Title
An Empirical Network Formation Model with Incomplete Information

Author
Zhou, Wenyu

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

An Empirical Network Formation Model with Incomplete Information

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Wenyu Zhou

2019

ABSTRACT OF THE THESIS

An Empirical Network Formation Model with Incomplete Information

by

Wenyu Zhou
Master of Science in Statistics
University of California, Los Angeles, 2019
Professor Yingnian Wu, Chair

This thesis studies a network formation model with incomplete information, which introduces the neighborhood effect into the analysis of network formation. We show that the model is identified under some mild conditions. To overcome the computational burden, we propose to use the nested pseudo-likelihood algorithm to estimate the parameters of interest. Finite sample performance of the NPL estimation method is investigated through several Monte Carlo experiments. We find that a positive neighborhood effect makes agents more likely to form links, which can increase the network density. Besides, we also discuss three potential research directions.

The thesis of Wenyu Zhou is approved.

Arash A. Amini

Frederic Paik Schoenberg

Yingnian Wu, Committee Chair

University of California, Los Angeles

2019

To my parents . . .

for their love and support

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Prof. Yingnian Wu, who is my advisor of the statistics master program, for his precious guidance during my studies in the statistics master program at UCLA.

I would like to thank Prof. Arash. A. Amini and Prof. Mark Stephen Handcock for their help and excellent teaching.

I also want to thank all the friends I met when I took different statistics courses.

I have spent a wonderful time in the statistics master program at UCLA, which makes me believe that I will miss these days in the future.

CHAPTER 1

Introduction

Understanding network formation process is important for us to answer various real-world questions, such as "Which countries are more likely to trade with each other," and "Which type of friendship is more stable?" In addition, studying network formation process also has considerable policy relevance. Previous social science studies have shown that networks could play an important role in determining people's economic behavior including academic achievement, technology adoption, welfare participation, etc. Therefore, it is necessary for us to study network formation process in detail in order to make effective policies and facilitate policy implementation.

The main goal of studying network formation process is to develop theoretical models that can capture different statistical features existing in real networks, such as homophily, clustering, transitivity, sparsity, etc. Even though numerous efforts have been made by researchers in statistics, economics, sociology and many other disciplines, it is still difficult to achieve a good balance between tractability and comprehensiveness due to several reasons. First, it is challenging to obtain enough data for studying network formation. Under most cases, researchers only observe existing networks, which means modeling network formation process is similar to study a black box. Most currently available network data is static, which determines it is hard to infer factors that influence network formation from network dynamics. In addition, many usually unobserved variables, such as personalities, social status, and peer effects, may greatly affect network formation process. Thus, simply ignoring them can induce significant bias. Second, network formation process is extremely complicated. Besides a large number of variables can influence network formation, it is often the case that decisions of forming links are correlated with each other in most cases. If we regard each link as a random variable which takes value of 1 or 0, all these variables contained in a network could be dependent, which makes the problem much harder to deal with. Furthermore, researchers usually do not have prior information on the structure of interdependence within the network and need to infer it from the data. The third fundamental difficulty is that estimating a network formation model is usually computationally intensive and could even be intractable. This is mainly because the number of links grows at an exponential rate with the number of agents while modern network data sets could include hundreds of thousands of agents. At the same, many network formation models can admit multiple equilibria, which may further worsen the estimation.

In this paper, we develop a network formation model with incomplete information, which is able to capture several important features of some network formation processes in the real world. The first feature we aim to model is neighborhood effect that may influence network formation when two agents decide whether to form a link between each other. As I mentioned above, decisions of forming links are likely to be dependent while considering neighborhood effect may help to explain such interdependence within the network. When two agents decide whether to form a link between each other, besides considering dyadic variables, it is likely the decision can also be influenced by the decisions of their neighbors. There are several potential explanations for the existence of neighborhood effect, such as conformity, imitation and social norms, making people tend to conform to behaviors that are common among their neighbors when there is neighborhood effect. In the network formation setting, the neighborhood effect means agents may be more likely or less likely to form links conditional on their neighbor's behavior. For example, agents in a network would like to form links with a higher probability if they expect their neighbors also tend to link to each other. Adding neighborhood effect into the network formation process may help to model some realistic scenarios, such as those online social networks. The second feature that is captured by our model is the homophily effect, which has long been recognized by previous studies. The homophily effect implies that agents are more likely to form links with others who share similar individual characteristics. In our model, the difference between two agents is modeled by the distance between two vectors of individual characteristics. Third, our model provides an explicit definition of neighborhood. Confining the interdependence of link formation decisions within a certain neighborhood can greatly simplify our analysis and computation, which also facilitates asymptotic analysis. The main intuition is that decisions made by agents that are far away from each other can be thought as nearly independent. Even though we assume that the size of neighborhood is prior information in this paper, it may be directly estimated from network data, and we leave this extension as a future research direction.

In this paper, the network formation process is modeled by a network game of incomplete information. Two agents i and j will link to each other if they expect the overall utility gained from forming the link is positive, as we assume that the utility from forming the link is transferable. We assume that all links are simultaneously formed in a one-period network formation game. Agents i and j observe the neighborhood structure, which are determined by their location in the network Z and the neighborhood size d, as well as the individual characteristics X of other agents in whole network while the link-level idiosyncratic shock ϵ_{ij} is treated as private information. We use a logistic regression model to incorporate different variables that may influence the network formation process, including individual characteristics, neighborhood effects and idiosyncratic shocks. We show that under some mild assumptions, there exists a unique Bayesian Nash Equilibrium (BNE) for the network formation game and the model can be identified. Following the literature in game theory, we propose to use the nested pseudo likelihood (NPL) algorithm to estimate the model and finite performance of the estimation procedure is investigated through some Monte Carlo simulations.

The rest of this thesis is organized as follows. In Chapter 2, we conduct a brief literature review on network formation models. In Chapter 3, we develop the model, discuss the identification issue and outline the estimation strategy. In Chapter 4, we investigate the finite sample performance of the estimation strategy through Monte Carlo simulations. In Chapter 5, we discuss future research directions. Finally, Chapter 6 concludes.

CHAPTER 2

Literature Review

There is large growing literature related to network formation in statistics and economics. Network formation models studied in statistics literature are usually highly abstracted and simplified, such as the Erdős and Rényi (1959). model, stochastic block model and the exponential random graph models (ERGM) model, which enables researchers to apply tools from probability theory to analyze the network formation process. These statistical network formation models have been proved to work well in some specific scenarios but may not capture many important features of social networks. Instead, economists prefer make their models more close to the reality, especially when they study those social networks existing in the real world. The main difficulty of this approach, as I mentioned above, is that the network formation model could become extremely complicated and intractable due to the interdependence of links. To gain a better understanding of the topic, in this chapter, we conduct a brief literature review on papers that are most related to this thesis. The first well-known network formation model can be traced back to the seminal work of Erdös and Rényi (1959). In this influential model, links are treated as independent random variables which follows a binomial distribution. Let A_{ij} represent the link between agent i and j, $A_{ij} = 1$ if i links to j and $A_{ij} = 0$ otherwise. The Erdős and Rényi model thus assumes $\mathbb{P}(A_{ij} = 1) = p$ for all $i, j \in \{1, ..., n\}$, where p is a constant between 0 and 1. Since all links are independent, classical results, such as law of large numbers (LLN) and central limit theorem (CLT) can be easily applied to study properties of the model. One popular and important network formation model, which is a natural extension of the Erdős and Rényi model, is the so-called stochastic block model (SBM), that was first introduced by Holland et al. (1983) and Wang and Wong (1987). The model assumes that all n agents have their own community identities, i.e., there exists a disjoint partition $C_1, ..., C_K$ of the set of agents $\{1, ..., n\}$. The probability of observing a link between agents i and j is only determined by their communities identities. For example, if agent i belongs to community C_m and agent j belongs to community C_n, the SBM assumes $\mathbb{P}(A_{ij} = 1) = p_{mn}$. The edge probabilities between these K communities thus can be represented by a $K \times K$ symmetric probability matrix $\mathbf{P}$. Estimating the stochastic block model is of great research interest in statistics. Snijders and Nowicki (1997) studies the estimation problem of the stochastic block model by using Gibbs sampling in undirected networks, while Nowicki and Snijders (2001) further extends their results by allowing for directed networks and an arbitrary number of classes. Bickel and Chen (2009) develops a general theory for checking the consistency of community detection methods that replies on optimizing certain criteria in SBM. Amini et al. (2013) establishes the consistency of using pseudo-likelihood method to estimate SBM. In addition, there is a large literature on estimating the stochastic block model using spectral clustering. See Rohe et al. (2011), Fishkind et al. (2013), Lei et al. (2015) for more details.

Besides SBM, Holland and Leinhardt (1981) proposes the exponential random graph models, which is a family of probability distributions on graphs. The main idea of this model is that the probability of observing certain graph is determined by its nodal attributes. Hoff et al. (2002) proposes a class of random graph models where the probability of forming a link between agents i and j is determined by their observed characteristics and their latent positions in the "social space". Their approach of modeling social networks shares similar intuitions with subsequent economics research on this topic. Network formation process has also attracted much attention in economics research, especially in microeconomic theory and theoretical econometrics. Researchers are interested in developing network formation models that can explain certain economic behaviors and are consistent with network data. Leung (2015) develops a network formation model with incomplete information and proposes a two-step estimation method. However, the consistency of the estimation method replies on the assumption that the individual characteristics of agents must be discrete in order to construct a consistent first-stage estimator. This strong assumption may restrict the applicability of the model. Miyauchi (2016) develops a model which has non-negative externality and his approach differs from ours. Graham (2017) proposes a network formation model with unobserved agent-level fixed effects as follows:

$$D_{ij} = \mathbf{1}(W_{ij}^{'}\beta_0 + A_i + A_j - U_{ij} \geq 0)$$

where $D_{ij} = 1$ if there a link between agents i and j. A_i and A_j denote the unobserved agent-level fix effects, W_{ij} is a known function of the individual characteristics of agents i and j, and U_{ij} is link-level idiosyncratic shock. The main advantage of this model is that it can partially overcome the omitted-variable bias and works well in sparse social networks. But this network formation model doesn't consider neighborhood effect and all decisions of forming links are independent of each other. Sheng (2014) studies a network formation model with complete information, in which the network formation process is modeled by a simultaneous-move game. The model thus can admit multiple equilibria which leads to failure in terms of point identification of parameters. Instead, this paper adopts a partial identification approach and innovatively utilizes subnetworks to derive bounds. Boucher and Mourifié (2017) studies a class of exponential random graph models (ERGMs) in which they recover parameters of interest using a simple logit-based estimation method. The model studied in this paper is also very different from ours. Another type of network formation models focuses on dynamic settings, including Hsieh and Lee (2017), Badev (2017), Mele and Zhu (2017), etc. Most of these dynamic network formation models need to maximize likelihood functions, which are often computationally infeasible. One possible method to estimate the model is to adopt Bayesian approach such as MCMC, which is also fundamentally different from the method used here.

This thesis is also closely related to the literature on network games and general games with incomplete information. Xu (2018) considers estimation of social interaction effects in a large network game with incomplete information. Our thesis build upon the model studied in Xu (2018), as we introduce the framework into the analysis of network formation. Other related literature include Brock and Durlauf (2001), Liu and Zhou (2017), Lee et al. (2014), etc. These papers study the identification and estimation of binary choice model with social interaction effects.

CHAPTER 3

Methodology

3.1 Model

3.1.1 Setup

We consider a simultaneous network formation game with incomplete information. Researchers observe network data $G = (V, E)$, where $V = \{1, ..., n\}$ is the set of agents and E is the set of observed links. Besides using graph notation, the network data can also be represented by an $n \times n$ adjacency matrix A, in which $A_{ij} = 1$ if there is a link between agents i and j. In this paper, we focus on undirected network, which implies the adjacency matrix A is symmetric. Following the literature, we set $G_{ii} = 0$ for all $i \in V$.

Each agent i has public information $\mathcal{I}_i = (X_i', Z_i')'$, which is also observed by researchers and all agents in the network. Here X_i is a $p \times 1$ vector that denotes the individual characteristics, while Z_i is a $q \times 1$ vector that represents agent i's geographical information in the network. The neighborhood of agent i, which is denoted by $N(i)$ is defined as follows:

$$N(i) \equiv \{k \mid \ \|Z_k - Z_i\|_2 < d,\ k \in V,\ k \neq i\}$$

where $\|\cdot\|$ is the Euclidean norm. The definition of $N(i)$ implies that j is i's neighbor if and only if the network distance between these two agents is small enough. In the real world, Z_i could either directly denote the geographical location of agent i or some measure of i's location in the social network. A good example would be the social network in high school. Agent i's decisions of forming links are more likely to be influenced by his classmates in the same classroom only rather than students in other classrooms.

The decision of forming a link between agents i and j must be agreed by both sides together.

Specifically, we assume the utility gained from forming a link can be arbitrarily transferred between two sides. This assumption ensures that $A_{ij} = 1$ if and only if the total utility gained from forming the link between i and j is positive. The utility from forming the link (i, j) is given by

$$U_{ij}(1) = W_{ij}'\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l)\in N_{ij}} A_{kl} - \epsilon_{ij}. \tag{3.1}$$

Here W_{ij} is a $l \times 1$ vector such that

$$W_{ij} = f(X_i, X_j), \tag{3.2}$$

where $f(\cdot,\cdot)$ is some measurable function. Following the literature, $f(\cdot,\cdot)$ is symmetric in its two arguments, so we have $W_{ij} = W_{ji}$. The function $f(X_i, X_j)$ measures the degree of homophily between agents i and j. If these two agents have much similarity, which means the homophily between them is large, so the probability of forming a link between i and j is large, i.e., $f(X_i, X_j)$ should be small. One potential functional form of $f(\cdot,\cdot)$ is

$$f(X_i, X_j) = -\|X_i - X_j\|_2 \tag{3.3}$$

for all $i, j \in V$, where $\|\cdot\|_2$ is the Euclidean norm. In the equation (3.1.1), λ is a parameter that measures the strength of the neighborhood effect on agents i and j's joint decision. We let N_{ij} to be the set of neighboring decisions of forming links that will influence agents i and j, and it is defined by

$$N_{ij} = \{(r, s) \mid r < s, \quad r \neq i, \quad s \neq j, \quad r, s \in N(i) \cap N(j)\ \},$$

and $|N_{ij}|$ denote the number of elements in set N_{ij}. The definition implies only decisions made by pairs of agents that are in both the neighborhood of i and the neighborhood of j will influence the decision made by agents i and j. A_{kl} is the $(k,l)^{\text{th}}$ element in the adjancy matrix A. In this thesis, we assume that the idiosyncratic shock on agents i and j's joint decision, denoted as ϵ_{ij}, is independent and identically distributed across all links and has the following probability density function

$$f(\epsilon_{ij}|I) = \frac{e^{\epsilon_{ij}}}{(1 + e^{\epsilon_{ij}})^2}, \tag{3.4}$$

for all $i, j \in V$, which simply implies that conditional on the data I, the shock ϵ_{ij} has a logistic distribution. It is worth noting that it is possible to relax the assumption of having logistic distribution. Instead, ϵ_{ij} could also follow other common distributions such as Gaussian distribution. The main reason we adopt the assumption of having logistic distribution is that such assumption could significantly reduce the burden of mathematical derivation and enables us to get straightforward expressions.

We assume the utility of not forming a link equals 0, i.e.,

$$U_{ij}(0) = 0, \tag{3.5}$$

for all $i, j \in V$.

3.1.2 Information Structure

We assume that ϵ_{ij} belongs to private information, which implies its value is only known to agents i and j when they jointly decide whether to form a link A_{ij} between each other. Agents' individual characteristics X_i, geographical location in the social network Z_i as well as the size of their neighborhoods d is assumed to be public information. So, we can now define the information set for agents i and j that they have when making the decision of forming the link $A_i j$ for all agents $i, j \in V$

$$\mathcal{F}_{ij} = \{\ X,\ Z,\ d,\ \epsilon_{ij}\ \}, \tag{3.6}$$

and we can also define the information set for agent i for all $i \in V$

$$\mathcal{F}_i = \bigcup_{j}^{n} \mathcal{F}_{ij} \tag{3.7}$$

3.1.3 Equilibrium

We now define the equilibrium for this network formation game with incomplete information. Let θ denote the parameters of interest, i.e., $\theta = (\beta', \lambda)'$. According to the setup of the model, agents i and j will link to each other if

$$\mathbb{E}[\ U_{ij} \mid \mathcal{F}_{ij},\ \theta\ ] > 0, \tag{3.8}$$

which is equivalent to

$$W'_{ij}\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l)\in N_{ij}} \mathbb{P}(A^*_{kl} = 1 \mid \mathcal{F}_i,\ \mathcal{F}_j,\ \theta) - \epsilon_{ij} > 0 \tag{3.9}$$

Here A^*_{kl} denote agents k and l's equilibrium strategy regarding the formation of link (k, l) for all $k, l \in N_{ij}$. So, for all agents $i, j \in V$ in this network formation game, their pure equilibrium strategy is

$$A^*_{ij}(\epsilon_{ij} \mid \mathcal{F}_i,\ \mathcal{F}_j,\ \theta) = \begin{cases} 1 & \text{if } \epsilon_{ij}\ <\ W'_{ij}\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l)\in N_{ij}} \mathbb{P}(A^*_{kl} = 1 \mid \mathcal{F}_i,\ \mathcal{F}_j,\ \theta) \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

for all $i, j \in V$, which also defines a system of simultaneous equations, i.e., $(A^*_{12}, ..., A^*_{n,n-1})$. Because we assume that ϵ_{ij} follows a logistic distribution for all $i, j \in V$ and is independent across different links, so equation 3.10 defines a sequence of equilibrium choice probabilities. Let $\sigma^*_{ij}(\mathcal{I}, \theta) = \mathbb{P}(A^*_{ij}(\epsilon_{ij} \mid \mathcal{F}_i,\ \mathcal{F}_j,\ \theta) = 1 \mid \mathcal{I}\ )$ be the equilibrium probability of forming the link (i, j) for all agents $i, j \in V$, where $\mathcal{I} = \{\mathcal{I}_1, ..., \mathcal{I}_n\}$. In addition, let

$\Gamma^*(\mathcal{I},\theta) = \{\sigma^*_{12}, ..., \sigma^*_{n-1,n}\}$ be the set of equilibrium probabilities for all the links in this network. According to equation 3.10, we have

$$\sigma^*_{ij}(\mathcal{I},\theta) = \frac{\exp\left[W'_{ij}\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma^*_{kl}(\mathcal{I},\theta)\right]}{1+\exp\left[W'_{ij}\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma^*_{kl}(\mathcal{I},\theta)\right]} \tag{3.11}$$

for all $i, j \in V$. Equation 7 is derived from the logistic cumulative distribution function of the logistic distribution. It is worth noting that equation 7 also gives us a system of simultaneous equations, in which there are $n(n-1)/2$ unknown variables that we need to solve. In fact, solving $(\mathcal{I},\theta)\}$ from the system of equations is equivalent to solving the system defined by 3.10, as pointed out by Bajari et al. (2010) and Xu (2018). Based on the results shown above, we have the following proposition on the equilibrium solution to equation 3.10.

Proposition 1 *There exists at least one equilibrium probability profile $\Gamma^*(\mathcal{I},\theta)$ that solves the system of simultaneous equations defined by equation 3.10.*

Proof: Notice that $\sigma^*_{ij} \in [0,1]$ for all $i, j \in V$, the right hand of equation 3.10 is a function mapping from $[0,1]^{n(n-1)/2}$ to $[0,1]^{n(n-1)/2}$, which is continuous. Because $[0,1]^{n(n-1)/2}$ is nonempty, compact and convex, by Brouwer's fixed point theorem, there exists at least one equilibrium probability profile $\Gamma^*(\mathcal{I},\theta)$ that solves the system.

3.2 Identification

In this section, we discuss the identification issue of the model. First, we consider the uniqueness of the equilibrium in this network formation game. Proposition 1 ensures the existence of Bayesian Nash Equilibrium in this network formation game, but there may be multiple equilibrium, which would lead to the failure of identification of the parameters θ. To establish the uniqueness of the equilibrium, we impose the following additional assumption.

Assumption 1 *In the network formation model discussed in this thesis, $\lambda < 4$.*

The Assumption 1 basically assumes that the scale of the neighborhood effect should be reasonable. The intuition of this assumption is also straightforward. More specifically, as the neighborhood effect becomes larger, the dependence between agents i and j's decision on the link (i, j) and their neighbor's decisions will increase. In this case, there may exist multiple equilibria as they can all solve the system of simultaneous equations. Under this assumption, we have the following result.

Proposition 2 *If Assumption 1 is satisfied, there only exists a unique equilibrium probability profile* $\Gamma^*(\mathcal{I}, \theta)$ *that solves the system of simultaneous equations defined by equation 3.10.*

Proof: See the Appendix.

Proposition 2 is important for us to discuss the identification of the parameters θ. If there are more than one equilibrium, we must impose extra equilibrium selection mechanisms to determine the one that fits the data. Otherwise, the model will be incomplete and we can't consistently estimate the parameters of interest in this model. There are some theoretical literature dealing with the multiple equilibria problem, such as Sheng (2014), Brock and Durlauf (2001), Leung (2015), Tamer (2003) and Tamer (2010), and we leave this question as a future research direction.

The parameters that we aim to identify in this thesis is $\theta = (\beta', \lambda)' \in \mathbb{R}^{l+1}$. We first consider the following transformation of $\sigma^*_{ij}(\mathcal{I}, \theta)$. By equation 7, we have

$$\ln \sigma^*_{ij}(\mathcal{I}, \theta) = W'_{ij}\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} \sigma^*_{kl}(\mathcal{I}, \theta) - \exp \left[1 + \exp \left(W'_{ij}\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} \sigma^*_{kl}(\mathcal{I}, \theta) \right) \right]$$

After some simple algebra, we have

$$\ln \sigma^*_{ij}(\mathcal{I}, \theta) - \ln(1 - \sigma^*_{ij}(\mathcal{I}, \theta)) = W'_{ij}\beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} \sigma^*_{kl}(\mathcal{I}, \theta) \tag{3.12}$$

We will use equation 3.12 to identify the parameters of interest θ. Notice the LHS of the above equation is identified according to the definition of identification. Furthermore,

$\ln \sigma^*_{ij}(\mathcal{I}, \theta)$ is also identified, again by the definition of identification. So, the identification problem of the parameters in equation 3.12 in fact can be reduced to the similar problem in linear regression models, i.e., the matrix of independent variables should have full rank. Next, we propose the sufficient condition to establish the identification of the parameters. Let

$$S_{ij} = \left[W'_{ij}, \ \frac{1}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} \sigma^*_{kl}(\mathcal{I}, \theta) \right]'$$

Now we give the full rank condition as follows.

Assumption 2 *The matrix* $\mathbb{E}[S_i S'_i]$ *has full rank, i.e., Rank*$(\mathbb{E}[S_{ij} S'_{ij}]) = l + 1$.

The full rank assumption is very common in the literature. This is in fact a quite weak assumption in the network environment. For example, the full rank condition can be satisfied easily if the function $f(\cdot, \cdot)$ which maps X to W is nonlinear and there are enough variations regarding agents' neighborhood. Many empirical studies have shown that there exists significant heterogeneity in the neighborhood of agents in the network, which is likely to ensure the validity of the full rank condition. Based on the above assumption, we can now establish our main identification result.

Proposition 3 *Under Assumption 1 and Assumption 2, the parameters* $\theta = (\beta', \lambda)'$ *of the network formation discussed in this thesis is identified.*

Proof: By equation 3.12, we have

$$\mathbb{E}[S_{ij}[\ln \sigma^*_{ij}(\mathcal{I}, \theta) - \ln(1 - \sigma^*_{ij}(\mathcal{I}, \theta))]] = \mathbb{E}[S_{ij} S'_{ij} \theta]$$

Let

$$R_i = \ln \sigma^*_{ij}(\mathcal{I}, \theta) - \ln(1 - \sigma^*_{ij}(\mathcal{I}, \theta)).$$

Therefore, we have

$$\theta = (\mathbb{E}[S_{ij} S'_{ij}])^{-1} E[S_{ij} R_i]$$

Since $\mathbb{E}[S_i S_i']$ has full rank and the equilibrium is unique, θ is identified.

3.3 Estimation

In this section, we discuss the estimation of the network formation model. Notice that agents' equilibrium strategies A_{ij}^* are independent of each other conditional on $\mathcal{I}$ and $\mathbb{P}(A_{ij}^* = 1|\mathcal{I}) = \sigma_{ij}^*(\mathcal{I}, \theta)$ for all $i, j \in V$. Therefore, we can derive the conditional log-likelihood function from the perspective of researchers, i.e.,

$$\hat{L}(\theta) = \frac{2}{n(n-1)} \sum_{i<j, i \neq j} \left[A_{ij}^* \ln \sigma_{ij}^*(\mathcal{I}, \theta) + (1 - A_{ij}^*)(1 - \ln \sigma_{ij}^*(\mathcal{I}, \theta)) \right] \tag{3.13}$$

where $\{\sigma_{ij}^*(\mathcal{I}, \theta)\}_{i<j}$. satisfies the system of simultaneous equations defined by equation 7. And we have

$$\hat{\theta}_{\text{MLE}} = \underset{\theta \in \mathbb{R}^{l+1}}{\arg\max} \quad \hat{L}(\theta)$$

The MLE estimator is known to be consistent under some additional mild conditions. However, it's usually very difficult to calculate the MLE estimator because the computational burden. When there are n agents in the network, there will be $n(n-1)/2$ decisions needed to be estimated and there are no closed expressions for $\{\sigma_{ij}^*(\mathcal{I}, \theta)\}_{i<j}$, which make the optimization problem almost impossible to solve. To overcome this computational problem, we propose to use the nested pseudo likelihood algorithm (NPL) to calculate the estimated parameters. The nested pseudo likelihood estimator has been widely used in literature to estimate dynamic discrete games, such as Aguirregabiria and Mira (2007), Liu and Zhou (2017) and Lin and Xu (2017). We first give the some definitions. Denote $\sigma = \{\sigma_{12}, ..., \sigma_{n-1,n}\}$ to be some arbitrary probability profile. Let

$$S_{ij}(\sigma) = \left[W_{ij}', \ \frac{1}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} \sigma_{kl} \right]' \tag{3.14}$$

to be the link-level characteristics that influence the formation of the link (i, j). Also, let

$$\Gamma_{ij}(\sigma, \theta) = \frac{\exp\left[S_{ij}(\sigma)'\theta\right]}{1 + \exp\left[S_{ij}(\sigma)'\theta\right]} \tag{3.15}$$

and

$$\hat{L}_{\text{NPL}}(\theta, \sigma) = \frac{2}{n(n-1)} \sum_{i<j, i\neq j} \left[A^*_{ij} \ln \Gamma_{ij}(\sigma, \theta) + (1 - A^*_{ij})(1 - \ln \Gamma_{ij}(\sigma, \theta))\right]. \tag{3.16}$$

The nested pseudo likelihood algorithm (NPL) is given as follows.

Algorithm 1 Nested Pseudo Algorithm

1: $\sigma^{[0]} = [0.5, ..., 0.5] \in \mathbb{R}^{n(n-1)/2}$, $\hat{\theta}^{[0]}_{\text{NPL}} = 0$ and $k = 1$;
2: $\hat{\theta}^{[1]}_{\text{NPL}} = \arg\max_{\theta} \ \hat{L}_{\text{NPL}}(\theta, \sigma^{[0]})$;
3: **while** $\left\|\hat{\theta}^{[k]}_{\text{NPL}} - \hat{\theta}^{[k-1]}_{\text{NPL}}\right\| > 10^{-3}$ **do**
4: $\quad k = k + 1$;
5: $\quad \hat{\theta}^{[k]}_{\text{NPL}} = \arg\max_{\theta} \ \hat{L}_{\text{NPL}}(\theta, \sigma^{[k-1]})$;
6: $\quad \sigma^{[k]} = \Gamma(\sigma^{[k-1]}, \hat{\theta}^{[k]}_{\text{NPL}})$;
7: **end while**

Under some additional mild conditions, the nested pseudo likelihood estimator is consistent and asymptotically normal. See Aguirregabiria and Mira (2007), Kasahara and Shimotsu (2012) and Lin and Xu (2017) for more details.

CHAPTER 4

Monte Carlo Simulations

In this section, we conduct limited Monte Carlo simulations to test the effectiveness of the NPL estimation method proposed in Chapter 3 and the validity of the network formation model discussed in the thesis. We first discuss the data generating process for the individual characteristics $(X_i', Z_i')'$ and the adjacency matrix A of the network.

4.1 DGP for Individual Characteristics

To reduce the computational burden, we assume $\{X_i\}_{i=1,...,n}$ is a scalar which is drawn from an *i.i.d.* standard normal distribution for all $i = 1, ..., n$.

$$X_i \sim N(0,1)$$

For the geographical location of agent i in social space, we assume Z_i is a 2×1 vector and has a *i.i.d.* 2-dimensional uniform distribution with support $[0, 10]$, for all agent $i \in V$, i.e.,

$$(Z_{i1}, Z_{i2}) \sim U[0,10] \times U[0,10].$$

We use this setting as we aim to mimic the geographical distribution of agents in social networks in reality.

4.2 DGP for the Adjacency Matrix A

For different agents i and j in the network, their distance is determined by

$$d_{ij} = \|Z_i - Z_j\|_2$$

If $d_{ij} < \bar{d} = 2$, we assume agent j belongs to agent i's neighborhood, and vice versa. $\{N(i)\}_{i=1,\dots,n}$ and $\{N_{ij}\}_{i<j}$ are constructed according to the definitions in Chapter 3. The following figures describe the geographical location of three simulated networks.

For W_{ij}, we assume

$$W_{ij} = -|X_i - X_j|$$

for all agents $i, j \in V$. So, the utility of forming the link (i, j) for all agents $i, j \in V$ is given by

$$U_{ij}(1) = -|X_i - X_j| \cdot \beta + \frac{\lambda}{|N_{ij}|} \sum_{(k,l) \in N_{ij}} A_{kl} - \epsilon_{ij}.$$

In this thesis, I set $\beta = 1$ and consider three different cases regarding the value of λ: (1) $\lambda_1 = 0$; (2) $\lambda_2 = 0.5$; (3) $\lambda_3 = 2$. The three choices of λ correspond to three different network formation scenarios in the real world. The first case, in which $\lambda_1 = 0$, means there is no neighborhood effects. The second case, in which $\lambda_2 = 0.5$, represents the situation where there is a relatively weak neighborhood effect that influences the network formation process, while the third case with $\lambda_3 = 2$ corresponds to the situation where there is a strong neighborhood effect.

To generate the simulated adjacency matrix from the network formation model discussed in this thesis, I first solve for the equilibrium probabilities defined by 7, i.e., $\sigma^* = \{\sigma^*_{12}, ..., \sigma^*_{n-1,n}\}$. After getting the equilibrium probabilities, I draw ϵ_{ij} from a *i.i.d.* standard logistic distribution for all $i, j \in V$. Finally, the A_{ij} is determined by comparing $U_{ij}(1)$, which is calculated from equation 4.2, and zero. If $U_{ij}(1) > 0$, $A_{ij} = 1$, otherwise $A_{ij} = 0$, for all agents $i, j \in V$.

I consider different network sizes: (1) small-sized network with $n_1 = 50$; (2) medium-sized network with $n_2 = 100$; (3) large-sized network with $n_3 = 200$. For each size of network, I repeat each numerical experiment for 1000 times and collect the mean and standard error of the NPL estimators.

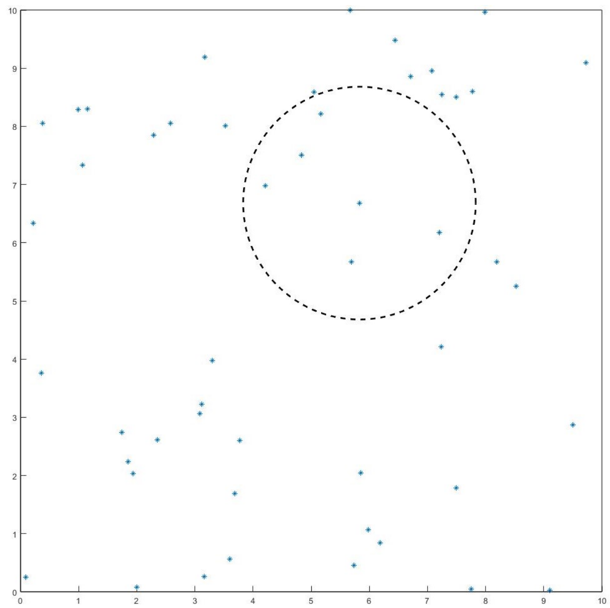

Figure 4.1: Geographical Distribution of A Simulated Network with 50 Agents

True Value of λ_0	Parameters	Mean	S.D.
0	λ	0.0074	0.1378
	β	1.0031	0.1114
0.5	λ	0.5067	0.1162
	β	1.0042	0.9920
2	λ	2.0039	0.1093
	β	1.0034	0.9897

Table 4.1: Results of Monte Carlo Simulation for Networks with 50 Agents

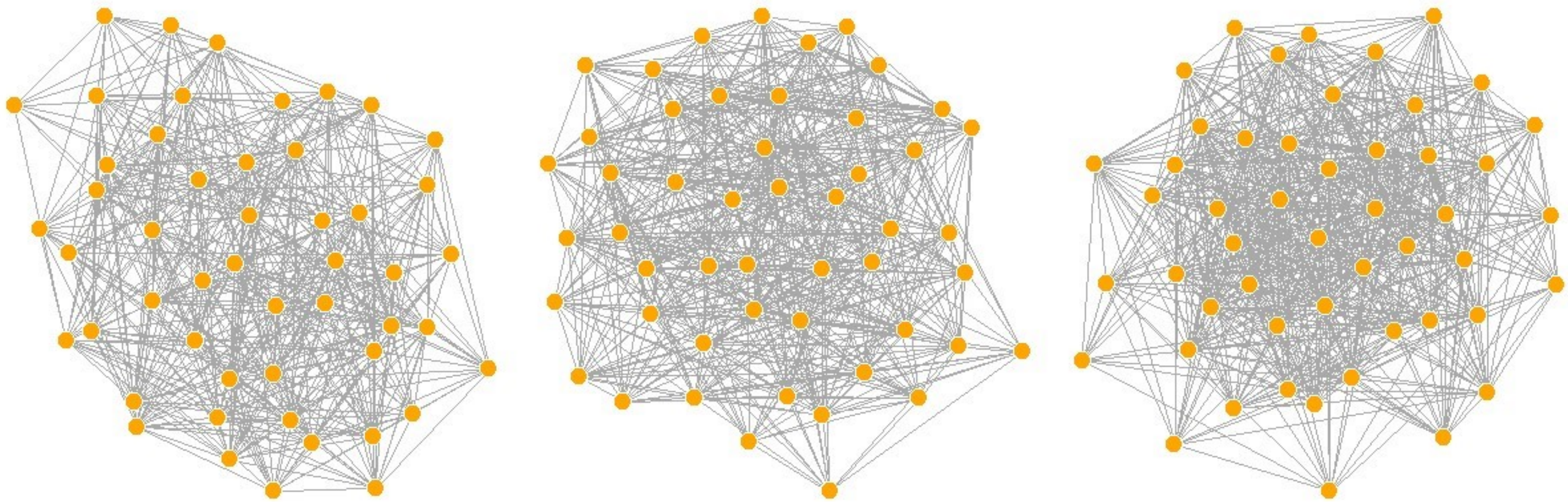

Figure 4.2: Graph of Simulated Networks with $n = 50$

The graph on the left corresponds to the case in which $\lambda = 0$, the graph in the middle corresponds to the case in which $\lambda = 0.5$, and the graph on the right corresponds to the case in which $\lambda = 2$.

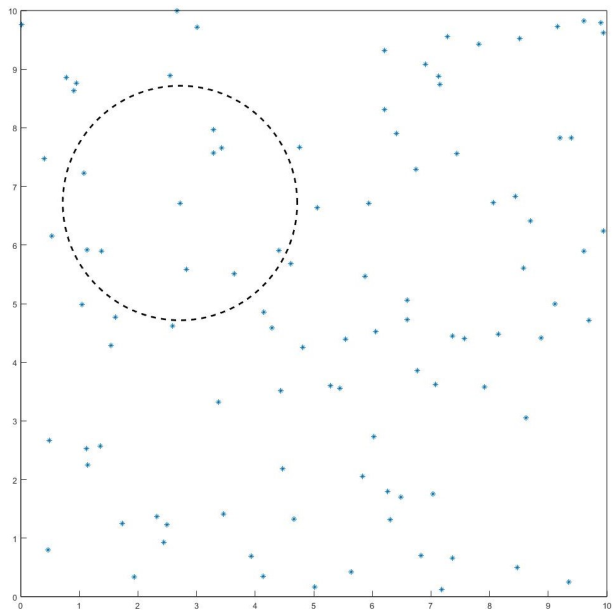

Figure 4.3: Geographical Distribution of A Simulated Network with 100 Agents

True Value of λ_0	Parameters	Mean	S.D.
0	λ	0.0056	0.0723
	β	0.1027	0.0606
0.5	λ	0.5043	0.0642
	β	1.0019	0.0613
2	λ	2.0051	0.0621
	β	1.0022	0.0584

Table 4.2: Results of Monte Carlo Simulation for Networks with 100 Agents

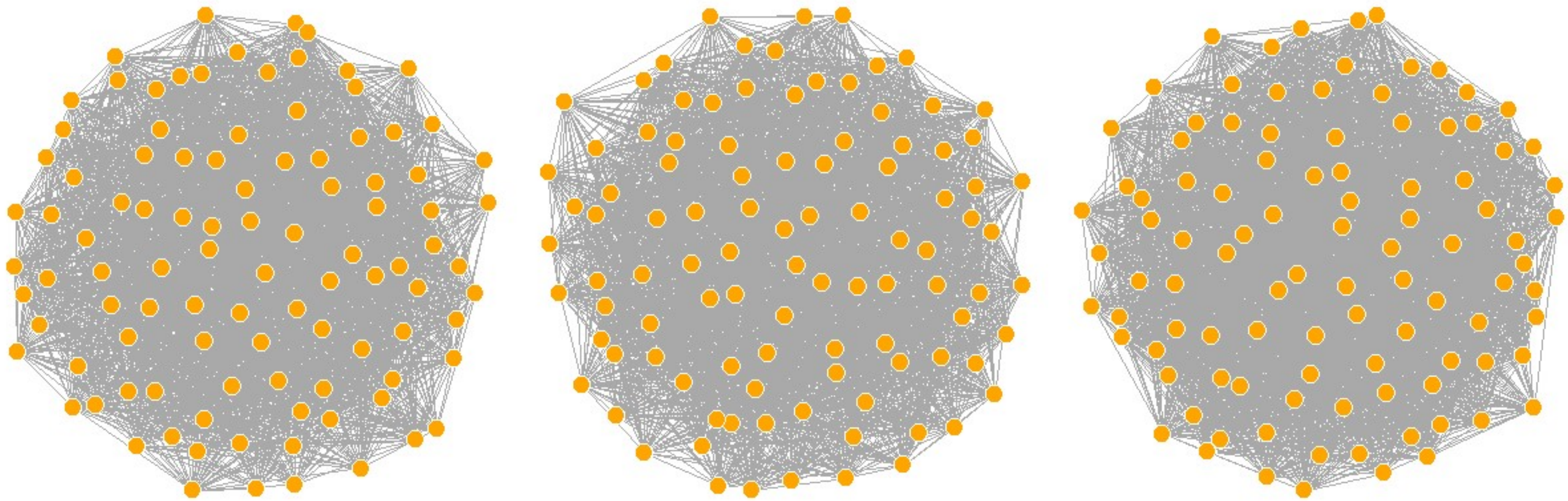

Figure 4.4: Graph of Simulated Networks with $n = 100$

The graph on the left corresponds to the case in which $\lambda = 0$, the graph in the middle corresponds to the case in which $\lambda = 0.5$, and the graph on the right corresponds to the case in which $\lambda = 2$.

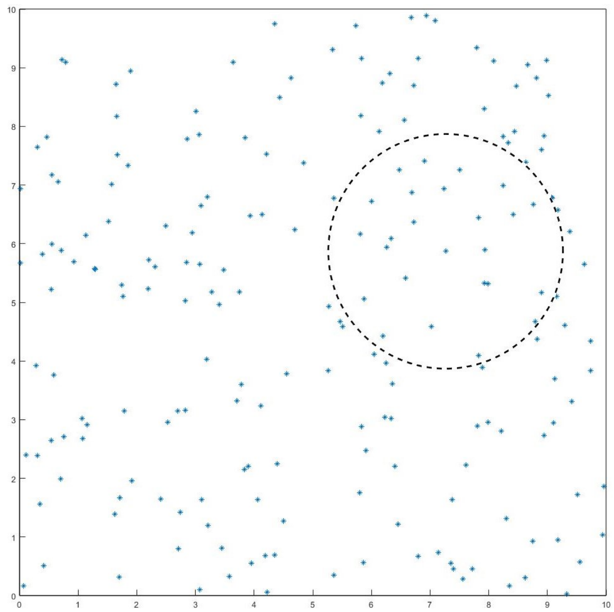

Figure 4.5: Geographical Distribution of A Simulated Network with 200 Agents

True Value of λ_0	Parameters	Mean	S.D.
0	λ	0.0035	0.0340
	β	0.9993	0.0312
0.5	λ	0.5016	0.0315
	β	1.0018	0.0296
2	λ	2.0029	0.0289
	β	1.0009	0.0282

Table 4.3: Results of Monte Carlo Simulation for Networks with 200 Agents

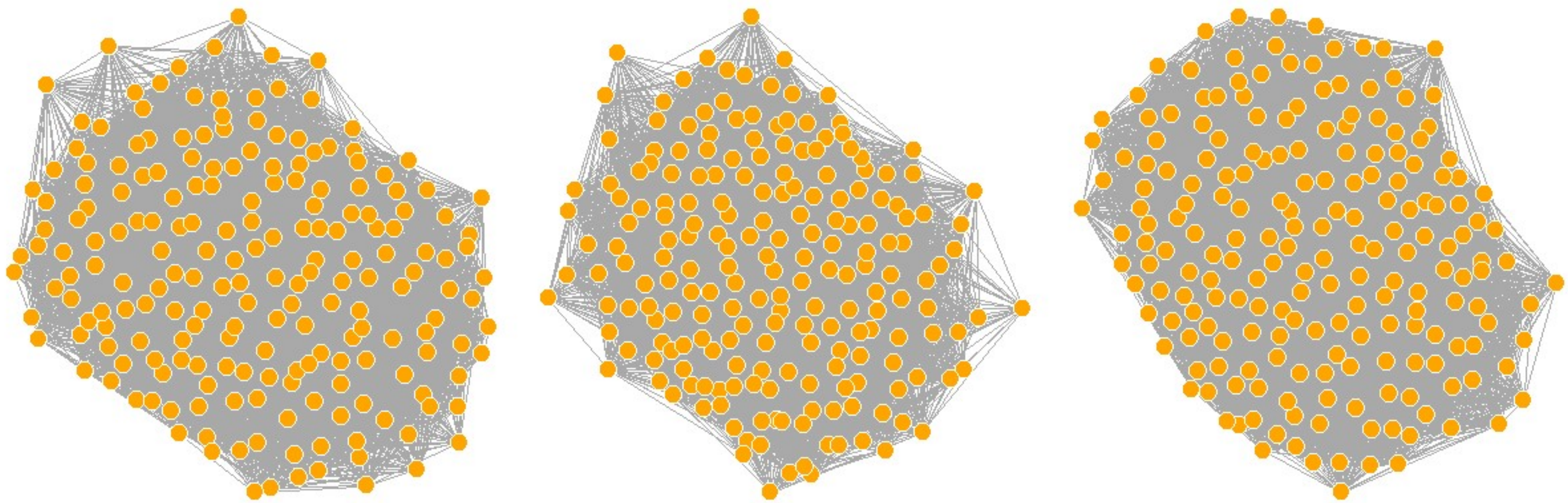

Figure 4.6: Graph of Simulated Networks with $n = 200$

The graph on the left corresponds to the case in which $\lambda = 0$, the graph in the middle corresponds to the case in which $\lambda = 0.5$, and the graph on the right corresponds to the case in which $\lambda = 2$.

The simulation results show that the nested pseudo likelihood algorithm (NPL) works well for estimating the parameters of the simultaneous network formation model discussed in this paper. The estimated values of parameters are essentially unbiased. The simulation results also imply that agents i and j are more likely to link to each other if the neighborhood effect becomes stronger, which is consistent with our intuitions.

CHAPTER 5

Future Research Directions

In this chapter, I would like to briefly discuss the future research directions of the network formation model studied in this thesis as well as the topic of network formation process in general.

The first interesting extension of the model is to estimate the neighborhood size d directly from the realized network data. In this thesis, we assume that the neighborhood size d is known to researchers, which is a high-level assumption, and we have to admit that in most cases, researchers may not have information about agents' neighborhood size d. However, if researchers pick a wrong value of d, the model is likely to be misspecified and can generate misleading conclusions. To see this, we can consider two extreme cases. First, notice that there will be no agents in agent i's neighborhood for all $i \in V$ if $d \to 0$. In this case, the benchmark model degenerates into the simplest bilateral network formation model, which is essentially a logistic regression model. In the second case, if we assume the support of $\{X\}_i$ is finite and $d \to \infty$, the benchmark model will converges to a stochastic block model when the number of agents n grows. A true neighborhood size d is likely to make the model somewhere in between the simplest model and the stochastic block model. Estimating d from network is in fact possible. The intuition is that if there is enough variation in the geographical distribution of all agents, a misspecified d will lead to a lower likelihood. However, since the number of agents in the neighborhood is a discrete, there may exist multiple d which can fit the data, which implies d could only be partial identified if we don't impose strong assumptions.

A second potential research direction is to consider use more complicated utility functions. It is possible that the homophily, which is represented by W_{ij} in the benchmark model, is a nonlinear function of individual characteristics of two agents. In the benchmark model, we assume the function is known to simply our analysis, which may lead to bias in reality. One possible way to overcome this problem is to use semiparametric or nonparametric utility function, which is left for future studies.

The last research direction is to adopt machine learning methods into the economic analysis of network formation process. The network formation model is essentially a supervised learning problem, in which the input include variables that can influence the formation of the link between agents i and j, and output is 1 if the link is formed and 0 otherwise. Machine learning methods, including neutral network (NN), deep neutral network (DNN) and random forests, have achieved great success in many supervised learning problems. These machine learning techniques may also be applied to develop empirical network formation models, which may have wide application in the real world. To my knowledge, some statisticians and researchers of computer science have already applied machine learning methods into developing models for link prediction. For example, Li et al. (2014) developed a deep learning framework to study the link formation process in a dynamic environment. As mentioned in the first chapter of this thesis, it is very difficult to consider the interdependence structure contained in social networks. Analytical models can barely catch every aspect since the interdependent relations are largely remain unknown and are usually nonlinear. Machine learning methods are advantageous in dealing with nonlinear and complicated functions, making them promising in the economic analysis of network formation process.

CHAPTER 6

Conclusion

In this thesis, I studies a simultaneous network formation model with incomplete information. The model can help to explore how neighborhood effect influences the network formation process. We successfully show that the model is identified and can be consistently estimated using the nested pseudo-likelihood algorithm (NPL). I conduct various Monte Carlo simulations to investigate the finite sample performance of the NPL estimator. The simulation results implies that the neighborhood effect may play an important role in determining agents' decisions of forming links: a positive neighborhood effect can motivate agents to conform to the behavior of their neighbors. To my best knowledge, this thesis is the first work introduces the neighborhood effect into the analysis of network formation process. I also discuss three potential research directions in the thesis and I hope to conduct the related research in the future.

CHAPTER 7

Appendix

Proof of Proposition 2: The proposition 2 is proved by contradiction. First, for a strategy profile $\sigma = \{\sigma_{12}, ..., \sigma_{n-1,n}\}_{i<j}$ and the set of information $\mathcal{I}$, define

$$g_{ij}(\sigma, \mathcal{I}, \theta) = \frac{\exp\left[W_{ij}'\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma_{kl}\right]}{1 + \exp\left[W_{ij}'\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma_{kl}\right]}$$

Fix $\mathcal{I}$ and the number of agents n. Suppose there are two different equilibria, which are denoted by σ^* and $\sigma^\dagger$. There exists some pair (i,j) such that

$$\sigma_{ij}^* - \sigma_{ij}^\dagger = \sum_{(k,l)\in N_{ij}} \frac{\partial g_{ij}(\tilde{\sigma}, \mathcal{I}, \theta)}{\partial \sigma_{kl}}(\sigma_{kl}^* - \sigma_{kl}^\dagger),$$

where the equality comes from mean value theorem. Without loss of generality, we assume $\sigma_{kl}^* \leq \sigma_{kl}^\dagger$. We have $\sigma_{kl}^* \leq \tilde{\sigma}_{kl} \leq \sigma_{kl}^\dagger$, for all $(k,l) \in N_{ij}$. Notice that

$$\ln g_{ij}(\sigma, \mathcal{I}, \theta) = W_{ij}'\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma_{kl} - \ln\left(1 + \exp\left[W_{ij}'\beta + \frac{\lambda}{|N_{ij}|}\sum_{(k,l)\in N_{ij}} \sigma_{kl}\right]\right)$$

which implies

$$\frac{\partial g_{ij}(\sigma, \mathcal{I}, \theta)}{\partial \sigma_{ij}} = \frac{\lambda}{|N_{ij}|}\left[g_{ij}(\sigma, \mathcal{I}, \theta)(1 - g_{ij}(\sigma, \mathcal{I}, \theta)\right].$$

So,

$$
\begin{aligned}
\sigma_{ij}^* - \sigma_{ij}^\dagger &= g_{ij}(\sigma^*, \mathcal{I}, \theta) - g_{ij}(\sigma^\dagger, \mathcal{I}, \theta) \\
&= \sum_{N_{ij}} \frac{\lambda}{|N_{ij}|} \left[g_{ij}(\tilde{\sigma}, \mathcal{I}, \theta)(1 - g_{ij}(\tilde{\sigma}, \mathcal{I}, \theta) \right] (\sigma_{kl}^* - \sigma_{kl}^\dagger) \\
&\leq \frac{1}{4} \lambda \max_{(k,l) \in N_{ij}} |\sigma_{kl}^* - \sigma_{kl}^\dagger|
\end{aligned}
$$

This implies

$$
\begin{aligned}
\max_{i,j \in V} |\sigma_{ij}^* - \sigma_{ij}^\dagger| &\leq \frac{1}{4} \lambda \max_{i,j \in V} \max_{(k,l) \in N_{ij}} |\sigma_{kl}^* - \sigma_{kl}^\dagger| \\
&\leq \frac{1}{4} \lambda \max_{k,l \in V} |\sigma_{kl}^* - \sigma_{kl}^\dagger|
\end{aligned}
$$

If $\lambda < 4$, there will be a contradiction. So, there is only a unique equilibrium if $\lambda < 4$.

Bibliography

Aguirregabiria, Victor and Pedro Mira, “Sequential estimation of dynamic discrete games,” Econometrica, 2007, 75 (1), 1–53.

Amini, Arash A, Aiyou Chen, Peter J Bickel, Elizaveta Levina et al., “Pseudo-likelihood methods for community detection in large sparse networks,” The Annals of Statistics, 2013, 41 (4), 2097–2122.

Badev, Anton, “Discrete Games in Endogenous Networks: Equilibria and Policy,” arXiv preprint arXiv:1705.03137, 2017.

Bajari, Patrick, Han Hong, John Krainer, and Denis Nekipelov, “Estimating static models of strategic interactions,” Journal of Business & Economic Statistics, 2010, 28 (4), 469–482.

Bickel, Peter J and Aiyou Chen, “A nonparametric view of network models and Newman–Girvan and other modularities,” Proceedings of the National Academy of Sciences, 2009, pp. pnas–0907096106.

Boucher, Vincent and Ismael Mourifié, “My friend far, far away: a random field approach to exponential random graph models,” The Econometrics Journal, 2017, 20 (3), S14–S46.

Brock, William A and Steven N Durlauf, “Discrete choice with social interactions,” The Review of Economic Studies, 2001, 68 (2), 235–260.

Erdös, Paul and Alfréd Rényi, “On random graphs, I,” Publicationes Mathematicae (Debrecen), 1959, 6, 290–297.

fei Lee, Lung, Ji Li, and Xu Lin, “Binary choice models with social network under heterogeneous rational expectations,” Review of Economics and Statistics, 2014, 96 (3), 402–417.

Fishkind, Donniell E, Daniel L Sussman, Minh Tang, Joshua T Vogelstein, and Carey E Priebe, “Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown," SIAM Journal on Matrix Analysis and Applications, 2013, 34 (1), 23–39.

Frank, Ove and David Strauss, "Markov graphs," Journal of the American Statistical Association, 1986, 81 (395), 832–842.

Graham, Bryan S, "Methods of Identification in Social Networks," Annual Review of Economics, 2015, 7 (1), 465–485.

_ , "An econometric model of network formation with degree heterogeneity," Econometrica, 2017, 85 (4), 1033–1063.

Hoff, Peter D, Adrian E Raftery, and Mark S Handcock, "Latent space approaches to social network analysis," Journal of the American Statistical Association, 2002, 97 (460), 1090–1098.

Holland, Paul W and Samuel Leinhardt, "An exponential family of probability distributions for directed graphs," Journal of the american Statistical association, 1981, 76 (373), 33–50.

_ , Kathryn Blackmond Laskey, and Samuel Leinhardt, "Stochastic blockmodels: First steps," Social networks, 1983, 5 (2), 109–137.

Hsieh, Chih-Sheng and Lung-Fei Lee, "Specification and estimation of network formation and network interaction models with the exponential probability distribution," 2017.

Jackson, Matthew O, "The economics of social networks," 2005.

Kasahara, Hiroyuki and Katsumi Shimotsu, "Sequential estimation of structural models with a fixed point constraint," Econometrica, 2012, 80 (5), 2303–2319.

Lei, Jing, Alessandro Rinaldo et al., "Consistency of spectral clustering in stochastic block models," The Annals of Statistics, 2015, 43 (1), 215–237.

Leung, Michael P, "Two-step estimation of network-formation models with incomplete information," Journal of Econometrics, 2015, 188 (1), 182–195.

Li, Xiaoyi, Nan Du, Hui Li, Kang Li, Jing Gao, and Aidong Zhang, "A deep learning approach to link prediction in dynamic networks," in "Proceedings of the 2014 SIAM International Conference on Data Mining" SIAM 2014, pp. 289–297.

Lin, Zhongjian and Haiqing Xu, "Estimation of social-influence-dependent peer pressure in a large network game," The Econometrics Journal, 2017, 20 (3), S86–S102.

Liu, Xiaodong and Jiannan Zhou, "A social interaction model with ordered choices," Economics Letters, 2017, 161, 86–89.

Mele, Angelo and Lingjiong Zhu, "Approximate variational estimation for a model of network formation," arXiv preprint arXiv:1702.00308, 2017.

Miyauchi, Yuhei, "Structural estimation of pairwise stable networks with nonnegative externality," Journal of Econometrics, 2016, 195 (2), 224–235.

Nowicki, Krzysztof and Tom A B Snijders, "Estimation and prediction for stochastic blockstructures," Journal of the American Statistical Association, 2001, 96 (455), 1077–1087.

Ridder, Geert and Shuyang Sheng, "Estimation of Large Network Formation Games," Technical Report, Working papers, UCLA 2015.

Rohe, Karl, Sourav Chatterjee, Bin Yu et al., "Spectral clustering and the high-dimensional stochastic blockmodel," The Annals of Statistics, 2011, 39 (4), 1878–1915.

Sheng, Shuyang, "A structural econometric analysis of network formation games," Unpublished, http://www. econ. ucla. edu/people/papers/Sheng/Sheng626. pdf, 2014.

Snijders, Tom AB and Krzysztof Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," Journal of classification, 1997, 14 (1), 75–100.

Tamer, Elie, "Incomplete simultaneous discrete response model with multiple equilibria," The Review of Economic Studies, 2003, 70 (1), 147–165.

_ , "Partial identification in econometrics," Annu. Rev. Econ., 2010, 2 (1), 167–195.

Wang, Yuchung J and George Y Wong, "Stochastic blockmodels for directed graphs," Journal of the American Statistical Association, 1987, 82 (397), 8–19.

Xu, Haiqing, "Social interactions in large networks: A game theoretic approach," International Economic Review, 2018, 59 (1), 257–284.