recently, starting perhaps in the early to mid 1990s, there has been an explosion of interest in networks and network-based approaches to modeling and analysis of complex systems. Much of the impetus for this growth derives from work by researchers in two particular areas of science: statistical physics and computer science. To the former can be attributed a seminal role in encouraging what has now become a pervasive emphasis across the sciences on understanding how the interacting behaviors of constituent parts of a whole system lead to collective behavior and systems-level properties or outcomes. Indeed the term complex system was coined by statistical physicists, and a network-based perspective has become central to the analysis of complex systems. To the latter can be attributed much of the theory and methodology for conceptualizing, storing, manipulating, and doing computations with networks and related data, particularly in ways that enable efficient handling of the often massive quantities of such data. Moreover, information networks (e.g, the World Wide Web) and related social media applications (e.g., Twitter), the development of which computer scientists have played a

key role, are examples of some of the most studied of complex systems (arguably reflecting our continued fascination with studying ourselves!).

More broadly, a network-based perspective recently has been found to be useful in the study of complex systems across a diverse range of application areas. These areas include computational biology [85, 92, 96] (e.g., studying systems of interacting genes, proteins, chemical compounds, or organisms), engineering [35, 66] (e.g., establishing how best to design and deploy a network of sensing devices), finance [1, 73] (e.g., studying the interplay among, say, the world's banks as part of the global economy), marketing (e.g., assessing the extent to which product adoption can be induced as a type of contagion), neuroscience [33, 54] (e.g., exploring patterns of voltage dynamics in the brain associated with epileptic seizures), political science [105] (e.g., studying how voting preferences in a group evolve in the face of various internal and external forces), and public health [30] (e.g., studying the spread of infectious disease in a population, and how best to control that spread).

In general, two important contributing factors to the phenomenal growth of interest in networks are (i) an increasing tendency towards a systems-level perspective in the sciences, away from the reductionism that characterized much of the previous century, and (ii) an accompanying facility for high-throughput data collection, storage, and management. The quintessential example is perhaps that of the changes in biology over the past 10-20 years, during which the complete mapping of the human genome, a triumph of computational biology in and of itself, has now paved the way for fields like systems biology to be pursued aggressively, wherein a detailed understanding is sought of how the components of the human body, at the genetic level and higher, work together.

1.1.2 Motivating examples of networks in science

In order to better appreciate the nature of the statistical foundation emerging in the analysis of network data, it is useful to have some initial sense of the contexts in which networks arise, the scientific questions being asked, and the measurements being taken. For convenience, and following [75], the presentation is organized loosely into four classes of networks: technological, social, biological, and informational. These divisions are intended to be soft, and not hard, as many networks can be said to fall into more than one category.

Technological Networks

Figure 1.1: Technological Network.

Arguably the networks most familiar to us are those of a technological nature (i.e., human constructions consciously created in a network form). Examples include communication networks (e.g., telephone networks or the Internet like in figure 1.1), transportation networks (e.g., networks of roads or rails, or networks of airline routes), and energy networks (e.g., networks for delivery of electricity or gas, or electrical circuits). Consider the rather celebrated example of the Internet, which is essentially a network of digital devices communicating over wired and wireless connections via a set of communication protocols. Network-oriented questions regarding the Internet tend to focus on those relating to its topology, the traffic it carries, the interaction of the two, and in turn the interaction of those with social and economic factors. For example, in regards to topology we may ask: What does the Internet look like, how big is it and what are its structural characteristics. In terms of traffic, questions include: How much traffic is flowing across the network, how can I distinguish between normal and anomalous traffic and does my network have the capacity to meet anticipated demands.

Social Networks

Figure 1.2: Social Network.

Specific examples of social networks (figure 1.2) include networks of friendships among school children, sexual contacts within a community, corporate alliances among businesses, email exchanges between individuals, co-authorship on scientific articles, and trade agreements among nations. The study of such networks is of particular interest to, and has traditionally been the province of, researchers in social sciences like sociology, anthropology, and psychology, although this interest is increasingly shared now by researchers in a number of other areas, such as business and public health. The focus in these areas typically is on social structure and the quantitative characterization and analysis of such structures. Questions of interest include: Who interacts with whom and what factors influence the tendency to interact, which interactions are mutual, whether there are friends of friends also friends, what social groups, if any, exist in the network, who are the power brokers, who is central to the network and who is peripheral and which actors are similar in the roles they play. In recent years, Internet has begun to have a fascinating impact on the field of social network analysis, due both to the potential for large-scale data acquisition and storage and the actual types of social interactions facilitated by the Internet.

Biological Networks

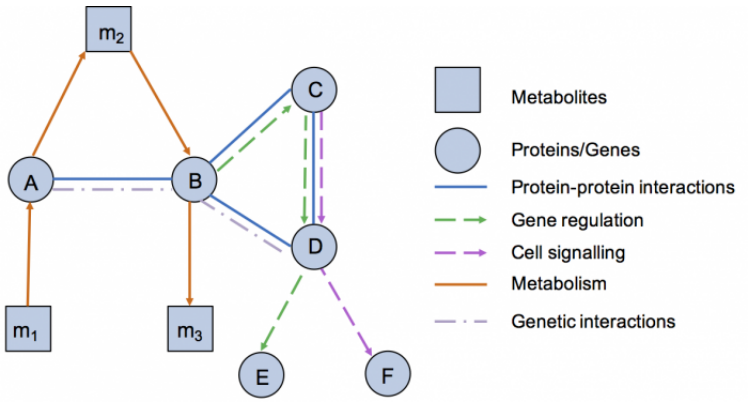

Figure 1.3: Biological Network

Networks are a natural and commonly used tool for representing the internal workings of biological systems, at all different scales. For example, intra-cellular networks of interest include those describing the regulatory behavior among genes, the physical affinity for binding among proteins, the participation of metabolites together in biochemical processes, and combinations thereof (figure 1.3). Similarly, a well-known example of an inter-cellular network is a network of neurons. On the other hand, networks describing interactions among complete organisms include ecological networks, such as those describing predator-prey relationships, and epidemiological networks, characterizing the spread of disease in a population. Not surprisingly, the nature of the data collected on biological networks and the manner in which they are analyzed and used vary widely with the nature of the underlying biological system being studied and our ability to obtain relevant measurements.

Information Networks

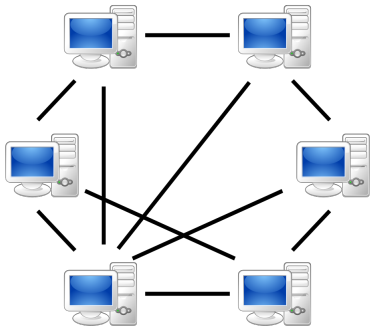

Figure 1.4: Peer to Peer Network.

Of particular use in this modern information age, although by no means new, are information networks (i.e., networks describing relationships among elements of information). Standard examples include networks of citations between academic journals or chapters, networks of co-authorship on chapters, or networks indicating semantic relationships (e.g., synonym, antonym, etc.) between words or concepts. In addition, the Internet has helped spawn a number of well-known classes of information networks. The preeminent example is the World Wide Web (WWW), in which nodes typically are web pages and edges indicate the referencing of one page by another. Another class of Internet-related information networks are peer-to-peer (i.e., P2P, figure 1.4), networks, in which nodes are typically Internet users and links indicate the exchange of content (e.g., music or movies) through an associated network protocol (e.g., Napster, Gnutella, KaZaa, etc.). As an illustration of an information network, consider the network depicted in figure 1.4, which is an example of an important class of sub-networks of the WWW called web-logs or simply blogs. Additionally, there is generally strong interest in questions regarding the structure of such networks, including which nodes are linked to many other nodes (e.g., Who are the most highly cited authors within the mathematical sciences literature?), whether certain tightly inter-woven subgraphs may be found (e.g., How does the content of web pages induce clustering on the

WWW?), and the manner in which network size and structure change over time (e.g., What are the dynamics of the lifetime of a scientific innovation?).

1.2 Brief review of network modeling

The review or survey in this section provides content for computational approaches regarding network modeling. Network modeling applications is proved to be a fertile ground for researchers in statistics to make advances. There has been a lot of effort to provide models in network data. Some of them described explicitly in the next chapter are Erdös-Rényi models [31, 32], Stochastic Block models [8, 47, 77], Exponential Random Graph Models [34, 89], small world models [111], preferential attachment models [13] and Latent space models [46]. Fitting those models gave rise to computational challenges, like computing Bayes Factors, computing intractable integrals, scaling of models etc.. Here we present literature that combines Bayesian computational methods on networks. This literature review includes articles in the modeling and the computational side regarding network data and does not mean to be comprehensive.

In [4] the authors introduce a class of variance allocation models for pairwise measurements: mixed membership stochastic blockmodels. These models combine global parameters that instantiate dense patches of connectivity (blockmodel) with local parameters that instantiate node-specific variability in the connections (mixed membership). They develop a general variational inference algorithm for fast approximate posterior inference. They demonstrate the advantages of mixed membership stochastic blockmodels with applications to social networks and protein interaction networks. Moreover, modeling relational data is an important problem for modern data analysis and machine learning. In [5], they propose a Bayesian model that uses a hierarchy of probabilistic assumptions about the way objects interact with one another in order to learn latent groups, their typical interaction patterns, and the degree of membership of objects to groups. Their model explains the data using a small set of parameters that can be reliably estimated with an efficient inference algorithm.

In the same spirit, in [67], an efficient MCMC algorithm is presented to cluster the nodes of a network such that nodes with similar role in the network are clustered together. This is known as block-modeling or block-clustering. The model

is the stochastic blockmodel (SBM) with block parameters integrated out. The resulting marginal distribution defines a posterior over the number of clusters and cluster memberships. Sampling from this posterior is simpler than from the original SBM as transdimensional MCMC can be avoided. The algorithm is based on the allocation sampler. It requires a prior to be placed on the number of clusters, thereby allowing the number of clusters to be directly estimated by the algorithm, rather than being given as an input parameter.

Embedding dyadic data into a latent space has long been a popular approach to modeling networks of all kinds. While clustering has been done using this approach for static networks, this chapter gives two methods of community detection within dynamic network data, building upon the distance and projection models previously proposed in the literature. In [95], the authors proposed approaches capture the time-varying aspect of the data, can model directed or undirected edges, inherently incorporate transitivity and account for each actor's individual propensity to form edges.

Many networks of scientific interest naturally decompose into clusters or communities with comparatively fewer external than internal links; however, current Bayesian models of network communities do not exert this intuitive notion of communities. In [71], the authors formulate a non parametric Bayesian model for community detection consistent with an intuitive definition of communities and present a Markov chain Monte Carlo procedure for inferring the community structure. Similarly, in [104] introduce a Bayesian estimator of the underlying class structure in the stochastic block model, when the number of classes is known. The estimator is the posterior mode corresponding to a Dirichlet prior on the class proportions, a generalized Bernoulli prior on the class labels, and a beta prior on the edge probabilities.

Scientists have shown that network motifs are key building block of various biological networks. Most of the existing exact methods for finding network motifs are inefficient simply due to the inherent complexity of this task. In recent years, researchers are considering approximate methods that save computation by sacrificing exact counting of the frequency of potential motifs. However, these methods are also slow when one considers the motifs of larger size. In [93], they propose two methods for approximate motif finding based on Markov Chain Monte Carlo (MCMC) sampling. Both the methods are significantly faster than the best of the existing methods, with comparable or better accuracy.

Exponential random graph models are a class of widely used exponential family models for social networks. The topological structure of an observed network is modeled by the relative prevalence of a set of local sub-graph configurations termed network statistics. One of the key tasks in the application of these models is which network statistics to include in the model. This can be thought of as statistical model selection problem. This is a very challenging problem-the posterior distribution for each model is often termed "doubly intractable" since computation of the likelihood is rarely available, but also, the evidence of the posterior is, as usual, intractable. The contribution of [23] is the development of a fully Bayesian model selection method based on a reversible jump Markov chain Monte Carlo algorithm extension of their previous algorithms which estimates the posterior probability for each competing model.

Usually we are dealing with situations where we have networks that we can not fully observe. Thus, we have networks that we partially observe due to a sampling mechanisms. Some of the most widely used sampling mechanisms that propagate through a social network are defined in terms of tuning parameters. In [6, 63] the authors are interested in the problem of optimizing these tuning parameters with the purpose of improving the inference of a population quantity, where such quantity is a function of the network. In [7], this is done by formulating the problem in terms of Decision Theory. The optimization procedure for different sampling mechanisms is illustrated via simulations in the fashion of the ones used for Bayesian clinical trials.

As we mentioned, there has been a lot of effort in modeling network data and formulating network models. In addition to that there are computational challenges that arise from fitting these models. There are additional challenges that need to be solved, as well. In this thesis, we are combining networks with computational methods in order to approach and tackle three challenging problems. MCMC (or variational Bayes) is required when the likelihood cannot be computed analytically. This is why Bayesian probability and statistics fell out of favor (and even view) for a long time. When most people think about "classical" statistics, they think about frequentist methods. But they came later, simply because at the time, Bayesian methods that were realistic were very hard to fit without either too many assumptions or computers. There are many things MCMC can offer. Even if our integral is computable (and with modern methods and computers, it's easier to compute difficult integrals), if there is high dimensionality, it will take

a lot longer to compute an integral exactly than it will to use MCMC, with only slight degradation in the solution if it's done correctly. Indicative examples from the literature are presented in the next subsection.

1.3 Contributions of the thesis and their context

As a first new method we compare Sampling Designs on random networks via information theory. This problem elaborate sampling techniques on networks. Most of the times we can not observe the whole network but only part of it. Therefore we have a partially observed network. There is a literature that deals with those kinds of networks. This set of problems is new in research and it has not receive as much attention as it should. The main reason why we deal with this problems is that there are situations where the practitioner have some control in the sampling process. We are interested in providing him/her with some insight about how to make better decisions for his/her objectives, regarding the appropriate combination of the network models and the sampling designs. In chapter 3 we are dealing with this problem and we resort to information theory tools to provide a principle method to compare and consequentially to suggest suitable sampling designs for different random network models.

The second problem we tackle, in chapter 4, is to provide useful statements for random network features. We discuss about and why it is important to relate statements about features of fully observed random networks and of partially observed random networks. Approaches that invokes only conditional statements are not effective because they do not combine the two levels of uncertainty regarding the network model and the sampling design. They allow us to make only separate statements. To be able to compute statements about the fully and partially observed graph we relate the uncertainty of the model and the sampling design. The theoretical results we provide are based on the algorithm of chapter 3 and can be extended easily to coarsening data [45] following the same logic. The connection between those statements is useful and helpful for practitioners in order to get and insight how different features are related in different random networks and sampling designs. We are able to answer questions like: How is community structure related with the degree distribution of a random network when we have a partially observed network and what can we say about the fully observed network (and via versa)? All the above provide general guidelines of how to construct those statements.

The last problem we are dealing with, in chapter 5, is the development of a framework for robustness on exchangeable networks based on [110]. Suppose we have our model, we fit it to network data and perform inference with it. Though, we do not take our parametric assumptions for granted. We put our assumptions into question. Therefore, we are willing to check how the quality of inference degrades if the model is misspecified. In order to achieve this, we combine recent developments in robustness and exchangeable random network models.

Chapter 2

Preliminaries and Definitions

In this chapter, we introduce the terminology and notation regarding network data, which is will to be used in the rest of the thesis. In the first section networks and their properties are presented. In the second section random network model, which are distributions over network data, and their features are presented. Specifically, those features are properties of the distribution and not properties of a realization of a network.

2.1 Networks and Properties

We define a $graph$ (i.e., $network$) $G = (\mathcal{V}, \mathcal{E})$ as a tuple of the set of nodes $\mathcal{V}$ and the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. We call the number of nodes $n = \mid V \mid$ the size of the graph and the $density$ denotes the proportion of observed edges N over the number of possible edges (assuming no loops):

$$deg(G) = \frac{N}{n(n-1)/2} \tag{2.1}$$

The density of a network lies between 0 and 1, with 0 being the empty network of no edges and 1 if there is an edge between all pairs of nodes. Representing by $A_{ij} \geq 0$ an edge between nodes i and j, we can describe the entire network using its $adjacency\ matrix\ A = (A_{ij})\ i, j = 1, \ldots, n$. We call a graph binary if two nodes i and j are either connected ($A_{ij} = 1$), or not ($A_{ij} = 0$). Figure 2.1 illustrates how to turn a binary network into an adjacency matrix for a toy example. We will see that adjacency matrices make generalizations of graphs easy and are useful for the analyses of networks.

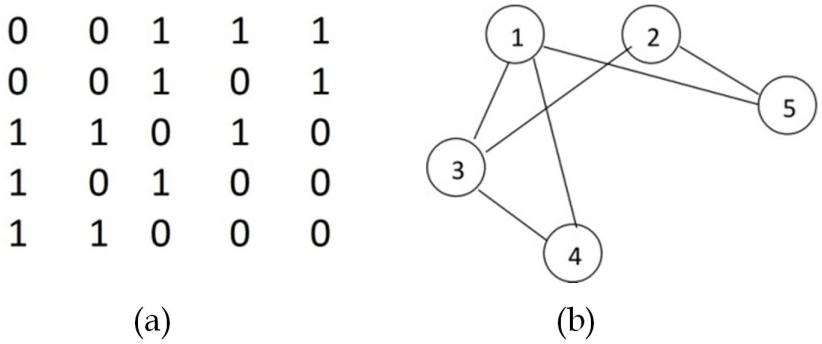

Figure 2.1: Adjacency Matrix (a) to Network(b).

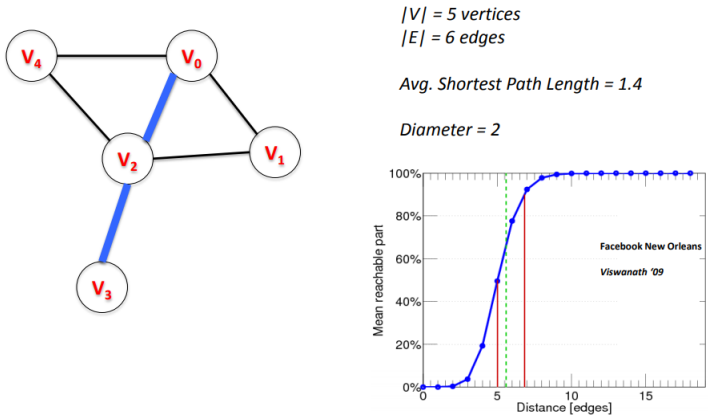

Figure 2.2: Example of a network characteristics. Average shortest path length is the the average of all the shortest paths from one node to another. Diameter is the largest path between two nodes of the network.

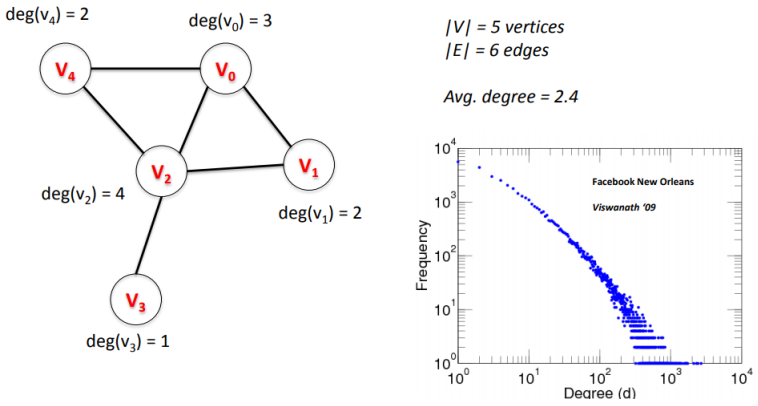

Figure 2.3: Another example of network characteristics. Average degree is the average of all nodes degrees of the network.

Each of the networks in the introduction can formally be described as a $simple$ graph. We group these networks by the nature of their relationships. A prominent binary graph is a simple network where we assume in addition to $A_{ij} \in \{0, 1\}$: the relationships are $symmetric$ ($A_{ij} = A_{ji}, \forall i, j$); and there are no self-loops, i.e., a node cannot connect to itself ($A_{ii} = 0, \forall i$). Friendship networks for instance are often modeled as simple graphs. Networks where two nodes can have more than one edge are called multi-edge networks; e.g., an email interaction network. When the connections between nodes i and j are quantified with a weight we call the network $weighted$; and networks where the relationships are not symmetric are called $directed$ networks. For the scope of this thesis, we concentrate on $undirected$ networks without self-loops (i.e., A is symmetric and $A_{ii} = 0, \forall i$), unless otherwise specified.

The degree $d_i = \sum_{i \neq j} A_{ij}$ denotes the number of connections of node i, as illustrated in figures 2.2 and 2.3. The degree plays a central role for this work as we will see later. In practice, scientists often analyze the degree sequence of an observed network, which is a vector of all degrees sorted in non-decreasing order. To discuss community structure, we partition nodes into groups (i.e., communities). The function g denotes the community assignment of the network such that $g(i)$ denotes the group of node i.

A $walk$ on a graph is a sequence of alternating nodes and edges ($v_0, e_1, v_1, e_2, v_2,$

$\ldots, v_l$);where the edge e_{i+1} between nodes v_i and v_i+1 needs to be present in the network for $i = 0, \ldots, l-1$. The $length$ of this walk is said to be l. A $cycle$ is a walk of length at least three that starts and ends at the same node but does not pass through any other node twice. A $path$ is a walk without repeated nodes and edges. The $distance$ between two nodes is the length of the shortest path connecting them where for weighted networks we calculate the sum of the weights. The $diameter$ (figure 2.2) of a graph is the longest distance between any two nodes in the graph. A graph is called $connected$ if there exists a walk from every node to every other node. A $component$ is a maximally connected subgraph; i.e., adding any other node to this subgraph would break the connectedness. The component of a graph that includes the largest number of nodes is called the largest component. A graph where there is an edge between every two nodes is called complete and a complete subgraph is called a $clique$. In $regular$ graphs, every node has the same degree.

2.2 Random Networks Models and Features

Random graph is the general term to refer to either probability distributions over graphs or to a random process which generates them. The theory of random graphs lies at the intersection between graph theory and probability theory. From a mathematical perspective, random graphs are used to answer questions about the properties of typical graphs. Its practical applications are found in all areas in which complex networks need to be modeled - a large number of random graph models are thus known, mirroring the diverse types of complex networks encountered in different areas. For the rest of this thesis, in order to be consistent with notation, we denote a random network by the symbol $\mathcal{G}$ and the network realization that is produced with $\mathcal{G}(\omega)$.

2.2.1 Random Network Features

Here we provide a brief description of all the features we are dealing with in this thesis. More extensive reviews can be found in [53, 75, 112]. For sake of simplicity we denote $\tau(\mathcal{G})$ every feature of random model $\mathcal{G}$. Our main goal is to:

- Provide the intuition behind their usefulness
- Briefly, provide notation which is going to be consistent with the notation in chapters 3-5

- Make the reader understand what kind of questions do they answer
- Show how they are computed

Centrality

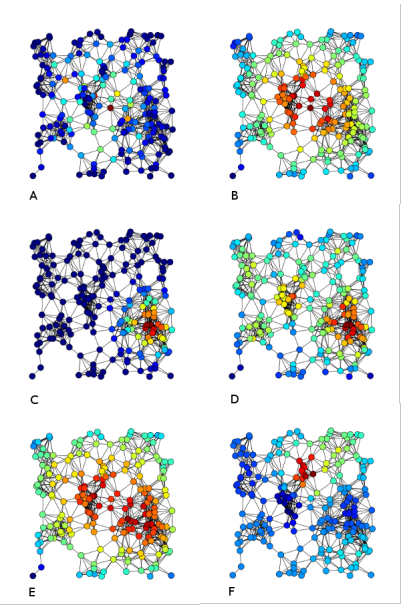

Figure 2.4: Examples of A) betweenness centrality, B) closeness centrality, C) eigenvector centrality, D) degree centrality, E) harmonic Centrality and F) katz centrality of the same graph. From dark blue to deep red are depicted the nodes in increasing order of how cetralized they are.

$Centrality$ indices are answers to the question: ”What characterizes an important vertex?”. The answer is given in terms of a real-valued function on the vertices of a graph, where the values produced are expected to provide a ranking which identifies the most important nodes.

The word ”importance” has a wide number of meanings, leading to many different definitions of centrality. Two categorization schemes have been proposed. ”Importance” can be conceived in relation to a type of flow or transfer across the network. This allows centralities to be classified by the type of flow they consider

important."Importance" can alternatively be conceived as involvement in the cohesiveness of the network. This allows centralities to be classified based on how they measure cohesiveness. Both of these approaches divide centralities in distinct categories. Restricting consideration to this group allows for a soft characterization which places centralities on a spectrum from walks of length one (degree centrality) to infinite walks (eigenvalue centrality). The observation that many centralities share this familial relationships perhaps explains the high rank correlations between these indexes.

Degree

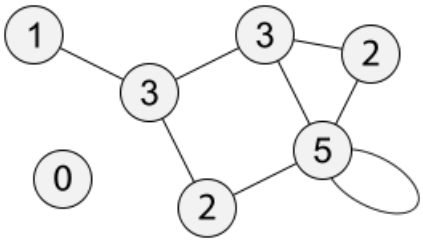

Figure 2.5: A graph with vertices labeled by degree

In graph theory, the $degree$ of a vertex of a graph is the number of edges incident to the vertex, with loops counted twice. The degree of a vertex v is denoted $\deg(v)$ or $\deg v$. The maximum degree of a graph G, denoted by $\Delta(G)$, and the minimum degree of a graph, denoted by $\delta(G)$, are the maximum and minimum degree of its vertices. In the graph on the right, the maximum degree is 5 and the minimum degree is 0.

Subgraphs-Motifs

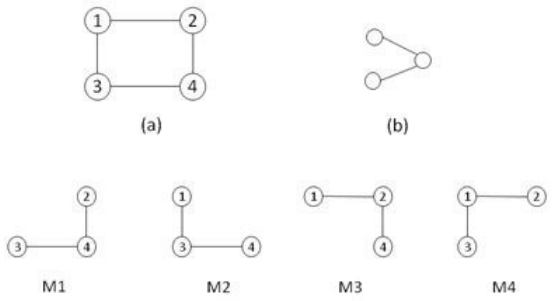

Figure 2.6: Different occurrences of a sub-graph in a graph. (M1-M4) are different occurrences of sub-graph (b) in graph (a). For frequency concept F1, the set M1, M2, M3, M4 represent all matches, so F1 = 4. For F2, one of the two set M1, M4 or M2, M3 are possible matches, F2 = 2. Finally, for frequency concept F3, merely one of the matches (M1 to M4) is allowed, therefore F3 = 1.

Network $motifs$ are subgraphs that repeat themselves in a specific network or even among various networks. Each of these subgraphs, defined by a particular pattern of interactions between vertices, may reflect a framework in which particular functions are achieved efficiently. Indeed, motifs are of notable importance largely because they may reflect functional properties. They have recently gathered much attention as a useful concept to uncover structural design principles of complex networks. Although network motifs may provide a deep insight into the network's functional abilities, their detection is computationally challenging.

Degree distribution

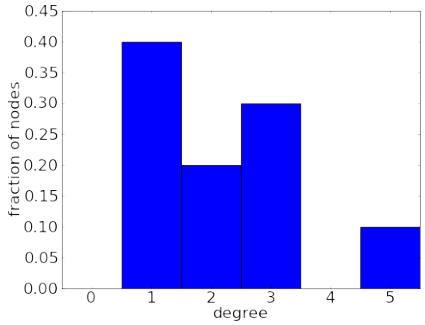

Figure 2.7: Degree distribution of an undirected network

In the study of graphs and networks, the degree of a node in a network is the number of connections it has to other nodes and the $degree\ distribution$ is the probability distribution of these degrees over the whole network. The degree distribution $P(k)$ of a network is then defined to be the fraction of nodes in the network with degree k. Thus if there are n nodes in total in a network and n_k of them have degree k, we have $P(k) = \frac{n_k}{n}$.

Assortativity

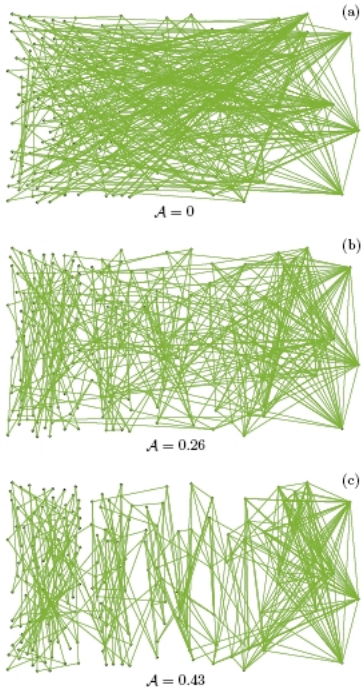

Figure 2.8: Scale-free networks for different degrees of assortativity: (a) A = 0 (uncorrelated network), (b) A = 0.26, (c) A = 0.43, where A indicates r (the assortativity coefficient, as defined in this sub-section).

$Assortativity$, or $assortative\ mixing$ is a preference for a network's nodes to attach to others that are similar in some way. Though the specific measure of similarity may vary, network theorists often examine assortativity in terms of a node's degree. The addition of this characteristic to network models more closely approximates the behaviors of many real world networks.

Correlations between nodes of similar degree are often found in the mixing patterns of many observable networks. For instance, in social networks, nodes tend to be connected with other nodes with similar degree values. This tendency is referred to as assortative mixing, or assortativity [74]. On the other hand, technological and biological networks typically show disassortative mixing, or disas-

sortativity, as high degree nodes tend to attach to low degree nodes.

Distance

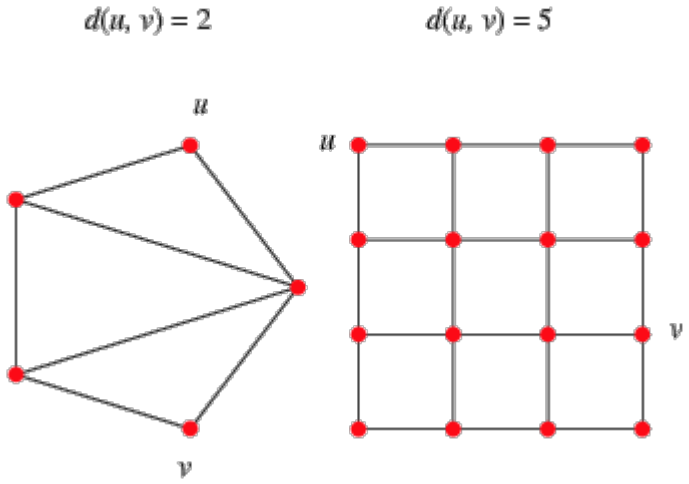

Figure 2.9: Distance

The $distance$ between two vertices in a graph is the number of edges in a shortest path (also called a graph geodesic) connecting them. This is also known as the geodesic distance. Notice that there may be more than one shortest path between two vertices. If there is no path connecting the two vertices, i.e., if they belong to different connected components, then conventionally the distance is defined as infinite.

In the case of a directed graph the distance $d(u, v)$ between two vertices u and v is defined as the length of a shortest directed path from u to v consisting of arcs, provided at least one such path exists. Notice that, in contrast with the case of undirected graphs, $d(u, v)$ does not necessarily coincide with $d(v, u)$, and it might be the case that one is defined while the other is not.

Modularity

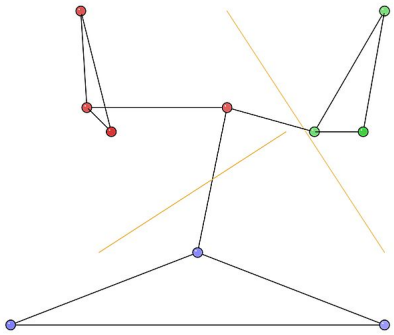

Figure 2.10: Sample Network corresponding to the Adjacency matrix with 10 nodes, 12 edges. Network partitions that maximize Q. Maximum Q=0.4896

The $Modularity$ was designed to measure the strength of division of a network into modules (also called groups, clusters or communities). Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules. Modularity is often used in optimization methods for detecting community structure in networks.

Let us consider a graph with n nodes and m links (edges) such that the graph can be partitioned into two communities using a membership variable s. If a node v belongs to community 1, $s_v = 1$, or if v belongs to community 2, $s_v = -1$. Let the adjacency matrix for the network be represented by A, where $A_{vw} = 0$ means there is no edge (no interaction) between nodes v and w and $A_{vw} = 1$ means there is an edge between the two. Also for simplicity we consider an undirected network. Thus $A_{vw} = A_{wv}$. (It is important to note that multiple edges may exist between two nodes, but here we assess the simplest case).

The modularity, often denoted by Q_s, is then defined as the fraction of edges that fall within group 1 or 2, minus the expected number of edges within groups

1 and 2 for a random graph with the same node degree distribution as the given network.

2.2.2 Random Network Models

Here we provide an intuition of the usefulness of the random network models models and we show how we calculate their likelihoods.

Erdös-Rényi

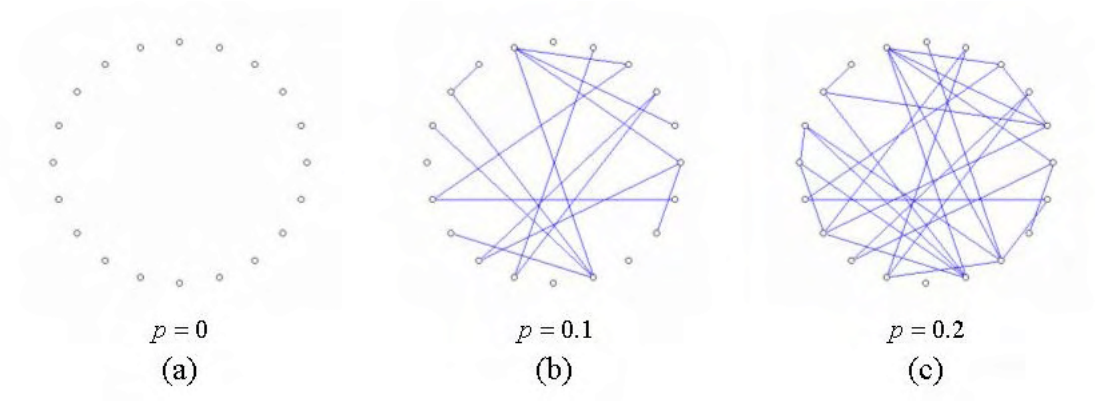

Figure 2.11: Erdös-Rényi Network.

For the model of $Erdös - Rényi$, [31, 32, 36] , all graphs on a fixed vertex set with a fixed number of edges are equally likely; in the model introduced by Gilbert, each edge has a fixed probability of being present or absent, independently of the other edges. All graphs with n nodes and M edges have equal probability of:

$$p^M(1-p)^{\binom{n}{2}-M}. \tag{2.2}$$

In the model $G(n,p)$, a graph is constructed by connecting nodes randomly. Each edge is included in the graph with probability p independent from every other edge. The expected degree of $G(n,p) = (n-1)p$. The parameter p in this model can be thought of as a weighting function; as p increases from 0 to 1, the

model becomes more and more likely to include graphs with more edges and less and less likely to include graphs with fewer edges. In particular, the case $p = 0.5$ corresponds to the case where all $2^{\binom{n}{2}}$ graphs on n vertices are chosen with equal probability.

Barabási Albert

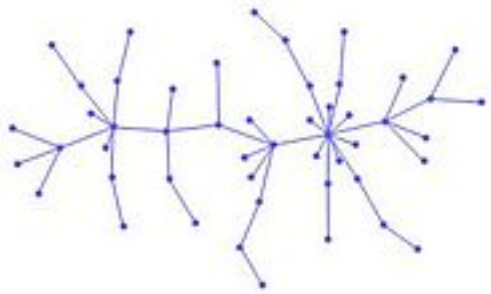

Figure 2.12: Barabási Albert Network.

The $Barabási - Albert$ (BA) model [13] is an algorithm for generating random scale-free networks using a preferential attachment mechanism. Several natural and human-made systems, including the Internet, the world wide web, citation networks, and some social networks are thought to be approximately scale-free and certainly contain few nodes (called hubs) with unusually high degree as compared to the other nodes of the network.

The network begins with an initial connected network of m_0 nodes. New nodes are added to the network one at a time. Each new node is connected to $m \leq m_0$ existing nodes with a probability that is proportional to the number of links that the existing nodes already have. Formally, the probability p_i that the new node is connected to node i is $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is the degree of node i and the sum is made over all pre-existing nodes j (i.e. the denominator results in twice the current number of edges in the network). Heavily linked nodes ("hubs") tend to quickly accumulate even more links, while nodes with only a few links are unlikely to be chosen as the destination for a new link. The new nodes have a "preference" to attach themselves to the already heavily linked nodes. This algo-

rithm computationally provides as with the likelihood of the model.

Watts-Strogatz

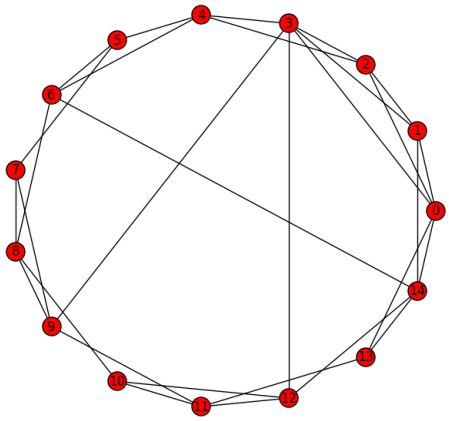

Figure 2.13: Watts-Strogatz Network.

The $Watts - Strogatz$ model [111] is a random graph generation model that produces graphs with small-world properties, including short average path lengths and high clustering.

Given the desired number of nodes N, the mean degree K (assumed to be an even integer), and a special parameter β , satisfying $0 \leq \beta \leq 1$ and $N \gg K \gg \ln N \gg 1$, the model constructs an undirected graph with N nodes and $\frac{NK}{2}$ edges in the following way:

Construct a regular ring lattice, a graph with N nodes each connected to K neighbors, $\frac{K}{2}$ on each side. That is, if the nodes are labeled $n_0 \ldots n_{N-1}$, there is an edge (n_i, n_j) if and only if $0 < |i-j| \bmod \left(N - 1 - \frac{K}{2}\right) \leq \frac{K}{2}$. For every node $n_i = n_0, \ldots, n_{N-1}$ take every edge connecting n_i to its $K/2$ rightmost neighbors, that is every edge $(n_i, n_j \bmod N)$ with $n_i < n_j \leq n_i + K/2$, and rewire it with probability β . Rewiring is done by replacing $(n_i, n_j \bmod N)$ with (n_i, n_k) where

k is chosen uniformly at random from all possible nodes while avoiding self-loops ($k \neq i$) and link duplication (there is no edge $(n_i, n_{k'})$ with $k' = k$ at this point in the algorithm). This algorithm computationally provides as with the likelihood model.

Stochastic Block Model

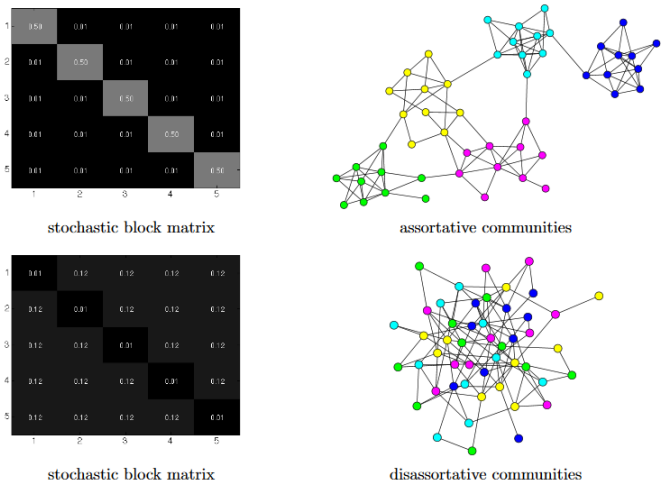

Figure 2.14: Stochastic Block Model

The *stochastic block model* is a generative model for random graphs. This model tends to produce graphs containing communities, subsets characterized by being connected with one another with particular edge densities. For example, edges may be more common within communities than between communities.

A SBM with K blocks on n nodes is defined as follows. A vector of latent labels $C = (C_1, \ldots, C_n)$ is generated with C_i taking integer values from $[K] = \{1, \ldots, K\}$ governed by a multinomial distribution with parameters $P = (\pi_1, \pi_2, \ldots, \pi_K)$. Given $C_i = \alpha$, $C_j = b$, an adjacency matrix A is generated with:

$$A_{i,j} \mid (C_i = \alpha, C_j = b) \sim Bernoulli(P_{\alpha,b}), i \neq j. \tag{2.3}$$

We consider a symmetric A with zero diagonal entries corresponding to an undirected graph, although our arguments generalize easily to directed graphs. Let P

is a $K \times K$ symmetric matrix describing the connectivities within and between blocks. We denote the model parameters $\theta = (\pi, P)$ and let Θ_K be the parameter space of a K-block model,

$$\Theta_K = \{\theta \mid \pi \in (0,1)^K, \sum_{\alpha=1}^{K} \pi_\alpha = 1, P \in (0,1)^{K \times K}\} \tag{2.4}$$

We assume $\theta^* \in \Theta_K$ and P^* has no identical columns, meaning the underlying model has K blocks and it is identifiable in the sense that it cannot be further collapsed to a smaller model. $c = (c_1, \times, c_n) \in [K^{'}]^n$ represents another set of labels under a $K^{'}$-block model with $K^{'}$ not necessarily equaling K. $g(A;\theta)$ is the likelihood function describing the distribution of A with parameter $\theta \in \Theta_{K^{'}}$ and can be written as the sum of the complete likelihood function $f(c, A; \theta)$ associated with the labels $c \in [K^{'}]^n$:

$$g(A;\theta) = \sum_{c \in [K^{'}]^n} f(c, A; \theta), \tag{2.5}$$

where $f(c, A; \theta)$ takes the form:

$$f(c, A; \theta) = (\prod_{\alpha=1}^{K^{'}} \pi_\alpha^{n_\alpha(c)})(\prod_{\alpha=1}^{K^{'}} \prod_{b=1}^{K^{'}} P_{\alpha,b}^{O_{\alpha,b}(c)} (1 - P_{\alpha,b})^{n_{a,b}(c) - O_{a,b}(c)})^{1/2} \tag{2.6}$$

with count statistics:

$$n_\alpha(c) = \sum_{i=1}^{n} \mathbb{1}(c_i == \alpha), \tag{2.7}$$

$$n_{\alpha,b}(c) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}(c_i == \alpha, c_j == b), \tag{2.8}$$

$$O_{a,b}(c) = \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{1}(c_i == \alpha, c_j == b) A_{i,j}, \tag{2.9}$$

g and f are invariant with respect to a permutation on the block labels, $\tau : [K^{'}] \to [K^{'}]$, and its corresponding permutations on the node labels c and the parameters θ.

Latent Space models

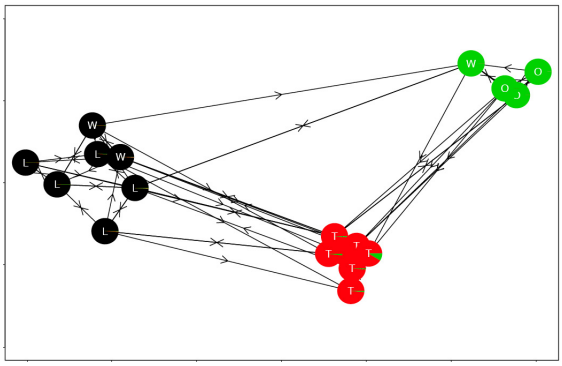

Figure 2.15: Latent Space Model

In *latent variable models* [46] the entries of the adjacency matrix are dependent on a set of unobserved or latent variables. Observed variables assumed to be conditionally independent given latent variables. Adjacency matrix A is invariant to row and column permutations. Although, Aldous-Hoover theorem implies existence of a latent variable model of form $A_{ij} = h(\theta, z_i, z_j, \epsilon_{ij})$ for iid latent variables z_i and some function h, latent variable models allow for both for homogeneity (most nodes or nodes among clusters have the same number of connections) and heterogeneity (most nodes have not the same number of connections) of nodes in social networks. This mean. Each node (actor) has a latent variable z_i. Probability of forming edge between two nodes is independent of all other node pairs given values of latent variables.

$$p(A \mid Z, \theta) = \prod_{\neq} p(A_{ij} \mid z_i, z_j, \theta) \tag{2.10}$$

The log likelihood of α, β and the z_is for the latent space model is as follows:

$$\eta_{\mathrm{i,j}} = \mathrm{logodds}(A_{\mathrm{i,j}} = 1 \mid z_{\mathrm{i}}, z_{\mathrm{j}}, x_{\mathrm{i,j}}, \alpha, \beta) = \alpha + \beta' x_{\mathrm{i,j}} - \mid z_{\mathrm{i}} - z_{\mathrm{j}} \mid .$$

where $\eta_{i,j} = \alpha + \beta^{'} x_{i,j} - \mid z_i - z_j \mid$.

Ideally latent variables should provide an interpretable representation (continuous) latent space model. The motivation behind them are homophily or assortative mixing. Probability of edge between two nodes increases as characteristics of the nodes become more similar. They represent nodes in an unobserved (latent) space of characteristics or social space. Small distance between 2 nodes in latent space means high probability of edge between nodes. Furthermore they induce transitivity: observation of edges (i, j) and (j, k) suggests that i and k are not too far apart in latent space which is more likely to also have an edge. The (continuous) latent space model (LSM) were proposed by [37]. Specifically, each node has a latent position $z_i \in \mathbb{R}^d$, the Probabilities of forming edges depend on distances between latent positions and they define pairwise affinities $\psi_{i,j} = \theta - \mid\mid z_i - z_j \mid\mid_2$. The practitioner sample node positions in latent space $z_i \sim$ Gaussian(0,kI) compute affinities between all pairs of nodes $\psi_{i,j} = \theta - \mid\mid z_i - z_j \mid\mid_2$ and sample edges between all pairs of nodes $P(A_{ij} = 1 \mid \psi_{ij}) = \sigma(\psi_{ij})$.

Advantages of latent space model are that they are visual and interpretable spatial representation of networks and models homophily (assortative mixing) well via transitivity. The disadvantages of latent space model include the following: the statistician has to fix the dimension beforehand without knowing a priori what the actual dimension is. Moreover, they can not model disassortative mixing (people preferring to associate with people with different characteristics).

Chapter 3

Comparing Sampling Designs on Random Networks via Information Theory

In this chapter, we propose a general approach for comparing sampling designs on networks. Our approach is based on the concept of data compression from information theory. The criterion for comparing sampling designs is formulated so that the results prove to be robust with respect to some of the most widely used loss functions for point estimation and prediction. The rationale behind the proposed approach is to find sampling designs such that preserve the largest amount of information possible from the original data generating mechanism. Our approach is inspired by the same principle as the reference prior, with the difference that, for the proposed approach, the argument of the optimization is the sampling design rather than the prior. The information contained in the data generating mechanism can be encoded in a distribution defined either in parameter's space (posterior distribution) or in the space of observables (predictive distribution). In our simulation studies we consider both cases.

For applications involving network data, such as epidemiology, it is often the case that practitioners can only observe the network partially via a sampling design. Examples in Epidemiology include case studies, when practitioners recruit individuals from a hard-to-reach population with the aim to infer the prevalence of HIV in that population. Respondent-Driven Sampling (RDS) is a design that takes advantage of the social network structure to solve this problem. Under the assumption that the random graph (statistical network) model adopted by the statis-

tician is reasonable for the application at hand, consider the scenario where there are multiple sampling designs that could be adopted: RDS with different tunning parameters, Snowball sampling with different allocation schemes for the seeds. To determine which option of data collection is the most suitable, we could resort to a decision theory approach [63]. Another, low level intuitive example involves the following setting: A statistician wants to send a network data set which is an output of an MCMC. This MCMC samples large network data e.g. graphical models (probabilistic models for which a graph expresses the conditional dependence structure between random variables). Do the statistician have to store all of them? For memory issues he/she can compress them and the price to pay is the information he lose. Which one is the best way to store those messages, in order to maximize the amount of information he will send? In this chapter we propose an alternative: to compare sampling designs on networks in terms of the amount of information preserved. The motivation behind this set of ideas is to provide a principled procedure for ranking the set of possible designs that proves to be robust to different choices for the loss function.

The approach we propose is based on two information theory concepts: data compression and decompression. More precisely: sampling is modeled as a process that compresses information regarding the probabilistic model that generates the full network, while computing the posterior is modeled as a process that decompresses that information. The procedure for comparing sampling designs on networks proposed in this chapter is based on the previous idea and on the rationale behind the computation of a reference prior. The core idea can be phrased as follows: instead of performing optimization with respect the prior, which would ensure a maximum gain of information, we optimize with respect to the sampling mechanism, which would ensure a maximum amount of preserved information. This construction implies a distance or a divergence between probability distributions. Thus, information theory enable us to compare, in a principled way, sampling designs on random network models.

The literature of sampling designs has evolved from discussing sampling designs on networks as algorithms ([44] and [107]) to approaches where: i) some sources of uncertainty are modelled explicitly ([38], [12]) and ii) likelihood and fully Bayesian approaches ([63]). The literature dealing with the comparison of sampling mechanisms includes simulation studies based on heuristic arguments ([21]) and the approach by [14], which is based on ideas from Bayesian experimental design [61] and [26]. Our work borrows ideas from the computation of the

reference prior [18] and the concept of data compression, both of which rely on information theory.

The main contribution of our chapter can be phrased as follows: Current approaches involve evaluating different sampling designs with respect to a loss function ([61] and [26]). Whilst this can work well, the choice of sampling design is sensitive to the choice of loss function. Here, we propose methodology for ranking sampling mechanisms on networks such that the top designs in the ranking tend to produce posteriors that preserve more information about the data generating mechanism. The ranking of sampling designs is obtained without the need to specify a loss function . We make the case that the rankings of sampling designs implied by our approach are reasonably consistent with rankings implied by the most widely used loss functions for estimation and prediction.

The chapter proceeds as follows: In Section 3.1, we describe settings of the problem, we formulate them and give notation and definitions of networks, random networks and sampling designs. Furthermore, we present information theory tools that are useful in the next sections. Then, in section 3.2, we focus in our main purpose of this chapter which is how we use and compare sampling mechanisms using information theory in order to compress and decompress random networks. Conceptually and computationally, our methodology is presented. In section 3.3, for many random network models data analysis that gives experimental results involving compression, decompression and model misspecification is conducted showing the results of our approach. Finally, in section 3.4, we present with more details the future work involving overlapping research areas.

3.1 Preliminaries

3.1.1 Random graphs

We define a network as a pair $\mathcal{G}(\omega) = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes the set of nodes, and $\mathcal{E}$ the set of edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let N denote $|\mathcal{V}|$, the number of nodes and let $e_{i,j}$ denote the element of $\mathcal{E}$ connecting nodes i and j, $1 \leq i, j \leq N$. A network is called simple if at most one edge exists between each pair of nodes and no self-loops are allowed. We denote by $A_{\mathcal{G}(\omega)}$ the adjacency matrix of $\mathcal{G}(\omega)$; for simple

graphs, the entries of this matrix can be defined as follows:

$$A_{\mathcal{G}(\omega)}(i,j) = \begin{cases} 1 & \text{if } e_{i,j} \in \mathcal{E}, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i, j \leq N$. A network is called undirected if the corresponding adjacency matrix is symmetric.

A random network (or random graph model) is a probability model on the space of adjacency matrices. In this chapter we consider random network models in the space of simple undirected networks, *i.e.*, a distribution on the space of binary symmetric adjacency matrices. We use $\mathcal{G}(\tau(\mathcal{G}))$ to denote a specific feature of the random network, with $\tau(\cdot)$ the process of extracting a specific feature from a network (*e.g.* degree distribution, modularity).

The random graph models we will use to illustrate our method include, as already described in chapter 2:

- The Stochastic Block model, where the partition of the nodes set $\{1, \ldots, N\}$ into disjoint subsets $C_1, \ldots, C_{\mathrm{b}}$ is called communities and for the symmetric $b \times b$ matrix P of edge probabilities we have:

$$\Pr\{A_{\mathcal{G}}(i,j) = 1 \mid i \in C_{\mathrm{u}}, j \in C_{\mathrm{v}}\} = P_{\mathrm{u,v}}, \quad P_{\mathrm{u,v}} \in (0,1)$$

- The Latent Space model with Euclidean distance, [110], which is a particular case of latent position models where each node has an associated latent position. Nodes with nearby latent positions are likely to form ties. The parameterization of $P(A_{\mathrm{i,j}} \mid z_{\mathrm{i}}, z_{\mathrm{j}}, x_{\mathrm{i,j}}, \theta)$ is the logistic regression model in which the probability of a tie depends on the Euclidean distance between z_i and z_j, as well as on observed covariates $x_{i,j}$ that measure characteristics of the dyad,

$$\eta_{\mathrm{i,j}} = \mathrm{logodds}(A_{\mathrm{i,j}} = 1 \mid z_{\mathrm{i}}, z_{\mathrm{j}}, x_{\mathrm{i,j}}, \alpha, \beta) = \alpha + \beta' x_{\mathrm{i,j}} - \mid z_{\mathrm{i}} - z_{\mathrm{j}} \mid .$$

3.1.2 Sampling Designs

Sampling Designs on Networks

Here, we introduce sampling designs on one realization network, explain why they are useful and provide some examples. In sampling, we are typically interested in using field point data to derive inferences, we need enough samples to be

confident that they approximate the target population. In the case of calibrating a laboratory device, we might only need two measurements, each at opposite ends of the measurement scale. This illustrates the point that sample size is closely related to the inherent variability in the data. The number of samples required increases with increasing variability. Also, the more samples we have, the greater the confidence level we can achieve. For example, sampling at an 85 percent confidence level is less intensive than sampling at a 95 percent confidence level.

In this section, we consider the conceptual and computational theory of network sampling. There is a substantial literature on network sampling designs. Our development here follows [101], [102], [103] and [2]. Let denote a network with n nodes. Note that in most network samples, the unit of sampling is the node, while the unit of analysis is typically the dyad. Let $\mathcal{G}(\omega)$ be the $n \times n$ binary matrix indicating if the corresponding element of the adjacency matrix was sampled or not. The value of the i, jth element is 0 if the (i, j) ordered pair was not sampled and 1 if the element was sampled.

Under many sampling designs the set of sampled dyads is determined by the set of sampled nodes. The sampled network a binary n-vector indicating a subset of the nodes, where the ith element is 1 if the ith node is part of the set, and is 0 otherwise. For example, consider an undirected network where the set of observed dyads are those that are incident on at least one of the sampled nodes. A primary example of this is where people are sampled and surveyed to determine all their edges.

Computational complexity makes network analysis task difficult for very large graphs. By network analysis task here we refer to:

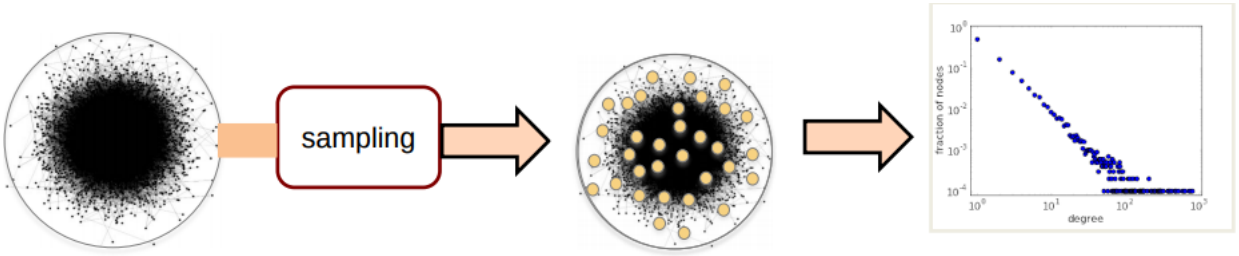

Figure 3.1: Estimation of network characteristics by sampling vertices (or edges) from the original networks.

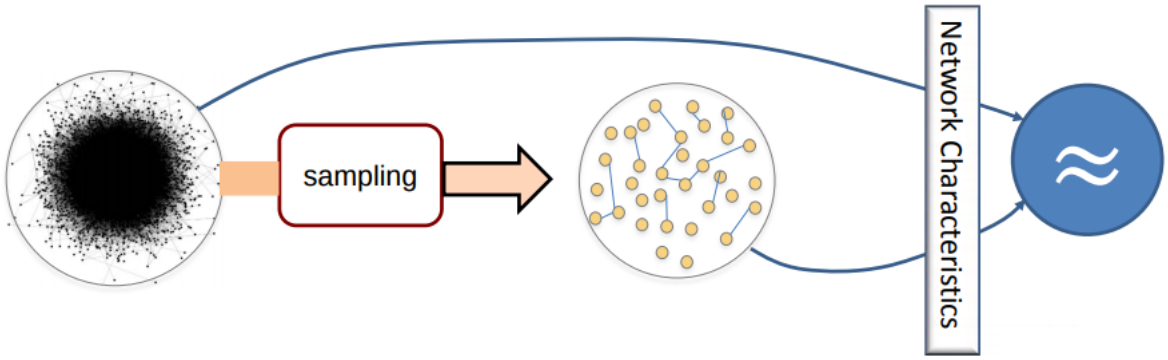

Figure 3.2: Study of the connectivity structure of networks and investigation of the behavior of processes overlaid on the networks.

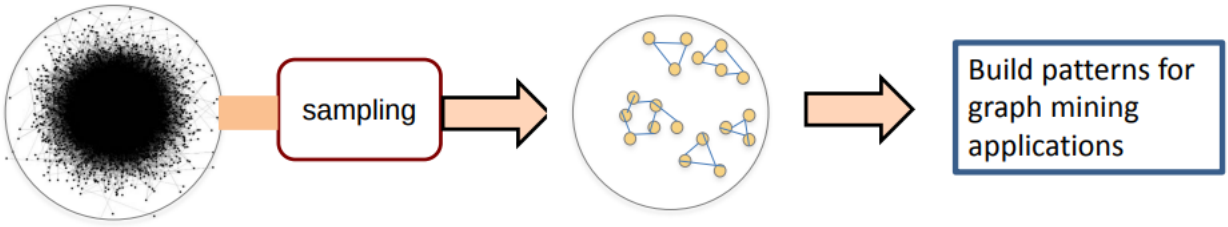

Figure 3.3: Study of local topologies and their distributions to understand local phenomenon.

- The study of the node/edge properties in networks e.g., Investigate the correlation between attributes and local structure, estimate node activity to model network evolution and predict future links (what is the probability that u and v will be connected in future?) and identify hidden links. For this task, statisticians estimate network characteristics by sampling vertices (or edges) from the original networks, population is the entire vertex set (for vertex sampling) and the entire edge set (for edge sampling) and sampling is usually with replacement (figure 3.1).

- The study of the connectivity structure of networks and investigate the behavior of processes overlaid on the networks e.g., estimate centrality and distance measures in communication and citation networks, identify communities in social networks and study robustness of physical networks against attacks. For this task, from network we sample a subgraph with k nodes which preserves the value of key network characteristics of the network, such as degree distribution, diameter, centrality, and community structure through modularity. Note that here, the sampled network is smaller, so there is a scaling effect on some of the statistics; for instance, average degree of the sampled network is smaller and statisticians consider the population of all subgraph of size k (figure 3.2).

- The study of local topologies and their distributions to understand local phenomenon e.g., discovering network motifs in biological networks and counting triangles to detect Web (i.e., hyper link) spams. Now, statisticians sample sub-structure of interest and find frequent induced subgraph (network motif) and sample sub-structure for solving other tasks, such as counting, modeling, and making inferences (figure 3.3).

Best time complexities for various tasks are vertex count $O(n)$, edge count $O(m)$, centrality metrics $O(mn)$, Community Detection using Girvan-Newman Algorithm $O(m^2n)$, triangle (motif) counting $O(m^{1.41})$ etc.

More specifically, with sampling on networks we can sample a set of vertices (or edges) and estimate nodal or edge properties of the original network e.g., average degree and degree distribution. Instead of analyzing the whole network, we can sample a small subnetwork similar to the original network. The goal here is to maintain global structural characteristics as much as possible e.g., degree distribution, clustering coefficient, community structure though modularity etc. Finally we can also sample local substructures from the networks to estimate their relative frequencies or counts e.g., sampling triangles, or network motifs. Types of sampling designs on networks include: snowball sampling, stratified sampling, Breadth-First Search (BFS), or Depth-First Search (DFS), Forest Fire (FF), Random walk techniques (exploration with replacement) and Respodent Driven Sampling. In this thesis, we distinguish sampling designs in ignorable and non-ignorable (appendix) and use as an example of ignorable sampling designs snowball sampling and RDS for non-ignorable sampling designs.

We introduce further notation to the next section where we generalize what is considered for a network realization to a distribution or networks.

Sampling designs on Random Networks

In contrast with before, the variability is in the network. We are dealing with a population of networks which are generated by a random network model or a random network mechanism. Let $\mathcal{G}$ denote a random network and let I denote a sampling design that propagates through the network. A realization of $I(\mathcal{G})(\omega)$ implies a partition of the network realization $\mathcal{G}(\omega)$ into $\mathcal{G}_{\mathcal{INC}}(\omega)$ and $\mathcal{G}_{\mathcal{EXC}}(\omega)$, which denote, respectively, the observed and unobserved parts of the random network realization. To keep the notation simple, we write $\mathcal{G}_{\mathcal{INC}}$ and $\mathcal{G}_{\mathcal{EXC}}$ when referring to all the realizations; for describing computations and presenting definitions, this is all that is needed. In analogous manner, all the realizations of the adjacency matrices $A_{\mathcal{G}}$ can be written as $A_{\mathcal{G}_{\mathcal{INC}}}$ and its completion $A_{\mathcal{G}_{\mathcal{EXC}}}$.

In order to perform computations, we need to define $\mathcal{G}_{\mathcal{INC}}$ and $\mathcal{G}_{\mathcal{EXC}}$ with more detail: $\mathcal{G}_{\mathcal{INC}}$ is given by a set of n nodes $\{1, 2, \ldots, n\}$ and a set of edges and non-edges between these n nodes. For N specified, the maximum allowed size for an

imputed network, $\mathcal{G}_{\mathcal{EXC}}$ is given by a set of nodes $\{n+1, n+2, \ldots, N\}$, where $n \leq N$ and a set of edges for an adjacency matrix $A_{\mathcal{G}}$ with N nodes, such that: (i) the edges incident to at least one node in $\{n+1, n+2, \ldots, N\}$, are not included in $\mathcal{G}_{\mathcal{INC}}$, (ii) each of the connected components of the network $\mathcal{G}$ obtained by adding the edges of $\mathcal{G}_{\mathcal{EXC}}$ to $\mathcal{G}_{\mathcal{INC}}$ has a non-empty intersection with $\{1, 2, \ldots, n\}$. We will use the notation $\mathcal{G}_{\mathcal{EXC}} \sim \mathcal{G}_{\mathcal{INC}}$ to denote the fact that a specific $\mathcal{G}_{\mathcal{EXC}}$ serves to complete a specific $\mathcal{G}_{\mathcal{INC}}$ to form an adjacency matrix.

To perform Bayesian inference, it is necessary to specify the likelihood correctly. The concept of ignorability [87] helps on this task by providing criteria for deciding if the uncertainty due to the sampling mechanism needs to be modeled explicitly in the likelihood (Appendix). In this chapter, we use one ignorable sampling design, called snowball and one non-ignorable sampling design, called Respondent Driven Design (RDS).

More specifically, in snowball design we follow the algorithm 1:

Algorithm 1 Snowball Design

- Select l individuals (seeds) at random.
- Observe all dyads involving the selected individuals.
- Identify r individuals (referrals) reported to have at least one relation with the initial sample
- Observe all dyads involving the newly selected individuals.
- Repeat the last three steps either k times (waves) or until all recruited nodes are collected.

and RDS design algorithm 2:

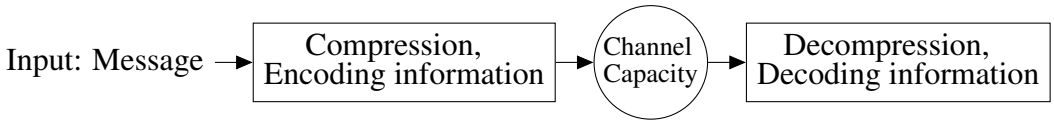

Figure 3.4: This is an schematic description of the data compression process (source [68]).

Algorithm 2 RDS Design

- Select l individuals (seeds) at random.
- Observe all dyads involving the selected individuals.
- Identify r individuals (referrals) reported to have at least one relation with the initial sample.
- Observe all dyads involving the newly selected individuals.
- Repeat the last three steps k times (waves) or until all recruited nodes are collected..
- Observe the only the degrees, both from observed and unobserved part of the network, of each observed node only for the last wave.

3.1.3 Information Theory

The problem of data compression [68] can be stated as follows: A message is generated by a probability distribution $p(\cdot)$, defined over a sample space $\mathcal{X}$, and, to be transmitted, it is mapped into a space of lower dimension (traditionally $\{0, 1\}$). After the message has been transmitted, it is mapped back to the original space. This process may entail a loss of information. The action of mapping the message to a low dimensional space is called *compression*, while the action of mapping the message back to the original space is known as *decompression*. This process is illustrated in Figure 3.4. Usually, the problem of data compression involves a loss of information during transmission; this loss is quantified in what is known as the *channel capacity*. In this chapter we will assume that any loss of information occurs during compression and none during transmission.

To measure the loss of information incurred during data compression, it is necessary to have either a divergence or a metric between the distribution that

generated the message $p(\cdot)$ an the distribution implied by the decompression. In this chapter, the Hellinger distance will play that role. The Hellinger distance belongs to the family of f-divergence.

3.2 Methodology

3.2.1 General Concepts

In this section, we present methodology for comparing sampling designs on networks. Given a set of potential sampling designs $\{I_1, I_2, \ldots, I_{\mathrm{k}}\}$ to be applied to a random graph model, we want to produce a ranking of these designs. The ranking should be such that designs with higher positions in the ranking tend to lead to posteriors that preserve more information about the probabilistic mechanism that generates the data. The criterion to make the comparison is based on the following rationale: the process of applying a sampling mechanism and inferring the random graph model via computing a posterior distribution, can be cast as a problem of data compression. The optimal design is the one that minimizes the loss of information. This is under the assumption that the full network is a realization of a random graph model correctly specified.

Let θ the vector parameter of random network model $p(\mathcal{G} \mid \theta)$. In order to illustrate the main ideas, we first assume that θ is specified. The next step is to apply the sampling design I to $\mathcal{G}(\tau(\mathcal{G}))$ to get the compressed network realization $\mathcal{G}_{\mathcal{INC}}(\tau(\mathcal{G}))$ which implies posterior distribution $p(\theta \mid \mathcal{G}_{\mathcal{INC}}(\tau(\mathcal{G})))$. This process is illustrated in Figure 3.5. First, the network model $p(\mathcal{G} \mid \theta)$ generates $\mathcal{G}$ and then the sampling mechanism is applied to produce a compressed version of the full network; this process is denoted by $g(\mathcal{G}_{\mathcal{INC}} \mid I, \mathcal{G})$. Given the observed (compressed) network, the posterior distribution $p(\theta \mid \mathcal{G}_{\mathcal{INC}})$ and the posterior predictive $p(\tau(\mathcal{G}) \mid \mathcal{G}_{\mathcal{INC}})$ can be computed. The next step is to compute the Hellinger distance between the prior predictive and the posterior predictive implied by $\mathcal{G}_{\mathcal{INC}}(\tau(\mathcal{G}))$, *i.e.*,

$$H(p(\tau(\mathcal{G}) \mid \theta), p(\tau(\mathcal{G}) \mid \mathcal{G}_{\text{INC}})).$$

By averaging over the uncertainty of $p(\mathcal{G} \mid \theta)$, we obtain the average of those

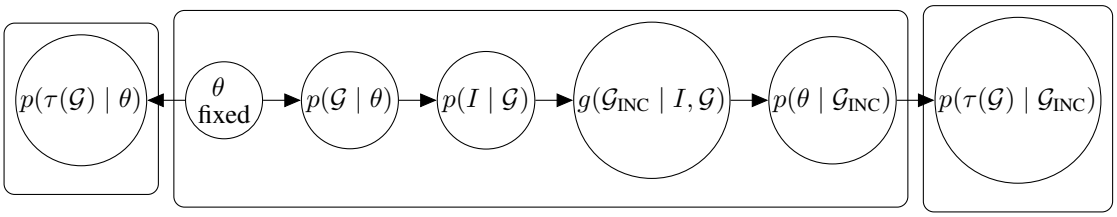

Figure 3.5: Illustration of how to cast the process of performing Bayesian inference from a partially observed network as a data compression process assuming θ is fixed but unknown.

distances:

$$\psi_{\bigstar}(I, \theta, \tau(\cdot)) = \sum_{\mathcal{G}_{\mathcal{INC}}} \sum_{\mathcal{G}_{\mathcal{EXC}} \sim \mathcal{G}_{\mathcal{INC}}} H(p(\tau(\mathcal{G}_{\mathcal{INC}}), \mathcal{G}_{\mathcal{EXC}} \mid \theta), p(\tau(\mathcal{G}_{\mathcal{INC}}), \mathcal{G}_{\mathcal{EXC}} \mid \mathcal{G}_{\mathcal{INC}})). \tag{3.1}$$

This will be the score associated to the amount of information preserved by the design.

Now, in Figure 3.5 we proceed to explain how to compare $\{I_1, I_2, \ldots, I_{\text{k}}\}$. The distribution specified by θ entails a predictive distribution $p(\tau(\mathcal{G}) \mid \theta)$. The model for the full network $p(\mathcal{G} \mid \theta)$ in conjunction with the sampling mechanism that propagates through the network $p(I \mid \mathcal{G})$ produce the observed part of the network $\mathcal{G}_{\mathcal{INC}}$ through a deterministic mapping $h(\cdot; \cdot)$. The observed data serves as input for the procedure that computes the posterior $p(\theta \mid \mathcal{G}_{\mathcal{INC}})$; this posterior entails a posterior predictive distribution $p(\tau(\mathcal{G}) \mid \mathcal{G}_{\mathcal{INC}})$. In the same spirit as in Bayesian experimental design approaches, we incorporate uncertainty on θ by assuming that it was sampled from the prior $p(\cdot)$. The uncertainty induced by $p(\cdot)$ propagates trough the process described in Figure 3.5. All the uncertainty is in the prior, assuming the prior is specified in that way that is capturing where the true value of the parameter is living, so when we use the Bayesian experimental design we get a reasonable answer. Then we proceed following decision theory tools. To compute the score under this setup, we need to average the Hellinger distances

$$\psi(I, p(\cdot), \tau(\cdot)) = \int_{\theta \in \Theta} \psi_{\bigstar}(I, \theta, \tau(\cdot)) p(\theta) d\theta \tag{3.2}$$

As in section 2.2, $\tau(\cdot)$ denotes the operation of extracting a specific feature from a network (*e.g.* degree distribution, modularity) and $p(\tau(\mathcal{G}))$ is the distribu-

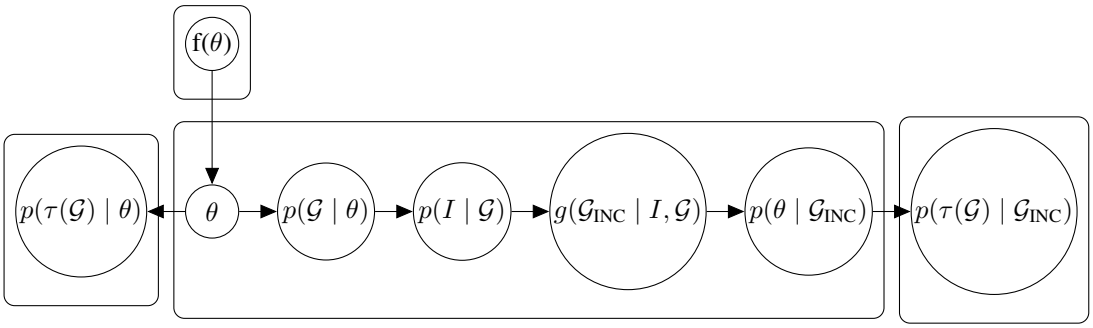

Figure 3.6: Illustration of how to cast the process of performing Bayesian inference from a partially observed network as a data compression process when θ is not specified in the space of observables.

tion for the feature implied by the prior predictive. This process is described in Figure 3.6.

The previous strategy enables the statistician to compare designs in terms of distances between predictive distributions. If the statistician requires to conduct the comparison of sampling designs in the space of parameters, then an interesting challenge arises: we still need to introduce a probabilistic distribution for θ, since its value is unknown. This distribution cannot be the prior, since it is the distribution for which information should be preserved and it will be unknown to the practitioner when fitting the random graph model. We view the prior distribution as a tool for decompressing information. Simply, if the prior is misspecified, e.g. the statistician setup a prior which completely misses out where the true value is located then the results are misleading. To protect himself from that case that is when a distribution of the statistician prior belief's is considered which is different from the distribution where the parameters live. This process is illustrated in Figure 3.7. The only difference with respect to the process described in Figure 3.5 is that θ is a realization from a random variable with distribution given by $f(\cdot)$. Moreover, in figure 3.7 the main difference with respect to the process described in figure 3.6 is that the distributions to be compared are defined on parameter space, rather than on the space of observables. Using features instead of full networks obtained from the predictive will make the computation of the scores feasible.

The idea of using such distributions for the prior and the idea of comparing designs based on the predictive distribution of a feature can be combined, so the

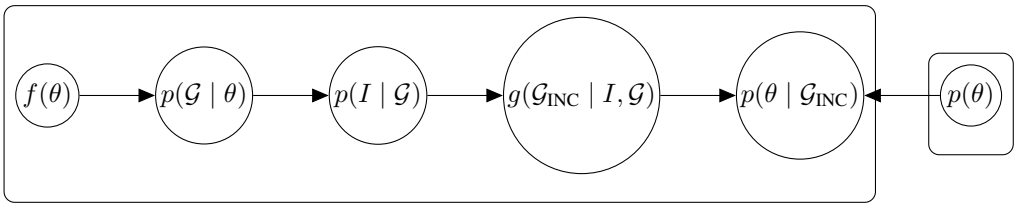

Figure 3.7: Illustration of how to cast the process of performing Bayesian inference from a partially observed network as a data compression process when θ is not specified in the space of parameters.

score is computed in terms of the Hellinger distance between predictive distributions, *i.e.*,

$$H(p(\tau(\mathcal{G})), p(\tau(\mathcal{G}) \mid \mathcal{G}_{\mathcal{INC}})),$$

where $\tau(\cdot)$ denotes the operation of extracting a specific feature from a network (*e.g.* degree distribution, modularity) and $p(\tau(\mathcal{G}))$ is the predictive distribution for the feature implied by this prior distributions.

3.2.2 Bayesian Computation

We obtain samples of the posterior distribution $p(\theta, \mathcal{G}_{\mathcal{EXC}} \mid \mathcal{G}_{\mathcal{INC}})$ via a Gibbs sampler scheme based on the full conditionals

$$p(\theta \mid \mathcal{G}_{\mathcal{INC}}, \mathcal{G}_{\mathcal{EXC}}) \qquad \text{and} \qquad p(\mathcal{G}_{\mathcal{EXC}} \mid \mathcal{G}_{\mathcal{INC}}, \theta).$$

Once enough posterior samples $\theta^{(i)}$ from the posterior have been obtained, we compute either

$$H(f(\theta), p(\theta \mid \mathcal{G}_{\mathcal{INC}})),$$

or

$$H(p(\tau(\mathcal{G})), p(\tau(\mathcal{G}) \mid \mathcal{G}_{\mathcal{INC}})),$$

for which we need to perform the additional step of sampling from the predictive distribution.

3.2.3 Consistency of F-divergence with Decision theory

F-divergences are a general class of divergences (indexed by convex functions f) that include the KL divergence and Hellinger distance as special cases. The

f-divergence between two probability distributions P and Q is characterized by non-negativity, monotonicity and joint convexity. So, by considering the class of all f-divergences every divergence behave in the same monotonic way and each divergence gives the same ranking.

3.3 Simulation Studies

3.3.1 Simulation Set Up

To investigate the behavior of the proposed approach, we conducted a simulation study. The objective of the simulation is to assess if the ranking of the sampling designs implied by our approach is compatible with the rankings implied by an array of widely used loss functions (Table 3.3). The regimes of the simulation study are given by: the random graph model, the corresponding vector of parameters (Table 3.1), the functional form of the sampling design, the tuning parameters (Table 3.2) and the sample size. To infer the stochastic block model parameters we use [57] approach, not using gibbs sampling since we need Reversible Jump MCMC, by using the same notation for number of blocks (K), and inclusion probabilities (λ, ϵ). To infer the latent space model parameters, using logit, we use a bivariate normal distribution for the covariance matrix for the latent positions with mean $(0, 0)$ and a covariance matrix with dependencies. Moreover, the second defined prior distributions in subsection 3.2.1. used here for intercept and regression coefficient are both one dimensional normal distributions. The graph features $\tau(\mathcal{G})$ we considered were: the number of communities of SBM and the regression coefficient α from LSM. When we implement our method we consider 1000 samples of the desired parameters from the desirable prior distributions and 100 networks given the value of the parameters.

As discussed in Section 3.2, our approach associates a score to each design. Such score is given by the mean of the Hellinger distances between the desirable prior distributions and the posterior. The mean is computed with respect to the ground truth distribution.

3.3.2 Empirical Results

We explore the performance of our method when the loss of information is

Random Graph Model	Parameter Specification	Features
SBM	K=10, $\lambda = 0.9$, $\epsilon = 0.1$, $N = 100$	K
LSM	$\alpha = 0.5$, $\beta = 1$, Cov. matrix=[(0,0), (1,1), 0.3] and $N = 100$	α

Table 3.1: Random graph models, parameter vectors and graph features considered for setting up simulation regimes.

Sampling Design	Tuning Parameters
Snowball	(l,r,k)= $(2, 2, 2)$, $(3, 3, 2)$, $(3, 2, 3)$
Respondent-Driven Sampling	(l,r,k)= $(2, 2, 2)$, $(3, 3, 2)$, $(3, 2, 3)$

Table 3.2: Sampling designs on networks and tuning parameters considered for setting up simulation regimes.

Loss Function	Expression	Type of Inference
Hellinger Distance	$H(p(\tau(\mathcal{G})), p(\tau(\mathcal{G}) \mid \mathcal{G}_{\mathcal{INC}}))$	Point Prediction
Quadratic Loss	$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$	Point Estimation and Prediction
Absolute Loss	$L(\theta, \hat{\theta}) = \lvert\theta - \hat{\theta}\rvert$	Point Estimation and Prediction

Table 3.3: Loss functions considered to compute rankings of sampling designs based on expected loss.

measured, both in the space of parameters and the space of graph features $\tau(\mathcal{G})$. For computing the posterior $p(\theta \mid \mathcal{G}_{INC})$, we consider the two possibilities outlined in Section 3.2 (Figures 3.6 and 3.7). All posterior samples for random network parameters were obtained via MCMC with a burn-in of 5000 samples and 1000 posterior samples.

We present the results in table 3.4 and figures 3.8 and 3.9 in terms of Hellinger distances distributions means [18], and the expected loss ([7]). Both simulation studies present empirical evidence why our approach could resembles decision theory methods ([61] and [26]) and generally the type of f-divergence function on information theory, provides evidence that our method produces robust (with respect to the choice of the loss function) and reasonable rankings of sampling designs by comparing the rankings. We observed that the results derived from both perspectives are compatible even if they rely on different assumptions.

3.4 Discussion

As far as we know, our methodology is the first attempt for comparing sampling designs on networks based on information theory. Sampling on networks is important for certain inferential problems. Often we get to choose how we sample them, we need criteria for choosing a good sampling design. Existing methods exist based on minimizing loss functions; but these can be sensitive to choice of loss. In this chapter, we present a new criterion, based on information-loss, which is more reliable. The main advantage of our method is that it provides the statistician with a conceptual framework that enables him/her to compare sampling designs without the need to specify the inference beforehand, and therefore, the loss function. The results suggest that the ranking obtained from our method is reasonable when compared to rankings implied by the most widely used loss functions.

The proposed approach follows a frequentist perspective in the following sense: the second defined prior distribution in subsection 3.2.1 is treated as a fixed but unknown component of the data generating mechanism. The computation of the posterior is employed as a device to retrieve the ground truth. Finally, the evaluations of our approach in Section 3.3 are all frequentist measures of performance.

We want to emphasize that, while we used the MCMC schemes proposed by [63] to compute the posterior for the graph parameters, this does not have to be

Model	Feature	SD	MHD	MSE (P.P.)	MAE (P.P.)	MSE (P.)	MAE (P.)
SBM	K	S (2,2,2)	0.4432	82.917	9.238	79.297	9.137
SBM	K	S (3,3,2)	0.2934	37.081	6.153	36.028	6.253
SBM	K	S (3,2,3)	0.1122	16.721	4.065	15.729	4.153
SBM	K	RDS (2,2,2)	0.2752	33.291	5.812	32.876	5.974
SBM	K	RDS (3,3,2)	0.2192	20.917	5.249	20.385	5.298
SBM	K	RDS (3,2,3)	0.1974	21.021	5.397	21.746	5.464
LSM	α	S (2,2,2)	0.3982	70.002	8.085	68.682	7.901
LSM	α	S (3,3,2)	0.2464	32.997	5.064	31.827	5.003
LSM	α	S (3,2,3)	0.1334	13.723	3.032	12.829	3.237
LSM	α	RDS (2,2,2)	0.2381	31.769	5.902	30.076	5.902
LSM	α	RDS (3,3,2)	0.1527	18.177	3.979	16.994	3.007
LSM	α	RDS (3,2,3)	0.1401	18.922	4.092	17.375	3.394

Table 3.4: Means of Hellinger Distances Distribution (MHD) and means of Predictive Posterior (P.P), for point prediction, and Posterior (P.) Quadratic and Absolute Mean Distribution (MSE and MAE), for point estimation, for six different sampling designs in the settings of number of communities in SBM (K) and regression coefficient in latent space model (α).

the case and other approaches can be used to perform that computation.

The limitations of the approach in its current form include: i) All our experiments are obtained by using Hellinger distance to compare distributions. Other choices from f-divergence family tool, like the Kullback-Leibler divergence, could have been used instead. Though, all the well-known corresponding $f(t)$ functions from the f-divergence family (Kullback-Leibler, Total variation distance, x^2-divergence) behave the same in terms of monotinicity. ii) Compared to a decision theory approach, we have less conceptual tools for dealing with model misspesification. The work by [110] provides a framework for evaluating how the comparison of sampling designs would suffer under a decision theory approach ([61] and [26]).

Future work includes: i) To take into account additional sources of uncertainty such as missing and coarsening data. By doing this, we will need to incorporate the concept of channel capacity in our formulation. Our approach can be applied either for comparing mechanisms for coarsening data [45] and mapping methods (e.g. Isomap) on networks by optimizing regarding those mechanisms or mapping methods, respectively, instead of sampling designs. ii) To investigate the behavior of our method under model misspecification.

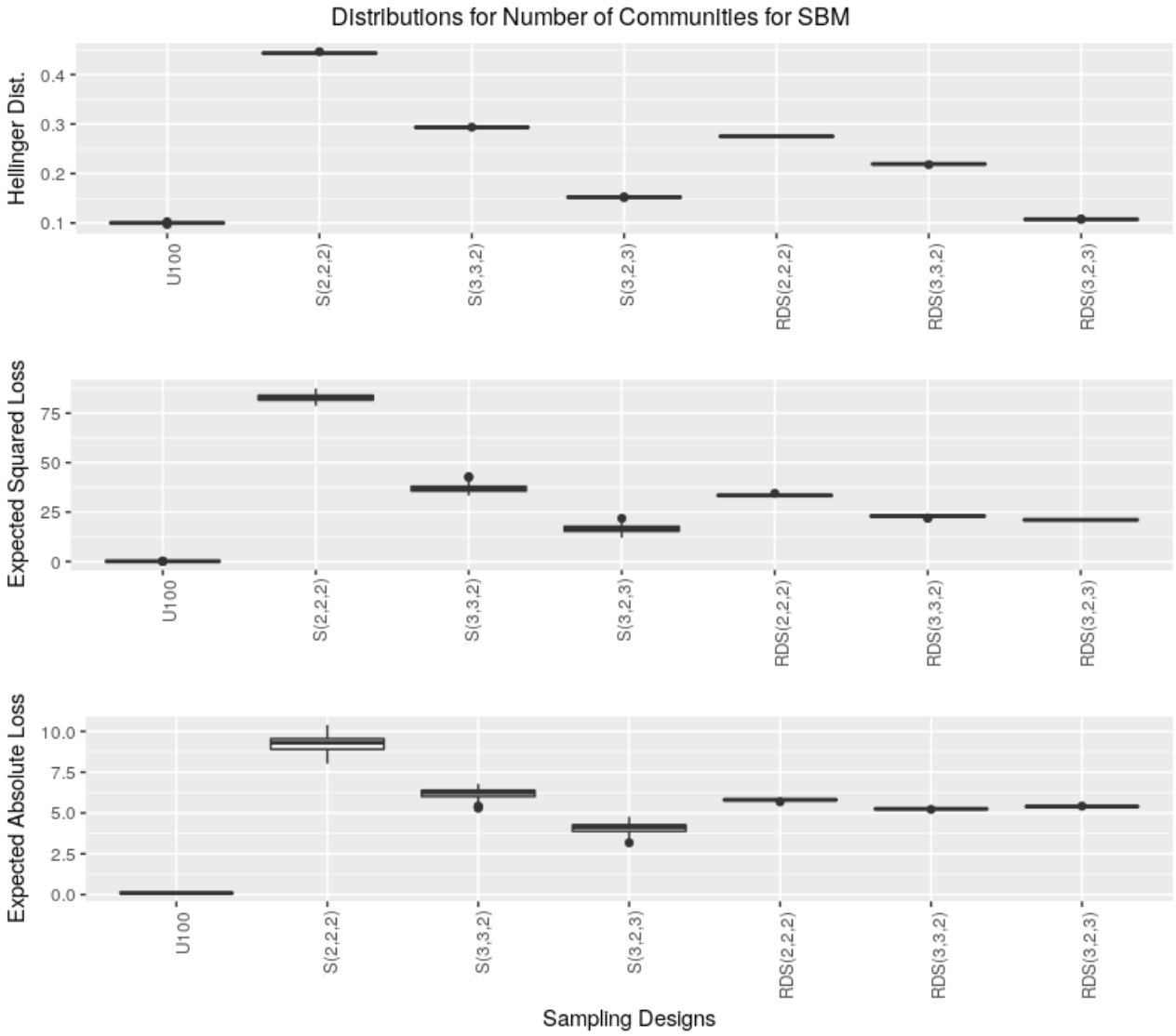

Figure 3.8: Upper: The lower the value of the mean of the Hellinger distance distribution the better for a sampling designs to be recruited regarding K, which is the number of the communities. Middle: The lower the value of the mean of the expected squared loss distribution of the predictive distribution, $\mathbb{E}_K(K - \hat{K})^2$, the better for a sampling designs to be recruited regarding community number. Down: The lower the value of the mean of the expected absolute loss distribution of the predictive distribution, $\mathbb{E}_K(\mid K - \hat{K} \mid)$, the better for a sampling designs to be recruited regarding communities. U100 means we sample the all 100 nodes and the Hellinger distance is essentially 0.

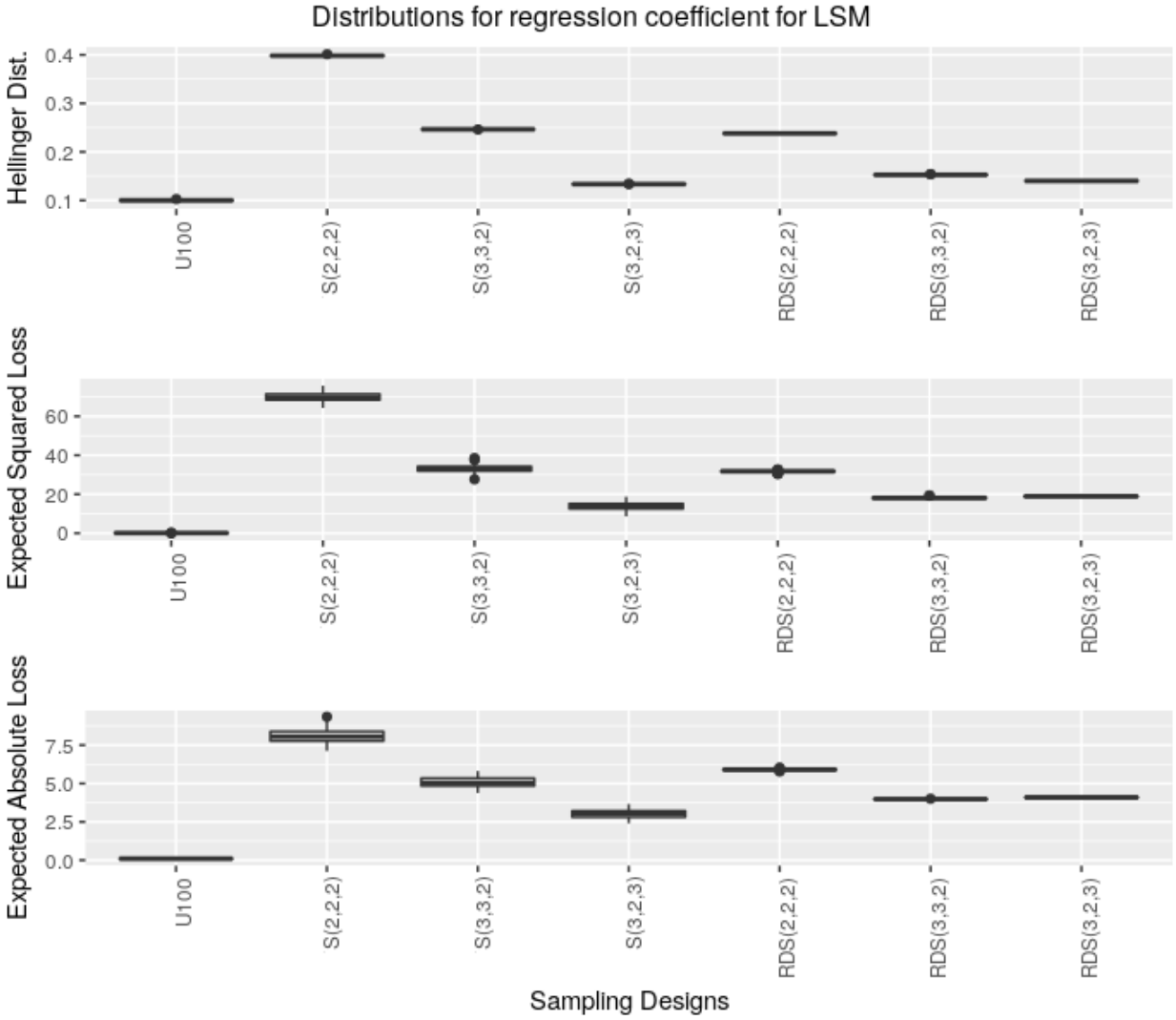

Figure 3.9: Upper: The lower the value of the mean of the Hellinger distance distribution the better for a sampling designs to be recruited regarding α, which is the regression coefficient. Middle: The lower the value of the mean of the expected squared loss distribution of the predictive distribution, $\mathbb{E}_\alpha(\alpha - \hat{\alpha})^2$, the better for a sampling designs to be recruited regarding the regression coefficient. Down: The lower the value of the mean of the expected absolute loss distribution of the predictive distribution, $\mathbb{E}_\alpha(\mid \alpha - \hat{\alpha} \mid)$, the better for a sampling designs to be recruited regarding the regression coefficient.

Chapter 4

Inference on network parameters and features involving sampling mechanisms

In this chapter we investigate to what extent a statement that we can make for a partially observed network can be translated regarding a statement to a fully observed network and via versa. There has been a previous effort in [9] where the authors make statements using conditional distributions. Their reasoning leads to wide-ranging and challenging problems [9]. Here, we prove that it is not always possible to connect such statements for all values of fully and partially observed random network features for inference purposes. In this chapter, we examine how those statements should be connected. We make the case that this problem should be formulated in terms of joint distribution of the underlying network and the sampling design. The statements allow them to constrain features and establishing some desirable features for either the fully or the partially observed random network.

Usually, with random graph models we are interested in performing inferences on a specific feature. For instance, assume that we have an Erdös-Rényi network model. Giant components are a prominent feature of the Erdös-Rényi model of random graphs, in which each possible edge connecting pairs of a given set of n vertices is present, independently of the other edges. According to the coupon collector's problem [86], $\Theta(n \log n)$ edges are needed in order to have high probability that the whole random network is connected. We are interested in the following problem: if we observe a random network that has a highly connected (in

some sense) giant vertex set of value α, can we infer that whole network has a similarly highly connected giant vertex set of value $\beta(\alpha)$? To address problems like, this we provide statements which involve original and observed random network models and a sampling design, in order to perform inference in their parameters or features. Ideally, those statements will be both meaningful and applicable for practitioners on networks.

Recent related literature covers a wide range of topics. More specifically, in [78], the author specify a randomized algorithm that, given a very large network, extracts a random subnetwork. The aim of the author is to examine what can we learn about the input network from a single subsample. That is why they derive laws of large numbers for the sampler output, by relating randomized subsampling to distributional invariance, by assuming the sample has been generated by a specific algorithm. In [114], the authors propose a new sequential importance sampling method for sampling networks with a given degree sequence. These samples can be used to approximate closely the null distributions of a number of test statistics involved in such networks and provide an accurate estimate of the total number of networks with given vertex degrees. In [115], the authors study the problem of how to estimate the degree distribution a true underlying network from its sampled network and show that this problem can be formulated as an inverse problem. Overall, their results show that the true degree distributions from both homogeneous and inhomogeneous networks can be recovered with substantially greater accuracy than reflected in the empirical degree distribution resulting from the original sampling. In [28], they describe how analysts can reconstruct topological features of networks that are partially revealed by diffusion processes. The framework is general and applies to network data that arise from a variety of missing-data mechanisms. In [97], they examine the following problem: given an incomplete network, which b nodes should be probed to bring the largest number of new nodes into the observed network? Many graph-mining tasks require having observed a considerable amount of the network. In [91], they address the problem of missing data in information networks. Specifically, given only a fraction of the complete network, their goal is to estimate the properties of the complete network.

This chapter is the idea of relating features of partially and fully observed networks respectively with the sampling method developed in chapter 3. We envisage an arbitrary random network model for the network we can not fully observe and devise a probability model for observed sampled network. Our contributions include: i) We show that generally statements that relate all values of a certain

random network model feature can not hold. Our aim is to prove that there is not an analytical solution for statements on how we can estimate a feature of interest of all random network model. Therefore, we propose more general statements, based on a Bayesian numerical solution to perform inference on features, which holds in every case of uncertainty regarding random network model, sampling design and network features. Analytical solutions might exist by constraining the model (e.g. for certain random network models). ii) We provide a tool to practitioners to check whether they are able to tweak either parameters or features of the fully observed random network models through a sampling mechanism to influence and affect other parameters or features of the partially observed network or the opposite. For this, the joint distribution of a certain feature regimes for the fully observed random network and a feature regimes for the partially observed random network is necessary.

The chapter proceeds as follows: In section 4.1, we give notation and definitions of networks, random networks and sampling designs and present statements that combine fully and partially observed features for one random network. In section 4.2, we prove symbolically that those kind of statements does not hold. Thus, we prove that generally it is enough to adopt our proposed statements, which are presented in section 4.3, in order to infer them by an numerical procedure. Conceptually and computationally, our methodology is presented. In section 4.4, for many random network models, data analysis that gives experimental results is conducted, showing the results of our approach. Finally, in section 4.5, we present with more details the future work involving overlapping research areas.

4.1 Preliminaries

We define a network $\mathcal{G} = (V, E)$ as a tuple of the set of N nodes V and the set of edges $E \subseteq V \times V$. Moreover, we denote by $A_{\mathcal{G}_{ij}}$ the $N \times N$ adjacency matrix of $\mathcal{G}$ with elements $\{i, j\}$. A random network is a probability model on the space of adjacency matrices. On this chapter we consider random network models in the space of binary symmetric adjacency matrices. To simplify the notation we are using the same letter for a network and a random network network, $\mathcal{G}$. In case we need to distinguish them we use $\mathcal{G}(\omega)$ as a realization.

Given $\mathcal{G}(\omega)$, a realization of the random network model, let I denote a sampling design that propagates through the network. A realization of $\mathcal{G}_{\mathcal{INC}}$ implies a

partition the network realization $\mathcal{G}(\omega)$ into $\mathcal{G}_{\mathcal{INC}}(\omega)$ and $\mathcal{G}_{\mathcal{EXC}}(\omega)$, which denote, respectively, the observed and unobserved parts of the random network realization. In analogous manner, all the adjacency matrix realizations $A_{\mathcal{G}}$ can be written as $A_{\mathcal{G}_{\mathcal{INC}}}$ and its completion $A_{\mathcal{G}_{\mathcal{EXC}}}$.

4.2 Potential Statements

Here we adopt the notation in [9]. Given a random network model, or a network which is an instance of this random network model, that has a certain feature of interest and a sampling design how likely is the produced observed random network to has this or another feature? More specifically, we want to infer the following function $\beta()$:

Statement: if $\mathcal{G}$ has a feature of value δ, then we can infer that $\mathcal{G}_{\mathcal{INC}}$ has a similar feature of value $\beta(\delta)$.

Likewise, given a partially observed random network model which was produced by a sampling design, how likely is to be produced by a random network which has another or the same feature? More specifically, like before we want to infer function $\phi()$:

Statement: if we observe $\mathcal{G}_{\mathcal{INC}}$ has a feature value δ^*, then we can infer that $\mathcal{G}$ has a similar feature value $\phi(\delta^*)$.

Some examples of the last two statements involve:

- If a certain partially observed random network has a highly connected giant vertex set of value α, then can we infer that the fully observed random network has a similarly highly connected giant vertex set of value $\beta(\alpha)$?
- If a certain partially observed random network has average degree n or say 13, then can we infer that the fully observed random network has a similarly average degree value $\beta(n)$?

If there were monotone invertible functions β or ϕ connecting those features and maps one to the other, then we are able to infer the one feature from the other. In that case, we say that those features are related. If there is no such monotone,

invertible functions then those features are not related and you can not retrieve the feature from the other. The logic described above proposed in [9], the authors provided a digression to prove this inference assertion. More specifically, suppose $\mathcal{G}(\omega)$ is conditional (a known realization network) to a random network $\mathcal{G}$. The authors claimed that if we have the following statement:

Statement: if $\mathcal{G}(\omega)$ has property F^*, we claim with probability $\geq p$ that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature Q.

or:

Statement: if $\mathcal{G}$ has feature Q^*, we claim with probability $\geq p$ that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature Q.

How can we restate this as an inference procedure of the format:

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, does not have feature Q then with probability $\geq p^*$, $\mathcal{G}$ does not have feature Q^*.

While in [9], the authors try to make statements in terms of confidence intervals, we formulate our in terms of posterior and predictive posterior probabilities in terms of parameters and features of the partially and fully observed networks, respectively. So, we model through predictive posterior the uncertainty of the underlying network combined with the uncertainty of sampling design.

We will show, through this assertion, that in some cases $\beta(\delta)$ and δ (or $\phi(\delta^*)$ and δ^*) are not related for every combination of random network model, its parameters or features and sampling design in such statements. Therefore, we propose another more general statements than [9] but we are not dealing with cases where $\beta(\delta)$ and δ are not related. In the following subsections we are going to give counterexamples for some well know random networks to show that there are features that $\mathcal{G}$ and $\mathcal{G}_{\mathcal{INC}}$ are not related.

4.3 Collapsing Potential Statements

Here, we will show via a counterexample the way to combine/relate the statements proposed by [9] can not be that general. This logic can be stated in the following

format:

$$(\mathcal{G} \in A \wedge \mathcal{G}_{\mathcal{INC}} \in B) \Rightarrow (\neg \mathcal{G}_{\mathcal{INC}} \in B \Rightarrow \neg \mathcal{G} \in A) \tag{4.1}$$

which means that $\mathcal{G}$ is an event A with some features and $\mathcal{G}_{\mathcal{INC}}$ is and event B with some other features.

The reasoning why this logic collapses is based on a counterexample which covers the first implication but not the second. The intuition behind this counterexample lies behind the simple fact that we can not infer one conditional $p(B \mid A)$ just from the following conditional $p(A \mid B)$. The joint distribution $p(A, B)$ of the features of fully and partially observed random networks, which we will use in section 4.4, is needed through the Bayes theorem. From Bayes theorem we know that:

$$p(B \mid A) = \frac{p(A, B)}{p(A)} = \frac{p(A \mid B) \times p(B)}{p(A)} \tag{4.2}$$

and

$$p(A \mid B) = \frac{p(A, B)}{p(B)} = \frac{p(B \mid A) \times p(A)}{p(B)} \tag{4.3}$$

The information that an event occurs is based on the other event occurring and consequently on the joint distribution of the two events. The only case that we can have valid connected statements from the previous cases, where the function of A, B of the quotient $\frac{p(B)}{p(A)}$ is known through a very strong assumption, e.g. it is constant.

Counterexample: Suppose A is a sure event. Then if we have that $\mathcal{G}_{\mathcal{INC}}$ comes with probability greater than p of the times. Then the probability of B has happened given A has happen is greater than p. This fulfills the premise. If B does not happen then A does not happen. But here if B does not happen it can not happen that A does not happen because A covers the whole space. So that means the second implication breaks so the whole statement breaks.

Concrete examples of particular random network models include:

- **Small World, $\mathcal{G}$ always happens, RDS**: Assume a small world graph where $\mathcal{G}$ has always one connected component. For I we select a link tracing design, with large sample size and large number of seeds, and B is the number

of connected components of $\mathcal{G}_{\mathcal{INC}}$. $\mathcal{G}_{\mathcal{INC}}$ does not have one connected component with high probability. By design the second implications breaks.

- **SBM, $\mathcal{G}_{\mathcal{INC}}$ always happens, Snowball**: Assume a stochastic block model where two blocks one of which is quite dense and we can control the probability for interblocks. For I we select snowball such that it has one seed and that seed is always in the same dense block. $\mathcal{G}_{\mathcal{INC}}$ will always have one connected component. Since we can control the interconnected probability, $\mathcal{G}$ can not have more than one connected component with high probability. By design the second implications breaks.

If we want to combine these statements and go back and forth we have to relate the uncertainty of the population of the underlying network and the uncertainty we have through the sampling mechanism. One type of statements is conditional on the full random graph and use the uncertainty of sampling networks to be indicative regarding a feature of the observed graph. For the other type of uncertainty we have the partially observed random network and we would like to perform inference about features in the space of possible fully random networks. Of course, the practitioner can have separately conditional statements. Though, in case his/her objective is to use one statement of one level to learn something about a statement in a different level and via versa he/she needs to relate the uncertainties of underlying networks and sampling design.

In case we want to find such relationships we have to escape from the above constraints. Finding other similar conditions or analytical $\beta()$ functions are problems for further research. As a general approach to find relationships between $\mathcal{G}$ and $\mathcal{G}_{\mathcal{INC}}$ we follow the numerical procedures in chapter 3.

4.4 Proposed statements for Inference

Here, we suggest a more general type of statements, than the aforementioned logic, that holds for every source of uncertainty concerning every random network model, sampling design and feature and can be computed numerically, in the space of observables, through figures 4.1 and 4.2.

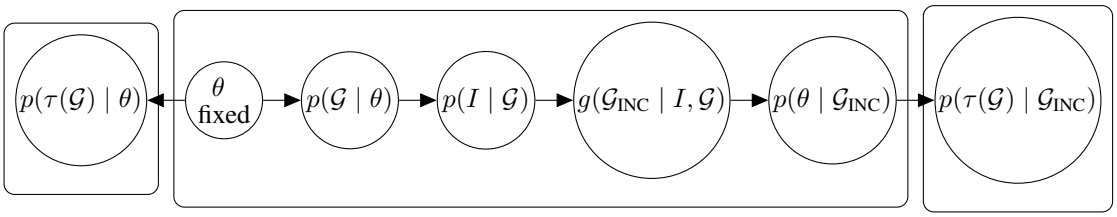

Figure 4.1: Getting from fully observed random network feature to statements about partially observed network parameter or feature given a sampling design I (the same holds for a sampling mechanism on edges as well).

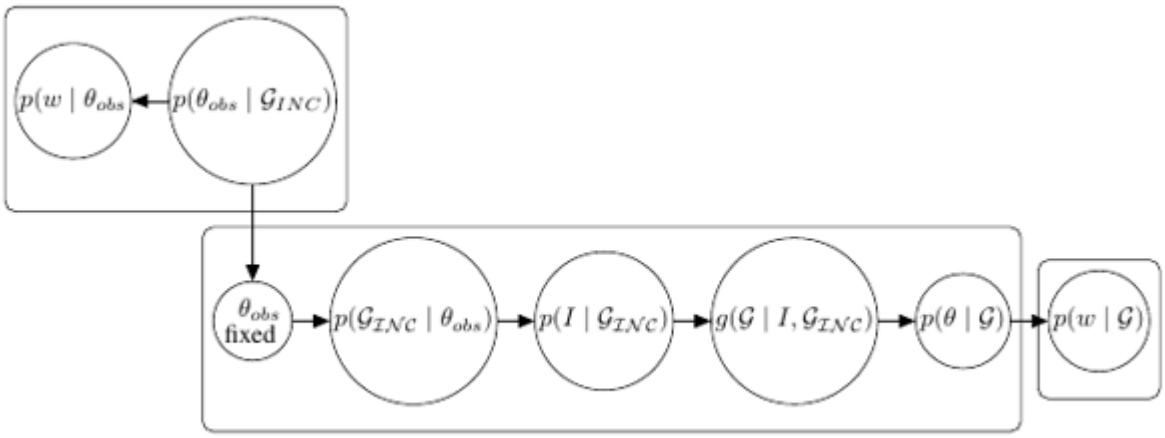

Figure 4.2: Getting from partially observed random network feature to statements about fully observed network parameter or feature given a sampling design I (the same holds for a corrupting mechanism on edges as well).

Statement: if $\mathcal{G}(\omega)$ has property W, we claim with probability $\geq p$ that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature Q.

or:

Statement: if $\mathcal{G}$ has feature Q^*, we claim with probability $\geq p$ that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature Q.

Is it possible to restate it as an inference procedure of the format:

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, does not have feature Q then with p^*, $\mathcal{G}$ does not have feature Q^{**}.

We can state the inference in positive terms, so we negate the features and restate as follows.

If we wish to justify a statement of the format:

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature P, we claim with probability $\geq p$ that $\mathcal{G}$ has feature P^*.

then we need to prove a theorem of the format:

Statement: if $\mathcal{G}$ has not feature P^*, then with $\geq p$ probability $\mathcal{G}_{\mathcal{INC}}$,, which was produced from a suitable sampling design I, does not have feature P^{**}.

Those statements can be helpful in the context of providing a practitioner with the intuition of how two features are related when a sampling design is involved. From [75] and more particularly from [99], it is proven that a subnetwork of a scale-free model is not necessarily scale free, so the power exponent is distorted through some sampling mechanisms. On the other hand, from [31] we know that below the threshold of $\frac{1}{n}$, the largest component of the graph includes no more than a factor times $log(n)$ of the nodes. Above the threshold of $\frac{1}{n}$, a giant component emerges, which is the largest component that contains a nontrivial fraction of all nodes. The giant component grows in size until the threshold of $\frac{logn}{n}$, at which point the network becomes connected. This is illustrated in figure 4.3 for 50 nodes and in figure 4.4 for 500 nodes and is in the heart of what we want to achieve. Sometimes, we can not related features of partially and fully observed network data when the random network model and the sampling design are mismatched. Our main goal is to leverage the theoretical results of collapsing the logic in [9] in order to encapsulate the all the information in order the practitioner to learn as much as he can about the features he/she is interested in.

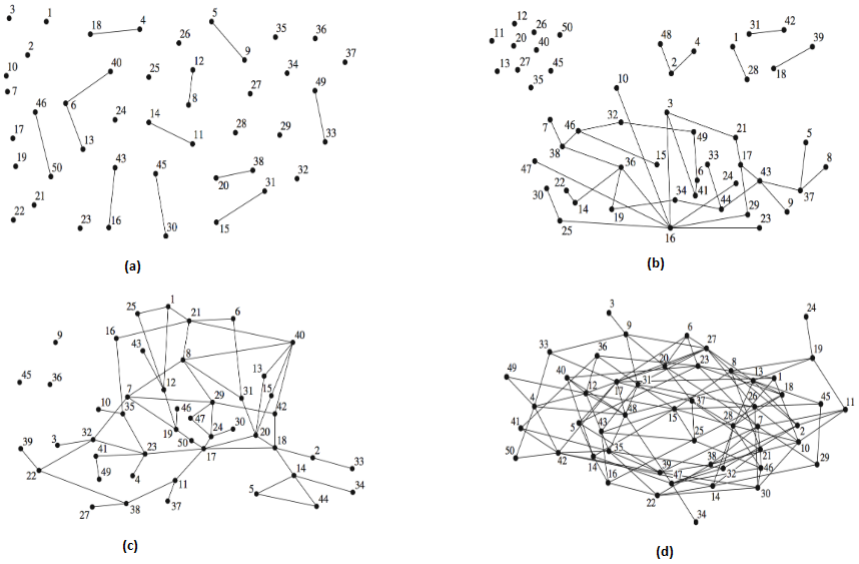

Figure 4.3: d is the degree density. (a) A first component with more than two nodes: a random network on 50 nodes with $d = 0.01$. (b) Emergence of cycles: a random network on 50 nodes with $d = 0.03$. (c) Emergence of a giant component: a random network on 50 nodes with $d = 0.05$ and (d) Emergence of connectedness: a random network on 50 nodes with $d = 0.10$.

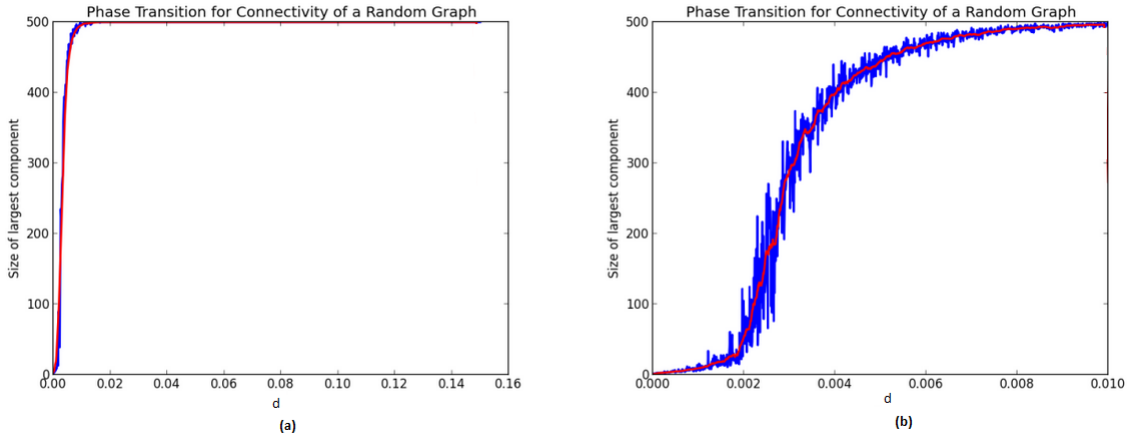

Figure 4.4: The transition starts at 0.002 and ends at 0.01 after which threshold we have one giant component and the network is connected.

We consider the case when both partially and fully observed random networks are related through a sampling mechanism. Given that the way we formulate the statements we can investigate whether they are useful via joint distribution. The joint distribution implied by the full random graph network and the sampling design distribution, can provide a principle tool for constraining information of the one feature that can be translated in constraining information in another feature. The point behind the following logic is that it can inform practitioners whether and how two features are connected and quantify their association.The joint distribution is used as a bridge between those two (or more) features and shows how they work together.

Making statements using joint distributions for different types of features in the levels of partially and fully observed networks sometimes is very hard. We are aware that for some combinations of features in those two levels it is hard to convey information concerning the relationship between those two features. Some choices of features might be completely non-sense. The above statements are very general and might not be useful for some particular applications. They do not provide a meaningful answer in many cases, since the values of Q^* and Q^{**} might be very different. That is why we need to investigate these statements through simulation studies.

We assume the following setup: we have the joint distributions of a certain feature interval Q for the fully observed random network and a feature interval W for the partially observed random network, respectively. Either the fully observed network or the partially observed network was given with the sampling design.

Statement: if $\mathcal{G}$ has feature $Q^* \in Q$, we claim with probability $\geq p$ that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature $W^* \in W$.

It is possible to restate it as an inference procedure of the format:

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, does not have feature $W^* \in W$ then with probability p, $\mathcal{G}$ does not have feature $Q^* \in Q$.

We can state the inference in positive terms, so we negate the features and restate as follows.

If we wish to justify a statement of the format:

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable sampling design I, has feature P, we claim with probability $\geq p$ that $\mathcal{G}$ has feature P^*.

then we need to prove a theorem of the format:

Statement: if $\mathcal{G}$ has not feature P^*, then with $\geq p$ probability $\mathcal{G}_{\mathcal{INC}}$,, which was produced from a suitable sampling design I, does not have feature P.

As we have stated in order to get those statements the joint distribution of the features of fully and partially observed network is required. From this distribution which is the connection between the uncertainties of the fully observed network and the partially observed network's features we can derive both types of the above statements. In that way, we can calibrate either Q^* or W^* to constrain them respectively and get valid and useful relationships between features.

To provide solutions to each one of the above statements we use the algorithms that was created in [6] and used in [7] for both ignorable and non-ignorable sampling designs.

4.5 Simulations

We investigate how a feature of the partially observed random network is connected with a feature of fully observed random network via a sampling design. In this section, we consider the Erdös-Rényi model [31, 32] and degree and clustering coefficient for transitivity. In Erdös-Rényi model the degree and the transitivity are the same. We will perform two simple cases creating statements about features regarding their conditional distributions (proposed statements in previous section) first and then the features joint distribution. With those simulations we want to show that there are cases where we can use the conditionals instead of joint distribution because they provide the same information (so the quotient of $\frac{P(A)}{P(B)}$ is constant), in addition with the well known examples such as: i) subnetworks of scale free distributions are not necessarily scale free ii) Erdös-Rényi phase transition described in section 4.3.

Figures 4.5 and 4.6 are interpreted as follows: The left subfigure is always

Random Network Model	Features	Sampling Design
Erdös-Rényi model	Degree Density and Transitivity	Ignorable-Snowball design I

Table 4.1: Random graph model, features and Snowball sampling design $I = S(2, 3, 3)$ with N=100 nodes considered for setting up simulation regimes.

illustrating the joint distribution of having a network model and applying to it a sampling design. It is used to extract all the statements concerning joint distributions and their negations. On the other hand, every right subfigure show how the joint distribution of having a partially observed network and "guessing" the true ground truth model which generated it (without knowing it beforehand). Using both subfigures we can obtain the statements about the conditionals and their negations.

For the degree and transitivity in the setting of table 4.1 and the conditional distributions we have the following statements:

Statement: if $\mathcal{G}$ has degree $Q^* \in [0.188, 0.216]$, we claim with probability 90% that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a snowball design with $I = (2, 3, 3)$, has transitivity $W^* \in [0.2015, 0.2025]$.

Is it possible to restate it as an inference procedure in figure (as in figure 4.5 right):

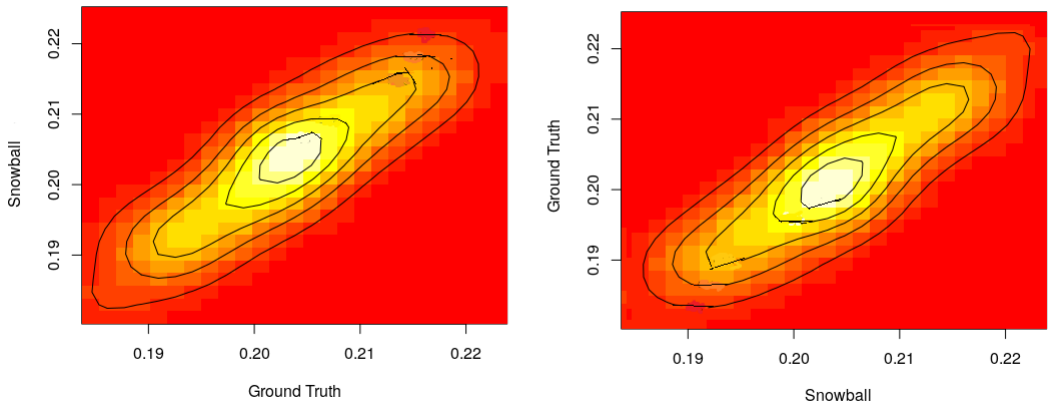

Figure 4.5: Illustration of statements connecting the Degrees and Transitivity values in a random network through an Ignorable Sampling Design $I = S(2, 3, 3)$.

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a snowball design with $I = (2, 3, 3)$, does not have transitivity of $W^* \in [0.2015, 0.2025]$ then with 90%, $\mathcal{G}$ does not have degree $Q^* \in [0.189, 0.213]$.

For the degree and transitivity in the setting of table 4.5 (left) and the joint distribution we have the following statements:

Statement: if $\mathcal{G}$ has degree $Q^* \in [0.188, 0.216]$, we claim with probability 90% that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a snowball design with $I = (2, 3, 3)$, has transitivity $W^* \in [0.2015, 0.2025]$.

It is possible to restate it as an inference procedure in figure 4.5 (left):

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a snowball design with $I = (2, 3, 3)$, does not have transitivity $W^* \in [0.2015, 0.2025]$ then with 90%, $\mathcal{G}$ does not have degree $Q^* \in [0.188, 0.216]$.

Moreover, in this section we will perform two more simple calculations creating statements about table 4.2.

For the degree and transitivity in the setting of table 4.2 and the conditional

Random Network Model	Features	Sampling Design
Erdös-Rényi model	Degree Density and Transitivity	Non-Ignorable-RDS design I

Table 4.2: Random graph model, features and sampling design $I = RDS(2, 3, 3)$ with N=100 nodes considered for setting up simulation regimes.

distributions (figure 4.6) we have the following statements:

Statement: if $\mathcal{G}$ has degree $Q^* \in [0.188, 0.216]$, we claim with probability 90% that $\mathcal{G}_{\mathcal{INC}}$, which was produced from $I = RDS(2, 3, 3)$, has transitivity $W^* \in [0.201038, 0.202726]$.

Is it possible to restate it as an inference procedure in figure (as in figure 4.6 right):

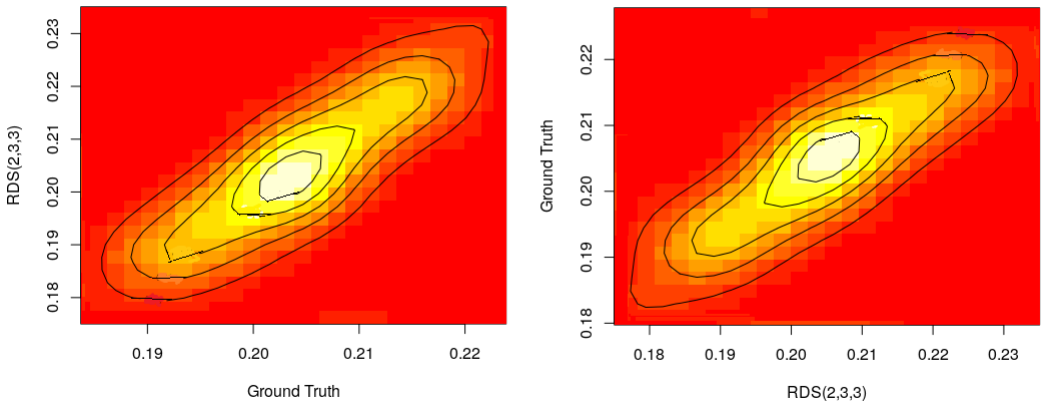

Figure 4.6: Illustration of statements connecting the Degrees and Transitivity values in a random network through an Non Ingorable Sampling Design $I = RDS(2, 3, 3)$.

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable non-ignorable sampling design $I = RDS(2, 3, 3)$, does not have transitivity of $W^* \in [0.201038, 0.202726]$ then with 90%, $\mathcal{G}$ does not have degree $Q^* \in [0.188, 0.215]$.

For the degree and transitivity in the setting of figure 4.6 (left) and the joint distribution we have the following statements:

Statement: if $\mathcal{G}$ has degree $Q^* \in [0.188, 0.216]$, we claim with probability 90% that $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable non-ignorable sampling design $I = RDS(2, 3, 3)$, has transitivity $W^* \in [0.201038, 0.202726]$.

It is possible to restate it as an inference procedure in figure 4.6 (left):

Statement: if $\mathcal{G}_{\mathcal{INC}}$, which was produced from a suitable non-ignorable sampling design $I = RDS(2, 3, 3)$, does not have transitivity $W^* \in [0.201038, 0.202726]$ then with 90%, $\mathcal{G}$ does not have degree $Q^* \in [0.1188, 0.216]$.

The proposed statements for inference are confirmed in this section. We can easily see, from all two examples, that the values of the features of the fully and the partially observed networks are connected when we consider both the conditional distributions of the network model and the sampling design and their joint distribution, something that was expected to be shown. The way we can use a statement of partially observed network and translate it to a statement of fully observed network depends on the joint distribution of the underlying network and the sampling mechanism. The only case that relaxes this assumption is extra assumptions on the function of the quotient $\frac{P(A)}{P(B)}$ (e.g. to be constant). If the practitioner only have the conditionals this is the only reliable way to connect those statements. Generally, we can not retrieve/extrapolate the results because of the part of information not encoded in the conditionals. The same can be performed with every combination of: a) random network model, b) a couple of their features and c) any sampling design.

4.6 Discussion

We conclude, through the assertion in section 4.1, that generally for $\mathcal{G}$ and $\mathcal{G}_{\mathcal{INC}}$ random networks we can not make the following conditional statements of the form:

- if $\mathcal{G}$ has a feature of size δ, then we can infer that $\mathcal{G}_{\mathcal{INC}}$ has a similarly feature of value $\beta(\delta)$, or

- if we observe $\mathcal{G}_{\mathcal{INC}}$ has a feature of value δ^*, then we can infer that $\mathcal{G}$ has a feature of value $\phi(\delta^*)$.

The question that arises naturally is how useful are the statements of the suggested logic for all random networks. An approach for tackling this question needs to take into account the following sources of uncertainty:

- Random Network Model
- Sampling Mechanisms.
- Probability of the statement (the only factor that we can certainly state that when it decreases the range of values of features will not increase).
- Features

The results confirm our initial intuition that for different random network models, different features can behave differently regarding the sampling designs. In order to connect such statements the joint distribution of both the random networks and the sampling design is required, except the case where the conditional distributions are identical. It is reasonable to say that there are cases that those statements are insightful and others that are not.

One more direction for future work involves data coarsening. Data coarsening is a statistical framework for statistical data including cases such as rounded, heaped, censored, partially categorized and missing data [45]. The type of incompleteness that is most commonly studied is missing data are the cases where the data are either perfectly known or entirely unknown. Though, in common situations, data are neither entirely missing nor perfectly present. Instead, we observe only a subset of the complete-data sample space of $\mathcal{G}$ where the unobservable data lie and we refer to this kind of incomplete data as coarse data.

Chapter 5

Robustness on Exchangeable networks

In this chapter, we propose an approach for robustness on exchangeable random networks. We assume an exchangeable model for performing inference on a feature of a random network. The problem is to assess how the quality of that inference gets degraded if the model is slightly modified. In that way, we are able to check, for an assumed network model which we can fit, how bad can a specific inference for that feature be. We consider decision making methods under model misspecification by quantifying stability of optimal actions to perturbations to the approximating/assumed/working model (the model that used to fit the data) within a neighborhood in model space. This neighborhood is consisted by a ball in a model space with radius defined by an information (Kullback-Leibler) divergence around the graphon of the assumed model. Our approach is inspired by recent developments in the context of robustness and recent works in the robust control, macroeconomics and financial mathematics literature and more specifically is based on the concept of graphon approximation through a stochastic block model.

This chapter presents a new method in robust decision making in random network models from approximate statistical models. We adopt Bayesian inference on networks and particularly we focus on inferences that can be phrased in terms of decision theory. One examples of inference includes point estimation for a specific feature. The setup is the following: how model misspecification degrades the quality of the inference. In order to achieve that, we cast this problem in the framework of [110] in network context. To be clear and consistent with the con-

text there are two models we are interested in throughout this thesis: the model that generates the data, which we called either the true or the generative model and the model we use to fit the data which we call either approximating or working or assumed model. For simplicity, we will use either the word approximating or assumed model.

As a trivial illustration consider the following two motivational examples: The first example involves different departments of a university where co-authorship chapters are published. For the most medical departments the researchers collaborate with epidemiologists/statisticians. Considering the statistical department, there are many pure statisticians authors and coauthors from the people of different departments. As a result, some people of the departments are more productive than others and some of the statisticians are more proactive than others so we can have different connection nodes (edges going to medical, gastronomic, pharmacology from statistical department etc). To fit a model, describing the structure of those relationships, makes sense because some statisticians are assigned to these departments, such as many statistician potentially will connect with different department with different probabilities. In higher level; this can be interpreted as a stochastic block model (exchangeable random network model) but some assumptions can be into question. Can the average density of this network be formulated by a simple model? The second example is associated with epidemiology (example in [6]). Practitioners want to learn about the population of Drug abusers or sex workers and we want to perform inference in the prevalence of HIV. They can setup a model for the network and they can have a lot of interest in the degree distribution. The quality of interest (prevalence of HIV) depends on how well they can learn the degree distribution. If they assume again an exchangeable random network model for describing the underlying network and if they know the model that generated the data they can ask how much harm there is by using the assumed model in the estimation of degree distribution and answer how much harm is done?

The approach we propose is based on three main different statistical concepts: robustness approach presented by [110], graphon approximation and stochastic optimization. First, we use the approximation the true/generated exchangeable random network representation that produces/generated the data by a parametric model in order to handle it easier. Then we propose another approximating exchangeable network model that it is easier to fit to the true data. We want to see how much harm is the last approximating model doing to the data when it comes

to a specific inference in a network feature. The core idea is to construct a neighborhood in model space, find the worst case scenario for the inference. The main challenge will be in terms casting and implementing this theory to random networks.

The literature of exchangeability of infinite binary arrays - which encompasses networks - is very well understood: see for example [10, 11, 48], [59, 60] and [51]. Of particular relevance here is the work of [29] (but see also [79]), which details the connections between exchangeability of random graphs and the notion of graph limits developed in [62]. In [80], [5] and [57] the authors describe methods of how to approximate network limits (graphons with a stochastic block model. As an extend in [81] the authors present the network histogram, a version of which we are will use in this chapter. Last, in the literature the scientific theory behind robustness is well established and being used from the early 80s (e.g. [16]) until the 1990s and 2000s when computational advances and hierarchical models broadly outpaced the complexity of data sets being considered by statisticians. In more recent times, very high-dimensional data are becoming common, the so called big data era, whose size and complexities prohibit application of fully specified carefully crafted models. In the chapter of [110] the theory is recapped and extended.

The main contribution of our chapter can be phrased as follows: we provide a methodology to examine whether and how much an approximating random network model is suitable for describing a true random network model in terms of a specific feature. We want to see whether a model we are using for inference is useful and can describe the data of the generating model. Recent literature, provides a bridge between graph limits called graphons, in model space, with stochastic block models for exchangeable random network model. Our main challenge that we encounter is that we connect tools like stochastic optimization and graph limits to explore the model space. Suppose we have a SBM for prediction which is used to code the community structure. We might have an application with communities with some unambiguity on how data are generated. We are putting the assumptions into question. For instance, when our assumptions are not clear then in that setup maybe something that generated the data is something like a block model but not exactly. We want to infer the number of communities or a feature of community e.g. state that statisticians assume a model for this applied problem and they perform this inference. How sensitive could be the quality of that inference if the data did not come exactly from a SBM but something else that is approxi-

mated by a way we do not know? What would be the harm if the model was not coming from SBM? Here, we answer these questions by looking in the inference of the exchangeable random network models feature (e.g. number of blocks) by stating it as a decision theory problem. In order to perform all the above we make a key assumption: the model that generates the data is in the neighborhood of the model used to fit the data.

The chapter proceeds as follows: In Section 5.2, we describe settings of the problem, we formulate them and give notation and definitions of exchangeable networks, exchangeable random networks, graphons and model space. Then, in section 5.3, we focus in our main purpose of this chapter which is how to use and apply current tools on random networks to see how robust a random network is regarding the inference of a specific feature. Conceptually and computationally, our methodology is presented. In section 5.4, for random network models data analysis that gives experimental results involving graphons, stochastic block models, simulated annealing,robustness and model misspecification is conducted showing the results of our approach. Finally, in section 5.5, we present with more details the future work involving overlapping research areas.

5.1 Preliminaries

5.1.1 Robustness

Robust Bayesian analysis investigates the robustness of answers from a Bayesian analysis to uncertainty about the precise details of the analysis, e.g. the misspecification of the prior. People in robust optimization deal with optimization problems in which a certain measure of robustness is sought against uncertainty that can be represented as deterministic variability in the value of the parameters of the problem itself and/or its solution. Bayesian robustness and robust optimization are the predecessors of [110] who presents an overview of recent natural developments in the area. Related to this, approximate probabilistic inference techniques that are misspecified by design have emerged as important tools for applied statisticians tackling complex inference problems.

Sometimes the generative model for data is not contained in the family of the models that used to fit the data and statisticians are interested in performing inference on those data. What is the impact of model misspecification on the quality of

inference? We assume that the object of specific inference is in terms of specific feature and we suspect and want to rule out the possibility that possible model misspecification may affect out capability of inferring that feature. By robustness we focus on inference that we are interested a priori. We are not entering the discussion of a model trying to capture every feature of the random network model in a satisfactory way(e.g. diagnostics). For example, suppose that the practitioner is interested in the mean of a distribution and he assumes a very well simple model for the tail. It is explaining very well the tails but is complex? Was it worth it? He has better fit in the tails but worse in the mean in which he is interested in.

In [110] and their discussants [16, 43, 65] review recent Bayesian decision theory research based on a local-minimax approach. They assume a prior and likelihood are specified, but then consider decisions that are minimax over all distributions within a given (small) Kullback-Leibler divergence from the posterior. They describe methods for estimating the sensitivity of a model with respect to the loss function by analyzing the effect of local perturbations in neighborhoods centered at the approximating model (in a Bayesian context this would be the posterior distribution). These neighborhoods are defined using the Kullback-Leibler divergence. This approach provides a bridge between the two dominant paradigms in decision theory: Walds minimax [106, 108] and Savages expected loss criterion. Two key features of this framework are that the solution is analytical, and it unifies other well known methods in Statistics such as predictive tempering, power likelihoods and Gibbs posteriors. It also offers an interesting solution to the Ellsberg paradox. Another application of their work is in the area of computational decision theory where the statistician only has access to the model via a finite set of samples. In this context, the methods can be used at very little extra computational cost. Moreover, they consider non parametric extensions to the approximating reference model. In particular, they look at the Pòlya tree process, the Dirichlet process and bootstrap procedures. Again using the Kullback-Leibler divergence, it is possible to characterize random samples of these non parametric models with respect to the base model, and therefore understand the effect of local perturbations on the distribution of loss of the approximating model. A series of diagnostic plots and summary statistics are presented. These complete the framework of post-hoc assessment of model stability and allow the user to understand why the model might be sensitive to misspecification. Graphical displays are an essential part of statistical analyses, indeed the point of departure for any serious data analysis. Their use in model exploration in the context of decision theory, however, is not common. They borrow some ideas from finance and econometrics

as a basis of exploratory decision-system plots.

Our approach is focused in robustness regarding the misspecification of the likelihood whereas there are other approaches like the misspecification of the prior which we are not dealing with. Authors in [16] provide a thorough review while [110] consider this in a decision focused manner. Approaches based on tilted likelihoods, which we focus and use in this chapter, can be found in [42, 69] and mainly in [110], the theory of whom we are using, among others. Specifically, in our case, in robustness the quality of that inference is encoded by a loss function. In order to do that a neighborhood, Γ, of the model space ($\mathcal{M}$) is constructed, which is given by a sphere-ball with a center, G^* and radius C. The center is the approximating model and the topology of the space is defined by some distance. Then exploration and exploitation of the space to find the worst case scenario, which is given by the maximum expected loss function in that neighborhood, is conducted. Usually practitioners resort to information theory to define the space of the potential exchangeable random networks that could fit the data. For that reason, every move in the space of potential models is performed by using Kullback-Leibler divergence or either metric spaces (e.g. l^1). Moreover, decision theory helps practitioners to check how much impact has a potential model in that space regarding a specific feature (e.g. density, degree distribution in networks etc). More specifically, after the move is performed, is common, squared Expected Loss of a feature of the exchangeable random network model to check how this model is behaving is used to show how well is characterizing the true data, in terms of that feature.

5.1.2 Exchangeable Random Networks and Graphons

There are two types of exchangeability network models. The first one concerns a node exchangeable random graph (exchangeable random graph) is a random graph on labeled nodes such that any (fixed) permutation of the labels yields a random graph with the same distribution. This is natural if the labels are just labels without intrinsic significance. An example is a stochastic block model. The second type of exchangeability is the edge exchangeable random graphs which were introduced in [27]. An equivalent model, using somewhat different formulations, was given by [22] and [24]. The idea is that we have a fixed (labelled) vertex set, and add a sequence of edges (regarded as pairs of vertices). Repetitions are allowed, so we construct a multigraph. The sequence of edges is supposed to be exchangeable. By De Finettis theorem, this is equivalent to the following:

Let $\mathcal{V}$ be a finite or infinite set, and let μ be a deterministic or random probability measure on the edges of the complete graph on $\mathcal{V}$.

- Given μ, take N i.i.d. edges with distribution μ.
- Delete all isolated vertices.

There are some similarities with vertex exchangeable random graphs with a discrete type space N, but this type of exchangeability is quite different. An example is: let (q_i) be a probability distribution on $\mathcal{N}$. For each edge, just pick its two endpoints independently with this distribution. Thus $\mu(ij) = q_i q_j$.

For this chapter when we refer to exchangeable models for networks, we refer to node exchangeability. More specifically, a common feature shared by many network models is that of invariance to the relabeling of the network units, or (finite) exchangeability, whereby isomorphic graphs have the same probabilities, and are therefore regarded as statistically equivalent. Exchangeability (from De Finetti's theorem) is a basic form of probabilistic invariance, but also a natural and convenient simplifying assumption to impose when formalizing statistical models for random networks. Examples of popular network models which rely on exchangeability include many exponential random graph models [98], the stochastic block model, latent space models [84], to name a few.

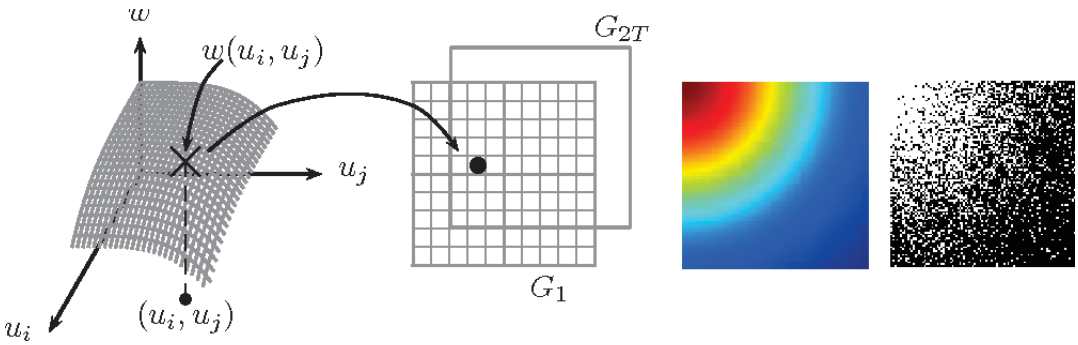

Figure 5.1: Example of graphon representations ([5]).

Graphons (figure 5.1) arise as the fundamental objects in two areas: as the defining objects of exchangeable random graph models and as a natural notion of limit for sequences of dense graphs. Specifically, a graphon is a symmetric

measurable function $f : [0,1]^2 \to [0,1]$ which integrates to 1. Usually a graphon is understood as defining an exchangeable random graph model according to the following scheme: Each vertex j of the graph is assigned an independent random value $u_j \sim U[0,1]$. Edge (i,j) is independently included in the graph with probability $f(u_i, u_j)$. A random graph model is an exchangeable random graph model if and only if it can be defined in terms of a graphon in this way.

The simplest example of a graphon is $f(x,y) \equiv p$ for some constant $p \in [0,1]$. In this case the associated exchangeable random graph model is the Erdös-Rényi model that includes each edge independently with probability p. The Erdös-Rényi model can be generalized by Stochastic Block Model: Divide the unit square into $K \times K$ block, not necessarily of equal size. Let f equal p_{lm} on the l, m th block. In this chapter, we are considering undirected random networks, so symmetric graphons.

5.1.3 Simulated Annealing

Simulated annealing (SA) is a probabilistic optimization technique for approximating the global optimum of a given function. Therefore, rescaling partially avoids the trapping attraction of local maximum. Given a temperature parameter $T > 0$, a sample $\theta_1^T, \theta_2^T, \ldots$ is generated from the distribution:

$$\pi(\theta) \propto \exp(h(\theta)/T) \tag{5.1}$$

and can be used to come up with an approximate maximum of h. As T decreases toward 0, the values simulated from this distribution become concentrated in a narrower and narrower neighborhood of the local maxima of h.

Algorithm 3 Simulated Annealing

1. Simulate ζ from the distribution of $\pi(\theta)$.
2. Accept $\theta_{i+1} =$ with probability $\rho_i = \exp(\Delta h_i/T_i) \wedge 1$ where $\Delta h = h(\zeta) - h(\theta_i)$ for $i = 0 \ldots I$, I the number of iterations and

$$\theta_i = \begin{cases} \zeta \text{ with probability } \rho = \exp(\Delta h/T) \wedge 1 \\ 0 \text{ with probability } 1 - \rho \end{cases}$$

3. Update T_i to T_{i+1}.

Therefore, if $h(\zeta) \geq h(\theta_0)$, ζ is accepted with probability 1; that is, θ_0 is always changed into ζ. On the other hand, if $h(\zeta) \leq h(\theta_0)$, ζ is still be accepted with probability $p \neq 0$ and θ_0 is then changed into ζ. This property allows the algorithm to escape the attraction of θ_0 if θ_0 is a local maximum of h, with a probability which depends on the choice of the scale T, compared with the range of the distribution of π.

5.2 Methodology

We cast the ideas proposed by [110] in the context of network data, mentioned above, on how to perform the computations and implementations on network. We do not provide additional theoretical results from those in [110] but a computational tool for robustness on exchangeable network models. Specifically, the final outcome we propose presents the robustness of the exchangeable random network model which is given by using stochastic optimization algorithms in the space of non parametric models in order to find the worst possible model that maximize the expected loss function in the model space. The main challenges are to use simulated annealing as optimization algorithm, use graphons as non parametric models and define the model space and the radius of the space. Algorithm 4, below, describe explicitly all the steps we follow:

Algorithm 4 Algorithm for Robustness

1. Get the graphon of the generating model and the approximated model.
2. Discretize the graphon, by using an $n \times n$ grid, get a SBMs and compute the KL between them in order to find the radius C of the sphere which includes all the exchangeable random network models (fulfilling the assumption made).
3. Sample the ball to get an appropriate value of the parameter T of the simulated annealing.
4. The function h of the simulated annealing is the expected loss of a specific feature .
5. Move inside the ball using the perturbing and rescaling moves and find the maximum expected loss of that specific feature using simulated annealing.

Firstly, in order to to be consistent with [110] we define the space of the

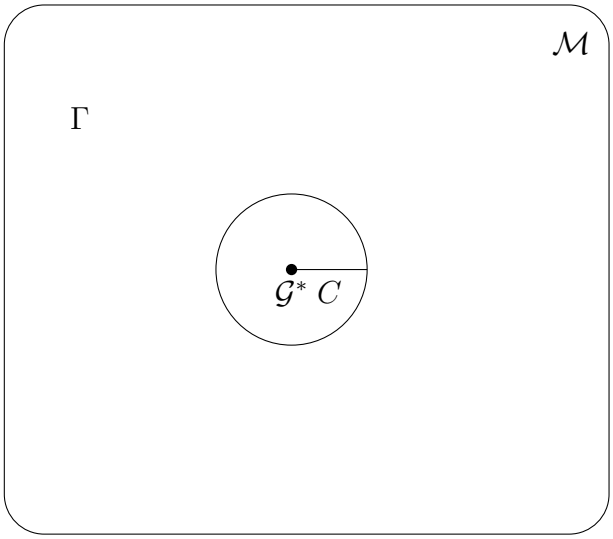

Figure 5.2: Space of Models

exchangeable random network models the approximating model, the generating model and how we compute the radius of the ball. The space of models we are dealing in this chapter is the space of all graphons of the exchangeable random network models, here denoted by $\mathcal{M}$. The Kullback-Leibler divergence defines the space where the modified models are located. The model which generates the data is denoted by G and the approximating model $\mathcal{G}^*$. The radius C of the sphere that contains all the exchangeable models of our interest, as in [110], is defined by the Kullback-Leibler between the true and the approximating model. Kullback-Leibler is a premetric and generates a topology on the space of probability distributions. So we define the closed ball with center $\mathcal{G}^*$ and radius C such that:

$$\Gamma(\mathcal{G}^*) = \{\mathcal{G} : KL(\mathcal{G}^*, \mathcal{G}) \leq C\} \tag{5.2}$$

Usually, one important example of a neighborhood defined by C is the ϵ-contamination neighborhood from [15] formed by the mixture model,

$$\Gamma = \{\mathcal{G}_{current} = (1-\epsilon)\mathcal{G}^* + \epsilon q, q \in Q\}, \tag{5.3}$$

where ϵ is the perceived contamination error, which in our case is provided by two moves, in $\mathcal{G}^*$ and $\mathcal{Q}$ is a class of contaminant distributions - perturbed versions of graphons (figure 5.2).

The next challenge is to explore and exploit that space. For any approximating exchangeable network model, assumed by the statistician/practitioner to fit the data (e.g. Erdös-Rényi or Stochastic Block Model), initially, we find its graphon (e.g. flat Erdös-Rényi, piecewise constant for Stochastic Block Model) and discretized it like in [5, 57, 80] by splitting the unit square in a large number (n^2 of equal squared blocks. In order to move into the sphere we need to shuffle and to change the heights of the grid. The grid is fixed, with as we mentioned $n \times n$ equal cells, which give the approximated SBM of a graphon. We follow two moves, which are described below, are the perturbing move and the rescaling move:

- In order to perturb the discretized graphon and move in the neighbors models inside the ball that are defined by KL between the approximate model and the perturbed models like in [2], we use a simplex like below:

$$\{K \in \mathbb{R} : K_{1,1} + \cdots + K_{n,n} = \alpha, K_{i,j} \geq 0, i, j = 0, \ldots, n\} \tag{5.4}$$

  Initially, parameter α above is 1 but changes its value less than 1 due to the second move. By changing the probabilities of K_i, we perturb each current model every time. (assumptions: same sized squares, graphon is symmetric and simplex). For this:

  Draw n^2 independent random samples $y_{1,1}, \ldots, y_{n,n}$ from Gamma distributions each with density:

$$\text{Gamma}(\alpha_{i,j}, 1) = \frac{y_{i,j}^{\alpha_{i,j}-1}\, e^{-y_{i,j}}}{\Gamma(\alpha_{i,j})} \tag{5.5}$$

  where , $\alpha_{i,j}$ which denotes the counts in each cell, and then set

$$K_{i,j} = \frac{y_{i,j}}{\sum_{i,j=1}^{n} y_{i,j}} \tag{5.6}$$

  If $y_{i,j}$ are independent $\text{Gamma}(\alpha_{i,j}, 1)$, for $i, j = 1, \ldots, n$ then:

$$(K_{1,1}, \ldots, K_{n,n}) = \left(\frac{y_{1,1}}{\sum_{i,j=1}^{n} y_{i,j}}, \ldots, \frac{y_{n,n}}{\sum_{i,j=1}^{n} y_{i,j}} \right) \sim \text{Dirichlet}(\alpha_{1,1}, \ldots, \alpha_{n,n})\,. \tag{5.7}$$

- Scaling the graphon: E.g. for the Erdös-Rényi model the initial graphon (the one which has equal degree density and be flat) is going to be equally

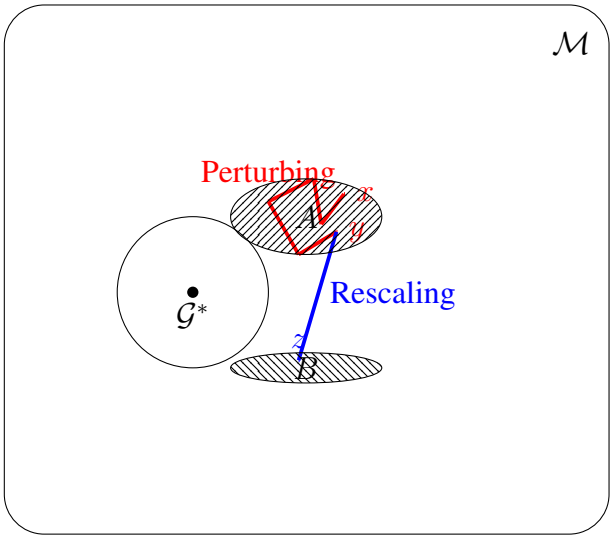

Figure 5.3: Exploring and exploiting the Models in the sphere. The red trajectory inside a subspace A illustrates the perturbing move which is created by changing the heights of two cells of the Stochastic Block Model. With the blue line, the rescaling move is illustrated, jumping to another subspace B of the sphere.

to one. Then to get the desired surface we multiply the SBM heights that approximate the graphon with ρ. When we scale it, it is no longer a graphon, it is a scaled graphon. In our case ρ is randomized with $1 \pm \delta$, where δ has a small value that varies.

Those two moves show we how different we are with respect to a constant and to scaling. The adjustment of the heights are coming from those two moves. We perturb a certain number of times and then we rescale and check if we the next model is included in the KL ball. Those two moves need to be inside the ball and satisfy $d(W1, W2) < C$ (distance metric e.g L^2 or L^∞ getting moves within this ball, for the computations).

The Kullback-Leibler between the graphon approximations of the two random graph models needs to be approximated:

$$KL(p \mid\mid q) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(x,y) log \frac{p(x,y)}{q(x,y)} dxdy \tag{5.8}$$

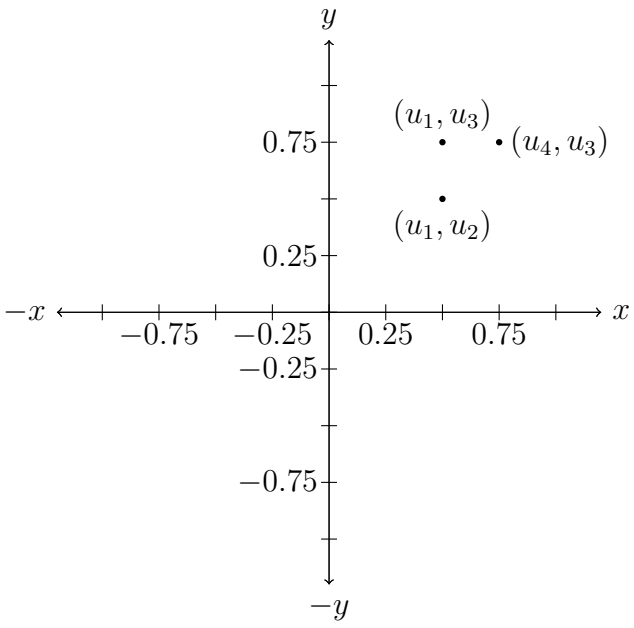

Figure 5.4: Latent positions projected in x,y-axis.

conditional on the latent positions.

One computational problem that arises in network data is regarding the entries of the adjacency matrix associated with u_1 and u_2 and the entries associated with u_1 and u_3 in figure 5.4. This happens because the latent positions (u_1, u_2) and (u_1, u_3) are dependent as they have the coordinate u_1 in common. It follows that the events of having an 0 or 1 on entries of a network adjacency matrix associated with those latent positions are dependent if we do not condition on the values of these latent positions. In that way, we average out these dependencies. The same happens with (u_1, u_3) and (u_4, u_3) because of u_3. However, this is not occurring with points (u_1, u_2) and (u_4, u_3) which are independent, since they do not have a coordinate in common (figure 5.4).

Algorithm 5 MC algorithm for approximating Kullback-Leibler

1. Draw a uniform and then draw a point pattern correspond to that uniform, by reflecting the points of x-axis to y-axis. This gives us a coordinate of point.
2. Since that grid is fixed essentially we count, take those points that they are occupied by the point pattern get the KL_i for those in the upper triangle and then add them.
3. Repeat 1 and 2 several times to get $KL_1, \ldots, KL_n$.
4. Average all KL results that we got and get an approximation of KL.

Algorithm 5 gets into account the dependence and integrated them out. When we average of all of this we get an approximation and we take into account the dependencies due to the point pattern because we have incorporated them in all simulation.The MC algorithm 5 gives an average of Kullback-Leibler divergences over their dependencies which is a good approximation. It preserves the structure and averages out where the possible latent positions are falling. Because we sample from those uniforms over and over again, the latent positions are not taken into account by averaging out.

We need to define an expected loss function (objective function), for a specific inference (predict the present of certain edges, density of the graph etc), which we will maximize with respect of expected loss by using Simulated annealing. This has to applied to Erdös-Rényi and Stochastic Block Model. Here Squared loss function is used:

$$E[(\hat{\tau}(\mathcal{G}) - \tau(\mathcal{G}))^2] = E[(\hat{\tau}(\mathcal{G}) - E[\hat{\tau}(\mathcal{G})])^2] + (E[\hat{\tau}(\mathcal{G})] - \tau(\mathcal{G}))^2 = \\ = V(\hat{\tau}(\mathcal{G})) + (E[\hat{\tau}(\mathcal{G})] - \tau(\mathcal{G}))^2 \quad (5.9)$$

The expected value has to be sample many times the realizations of networks of the surface. $\tau(\mathcal{G})$ is true value of the current model feature and $\hat{\tau}(\mathcal{G})$ is estimator of that feature.

Algorithm 6 Selecting T parameter for Simulated Annealing

1. Sample the sphere regarding the expected losses of the feature, combined with values of the sphere boundaries and approximated model $\mathcal{G}^*$.
2. Find the median of those samples.
3. Give T_0 this value.

Another challenge is to propose a method how to select the initial value of parameter T from the simulated annealing, the stochastic optimization technique we use. This value has to be calibrated according to the values of the Expected Losses of the feature. We select T_0 by algorithm 6. T_0 and the losses have to be in the same scale and in order to select T_0 we sample randomly the sphere.

5.3 Simulation Studies

This section is divided into two subsections. The objective of the first subsection is to show that the simulated annealing is reliable and the objective of the second section is to present actual results. For both subsections we consider the following: we have the true network model and the approximating exchangeable random network model. We fit the graphon to the generating model and sample the graphon to see what is the best approximation using a SBM, with certain number of blocks. That KL gives us an intuition of the size of the ball, by the radius C. To investigate the behavior of the proposed approach, we conducted three simulation studies. The objective of the simulations is to examine how harmful is the approximating exchangeable random network model when we use it to fit the data by the generating model for inferring a specific feature. The regimes of the simulation study are given by: the random graph model, the corresponding vector of parameters (Table 5.1) and the sample size.

To infer the generating model we use [80] approach. The graph features we considered were: the density of the networks and number of communities for an Erdös-Rényi model and two Stochastic Block Models. The first Stochastic Block Model (SBM_1) is selected a random Stochastic Block Model after 100 iterations of moves of Erdös-Rényi simulations. To infer second Stochastic Block Model (SBM_2) parameters we use [58] approach, by using the same notation for number of blocks (K), and inclusion probabilities (λ, ϵ). When we implement our

Approximating Exchangeable RGN	Parameter Specification	Features
ER	K=1, $N = 100$ (fig. 5)	Blocks and Density
SBM_1	Point of ER after 100 moves (fig. 6)	Blocks and Density
SBM_2	$\lambda = 0.5$, $\epsilon = 0.5$, $N = 100$	Blocks and Density

Table 5.1: Approximating Exchangeable Random graph models, parameter vectors and graph features considered for setting up simulation regimes.

method we consider 1000 samples of the desired parameters θ_i from the models inside the sphere and 100 networks instances given the value of the parameters.

5.3.1 Variability of Stochastic Optimization process and data sets

Our goal is to check whether the proposed approach reaches the worst case scenario (maximum expected loss). In other words, we want to investigate how different are the models from the actual worst case scenarios. For that reason, we explore the space of the balls by exhausting as many as possible models exploring the ball. The three settings mentioned above are considered. We examine the variability of the simulated annealing by running it 10 times and extracting the mean and the variance compared with the ground truth. This ground truth comes from a brute force algorithm which considers 1.000.000.000 models in the balls. In each setting the variability of the stochastic optimization process is conditional to the data. That is why we repeat this procedure for each of these different three settings considering the variability of the data sets, as well.

Approximating Exchangeable RGN for Density	Brute Force	Mean	Variance
ER	7.65817e-7	7.65828e-7	9.1742e-17
SBM_1	7.65817e-7	7.65831e-7	9.1765e-17
SBM_2	0.000929188	0.000929213	8.9826e-8

Table 5.2: Expected loss of worst case scenario with brute force (ground truth) compared with the mean of expected loss of the worst case scenario for the density of the three different models providing the variance of Expected loss of worst case scenario

Approximating Exchangeable RGN for Blocks	Brute Force	Mean	Variance
ER	50.2342	50.2748	0.00281
SBM_1	48.9342	48.8281	0.00301
SBM_2	1826.2342	1827.1977	1.28392

Table 5.3: Expected loss of worst case scenario with brute force (ground truth) compared with the mean of expected loss of the worst case scenario for the number of blocks of the three different models providing the variance of expected loss of worst case scenario.

As expected theoretically, Tables 5.2 and 5.3 confirm the reliability of our approach since the expected losses of the exhausting method is almost the same with the mean of the expected losses of our approach. The calculated variances are insignificant with very small values.

5.3.2 Results

Here, we present and illustrate the results of the proposed method for each one of the three different settings. As discussed in Section 3, our approach associates a score in each case. Such score is given by the maximum expected loss given by the inference to the feature, while exploring and exploiting the sphere-ball. Figures 5.5, 5.6 and 5.7 are the approximating models which are located in the center of the ball.

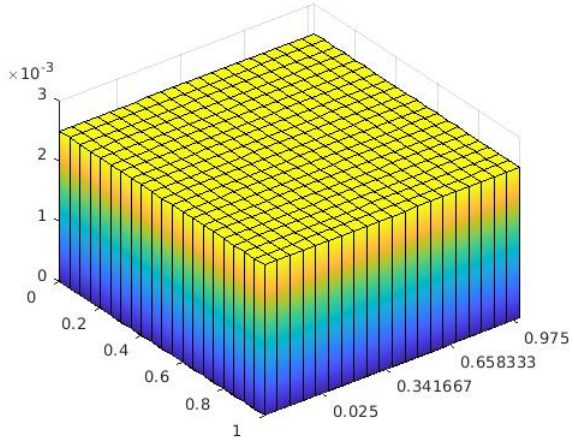

Figure 5.5: Erdös-Rényi model, 20 $\times$ 20 cells.

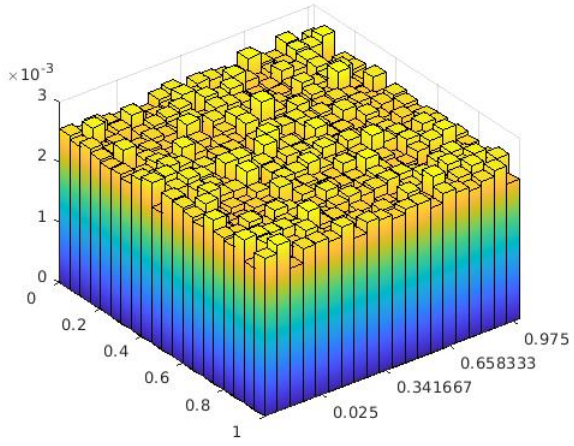

Figure 5.6: Stochastic Block model, 20 $\times$ 20 cells, after applying the moves. This stochastic block model is used as an approximating model SBM_1 for the second experimental design, as well.

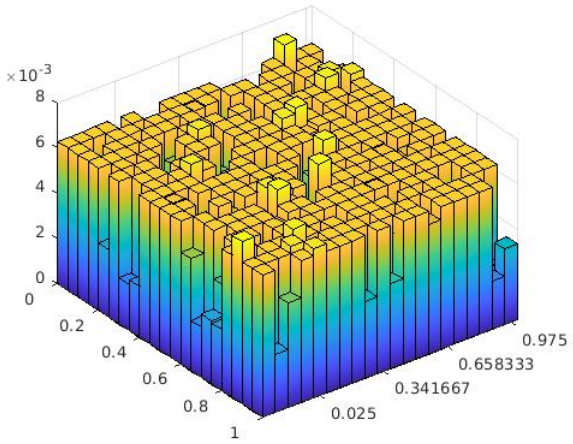

Figure 5.7: True approximation of the graphon represented by a Stochastic Block model.

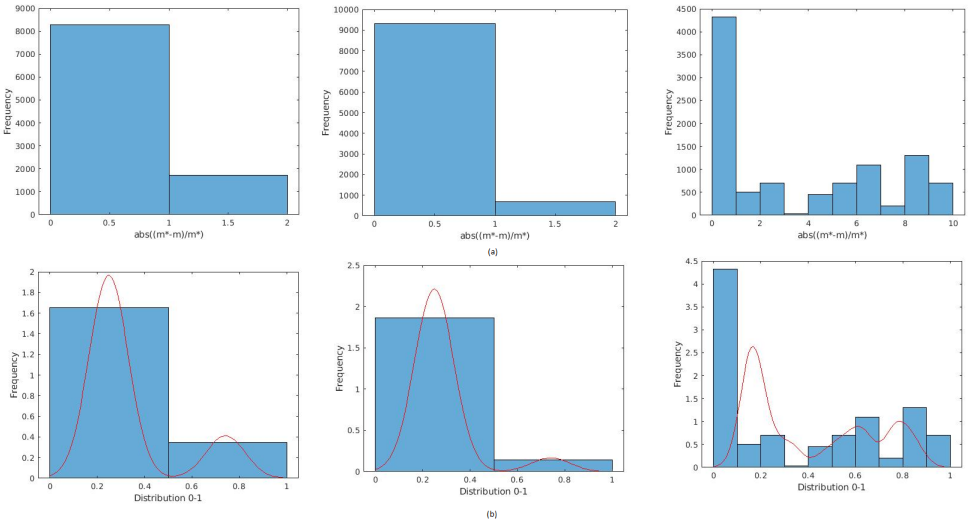

Figure 5.8: (a) Values of frequencies (Expected Loss (model)-Expected Loss(center model))/Expected Loss(center model) for the density. (b) Values of frequencies for the scaled distribution between 0 and 1 in x-axis.

Approximating RNM ($\mathcal{G}^*$)	Radius C	A	B
ER	0.921	1.87	1.98
SBM_1	0.887	1.56	1.75
SBM_2	1.192	9.43	14.87

Table 5.5: Results for three models. Radius C given by Kullback-Leibler divergence is given and the maximum expected loss for one Erdös-Rényi model and two different Stochastic block models are presented. The fist two model are reasonable to be fitted but the last Stochastic block model is very robust in terms of inference for Density and Number of Communities.

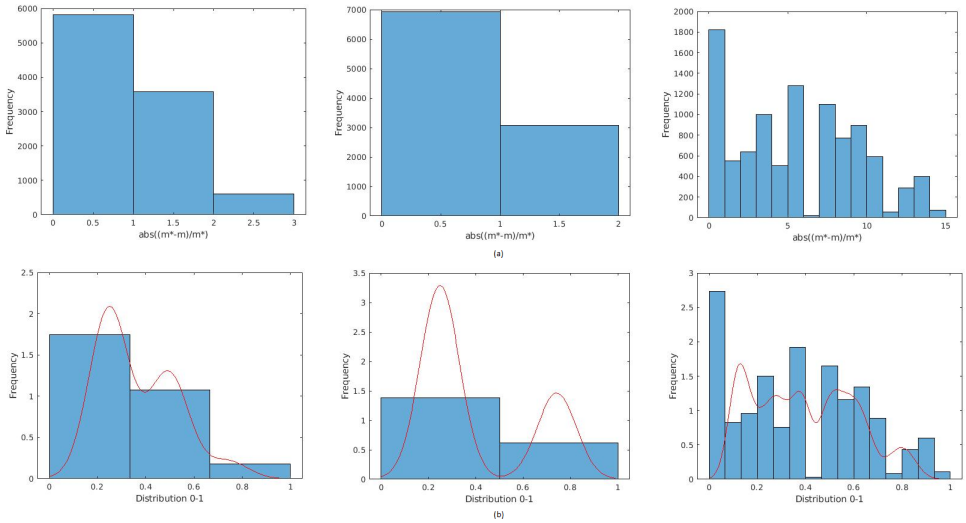

Figure 5.9: (a) Values of frequencies (Expected Loss (model)-Expected Loss(center model))/Expected Loss(center model) for the community blocks. (b) Values of frequencies for the scaled distribution between 0 and 1 in x-axis. The gap in the right figures are due to the nature of the random networks and their features.

In figure 5.8 and 5.9 we criticize the approximating models regarding their density and community blocks. The first two models behave well and capture well the interested features regarding the true data. This can be shown by the fact that the mass of the distribution is centered and compact in low values

so all the models and the approximating models do not differentiate too much regarding those features in that ball C. On the other hand, the third model does not behave well. This can be shown by the fact that its values are dispersed and the majority of the models in the ball are completely different. A slight perturbation in the model might end up with a huge perturbation in expected value quantities A and B as defined in table 5.4. Finally, all the results are shown in table 5.5, confirming the above criticism.

5.4 Discussion

In this chapter, we present a new principle approach for robustness for inferences on exchangeable random network models, based on graphons. To the best of our knowledge, this is one of the first efforts, in the literature, graphons are used to deal with practical problems. The main advantage of our method is that it provides the statistician with a conceptual framework that enables him/her to check whether an specific inference of an approximating model is robust with respect to model misspecification regarding the generating model. The results suggest that our method gives reasonable answers when we compare different kind of SBMs to examine how well they are performing.

Three limitations of our work are the following: The first limitation is that we do not provide a general way to represent a flexible way to perturb non-exchangeable models. Here, we used the power of graphons. Graphon are able to give us a general way to perturb them. If we fit a SBM approximation of a graphon in non-exchangeable random network model there might be a misspecification on the model. Until now, there is no way to characterize this misspecification. Therefore, to perform a similar method for non exchangeable models we need a way to define and to move to the space of models that take into account non-exchangeability. Secondly, we are dealing with the biasness of estimators combining network models and network realizations for specific features e.g. the average distance. We need good estimators to calculate the generating model specific feature of the random network vs the properties of the network in order to estimate the expected loss. Otherwise, there might be a misspecification on the model. Finally, the third limitation is that we do provide a methodology to express and perturb

the graphon and the model we are interested in explicitly with an analytical expression. In contrast, we resort to computational and approximation approaches which come with a (small) error.

For future work we propose the following three directions: The first two have to do with the limitations mentioned above. First, we need to find estimators that can be useful to connect random networks with networks in terms of features and properties, respectively. The second direction is extending our method to non-exchangeable models. One way is to define and use a very flexible and hard to interpret family of models and compute the distance with the approximating model [72]. Since, there are not such non exchangeable models we have to resort to another space e.g. to move from a parametric space to a non-parametric space, maybe observable space and then apply [72] techniques. The last direction is to develop a method that can be applied to networks and (non) exchangeable random network models adopting the robustness setting in [82].

For the last two directions, one challenge is that the neighborhood is centered at the model we try to fit and specify and we assume it is correct. We are provided by the size of the neighborhood, in which we are going to find the worst case scenario. One of the challenges is what should this size of this ball be? In [110], the authors do not provide any answer. If we have a goodness of fit of Bayesian Models, how reasonable is my model? Combining the logic of [72] can provide us with a sense of how far is the data generating mechanism from the model we want to use in the first place and interpret.

Chapter 6

Summary, discussion, and future work

Here we provide a summary of the contributions of the thesis (Chapters 3-5). We then present a brief discussion of these and other prominent challenges in network modeling, along with possible avenues for future work (thereby extending Chapter 3-5).

6.1 Summary of our contributions

In this thesis, we contributed to the field of statistical inference for networks by proposing Bayesian methods to improve our understanding of some widely used network models regarding their behavior of their features. We developed three different framework that enables us to investigate different practical aspects of network models.

In chapter 3, our objective was to develop a new method for comparing sampling design on network data by using information theory. We make the case that different designs are more suitable for different random network models. At the end of the day, a practitioner can use our framework to select sampling designs which are more informative for the random networks of his choice. Regarding the theory of this section, we associated the problem with the rationality behind the statistical interpretation of the reference priors. As a consequence, it is showed that following the same assumptions of the reference priors, a framework for comparing sampling designs on network data

can be derived. We examined the usefulness of this approach by simulation studies and the results indicate that the ranking of many different sampling designs is consistent with decision theory approach without the need of specifying a loss function [7].

Subsequently, in the second part of this thesis we used the Bayesian algorithm regarding sampling designs from section 3 to infer reasonable statements that combine both random networks, their features and sampling designs. More specifically, we provided valid and generic ways to translate statements for partially observed networks to fully observed networks and via versa. The most interesting thing we want to investigate is to theoretically and practically understand what information is required to enable us to use statements at the level of partially observed networks and turn them into statements for fully observed networks that they produced them. In the same spirit other coarsening techniques [45] can be used. In that context our goal was to investigate to what extend a statement that we can make for a partially observed network can be translated to a statement to a fully observed network. We proved that in the general case the answer is that the joint distribution of their features and the sampling designs is required in order to create relevant statements. In the simulation studies we used Erdös-Rényi model, degree and transitivity which leaded us with two detailed and intuitive examples of how two features are related when a sampling design is involved. We encountered a case in which conditional distributions regarding network data feature and sampling designs provided enough information, compared with other cases in the literature, to build such statements confirming our point.

In the third project, we adopted the statistical framework of [110] to provide tools to the modeler to evaluate how the quality of inference for a specific feature of a random network model is degraded when the approximating model is misspecified. We tried to answer how sensitive could be the quality of an inference be when the data is not coming exactly from the true exchangeable model. More specifically, we provided methodology to examine whether and how much an approximating random network model is suitable for describing a true random network model in terms of a specific feature. In terms of methodology, our main challenge was to combine stochastic optimization and graph limits tools to explore the model space. To test the effectiveness of our approach we performed simulation studies that

show how harmful is the fact that a model is not coming from the true model.

6.2 Discussion

Outlining what we have learned in light of theory, experiments and simulation studies we can state the following:

- The proposed approach in chapter 3 which is based on information theory provides a ranking of sampling design which is consistent with the results that are obtained with some of the most important loss function for estimation and prediction in Bayesian inference.
- In chapter 4, we provided the practitioners with the theory concerning how to develop and connect statements which involves uncertainty on random network models, features and sampling designs (inline with coarsening data in [45]). Those statements concern the features of fully and partially observed networks and provide the practitioner with the intuition of how those features are related and whether they provide useful information when working with those networks.
- We casted the paper of [110] on robustness for exchangeable random network models. The main advantage of our method is that it provides the statistician with a conceptual framework that enables him/her to check whether an specific inference of an approximating model is reasonable regarding the generating model. The results suggest that our method gives reasonable answers when we compare different kind of SBMs to examine how well they are performing.

6.3 Extension and future work

6.3.1 Network models with higher dimensionality

In our case the curse of dimensionality is reflected in chapters 3-4. The MCMC algorithm we use is slow but can be easily scaled out (appendix). Though, new faster algorithms based on other MCMC implementations or variational methods can be constructed. Generally, the curse of dimensionality is a big issue in many computational problems. Due to technological

advances we are able to collect data that are increasingly large and diverse in structure. To fully exploit these rich data, there is a strong need for network models to catch up in their dimensionality, and for us to derive the asymptotic properties of these models such that we can deliver statistical guarantees. In contrast to classical statistics, the observations may neither be identical nor independent; and there is no natural ordering inherited in the data and no means of geometry, as is the case for time series or spatial statistics. As a result, a key challenge here is to introduce high-dimensional models that reflect the unique structure inherited in networks.

6.3.2 Imperfectly observed networks

Broad topics around imperfectly-observed networks have been studied from many different viewpoints e.g. an overview can be gleaned from the talks at the workshop [113]. However, there is no probability model involved; different algorithms are compared experimentally by taking a real-world network, randomly deleting a proportion of edges to create a synthetic observed graph, and comparing the algorithms effectiveness in predicting the deleted edges. Clearly, their research area of identifying specific useful expressions, correlations and connections regarding the features of partially and observed networks is fruitful. Three possible future direction are:

- Construct faster computational algorithms.
- Propose computational algorithms for specific circumstances (e.g. particular feature) looking for informative ways at a level of underlying network.
- Investigate for analytical solutions when possible.

6.3.3 Exchangeability on Networks

A fundamental question in modern network statistical science is how we can escape exchangeability. A reasonable and desirable property of generative models for networks is that they should not depend on the order in which we observe data, i.e., that our models are exchangeable. Network data, presented as a (random) adjacency matrix, requires joint exchangeability, so row and column identities in the matrix are jointly preserved under permutations

of the data.

Exchangeability as a requirement leads to representation theorems. For exchangeable sequences, de Finettis theorem implies that they can be represented by an underlying i.i.d. mixture of random variables. From this perspective, exchangeability for network data implies the existence of a latent variable generative model, and furthermore that in such a model, edges are conditionally independent. That is, exchangeability of a network model implies that we are in the regime of dense networks. This is problematic because most real-world networks, in fact, are sparse (vertices have constant or very slowing-growing degree; graphs have $O(n)$ edges).

Most of the models discussed in our walking tour fall within the jointly exchangeable framework of Aldous-Hoover. This would seem to imply that all generative models for networks are misspecified, despite their many successful practical applications. One solution to this fundamental problem is to abandon exchangeability in favor of alternative properties, or to attempt to escape the Aldous-Hoover representation entirely. While networks enable us to model complex dependencies between entities, the lack of techniques to model non exchangeable network models makes the results unreliable.

Here we can do not consider exchangeable models. The ball is constructed in a required way. Other constructions apart/alternatives from the graphon construction. For example, in [25] they consider exchangeable random measures. How this ball is going to be built for theoretical and computational-probably scalable-reasons? Another direction is find alternative ways to determine the radius.

Supplementary material for Chapter 3

Scalability

Procedures in figures 3.5, 3.6 and 3.7 are scalable in terms of samples we are collecting either from the fixed parameter or from ground truth distribution. The algorithms are computationally expensive even for random matrices with 100 nodes. Figure 6.1 shows how samples can be combined together taking into advantage the map-reduce scheme to decrease the time of the computa-

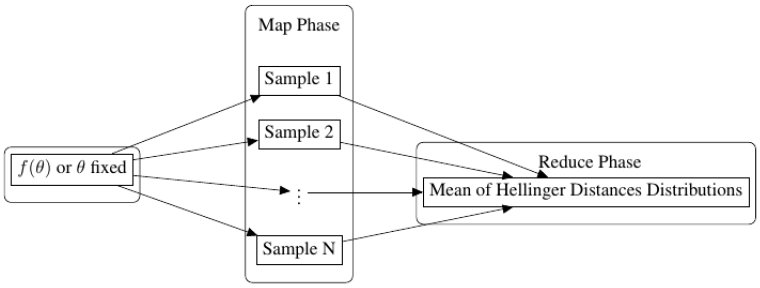

Figure 6.1: Parallelizing Figures 3.2, 3.3 and 3.4, using a cluster with N units in Map phase.

tions.

Moves for Non-Ignorable designs

Let $\mathcal{V}_{\mathcal{INC}}$ and $\mathcal{V}_{\mathcal{EXC}}$ the set nodes of $\mathcal{G}_{\mathcal{INC}}$ and its complement. Denote by A_H the submatrix of A_G (the adjacency matrix for G) obtained by taking only the rows associated to elements of $\mathcal{V}_{\mathcal{INC}}$. We denote by b_O the number of zeros in submatrix of A_G obtained by taking the rows associated to $\mathcal{V}_{\mathcal{INC}}$ and the columns associated to $\mathcal{V}_{\mathcal{EXC}}$. Denote by j_I the number of edges included in $\mathcal{G}_{\mathcal{INC}}$.

To sample from $p(\theta, \mathcal{G}_{\mathcal{EXC}} \mid \mathcal{G}_{\mathcal{INC}})$ we implemented a Metropolis step based on a mixture of four kernels.

The first kernel corresponds to a proposal where a vertex $v \in \mathcal{V}_{\mathcal{INC}}$ is picked uniformly at random, and all the edges connecting it to elements of $\mathcal{V}_{\mathcal{EXC}}$ are re-wired, so the number of neighbors of v belonging to $\mathcal{V}_{\mathcal{EXC}}$ remains constant. The Metropolis ratio implied by these choices has the form

$$H^{(t)} = \frac{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t)}_{EXC})}{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t-1)}_{EXC})} \tag{6.1}$$

where $p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{()}_{EXC})$ is the value of that results from using the imputation of the network implied by $G^{()}_{EXC}$. Note that, this move keeps the

number of edges constant, therefore the terms corresponding to the Erdös-Rényi probability mass function cancel out. Clearly, the move that reverses the proposed one implies picking the same $v \in \mathcal{V}_{\mathcal{INC}}$. By conditioning on this event, the proposal becomes a uniform over the subsets of $\mathcal{V}_{\mathcal{EXC}}$ that have as many elements as v has neighbors in $\mathcal{V}_{\mathcal{EXC}}$. This last statement implies that the terms corresponding to the proposal also cancel out.

The second and third kernels should be seen as dual: the second kernel corresponds to the proposal where an edge connecting to vertices $(v, w) \in \mathcal{V}_{\mathcal{INC}} \times \mathcal{V}_{\mathcal{INC}}$ is chosen uniformly at random and then substituted by two edges, each of them connecting a different element of $\{v, w\}$ with an element of $\mathcal{V}_{\mathcal{EXC}}$ (not necessarily the same one) picked uniformly at random. The third kernel allows for the opposite move: it takes two vertices $(v, w) \in \mathcal{V}_{\mathcal{INC}} \times \mathcal{V}_{\mathcal{INC}}$ such that, each of them has at least one edge connecting it to an element of $\mathcal{V}_{\mathcal{EXC}}$, then, two of such edges are chosen (one incident to v and one incident to w) uniformly at random and then replaced by an edge connecting v and w. Let $h_I^{()}$ be the number of edges of the form $(v, w) \in \mathcal{V}_{\mathcal{INC}} \times \mathcal{V}_{\mathcal{INC}}$ that are in the current version of the network due to imputation (i.e., these edges were not observed). The Metropolis ratio corresponding to the second kernel is of the form

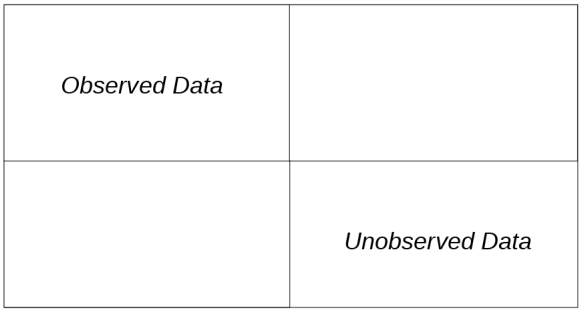

Figure 6.2: We arrange the unobserved data with the information we have from the observed data.

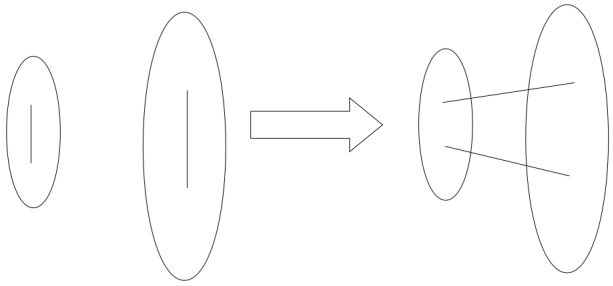

Figure 6.3: We add two nodes that link the observed and unobserved nodes and connect the two observed nodes.

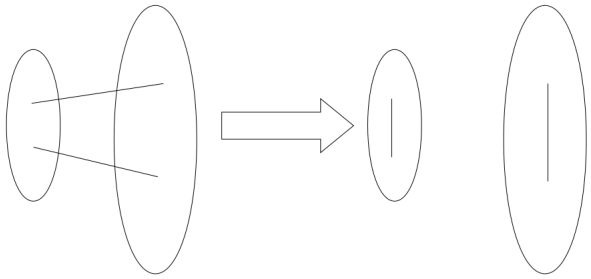

Figure 6.4: We remove one edge from two observed nodes and we connect those two nodes with another two unobserved nodes

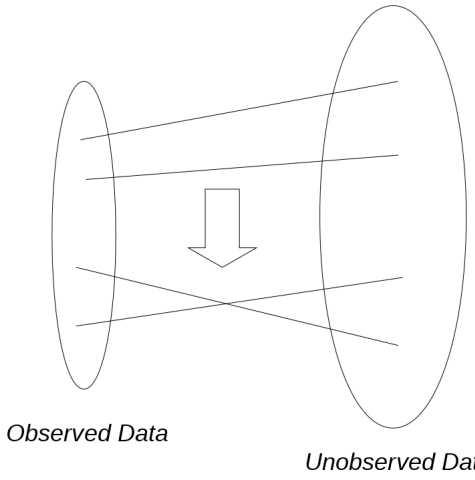

Figure 6.5: We have two observed nodes connected with two unobserved and we rewire them.

$$H^{(t)} = \frac{\alpha}{1-\alpha} \times \frac{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t)}_{EXC})}{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t-1)}_{EXC})} \times \frac{h_I^{(t)} \binom{b_O^{(t)}}{2}}{(\frac{n(n-1)}{2} - j_i - h_I^{(t)} + 1)\binom{(n(N-n)-b_O^{(t)}+2)}{2}} \tag{6.2}$$

while the Metropolis ratio for the third kernel is given by

$$H^{(t)} = \frac{1-\alpha}{\alpha} \times \frac{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t)}_{EXC})}{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t-1)}_{EXC})} \times \frac{(\frac{n(n-1)}{2} - j_i - h_I^{(t)})\binom{(n(N-n)-b_O^{(t)})}{2}}{(h_I^{(t)} + 1)\binom{b_O^{(t)}+2}{2}}$$

(6.3)

The fourth kernel corresponds to the proposal where the submatrix of A_G with rows and columns associated to V_{EXC} is imputed using independent draws from a Bernoulli with probability of success α. This proposal implies the Metropolis ratio

Random Graph Model	Parameter Specification	Features
Erdös-Rényi	$\theta = 0.2$, $N = 100$	θ and Trans.

Table 6.1: Random graph models, parameter vectors and graph features considered for setting up simulation regimes.

$$H^{(t)} = \frac{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t)}_{EXC})}{p(I \mid \mathcal{G}_{\mathcal{INC}}, G^{(t-1)}_{EXC})} \tag{6.4}$$

since the terms corresponding to the proposal and those corresponding to the random graph distribution (given α) cancel out.

Erdös-Rényi model

Here, we use another, more simpler, random graph model to illustrate our method:

- The Erdös-Rényi model:

$$\Pr\{A_{\mathcal{G}}(i,j) = 1\} = p, \quad p \in (0,1)$$

The graph features $\tau(\mathcal{G})$ we considered were: density and transitivity for Erdös-Rényi:

Ignorability

To perform Bayesian inference, it is necessary to specify the likelihood correctly. The concept of ignorability [**?**,]]Rubin helps on this task by providing criteria for deciding if the uncertainty due to the sampling mechanism needs to be modeled explicitly in the likelihood. Let $p(\mathcal{G})$ denote the distribution of the full network data. We follow the convention by [**?**,]]Rubin and write the joint distribution of $(\mathcal{G}, I)$ as

$$p(\mathcal{G}, I, \eta) = p(\mathcal{G})p(I \mid \mathcal{G}, \eta), \tag{6.5}$$

where η represents the vector of tuning parameters of the sampling mechanism, which can be specified by the statistician.

A sampling design I is ignorable if:

$$p(I \mid \mathcal{G}, \eta) = p(I \mid \mathcal{G}_{\mathcal{INC}}, \eta), \tag{6.6}$$

where η are the parameters for the sampling design and the full data τ are distinct.

If a sampling design is ignorable, then the term corresponding to the distribution of I is omitted from the likelihood:

$$p(\mathcal{G} \mid \theta) \propto \int_{\mathcal{G}_{\mathcal{EXC}}} p(\mathcal{G}_{\mathcal{INC}}, \mathcal{G}_{\mathcal{EXC}} \mid \theta) d\mathcal{G}_{\mathcal{EXC}}. \tag{6.7}$$

A sample mechanism that does not fulfill the definition above is called non-ignorable. A consequence of a sampling design being non-ignorable is that the likelihood is not constant with respect to missing data, and therefore, it has to be imputed in a way that reflects the changes in the values for the likelihood:

$$p(\mathcal{G} \mid \theta) \propto \int_{\mathcal{G}_{\mathcal{EXC}}} p(I \mid \mathcal{G}_{\mathcal{INC}}, \mathcal{G}_{\mathcal{EXC}}) p(\mathcal{G}_{\mathcal{INC}}, \mathcal{G}_{\mathcal{EXC}} \mid \theta) d\mathcal{G}_{\mathcal{EXC}}. \tag{6.8}$$

In the first case the part $p(I \mid \mathcal{G}_{\mathcal{INC}}, \mathcal{G}_{\mathcal{EXC}})$ becomes constant with respect to $\mathcal{G}_{\mathcal{EXC}}$.

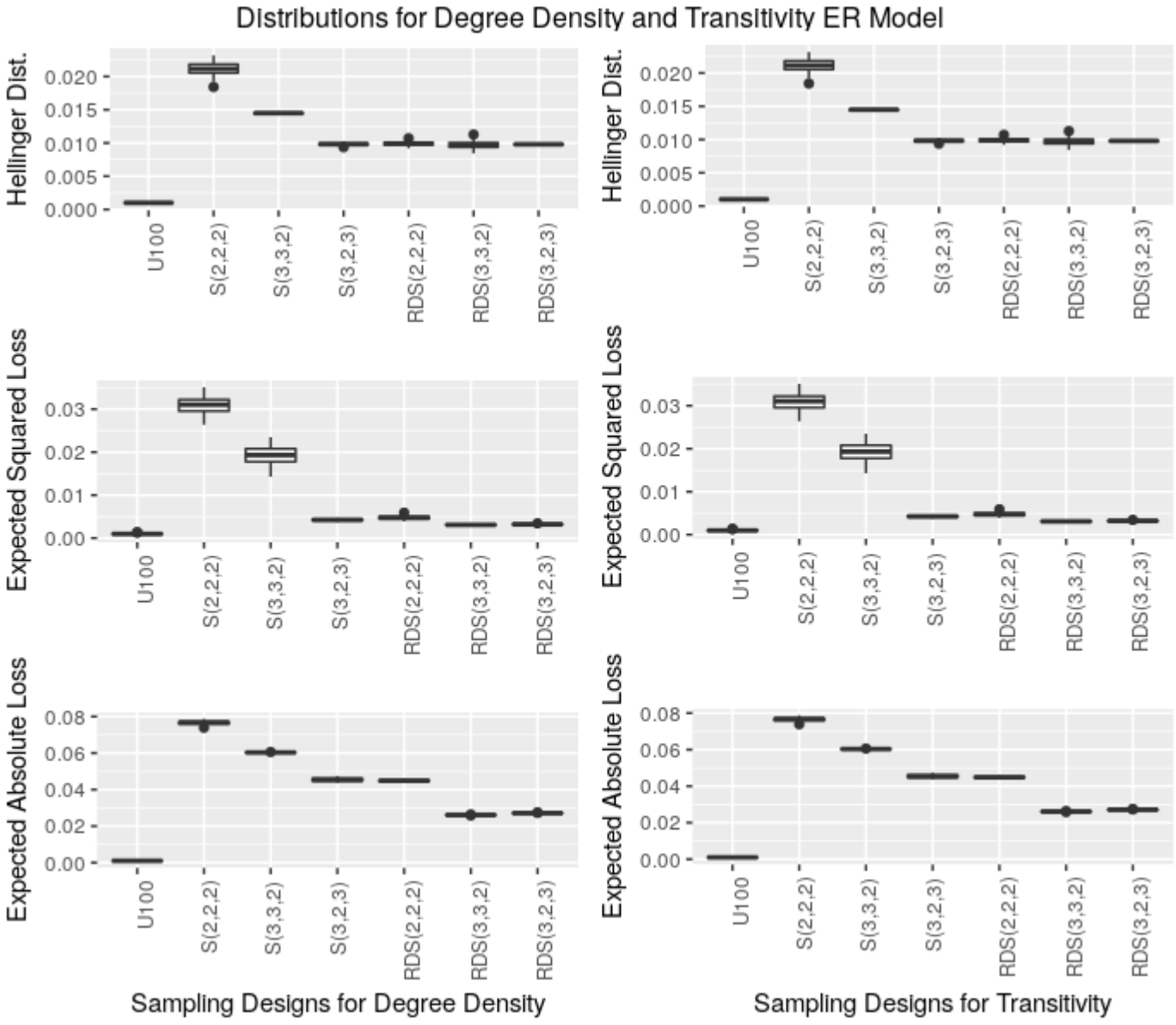

Figure 6.6: Upper: Comparison of sampling designs regarding the mean of the Hellinger distance distribution regarding degree density and transitivity. Middle: Comparison of sampling designs regarding the mean of the expected squared loss distribution of the predictive distribution, $\mathbb{E}_\theta(\theta - \hat{\theta})^2$ where θ is the either the degree density or the transitivity. Down: Comparison of sampling designs regarding the mean of the expected absolute loss distribution of the predictive distribution, $\mathbb{E}_\theta(\mid \theta - \hat{\theta} \mid)$ where θ is the either the degree density or the transitivity.

Model	Feature	SD	MHD	MSE (P.P.)	MAE (P.P.)	MSE (P.)	MAE (P.)
ER	θ	S (2,2,2)	0.0212	0.0306	0.0766	0.0308	0.1758
ER	θ	S (3,3,2)	0.0095	0.0193	0.0403	0.0189	0.1372
ER	θ	S (3,2,3)	0.0040	0.0072	0.0274	0.0071	0.0845
ER	θ	RDS (2,2,2)	0.0099	0.0048	0.1498	0.0021	0.0449
ER	θ	RDS (3,3,2)	0.0097	0.0031	0.0865	0.0007	0.0261
ER	θ	RDS (3,2,3)	0.0098	0.0032	0.0935	0.0017	0.0409
ER	Trans.	S (2,2,2)	0.0215	0.0314	0.0743	-	-
ER	Trans.	S (3,3,2)	0.0093	0.0188	0.0416	-	-
ER	Trans.	S (3,2,3)	0.0041	0.0073	0.0272	-	-
ER	Trans.	RDS (2,2,2)	0.0099	0.0048	0.1502	-	-
ER	Trans.	RDS (3,3,2)	0.0093	0.0032	0.0878	-	-
ER	Trans.	RDS (3,2,3)	0.0095	0.0038	0.0965	-	-

Table 6.2: Means of Hellinger Distances Distribution (MHD) and means of Predictive Posterior (P.P), for point prediction, and Posterior (P.) Quadratic and Absolute Mean Distribution (MSE and MAE), for point estimation, for six different sampling designs in the settings of degree density and transitivity on Erdös-Rényi model.

Bibliography

9

[1] Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. Systemic Risk and Stability in Financial Networks. American Economic Review **2015**, 105(2): 564608.

[2] Nesreed K.Ahmed, Jennifer Neville and Ramana Kompella. Network Sampling: From Static to Streaming Graphs. Journal ACM Transactions on Knowledge Discovery from Data (TKDD), Volume 8 Issue 2, Article No. 7 **2013**.

[3] E. M. Airoldi, T. B. Costa, and S. H. Chan, Stochastic blockmodel approximation of a graphon: Theory and consistent estimation, Advances in Neural Information Processing Systems. **2013**.

[4] Edoardo M. Airoldi, David M. Blei, Stephen E. Fienberg and Eric P. Xing. Mixed Membership Stochastic Blockmodels. Journal of Machine Learning Research, 9(Sep):1981–2014, **2008**.

[5] Edoardo Airoldi, David Blei, Eric Xing and Stephen Fienberg. A Latent Mixed Membership Model for Relational Data. Proceeding LinkKDD '05 Proceedings of the 3rd international workshop on Link discovery Pages 82-89 **2005**.

[6] Edoardo M. Airoldi, Simon Lunagomez. Valid inference from non-ignorable network sampling mechanisms, **2014**, arXiv no. 1401.4718.

[7] Edoardo M. Airoldi, Simon Lunagomez. Sampling on Social Networks from a Decision Theoretic Perspective (In preparation). **2016**.

[8] Anderson, C. J., Wasserman, S. and Faust, K. Building stochastic blockmodels. Social Networks, 14, 137161. **2012**.

[9] David Aldous, Li, X. A framework for imperfectly observed networks. **2017**.

[10] David J. Aldous. Exchangeability and related topics. In Ecole d et e de probabilites de Saint-Flour, XIII-1983, volume 1117 of Lecture Notes in Math., pages 1-198. Springer, Berlin. **1985**.

[11] David J. Aldous, Representations for partially exchangeable arrays of random variables. J. Multivariate Anal., 11(4):581-598. **1981**.

[12] Aronow, P.M. and Crawford, F.W. Nonparametric Identification for Respondent-Driven Sampling, Statistics and Probability Letters, 106, 100–102. **2015**.

[13] Barabsi, Albert-Lszl; Albert, Rka. Emergence of scaling in random networks. (October **1999**).

[14] Baraff A., McCormick T. and Raftery, A.E. Estimating uncertainty in respondent-driven sampling using a tree bootstrap method, Proceedings of the National Academy of Sciences, Vol. 113, No. 51. **2005**.

[15] Berger, J. and Berliner, L. M. Robust Bayes and empirical Bayes analysis with ϵ-contaminated priors. Ann. Statist. 14 461-486. **1986**.

[16] James O. Berger. Statistical Decision Theory and Bayesian Analysis, Springer Series in Statistics, 2nd Edition, **1985**.

[17] James O. Berger and Luis R. Pericchi. The Intrinsic Bayes Factor for Model Selection and Prediction. Journal of the American Statistical Association Vol. 91, No. 433 (Mar., **1996**), pp. 109-122.

[18] Berger, J. and Bernardo, J. On the developments of Reference Priors. Oxford university press, Bayesian Statistics, 4, 35-60. **1992**.

[19] Berger, J. and Bernardo, J. Tutorials on Objective Bayesian Analysis.**2005**.

[20] Bernardo, J.M. and Smith, A.F.M. Bayesian Theory. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley and Sons. **1994**.

[21] Blitzstein, J. and Nesterko, S. Bias-variance and breath-depth tradeoffs in respondent-driven sampling, J R Stat Soc Ser A Stat Soc. 2015 Jan; 178(1): 241269. May **2012**.

[22] Tamara Broderick, Diana Cai. Edge-exchangeable graphs and sparsity. Advances in Neural Information Processing Systems 29 (NIPS 2016).**2016**.

[23] Alberto Caimo, Nial Friel. Bayesian model selection for exponential random graph models, HHS Author Manuscripts, Stat Interface. 6(4): 559576.**2012**.

[24] Trevor Campbell, Diana Cai, Tamara Broderick. Exchangeable Trait Allocations. Electronic Journal of Statistics Vol. 12 (2018) 22902322 ISSN: 1935-7524 **2016**.

[25] Francois Caron Emily B. Fox. Sparse graphs using exchangeable random measures. Journal of the Royal Statistical Society: Series B (Statistical Methodology) Volume 79, Issue 5. **2017**.

[26] Chaloner, K. and Verdinelli, I. Bayesian experimental design: A review. Statistical Science, 10, 273–304. **1995**.

[27] Harry Crane, Walter Dempsey. Edge exchangeable models for network data, Journal of the American Statistical Association Volume 113, 2018 - Issue 523**2016**.

[28] Forrest W. Crawford. The Graphical Structure of Respondent-driven Sampling. Sociological Methology First Published April 25, **2016**.

[29] Persi Diaconis and Svante Janson. Graph limits and exchangeable random graphs, January **2008**.

[30] Brian Dixon. Health Information Exchange: Navigating and Managing a Network of Health Information Systems. **2017**.

[31] Paul Erdös and A. Rényi. The evolution of random graphs. Magyar Tud. Akad. Mat. Kutato Int. Kolz, 5:17-61, **1960**.

[32] P. Erdös and A. Rényi. On random graphs, i. Publicationes Mathematicae (Debrecen), 6:290297, **1959**.

[33] Flavio Fröhlich. Network Neuroscience. **2016**

[34] Fienberg, S. E.; Wasserman, S. (1981). Discussion of An Exponential Family of Probability Distributions for Directed Graphs by Holland and Leinhardt. Journal of the American Statistical Association. 76: 5457. 1981.

[35] Nikolaos Fountoulakis, Anna Huber, Konstantinos Panagiotou. The Speed of Broadcasting in Random Networks: Density Does Not Matter.**2009**.

[36] E.N. Gilbert. Random graphs. The Annals of Mathematical Statistics, 30(4):11411144, **1959**.

[37] Mark S. Handcock and Krista J. Gile. Modelling social networks from sampled data. Ann Appl Stat., 4(1): 525. **2010**.

[38] Gile, K. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. Journal of the American Statistical Association, 106, 135–146. **2011**.

[39] Krista J. Gile, Lisa G. Johnston and Matthew J. Salganik. Diagnostics for respondent-driven sampling. August **2013**.

[40] Mark S. Handcock, Krista J. Gile. On the Concept of Snowball Sampling. August 2, **2011**.

[41] Krista J. Gile. Respondent-Driven Sampling: An Assessment of Current Methodology. Sociol Methodol. **2010** August ; 40(1): 285327.

[42] Grunwald, P. and van Ommen, T. Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. arXiv preprint arXiv:1412.3730. **2014**.

[43] Hansen, L.P. and Sargent, T.J. Robustness. Princeton university press.**2008**.

[44] Heckathorn, D. Respondent-driven sampling: A new approach to the study of hidden populations. Social Problems, 44, 174–199. **1997**.

[45] Heitjan, D.F. and Rubin D.B. Ignorability and Coarse Data. Annals of Statistics, Vol 19, No. 4, 2244-2253. **1991**.

[46] Hoff, P.D., Raftery, A.E., Handcock, M.S., **2002**. Latent space approaches to social network analysis. Journal of the American Statistical Association 97, 10901098.

[47] Holland, P., Laskey, K. B., Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. Social Networks, 5, 109137.

[48] D. N. Hoover. Relations on probability spaces and arrays of random variables. Preprint, Institute for Advanced Study, Princeton, NJ, **1979**.

[49] Ivona Bezakova, Adam Kalai, Rahul Santhanam. Graph Model Selection using Maximum Likelihood. ICML '06 Proceedings of the 23rd international conference on Machine learning Pages 105-112 **2006**.

[50] Kadane, J.B. and Srinivasan, C. Discussion of Berger, J.O. An overview of robust Bayesian analysis with discussion. Test, 3(1), 116120. **1994**.

[51] Olav Kallenberg, Probabilistic Symmetries and Invariance Principles, **2005**.

[52] Karrer, B. and Newman, M. E. J. Stochastic blockmodels and community structure in networks. Phys. Rev. E 83 016107. **2011**.

[53] Eric D. Kolaczyk. Statistical Analysis of Network Data: Methods and Models (Springer Series in Statistics) **2009**.

[54] Mark A. Kramer. A brief primer on networks in neuroscience.

[55] Pierre Latouche, Etienne Birmel, and Christophe Ambroise. Model selection in overlapping stochastic block models. Electronic Journal of Statistics, Volume 8, Number 1 (**2014**), 762-794.

[56] Latouche, P., Birmele, E. and Ambroise, C. Variational bayesian inference and complexity control for stochastic block models. Statistical Modelling, 12, 93115. **2012**.

[57] Pierre Latouche, Stephane Robin. Variational Bayes Model Averaging for Graphon Functions and Motif Frequencies Inference in W-graph Models. **2015**.

[58] Pierre Latouche, Etienne Birmele, Christophe Ambroise. Variational Bayesian Inference and Complexity Control for Stochastic Block Models, **2010**.

[59] Steffen Lauritzen. Exchangeable Matrices and Random Networks, University of Oxford, **2013**.

[60] Steffen L. Lauritzen. Exchangeable Rasch Matrices, University of Oxford, September 17, **2007**.

[61] Lindley, D. Bayesian Statistics: A Review. SIAM. **1972**.

[62] L. Lovasz. Large networks and graph limits, volume 60 of Amer. Math. Soc. Colloq. Publ. Amer. Math. Soc., Providence, RI, **2012**.

[63] Lunagómez, S. and Airoldi, E. Valid inference from non-ignorable network sampling designs. **2016**.

[64] Pierre-Andre G. Maugis, Sofia C. Olhede Patrick J. Wolfe. Topology reveals universal features for network comparison. **2018**.

[65] Maccheroni, Fabio, Massimo Marinacci, and Aldo Rustichini. Ambiguity Aversion, Robustness, and the Variational Representation of Preferences. Econometrica 74 (6):14471498. **2006**.

[66] Massimo Franceschetti, Ronald Meester. Random Networks for Communication: From Statistical Physics to Information Systems (Cambridge Series in Statistical and Probabilistic Mathematics) 1st Edition. **2007**.

[67] Aaron F. McDaid, Thomas Brendan Murphy, Nial Friel and Neil J. Hurley. Clustering in networks with the collapsed Stochastic Block Model. Computational Statistics and Data Analysis (Elsevier) **2012**.

[68] Mezard and Montanari. Physics, information and Computation.**2009**.

[69] Miller, J. W. and Dunson, D. B. Robust Bayesian inference via coarsening. arXiv preprint arXiv:1506.06101.**2015**.

[70] Antonietta Mira, Garry Robins, Alessandro Lomi. Scalable MCMC algorithm for the accurate estimation of Exponential Random Graph Models. **2017**.

[71] Morten Morup and Mikkel N. Schmidt. Bayesian Community Detection. April 17, **2012**.

[72] Jared S. Murray, Carlos M. Carvalho, under preparation **2018**.

[73] Jouchi Nakajima and Mike West. Bayesian Analysis of Latent Threshold Dynamic Models. August **2012**.

[74] M. E. J. Newman. Assortative mixing in networks. Physical Review Letters. 89 (20): 208701. **2002**.

[75] Mark Newman. Networks. Oxford University Press. **2018**.

[76] Neville, J., Gallagher, B., Eliassi-Rad, T. and Wang, T. Correcting evaluation bias of relational classifiers with network cross validation. Knowledge and information systems, 30 (1), 3155. **2012**.

[77] Nowicki, K. and Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. Journal of the American Statistical Association, 96, 10771087. **2001**.

[78] Peter, Orbanz. Subsampling large graphs and invariance in networks. **2017**.

[79] Peter Orbanz, Daniel M. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. IEEE Trans. Pattern Anal. Mach. Intell., 37(2):437461, **2015**.

[80] Patrick J. Wolfe and Sofia C. Olhede. Nonparametric graphon estimation. **2013**.

[81] Sofia C. Olhede and Patrick J. Wolfe. Network histograms and universality of blockmodel approximation. **2014**.

[82] Simon Lunagomez, Sofia C. Olhede and Patrick J. Wolfe. A symmetric Multivariate Model for Network Data Based on Graph distance. **2018**.

[83] Piexoto. Nonparametric Bayesian inference of the microcanonical stochastic block model. **2018**

[84] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent Space Approaches to Social Network Analysis. **2002**.

[85] Rka Albert. Scale-free networks in cell biology. Journal of Cell Science **2005** 118: 4947-4957.

[86] Bengt Rosen. On the coupon collectors waiting time. Ann. Math.Statist., 41:19521969, 1970.

[87] Rubin, D. B. Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons. **1987**.

[88] Rubin, D. B. (**1987**). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons.

[89] Robins, G.; Snijders, T.; Wang, P.; Handcock, M.; Pattison, P. Recent developments in exponential random graph (p*) models for social networks. Social Networks. 29: 192215. 2007.

[90] Ruggeri, F., Ros Insua, D. and Martn, J. Robust Bayesian Analysis. Handbook of statistics, 25, 623667. **2005**.

[91] Eldar Sadikov, Montserrat Medina, Jure Leskovec. Hector Garcia-Molina. Correcting for Missing Data in Information Cascades. **2011**.

[92] Pratha Sah, Lisa O. Singh, Aaron Clauset and Shweta Bansal. Exploring community structure in biological networks with random graphs. Dec. 22, **2013**.

[93] Tanay Kumar Saha and Mohammad Al Hasan. Finding Network Motifs Using MCMC Sampling.

[94] G. Schwarz, Estimating the dimension of a model, The annals of statistics, vol. 6, no. 2, pp. 461-464, **1978**.

[95] Daniel K. Sewell and Yuguo Chen. Latent Space Approaches to Community Detection in Dynamic Networks Bayesian Anal. Volume 12, Number 2 (**2017**), 351-377.

[96] Shweta Bansal, Shashank Khandelwal and Lauren Ancel Meyers. Exploring biological network structure with clustered random networks. BMC Bioinformatics **2009** 10:405.

[97] Sucheta Soundarajan, Tina Eliassi-Rad, Brian Gallagher, Ali Pinar. MaxOutProbe: An Algorithm for Increasing the Size of Partially Observed Networks. **2015**.

[98] Strauss, D., On a general class of models for interaction, SIAM Review 28, 513527 (**1986**).

[99] Michael P. H. Stumpf, Carsten Wiuf, and Robert M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. PNAS March 22, **2005**.

[100] Tianxi Li, Elizaveta Levina, Ji Zhu. Network cross-validation by edge sampling. DSW 2019, 2019 IEEE Data Science Workshop — 2-5 June 2019, **2018**.

[101] Thompson, S. K., and Seber, G. A. F. Adaptive Sampling. John Wiley and Sons, New York, **1996**.

[102] Thompson, S.K. and Frank, O. Model-based estimation with link-tracing sampling designs. Survey Methodology 26 87-98. **2000**.

[103] Thompson, S.K. and Collins, L.M. Adaptive sampling in research on risk-related behaviors. Drug and Alcohol Dependence 68 S57-S67. 2002.

[104] S.L. van der Pas and A.W. van der Vaart. Bayesian Community Detection. **2016**.

[105] Jennifer Nicoll Victor, Alexander H. Montgomery, and Mark Lubell. Network Theory and Political Science. The Oxford Handbook of Political Networks. **2017**.

[106] Vidakovic, B. Γ-minimax: a paradigm for conservative robust Bayesians. Pages 241-259 of: Rios Insua, D., Ruggeri, F. (eds), Robust bayesian analysis. Springer. **2000**.

[107] Volz, E. and Heckathorn, D. Probability based estimation theory for respondent driven sampling. Journal of Official Statistics, 24, 79–97. **2008**.

[108] Wald, A. Statistical Decision Functions. Wiley, New York. **1950**.

[109] Y.X. Rachel Wang, Peter J. Bickel. Likelihood-Based model selection for Stochastic Block Model, Annals of Statistics, **2016**.

[110] James Watson and Chris Holmes. Approximate Models and Robust Decisions, Statist. Sci. Volume 31, Number 4 (**2016**), 465-489.

[111] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. Nature, 393:440-442, **1998**.

[112] Duncan J. Watts. Small Worlds: The Dynamics of Networks between Order and Randomness (Princeton Studies in Complexity) Paperback December 14. **2003**.

[113] WIND16. Workshop on incomplete networked data. eliassi.org/WIND16.html. Abstracts for March **2016** workshop.

[114] Jingfei Zhang and Yuguo Chen. Sampling for Conditional Inference on Network Data. Pages 1295-1307, 01 Apr **2012**,Journal of the American Statistical Association.

[115] Yaonan Zhang, Eric D. Kolaczyk, and Bruce D. Spencer. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. Ann. Appl. Stat. Volume 9, Number 1 (**2015**), 166-199.