Sampling and Inference in Social Networks

by

Don Buddhika Wijayantha Nettasinghe

B.Sc. Eng., University of Peradeniya, Sri Lanka, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

The Faculty of Graduate and Postdoctoral Studies

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

December 2016

Abstract

This thesis explores three practically important problems related to social networks and proposes solutions utilizing tools from statistical signal processing and applied mathematics.

In the first problem, we consider the maximum likelihood estimation of the degree distribution of a network from a sampled network. We are interested in the variance of the estimates for two widely used degree distribution models of large scale networks: (1) exponential model (2) power-law model. We find that the variance of the estimate is approximately an order of magnitude smaller for exponential model when compared to the power-law model. The intuition behind this interesting property is conjectured to follow from the second order stochastic dominance of the probability distributions.

In the second problem, we consider a social network where each node has a vector of parameters (interests) associated with it. Only the partial graph and parameters of a subset of nodes are known. Our aim is to identify the parameters associated with each node. Assuming the *Homophily behaviour* of the social network, we propose a Cayley-Menger Determinant based optimization approach combined with a previously proposed spectral clustering method to solve the interest identification problem.

In the third problem, we study interest influencing in social networks where the aim is to find the minimum set of missing edges in order to maximize the influential power over the interests of the individuals in the network. This problem is formulated as a classic problem in structural rigidity theory using the characteristic *Contagion behaviour* of social networks. We propose a solution approach using the Laman's theorem which characterizes minimally rigid graphs in the two dimensional space.

Preface

The following publications have resulted from the research presented in this thesis:

- D. B. W. Nettasinghe, V. Krishnamurthy, and V. K. Bhargava, "Identification of Individual Interests in Online Social Networks Using Cayley-Menger Determinants," accepted for presentation at *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* (Linked to Chapters 3)

Statement of Authorship

I am the primary author for the publications listed above. I have been responsible for developing original ideas, deriving mathematical solutions, and generating simulation results for these publications with valuable frequent suggestions, advice and feedback from Prof. Vikram Krishnamurthy and Prof. Vijay K. Bhargava.

A version of the content presented in the Appendix A has been submitted as a course project for *STAT 547C: Topics in Probability* course.

Table of Contents

Appendix

List of Figures

List of Abbreviations

AIC	:	Akaike Information Criterion
BFS	:	Breadth-First Sampling
CDF	:	Cumulative Distribution Function
CRLB	:	Cramèr-Rao Lower Bound
ID	:	Identification Number
MISC	:	Missing node Identification by Spectral Clustering
ML	:	Maximum Likelihood
MSE	:	Mean Squared Error
RDN	:	Random Degree Node
RW	:	Random Walk
SOSD	:	Second Order Stochastic Dominance

Acknowledgements

First and foremost, I would like to thank my advisors, Professor Vikram Krishnamurthy and Professor Vijay K. Bhargava, for their support, guidance and and generous financial support. I am grateful to Prof. Vikram Krishnamurthy for always making time to have long discussions related to various aspects of statistical signal processing that ignited my passion towards this research area and for guiding me to explore the various tools and techniques studied in this thesis. I am grateful to Professor Vijay K. Bhargava for the academic guidance he provided from the beginning of my MASc and for allowing and facilitating me to pursue the areas that I have always been passionate and curious about, and thus making my MASc an amazing learning experience. This thesis would not have been possible without their support, advice, and encouragement.

Secondly, I would like to extend my sincere thanks to all the excellent colleagues at the Information Theory and Systems laboratory, who offered me genuine and friendly support to carry out my research. I should particularly thank K.N.R. Surya Vara Prasad for his friendship and advice in making many academic decisions during my MASc. I would also like to thank Duong Nguyen, Mark Adams, Sudha Lohani, Shankhanaad Mallick, Peiran Wu, Jun Zhu, Rudi Plesch, Anup Aprem, Shahrear Tanzil, Nandinee Fariah, Sujay Bhatt and Tavis Pederson for their friendship and support during my MASc research.

Lastly, I am profoundly indebted to my parents, brother and friends in Sri Lanka for their unconditional love and blessings. Without their continuous support and encouragement, I could not have completed this thesis.

Dedication

To my father Don Chandrapala Padmasiri Nettasinghe, mother Gunasiri Rajapaksa Manike, and brother Buddhima Nettasinghe.

With gratitude for your love and support.

Chapter 1

Introduction

1.1 Overview

Social networks have become a crucial part of the modern society: people use social networks to connect with family and friends, to share news and information, to get updates on events, to obtain the views of others regarding various issues, etc. This integration of social networks into the day to day lives of people has made drastic changes in the way people make decisions in voting, purchase goods, adapt new beliefs or views, etc.

These drastic effects of the social networks on almost every aspect of modern society has given rise to a growing interest in the research community to model and analyze the social networks in order to understand their dynamics in an abstract manner. Further, since a social network is a highly complex and dynamic network that continuously evolves, tools and techniques from a wide variety of research areas such as electrical engineering, computer science, microeconomic theory and statistics can be employed to model various aspects of social networks. For example, probability theory and statistical inference is used to estimate various characteristics of a social network such as degree distributions [24], tools arising from microeconomics such as game theory and revealed preference are used to study how individuals with limited awareness of the network can achieve global coordination and to model various human behaviours [19, 34] and concepts from machine learning theory are used to learn and update the operating policies of individuals in social networks.

In this thesis, we explore three problems related to social networks that are of high practical importance, namely, model based estimation of the degree distribution of a social

network, identification of individual interests in a partially known network and interest influencing in a social network. In the rest of this section, we provide a brief introduction about each these problems and explain the motivation to explore that problem. Each of these problems will be discussed in detail in the subsequent chapters.

1.1.1 Effect of the Underlying Model on Estimating the Degree Distribution of a Network Under Sampling

One of the most important characteristics of a network is the degree distribution D_k which gives the probability of a node chosen uniformly at random having k number of neighbours. Key properties of a network such as connectivity of the graph, information diffusion, existence of a giant component can be analyzed using the degree distribution of a network [34]. Therefore, the estimation of the degree distribution is of very high importance in analyzing a social network.

Two of the most widely used models for the degree distributions of large scale networks are exponential model (with degree distribution $C_{exp}e^{-\lambda k}$ characterized by the parameter λ) and power-law model (with degree distribution $C_{pl}k^{-\gamma}$ characterized by the parameter γ). The exponential model is used for real world networks such as Maritime Transportation Network, Email Networks and Power Grid Networks [12] with parameter $\lambda \in (0, 1)$. Further, such a degree distribution also holds for an Erdős-Rényi random graph [52]. The power-law model is used for scale-free networks such as social networks, World Wide Web [12] with parameter $\gamma \in (2, 3)$ [8, 10].

However, since most of the real world networks currently consists of millions of nodes, it is not feasible to analyze the complete network in order to identify the parameters of the underlying degree distribution model (this will be further explained in Section 2.1). Therefore, the most commonly used approach to estimate the degree distribution of a social network is to utilize a survey based sampling approach instead of analyzing the whole network. A

survey based sampling method includes a sampling process to select a subset of nodes and a querying method to obtain the parameter of interest (the true degree of the node in this case) from each sampled node. These two steps can be thought of as the step of selecting people for an experiment (sampling step) and asking questions from the selected people (querying step).

However, this survey based sampling introduces two types of noises (or random errors) into the process of estimating the degree distribution, namely, sampling noise and measurement (querying) noise. Sampling noise represents the randomness introduced from the sampling process (the process with which we select the participants of the survey). Measurement noise is introduced from the querying process. For example, consider the scenario of using a Facebook survey to ask every sampled node about his/her number of friends. Measurement noise accounts for the likely situation where a person does not actually provide his/her correct answer in such a survey (most work in literature assume that there is no measurement noise i.e. all queries of sampled nodes yield the true degree of that node in the original network). Due to these two types of noises, the accurate estimation of the degree distribution of the underlying network becomes a non-trivial problem.

Motivated by the above facts, in Chapter 2, we consider the problem of maximum likelihood estimation of a degree distribution from the degree distribution of a sampled network for two distinct cases: 1) The underlying graph is known to have an exponential distribution 2) The underlying graph is known to have a power-law distribution. The primary aim of this chapter is to answer the question, ***"Does the underlying degree distribution of a network have an effect on the accuracy of the maximum likelihood estimation?"***. Our main result in this section shows that the maximum likelihood estimation has a better performance in exponential graphs when compared to the power-law graphs. Specifically, we show that the variance of the maximum likelihood estimate is approximately an order of magnitude smaller for exponential graphs compared to the power-law graphs. We further conjecture that the reason for this behaviour is the second order stochastic dominance of the

probability distributions.

1.1.2 Interest Identification in Online Social Networks

In a social network with graph $G = (V, E)$, each individual has a set of parameters that are unique to him/her which are represented by the vector $h(v_i)$ for each node $v_i \in V$. These parameters could be related to interests or beliefs, shopping preferences, etc [1]. The knowledge about such parameters of the people in a social network may be of very high importance and value for purposes such as marketing, security, etc. However, the parameters of all nodes in an online social networks may not be visible to everyone. There may exist some nodes whose social structure (how they connect to other nodes in the social network) and the associated parameters are not known. The set of such nodes, $V_m \subset V$, are called missing nodes. For example, consider a company called ABC that has a set of customers subscribed in Facebook (a similar example is considered in [14, 15] for a different problem). The profiles of the set of people $V_k \subset V$ who have subscribed to the services of the company are visible to the company. Hence, the parameters $h(v_i)$ of each customer $v_i \in V_k$ are known to the company. However, ABC is interested in knowing the parameters associated with the set of people V_m who are not its subscribers since ABC can use the knowledge of such parameters to identify the best set of potential customers who would be interested in the services of ABC. For example, if ABC is a company that sells sports cars, the parameter which indicates if a person (who is not a subscriber of ABC) is interested in sports cars would be of interest to ABC in order to identify potential customers. This motivates us to pursue the aim of the chapter which is formally stated below.

We consider a social network represented by the undirected graph $G = (V, E)$ in which there is an interest vector $h(v_i) \in \mathbb{R}^d$ associated with each node $v_i \in V$. We assume that

[1] In Facebook, these parameters include date of birth, gender, religion, political views, current and past occupations, attended university/high school, current and previous places of residence, spoken languages, family and relationship information, shopping preferences, music and movie preferences, travelled places, attended events, hobbies.

the set of nodes $V_m \subseteq V$ are missing, that is, we know neither how they connect to the rest of the nodes nor their interests. We only know part of the network that is denoted as $G_k = (V_k, E_k)$. Therefore, $V = V_k \cup V_m$ and $E = E_k \cup E_m$. In this context, the aim of the Chapter 3 is **Interest Identification**: Given G_k, the cardinality of V_m, i.e. $|V_m|$, and the interests associated with each known node $v_i \in V_k$, estimate the interests associated with each missing node $v_j \in V_m$.

As the main results of the Chapter 3, we introduce a method to solve the problem of interest identification of missing nodes utilizing a Cayley-Menger Determinant based optimization approach.

1.1.3 Interest Influencing in Online Social Network

The interests of the people in a social network may tend to vary over the time. This could be beneficial or not beneficial to certain entities. For example, consider our previous example of ABC company. ABC company wants to make sure that it does not lose its existing customer base. Hence, ABC company should make sure that the interests of its customers always stay positive towards the company. If strong ties could be built between customers who tend to leave the company and company's most loyal customers, there could be a possibility that the customers who are about to leave would be motivated to stay further with the company. For example, the ABC company can introduce a customer that is about to leave it to few of its loyal customers. If these new social ties are strong, there is a high possibility that the said customer will decide to continue to stay with the company.

In the previous example, we build connections between people in order to stop an undesirable change of interests. This is based on the concept of social contagion behaviour of social networks. Contagion/influence causes a strong friendship between two persons to pass attributes of one person to another person (induction) [9, 46]. Results given in [21] specifically show that shopping habits and leisure activities are contagious in social networks. Recently, [36] concluded that the individual behaviour of people in a social network depends

more on their local network structure (how they are connected to the rest of the nodes), than the actual behaviour of the whole network, a phenomenon which is called "the majority illusion". A very interesting article, [38], which recently appeared in Harvard Business Review discusses this further and says that "*you only need handful of influencers to give the impression that everyone is talking about your brand*" due to this majority illusion paradox.

Motivated by the above examples and facts, in Chapter 4, we explore this idea from a mathematical point of view to find out how to create a network structure in order to gain a maximum influential power over the interests of people in the network. The primary question answered here is ***"How can we introduce new connections between individuals of a network in a manner such that the new connections will prevent the preferences of individuals from changing?"***. As our solution approach to this problem, we introduce a method to identify the minimal set of new connections that should be added to the network in order to influence the interests of the nodes of a social network. We achieve this by viewing the interests associated with the nodes of the social network as a bar and joint framework residing in the 2-dimensional Euclidean space and then utilizing concepts from the mathematical branch of rigidity theory.

1.2 Main Contributions

In the following subsections, a brief list of contributions of each chapter of this thesis is provided. Detailed discussions of each contribution are provided in the corresponding chapters.

1.2.1 Estimating the Degree Distributions of Networks Under Noise

As explained in Section 1.1.1, Chapter 2 compares the maximum likelihood estimates of degree distributions for exponential and power-law models. The main contributions of this chapter are:

1. A complete formulation of the maximum likelihood estimation problem of the true degree distribution from the sampled (noisy) distribution for both exponential and power-law networks.

2. Showing that the variance of the maximum likelihood estimate is approximately an order of magnitude smaller for the exponential model compared to the power-law model (in the respective regions of interests of the parameters) from both numerical as well as theoretical analysis.

3. An intuitive explanation of the main result(contribution 2) using the second order stochastic dominance concept from microeconomic theory.

Apart from the above main technical contributions, a brief survey about sampling methods for social networks is also provided in Appendix A for the completion of the Chapter 2.

1.2.2 Identification of Individual Interests in Online Social Networks

Based on the motivation provided in Section 1.1.2, Chapter 3 explores the problem of interest identification in online social networks. The main contributions of this chapter are:

1. A complete formulation of the problem of interest identification assuming that the social network exhibits the *Homophily Behaviour* (friends have similar interests)

2. An optimization approach to solve the problem of interest identification subject to constraints obtained using Cayley-Menger Determinant based approach.

1.2.3 Interest Influencing in Online Social Networks

Based on the motivation and the examples provided in Section 1.1.3, Chapter 4 considers the problem of interest influencing. The main contributions of the Chapter 4 are:

1. A complete mathematical formulation of the problem of interest influencing.

2. A method to solve the problem of interest influencing by using the tools from the mathematical branch of *Rigidity Theory*.

1.3 Outline of the Thesis

The outline of the rest of this thesis is as follows.

Chapter 2 considers the maximum likelihood estimation of degree distribution of networks for two models: exponential model and power-law model. It begins by emphasizing the importance of sampling in analyzing a social network and explains the effect of sampling on estimating a parameter of a social network. It then defines the aim of the chapter which is to find out if the accuracy of maximum likelihood estimates of the degree distribution depends on the underlying model (exponential or power-law). Then, we derive the maximum likelihood estimates of the degree distributions for the two models and theoretically compare their variances using Cramèr-Rao Lower Bounds to prove the main result (the variance of the estimate for the exponential model is approximately an order of magnitude smaller compared to the power-law model). Then, a comparison of the mean squared errors of the estimators of the two models is provided in order to numerically verify the main result. Finally, a conjecture utilizing the Second Order Stochastic Domination of probability distributions is provided to intuitively explain the main result of the section.

Chapter 3 considers the problem of interest identification in social networks. We start by providing a brief explanation about the related work and then define the problem of interest identification. We also provide a detailed explanation of the main assumption of our work, *Homophily behaviour*, and how we mathematically model it. Then, we explain our solution to the interest identification problem which is a Cayley-Menger determinant based optimization approach and also provide a numerical example to illustrate the important steps of the proposed method.

Chapter 4 considers the problem of interest influencing and begins by formulating the problem based on the *contagion behaviour* of individuals in a social network. Then, the proposed method of solving the problem of interest influencing using the concept of rigidity theory is explained. Finally, some numerical examples are provided to demonstrate main steps of the proposed work.

Finally, Chapter 5 provides a conclusion of the work presented in this thesis and some future research directions.

Chapter 2

Estimating the Degree Distributions of Networks Under Noise

2.1 Why do we Sample Networks?

As described in Chapter 1, online social networks provide an interesting platform for studies in various domains. However, studying such social networks is difficult due to two primary reasons. The first reason is the massive size of these social networks. For example, in 2016, Facebook reports 1.71 billion monthly active users [1] and Twitter reports 310 million monthly active users [2]. Due to these massive sizes of the online social networks, it is impossible to perform any analysis on the actual network. The second difficulty arises in acquiring the data of each and every person in these massive social networks since most of the network administrators are not willing to provide their data for free [20].

A practical solution to the above difficulties in social networks analysis is to select a smaller (compared to the original network) subnetwork from the original social network which best represents the original network. This approach is called network sampling. The main decisions [37] that we have to make in the network sampling can be listed as:

- What is the sampling method?
- What is the sample size?
- How good is the selected sample?

In general, sampling methods for networks can be classified in to three main categories:

1. Sampling by Random Node Selection
2. Sampling by Random Edge Selection
3. Sampling by Exploration.

A brief discussion about each of the above sampling methods and their pros and cons is provided in the Appendix A.

It should be noted that a parameter of a sampled network will be a noisy estimate of the same parameter of the original network since the sampling process may not be able to capture the original parameter accurately (as emphasized in Section 1.1.1 as well). This parameter in concern for this chapter is the degree degree distribution of the network and we focus on estimating it under sampling noise in the following sections.

The organization of the rest of the chapter is as follows. Section 2.2.1 first defines the problem and then the Section 2.2.2 provides a brief survey of the key work that has been done in this context and explains how our work differs from them. In Section 2.3, we state our assumptions and formulate the maximum likelihood estimation problem for the two models and theoretically compare the variances of the estimates. In Section 2.4, the numerical results are provided for Mean Squared Error (MSE) to verify the theoretical analysis. Further, we provide an intuitive explanation of the main result using the concept of second order stochastic dominance and risk aversion in Section 2.5. Finally, in Section 2.6, we provide a summary of this chapter and some interesting future research directions of the presented work.

2.2 Estimating Degree Distributions from Sampled Networks: A Model Based Comparison

2.2.1 Problem Definition and Main Result

Consider a network represented by the undirected graph $G = (V, E)$. Our aim is to compare the variance of the estimates of the degree distribution D_k of G, using the degree distribution D'_k of the sampled network $G' = (V', E')$ for two distinct cases:

1. The underlying graph G is known to have an exponential degree distribution with a parameter λ

2. The underlying graph G is known to have a power-law degree distribution with a power law exponent γ.

The main result is that the variance is approximately an order of magnitude smaller for the exponential model compared to the power law model in the respective regions of interests of the parameters from both numerical as well as theoretical analysis. This result has not been presented in the existing literature to the best of our knowledge and this result emphasizes that more accurate techniques should be employed when estimating the power-law degree distributions compared to exponential degree distributions.

2.2.2 Related Work

Recently, [54] has analyzed the problem of estimating the degree distribution of networks using a sampled network. Their work is based on the findings in [16, 17] which says that, for some sampling methods (with no observation noise), the sampled distribution (D') and the true distribution (D) are related according to,

$$\mathbb{E}[D'] = PD \tag{2.1}$$

where P is a matrix that depends fully on the sampling scheme. They claim that the operator P is not invertible in practice (ill-conditioned) and hence use a weighted least squares based approach to solve the problem of estimating the true degree distribution, D. The main difference between this work and the method that we present lies in the fact that our approach is model-based, whereas the the work in [54] is design-based. In other words, we consider our observation to be a realization from a stochastic process that depends on a fixed parameter, whereas the randomness in design-based methods is assumed to be only due to sampling approach [25]. Further, [25] specifically concludes that design based methods suffer from severe limitations in network inference and hence model-based methods are preferred.

[4] considers the problem of estimating the parameter of a power law distribution (not necessarily related to networks) using maximum likelihood based methods. However, in this work, the likelihood of the data given the model is considered with the assumption of no noise, whereas the noise is taken into account in our model. Further, the main goal of [4] is to compare the maximum likelihood based method with other methods such as least square based methods and graphical methods, whereas our intention is to compare the performance of maximum likelihood estimation for two models. Similarly, [10] also utilizes maximum likelihood based approach to analyze data following power law distributions. Similar to [4], authors of [10] also mention that maximum likelihood based approaches are better when compared to other methods and attempt to detect and characterize power law distributions.

2.3 Problem Formulation

2.3.1 Maximum Likelihood Estimation: A brief Introduction

Consider a system characterized by a deterministic parameter, θ^* and a set of measurements $Y_N = (y_1, \ldots, y_N)$. Intuitively, the aim of Maximum Likelihood (ML) Estimation is to compute the most plausible value of the parameter which could have produced Y_N. This

plausibility is measured using the likelihood function, $L(\theta, N)$, which is defined as,

$$L(\theta, N) = \mathbb{P}(Y_N|\theta), \theta \in \Theta \tag{2.2}$$

where, Θ is the parameter space and the θ characterizes the parameterized model. The ML estimate $\hat{\theta}_N$ of θ^* is then defined as,

$$\hat{\theta}_N = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(\theta, N). \tag{2.3}$$

Often, it is more convenient to maximize log-likelihood,

$$\mathcal{L}(N, \theta) = \ln L(\theta, N) \tag{2.4}$$

since maximizers of both functions are the same.

The reasons for using ML estimation in our work are the following two properties [6, 26].

1. **Strong Consistency:** Under reasonable conditions, the ML estimate, $\hat{\theta}_N$, converges to the true parameter, θ^*, almost surely i.e.

$$\mathbb{P}\left(\lim_{N \to \infty} \hat{\theta}_N \to \theta^* \right) = 1 \tag{2.5}$$

2. **Asymptotic Normality:** The ML Estimate, $\hat{\theta}_N$, is asymptotically normally distributed about the true parameter, θ, with a variance equal to the inverse of the Fisher Information, I_{θ^*} [29]

A more detailed discussion about using ML estimation and its applications in various system models (linear Gaussian state space, hidden Markov model, etc) can be found in [33].

2.3.2 Deriving Maximum Likelihood Estimators

The degree distribution of an exponential graph with a parameter λ is of the form,

$$D_{exp,k} = (1 - e^{-\lambda})e^{\lambda k_{min}}e^{-\lambda k} \tag{2.6}$$

where k_{min} is the minimum degree of the graph.

Similarly, the degree distribution of a power-law graph (in the range from k_{min} to k_{max}) with parameter γ is given by,

$$D_{pl,k} = \frac{k^{-\gamma}}{\zeta(\gamma, k_{min}) - \zeta(\gamma, k_{max})}, \tag{2.7}$$

where $\zeta(\gamma, k)$ is the Hurwitz zeta function,

$$\zeta(\gamma, k) = \sum_{n=0}^{\infty}(n + k)^{-\gamma}. \tag{2.8}$$

We assume that the noise introduced from the measurement process and/or the sampling process into the measured distribution, D', to be additive Gaussian noise. [54] shows that sampling noise can be approximated using a Gaussian model. If the sample size is considerably large, the measurement noise can also be approximated by a Gaussian random variable by the central limit theorem, irrespective of the actual underlying distribution of the noise associated with each query.

Hence, it is fair to assume that the combined effect of both noises can be modelled by a normal random variable. Therefore, the measured distributions can be expressed as,

$$D'_{exp,k} = D_{exp,k} + \eta_{exp,k} \tag{2.9}$$

$$D'_{pl,k} = D_{pl,k} + \eta_{pl,k} \tag{2.10}$$

where $\eta_{exp,k}$ and $\eta_{pl,k}$ are both independent (in k) normal random variables with zero mean and variance σ^2. For simplicity, we assume that $k_{min} = 1$ and $k_{max} = N$.

Then, using (2.6) and (2.9), the log-likelihood function for the exponential parameter, $\mathcal{L}_{exp}(\lambda) = \ln \prod_{k=1}^{N} \mathbb{P}(D'_{exp,k}|\lambda)$, can be expressed as,

$$\mathcal{L}_{exp}(\lambda) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(D'_{exp,k} - (1-e^{-\lambda})e^{-\lambda(k-k_{min})}\right)^2. \tag{2.11}$$

Similarly, using (2.7) and (2.10), the log-likelihood function for the power-law parameter, $\mathcal{L}_{pl}(\gamma) = \ln \prod_{k=1}^{N} \mathbb{P}(D'_{pl,k}|\gamma)$, can be expressed as,

$$\mathcal{L}_{pl}(\gamma) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{N}\left(D'_{pl,k} - \frac{k^{-\gamma}}{\zeta(\gamma,k_{min}) - \zeta(\gamma,k_{max})}\right)^2. \tag{2.12}$$

The maximum likelihood estimators, $\hat{\lambda}$ and $\hat{\gamma}$, for the parameters λ and γ can then be obtained as,

$$\hat{\lambda} = \underset{\lambda}{\text{argmax}}\,[\mathcal{L}_{exp}(\lambda)] \tag{2.13}$$

$$\hat{\gamma} = \underset{\gamma}{\text{argmax}}\,[\mathcal{L}_{exp}(\gamma)]. \tag{2.14}$$

2.3.3 Cramèr-Rao Bound Comparison

In order to theoretically compare the ML estimates for the two models, we analyze the Cramèr-Rao Lower Bounds (CRLBs). We consider a degree distribution with $N = 30$ for each model. Further, we assume that the noise variance, $\sigma^2 = 0.05$ for both exponential and power-law models.

CRLB is defined as the reciprocal of the Fisher Information [29]. The significance of the CRLB is that the variance of any unbiased estimator is bounded below by the CRLB. Based on the asymptotic normality of the maximum likelihood estimators, we can assume

that our estimators are unbiased (due to the large number of measurements) and thus, permitting the use of CRLBs for the comparison of the estimators. Further, ML estimators are asymptotically efficient i.e. when the number of measurements tends to infinity the variance of the ML estimator will reach the CRLB [33].

The CRLBs, $\hat{\lambda}_{CRLB}$ and $\hat{\gamma}_{CRLB}$, of the two estimators, $\hat{\lambda}$ and $\hat{\gamma}$, yields,

$$var(\hat{\lambda}) \geq \hat{\lambda}_{CRLB} = -\left[\mathbb{E}\left[\frac{\partial^2 \mathcal{L}_{exp}(\lambda)}{\partial^2 \lambda}\bigg|_{\lambda=\lambda^*}\right]\right]^{-1} \tag{2.15}$$

$$var(\hat{\gamma}) \geq \hat{\gamma}_{CRLB} = -\left[\mathbb{E}\left[\frac{\partial^2 \mathcal{L}_{pl}(\gamma)}{\partial^2 \gamma}\bigg|_{\gamma=\gamma^*}\right]\right]^{-1} \tag{2.16}$$

where λ^* and γ^* are the true parameters. The second order partial derivatives inside the expectation operator in (2.15) and (2.16) can be calculated easily and then the expected values can be obtained using a Monte Carlo simulation over a large number of independent measurements. The resulting plots of the CRLBs for the two estimators are shown in Fig. 2.1. It can be clearly seen that the CRLB of the estimator for the exponential model is smaller compared to the CRLB of the estimator for the power-law model.

2.4 Numerical Results

In order to verify the theoretical comparison in Section 2.3.3, we compare the MSEs of the estimators for the two same models ($N = 30$ and $\sigma^2 = 0.05$)

The maximum likelihood estimation of the parameters was performed by numerically evaluating (2.13) and (2.14). Then, the MSEs, $\hat{\lambda}_{MSE}$ and $\hat{\gamma}_{MSE}$, of the estimators $\hat{\lambda}$ and $\hat{\gamma}$, were obtained by a Monte Carlo method for different values of the true parameters, λ and γ, within their respective regions of interests. Fig. 2.2 shows the resulting plots in the same figure. It can be clearly seen from Fig. 2.2 that the MSE associated with the $\hat{\lambda}$ is significantly smaller compared to the MSE associated with the $\hat{\gamma}$ in the respective regions of interest. Hence, the MSEs associated with the two models show that the maximum

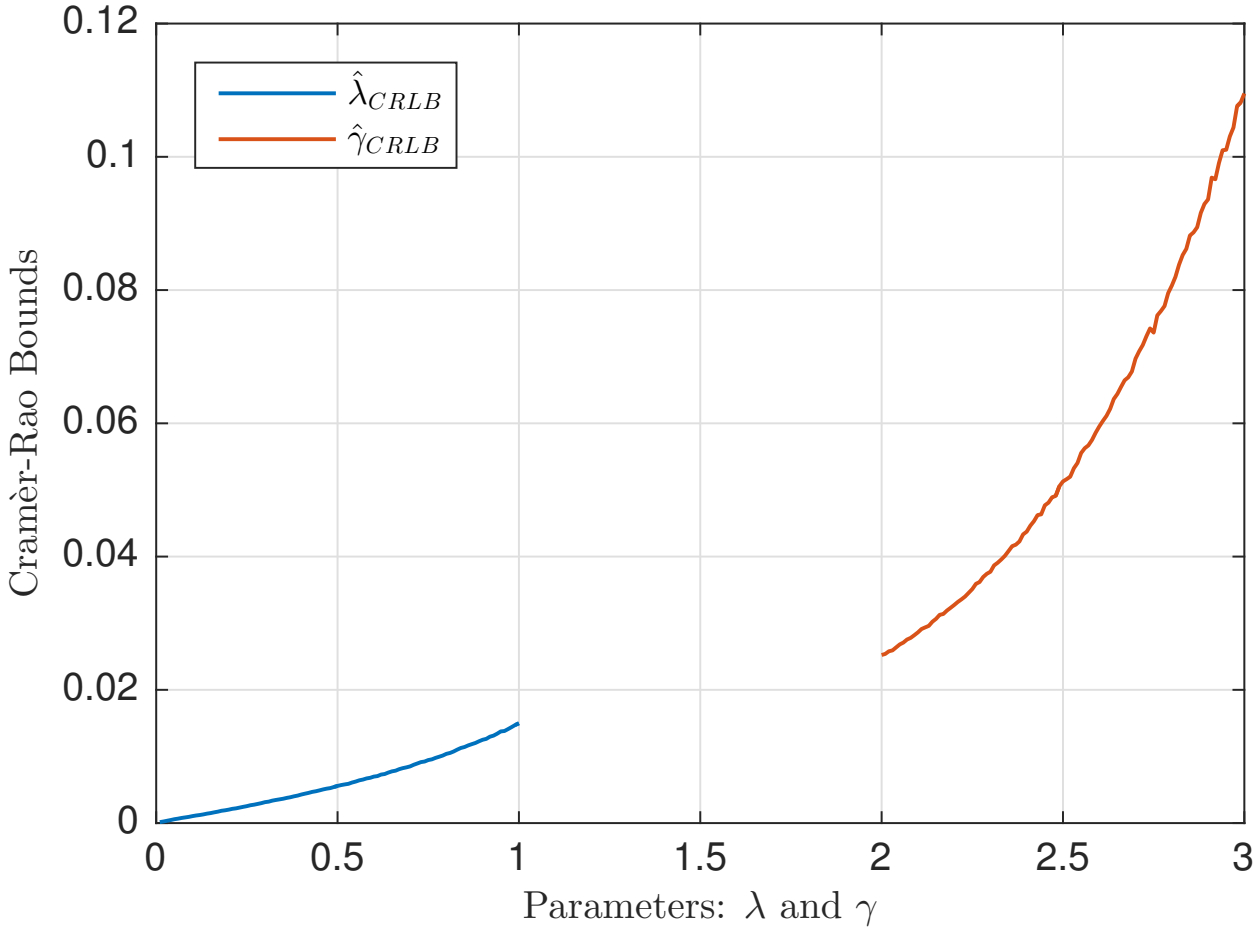


Figure 2.1: Variation of the CRLBs of the estimators, $\hat{\lambda}$ and $\hat{\gamma}$, with the true parameters λ and γ, in their respective regions of interest

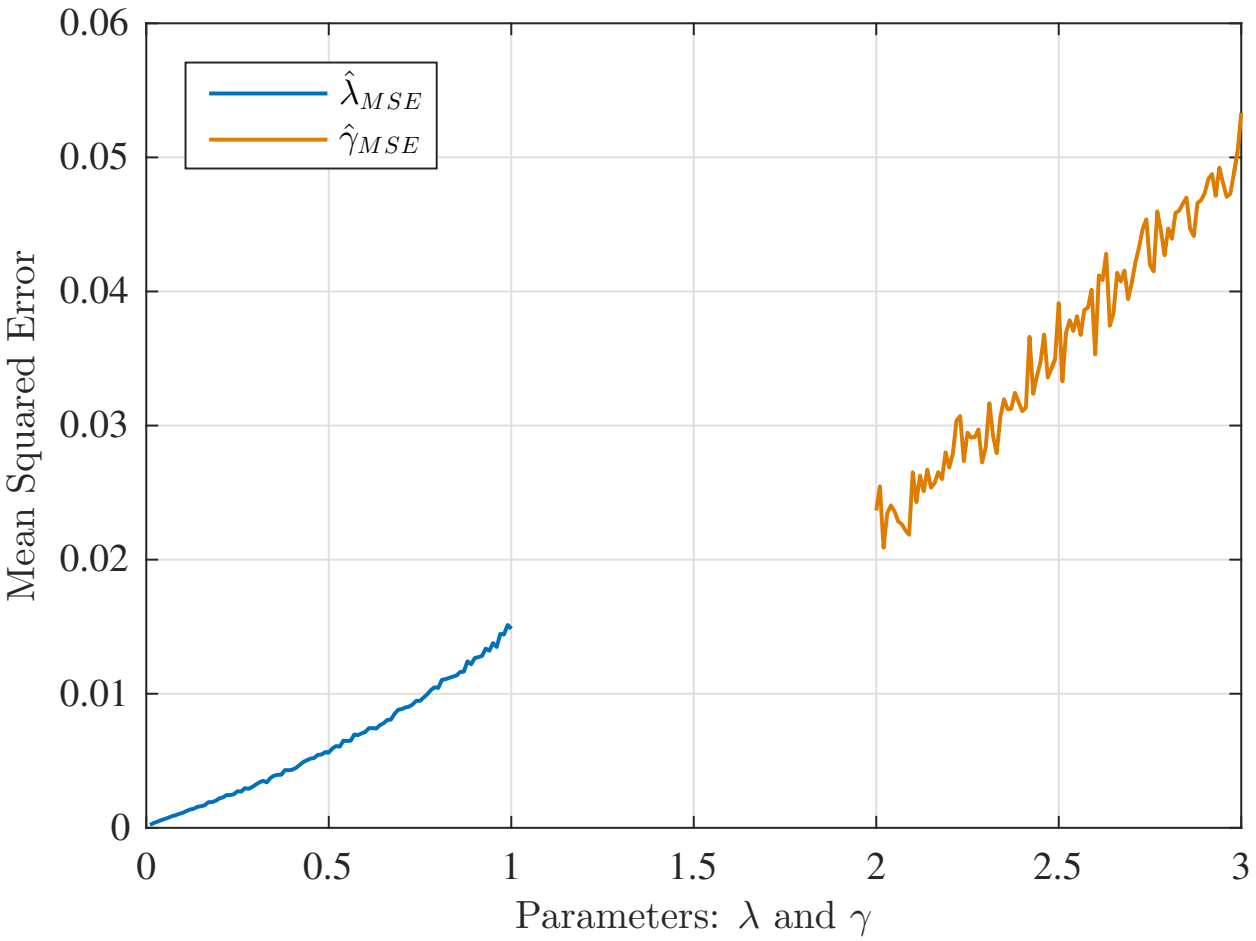


Figure 2.2: Variation of the MSEs of the estimators, $\hat{\lambda}$ and $\hat{\gamma}$, with the true parameters, λ and γ, in their respective regions of interest

likelihood estimation has a significantly better accuracy for exponential model compared to the power-law model.

By comparing Fig. 2.2 and Fig. 2.1, it can be seen that this theoretical analysis agrees with the MSE comparison and thus, confirming the fact the maximum likelihood estimation has a smaller variance (approximately an order of magnitude smaller in the middle of the regions of interests) in the case of exponential distribution compared to the power-law distribution.

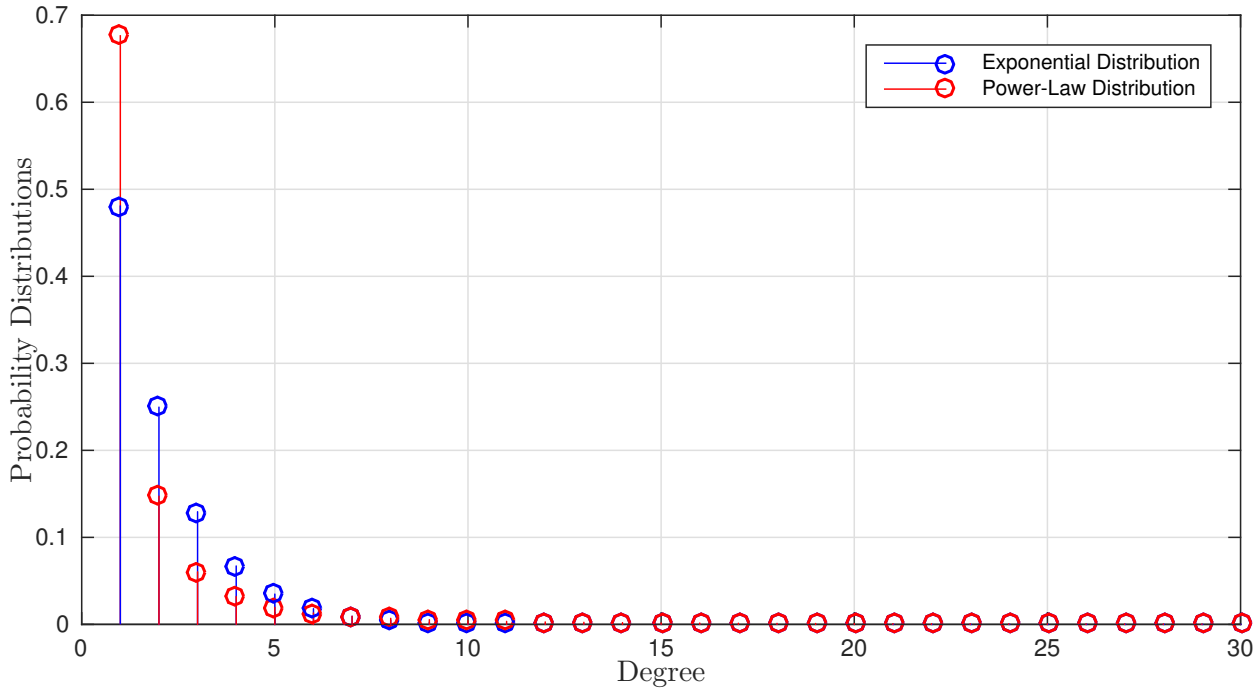
2.5 Intuitive Explanation: Second Order Stochastic Dominance

In Section 2.3.3 (and numerically in Section 3.4), we showed that the maximum likelihood estimator has a higher variance for power-law distributions compared to the exponential distributions with both numerical and theoretical analysis. However, the intuition behind this interesting result is still not completely clear at this point. Hence, we attempt to provide an intuitive explanation (as a conjecture) of this result using the concept of Second Order Stochastic Dominance (SOSD) [42].

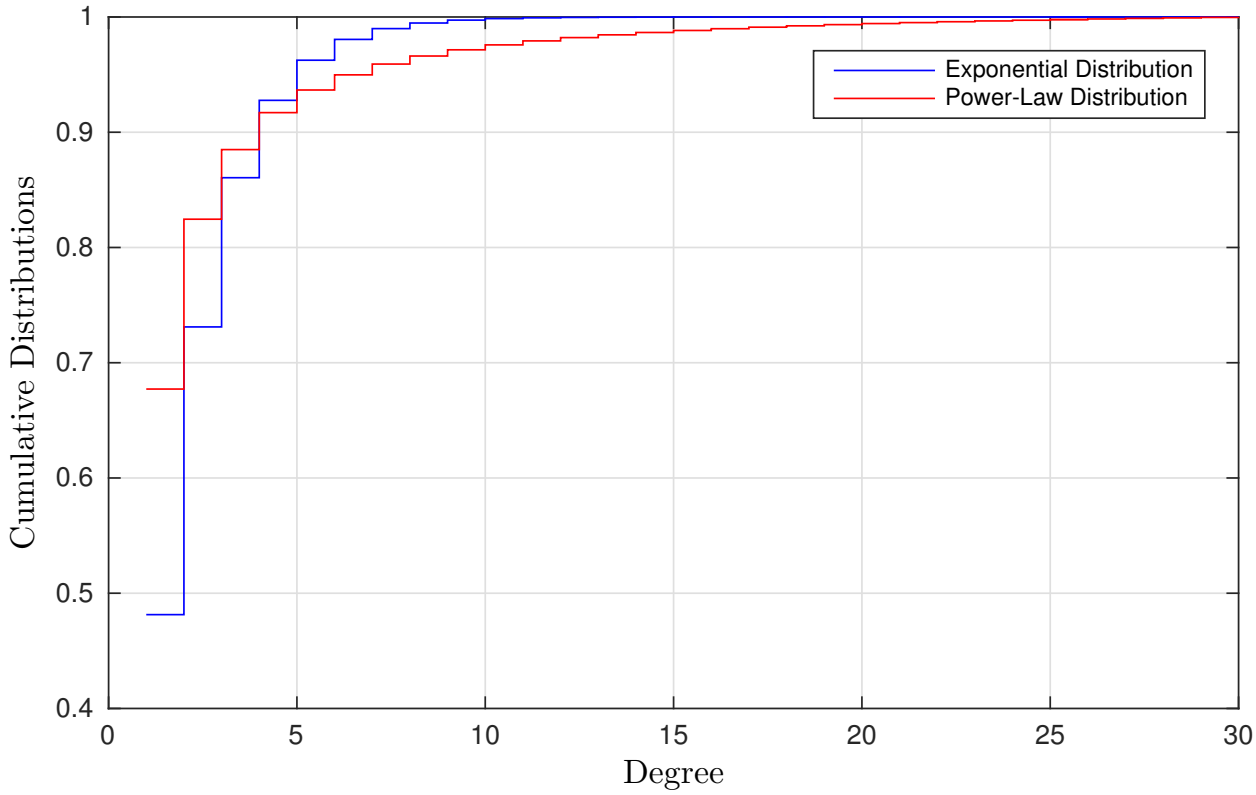
Consider the two Cumulative Distribution Functions (CDFs), F and G. F second-order stochastically dominate G, denoted $F \succeq_{SOSD} G$, if

$$\int_{-\infty}^{x} F(t)\,dt \leq \int_{-\infty}^{x} G(t)\,dt, \quad \forall x \in \mathbb{R}. \tag{2.17}$$

This concept of SOSD is used in microeconomic theory for risk averse decision making. As a simple example, consider a person who has to make a choice between betting money on two lotteries with the same expected winnings and characterized by the two distributions F and G where $F \succeq_{SOSD} G$. If the person is risk averse (a person who tries to avoid risk) he should choose F. Further, the variances of the two random variables also are related to the SOSD in the following way. If the two distributions F and G are equal in mean, then $F \succeq_{SOSD} G$ implies that F has a smaller variance compared to G (the converse is not true). Fig. 2.3 shows an exponential distribution function ($\lambda = 0.657$) and a power-law distribution function ($\gamma = 2.2$) with equal expected values (2.0772), their CDFs and the integrals of the CDFs (the integrals in (2.17) are replaced by summations for this discrete case). It can be seen from Fig. 2.3(c) that the exponential distribution second order stochastically dominate the power-law distribution. Hence, we conjecture that SOSD can be the reason behind the significant difference of the variances of the two estimators. In other words, this means that



(a) The exponential distribution ($\lambda = 0.657$) and power-law distribution ($\gamma = 2.2$) with equal means



(b) The CDFs of the two distributions

Figure 2.3: The distributions considered to depict the SOSD

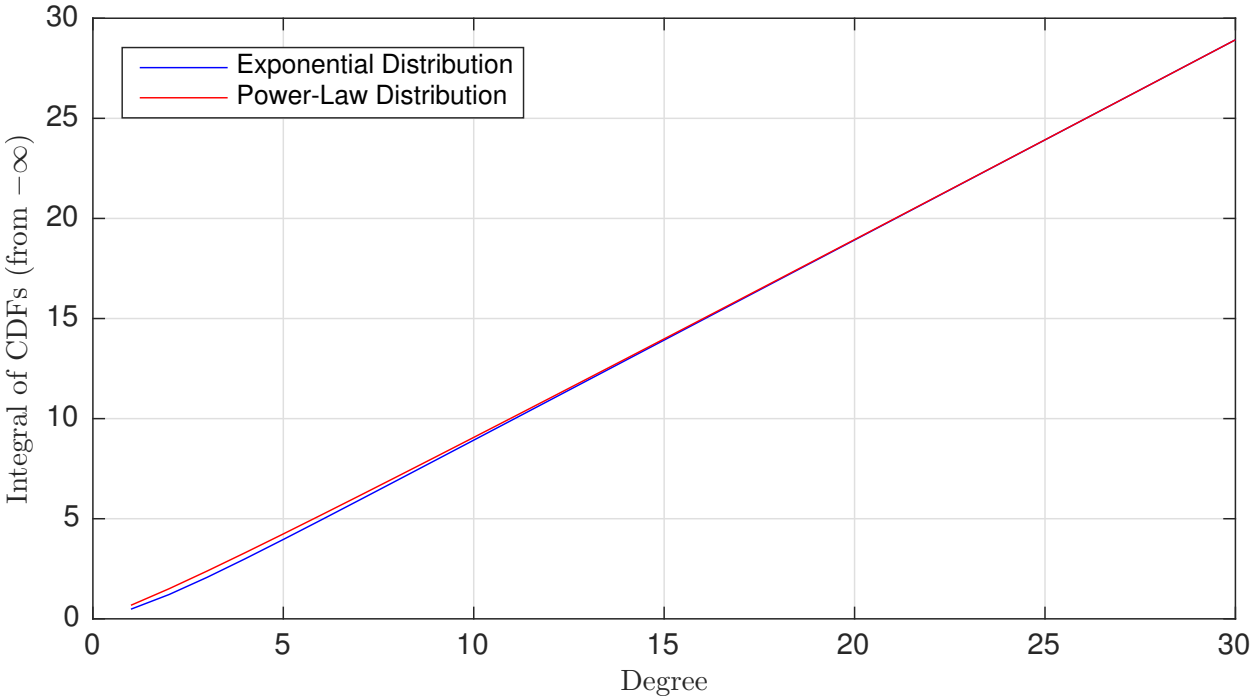


Figure 2.4: The integral of the CDFs from negative infinity showing the SOSD

it is less accurate to estimate parameters associated with a mean preserving spread of a distribution compared to the original distribution. This agrees with the intuition because it is always difficult to estimate a mean preserving spread, because it is a more dispersed distribution compared to the original distribution.

2.6 Summary

This chapter presented a comparison between the maximum likelihood estimation of network degree distributions from degree distributions of sampled networks for two cases: (1) the underlying network has an exponential degree distribution (2) the underlying network has a power-law degree distribution. It was assumed that the noise introduced from sampling/measurement process follows a zero mean Gaussian distribution (as a result of the central limit theorem) and that we know the type of the underlying distribution. Maximum likelihood estimators were derived for both models and then their variances were analyzed

using both mean squared errors and the Cramèr-Rao Bounds. The results show that the maximum likelihood estimator of the exponential model has a significantly smaller variance compared to the maximum likelihood estimator of the power-law model. Hence, maximum likelihood estimation has a better accuracy when the underlying network has an exponential degree distribution compared to a power-law degree distribution. The reason for this interesting behaviour was then conjectured to be the Second Order Stochastic Dominance of probability distributions. This was illustrated by considering a situation where the two distributions have the same mean.

Chapter 3

Identification of Individual Interests in Online Social Networks

3.1 Related Work

The first step of interest identification is to identify the missing nodes and how they connect to the rest of the nodes. For this step, we use the missing node identification by spectral clustering (MISC) algorithm proposed in [14, 15] which returns the prediction of the full network graph $\hat{G} = (\hat{V}, \hat{E})$ using only the number of missing nodes and an indication for the existence of missing nodes and edges. For example, this indication could be a Facebook post by an existing customer of the ABC Sports Car company mentioning about a friend who also might be interested in sports cars, but whose profile is not visible to the company. Similarly, these indications could also be Facebook posts, photos, tweets, etc. depending on the application. Image and text recognition software can be utilized to identify such indications in a practical implementation scenario. Readers are highly encouraged to read [14, 15] for a full description about the MISC algorithm and a comparison with the other network inference methods. Various other algorithms that are based on observations of a diffusion process are also presented in [11, 32, 44]. However, we use the MISC algorithm as the first step of our proposed theoretical framework since it fits our assumptions and model perfectly and then build the interest identification method from there.

Cayley-Menger Determinants are used to express the volumes of simplexes in Euclidean spaces in terms of the distances between the nodes [5, 48, 49]. This concept has been used

previously for applications in sensor networks, localization and robotics such as [30, 31, 51]. In our work, we use Cayley-Menger Determinants to obtain constraints relating the distances between the known and unknown interest vectors in order to formulate an optimization problem similar to the method proposed in [7].

The problem of the interest identification in a social network from a partial graph and partial set of interests has not been addressed before in the published literature to the best of our knowledge. Further, the novel model proposed in our work allows the use of results orienting from distance geometry such as Cayley-Menger Determinants which have not been utilized previously in the context of social network research. Therefore, we believe that this work will open new avenues for social network research.

The organization of the rest of this chapter is as follows. Section 3.2 defines the problem of interest identification and states the main assumptions of the work. Then, Section 3.3 describes the proposed interest identification method and Section 3.4 provides a numerical example to illustrate the important steps of the proposed method. Finally, Section 3.5 summarizes the work presented in this chapter with some concluding remarks.

3.2 Problem Definition and Assumptions

The identified social network graph $\hat{G} = (\hat{V}, \hat{E})$ (output of the MISC algorithm as described in Section 3.1) can be used to estimate the interests of the set of newly identified nodes $\hat{V}_m = \hat{V} \backslash V_k$. We can define this problem of identifying interests of nodes as below.

Problem (Interest Identification) Definition *: given a social graph* $\hat{G} = (\hat{V}, \hat{E})$ *where* $\hat{V} = \{V_k, \hat{V}_m\}$ *is a partition of* $\hat{V}$*, and a function* $h : \hat{V} \rightarrow \hat{W} \subseteq \mathbb{R}^d$ *where only* $\hat{V}$ *and* $h(V_k)$ *are known, find* $h(\hat{V}_m)$*.*

Our primary assumption in order to solve this problem is that the social network is a micro environment (a small network) which exhibits the Homophily behaviour. Homophily

principle states that there exists a connection between the social network of a person and his/her interests. Simply put, this principle states that similarities between people causes connections to be made between them (there is a tendency of nodes to be attached to other nodes that have similar characteristics) [27, 40]. In other words, Homophily behaviour states that *"there exists a connection between how people are connected (the underlying spatial structure) and their similarity of interests"*. We represent this assumption as a function, f, that maps some graph theoretic distance metric, $g(v_i, v_j)$, between each pair of nodes, (v_i, v_j), to some value that is approximately equal to the Euclidean distance between interest vectors of those nodes. Such distance metrics (domain of f) that are purely based on the structural properties of the graphs have been studied before for similar contexts in [14, 15]. Similar to our assumption, [18] specifically assumes a dependency between a set of feature vectors associated with the nodes and the network topology. This assumption can be mathematically formulated as:

***Main assumption**: There exists a function $g : \hat{V} \times \hat{V} \to B$ and a function $f : B \to \mathbb{R}^+$ such that $B \subseteq \mathbb{R}$ and $f \circ g : \hat{V} \times \hat{V} \to \mathbb{R}^+$ is given by*

$$f \circ g(\hat{v}_i, \hat{v}_j) = \|h(\hat{v}_i) - h(\hat{v}_j)\|^2 - \varepsilon_{i,j}, \quad \forall (\hat{v}_i, \hat{v}_j) \in \hat{V} \times \hat{V} \tag{3.1}$$

for some small $\varepsilon_{i,j}$.

3.3 Interest Identification Using Cayley-Menger Determinants

The function g can be selected to be one of the widely used graph theoretic metrics such as Gaussian distance, inverse squared shortest path, relative common friends [15] or some other metric defined to suit the network in concern. Then, the function $f \circ g$ can be thought of as a noisy function which maps each pair of nodes to the squared Euclidean distance between

the interests of those nodes. We can cluster the squared distances between interest vectors of all the pairs of nodes $(\hat{v}_i, \hat{v}_j) \in V_k \times V_k$ that have the same metric $g(\hat{v}_i, \hat{v}_j)$ as a set

$$C_{g(\hat{v}_i,\hat{v}_j)} = \{d_{ij}^2 | (\hat{v}_i, \hat{v}_j) \in g^{-1}(\{g(\hat{v}_i, \hat{v}_j)\}) \cap V_k \times V_k\} \tag{3.2}$$

where,

$$d_{ij}^2 = \|h(\hat{v}_i) - h(\hat{v}_j)\|^2. \tag{3.3}$$

This would give $|g(V_k \times V_k)|$ number of distinct clusters of the Euclidean distances of interest vectors. Then, a discrete function $f_d : g(V_k \times V_k) \to \mathbb{R}^+$ can be defined as

$$f_d(g(\hat{v}_i, \hat{v}_j)) = \bar{C}_{g(\hat{v}_i,\hat{v}_j)}, \quad \forall \hat{v}_i, \hat{v}_j \in V_k \tag{3.4}$$

where, $\bar{C}_{g(\hat{v}_i,\hat{v}_j)}$ is the arithmetic mean of the cluster $C_{g(\hat{v}_i,\hat{v}_j)}$.

This process can be repeated in a prior training stage to determine the best graph theoretic metric, g, from a set of candidate metrics $g_1, g_2, \ldots, g_m$ by minimizing a function of the summation of weighted variances of each cluster as

$$g = \underset{g_i}{\operatorname{argmin}} \sum\nolimits_{C_{g_i(\hat{v}_i,\hat{v}_j)}} \lambda_{g_i(\hat{v}_i,\hat{v}_j)} \times var(C_{g_i(\hat{v}_i,\hat{v}_j)}) \tag{3.5}$$

where $\lambda_{g_i(\hat{v}_i,\hat{v}_j)}$ is the weight assigned to the cluster corresponding to the value $g_i(\hat{v}_i, \hat{v}_j)$.

The discrete data points of function f_d, obtained by (3.4) can now be used to obtain a continuous function f by fitting a polynomial of degree n_{opt}. Then, the domain of f will be $[sup(g(V_k \times V_k)), inf(g(V_k \times V_k))]$. The optimal degree, n_{opt}, can be determined by a model order determination method such as Akaike Information Criterion (AIC). We assume that the known part of the network is large enough to make sure $g(\hat{V} \times \hat{V}) \subseteq [sup(g(V_k \times V_k)), inf(g(V_k \times V_k))]$.

Then, the function $f \circ g : \hat{V} \times \hat{V} \to \mathbb{R}^+$ can be used to obtain a noisy estimate of the squared Euclidean distance between any pair of nodes $(\hat{v}_i, \hat{v}_j) \in \hat{V} \times \hat{V}$. The noise free distance is known only for all $(\hat{v}_i, \hat{v}_j) \in V_k \times V_k$.

In a d dimensional Euclidean space, we can calculate the coordinates of any unknown point, if the accurate distances from that point to $d+1$ known points are available. For example, the coordinates of any point in a plane ($d = 2$) can be calculated by triangulation if the distance from that point to 3 known non-collinear points are available. Hence, to estimate the interest vector $h(\hat{v}_i)$ of $\hat{v}_i \in \hat{V}_m$, we need the distances between $h(\hat{v}_i)$ and $d+1$ known nodes. These distances can be obtained with some error using the previously obtained composite function $f \circ g$. However, the actual distances are not independent of each other due to the fact that they reside in a Euclidean space. [7] proposes a method which describes the dependencies between the distances of points in a Euclidean space as a set of quadratic equality constraints. Since the actual distances are not independent, the errors that are associated with the distances obtained using $f \circ g$ are also not independent. Hence, these dependencies between the errors can be used to find the set of errors which would yield the actual distances as close as possible. Therefore, we find an elegant adaptation of the method proposed in [7] (for sensor network localization) combined with results from [49] to suit our previously described model and assumptions in order to find $h(\hat{V}_m)$.

This method is based on the Cayley-Menger Determinants which are used to find the hypervolumes of simplexes. The volume, $V(p_0, \ldots, p_k)$, of a k-dimensional simplex with $k+1$ points $p_0, \ldots, p_k$ can be expressed as

$$V^2(p_0, \ldots, p_k) = \frac{(-1)^{k+1}}{2^k (k!)^2} \, CM(p_0, \ldots, p_k) \tag{3.6}$$

where,

$$CM(p_0, \ldots, p_k) = \begin{vmatrix} 0 & d_{01}^2 & \ldots & d_{0k}^2 & 1 \\ d_{01}^2 & 0 & \ldots & d_{1k}^2 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{k0} & d_{k1}^2 & \ldots & 0 & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{vmatrix} \tag{3.7}$$

and d_{ij} is the Euclidean distance between points p_i and p_j of the simplex. $CM(p_0, \ldots, p_k)$ is called the Cayley-Menger determinant of the points $p_0, \ldots, p_k$ [49].

The following theorem is a popular result in distance geometry which gives the necessary and sufficient conditions for the embeddability of a distance matrix in a Euclidean space. The full proof of it is given in [5] and an excellent description is also given in [49].

Theorem 1 (Global Embedding Theorem)**.** *A distance matrix of* $h+1$ *points* $D(p_0, p_1, \ldots, p_h)$ *is embeddable in a Euclidean space* E^n *of dimension* n *if and only if*

i. There exists a submatrix $R(p_0, p_1, \ldots, p_n)$ *of* $n+1$ *points of* D *so that, for every* $k = 1, 2, \ldots, n$*, the sign of* $CM(p_0, \ldots, p_k)$ *is equal to* $(-1)^{k+1}$*;*

ii. For every pair $p_i, p_j (i, j = n+1, ..., h)$

$$CM(p_0, \ldots, p_n; p_i) = 0 \tag{3.8}$$

$$CM(p_0, \ldots, p_n; p_j) = 0 \tag{3.9}$$

$$CM(p_0, \ldots, p_n; p_i, p_j) = 0 \tag{3.10}$$

The first part of Theorem 1 states that there exists a reference object whose n-dimensional volume does not vanish if the set of points are embeddable in the Euclidean space of n dimension. In 2-dimensional space, this is equivalent to finding 3 points that are not co-linear so that the area of the triangle formed by the three points does not vanish. Similarly, in 3-dimensional space, this is equivalent to finding four points that are not co-planer so that the volume of the object formed by the formed by the four points does not vanish. We fulfill this condition in our method by finding a set $V_R \subseteq V_k$ of $d+1$ nodes such that the interest vectors of the nodes in V_R form a simplex in $\mathbb{R}^d$ whose d dimensional volume is not zero i.e. $CM(h(V_R)) \neq 0$. Let the elements of $\hat{V}$ be re-labeled so that the indices from 1 to $d+1$ are

assigned to elements of V_R, thus making $V_R = \{\hat{v}_1, \hat{v}_2, \ldots, \hat{v}_{d+1}\}$. We further define,

$$\bar{d}_{ij}^2 = f \circ g(\hat{v}_i, \hat{v}_j),\ \forall \hat{v}_i, \hat{v}_j \in \hat{V}. \tag{3.11}$$

Then, we have

$$\bar{d}_{ij}^2 = d_{ij}^2 - \varepsilon_{i,j} \tag{3.12}$$

where $\varepsilon_{i,j}$ is the error of the function $f \circ g$. From the second part of Theorem 1, we know that

$$CM(h(V_R); h(\hat{v}_i)) = 0, \forall \hat{v}_i \notin V_R. \tag{3.13}$$

Therefore, for any $\hat{v}_i \in \hat{V}_m$, from equation 3.13 we get

$$\begin{vmatrix} 0 & d_{i,1}^2 & \ldots & d_{i,d+1}^2 & 1 \\ d_{i,1}^2 & 0 & \ldots & d_{1,d+1}^2 & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ d_{i,d+1} & d_{1,d+1}^2 & \ldots & 0 & 1 \\ 1 & 1 & \ldots & 1 & 0 \end{vmatrix} = 0. \tag{3.14}$$

The above expression can be decomposed as

$$\begin{bmatrix} d_{i,1}^2 & \ldots & d_{i,d+1}^2 & 1 \end{bmatrix} E \begin{bmatrix} d_{i,1}^2 \\ \vdots \\ d_{i,d+1}^2 \\ 1 \end{bmatrix} = 0. \tag{3.15}$$

where,

$$E = \begin{bmatrix} 0 & d_{1,2}^2 & \dots & d_{1,d+1}^2 & 1 \\ d_{1,2}^2 & 0 & \dots & d_{2,d+1}^2 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ d_{1,d+1} & d_{2,d+1}^2 & \dots & 0 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix}^{-1} \tag{3.16}$$

as proven in [7]. The existence of the matrix E is guaranteed by our selection of nodes such that $CM(h(V_R)) \neq 0$.

Using (3.12) and (3.15), we get

$$\begin{bmatrix} \bar{d}_{i,1}^2 + \varepsilon_{i,1} & \dots & \bar{d}_{i,d+1}^2 + \varepsilon_{i,d+1} & 1 \end{bmatrix} E \begin{bmatrix} \bar{d}_{i,1}^2 + \varepsilon_{i,1} \\ \vdots \\ \bar{d}_{i,d+1}^2 + \varepsilon_{i,d+1} \\ 1 \end{bmatrix} = 0 \tag{3.17}$$

which defines a quadratic surface of the variables $\varepsilon_{i,j},\ j \leq d+1$ associated with $\hat{v}_i \in \hat{V}_m$.

This can be formulated as a least squares problem where we minimize the sum of errors, $J = \varepsilon_{i,1}^2 + \varepsilon_{i,2}^2 + \ldots + \varepsilon_{i,d+1}^2$, subject to the derived quadratic constraint obtained by (3.17) which is a well studied optimization problem that can be solved using methods such as multiplier method [7]. Then, the resulting set of errors $(\varepsilon_{i,1}^2, \varepsilon_{i,2}^2, \ldots, \varepsilon_{i,d+1}^2)$ could be associated with the erroneous distances $(\bar{d}_{i,1}^2, \bar{d}_{i,2}^2, \ldots, \bar{d}_{i,d+1}^2)$ obtained from the composite function $f \circ g$ to obtain the estimate of the true distances. These distances can then be used to easily calculate the interest vector of the node in consideration using triangulation.

It should be noted that further quadratic constraints can be obtained by selecting a different set of elements for V_R, fulfilling the non-zero hypervolume condition. Then, the optimization problem could be solved subject to multiple quadratic constraints and thus reducing the search space in the optimization problem and yielding a better accuracy. However, this should be done depending on the available computational capacity. The interest

vectors of all the missing nodes can be identified by repeating this process for all nodes in $\hat{V}_m$.

3.4 Formulation of the Optimization Problem: A Numerical Example

Consider a missing node $\hat{v}_i \in \hat{V}_m$ of the identified social network $\hat{G}$ where the interests of the nodes lie in the 2-dimensional Euclidean space ($d = 2$). Assume that our reference nodes $V_R = \{v_1, v_2, v_3 \in V_k\}$ have the interests $h(v_1) = \begin{bmatrix} 0 & 0 \end{bmatrix}^T, h(v_2) = \begin{bmatrix} 1 & 8 \end{bmatrix}^T, h(v_3) = \begin{bmatrix} 2 & 3 \end{bmatrix}^T$ and let the outputs of the composite function in equation (3.1) be $\bar{d}^2_{i1} = 16, \bar{d}^2_{i2} = 25, \bar{d}^2_{i3} = 9$ (noisy squared distances between the missing node and the reference nodes).

By substituting $h(v_1), h(v_2)$ and $h(v_3)$ in (3.6), we can see that $V^2(v_1, v_2, v_3) = 42.25$, which fulfills the non-zero hypervolume condition of the part (i) of theorem 1 (can be seen from Fig. 3.1 as well). Using (3), we can obtain $d^2_{1,2} = 65, d^2_{1,3} = 13$ and $d^2_{2,3} = 26$ and then, substituting in (3.16) and (3.17), we can obtain the constraint of the optimization problem as:

$$2\varepsilon^2_{i,1} + \varepsilon^2_{i,2} + 5\varepsilon^2_{i,3} - 44\varepsilon_{i,1} - 32\varepsilon_{i,2} + 24\varepsilon_{i,3} + 2\varepsilon_{i,1}\varepsilon_{i,2} - 6\varepsilon_{i,1}\varepsilon_{i,3} - 4\varepsilon_{i,2}\varepsilon_{i,3} - 176 = 0. \quad (3.18)$$

Then, the errors can be calculated by minimizing $J = \varepsilon^2_{i,1} + \varepsilon^2_{i,2} + \varepsilon^2_{i,3}$ subject to (3.18). Solving optimization problems of this form has been extensively studied [22, 55] and can be done using many methods. Further, [7] also arrives at a similar optimization problem in context of sensor networks where they use the Lagrangian multiplier method as the first step.

The calculated errors can now be associated with the noisy distances to obtain the actual distances according to (3.12). This gives us three distances, d_{i1}, d_{i2} and d_{i3}, to the unknown interest vector from the three known vectors, $h(v_1), h(v_2)$ and $h(v_3)$. Hence, the interest of

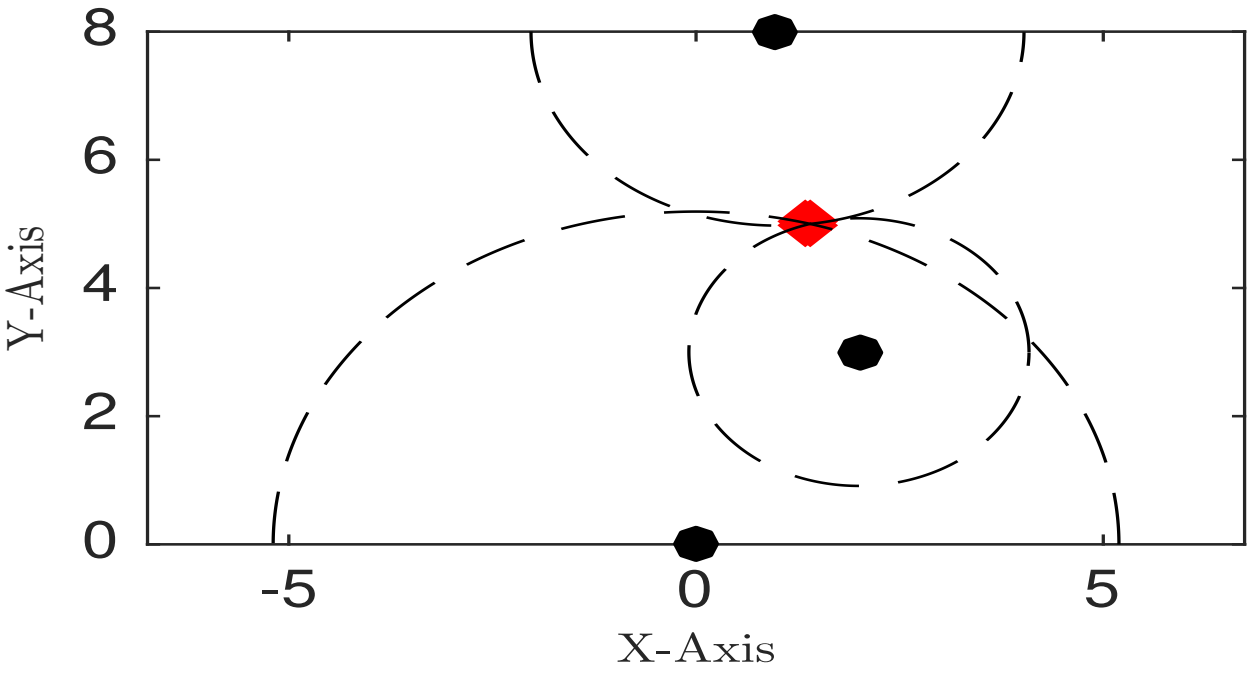


Figure 3.1: Illustration of the locations of the reference nodes (black dots) and triangulation process

the missing node, $h(v_i)$, can be calculated easily using triangulation as shown in Fig 3.1.

3.5 Summary

This chapter proposed a solution to the problem of interest identification in social networks. Only the partial graph and the interests of a subset of nodes are known, and thus making this problem non-trivial to solve.

Previous research has shown that there exists a relation between some graph theoretic distance metric and the difference of interests (Euclidean distance) of each pair of nodes due to the Homophily behaviour of social networks. This idea was utilized to obtain a composite function which gives the noisy interest distances from $d+1$ known nodes (in a d dimensional space) to a missing node. Then, equality constraints derived using Cayley-Menger Determinants were used to formulate an optimization problem in order to minimize noise and find the $d+1$ distances to the missing node which can then be used to triangulate

the interest associated with the missing node.

Chapter 4

Interest Influencing in Online Social Networks

4.1 Interest Influencing: Problem Definition and Assumptions

Consider the system model described in Section 1.1.2 again. In this chapter, we assume that $G = G_k$ (the network structure is known). As we pointed out in the Section 1.1.3, a connection between two persons can prevent a change of interests. We model this behaviour by assuming that a connection (an edge in the graph) in a social network fixes the distance between the interests of the two nodes involved in that connection. In other words, we assume that the interest distance between the nodes that are connected by an edge is not changing further. This assumption is particularly true in networks that are based on professional or research interests. In such networks, the similarity of interests (the fixed distance between interest vectors) between two nodes is what makes an edge exist in the graph between them. Hence, the interest distance should remain same for that edge to exist. This assumption can be mathematically formulated as:

***Assumption**: If $\langle \hat{v}_i, \hat{v}_j \rangle \in E$, then*

$$\frac{d\|h(\hat{v}_i) - h(\hat{v}_j)\|}{dt} = 0. \tag{4.1}$$

This assumption introduces distance constraints which fix the distances between the pairs of interests vectors associated with adjacent nodes. Therefore, the triple $G(h) = (V, E, h)$ can be viewed as a bar and joint framework where $G = (V, E)$ is the social graph and $h : V \to \mathbb{R}^d$ is the function that maps each node in the social network to its interest. Moving such a framework in the Euclidean space that it lives does not change the distances between the pairs of interest vectors that correspond to adjacent nodes. However, this does not necessarily ensure that the distances between non-adjacent vertices of the framework $G(h)$ will remain fixed. Hence, we are interested in finding a method to ensure that all interest vectors (whether they are adjacent or not) keep their inter-distances fixed. Based on the previous assumption, we can formally state this problem as:

Problem (Interest Influencing) Definition *: Given a bar and joint framework $G(h) = (V, E, h)$, find $E_n \subseteq V \times V$ so that $G_n(h) = (V, E \cup E_n, h)$ is a rigid framework.*

The term *rigid framework* in problem 2 definition refers to a framework where every motion preserves the distances between all pairs of vertices (*flexible* otherwise). Rigidity theory is the branch in mathematics which studies the rigidity and flexibility of bar and joint frameworks in Euclidean spaces [23, 45, 53]. As a simple example, a triangle formed by three rods and three joints is rigid in 2 dimensional space while a rectangle formed by four rods and four joints is flexible in 2 dimensional space since one rod of the rectangle can be moved while keeping another one fixed. The rectangle can be made rigid by adding a diagonal rod which will eliminate the additional degree of freedom and thus making the framework rigid. The rigidity theory develops a formal theoretical treatment for analysis of rigidity of bar and joint frameworks. It has been shown that the rigidity of frameworks in 2 dimensional Euclidean space can be analyzed purely by the properties of the underlying graph (neglecting the coordinates of joints). The concepts from rigidity theory have been utilized previously for sensor network localization purposes in [3, 13, 47]. In our work, we use rigidity theory to solve the second problem where the ultimate objective is to create

rigid graph in 2 dimensional space. For example, if the interests of the customers form such a rigid framework, we can fix the interests of all nodes by fixing the interests of few loyal customers.

The organization of the rest of this chapter is as follows. Section 4.2 describes the proposed method of solving the problem of interest influencing using the concept of rigidity theory. Section 4.3 provides numerical examples for the proposed method. Finally, Section 4.4 provides a summary of the work presented in this chapter and some concluding remarks.

4.2 Interest Influencing Using Rigidity Theory

Finding if a framework $G(h) = (V, E, h)$ is rigid in d dimensional space involves a solving set of $|E|$ quadratic equations with $d|V|$ unknowns that fixes the distances of the set of edges of the framework,

$$(h_i(t) - h_j(t))^T(h_i(t) - h_j(t)) = c_{ij}, \forall\langle\hat{v}_i, \hat{v}_j\rangle \in E \tag{4.2}$$

where we have denoted $h(v_i)$ by h_i. The framework is said to be to rigid, if all solutions to the system of quadratic equations preserve the distances between all points irrespective of whether they are adjacent or not. Since finding a solution to this quadratic system of equations is hard, the common approach is to linearize the problem by obtaining first derivatives (which simplifies the problem to finding the rank of a matrix). However, in the context of social networks, computation of the rank of such matrices may not be practical due to the large size.

As a solution to this, we can use Laman's theorem [35] which completely characterizes minimally rigid (not rigid after removing one of its edges) graphs in two dimensional space.

Theorem 2 (Laman's Theorem)**.** *Let a graph G have exactly $2|V|-3$ graph edges($|V| > 1$). Then, G is rigid in $\mathbb{R}^2$ if and only if no subgraph $G' = (V', E')$ has more than $2|V'|-3$ edges ($|V'| > 1$)*

A necessary (but not sufficient) condition for a graph to be rigid is to have at least $2|V|-3$ edges [50]. However, according to Laman's theorem, these edges have to be well distributed in order to make a rigid graph. Hence, Laman's theorem gives a purely combinatorial method to analyze the rigidity of frameworks in 2 dimensions based only on the underlying graph (without considering the coordinates of vertices). In other words, rigidity of a framework $G(h) = (V, E, h)$ in 2 dimensional space depends only on $G = (V, E)$. The readers encouraged to refer chapter 2 of [50] for an excellent description about Laman's theorem.

However, such a combinatorial type of a theorem to assess the rigidity of frameworks in dimensions of three and beyond is currently not available. Hence, the graph theoretic analysis of rigidity of frameworks in dimensions beyond two dimensions is not well understood and not very useful from a computational point of view due to the lack of polynomial time algorithms [50]. Hence, if we assume the number of parameters associated with nodes to be two ($d = 2$), the Laman's theorem and various combinatorial polynomial time algorithms that are based on it such as the pebble game algorithm [28] can be utilized to identify missing edges required to create a rigid network based on our system model and the assumption, thus giving the solution to the problem of interest influencing.

4.3 Assessing Rigidity in 2-Dimensional Space: Numerical Examples

Here, we provide 3 simple examples to illustrate how we can assess the rigidity of frameworks based on the Laman's theorem in a purely combinatorial manner.

Assume that our social network has 5 nodes and their interests lie in the two dimensional space as represented in Fig.4.1. In the case of Fig.4.1(a), it can be seen that the number of edges in the graph (6) is less than $2|V| - 3 = 7$. Hence, the first graph does not satisfy a necessary condition for rigidity and hence does not form a rigid structure. The structure in Fig. 4.1(b) is formed by adding an additional edge to the structure in Fig. 4.1(a) and

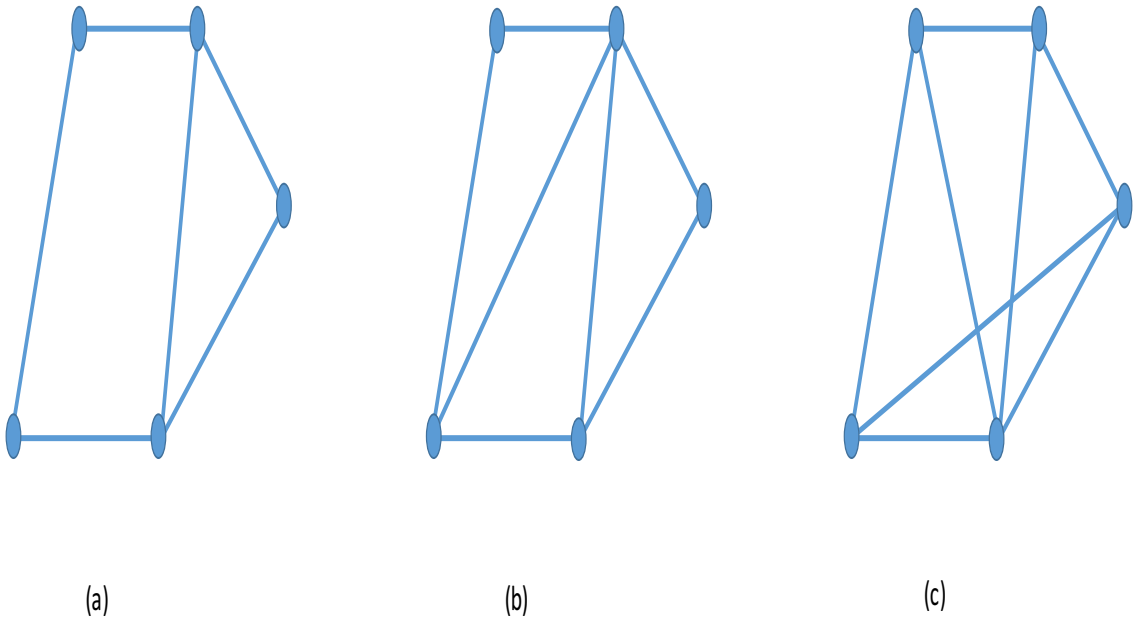


Figure 4.1: (a) Flexible graph (b) A minimally rigid graph (c) Redundantly rigid graph in two dimensional space

it can be seen that it satisfies the $2|V| - 3$ edge count. Further, no subgraph $G' = (V', E')$ of it has more than $2|V'| - 3$ edges. Therefore, the structure in Fig. 4.1(b) is minimally rigid. The structure in 4.1(c) has $8 > 2|V| - 3 = 7$ edges and hence satisfy the constraint for the minimal number of edges. If we remove the diagonal edge of the rectangle, then the subgraph formed by the rectangle has only 4 edges and hence it will make the structure non-rigid. However, if we remove the top edge of the lower triangle, then the graph will satisfy the minimal rigidity conditions and will be a rigid graph.

In the previous examples, we prefer to have the structures Fig.4.1 (b) or (c) in our social network since both of them are rigid. If our social network is of the form Fig.4.1(a), we would like to introduce the two nodes in order to add the diagonal edge which makes the structure rigid. The pebble game algorithm proposed in [28] is a method to do the same analysis we did above by assigning pebbles to a selected set of vertices initially. It is a polynomial time algorithm which is based on the Laman's theorem. In practice, such polynomial time algorithms can be utilized to do the same analysis we did above, especially when the network

is of large size.

4.4 Summary

This chapter proposed a solution to the problem of interest influencing in online social networks. The problem was first formulated as a classic problem in rigidity theory. Then, the Laman's theorem which completely characterizes a minimally rigid graph in 2 dimensional space was utilized to find the edges that should be added to make a rigid graph based on the assumption that the interests of the nodes are 2 dimensional vectors. This assumption was required due to lack of a Laman's type of a theorem to analyze rigidity of frameworks using only the combinatorial properties of the underlying graph.

Chapter 5

Conclusions and Future Work

This thesis explored three important problems related to the online social networks utilizing tools from statistical signal processing and applied mathematics.

The second chapter presented a comparison of the maximum likelihood estimates of the network degree distributions from the sampled network degree distributions for two cases: 1) the underlying network has an exponential degree distribution (2) the underlying network has a power-law degree distribution. Assuming that the noise introduced from sampling/measurement process follows a zero mean Gaussian distribution, maximum likelihood estimates were derived for both models and their variances were analyzed using the Cramèr-Rao Bounds and mean squared errors. Using the results of these theoretical and numerical comparisons, it was concluded that the maximum likelihood estimation has a significantly smaller variance (approximately an order of magnitude) for exponential model compared to the power-law model in the respective regions of interest. The reason for this interesting behaviour was then conjectured to be the Second Order Stochastic Dominance of probability distributions. This was illustrated by considering a situation where the two distributions have the same mean.

Devising a formal proof for the presented conjecture involving the Second Order Stochastic Dominance would be an interesting research direction for this work. Another interesting research direction of the presented work would be to compare the model based estimation of the degree distributions for different random graph models.

The third chapter proposed a solution to the problem of interest identification in social networks. The main reason for this problem to be non-trivial is the fact that only the

partial graph and the interests of a subset of nodes are known. Previous research has shown that there exists a relation between some graph theoretic distance metric and the difference of interests (Euclidean distance) of each pair of nodes due to the Homophily behaviour of social networks. This idea was utilized to obtain a composite function which gives the noisy interest distances from $d+1$ known nodes (in a d dimensional space) to a missing node. Then, equality constraints derived using Cayley-Menger Determinants were used to formulate an optimization problem in order to minimize noise and find the $d+1$ distances to the missing node which can then be used to triangulate the interest associated with the missing node.

An interesting future research direction of the presented work would be to extend the proposed framework to random environments where the parameters associated with nodes are random vectors and/or the underlying spatial structure is a random graph.

The fourth chapter considered the problem of interest influencing in an online social network i.e. how to create a network structure in order to gain a maximum influential power over the interests of people in the network? This problem was formulated as a classic problem in rigidity theory assuming the contagion behaviour of individuals in the network. This reduces the problem of interest influencing to finding the edges that should be added to the network to make it a minimally rigid graph. Then, the Laman's theorem which completely characterizes a minimally rigid graph in two dimensional space was utilized to find the edges that should be added to make a rigid graph based on the assumption that the interests of the nodes are two dimensional vectors. This assumption was required due to lack of a Laman's type of a theorem to analyze rigidity of frameworks using only the combinatorial properties of the underlying graph.

In future work, we intend to utilize real world data to further explore the theoretical framework proposed in this work and also to extend this work to random environments. Further, it would be interesting to explore how the rigidity theory concepts of body-bar frameworks (instead of bar and joint frameworks) can be related to social networks.

Bibliography

[1] Facebook, Inc. (2016, Oct. 10). *Company information: Stats* [Online]. Available: https://newsroom.fb.com/company-info/

[2] Twitter, Inc. (2016, Jun. 30). *Twitter usage* [Online]. Available: https://about.twitter.com/company

[3] J. Aspnes, T. Eren, D. K. Goldenberg, A. S. Morse, W. Whiteley, Y. R. Yang, B. Anderson, and P. N. Belhumeur, "A theory of network localization," *IEEE Transactions on Mobile Computing*, vol. 5, no. 12, pp. 1663–1678, Dec. 2006.

[4] H. Bauke, "Parameter estimation for power-law distributions by maximum likelihood methods," *The European Physical Journal B*, vol. 58, no. 2, pp. 167–173, Jul. 2007.

[5] L. M. Blumenthal, *Theory and Applications of Distance Geometry*. Chelsea House Pub, 1970.

[6] P. E. Caines, *Linear Stochastic Systems*. John Wiley & Sons, 1988.

[7] M. Cao, B. D. O. Anderson, and A. S. Morse, "Sensor network localization with imprecise distances," *Systems & Control Letters*, vol. 55, no. 11, pp. 887–893, Nov. 2006.

[8] K. Choromański, M. Matuszak, and J. Mikisz, "Scale-free graph with preferential attachment and evolving internal vertex structure," *Journal of Statistical Physics*, vol. 151, no. 6, pp. 1175–1183, Jun. 2013.

[9] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in medicine*, vol. 32, no. 4, pp. 556–577 , Feb. 2013.

[10] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, Nov. 2009.

[11] H. Daneshmand, M. Gomez-Rodriguez, L. Song, and B. Schoelkopf, "Estimating diffusion network structures: Recovery conditions, sample complexity & soft-thresholding algorithm," in *Proceedings of the International Conference on Machine Learning, ICML'14*, 2014, pp. 793-801.

[12] W. Deng, W. Li, X. Cai, and Q. A. Wang, "The exponential degree distribution in complex networks: Non-equilibrium network theory, numerical simulation and empirical data," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 8, pp. 1481–1485, Apr. 2011.

[13] T. Eren, O. Goldenberg, W. Whiteley, Y. Yang, A. Morse, B. Anderson, and P. Belhumeur, "Rigidity, computation, and randomization in network localization," in *Twenty-third Annual Joint Conference of the IEEE Computer and Communications Societies, INFOCOM'04*, 2004, pp. 2673–2684.

[14] R. Eyal, S. Kraus, and A. Rosenfeld, "Identifying missing node information in social networks," in *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence, AAAI'11*, 2011, pp. 1166–1172.

[15] R. Eyal, A. Rosenfeld, S. Sina, and S. Kraus, "Predicting and identifying missing node information in social networks," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 3, pp. 1–35, Jun. 2013.

[16] O. Frank, "Estimation of the number of vertices of different degrees in a graph," *Journal of Statistical Planning and Inference*, vol. 4, no. 1, pp. 45–50, Jan. 1980.

[17] O. Frank, "A survey of statistical methods for graph analysis," *Sociological methodology*, vol. 12, pp. 110–155, Jan. 1981.

[18] A. Freno, G. Garriga, and M. Keller, "Learning to recommend links using graph structure and node content," in *Neural information processing systems workshop on choice models and preference learning*, 2011.

[19] O. N. Gharehshiran, W. Hoiles, and V. Krishnamurthy, "Detection of homophilic communities and coordination of interacting meta-agents: A game-theoretic viewpoint," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 2, no. 1, pp. 84–101, Mar. 2016.

[20] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *29th Conference on Computer Communications, INFOCOM'10*, 2010, pp. 1–9.

[21] S. Goel and D. G. Goldstein, "Predicting Individual Behavior with Social Networks," *Marketing Science*, vol. 33, no. 1, pp. 82–93, Oct. 2013.

[22] G. H. Golub and U. Matt, "Quadratically constrained least squares and quadratic problems," *Numerische Mathematik*, vol. 59, no. 1, pp. 561–580, Dec. 1991.

[23] J. E. Graver, B. Servatius, and H. Servatius, *Combinatorial Rigidity*. American Mathematical Society, 1993.

[24] M. Hamdi, V. Krishnamurthy, and G. Yin, "Tracking a markov-modulated stationary degree distribution of a dynamic random graph," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6609–6625, Oct. 2014.

[25] M. S. Handcock and K. J. Gile, "Modeling social networks from sampled data," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 5–25, Oct. 2010.

[26] E. J. Hannan and M. Deistler, *The statistical theory of linear systems*. SIAM, 1988.

[27] M. O. Jackson, "Networks in the Understanding of Economic Behaviors," *Journal of Economic Perspectives*, vol. 28, no. 4, pp. 3–22, Nov. 2014.

[28] D. J. Jacobs and B. Hendrickson, "An algorithm for two-dimensional rigidity percolation: The pebble game," *Journal of Computational Physics*, vol. 137, no. 2, pp. 346 – 365, Nov. 1997.

[29] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Inc., 1993.

[30] U. A. Khan, S. Kar, and J. M. F. Moura, "Linear theory for self-localization: Convexity, barycentric coordinates, and cayley-menger determinants," *IEEE Access*, vol. 3, pp. 1326–1339, Aug. 2015.

[31] U. Khan, S. Kar, and J. Moura, "Distributed Sensor Localization in Random Environments Using Minimal Number of Anchor Nodes," *IEEE Transactions on Signal Processing*, vol. 57, no. 5, pp. 2000–2016, May 2009.

[32] M. Kim and J. Leskovec, "The network completion problem: Inferring missing nodes and edges in networks." in *Proceedings of the 2011 SIAM International Conference on Data Mining, SDM'11*, 2011, pp. 47–58.

[33] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[34] V. Krishnamurthy, O. Gharehshiran, and M. Hamdi, *Interactive Sensing and Decision Making in Social Networks*, ser. Foundations and Trends in Signal Processing. Now Publishers, 2014.

[35] G. Laman, "On graphs and rigidity of plane skeletal structures," *Journal of Engineering mathematics*, vol. 4, no. 4, pp. 331–340, Oct. 1970.

[36] K. Lerman, X. Yan, and X. Z. Wu, "The "majority illusion" in social networks," *PloS one*, vol. 11, no. 2, p. e0147617, Feb. 2016.

[37] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'06*, 2006, pp. 631–636.

[38] K. Libert. (2016, Feb. 23). *Your Network's Structure Matters More than Its Size* [Online]. Available: https://hbr.org/2016/02/your-networks-structure-matters-more-than-its-size

[39] L. Lovsz, "Random walks on graphs: A survey," 1993.

[40] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, Aug. 2001.

[41] S. Meyn and R. L. Tweedie, *Markov Chains and Stochastic Stability*, 2nd ed. Cambridge University Press, 2009.

[42] W. Ogryczak and A. Ruszczynski, "Dual stochastic dominance and related mean-risk models," *SIAM Journal on Optimization*, vol. 13, no. 1, pp. 60–78, May 2002.

[43] B. Ribeiro and D. Towsley, "Estimating and sampling graphs with multidimensional random walks," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 390–403.

[44] M. G. Rodriguez, D. Balduzzi, and B. Schlkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proceedings of the 28th International Conference on Machine Learning, ICML'11*, 2011, pp. 561–568.

[45] B. Servatius and H. Servatius, "Generic and abstract rigidity," in *Rigidity Theory and Applications*. Springer, 2002, pp. 1–19.

[46] C. R. Shalizi and A. C. Thomas, "Homophily and Contagion Are Generically Confounded in Observational Social Network Studies," *Sociological Methods & Research*, vol. 40, no. 2, pp. 211–239, May 2011.

[47] A. Singer and M. Cucuringu, "Uniqueness of low-rank matrix completion by rigidity theory," *SIAM Journal on Matrix Analysis and Applications*, vol. 31, no. 4, pp. 1621–1641, Feb. 2010.

[48] M. J. Sippl and H. A. Scheraga, "Solution of the embedding problem and decomposition of symmetric matrices." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 8, pp. 2197–2201, Apr. 1985.

[49] M. J. Sippl and H. A. Scheraga, "Cayley-menger coordinates," *Proceedings of the National Academy of Sciences*, vol. 83, no. 8, pp. 2283–2287, Apr. 1986.

[50] A. Sljoka, "Counting for Rigidity, Flexibility and extensions via the pebble game algorithm," Ph.D. dissertation, York University, Toronto, 2006.

[51] F. Thomas, E. Ottaviano, L. Ros, and M. Ceccarelli, "Performance analysis of a 3-2-1 pose estimation device," *IEEE Transactions on Robotics*, vol. 21, no. 3, pp. 288–297, Jun. 2005.

[52] R. Van Der Hofstad. (2016, May 01). *Random graphs and complex networks* [Online]. Available: http://www.win.tue.nl/rhofstad/NotesRGCN.pdf

[53] W. Whiteley, "Rigidity of Molecular Structures: Generic and Geometric Analysis," in *Rigidity Theory and Applications*. Springer, 2002, pp. 21–46.

[54] Y. Zhang, E. D. Kolaczyk, and B. D. Spencer, "Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks," *Annals of Applied Statistics*, vol. 9, no. 1, pp. 166–199, Mar. 2015.

[55] Z. Zhang and Y. Huang, "A projection method for least squares problems with a quadratic equality constraint," *SIAM Journal on Matrix Analysis and Applications*, vol. 25, no. 1, pp. 188–212, Jan. 2003.

Appendix A

Methods of Sampling Networks: A Brief Survey

In this Appendix, we provide a brief survey of the literature about the widely used methods for network sampling. A detailed description of each method can be found in the references provided.

A.1 Sampling by Random Node Selection

A.1.1 Random Node Sampling

This is the most basic method that falls under the Random Node Selection category. As the name itself implies, this method works by selecting a set of nodes uniformly. Let the social graph (undirected) be $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_N\}$ and let the set of sampled nodes be V_S. We need $|V_S| \leq N_{max}$ as a constraint. Then the Random Node sampling algorithm is as below.

1. Generate $X = Int(NU) + 1$ where $U \sim Unif(0, 1)$ and $N = |V|$. (The $Int(.)$ operator extracts the integer part of a number)
2. $V_S \longleftarrow V_S \cup v_X$
3. If $|V_S| \leq N_{max}$, go back to step 1.
4. The sampled graph, G_S is the graph induced by the set of nodes V_S

The first step in the above algorithm transforms the uniform distribution in to a step function (inverse transform approach). Many other variations of this method include methods obtained by using weighted sampling instead of uniform sampling. One such example is the Random Degree Node (RDN) sampling where probability of a node being selected is proportional to its degree. However, this method will have a very high bias towards nodes with higher degree [37]. Hence, this would not be a good choice if we want to estimate the degree distribution. Further, most of the social networks randomly sample the User IDs of the users in order to obtain random node samples. This is undesirable in the case where User IDs are sparsely populated (large number of valid user IDs are not assigned) [43]. In such cases, we have to use overly complicated versions of this method to reduce the large hit to miss ratio that may occur due to the unpopulated User IDs.

Further, it has also been proven that this method does not retain many interesting characteristics that are unique to social networks such as the power law in degree distribution [37].

A.1.2 Breadth-First Sampling (BFS)

Breadth-First sampling is another sampling algorithm that falls in to the category of sampling by random node selection. BFS uses two queues The steps of BFS algorithm are as below (using the same model as before).

1. Begin two empty queus: *Sampled* and *Processed*
2. Select $v_i \in V$ (seed) uniformly at random and insert it into *Sampled* and store the neighbours of v_i in *Processed*
3. Move the earliest explored node (last node in *Processed*) to the *Sampled* and insert all its neighbours (that are not already inserted) to *Processed*.
4. Repeat step 3 until length of *Sampled* reach N_{max}

According to [20], BFS is biased towards high degree nodes and also no statistical properties can be theoretically proven for this method.

A.2 Sampling by Random Edge Selection

Similar to the idea of sampling by random node selection, the idea behind sampling by random edge selection is to select edges uniformly at random and then obtain the subgraph created by that set of edges as the sampled graph.

However, a problem associated with this method in the context of social network sampling is that the sampled graph diameter will be extremely large (which does not agree with the social network community structure.). Since high degree nodes have more edges attached to them, this method is also biased towards the high degree nodes. A variation of Random Edge Selection which was proposed to eliminate this bias is Random Node Edge Sampling. In Random Node Edge Sampling, a node is first selected uniformly at random and then an edge incident to that node is picked uniformly at random [37]. However, if the edges do not have IDs (which is the practical scenario in social networks), then it may be hard and resource intensive to sample edges.

A.3 Sampling by Exploration

A.3.1 Random Walk Sampling

As pointed above, both random node and edge sampling have certain weaknesses that are undesirable in the context of social networks. An alternate, computationally less expensive method which also reduces most of the weaknesses in previous methods is to sample a network by means of a random walk [37, 43]. The Random Walk Sampling is explained next.

Assume the social network in concern is an undirected graph $G = (V, E)$ where $V = \{v_1, v_2, \ldots, v_n\}$. Let X_n be the random variable which gives the index of the node that the

random walker is on at the time step n. Then, clearly $\{X_n\}$ is a discrete time Markov chain and

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \begin{cases} \frac{1}{d(i)}, & \text{if}(i,j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{A.1}$$

where $d(i)$ denotes the degree of the node v_i. Hence, it can be seen that the Markov chain is also Homogeneous. In Random Walk (RW) sampling, we initiate a random walker at a node uniformly selected at random. Then, at each time step k, the walker moves from its current position v_i to v_j and the edge (v_i, v_j) is added to the sequence of edges that have already been sampled.

Under certain conditions, we can use RW sampling to create estimators of graph characteristics that are asymptotically unbiased (converge to the actual characteristic value when the sample size goes to infinity). This follows from an important property of random walks called detailed balance which is proven below (This is mentioned without a proof in [39]).

Theorem 3. *A random walk on a connected graph has a unique stationary distribution that satisfies the detailed balance.*

Proof. Assume the graph is connected. Therefore, the random walk is irreducible (since there is a path from each node to every other node). Then, by the law of large numbers for Markov chains, it has a unique stationary distribution which we denote by π. We must show,

$$P_{ij}\pi(i) = P_{ji}\pi(j) \ \forall i, j$$

. We proceed with proof by cases.

Case 1: $(i,j) \notin E$

Then $P_{ij} = P_{ji} = 0$ and therefore, the detailed balance equation is satisfied.

Case 2: $(i,j) \in E$

Let π be the vector that satisfies the detailed balance. Then,

$$P_{ij}\pi(i) = P_{ji}\pi(j)\ \forall i, j$$

$$\frac{\pi(i)}{d(i)} = \frac{\pi(j)}{d(j)} = k \text{ where } d(.) \text{ denotes the degree of the vertex.}$$

Further,

$$\sum_{i \in \chi} \pi(i) = 1.$$

Therefore,

$$k = \frac{1}{\sum_{v_i \in V} d(i)}.$$

Therefore, the Markov chain has a unique stationary distribution that satisfies the detailed balance.

□

Further, it should be noted that the Random Walk Sampling on connected, non-bipartite graphs sample edges uniformly at random when it is at the stationary distribution. In other words, sampling by random edge selection is similar to the random walk sampling at the stationary distribution. This is mentioned without proof in [43] and we present it as a formal statement with a proof.

Theorem 4. *A random walk on a connected non-bipartite graph sample edges uniformly at random as the step size reaches infinity.*

Proof. Assume that the random walk is on a connected, non-bipartite graph. Then, from Theorem 3, it has a unique stationary distribution. Further, since the graph is non-bipartite, the random walk is aperiodic. Then, by the Convergence of Marginals, the marginal distribution over nodes will reach the stationary distribution as the step size goes to infinity.

Assume the Markov chain is in stationary distribution at step n. Then, the probability of the random walker being in the node v_j is

$$(P^T\pi)_j = \frac{d(j)}{\sum_{v_i \in V} d(i)} \tag{A.2}$$

from Theorem 3.

Then, the probability of the edge sampled at step $n+1$ being (j,k), denoted by $\pi_e(j \longrightarrow k)$ is

$$\pi_e(j \longrightarrow k) = \frac{d(j)}{\sum_{v_i \in V} d(i)} \times \frac{1}{d(j)} \tag{A.3}$$

$$= \frac{1}{\sum_{v_i \in V} d(i)} \tag{A.4}$$

$$= \frac{1}{2m} \tag{A.5}$$

where m is the number of the undirected edges in the graph. Hence, the distribution over the edges is uniform which concludes the proof.

□

Next, with the aid of the above two theorems and a corollary of Law of Large Numbers for Markov Chains which was stated in [41], we show how Random Walk sampling can be used to create asymptotically unbiased estimators of various graph characteristics (which is one of the advantages of Random Walk sampling over the other discussed methods).

The following is a simpler version of the Theorem 17.2.1 of [41] (which is another version of the law of large numbers) with a simpler proof (than the original one given in [41]) which we will use to create unbiased estimators using random walk sampling.

Theorem 5. *For a random walk $\{X_n\}$, on a connected graph,*

$$\frac{\sum_{i=1}^{N} f(X_i)}{\sum_{i=1}^{N} g(X_i)} \xrightarrow{a.s.} \frac{\mathbb{E}_\pi[f(X)]}{\mathbb{E}_\pi[g(X)]} \tag{A.6}$$

where $E_\pi(.)$ denotes the expectation with respect to the probability measure π (stationary distribution)

Proof. Proof follows by noting that this is simply division of the terms of the law of large numbers for Markov chains.

□

[43] provides an extension of the above theorem as below with a full proof (which follows trivially from Theorem 5).

Theorem 6. *Let $E^* \subseteq E$ be nonempty. Let (u_i, v_i) be the i^{th} sampled edge such that $(u_i, v_i) \in E^*$; and let $B^*(B)$ be the number of such samples, where B is the number of total random walk steps. Then, for any function f, where $\sum_{(u,v)\in E^*} |f(u,v)| \leq \infty$,*

$$\lim_{B\to\infty} \frac{1}{B^*(B)} \sum_{i=1}^{B^*(B)} f(u_i, v_i) \xrightarrow{a.s.} \frac{1}{|E^*|} \sum_{\forall(u,v)\in E^*} f(u,v) \tag{A.7}$$

It can be seen that the Theorem 6 (which is an extension of Theorem 5) can be utilized to create asymptotically unbiased estimators. [43] also provides examples of four such estimators that estimates graph characteristics such as edge label density and global clustering coefficient by selecting the function f to suit the quantity that we are trying to estimate.

Hence, it can be seen that sampling graphs (especially social networks) using random walk based methods has more advantages compared to the other two sampling methods in terms of bias of the estimate and complexity.

A.3.2 Frontier Sampling

A recent advance in previously discussed Random Walk based methods in graph sampling is to use multiple dependent random walkers. This method, known as Frontier Sampling [43], can overcome the disadvantages of random walk sampling in cases where there are large number of disconnected components in the graph. In such cases, a single random walker

may get stuck in an isolated area and may provide estimates that represent only the said isolated area, thus causing large estimation errors. Frontier sampling reduces this weakness by employing multiple simultaneous random walkers to explore the graph, while retaining all the other nice properties of the random walk sampling methods [43].