UNSW
AUSTRALIA

EXPONENTIAL-FAMILY RANDOM GRAPH MODELS FOR EGOCENTRICALLY-SAMPLED DATE WITH DIRECTED RELATION

Haotian Guo

Supervisor: Dr Pavel Krivitsky

School of Mathematics and Statistics

UNSW Sydney

July 2021

Submitted in partial fulfilment of the requirements of the degree of

Master of Statistics

Plagiarism statement

I declare that this thesis is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere.

I acknowledge that the assessor of this thesis may, for the purpose of assessing it:

- Reproduce it and provide a copy to another member of the University; and/or,
- Communicate a copy of it to a plagiarism checking service (which may then retain a copy of it on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct, and am aware of any potential plagiarism penalties which may apply.

By signing this declaration I am agreeing to the statements and conditions above.

Signed: _____ Date: _____

Acknowledgements

By far the greatest thanks must go to my supervisor for the guidance, care and support they provided.

Haotian Guo, 11 July 2021.

Abstract

Contents

Chapter 1

Introduction

Since Georg.Simmel first addressed issues related to social networks in early 20th century, networks have become a focus of discussion in science and everyday life [1]. Scientists have also been highly interested in the problem of interconnected networks among individuals. The scientific study of networks, including information networks, social relation networks, and biological networks, has generated interesting discussions, and most of the research on these networks is interdisciplinary. For example, network models have been analyzed from the perspectives of economics, statistics and statistical physics [2]. Especially after 2000, the interest in the study of network data has intensified due to the emergence of social media. Statistical modeling of network data analysis has become an important topic in various fields.

When the goal is to investigate the whole network properties, the traditional approach used for studying networks is to investigate all the relationships of the network in the population of interest. A network containing N nodes (surveyed people) can in principle have N^2 edges, so the size of the network becomes a key factor affecting research in this area from both computation and data collection aspects [3]. Although the wide availability of computers has made it possible to collect and analyze networks on a bigger scale, it still makes the research very onerous. This is sometimes even impossible as the collection of data is not feasible in many empirical setting. In order to be able to study the network beyond the limits, the goal is to use statistical framework to obtain inference of network properties from subset of network data. So two ways of sampling, Link-Trace designs and

Egocentric designs, are used in research increasingly [4]. Such sub-data can also be used to study the properties of the complete networks, provided that the method is applied properly.

Link-Trace design is generally used to sample from the hard-to-reach population []. The sampling will start with a small number of seed nodes , each of which will nominate a number of "alters" based on the designed question, and these alters will nominate new alters as new respondents. Each new set of alters will be called " wave" or "generation". Within this sampling category, there are two variables that determine the different sampling methods. The first is how new alters are selected as respondents in new wave, such as investigator-driven and respondent-driven. The second is the number of generations of sampling. These investigations include snowball samples, chain referral, adaptive cluster sampling and so on [5]. Handcock and Gile [6] have used these sampling designs to analyze the statistics of the observed networks and have obtained some results.

In this paper, we will use the second sampling design: egocentric designs. This approach significantly lowers the empirical threshold for network research, and even in cases where link traced methods are also possible, this approach is also generally easier to collect information.[ ] Initial individuals (egos) are first selected by strict random sampling methods, then the information about their contacts (alters) is reported via interview. The "egos" will make a list ("alters") based on the questions, and in some designs the "egos" have to answer questions about the attribute of alters or the ties between alters. At this point, the "alters" are indirectly observed or recruited, and alters may not be identifiable. In addition, in some study design, the "alters" can be selected as new "egos" to be interviewed. Of course, this is only possible if the "alters" are identified in the study design.Variants arise from enumeration and description of respondents. This method is broadly used in social science [7].

While those designs greatly reduces the amount of work required for data collection, the statistical inferences we can obtain are also limited by the reduction

in information about network structure. Also, a variety of problems arise when estimating network models and inferring statistics from sampled date . A method to inference and estimation of a subset of Exponential-family Random Graph Models from egocentrically samples network data has been developed by Krivitsky and Morris [3].

Each individual in a social system has a different levels of predisposition, and individual predisposition is very important in forming networks [12]. For the ERGM model, the term "actor-attribute effect" indicates the tendency of actor attributes to influence relationship network formation. In the case of actors with the same attribute, homophily (nodematch) is observed in social networks because people tend to associate with people who are similar to themselves. In a previous paper, attribute-based degree homophily was shown to be estimable with egocentric sampling design [8].

In network relationships, the emergence of some relationships facilitates the formation of other relationships, and we call them network self-organization [13]. In networks invoking collaboration, network closure is a self-organizing effect that often arise naturally, and it will form triangulation. Triangulation reflects the human society's tendency for group collaboration, in which case more triangulations are observed more frequently. Krivitsky and Morris presented how one can infer triadic effects from egocentrically sampled network designs that include collection of the alter-alter matrix. However, the elicited alter-alter ties used to develop models are indirection ties, so direction-based configurations such as reciprocity , activity (out-degree) , and popularity (in-degree) has not been studied. This paper will show that if the directed relationship between alters are included in data, then reciprocity can also be estimated.

In this paper, a framework for estimation and inference of reciprocity from egocentrically sampled data with directed relations will be presented. As an access point for studying social networks. Egocentric data includes a a lot of information that has not yet been explored. With the right approach, many statistical inferences

about the complete network can be explored from it. The model is from the observed data (egocentric) and the assumption (the statistic terms) to define a class of models (the coefficient values). This framework based on a rigorous approach presented by Krivitsky and Morris[4][8]. We will review the basic theory about ERGMs and give some parameterization details in Section 2. In Section 3, we will describe the specific data used and provide a rigorous analysis of the data. In this section, everything will be put together: starting from the egocentric data, to estimating the model, testing the significance of the parameters, testing the goodness of fit of the model, and simulating the completed network from half the size of the data. The statistics of the simulated model will match the statistics of the full model with appropriately scaled. Finally, we finish the paper with a conclusion and a discussion of the limitations of this framework.

Chapter 2

The Basic Theory of Exponential-family random graph models

This section will introduce the background and the technical detail on ERGM estimation and inference method from egocentrically sampled data. We will introduce ERGM first, and on the top of that we will explain how egocentric data is implemented.

2.1 ERGMs

2.1.1 General Framework of ERGMs

The development of the ERGM family can be traced back to 1959. Erdos and Renyi proposed a strict method in 1959 that uses statistical models to obtain the characteristics of simple random graphs [9] [10] [11]. ERGM has long been one of the most important network models to analyse the structure and relationship in social networks over time. ERGM is a general class of models based in exponential-family theory specified by sufficient statistics mapped on the sample space. It expresses the probability of an observed graph y as below.

$$\Pr_{\mathrm{g}}(Y = y; x, \theta) \equiv \exp\left\{\theta^{\top} g(y, x)\right\} / \kappa_g(\theta, x), \quad y \in \mathcal{Y} \tag{2.1}$$

$\mathcal{Y}$ is the sample space of possible networks, and y is realization of the random variable for the specific state of the network. For example, Let y be a network with N nodes, which can be defined as a $N \times N$ adjacency matrix. The possible

network edges is Y_{ij}, where $i =, 1, 2, \ldots, N$ and $j =, 1, 2, \ldots, N$, which represents the connection between actor i and actor j. If $Y_{ij} = 1$, the relationship of two nodes is present, vice verse. There are two kinds of graphs: undirected and directed. In directed network, Y_{ij} is not equal to Y_{ji}. For example, in our study, although the actor A helped the Actor B, it does not mean that the actor B would also help the actor A back when A is needed. In undirected network, $Y_{ij} = Y_{ji}$

If we represent the numerator in log form, it becomes a more clearly understood linear pattern:

$$\log\left(\exp\left(\theta' g(y)\right)\right) = \theta_1 g_1(y) + \theta_2 g_2(y) + \ldots + \theta_p g_p(y) \tag{2.2}$$

Where p is the number of terms/ covariates in the model. θ is parameter vector associated with $g(y, x)$, which represents the size and direction of the effects of the covariates $g(y, x)$. $g(y, x)$ is a sufficient statistic vector of network statistics that related to the possible covariates (configurations) of the whole network. It represents network features like degrees, homophily, density and reciprocal. The configurations are features of interest that are considered to be representative of network relationships, where the ties within those configurations are assumed to be conditionally dependent/independent.

A distinction in covariates is dyad independence or dyad dependence. Dyad independent terms are generally used to indicate dependencies between nodal attributes rather than edges dependencies between nodes, such as node-match. In contrast, dyad dependent terms would be used to show the dependence of edges between nodes, for example the tie between i and k would depend on the existence of ties between i and j, j and k. This is called triad terms [15]. We will further explain their formula in the Section 2.3.1. $\kappa(\theta) = \sum_{y \in \mathcal{Y}} \exp\left(\theta_1 g_1(y) + \theta_1 g_2(y) + \cdots + \theta_p g_p(y)\right)$ is a normalizing quantity based on the graph space of networks that ensures the sum of probabilities equals to 1.

A fundamental concept that supports the ERGM model is the dependency between relationships. A certain pattern of relationships can not be formed without

the corresponding dependency. This pattern is called configuration. Thus, to investigate configurations by $g(y, x)$ is to investigate dependencies between network relationships. If all relationships in our model are not dependent on each other (i.e., $g(y, x) = 0$), then there is no opportunity to form a specific configurations because all graphs have the same probability of occurring. Once we have adopted some kind of dependency theory, that is, determined $g(y, x)$, then θ is actually our inference of the power of each configurations and a simulation of the principle of development of the complete network. Theoretically, the likelihood of the global network can be inferred by the sampled local network. Therefore,θ is our target of inference.

2.1.2 Conditional Probability and change statistics

The conditional probability on the binary model represents the probability that a tie (i, j) will occur or not occur (equals to 1 or 0), when the rest of network (Y_{-ij}) being fixed [13], defined as

$$\Pr\left(Y_{ij} = 1 \mid Y_{-ij} = y_{-ij}, \theta\right) = \frac{\exp\left(\delta_{ij} g\left(y_{ij}\right)\right)}{1 + \exp\left(\delta_{ij} g\left(y_{ij}\right)\right)}$$

The change statistics δ_{ij} is defined as the change in the statistics of this network if Y_{ij} is present and when Y_{ij} is not. In other words, y_{ij}^{+} represents the network with $y_{ij} = 1$, y_{ij}^{-} represents the network with $y_{ij} = 0$.

$$\delta_{ij}\left(y_{ij}\right) = g\left(y_{ij}^{+}\right) - g\left(y_{ij}^{-}\right)$$

For instance, in the simple model, the appearance of Y_{ij} simply means that the number of edges increases by one. But in a more complex model, it is possible that the appearance of Y_{ij} will cause a change in the number of homophily or a difference in in-degree / out-degree.

2.2 Notation for Egocentric Data

To better establish egocentric estimation with direction network, we first review some necessary notation for egocentrically sampled data.

2.2.1 Population network

For a given graph, Let N = $\{1, \ldots, |\mathrm{N}|\}$ be a large but finite set of actors in the population of interest, and $i \in N$ means that " i belongs to the set N." For each actor, let x_i be a vector of attributes, such as age, gender, religion and so on. X_N or just N being a design matrix with all attribute for all actors. Let $\mathbb{Y}(\mathbb{N})$ be the set of all possible relational ties for the node set N, and the number of this set is $\binom{n}{2} = \frac{n(n-1)}{2}$ for undirected network, and $n(n-1)$ for directed network. The difference in numbers is due to the fact that the undirected network there is no need to distinguish between (i, j) and (j, i). Let $\mathbb{Y}(N) \equiv \{\{i, j\} : (i, j) \in N \times N \wedge i \neq j\}$ be the set of potential connection (dyads) in an network of these actors. This set excludes pairs (i, i), because self-ties are meaningless. For any network, some edges in $\mathbb{Y}(\mathbb{N})$ may be present and absent. Then , let $\mathcal{Y}(N, x) \subseteq 2^{\mathbb{Y}(N)}$ be the set of realization networks of studying. $\mathbb{Y}(\cdot, \cdot)$ may incorporate exogenous constraints. In out case, the ties of egos are limited to 6. We will discuss in section 3.1. For a network $y \in \mathcal{Y}(N, x)$, let Y_{ij} be an indicator function of whether a tie between i and j is present and $y_i = \{j \in N : y_{i,j} = 1\}$, the set of i's network neighbors, with $|y_i|$ being their number.

2.2.2 Egocentric data and sampling

Let e_i be the " egocentric " view of network y from ego actor i. It consists of two parts, and we distinguish them each with the upper right corner marker e and a. The first part is $e_i^{\mathrm{e}} \equiv x_i, i \in N$, ,which records all the attributes reported by ego about them self and is a vector. The other part is $e_i^{\mathrm{a}} \equiv (x_j)_{j \in y}, j' \in \{1, \ldots, |y_i|\}$, which consists of a list of vectors, each vector j represents the attribute of j^{th} person who is nominated by ego i. In many cases, $e_i^{\mathrm{a}} \equiv (x_j)_{j \in y}$ will have many zeros, because ego usually does not know the alter as well as himself / herself, or is in the principle of privacy. Then , j^{th} element of $e_i^{\mathrm{a}}, e_{i,i'}^{\mathrm{a}} \equiv x_{i:j'}$. Also , the k^{th} attribute observed on them are $e_{i,k}^{\mathrm{e}} \equiv x_{i,k}$, $e_{i,j',k}^{\mathrm{a}} \equiv x_{i:j',k}$. Then e_N represents the

egocentric census. The information about y contained in an egocentric sample of actors $S \subseteq N$ can then be represented as $e_S \equiv (e_i)_{i \in S}$.

When we sample S, we can take the conventional sampling method without specific weight for each individual, or we can choose the designed method to sampling with weight. In out study, each individual in each stratum is given a different probability according to the number of housing units in each district to ensure the geographic dispersion between areas. In addition to stratification weight each ABS case in the project also has post stratification weight.

Most of the above notation we have followed the Krivitsky and Morris (2017) framework [8]. We describe a specific case for direct network here.

Egocentric data contains not only undirected connections, but also in some studies the tie between actor is directed. For example, in out study, ego did not only answer who would provide help to him/her when he/she was injured, but also answered which of his/her friends would provide help to him/her when he/she was injured. We define such data $e^{\text{ea}}_{i,j} = Y_{ij}$ and $e^{\text{ea}}_{j,i} = Y_{ji}$.

2.3 Egocentric ERGMs

If the sufficient statistics and sample space of an ERGM can be obtained from the egocentric census, then we call this ERGM of the form 2.1 egocentric ERGM. Next we will give more detail of the sufficient statistics and the sample space in order.

2.3.1 Egocentric statistics

A network sufficient statistic is called egocentric if it can be expressed as

$$g_k(y, x) \equiv \sum_{i \in N} h_k(e_i) \tag{2.3}$$

for some function $h_k(e_i)$ of egocentric information. In addition, egocentric statistics are usually divided into two categories. The first is dyadic-independent, these statistics are generally symmetric function in form of $g_k(y, x) = \sum_{(i,j) \in y} f_k(x_i, x_j)$. The statistic term is called dyadic-independent if the connection state of $y_{i,j}$ affects

only the change statistic of (i, j) [17]. Attributes for two actors inform a range of dyadic-independent terms. Another is dyadic dependent statistics that normally be expressed as $g_k(y, x) = \sum_{i \in N} f_k \{x_i, (x_j)_{j \in y_i}\}$ for some function of the attributes of an actor and their network neighbors. Some example are given in Table 2.3.1.

Examples of egocentric statistics for undirected networks. $x_{i,k}$ may be a dummy variable indicating i's membership in a particular exogenously defined group. $h_k(\boldsymbol{e}_i)$ that sum over ties are halved because each tie is observed egocentrically twice: once at each end

Statistic	$\boldsymbol{g_k(y, x)}$	$\boldsymbol{h_k(e_i)}$
General sum over ties	$\sum_{(i,j)\in \boldsymbol{y}} f_k(\boldsymbol{x}_i, \boldsymbol{x}_j)$	$\frac{1}{2}\sum_{z\in \boldsymbol{e}_i^{\mathrm{a}}} f_k(\boldsymbol{e}_i^{\mathrm{e}}, z)$
Number of ties in the network	$\lvert\boldsymbol{y}\rvert \equiv \sum_{(i,j)\in \boldsymbol{y}} 1$	$\frac{1}{2}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert$
weighted by actor covariate $x_{i,k}$	$\sum_{(i,j)\in \boldsymbol{y}} (x_{i,k} + x_{j,k})$	$\frac{1}{2}(e_{i,k}^{\mathrm{e}}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert + \sum_{z\in \boldsymbol{e}_{i,k}^{\mathrm{a}}} z)$
weighted by difference in $x_{i,k}$	$\sum_{(i,j)\in \boldsymbol{y}} \lvert x_{i,k} - x_{j,k}\rvert$	$\frac{1}{2}\sum_{z\in \boldsymbol{e}_{i,k}^{\mathrm{a}}} \lvert e_{i,k}^{\mathrm{e}} - z\rvert$
within groups identified by $x_{i,k}$	$\sum_{(i,j)\in \boldsymbol{y}} 1_{x_{i,k}=x_{j,k}}$	$\frac{1}{2}\sum_{z\in \boldsymbol{e}_{i,k}^{\mathrm{a}}} 1_{e_{i,k}^{\mathrm{e}}=z}$
General sum over actors	$\sum_{i\in N} f_k\{\boldsymbol{x}_i, (\boldsymbol{x}_j)_{j\in \boldsymbol{y}_i}\}$	$f_k(\boldsymbol{e}_i^{\mathrm{e}}, \boldsymbol{e}_i^{\mathrm{a}})$
Number of actors with d neighbors	$\sum_{i\in N} 1_{\lvert\boldsymbol{y}_i\rvert=d}$	$1_{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert=d}$
weighted by actor covariate $x_{i,k}$	$\sum_{i\in N} x_{i,k} 1_{\lvert\boldsymbol{y}_i\rvert=d}$	$x_{i,k} 1_{\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert=d}$

Table 2.3.1 Undirected-network statistics

The statistics that we can obtain generally depend on the way the egocentric study design is done. In out paper, we use basic egocentric design with ego reports of alter degree. The table below illustrate the difference between undirected and directed network under same design.

Basic (minimal) egocentric design with ego reports of alter degree	
Undirected	**Directed**
Basic (minimal) egocentric design with ego reports of alter degree	
Nodal Covariate/Factor effects	Nodal Covariate/Factor effects (in and out)
Homophily	Homophily
Degree distribution	Degree distribution (in and out)
	Reciprocal
With alter-alter ties	
Triangles	Triangles
	Transitive triple
	Cyclic triple

Table 2.3.2

So not all statistic terms can be measured under this sample design, such as transitive. Also, in previous studies most of the egocentric networks are undirected networks, while the object of our study is a directed network, so many statistics will be different. The statistics are given following:

Examples of egocentric statistics for directed networks. $x_{i,k}$ may be a dummy variable indicating i's membership in a defined group.

Statistic	$g_k(y)$	$h_k(e_i)$
General sum over ties	$g_k(y,x)=\sum_{i\in N} f_k\{x_i,(x_j)_{j\in y_i}\}$	$\sum_{z\in e_i^a} f_k(e_i^e,z)+\sum_{s\in e_i^e} f_k(e_i^a,s)$
Number of ties in the network	$\lvert\boldsymbol{y}\rvert\equiv\sum_{(i,j)\in\boldsymbol{y}}1$	$\lvert e_i^a\rvert+\lvert e_i^e\rvert$
weighted by actor covariate $x_{i,k}$	$\sum_{(i,j)\in\boldsymbol{y}}(x_{i,k}+x_{j,k})$	$\left(e_{i,k}^{\mathrm{e}}\lvert\boldsymbol{e}_i^{\mathrm{a}}\rvert+\sum_{z\in e_{i,k}^{\mathrm{a}}} z\right)+\left(e_{i,k}^{\mathrm{a}}\lvert\boldsymbol{e}_i^{\mathrm{e}}\rvert+\sum_{s\in e_{i,k}^{\mathrm{e}}} s\right)$
weighted by difference in $x_{i,k}$	$\sum_{(i,j)\in\boldsymbol{y}}\lvert x_{i,k}-x_{j,k}\rvert$	$\sum_{z\in e_{i,k}^{\mathrm{a}}}\lvert e_{i,k}^{\mathrm{e}}-z\rvert+\sum_{s\in e_{i,k}^{\mathrm{a}}}\lvert e_{i,k}^{\mathrm{e}}-z\rvert$
General sum over actors	$\sum_{i\in N} f_k\left\{\boldsymbol{x}_i,(\boldsymbol{x}_j)_{j\in\boldsymbol{y}_i}\right\}$	$f_k(e_i^e,e_i^a)+f_k(e_i^a,e_i^e)$
Number of actors with d neighbors	$\sum_{i\in N}1_{\lvert y_i\rvert=d}$	$1_{\lvert e_i^a\rvert=d}+1_{\lvert e_i^e\rvert=d}$

Table 2.3.3 Directed-network statistics

2.4 Theoretical basis for inference

The obstacles posed to the egocentric ERGM are many. First, we cannot directly observe the statistics of the complete network. Second, we do not know the size of the population network [8]. The population network size can be very large.

Because we cannot directly observe the statistics of the entire network but rather the sampled version. The first question is how to scale sample statistics up to the population level. For the dyadic independent statistic, our first reaction is to let them scale on a per capita basis [14]. This is because this density-style statistic does not change as the network size increases or decreases. For example, the mean degree is "degree per capita", which is $\frac{ties}{nodes}$. In directed network, if a network of 10 people has 20 ties, the mean degree is $\frac{ties}{nodes}=\frac{20}{2}=10$, so we would expect a network of 20 people to have $20\times 2=40$ ties. In undirected network, if a network of 10 people has 20 ties, the mean degree is $\frac{2\times ties}{nodes}=\frac{40}{10}=4$. Although it appears that only 20 ties are connected to the nodes, we know that these ties are sent both ways.

So we would expect a network of 20 people to have $20 \times (\frac{4}{2}) = 40$ ties. However, when we estimate the design-based estimator that relies on the per capita scaling assumption $g(y,x) \approx \tilde{g}(e_S) = \frac{|N|}{|S|} \sum_{i \in S} h_k(e_i)$, the second problem that arises is that we do not know the size of the network. For the network census, the network size is known and finite. By contrast, for a sampled network, we do not know the size of the completed network, or the network size may be too large. But if the statistics we observe in the egocentric network data vary with network size in a known regular way, then we can adjust for the regularity in our estimation, and we call the this parameter estimation "size invariant". In this paper, we will follow Krivitsky, Handcock, and Morris (2011) [**?**] ,who use offset to get a per capita size invariant parameters from degree-based terms. As they showed in paper if we just adjust estimation without this technique, the mean degree tends to infinity as N converges to infinity. We know that this is impossible to exist in reality. In such a case, we need to estimate the model that is invariant to network size.

2.4.1 Network size effects for dyadic-independent terms

We first consider the Bernoulli model. Any individual potential ties within dyads are independent. Without adding offset, the network density tends to infinity, however if α is mapped to $\alpha - \log N$, then as N increases, the mean degree converges to e^{α}. This mapping also reflects the fact that as long as it is costly to establish connections, the probability of individuals remaining linked decreases as the network increases. This is a network size adjustment made by KM in order to facilitate estimation While accompanying the increase in $|N|$. The offset keeps the distribution of degrees and the effect of co-variates consistent, regardless of changes in network size.

2.4.2 Network size effects for dyadic-dependent terms : reciprocal

But for dyadic-dependent terms,$-\log(|N|)|y|$ no longer works. Krivitsky and Kolaczyk (2015) [14] used the Bernouli model with reciprocity to show that if $-\log(|N|)|y|$ is used as offset, which means that alpha is mapped to $\alpha - \log N$,then the expected value of the number of reciprocated tie will be $\frac{e^{2\alpha+\beta}}{N}$. That is ,

if n tends to infinity the number will tend to 0. However, if we map $\beta \mapsto \beta + \log N$ and keep the mapping of alpha unchanged. In this case the expected value of the number of reciprocated ties will be $e^{2\alpha+\beta}$. Both parameters will have an impression on the behavior of the model, regardless of the size of the change in n.

Chapter 3

Modeling ERGMs for directed network

This study is based on the UCNets "Understanding how Personal Networks Change" study. Respondents were asked who you think would help you if you were injured or sick and who you would help if your friend was injured or sick. The survey contains not only information about social relationships but also some additional information about attribute, such as gender and age. We will use this data to answer the following question. (1) How strong is the reciprocal? (2) Is there a tendency for older people to be more likely to provide help to younger people in this network of help. (3) what are the patterns of help networks between genders. (4) what impact do these network features have on overall network connectivity ?

3.1 data

Our data for this study comes from a five-year study by the UC Berkeley Social Networks Study (UCNets), "Understanding how Personal Networks Change". To examine how networks and characteristics change over time, this study interviewed researchers three times over a five-year period respectively. It is worth noting that there is a break in the age of the respondents, the age of respondents was limited to 21-30 and 50-70.Thedistribution of age is given in Figure 3.1.1

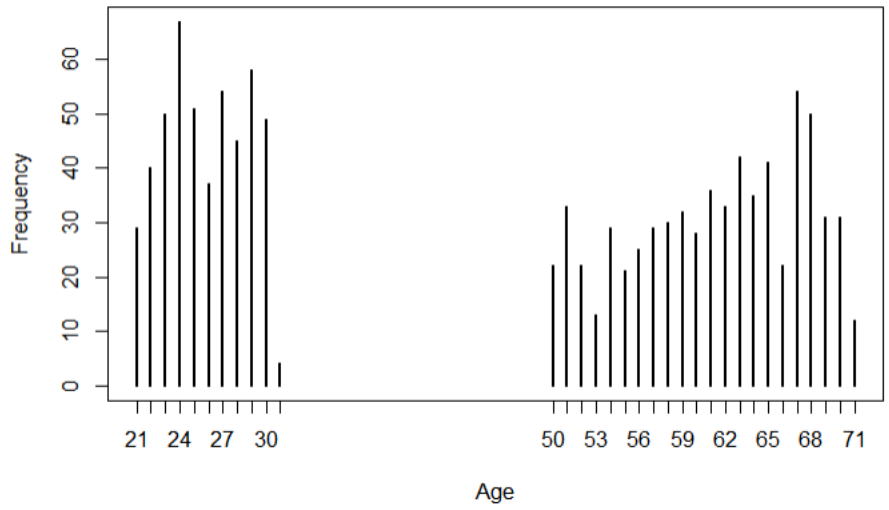

Figure 3.1.1 The distribution of age

Because researchers believe that people during those age range are more likely to experience different kinds of life events. Processing of missing actor data is a topic for future research. These individuals were sampled from the six-county San Francisco Bay Area. For each person asked, they were allowed to nominate up to six alters as answers. So the number of nominees does not obey normal distribution properly, as many egos who want to nominate more than six will be limited to six. A discussion on this point should also be put into future studies. These alters were not surveyed as ego's again. The questions were " If you were seriously injured or sick and needed some help for a couple of weeks with things such as preparing If you were seriously injured or sick and needed some help for a couple of weeks with things like preparing meals and getting around, who would you ask?" and "Who are the people that you help out practically, or with advice, or in other kinds of ways at least occasionally?" 1159 ego's participated in this survey, and 20,313 alters are nominated. Two weights were calculated for each individual to form a post-stratification weight, the first weight was related to the stratified clustered nature of households, and the second weight took into account the missing data in the survey.

In addition to the question of helping, two actor attribute will be added to our network study: age and gender. However, we found that many of the ego's answered the question about ties, but did not answer the question about the actor attribute of the alters.So we remove the alters without actor attribute and their corresponding egos.These missing actor attribute are mainly age attribute. The number of ego's after the deletion became 622 and the number of alters became 10417. After data cleaning, the distribution about gender is as table 3.1.2. We can see that there are still some alters whose gender attribute is 'NULL', this is because they are not linked to any other nodes, so we do not process them. And the age distribution after the cleaning has a very obvious change,as figure 3.1.3. An obvious phenomenon is that most of the ego's that do not give the attribute of alters are older generation. We will discuss how this phenomenon affects our model in the conclusion.

Table 3.1.2 Gender Frequency

	Male	Femalev	Null
Egos	208	414	—
Alters	2682	3730	4005

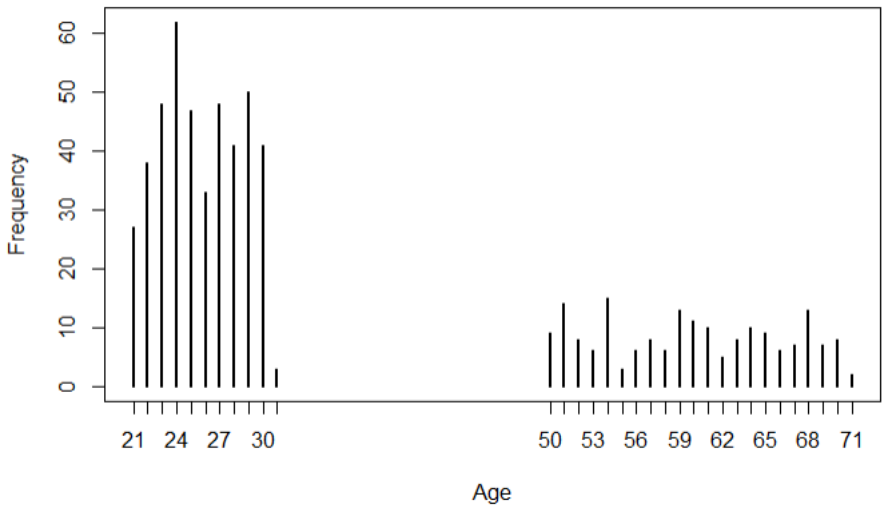

Figure 3.1.3

For age attribute, the problem we face is not only that the ego's age has gap, but we also don't know the exact age of the alters. The two questions that give

us information about the age of the alter are as follows: question 1– "Whether the age of the alter is within plus or minus 6 years of the age of the ego", and question 2– "Is the alter 6 years older than the ego and above ". By these two questions and using the logic as in table 3.1.4, we give alter three age labels "same" - the age difference between ego and alter is within 6 years, "older"—alter is older than ego by more than six years, and "younger"—alter is younger than ego by more than six years. The distribution of alters after label is as Figure 3.1.4. As we mentioned above for gender, if a node is not linked to other nodes, there nodal attribute may be 'Null'.

Logical judgment for age

1. If the answer to question 2 is yes, then we label the alter 'older' regardless of the answer to question 1.
2. If the answer to question 2 is no or null, and the answer to question 1 is yes, then we label the alter 'same'.
3. If the answer to question 2 is no, and the answer to question 1 is no, then we label the alter 'younger'

Table 3.1.4

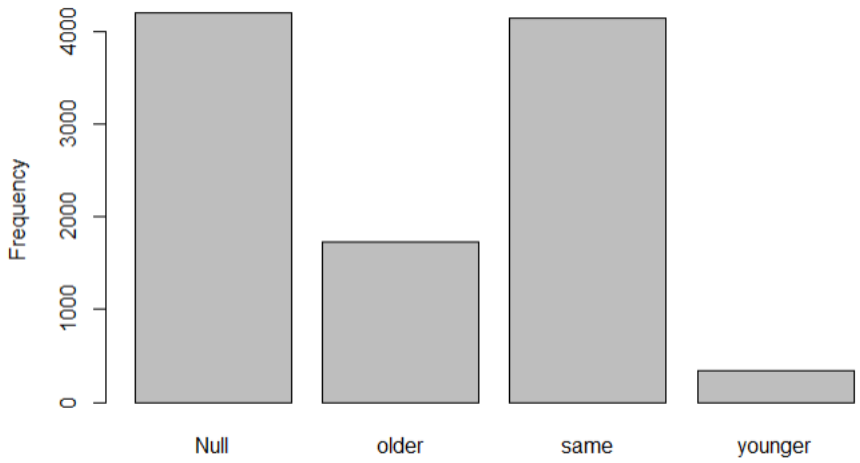

Figure 3.1.5 Age distribution of alters

3.2 Methods and models

Reciprocal is operationalized as the co-variate that equal to the number of dyads which ties in both direction exist. In the model we use mutual to represent this. To investigate the role of age in the network, for example whether older people are more likely to offer help to younger people, we cite network filters, which filters out ties (i, j) by given condition, and then observe the effect of age on the out-degree in-degree ties in these ties. There are three such filters in our model. We expect the values of out-degree and in-degree to be opposite, which suggests that younger/older people are more likely to help others/get help. The study of gender in helping networks falls on comparing male and female who are more likely to get help, and which gender is more likely to give help. We expect the same gender to have opposite performance in terms of out-degree and in-degree. A positive coefficient for one gender on the out-degree means that this gender is more likely to help others. We are fitting the following three model to data. And the network statistics are as table 3.2.1 below.

Model 1: This model serves as baseline, it has terms for the main effects of nodal attribute-gender. The number of edges that male appears as the node of origin/terminal of a directed tie are counted. In other words, this model uses females as the reference to test whether male have stronger propensity in helping network.

Model 2: This model adds terms for the main effects of nodal attribute-age. For data considerations (we do not know the exact ages of the alters), we used three separate filters to filter ties (i, j) for which i is more than six years older than j, i is more than six years younger than j, and i and j are within six years of each other. We tried to observe different results at different age configurations. And under each filter, we have two terms corresponding to the effects of age in the node of origin/terminal of a directed tie.

Model 3: This model tests the reciprocal ties in the network. We expect this terms to be dramatically positive. So this model adds 'mutual' on top of the terms mentioned above.

Statistics	legend
Edges	Total number of edges
Mutual	Total number of pairs of actors i and j for which (i,j) and (j,i) both exist.
Nodeifactor-Male	The number of times a node with attribute "Male" as the node of terminal.
Nodeofactor-Male	The number of times a node with attribute "Male" as the origin node.
F(edges, diff("Age")>6)	The number of edges (i,j) for which $Age(i) - Age(j) > 6$
F(edges, diff("Age")< −6)	The number of edges (i,j) for which $Age(i) - Age(j) < -6$
F(edges, absdiff("Age")≤6)	The number of edges (i,j) for which $\|Age(i) - Age(j)\| \leq 6$
F(nodeocov("Age") + nodeicov("Age"), diff("Age")>6)	Total value of attr(i)/attr(j) for edges (i, j) which $Age(i) - Age(j) > 6$, ego only
F(nodeocov("Age") + nodeicov("Age"), diff("Age")< −6)	Total value of attr(i)/attr(j) for edges (i, j) which $Age(i) - Age(j) > 6$, ego only
F(nodeocov("Age") + nodeicov("Age"), absdiff("Age")≤6)	Total value of attr(i)/attr(j) for edges (i, j) which $Age(i) - Age(j) > 6$, ego only

Table 3.2.1 Interpretation of network statistics

ERGM is built starting from out observation of local perspective (observations), and the the ERGM built from this observation will in turn explain how the network is generated globally. So we will next use methods that test whether the local model of the network "fits the data" , even though these properties are not terms in the model. But by testing whether a local fits will be a good observation of the global network. We evaluate the goodness of fit by the following criteria:

Reproducing observed degree distribution. This data is given by the gof function, which compares our observed degree distribution with the realizations. Where gof gives three ERGM terms as parameters, degree (both in and out for directed network) , esp (edge shared partner), and geodesic distance. However, due to data limitations and the direction of the research problem, we choose the degree distribution to test network connectivity. As an important determinant for studying network connectivity, we are looking forward to the results of its guide.

Robustness of conclusions to additional predictors. Even the complete model we built is not a real-scale network process. This is because it does not include many of the factors that are known to affect the help network. For example, the distance two people live, how long they have known each other, or even whether they attend the same school. In a way, of course, these factors can never be exhausted, and the question is whether the factors we study are robust to their inclusion. Take age for example, because it influences to some extent the time of acquaintance between the

nominator and the respondent. But if we add them into model 2 causing the weakening of the significant of gender term, which suggests that there are mechanisms of age as the basis for gender selection.

3.3 Results

Coefficients and standard errors for the three models. Coefficients reported are in the presence of an edge count offset and mutual offset

$$-log(11039) = -9.30919, log(11039) = 9.30919$$

Model	Terms	Estimatie	Std. Err
Model 1	edges	-2.95649***	0.02403
	nodeifactor.GENDER.male	-1.439329***	0.03426
	nodeofactor.GENDER.male	-1.412897***	0.03286
Model 2	nodeifactor.GENDER.male	-1.26149***	0.03486
	nodeofactor.GENDER.male	-1.15589***	0.03533
	F.diff.Age<6.edges	48.24168***	0.24896
	F.diff.Age<6.nodeocov.Age	-6.77560***	0.03794
	F.diff.Age<6.nodeicov.Age	6.75898***	0.03923
	F.diff.Age<-6.edges	20.65998***	0.96827
	F.diff.Age<-6.nodeocov.Age	2.96405***	0.1361
	F.diff.Age<-6.nodeicov.Age	-2.96401***	0.13601
	F.absdiff.Age<=6.edges	16.96493***	0.32151
	F.absdiff.Age<=6.nodeocov.Age	0.33876***	0.00828
	F.absdiff.Age<=6.nodeicov.Age	-1.19366***	0.01518
Model 3	mutual	2.134063***	0.027728
	nodeifactor.GENDER.male	-1.069271***	0.030475
	nodeofactor.GENDER.male	-1.021525***	0.063937
	F.diff.Age<6.edges	8.686847***	0.590623
	F.diff.Age<6.nodeocov.Age	-1.326887***	0.08095
	F.diff.Age<6.nodeicov.Age	1.322660 ***	0.081228
	F.diff.Age<-6.edges	31.903482***	3.197945
	F.diff.Age<-6.nodeocov.Age	4.705666***	0.453646
	F.diff.Age<-6.nodeicov.Age	-4.701149***	0.453943
	F.absdiff.Age<=6.edges	9.419477***	0.278886
	F.absdiff.Age<=6.nodeocov.Age	0.284571***	0.007019
	F.absdiff.Age<=6.nodeicov.Age	-0.808182***	0.012283

Figure 3.3.1

Model 1 is a little different from what we expected. Our previous guess was that a particular gender would have opposite coefficients on in-edges and out-edges (In our study this "particular gender" will be male, since we use female as the reference). However, the real result is that they both have negative coefficients for male. That is, male are less likely to either give help or receive help. First of

all, male may nominate fewer friends when answering this question, while female don't mind nominating more friends. The second possibility is that when male need help they may be more minded to get help from female, so they nominate people with same gender. Girls do not have such concerns, and they do not think too much about gender when nominating more people. Later in the paper I will explain what kind of follow-up studies we need to do to get a more accurate and reliable explanation. The equations are:

$$\log \frac{\Pr\left(Y_{ij}=1 \mid Y_{-ij}=y_{-ij}\right)}{\Pr\left(Y_{ij}=0 \mid Y_{-ij}=y_{-ij}\right)} = -2.9562 - 1.43911(\text{ nodeifacor.male }) - 1.41182(\text{nodeofacor.male})$$

The performance of nodeifactor and nodeifactor in model 2 does not change much by adding more terms, which proves their robustness to some extent. we then discuss the performance of the age factor in terms of receiving help and giving help under three type age differences, respectively. First is the case where i is more than six years older than j in directed tie (i, j). We can see from the results that the age of origin node has a negative effect on this type of connection, which means the probability of type connection decrease with the increasing age of origin node. Since only ego's age is recorded in our data, the older ego is the less likely to nominate himself/ herself as a helper, but will nominate himself/ herself as a helped, among all the ties that older helpers helping the younger. So the trend representing ego's understanding of the helping network is that younger people would help older people more. The second is the case in (i, j) where i is more than six years younger than j. It seems like that the age of origin node has a positive effect on this type of connection, which means the chance to form this type connection increase with the increasing age of origin node. So of all the younger helpers of the older tie, the older the ego is the more likely it is to nominate himself/ herself as the helper. It also means that more younger people are willing to help older people. The third is the case where the age difference between i and j in (i, j) is within six years. Under this condition, the older the person is, the more willing he or she is to be

a helper.The third is the case where the age difference between i and j in (i, j) is within six years. In this condition, the older the person is, the more willing he or she is to be a helper.

Finally, in Model 3, the significant positive coefficient of "mutual" term is also very consistent with our prediction. This is because if a person gets help from someone else, he/she is more likely to help them back. This also means that there is a deeper connection between each other.

We also can assess the goodness of fit by figure 3.3.1-3.3.6.Plots 3.3.2/3.3.4/3.3.6 show the out degree distribution, which is the distribution of nodes that tend to give help out to d person.3.3.1/3.3.3/3.3.5 show the in degree distribution, which is the distribution of nodes that receiving help from d person. Even the fit of the third model is not very good. There are three possible reasons for such results. First, our research can only nominate up to 6 alters per ego, while there is no restriction on this in our ERGM model. Second, since all alters will not be investigated again, their degree can only be 1. This is also not well simulated in the model. Finally, there is not enough terms in our model to allow model to simulate the exact degree distribution. We will discuss the possible causes of such results in the discussion and give directions for future research.

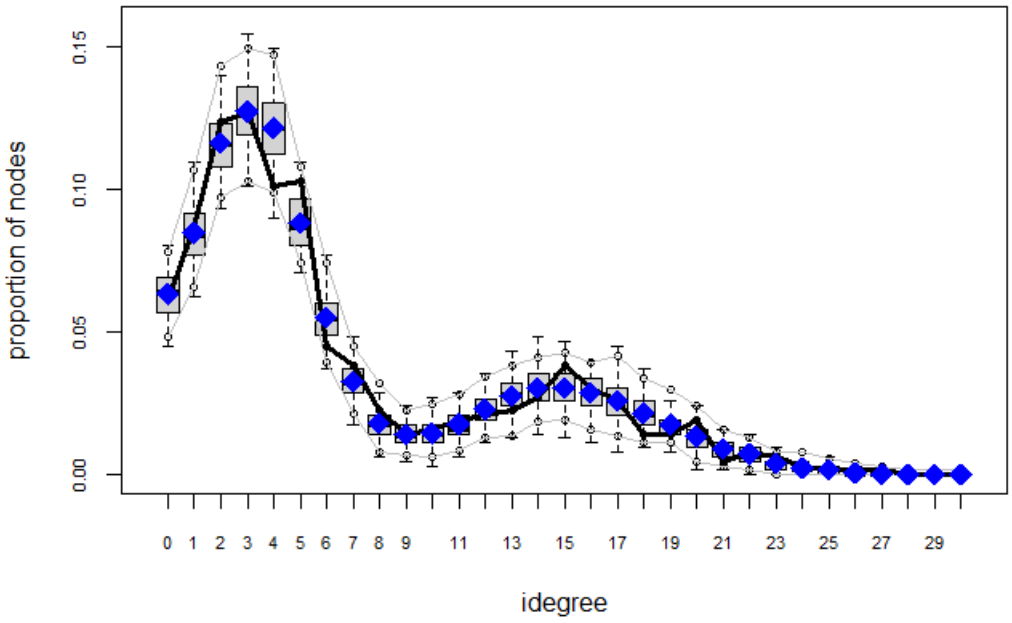

Figure 3.1: model 1 in-degree

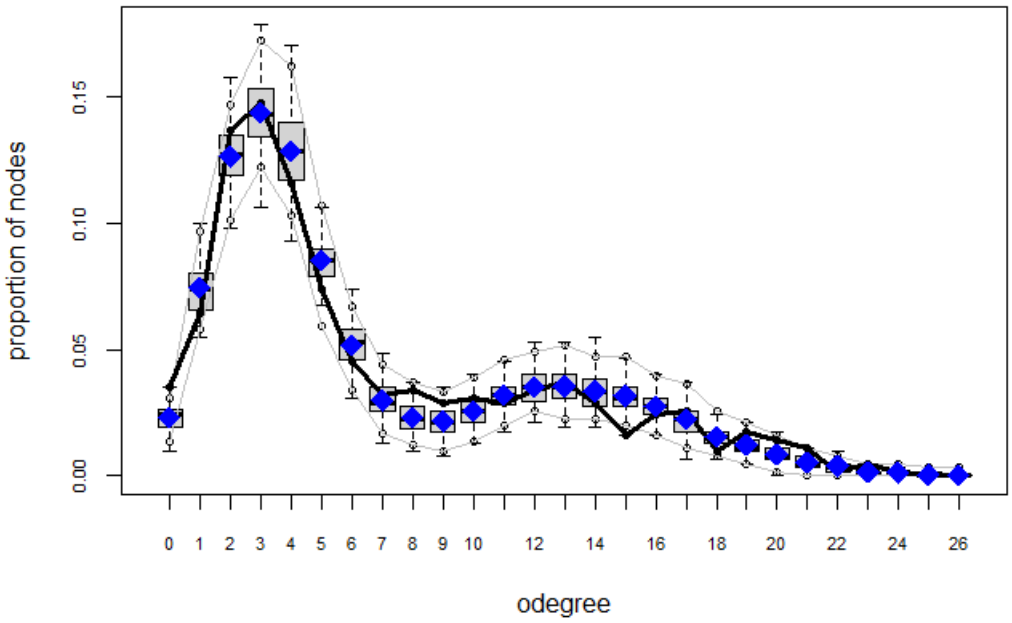

Figure 3.2: model 1 out-degree

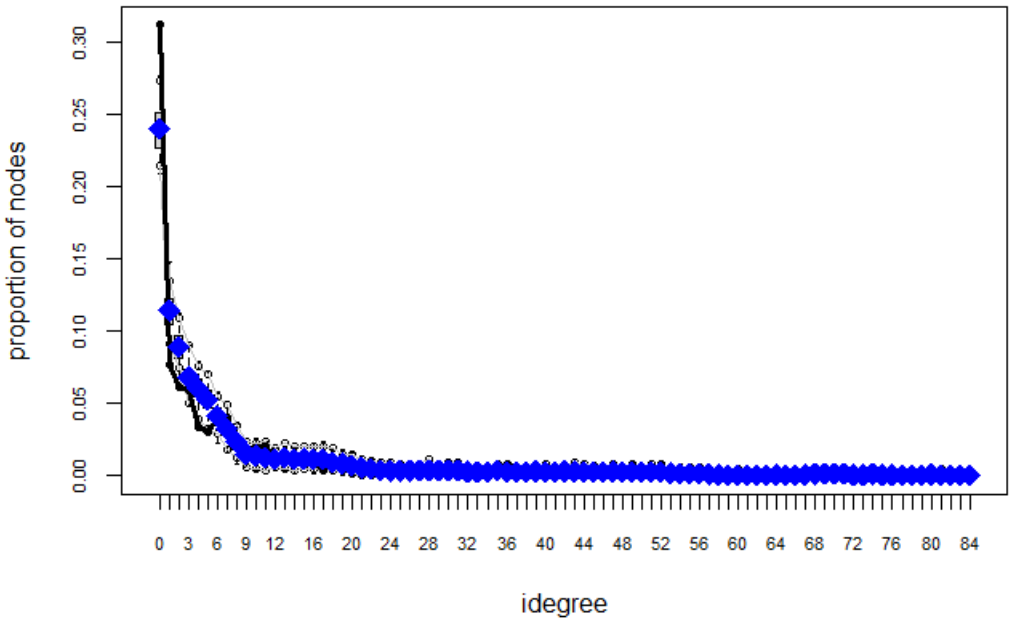

Figure 3.3: model 2 in-degree

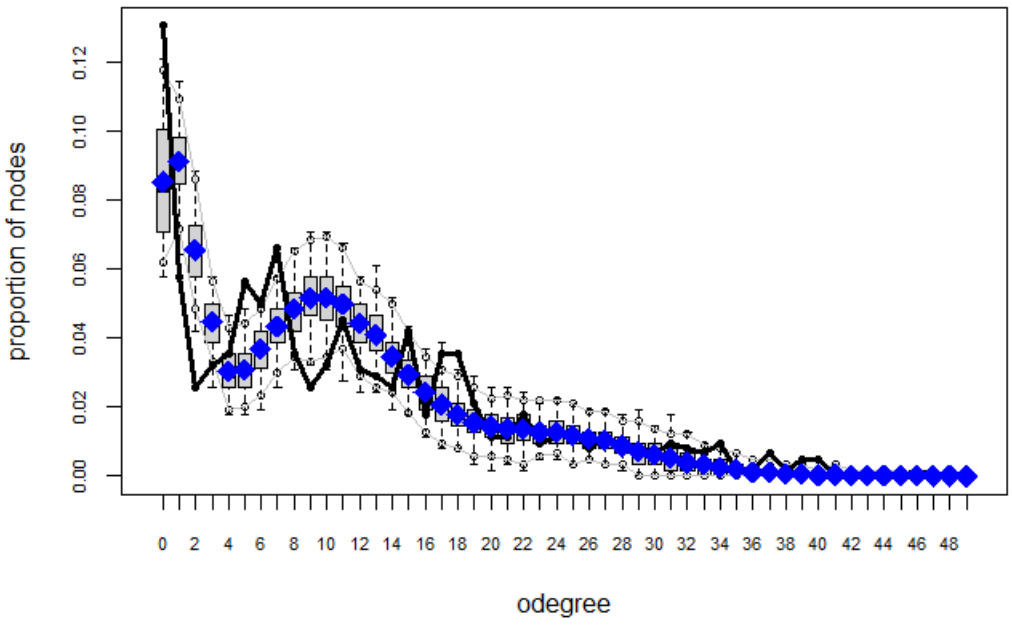

Figure 3.4: model 2 out-degree

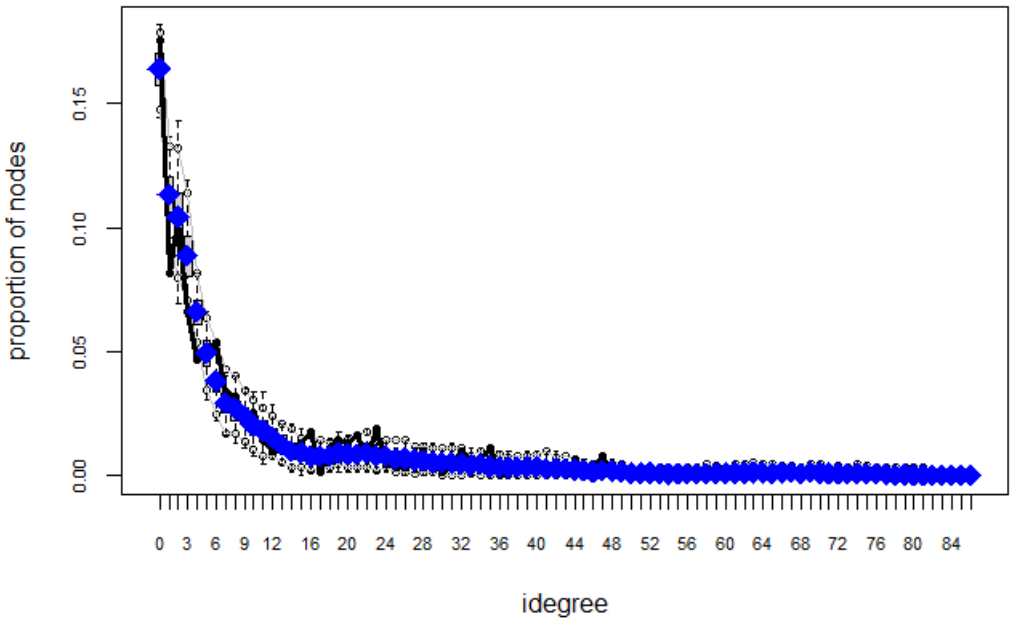

Figure 3.5: model 3 in-odegree

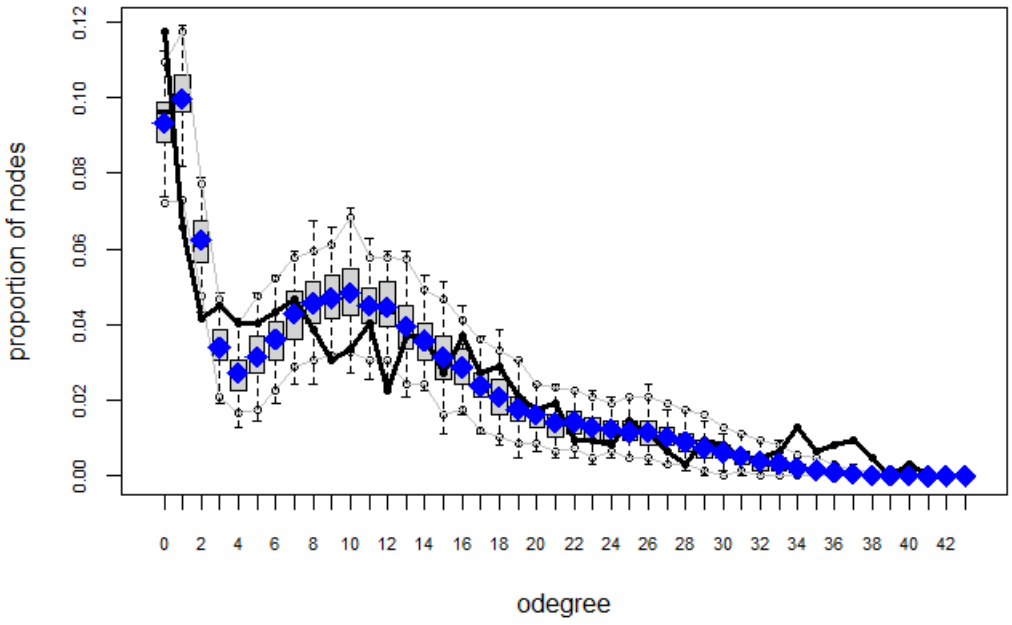

Figure 3.6: model 3 out-idegree

Chapter 4

Discussion

The study of Krivitsky and Morris [8] gets an extension towards directed networks and an application to filter in our paper. First, many statistics have changed when facing directed networks, as well as we can add reciprocal in the model now. The second point is how we can study them with filter when our data has significantly gap. Both of these points have a lot of practical applications in realistic model sampling. But on the other hand, our model has many directions that need further research. For example, when we look at the goodness of fit results, we also notice the inaccuracy of the degree distribution. The following discussion is about limitations and future directions.

4.1 model term

In model 1, I had provided realistic explanations for the observed fact that the coefficients of nodeifactor-male and nodeofactor-male are negative. For example, I mentioned that male are more likely than female to nominate the same gender. This can be verified in future studies through the observation of "nodematch". The second is about the degree distribution, without going into the discussion of the degree censoring, the performance of our model in goodness of fit also represents the possibility that more terms are needed to describe the degree distribution. Or on the other hand, maybe network has not been studied from enough aspects. We only considered the effect of gender and age, which needs more discussion in future studies.

4.2 degree censoring

In the data used in this paper, each respondent could only nominate up to six people. This is called fixed-choice design(FCD)[19], which is a phenomenon that is common in egocentric-data. However, there is no simulation of this point in our model. This affects our simulation of degree distribution in two ways. One, the goodness of fit of degree distribution assumes that the model has many nodes with $degree > 6$. Two, the case of degree = 6 would be more than expected because only six people can be nominated, so many respondents who want to nominate more than 6 also answer 6. The figure below shows a distribution of the number of nominations. We can notice that the bell curve in the first graph is broken at degree=6, because the ego that wants to nominate more people is all counted as 6. In the second graph, there is a significant increase in the number of degrees=6, compared to the previous more flat increase.

There are several possible solutions. The first is to directly impose a "in-degree" and "out-degree" restrictions to the model to test whether the model can simulate the network performance well under the constraint[18]. The second is to assign a value to the data based on a guess of the distribution of the degree. For example, most of the degree distributions obey normal distribution. However, whether the two methods can be used and which one is better need to be studied. Third method is called augmented fixed-choice design (AFCD) proposed by ott et al.(2017) [19]. This method re-samples m individuals ($m < |y_i|$) from the alters that have been nominated. e_i^s ,which is an ego data collected using AFCD, can be described by :

e_i^{as}: Attribute information collected on the m sampled alters.

$|y_i^s|$: The number of sampled alters.

Then, We can represent estimator for ego i as :

$$\widetilde{h}_k\left(\boldsymbol{e}_i^{\mathrm{s}}\right)=\frac{|\boldsymbol{y}_i|}{|\boldsymbol{y}_i^{\mathrm{s}}|}h_k\left(\boldsymbol{e}_i^{\mathrm{as}}\right).$$

For example, in a fixed choice design data, each ego can nominate up to 4 alters and we use m=3 to re-sample. And these three alters have actor attribute 1, 2, and

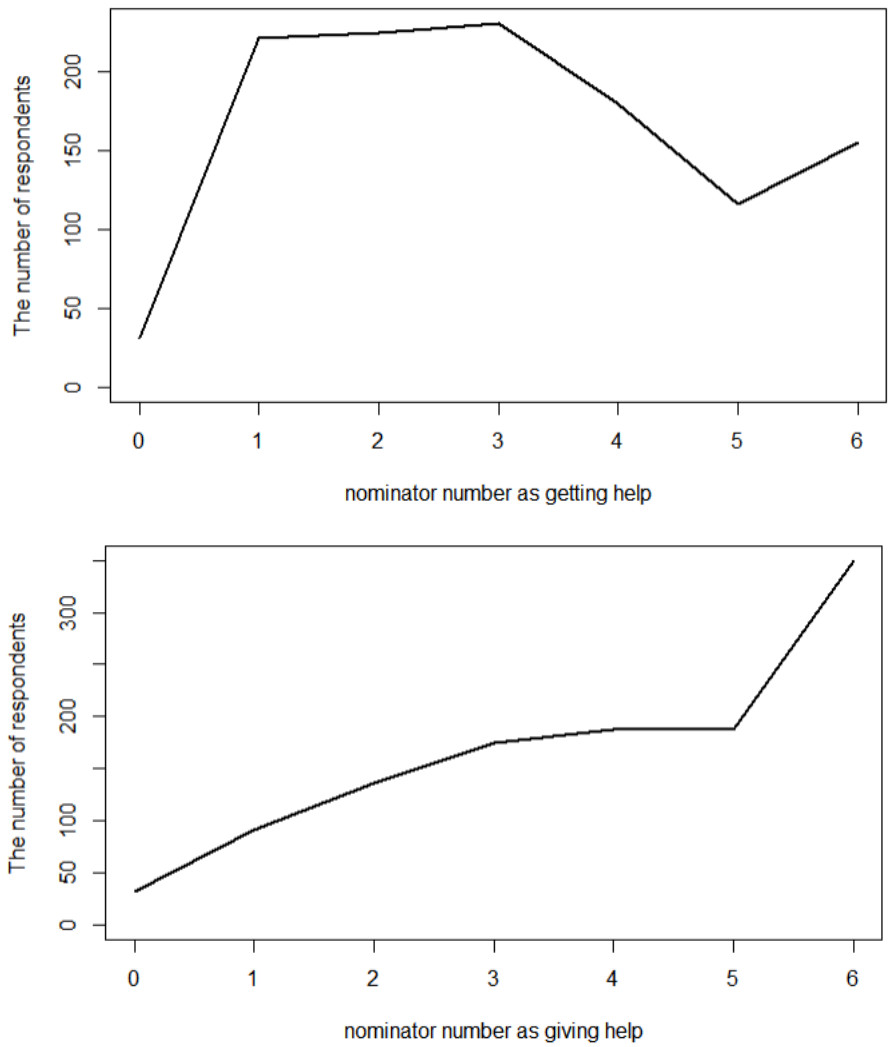

Figure 4.2.1 top: the frequency of in-degree;
bottom: the frequency of out-degree

3 respectively. Then, the AFCD estimate for ego's contribution would be,

$$\widetilde{h}_4\left(\boldsymbol{e}_i^{\mathrm{s}}\right)=\frac{4}{3}h_4\left(\boldsymbol{e}_i^{\mathrm{s}}\right)=\frac{4}{3}\cdot\frac{1}{3}(1+2+3)=\frac{8}{3}$$

Another limitation in this paper of ours is that each alter will not be recruited as a respondent again. This results in all alters having a degree of only = 0 or 1. This can also have a significant impact on the model.

4.3 Dyadic covariates

The data used this time reveals a serious problem, which is that sometimes we can observe dyadic covariate but we don't really know the actor attribute. For example, many of the attributes of the alters in the data we use are "Is the xxx attribute the same as ego?". Although we can get the same/different answer , we don't really know the actor attribute of the alter, which will limit our simulation of the network. There is even the possibility of reducing potential ties. If it is a binary attribute like gender, "is the same" will provide enough information. But if it is more selective, whether we can assign values to alters based on ego's attribute needs more study.

4.4 data gap and data distribution

Although sometimes we know the exactly attribute of both ego and alter, in some cases, limited by survey factors, some attributes are not continuous. For example, we use the age in data this time, is not available from 37 to 57 years. In this study, we use constraint to model though. But maybe there are other better ways to improve the accuracy of the model need to be investigated.

The second point worth noting is that we have removed some of the data that did not provide enough alter's attribute. However, there is a noteworthy pattern in these removed data, which is that most of them are older generation. The age distribution of these deleted ego's is as Figure 4.1.2. The question of whether this phenomenon will have an impact on our study, and what kind of impact it will have all need further research.

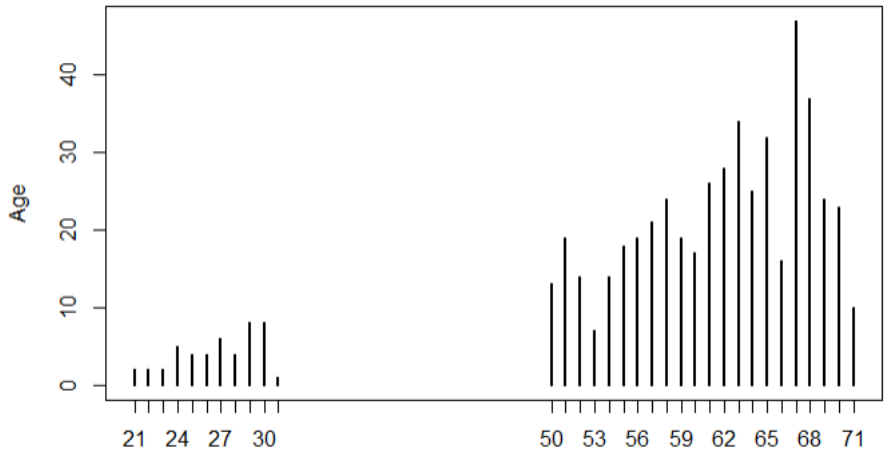

Figure 4.4.1

4.5 Repeated measure

There are some surveys that obtain several times at different times to explore the network variation over time, and estimation based on these longitudinal data is a research topic that can be carried out. Our data then consists of three waves, of which we studied only one this time.

References

[1] Harris, Jenine K. An introduction to exponential random graph modeling. Vol. 173. Sage Publications, 2013.

[2] Mark E. J. Newman. Networks: an introduction. Oxford University Press, 2010.

[3] Kolaczyk, Eric D., and Pavel N. Krivitsky. "On the question of effective sample size in network modeling: An asymptotic inquiry." Statistical science: a review journal of the Institute of Mathematical Statistics 30.2 (2015): 184.

[4] Krivitsky, P.N., Morris, M. and Bojanowski, M., 2019. Inference for Exponential-Family Random Graph Models from Egocentrically-Sampled Data with Alter–Alter Relations.

[5] Thompson, S.K., 1997. Adaptive sampling in behavioral surveys. NIDA Research Monograph, 167, pp.296-319.

[6] Handcock, M.S. and Gile, K.J., 2010. Modeling social networks from sampled data. The Annals of Applied Statistics, 4(1), p.5.

[7] Chung, K.K., Hossain, L. and Davis, J., 2005, November. Exploring sociocentric and egocentric approaches for social network analysis. In Proceedings of the 2nd international conference on knowledge management in Asia Pacific (pp. 1-8).

[8] Krivitsky, P.N. and Morris, M., 2017. Inference for social network models from egocentrically sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US. The annals of applied statistics, 11(1), p.427.

[9] Erdos, P. and Rényi, A., 1959. On random graphs. Publicationes Mathematicae, 6: 290–297

[10] Wasserman, S. and Pattison, P., 1996. Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika, 61(3), pp.401-425.

[11] Hunter, D.R. and Handcock, M.S., 2006. Inference in curved exponential family models for networks. Journal of Computational and Graphical Statistics, 15(3), pp.565-583.

[12] Parkhe, A., Wasserman, S. and Ralston, D.A., 2006. New frontiers in network theory development. Academy of management Review, 31(3), pp.560-568.

[13] Lusher, D., Koskinen, J. and Robins, G. eds., 2013. Exponential random graph models for social networks: Theory, methods, and applications (Vol. 35). Cambridge University Press.

[14] Kolaczyk, E.D. and Krivitsky, P.N., 2015. On the question of effective sample size in network modeling: An asymptotic inquiry. Statistical science: a review journal of the Institute of Mathematical Statistics, 30(2), p.184.

[15] Potter, G.E., Handcock, M.S., Longini Jr, I.M. and Halloran, M.E., 2012. Estimating within-school contact networks to understand influenza transmission. The annals of applied statistics, 6(1), p.1.

[16] HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008a). Goodness of fifit for social network models. J. Amer. Statist. Assoc. 103 248–258.

[17] Krivitsky, P.N., Handcock, M.S. and Morris, M., 2011. Adjusting for network size and composition effects in exponential-family random graph models. Statistical methodology, 8(4), pp.319-339.

[18] Hoff, P., Fosdick, B., Volfovsky, A. and Stovel, K., 2013. Likelihoods for fixed rank nomination networks. Network Science, 1(3), pp.253-277.

[19] Ott, M.Q., Harrison, M.T., Gile, K.J., Barnett, N.P. and Hogan, J.W., 2019. Fixed choice design and augmented fixed choice design for network data with missing observations. Biostatistics, 20(1), pp.97-110.