Structural Identification and Estimation of Peer Effects Using Network Data

Fei, WANG

Economic Department, University of Konstanz

School of Economics, University of Nottingham

April 17, 2019

Advisor: Professor Dr. Winfried Pohlmeier; Dr. Livia Shkoza

Student ID: 942870

Software: R

Contents

1 Introduction

The famous motto - *Correlation does not imply causality* - keeps haunting economists, especially when it comes to the identification and estimation of peer effects in social networks. On one hand, similarities (either on micro or macro-level) begets friendship. On the other hand, the interaction or formation of friendship within social network also affect behaviors of agents. A perfect way to identify and estimate the peer effects is to run a random experiment that can isolate the peer effects from other factors. An imagined experiment shown in figure 1.1 leads us to answer a very primary question: what do we mean when we talk about 'peer effects'?

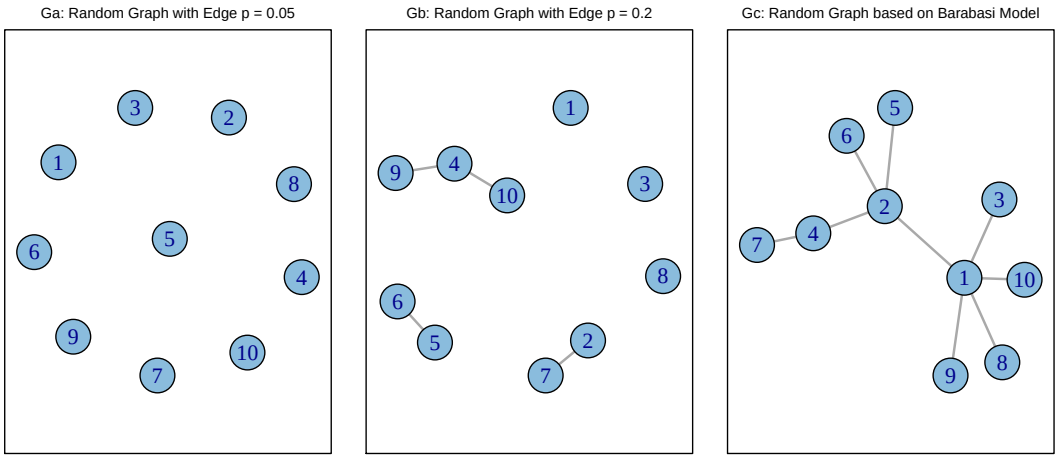

Figure 1.1: Random Graph generated with 10 nodes

According to *Cambridge Dictionary*, peer is a person who is the same age or has the same position or the same abilities as other people in a group. In terms of peer effects, we can have at least three ways of measurement based on graphs[1] in Figure 1.1 by assuming all agents represented by nodes are peers:

- Ga Vs Gb: Both networks are generated with independent edge probability, peer effects exist due to the random connected network
- Ga Vs Gc: Ga is generated with independent edge probability, whereas Gc is generated based on a preferential attachment mechanism (Albert and Barabási, 2002). In Gc, peer effects exist due to the self-selection and interaction of agents.
- Gb Vs Gc: The peer effects exist in Gb might push agents to break up with the old connection and then form a new one which can be in the format like Gc.

With the discourse on imagined experiment, it should be acknowledged that the examination on network formation is very important to identify and estimate peer effects. This paper will try to model the identification and estimation of peer effects with structural examination of network formation. First, it will do a short survey on peer effects. Second, it will present the structural model of peer effects in social networks based on

[1]The methods and models used for generating random graphs are explained in Appendix A.

the investigation by Calvó-Armengol et al. (2009). As usual the estimation and simulation will be given after the presentation of model and data description. In the end, the limitations of this paper and possible future research proposals will be discussed.

2 A Short Survey on Peer Effects

Although there are thousands of papers on Peer effects in education or human behaviors (like drinking or smoking) study, the concerns of those scholars are very unvaried. Almost all of them try to tackle the reflection problem analyzed by Manski (1993). This section will again go further to discuss this issue, examine the weak instruments problem when using the characteristics of second order friends as IVs, and then discuss how to disentangle the network formation.

2.1 The Reflection Problem of Peer Effects

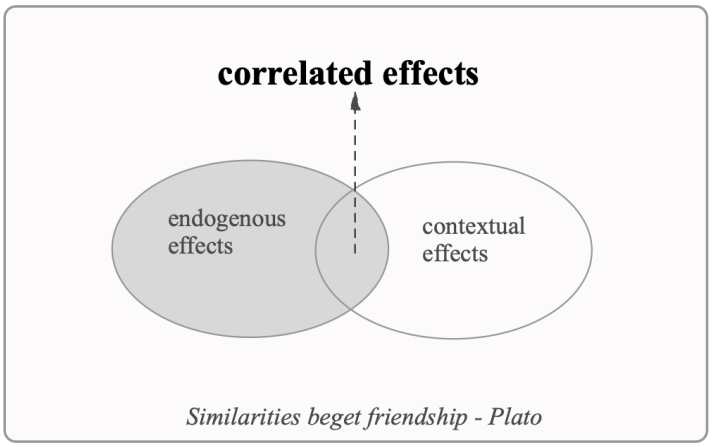

Figure 2.1: Illustration of Reflection Problem

According to Manski (1993), the reflection problem of peer effects can be explained by the venn diagram in Figure 2.1, where those three effects are defined in the following way:

(a) endogenous effects, wherein the propensity of an individual to behave in some way varies with the behavior of the group.

(b) exogenous (contextual) effects, wherein the propensity of an individual to behave in some way varies with the exogenous characteristics of the group.

(c) correlated effects, wherein individuals in the same group tend to behave similarly because they have similar individual characteristics or face similar institutional environment.

Intuitionally, table 2.1 explains why it is so difficult to differentiate endogenous effect and contextual effects[2]. Taking Network 1 as an example, researchers who are trying to

[2]Any study on causality, especially numerical identification of causality is trying to find the significant difference between the group with treatment and the one without treatment. This is a very ancient philosophy of doing science.

identify the peer effects in Network 1 can barely find the significant differences between endogenous and contextual effects with the existence of profound collinearity between all observations. For instance, suppose all observations have almost same income and same abilities, etc., then the achievements will be almost same with peer effects (connected in subgroup within network) and without peer effects (not connected in subgroup within the same network). Similar issue still exist for the Network 2. Although Network 3 would be preferred for researchers, it is the very rare case in social network formation.

Table 2.1: Network Examples with Different Correlations

Group	**Similarities**	**Collinearity**	**Differences between endogenous and contextual effects**
Network 1	Same position and abilities	Very High	Almost no way to identify
Network 2	Similar position and abilities	High	Weak identification
Network 3	Different positions and abilities	Low	Strong identification

How about we compare Network 1 (or 2) with Network 3? Will this search makes it easier to identify the peer effects. Unless we have the very strong overlapped covariate distributions among two groups, it is difficult to tell whether the differences of achievements among two networks are due to different abilities or different network connection behaviors. Or in other words, researchers is highly possible to end with the situation which a very large share of achievements differences can be explained by the contextual effects, wherein for instance, individuals from Network 1 (or 2) have better income (or education, or higher abilities) than those from Network 3, then have better achievement no matter they are connected or not (with or without peer effects).

With the detailed explanation on three cases, it is clear that how difficult it can be to identify the peer effects in social network. Even through the randomized experiment done by Sacerdote (2001), the common attributes of students from Dartmouth still yields some doubts on the identification of peer effects. More Rigorously, Manski (1993) shows that the linear-in-means model is not identified without knowing anything about the group structure. Considering the idea of Manski (1993) is such a seminal one and also self-named, I will present and discuss his model again.

Follow the setup by Manski (1993), let the agent in the network be characterized by a value for $(y, x, z, u) \in R^y \times R^x \times R^z \times R^u$. Here y is a scalar representing the achievement, x are attributes characterizing network where the agent belongs (e.g. school or ethnic group), and (z, u) are observable and unobservable exogenous attributes (e.g. socioeconomic status and ability). A researcher observes a random sample of realization of (y, x, z). Realizations of u are not observed.

Assume that

$$y = \alpha + \beta E(y|x) + E(z|x)'\gamma + z'\eta + u, \quad E(u|x, z) = x'\delta \tag{2.1}$$

Then the mean regression of y on (x, z) has the linear form

$$E(y|x, z) = \alpha + \beta E(y|x) + E(z|x)'\gamma + x'\delta + z'\eta \tag{2.2}$$

where $(\alpha, \beta, \gamma, \delta, \eta)$ is a parameter vector, and

(a) If $\beta \neq 0$, the linear regression (2.2) expresses an endogenous effect.

(b) If $\gamma \neq 0$, the model expresses an exogenous effect

(c) If $\delta \neq 0$, the model expresses correlated effects.

(d) The parameter η expresses the direct effect of z on y.

Manski (1993) proved that the composite parameters in the model (2.2) can be identified if the regressors $[1, E(z|x), x, z]$ are linearly independent in the population with $\beta \neq 1$, which means that the attributes characterizing network are linearly independent with exogenous attributes (e.g. socioeconomic status). This is highly unlikely to see from the social network data. Furthermore he also proved that in the model (2.2) with $\beta \neq 1$, the composite social-effects parameter is not identified if any of these conditions hold, almost everywhere.

(a) z is a function of x

(b) $E(z|x)$ does not vary with x

(c) $E(z|x)$ is a linear function of x

This means that in the context of the linear model (2), inference is possible only $E(z|x)$ varies non-linearly with x and $Var(z|x) > 0$.

To solve this reflection problem in the process of identifying endogenous social effects, the spatial autoregressive (SAR) model was proposed

$$y = \lambda Ay + X\beta + \varepsilon \tag{2.3}$$

, where A is adjacency matrix and X is a vector of exogenous variable, and the model assume:

$$\varepsilon_i | A, X \sim N(0, \sigma^2) \tag{2.4}$$

Assumption (2.4) means that the network attributes represented by the adjacency matrix A is independent with exogenous variables of members in network, which is not plausible. In other words, assumption (2.4) is equivalent to make $E(u|x, z) = 0$ in the model examined by Manski (1993), which basically claims that there is no correlation between network formation and attributes of agents (or players) in the network.

At this stage, it is quite clear that to identify the endogenous social effects (or peer effects) from contextual effects, we have the following solutions

(a) Assume $E(u|x, z) = 0$ (or $\varepsilon_i | A, X \sim N(0, \sigma^2)$ in SPA model), which is a very unplausible (or even blind) assumption.

(b) Find a possible and reasonable relationship to show that $E(z|x)$ varies non-linearly with x and $V(z|x) > 0$ as proposed by Manski (1993).

In section 2.2, we will show that even with assuming $\varepsilon_i | A, X \sim N(0, \sigma^2)$ in SPA model, the weak instruments/identification issue still exists, and to some extent, it is even fatal to identify the peer effects with this assumption. In section 2.3, it will present that is is possible to show $E(z|x)$ varies non-linearly with x and $V(z|x) > 0$ by disentangling the network formation.

2.2 Weak Instruments/Identification Issue

Now, start again from the SAR model in (2.3):

$$y = \lambda Ay + X\beta + \varepsilon$$

by assuming $\varepsilon_i|A, X \sim N(0, \sigma^2)$. Suppose $I_n - \lambda A$ is nonsingular, then the above equation can be rewritten as

$$y = (I_n - \lambda A)^{-1}X\beta + (I_n - \lambda A)^{-1}\varepsilon \tag{2.5}$$

Since $E[\varepsilon' Ay] \neq 0$, the spatial lag Ay is endogenous and we can show:

$$E[Ay] = A(I_n - \lambda A)^{-1}X\beta = AX\beta + \lambda A^2 X\beta + \lambda^2 A^3 X\beta + \cdots$$

, which leads to use $[AX, A^2X, A^3X, \cdots]$ as IVs for Ay. This method was first proposed by Bramoullé et al. (2009). Bramoullé et al. (2009) setup the model as

$$Y = \lambda Ay + X\beta + \delta AX + \varepsilon, \tag{2.6}$$

which using A^2X as IV. However, the instrument A^2X must be providing some information about AY, which has been discussed by Manski (1993). Or in other words, as everyone has more friends everyone's second-order friends start to look more and more similar, which leads to the issue of weak identification. Based on the model by Bramoullé et al. (2009), Startz and Wood-Doughty (2017) run the Monte Carlo simulation of 2SLS estimation using A^2X as IVs[3]. They showed that weak instruments issue results in biased standard error but only modestly biased coefficients. The estimation of peer effects λ is highly inconsistent with various network densities (Startz and Wood-Doughty, 2017). In addition, there is a dramatic decline in power of test when increasing the sample size and including more covariates, which should be expected considering the higher collinearity with larger sample size or more covariates in social network. Figure C.1 in Appendix C gives the simulation results for peer effects estimation $\hat{\lambda}$ ($\hat{\beta}$ in their paper) by Startz and Wood-Doughty (2017) .

2.3 Disentangle the Network Formation

The reflection problem with assumption $E(u|x, z) \neq 0$ (or $E(\varepsilon|A, X) \neq 0$) makes the separation of endogenous and contextual effects extremely difficult. The existence of weak instruments with high network density makes the estimation of peer effects inconsistent when assuming $E(u|x, z) = 0$. Like what it has discussed in introduction part based on Figure 1.1, it is fruitless to identify and estimate peer effects without investigating the formation of network, especially when assuming $E(u|x, z) \neq 0$. In this subsection, I will present a few of the workhorse models of static and dynamic network formations and then discuss the implications for identifying and estimating peer effects based on formation process.

[3]Both Bramoullé et al. (2009) and Startz and Wood-Doughty (2017) used the well-known `Add Health Dataset`.

2.3.1 Static Models of Random Networks

The *static* models of random networks refers to the case that all nodes in network are established at the same time and then links are drawn between them according to some probability rule Jackson (2010). Graph *Ga and Gc* in Figure 1.1 are this kind of network. By assuming edges (or links) in network are formed independently, the *degree distribution of a random network* can be proved to follow Poisson distribution

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!} \tag{2.7}$$

Equation (2.7) can help us to check whether the network fit the random graph or not if we know the average degree of a network λ. For more detailed derivations on both static models of network formation and dynamic one, please refer to Appendix A.

2.3.2 Dynamic Models of Network Formation

The poisson form offers a poor fit for the most network datasets in real world as it could not generate the hubs. The degree distribution of most real world networks are well approximated with

$$p_k \sim k^{-\gamma} \tag{2.8}$$

Equation (2.8) is called a *power law distribution* and the exponent γ is its *degree exponent*. If we take a logarithm of (2.8), we obtain

$$\log p_k \sim -\gamma \log k \tag{2.9}$$

If (2.8) holds, $\log p_k$ is expected to depend linearly on $\log k$, the slope of this line being the degree exponent γ. A *scale-free network* is a network whose degree distribution follows a power law.

Based on the scale-free network model, Barabási et al. (2016) proposed the *Barabasi-Albert* model by recognizing growth and preferential attachment coexisting in real networks. The Barabasi-Albert model is such of importance to economists as it is built based on preferential attachment, which means it aligns the assumed utility-based agents in economic models. More importantly, it provides the feasible way to measure the preferential attachment, which will be used in this paper to exam the network is random network or preferential attachment network.

In summary, we can fit the degree of distribution of network with Poisson or power law distribution to check whether the network of interest is static random network or scale-free network.

3 Structural Model of Peer Effects in Social Network

A short survey on peer effects in section 2 has shown that to identify the endogenous peer effects from contextual effects, we have two solutions:

(a) Assume $E(u|x,z) = 0$ (or $\varepsilon_i|A,X \sim N(0,\sigma^2)$ in SPA model), which is a very unplausible (or even blind) assumption.

(b) Find a possible and reasonable relationship to show that $E(z|x)$ varies non-linearly with x and $V(z|x) > 0$ as proposed by Manski (1993).

Solution (a) is plausible if the network is randomly generated, whereas solution (b) is plausible if the network is generated based on preferential attachment. This paper will model the peer effects structurally based on the pioneer work by Ballester et al. (2006). Ballester et al. (2006) modeled the network formation implicitly by analyzing the Nash equilibrium in the network game, which reflects the preferential attachment of network formation.

3.1 The Game

Following the setup by Ballester et al. (2006) and Calvó-Armengol et al. (2009), consider the network with a finite agents $N = \{1, \cdots, n\}$. The network can be represented by the adjacency matrix $G = [g_{ij}]$, where $g_{ij} = 1$ if i and j have mutual friendship, and $g_{ij} = 0$, otherwise[4]. Each player $i = 1, \cdots n$ selects an effort $x_i \geq 0$ and obtain the payoff

$$u_i(x_1, \cdots, x_n) = \alpha_i x_i - \frac{1}{2}x_i^2 + \lambda \sum_{j=1}^{n} g_{ij} x_i x_j \tag{3.1}$$

, which is strictly concave in own effort. Now, decompose the effort x_i into two parts: y_i the effort of individual i absent of any peer influence; z_i the peer effort that corresponds to peer efforts from friends. Each agent i decides how much y_i and z_i to put during interaction process. Then, the model (3.1) can be rewritten as

$$u_i(y, z; g) = \alpha_i(y_i + z_i) - \frac{1}{2}(y_i + z_i)^2 + \lambda \sum_{j=1}^{n} g_{ij} z_i z_j \tag{3.2}$$

, where the last term of (3.1) can be replaced with $\lambda \sum_{j=1}^{n} g_{ij} z_i z_j$ as $\lambda \sum_{j=1}^{n} g_{ij} y_i y_j = 0$. This condition holds when we assumed that two efforts y_i and z_i are indepedent. Equation (3.2) can be expanded as

$$u_i(y, z; g) = \alpha_i y_i - \frac{1}{2}y_i^2 + \alpha_i z_i - \frac{1}{2}z_i^2 - y_i z_i + \lambda \sum_{j=1}^{n} g_{ij} z_i z_j \tag{3.3}$$

Equation (3.3) shows that assuming y_i and z_i still generates correlated effects in utility level as

$$\begin{aligned}
&\frac{\partial u_i}{\partial y_i} = \alpha_i - y_i - z_i; && \Rightarrow \quad y_i^* = \alpha_i - z_i^* \\
&\frac{\partial u_i}{z_i} = \alpha_i - z_i - y_i + \lambda \sum_{j=1}^{n} g_{ij} z_j; && \Rightarrow \quad z_i^* = \alpha_i - y_i^* + \lambda \sum_{j=1}^{n} g_{ij} z_j^* \\
&\frac{\partial u_i}{\partial y_i \partial z_i} = -1;
\end{aligned}$$

Solve the above first order conditions can give the equilibrium behaviors. However, the solution is not straightforward as z_i^* not only depends on y_i^* but also z_j^*, wherein y_i^* also

[4]If the friendship is directed, $g_{ij} = 1$ if i list j as her friend, whereas $g_{ji} = 0$ when j does not list i as her friend. I will not model directed network, but it will be discussed in the coming context.

depends on z_i^*. This means the best response of player i in this network game depends on the best responses of all other players that are linked by g_{ij}. And each player has to decide how much y_i and z_i to put depends on the connections (or positions) in the network. The pioneer paper by Ballester et al. (2006) proved that the Nash Equilibrium exists and it is rooted in the vector of *Katz-Bonacich centralities* (Katz, 1953; Bonacich, 1987), which is defined in the following way.

Definition 3.1.1. (Katz, 1953; Bonacich, 1987) Given a vector $u \in \mathbb{R}_+^n$, and $\phi \geq 0$ a small enough scalar, the vector of Katz-Bonacich centralities of parameter ϕ in network g is defined as:

$$\boldsymbol{b}(g, \phi) = (I - \phi G)^{-1} u = \sum_{p=0}^{\infty} \phi^p G^p u \tag{3.4}$$

3.2 A network model of peer effects

Based on the model by Ballester et al. (2006), Calvó-Armengol et al. (2009) later simplify equation (3.3) as

$$u_i(y, z; g) = \alpha_i y_i - \frac{1}{2} y_i^2 + \mu g_i z_i - \frac{1}{2} z_i + \phi \sum_{j=1} g_{ij} z_i z_j \tag{3.5}$$

The part $y_i z_i$ in (3.3) is merged into the coefficient α_i in (3.5):

$$\alpha_i(x) = \sum_{m=1}^{M} \beta_m x_i^m + \frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} \gamma_m g_{ij} x_j^m$$

, where x_i^m is a set of M variables accounting for observable differences in covariate variables of players and β_m and γ_m are parameters. Again deriving the first order conditions gives the optimal efforts:

$$y_i^*(x) = \alpha_i(x) = \sum_{m=1}^{M} \beta_m x_i^m + \frac{1}{g_i} \sum_{m=1}^{M} \sum_{j=1}^{n} \gamma_m g_{ij} x_j^m \tag{3.6}$$

$$z_i^*(g) = \mu g_i + \phi \sum_{j=1}^{n} g_{ij} z_j^* \tag{3.7}$$

As the game is symmetric, in equilibrium the best responses z_i^* and z_j^* can be delineated in the Katz-Bonacich centralities of ϕ. Hence, the individual equilibrium outcome $E_i(x)$ is uniquely defined and give by:

$$\begin{aligned} E_i(x) &= y_i^*(x) + z_i^*(g) \\ &= \underbrace{\alpha_i(x)}_{exogenous} + \underbrace{\frac{\mu}{\phi} \boldsymbol{b}_i(g, \phi)}_{endogenous} \end{aligned} \tag{3.8}$$

The proof for equation (3.8) can be found in the paer by Calvó-Armengol et al. (2009).

Now, suppose there are K network components in the dataset, Calvó-Armengol et al. (2009) also provided the following empirical strategy:

$$y_{i,\kappa} = \sum_{m=1}^{M} \beta_m x_{i,\kappa}^m + \frac{1}{g_{i,\kappa}} \sum_{m=1}^{M} \sum_{j=1}^{n_\kappa} \gamma_m g_{i,j\kappa} x_{j,\kappa}^m + \eta_\kappa + \varepsilon_{i,\kappa} \tag{3.9}$$

$$\varepsilon_{i,\kappa} = \mu g_{i,\kappa} + \phi \sum_{j=1}^{n_\kappa} g_{i,j\kappa} \varepsilon_{j,\kappa} + v_{i,\kappa}, \quad i = 1, \cdots n; \ \kappa = 1, \cdots, K$$

In the model by Calvó-Armengol et al. (2009), they also consider the directed network. Here, I will only consider the undirected network. Then, equation (3.9) can be written in the matrix format:

$$Y = X^m \beta + DGX\gamma + \eta + \varepsilon \tag{3.10}$$

$$\varepsilon = \mu G\mathbf{1} + \phi G\varepsilon + v \tag{3.11}$$

, where Y is the vector of players' activities (e.g. education achievement), X^m is the observable differences in covariate variables of players, $D = diag(1/g_1, \cdots, 1/g_n)$ is a $n \times n$ matrix composed by the fraction of total links of each player ($g_1 = \sum_{j=1}^{n_k} g_{ij}$), η is a $n \times n$ diagonal matrix of network fixed effects, G is the adjacency matrix, v is random error. The model and empirical strategy proposed by Calvó-Armengol et al. (2009) are very similar to the spatial autoregressive (SAR) methods. However SAR does not provide the theoretical support to identify the peer effects, whereas Calvó-Armengol et al. (2009) claimed that the peer effects are identified if the structural parameters (μ, ϕ) uniquely determine the equation (3.11), which can be reduced as

$$\varepsilon = \mu(I - \phi G)^{-1} G\mathbf{1} + (I - \phi G)^{-1} v \tag{3.12}$$

Later, in estimation and simulations part, the differences between the model by Calvó-Armengol et al. (2009) and SAR methods will be discussed.

4 Dataset, Descriptive Statistics, and Primary Tests

4.1 Network Formation

The dataset used in this essay is the Panel Study of Cologne Gymnasium Students (KGP). The details on data collection process can be found in Hummell (1970). Thanks to the data cleaning by Dr Livia Shkoza, the final version of dataset has 3391 observations and 81 variables, which includes the friendship relationships and covariates of observations. Names and meanings of each variable in dataset can be found in Appendix B.

To get as much as network information, only rows that are all NA (Not available) values are removed at the beginning. In the end, there are 121 network components and they are represented by a big block matrix G (3385×3385). Table 4.1 gives the summary of network.

Table 4.1: Summary of Basic Network Measurements

	Directed Network		**Undirected Network**
	In	Out	Mutual
Average degree	5.826	5.826	2.754
Average density		0.0017	0.00077
Maximum degree	24	32	21

Table 4.1 is useful as we can use the average degree or density to simulate the random networks based on Erdos-Renyi model (the static model in Appendix A.1). After the simulation based on the average degree or density[5], we can check whether the network from data fits with simulated confidence interval or not in terms of different network properties. Figure 4.1 gives the goodness-of-fit by comparing the undirected network from data with 100 simulated network from exponential model. The algorithm behind the simulation can be found in the paper by Hunter et al. (2008). It can see that the network from KGP dataset deviates from the random networks in terms of edge-wise shared partners and minimum geodesic distance, whereas it fits with the random networks well in terms of triad census. Considering minimum geodesic distance and edge-wise shared partners are important properties of social network, it can be claimed that the undirected network from KGP is not formed randomly.

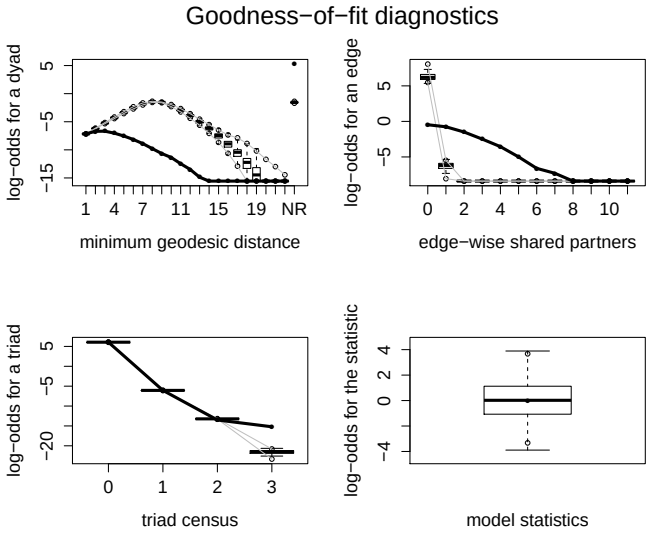

Figure 4.1: Fit with simulated random network (Hunter et al., 2008)

Now, the tidier dataset without NAs from covariate variables will be examined again. The exponential-family random graph models with exogenous covariate variables, such as grades or gender, etc., will be used to check how the networks are formed. During cleaning the dataset, removing NAs from covariate variables will lose the information on Network. For instance, once all NAs are removed from German grade in 1970, some observations

[5]To generate random network based on Erdos-Renyi model, we only need one indicator - average degree. With exponential probability model, one indicator (like average density) is also sufficient to generate a random network.

will be removed from the data, wherein the friends she/he list might not be included in the dataset. Hence, the network components will become smaller. Figure 4.2 gives an example from cleaned dataset and Figure C.2 in Appendix C gives the distributions of degree for directed and undirected networks.

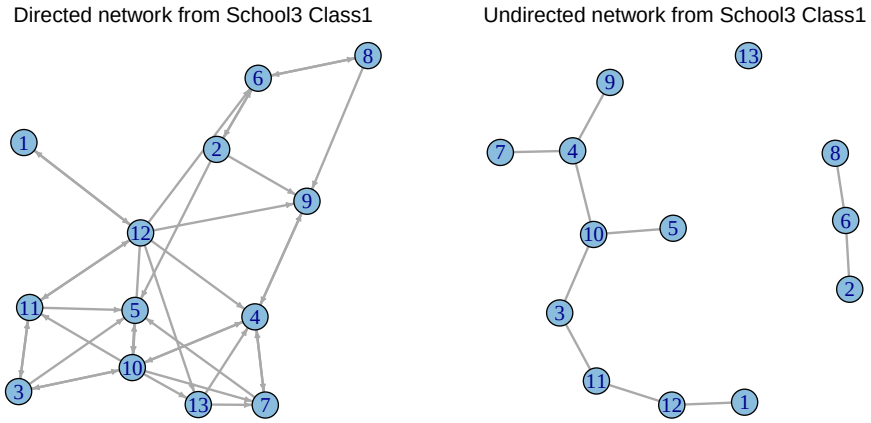

Figure 4.2: Network From School3 Class1

With cleaned dataset, the exponential random graph model (ERGM)[6] was fitted again to check the factors determining the formation of networks. Through ERGM, the weights of different factors on affecting network formation can be revealed. Table 4.2 gives the estimated results of EGRM for KGP dataset.

Table 4.2: ERGM fit for the KGP network with covariates

	Undirected Network			Directed Network		
Terms	Estimate	SE	p value	Estimate	SE	p value
edges	-9.02945^{***}	0.07457	<1e-04	-7.78916^{***}	0.02927	<1e-04
match.age	0.23427^{***}	0.03814	<1e-04	0.17132^{***}	0.01885	<1e-04
match.gender	2.64159^{***}	0.07508	<1e-04	2.12883^{***}	0.02972	<1e-04
match.prestige	0.36691^{***}	0.06843	<1e-04	0.25575^{***}	0.03520	<1e-04
match.gpa_68	0.90769^{***}	0.08779	<1e-04	0.86141^{***}	0.04421	<1e-04
match.IQ	0.03712	0.10553	0.725	0.10175^{*}	0.05030	0.0431
match.gpa_70	1.06089^{***}	0.08692	<1e-04	0.99540^{***}	0.04411	<1e-04
BIC	42455			161263		

The estimated coefficients from Table 4.2 can be interpreted as the conditional log-odds of the network depended on the corresponding covariates. For instance, the log-odds of the edge (or tie, or link) between two nodes is around 2.6416 when the gender of those nodes are matched in undirected network. According to Table 4.2, most estimates for matched terms are positive, which implies that some types of nodes are more likely to form ties than others. Among all terms, the gender match has the largest weight expect for the edges.

[6]The explanation on ERGM is given in Appendix D.

Figure C.3 and C.4 gives the goodness-of-fit based on the Monte Carlo simulations. Compared to the fit from Figure 4.1, Figure C.3 and C.4 shows there is no improvement in terms of minimum geodesic distance and edge-wise shared partners even though several important covariates have been added in Table 4.2. This is because ERGM assumes that edges in the network are considered conditionally independent if they don't share a node. However, in real word, social network is more like preferential attachment which is the type of scale-free type of networks. In other words, edges are endogenously formed when the later comer (she/he) is more willing to join an existed popular group. That's why we see the edge-wised shared partners from the data are much higher than those generated from ERGM model. In addition higher edge-wised shared partners means the exist of hubs in the social network, which can be generated only from preferential attachment models Albert and Barabási (2002).

In summary, once the network is relatively large enough, the network effects itself has more explaining power for edges formation than exogenous covariate variables. Take a very extreme example - Google, once it grew large enough, it will be used or connected by almost everyone no matter who you are, how old you are, etc,. Therefore, the preferential attachment properties of social network brings the serious challenges for ERGM models.

5 Estimation and Simulation

The primary tests from section 4 shows the exist of hubs in social network, which can be measured by centralities like Katz-Bonacich centralities (identifying the key players in network). This justifies the structural network model of peer effects discussed in section 3. As it has been mentioned in section 3, the model proposed by Calvó-Armengol et al. (2009) can be rewritten as the spatial autoregression model (SAR).

In this section, the reduced form of model discussed in section 3 will be estimated. Three main models will be estimated with different methods, wherein they will be represented in SPA format. Tables 5.1 gives the summary of estimated models.

Table 5.1: Summary of Estimated Models

Model	**Methods**	**Covariates**
Model 1a: $y = \phi Gy + X\beta + \alpha u + \varepsilon$	MLE without considering endogenous issue	previous gpa /previous score; gender dummy; age; IQ; prestige; father and mother education dummy (university)
Model 1b: $y = \phi Gy + X\beta + \alpha u + \rho\xi + \varepsilon$	MLE with considering endogenous issue	
Model 2: $y = \phi Gy + X\beta + GX\gamma + \alpha u + \varepsilon$	MLE with considering endogenous issue 2SLS with IVs 2SLS with IVs from estimated G	
Model 3a: $y = \phi Gy + X\beta + \alpha u + e;\ e = \rho Ge + \nu$ Model 3b: $y = \phi Gy + X\beta + GX\gamma + \alpha u + e; e = \rho Ge + \nu$	MLE	

For all models in table 5.1, four different academic performances, including English, German and Mathematic scores, and average GPA, will be used as the dependent variable. Table B.2 in Appendix B gives the descriptive statistics on main variables used in all models.

5.1 Results

Table 5.2 gives the estimations of different models with average GPA being dependent variables. By comparing the results of different models, it can be found that the estimation of peer effects (ϕ) is not consistent through different models. Almost all estimations of peer effect are not significant except for the one from 2SLS with original adjacency matrix. The estimated coefficients for fixed effect (α) and all other covariates are very close and consistent through all models. For instance, all estimates of previous GPA are in range from 0.65 to 0.69, and also significant.

Table 5.2: Estimated Results with Average GPA as Dependent Variable

	Model						
	(1a) MLE	(1b) MLE	(2) MLE	(2) 2SLS(G)	(2)2SLS ($\hat{G}$)	(3a)	(3b)
ϕ	-0.0005	-0.0005	0.0368	0.1106^{***}	-0.0151	0.0006	0.0034
	(0.0011)	(0.0011)	(0.0520)	(0.0286)	(0.0297)	(0.0025)	(0.0231)
α	1.0999^{***}	1.0992^{***}	1.1135^{***}	1.065^{***}	1.0564^{***}	1.0076^{***}	0.8903^{*}
	(0.1287)	(0.1287)	(0.1344)	(0.2921)	(0.2979)	(0.2460)	(0.3591)
Previous GPA	0.6710^{***}	0.6709^{***}	0.6577^{***}	0.6652^{***}	0.6832^{***}	0.6676^{***}	0.6984^{***}
	(0.0133)	(0.0133)	(0.0141)	(0.0147)	(0.0144)	(0.0256)	(0.0370)
Female	0.0102	0.0103	0.0153	0.0428	0.0581^{*}	0.0255	0.1319^{*}
	(0.0130)	(0.0130)	(0.0195)	(0.0273)	(0.0304)	(0.0283)	(0.0528)
age	0.0047	0.0047	0.0060	0.0043	0.0008	0.0156	0.0105
	(0.0073)	(0.0073)	(0.0077)	(0.0080)	(0.0080)	(0.0141)	(0.0204)
IQ	-0.0028^{***}	-0.0028^{***}	-0.0027^{***}	-0.0022^{***}	-0.0023^{***}	-0.0053^{***}	-0.0041^{*}
	(0.0007)	(0.0007)	(0.0007)	(0.0008)	(0.0008)	(0.0009)	(0.0020)
prestige	-0.0010^{*}	-0.0010^{*}	-0.0009^{*}	−0.0005	−0.0006	−0.0008	−0.0006
	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0005)	(0.0008)	(0.0013)
Mother edu	-0.0003	-0.0003	-0.0005	0.0005	0.0031	0.0279	0.0026
	(0.0144)	(0.0144)	(0.0147)	(0.0151)	(0.0151)	(0.0271)	(0.040)
Father edu	0.0066	0.0067	0.0074	0.0211	0.0203	−0.0150	−0.006
	(0.0153)	(0.0153)	(0.0155)	(0.0160)	(0.0162)	(0.0291)	(0.045)
Unobservable		0.0012	0.0021				
		(0.0039)	(0.0039)				
ρ						0.0552^{**}	0.056
						(0.0172)	(0.1317)
Contextual effects	no	no	yes	yes	yes	no	yes
AIC	976.7	978.61	973.18				

Note: $*p < 0.1; **p < 0.05; ***p < 0.01$

(1a)/(1b) is mode 1 without/with considering endogenous issue

(3a)/(3b) only uses the first 600/300 observations as it takes time to do MLE.

The pattern from table 5.2 still exists when different dependent variables are used in the model. The peer effect ϕ is estimated to be statistical insignificant most time across different models. All other estimated coefficients are quite similar from different models. Furthermore, the autocorrelation coefficient is significant in model (3a), which means

there is still network effects in the error term ($\rho = 0.0552^{***}$ with 600 observations; $\rho = .065^{*}$ with 300 observations). By considering the contextual effects, the autocorrelation coefficient is not significant any more in model (3b). However, model (3b) only takes the first 300 observations[7], further investigation is needed for using the large sample.

Table 5.3 gives the summary on selected estimated coefficients. As model (3b) takes more than half an hour to get the estimation in the standard laptop even with only 300 observations, it was dropped for estimating peer effects on different dependent variables. The full results for all estimated models can be found in Appendix C.

Table 5.3: Summary on selected estimated coefficients

Dependent		**Model**			
Variables	Coefficients	(2)MLE	(2) 2SLS(G)	(2) 2SLS($\hat{G}$)	(3a)
Math	ϕ	0.0449	0.0932	-0.0164	0.0021
	α	2.1300^{***}	2.4368^{***}	2.4939^{***}	1.8590^{***}
	Previous score	0.5027^{***}	0.4950^{***}	0.5149^{***}	0.5733^{***}
	IQ	-0.0058^{***}	-0.0060	-0.0063	-0.0064
	ρ				0.0675^{***}
English	ϕ	-0.1250	0.0862^{***}	0.0170	0.0086
	α	1.0980^{***}	1.0659	0.9671	1.6046^{*}
	Previous score	0.5252^{***}	0.5363^{***}	0.5393^{***}	0.5087^{***}
	IQ	-0.0104^{***}	-0.0112^{***}	-0.0110^{***}	-0.0100^{***}
	ρ				0.0263
German	ϕ	0.1273^{***}	0.1291^{***}	-0.0187	-0.0004
	α	0.4663^{*}	0.8343	0.9530	-0.1806
	Previous score	0.5072^{***}	0.5044^{*}	0.5050^{***}	0.5018^{***}
	IQ	-0.0023	-0.0029^{*}	-0.0029^{*}	-0.0042
	ρ				0.0444^{*}

Note: $*p < 0.1; **p < 0.05; ***p < 0.01$

model (3a) only uses the first 500 observations

5.2 Simulation on 2SLS

According to Startz and Wood-Doughty (2017), the weak instruments issue of estimating peer effects by 2SLS method will result in the inconsistent estimation. To check the consistency of estimations, the Monte Carlo simulations has been implemented for only 2SLS methods. In each round of simulation, 2SLS with original adjacency matrix and estimated adjacency matrix will be estimated. To keep the network integrated, each simulation takes 50 out of 101 networks in the dataset. This means the sample size varies each round, whereas, it has around 1100 observations [8] for every round of simulation.

[7]Running MLE for model (3b) takes more than 10 minutes even with 300 observations, and the running time increases exponentially when adding more observations (e.g., it takes more than 1 hour with 600 observations.

[8]The sample size of cleaned data is 2320, the summary of dataset can be found in Appendix B.

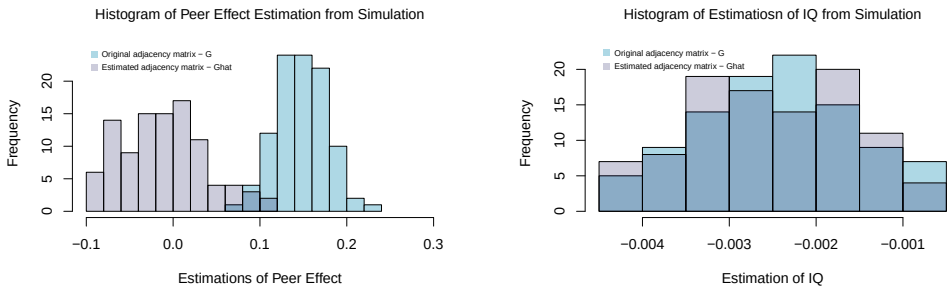

Figure 5.1: Selected Estimations from Simulation

Table 5.4: Rejection rate for selected estimations from 100 simulations

Estimations	2SLS(G)	2SLS($\hat{G}$)
ϕ	0.97	0.06
α	1.00	1.00
iq	0.65	0.61
Previous GPA	1.00	1.00

H_0 : Estimated coefficients = 0; rejection at $p < 0.1$

With 100 simulations, Figure 5.1 gives the distribution of peer effect and IQ effect estimations, and Table 5.4 gives the summary on rejecting the null hypothesis at 10% significant level. For estimation of peer effect, ϕ, the rejection rate of 2SLS with original adjacency matrix is around 97%, whereas the rejection rate of 2SLS with estimated adjacency matrix based on exponential random graph model is only 6%. In other words, the peer effect is estimated to be significant at 10% level with original adjacency level G in most cases, whereas it is estimated to be not significant at 10% level with estimated adjacency matrix $\hat{G}$ in most cases. For all other selected estimations (α, iq, previous GPA), the rejection rate is almost same for 2SLS with G and $\hat{G}$.

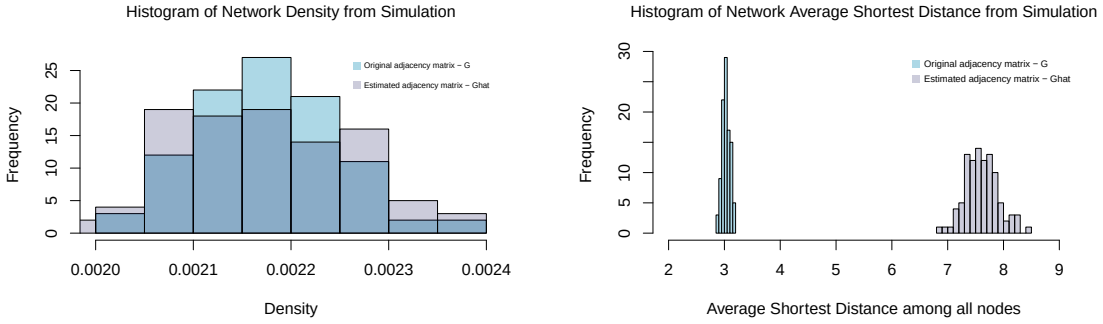

Figure 5.2: Summary on Networks from Simulation

Based on the summary from Figure 5.2, there is no significant difference for original adjacency matrix (G) and estimated adjacency matrix ($\hat{G}$) in terms of the network density distribution. However, when it comes to the network average shortest distance

distribution, the difference is significant. The average shortest distance among all nodes in estimated networks is much larger than that of original ones as there is no hubs from estimated networks[9]. Again the exponential random graph model could not generate hubs as it assumes the independence of edges (or links) formation.

Table 5.5: Peer Effect Correlation Matrix for selected Variables

	t-value(G)	t-value(Ghat)	Density(G)	Density(Ghat)	Distance(G)	Distance(Ghat)
t-value(G)	1	0.016	-0.272***	-0.222**	0.181*	0.102
t-value(Ghat)	0.016	1	-0.060	0.026	0.033	-0.009
Density(G)	-0.272***	-0.060	1	0.810	-0.338***	-0.487
Density(Ghat)	-0.222**	0.026	0.810	1	-0.240**	-0.720
Distance(G)	0.181*	0.033	-0.338***	-0.240**	1	0.127
Distance(Ghat)	0.102	-0.009	-0.487	-0.720	0.127	1

The $p-value$ for the correlation is in table C.15 of Appendix C; Distance is short for average shortest distance

The correlation matrix in Table 5.5 shows that there are significant correlations between the t ratio from 2SLS model with G and density and average shortest distance of original network G. However, for the 2SLS with $\hat{G}$, none of those correlation is significant. The correlation of t ratio from 2SLS(G) and density of network G is -0.272 (also significant), which implies that the size of the test against null hypothesis ($H_0 = 0$) declines with the increasing density of networks. This result is aligned with the finding by Startz and Wood-Doughty (2017).

In summary, the simulation from this section proves the existence of 'weak instruments' issue by comparing the estimations of peer effects with original network G and estimated $\hat{G}$. When the network does not have hubs, the mean of estimated peer effects ϕ is around zero. However, the estimations become significantly different from zero with the existence of hubs in network. Again the rejection rate presented in Table 5.4 shows the this dramatic change on estimated coefficients only happens to the peer effect estimation ϕ rather than others (such as α, etc.). Furthermore, the pattern of correlation between t ratios of estimated models and characteristics of network in Table 5.5 is also unique for ϕ, whereas, all other estimated coefficients don't have this pattern. For instance, same table for iq (in Table C.16, C.17) shows that there is no significant correlation between t ratio of 2SLS(G) and average shortest distance of network G.

5.3 Robustness Check

As so many models with different dependent variables have been estimated in this paper, it is not practical to do robustness check for all models. Simulation results from the last section have shown the inconsistent property of 2SLS. For the robustness checking, the directed network should be used to compare the results of selected models. Considering all estimations on main covariate variables except for peer effect ϕ show the super consistence, it is not unexpected that the estimations of those variables (such as previous score or IQ) with directed network are still super consistency. With the limited space, the estimations

[9]Think the hub and average shortest path like this way: if there is a hub in the network, and every node (player) is connected to this hub, then the shortest path for all nodes (players) is same, which is equal to 1. Then, average shortest path will also be 1.

with directed network are not reported in this essay as there is no insights that can be drawn from estimating all models with directed networks.

6 Discussions and Limitations

With fitting exponential random graph model, the primary tests in section 4 has shown that social network formation follows the preferential attachment mechanism. The evidence to support this claim includes that edged-wise shared partners are much higher in social networks in the dataset than the networks generated from exponential random graph model. Or the average shortest distance from social network in the dataset is much lower than the those generated from exponential random graph model. Because of this, the correlated effects in social network becomes relative large. It might be the case that the reflection problem in social network is more profound with the existence of hubs. This brings the challenge to identify and estimate the peer effects as the existence of hubs makes everyone's second-order friends become more similar.

The main limitations of this essay include but not limited to the followings. First, it didn't estimate spatial autocorrelation model with full sample size. Considering there is the strong connection between the existence of hubs and the existence spatial autocorrelation in the error terms, it is very valuable to use spatial autocorrelation model to investigate the identification and estimation of peer effects. Second, this paper didn't include the maximum likelihood estimations for running simulation. Future research might discover more by comparing results from MLE and 2SLS from Monte Carlo simulations. Third, this essay didn't apply Bayesian estimation on identifying and estimating peer effects. Further investigation based on Hsieh and Lee (2016) can be extended from this paper.

7 Conclusion

By analyzing and disentangling the social network formation, this essay shows the preferential attachment mechanism brings the challenge for identifying and estimating the peer effect in social network. Even though Calvó-Armengol et al. (2009) has proposed a structural model to identify the peer effects theoretically based on the key paper by Ballester et al. (2006), there is still a big leap to fill when it comes to empirical work. However, this does not mean 'much ado about nothing'.

A Derivation of Network Formation

A.1 Static Model

Consider a set of nodes $N = \{1, \cdots, n\}$, and let a link between any two nodes, i and j, be formed with probability p, where $0 < p < 1$. The formation of links is independent. For n nodes, the all possible combination (or links) are:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$$

Then, any given network that has m links on n nodes has a probability of

$$p^m(1-p)^{\frac{n(n-1)}{2}-m}$$

We can therefore write the probability that a particular realization of a random network has exactly L links as

$$p_L = \binom{\frac{N(N-1)}{2}}{L} p^L (1-p)^{N(N-1)/2-L} \tag{A.1}$$

Define the average degree of a random network as λ

$$\lambda = E[k] = p(N-1) \tag{A.2}$$

When $N \to \infty$, we can derive the *degree distribution of a random network* as the Poisson distribution

$$p_k = e^{-\lambda}\frac{\lambda^k}{k!} \tag{A.3}$$

For the detailed derivation of (A.1) to (A.3) , please refer to the textbook by Barabási et al. (2016).

A.2 Dynamic Model

Start with m_0 nodes, the links between which are chosen arbitrarily, as long as each node has at least one link. The network develops following two steps:

(a) Growth: at each timestep we add a new node with m ($\leq m_0$) links that connect the new node to m nodes already in the network.

(b) The probability $\Pi(k)$ that a link of the new node connects to node i depends on the degree k_i as

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{A.4}$$

Preferential attachment is a probabilistic mechanism: A new node is free to connect to any node in the network, whether it is a hub or has a single link. Equation (2.12) implies, however, that if a new nodes has a choice between a degree-two and a degree-four node, it is twice as likely that it connects to the degree-four node, which can provide more resources.

Two models extend Barabasi-Albert model by assuming different mechanism of the probability $\Pi(k)$: copying scale-free model and optimizing scale-free model. Copying scale-free model assumes that once we assign a node j from network to new node i, node i will copy the connection behaviours of node j and then connected them. This model can be interpreted as the mimic process. The optimizing scale-free model assumes that the new connected node i will chose the nodes to connect based on minimizing the certain cost function.

B Explore the Dataset

The original dataset I used is called `data_network2.csv`, which is cleaned and organized by Dr Livia Shkoza. The raw dataset has 3391 observations and 81 variables. The variable names and descriptions are listed in table A.1.

Table B.1: Dataset Descriptions

Variables	**Types**	**Descriptions**
`ques_1_1`	Indicator	Include the network relationships,
...		wherein each students list their
`ques_1_32`		friends based on ID numbers, such as $1, 2, 16, \cdots$
`id`	ID	Unique ID for each observation composed by interview wave, school, class and student IDs, e.g. 104102
`interview`	Indicator	Interview phase stage, all 1
`school`	ID	School ID, e.g. 13
`class`	Categorical	1 - class chosen based on school size 2 - class chosen based on school branch
`studentid`	ID	ID number represents each student, which is also used to indicate friendship
`classind`	ID	class ID number assigned within dataset
`classsize`	Discrete	Gives the size of each class
`idt`	ID	Tidy version of variable `id`
`st_V8`	Categorical	School branches, e.g. social science; 7 branches in total
`sch_score_68_3`	Discrete	Math grade in previous year (1968), 1 is the best
`sch_score_68_8`	Discrete	German grade in previous year (1968)
`sch_score_68_13`	Discrete	English grade in previous year (1968)
`sch_gpa_68`	Continuous	Average gpa in 1968, 1 is the highest
`sch_score_70_3`	Discrete	Math grade in 1970, 1 is the best
`sch_score_70_8`	Discrete	German grade in 1970
`sch_score_70_13`	Discrete	English grade in 1970
`sch_gpa_10`	Continuous	Average gpa in 1970, 1 is the highest
`age`	Discrete	Age of students
`iq`	Discrete	Proxy of student IQ
`edu_father`	Categorical	1-school, 2-university, 3-Traineeship, 5-employment, $\cdots$
`edu_mother`	Categorical	1-school, 2-university, 3-Traineeship, 5-employment, $\cdots$
`empmot`	Categorical	1, 2, 3, not clear on representation
`int_index`	Discrete	noncognitive skills, larger number is higher noncognitive ability
`int_5q`	Categorical	five quantiles by dividing `int_index`
`intdum1`	Dummy	Dummy variable to represent the categorical ones
...		
`intdum6`		
`iq_d1`	Dummy	Dummy variable to represent the categorical IQ
...		
`edu_mother1`	Dummy	Different dummy variables for giving parents' education
...		
`prestige`	Discrete	Standard International Occupational Score
`fem`	Dummy	Gender

Table B.2: Descriptive Statistics on Main Variable

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
sch_gpa_10	2,320	3.152	0.451	1.538	2.846	3.500	4.111
sch_score_70_3	2,320	3.416	0.739	1	3	4	5
sch_score_70_8	2,320	3.575	0.809	1	3	4	6
sch_score_70_13	2,320	3.433	0.924	1	3	4	6
sch_gpa_68	2,320	3.189	0.493	1.400	2.857	3.545	4.667
sch_score_68_3	2,320	3.452	0.759	1	3	4	5
sch_score_68_8	2,320	3.560	0.862	1	3	4	6
sch_score_68_13	2,320	3.470	0.935	1	3	4	6
fem	2,320	0.471	0.499	0	0	1	1
age	2,320	15.378	0.871	13	15	16	18
iq	2,320	40.308	9.147	9	34	46	68
prestige	2,320	48.154	13.091	19	39	58	82
edu_motherd2	2,320	0.251	0.434	0	0	1	1
edu_fatherd2	2,320	0.212	0.409	0	0	0	1
gpa_diff	2,320	−0.037	0.337	−1.650	−0.239	0.182	0.872

C Figures and Tables

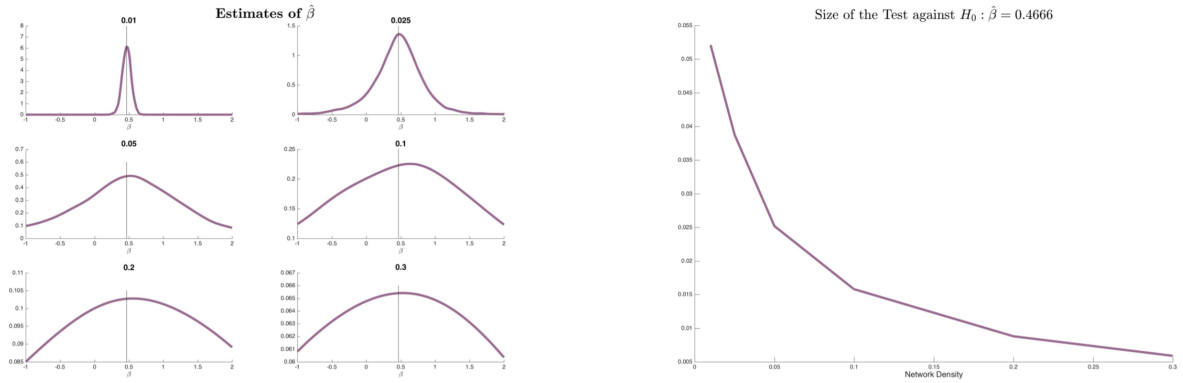

Figure C.1: 2SLS Simulations with Different Network Densities (Startz and Wood-Doughty, 2017)

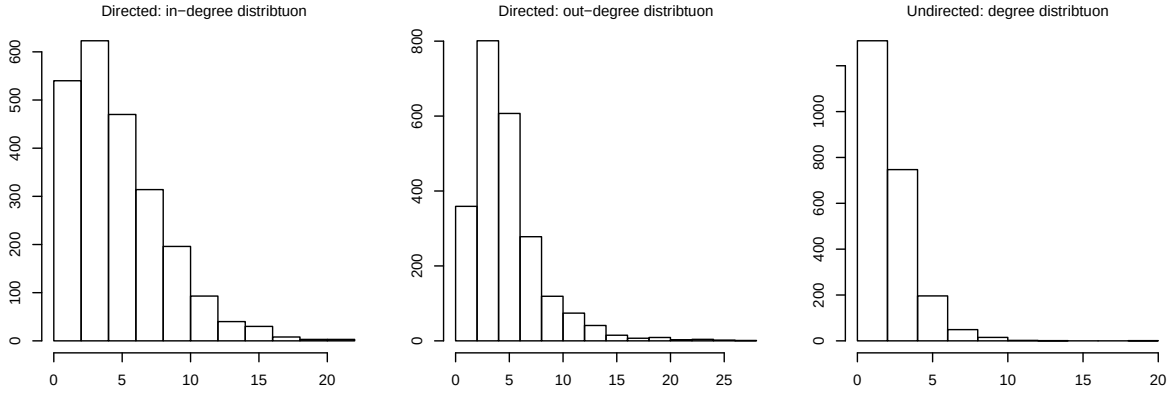

Figure C.2: Degree Distribution for the Network

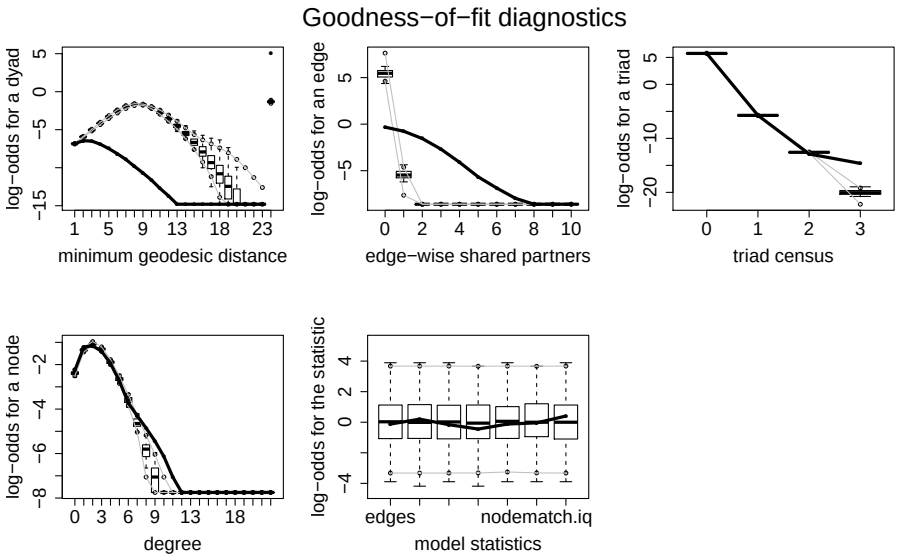

Figure C.3: Diagnostics for ERGM1 (undirected)

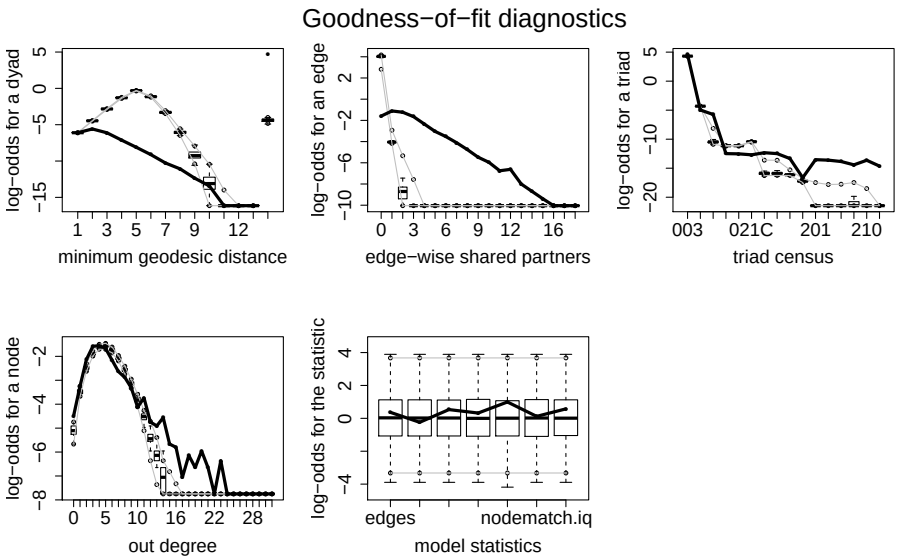

Figure C.4: Diagnostics for ERGM1 (Directed)

Table C.1: ERGM fit for Mode 1b

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0001	0.00003	-4.374	0.00001
exclusion_restriction	0.481	0.0002	2,226.615	0
sch_gpa_68	0.0001	0.0001	1.830	0.067
fem	0.0002	0.00002	8.357	0
age	0.0001	0.00002	3.325	0.001
iq	-0.0001	0.0001	-1.259	0.208
prestige	0.0002	0.00004	3.582	0.0003
edu_motherd2	-0.00004	0.00002	-1.723	0.085
edu_fatherd2	0.00002	0.00002	0.949	0.342

Table C.2: ERGM for Model 2 (MLE)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0001	0.00003	-2.2261	0.0260
exclusion_restriction	0.4805	0.0002	2,226.6080	0
sch_gpa_68	0.0001	0.0001	1.8629	0.0625
fem	0.0001	0.00002	6.3350	0
age	0.0001	0.00002	3.2472	0.0012
iq	-0.0001	0.0001	-1.2622	0.2069
prestige	0.0002	0.00004	3.5531	0.0004
edu_motherd2	-0.00004	0.00002	-1.7938	0.0728
edu_fatherd2	0.00002	0.00002	1.0846	0.2781
Gsch_gpa_68	-0.0009	0.0002	-4.9304	0.000001
Gfem	0.0001	0.00003	2.1598	0.0308
Gage	0.0003	0.0001	4.9942	0.000001
Giq	-0.0004	0.0001	-2.8166	0.0049
Gprestige	0.0001	0.0001	0.6219	0.5340
Gedu_motherd2	-0.0001	0.00002	-3.2631	0.0011
Gedu_fatherd2	-0.0001	0.00002	-2.8967	0.0038

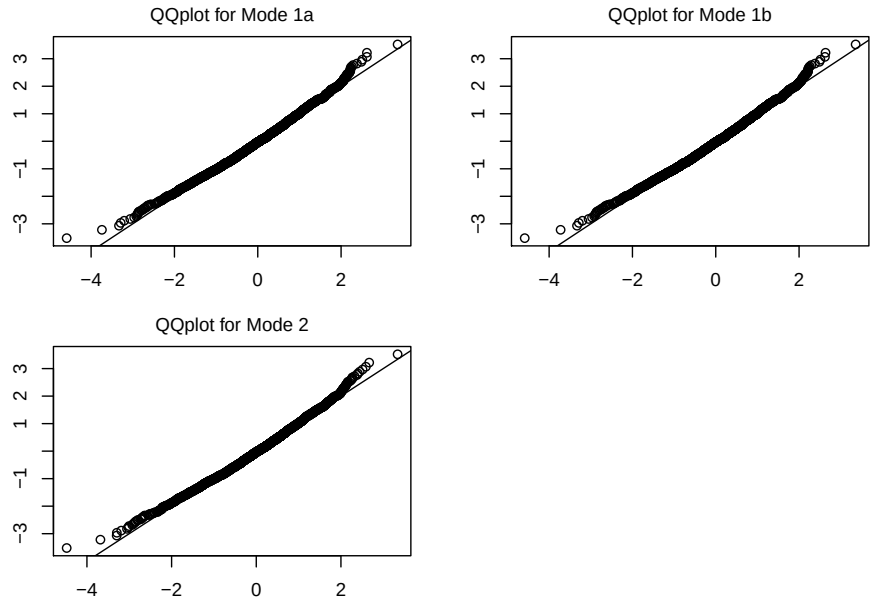

Figure C.5: Residual diagnostics on estimations in table 5.2

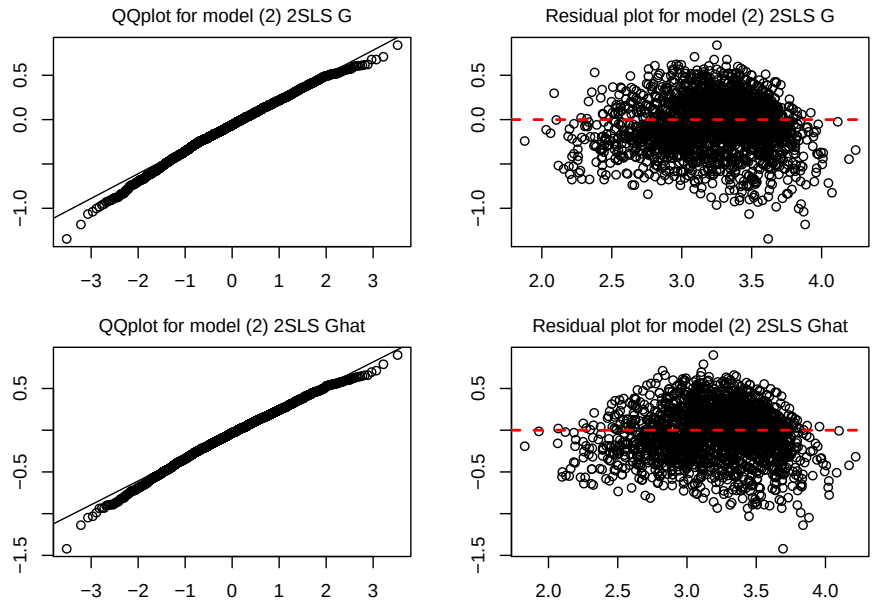

Figure C.6: Residual diagnostics on estimations in table 5.2

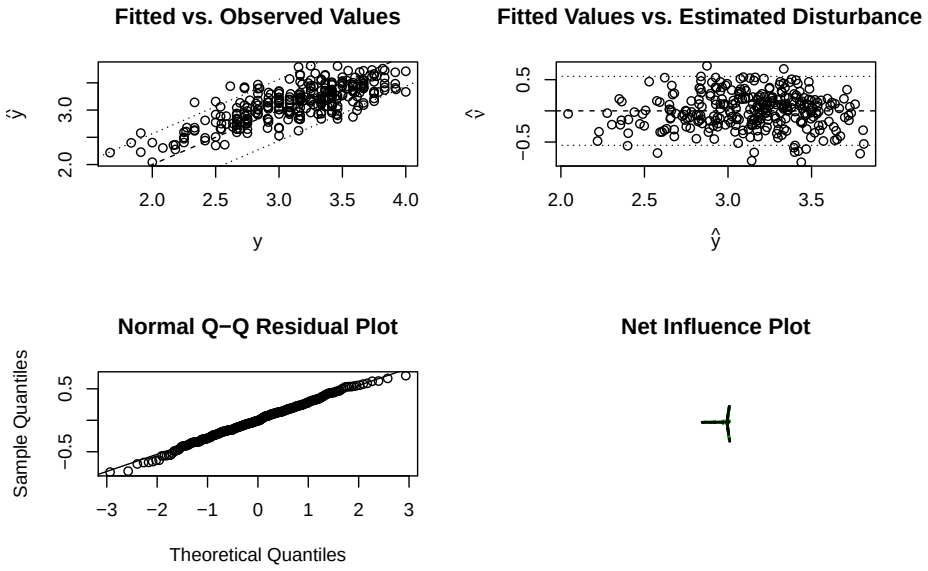

Figure C.7: Residual diagnostics on estimations in table 5.2: model (3a)

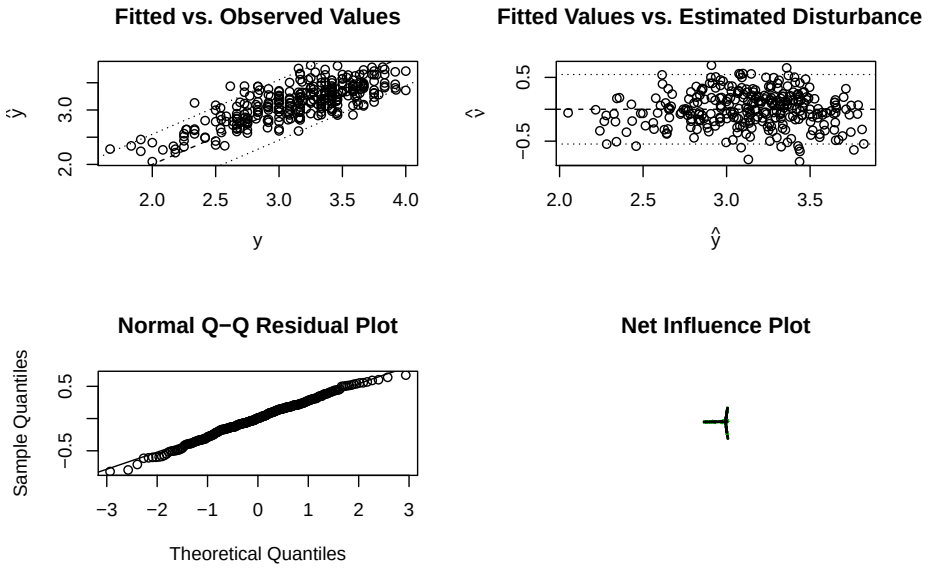

Figure C.8: Residual diagnostics on estimations in table 5.2: model (3b)

Table C.3: MLE Peer Effect in Math

	Estimate	Std. Error	t value	Pr(>|t|)
alpha	2.1312	0.2647	8.0522	0
beta_sch_score_68_3	0.5027	0.0184	27.2611	0
beta_fem	-0.0606	0.0396	-1.5280	0.1267
beta_age	-0.0079	0.0156	-0.5017	0.6159
beta_iq	-0.0058	0.0015	-3.9208	0.0001
beta_prestige	-0.0008	0.0010	-0.7941	0.4272
beta_edu_motherd2	-0.0067	0.0296	-0.2258	0.8214
beta_edu_fatherd2	0.0018	0.0317	0.0571	0.9545
beta_Gsch_score_68_3	0.0235	0.0314	0.7464	0.4555
beta_Gfem	-0.00001	0.0129	-0.0010	0.9992
beta_Gage	-0.0152	0.0047	-3.2052	0.0014
beta_Giq	0.0003	0.0007	0.4000	0.6892
beta_Gprestige	-0.0004	0.0005	-0.8158	0.4147
beta_Gedu_motherd2	0.0217	0.0178	1.2171	0.2237
beta_Gedu_fatherd2	-0.0307	0.0181	-1.6917	0.0908
phi	0.0449	0.0464	0.9685	0.3329
beta_unobservables	-0.0057	0.0080	-0.7189	0.4723

Table C.4: MLE Peer Effect in English

	Estimate	Std. Error	t value	Pr(>|t|)
alpha	1.0981	0.3213	3.4172	0.0006
beta_sch_score_68_13	0.5252	0.0176	29.8013	0
beta_fem	0.0602	0.0489	1.2311	0.2184
beta_age	0.0540	0.0186	2.9091	0.0037
beta_iq	-0.0104	0.0019	-5.5152	0.000000
beta_prestige	0.0004	0.0012	0.2988	0.7651
beta_edu_motherd2	0.0501	0.0380	1.3194	0.1872
beta_edu_fatherd2	0.0966	0.0394	2.4539	0.0142
beta_Gsch_score_68_13	0.0662	0.0588	1.1256	0.2604
beta_Gfem	0.0351	0.0239	1.4662	0.1427
beta_Gage	0.0138	0.0131	1.0559	0.2911
beta_Giq	-0.0002	0.0014	-0.1367	0.8913
beta_Gprestige	-0.0005	0.0007	-0.7577	0.4487
beta_Gedu_motherd2	0.0358	0.0240	1.4928	0.1356
beta_Gedu_fatherd2	0.0211	0.0264	0.7970	0.4256
phi	-0.1251	0.1087	-1.1510	0.2498
beta_unobservables	0.0105	0.0094	1.1156	0.2647

Table C.5: MLE Peer Effect in German

	Estimate	Std. Error	t value	Pr(>|t|)
alpha	0.4663	0.2423	1.9242	0.0544
beta_sch_score_68_8	0.5072	0.0173	29.3960	0
beta_fem	-0.1137	0.0394	-2.8856	0.0039
beta_age	0.0920	0.0149	6.1865	0
beta_iq	-0.0023	0.0016	-1.4583	0.1449
beta_prestige	0.0007	0.0011	0.5858	0.5581
beta_edu_motherd2	0.0610	0.0325	1.8758	0.0608
beta_edu_fatherd2	-0.0055	0.0350	-0.1557	0.8763
beta_Gsch_score_68_8	-0.0731	0.0150	-4.8644	0.000001
beta_Gfem	0.0223	0.0103	2.1724	0.0299
beta_Gage	-0.0101	0.0051	-1.9798	0.0478
beta_Giq	-0.0002	0.0006	-0.3273	0.7435
beta_Gprestige	-0.0009	0.0005	-2.0290	0.0426
beta_Gedu_motherd2	0.0199	0.0176	1.1263	0.2602
beta_Gedu_fatherd2	-0.0298	0.0167	-1.7839	0.0746
phi	0.1273	0.0285	4.4701	0.00001
beta_unobservables	0.0085	0.0080	1.0625	0.2881

Table C.6: 2SLS Peer Effect in Math (G)

	Estimate	stderror	t ratio
ϕ	0.0932	0.1502	0.6204
sch_score_68_3	0.4950	0.1095	4.5189
fem	-0.0523	0.3257	-0.1606
age	-0.0249	0.0940	-0.2652
iq	-0.0060	0.0092	-0.6555
prestige	-0.0010	0.0062	-0.1540
edu_motherd2	-0.0060	0.1776	-0.0339
edu_fatherd2	0.0122	0.1891	0.0643
sch_score_68_3.1	-0.0109	0.1153	-0.0949
fem.1	-0.0121	0.0942	-0.1280
age.1	-0.0184	0.0266	-0.6909
iq.1	0.0005	0.0054	0.0891
prestige.1	-0.0005	0.0035	-0.1305
edu_motherd2.1	0.0195	0.1195	0.1629
edu_fatherd2.1	-0.0194	0.1228	-0.1577
α	2.4368	0.5966	4.0848

Table C.7: 2SLS Peer Effect in Math ($\hat{G}$)

	Estimate	stderror	t ratio
ϕ	-0.0164	0.1958	-0.0836
sch_score_68_3	0.5149	0.1137	4.5296
fem	-0.0987	0.3935	-0.2508
age	-0.0318	0.0994	-0.3195
iq	-0.0063	0.0097	-0.6525
prestige	-0.0011	0.0067	-0.1676
edu_motherd2	-0.0078	0.1878	-0.0413
edu_fatherd2	0.0070	0.2003	0.0351
sch_score_68_3.1	0.0147	0.1289	0.1141
fem.1	0.0030	0.1050	0.0289
age.1	-0.0041	0.0316	-0.1312
iq.1	0.0003	0.0053	0.0627
prestige.1	0.0011	0.0038	0.3013
edu_motherd2.1	-0.0071	0.1185	-0.0598
edu_fatherd2.1	-0.0162	0.1236	-0.1311
α	2.4939	0.6099	4.0890

Table C.8: 2SLS Peer Effect in English (G)

	Estimate	stderror	t ratio
ϕ	0.0862	0.0302	2.8550
sch_score_68_13	0.5363	0.0187	28.7188
fem	0.0382	0.0672	0.5686
age	0.0531	0.0197	2.6927
iq	-0.0112	0.0020	-5.6612
prestige	0.0014	0.0013	1.0993
edu_motherd2	0.0249	0.0372	0.6688
edu_fatherd2	0.0874	0.0396	2.2076
sch_score_68_13.1	-0.0383	0.0210	-1.8184
fem.1	-0.0029	0.0203	-0.1439
age.1	-0.0100	0.0056	-1.7949
iq.1	0.0007	0.0012	0.6095
prestige.1	-0.0006	0.0008	-0.7625
edu_motherd2.1	0.0112	0.0250	0.4473
edu_fatherd2.1	-0.0048	0.0259	-0.1853
α	1.0659	0.7416	1.4375

Table C.9: 2SLS Peer Effect in English ($\hat{G}$)

	Estimate	stderror	t ratio
ϕ	0.0170	0.0321	0.5304
sch_score_68_13	0.5393	0.0187	28.8553
fem	0.0993	0.0755	1.3144
age	0.0557	0.0198	2.8182
iq	-0.0110	0.0020	-5.5332
prestige	0.0014	0.0013	1.0696
edu_motherd2	0.0286	0.0373	0.7659
edu_fatherd2	0.0921	0.0398	2.3128
sch_score_68_13.1	-0.0033	0.0203	-0.1609
fem.1	-0.0236	0.0214	-1.1042
age.1	-0.0048	0.0061	-0.7896
iq.1	0.0004	0.0012	0.3209
prestige.1	0.0008	0.0008	1.1127
edu_motherd2.1	0.0054	0.0239	0.2279
edu_fatherd2.1	-0.0179	0.0250	-0.7181
α	0.9671	0.7484	1.2921

Table C.10: 2SLS Peer Effect in German (G)

	Estimate	stderror	t ratio
ϕ	0.1291	0.0310	4.1666
sch_score_68_8	0.5044	0.0177	28.4568
fem	-0.1539	0.0616	-2.5004
age	0.0702	0.0179	3.9261
iq	-0.0029	0.0017	-1.7186
prestige	0.0013	0.0012	1.1020
edu_motherd2	0.0659	0.0335	1.9657
edu_fatherd2	0.0092	0.0355	0.2583
sch_score_68_8.1	-0.0538	0.0179	-2.9966
fem.1	0.0062	0.0177	0.3511
age.1	-0.0189	0.0059	-3.2007
iq.1	0.0007	0.0010	0.6692
prestige.1	-0.0004	0.0007	-0.6549
edu_motherd2.1	0.0104	0.0228	0.4577
edu_fatherd2.1	-0.0021	0.0232	-0.0902
α	0.8343	0.6544	1.2749

Table C.11: 2SLS Peer Effect in German ($\hat{G}$)

	Estimate	stderror	t ratio
ϕ	-0.0187	0.0294	-0.6368
sch_score_68_8	0.5050	0.0177	28.5642
fem	-0.1958	0.0687	-2.8496
age	0.0623	0.0178	3.5038
iq	-0.0029	0.0017	-1.7342
prestige	0.0011	0.0012	0.9785
edu_motherd2	0.0712	0.0332	2.1467
edu_fatherd2	0.0012	0.0354	0.0342
sch_score_68_8.1	-0.0025	0.0189	-0.1302
fem.1	-0.0005	0.0185	-0.0286
age.1	0.0096	0.0054	1.7683
iq.1	-0.0010	0.0010	-0.9886
prestige.1	-0.0007	0.0007	-0.9832
edu_motherd2.1	-0.0112	0.0218	-0.5135
edu_fatherd2.1	0.0098	0.0222	0.4395
α	0.9530	0.6666	1.4297

Table C.12: spatial autocorrelation Peer Effect in Math

	Estimate	Std.Error	Z value
sch_score_68_3	0.5733	0.0359	15.9735
fem	-0.1214	0.0647	-1.8751
age	-0.0064	0.0316	-0.2039
iq	-0.0064	0.0031	-2.0992
prestige	0.0001	0.0020	0.0285
edu_motherd2	0.1003	0.0590	1.6986
edu_fatherd2	-0.0114	0.0636	-0.1794
fix_effect	1.8594	0.5530	3.3625
ϕ	0.0022	0.0052	0.4118
ρ	0.0675	0.0191	3.5267

Table C.13: spatial autocorrelation Peer Effect in English

	Estimate	Std.Error	Z value
sch_score_68_13	0.5087	0.0360	14.1236
fem	0.1539	0.0705	2.1816
age	0.0194	0.0379	0.5127
iq	-0.0100	0.0038	-2.6358
prestige	-0.0009	0.0024	-0.3761
edu_motherd2	0.0772	0.0729	1.0588
edu_fatherd2	0.0652	0.0784	0.8317
fix_effect	1.6046	0.6541	2.4533
ϕ	0.0086	0.0054	1.5932
ρ	0.0263	0.0206	1.2767

Table C.14: spatial autocorrelation Peer Effect in English

	Estimate	Std.Error	Z value
sch_score_68_8	0.5018	0.0368	13.6226
fem	-0.0569	0.0685	-0.8312
age	0.1260	0.0355	3.5458
iq	0.0042	0.0035	1.2047
prestige	-0.0022	0.0022	-0.9790
edu_motherd2	0.1496	0.0676	2.2147
edu_fatherd2	-0.0438	0.0729	-0.6009
fix_effect	-0.1806	0.6124	-0.2949
ϕ	-0.0004	0.0053	-0.0820
ρ	0.0444	0.0198	2.2393

Table C.15: P-value for Correlation Matrix of selected variables: Peer effect ϕ

	t-value(G)	t-value(Ghat)	Density(G)	Density(Ghat)	Distance(G)	Distance(Ghat)
t-value(G)		0.878	0.006	0.027	0.072	0.311
t-value(Ghat)	0.878		0.554	0.794	0.744	0.931
Density(G)	0.006	0.554		0	0.001	0
Density(Ghat)	0.027	0.794	0		0.016	0
Distance(G)	0.072	0.744	0.001	0.016		0.208
Distance(Ghat)	0.311	0.931	0	0	0.208	

Table C.16: Correlation Matrix of selected variables: iq

	t-value(G)	t-value(Ghat)	Density(G)	Density(Ghat)	distance(G)	distance(Ghat)
t-value(G)	1	0.925	-0.190	-0.099	0.022	0.092
t-value(Ghat)	0.925	1	-0.178	-0.054	-0.0003	-0.016
Density(G)	-0.190	-0.178	1	0.810	-0.338	-0.487
Density(Ghat)	-0.099	-0.054	0.810	1	-0.240	-0.720
distance(G)	0.022	-0.0003	-0.338	-0.240	1	0.127
distance(Ghat)	0.092	-0.016	-0.487	-0.720	0.127	1

Table C.17: P-value for Correlation Matrix of selected variables: iq

	t-value(G)	t-value(Ghat)	Density(G)	Density(Ghat)	distance(G)	distance(Ghat)
t-value(G)		0	0.059	0.325	0.827	0.363
t-value(Ghat)	0		0.077	0.592	0.998	0.872
Density(G)	0.059	0.077		0	0.001	0
Density(Ghat)	0.325	0.592	0		0.016	0
distance(G)	0.827	0.998	0.001	0.016		0.208
distance(Ghat)	0.363	0.872	0	0	0.208	

D Exponential Random Graph Model

The exponential random graph model (ERGM) has the distribution over networks

$$p(Y = y|\theta) = \frac{1}{Z} e^{\theta^T} \phi(y) \tag{D.1}$$

, where y is observed network adjacency matrix, $\phi(y)$ features the properties of the network and θ are parameters to be estimated, Z is normalized as the constant $\sum_y e^{\theta^t} \phi(y)$. Each component of the θ vector may be interpreted as the increase in the conditional log-odds of the network, per unit increase in the corresponding component of $\phi(y)$. Edges in the network are considered conditionally independent if they don't share a node.

E Instruction on Source Code

All data cleaning and model estimations are implemented in R platform. The source code file `PeerEffect_942870.zip` includes:

- One instruction file: `readmefirst.pdf`, which is same with this instruction.
- One original dataset cleaned by Dr. Livia Shkoza: `data_network2.csv`
- Eight R programming scripts
- Seven Rdata files

Table E.1 gives the instructions on eight R programming scripts

File	Descriptions	Instruction
`PeerEffect_fun.R`	Self-written functions for this project	Instructions can be found in the file
`PeerEffect_matrix.R`	Clean the dataset again by removing NAs; Construct the block matrix for directed and undirected network; Primary test by fitting ERGM models	In the end, it will produce main dataset and save as: kgpclean.Rdata kgpcleanmat.Rdata kgpcleanmat_undirt.Rdata kgpcleanblock.Rdata kgpcleanblock_undirt.Rdata need load `PeerEffect_fun.R`
kgpclean.Rdata: main dataset includes all network information and all covariate variables kgpcleanmat.Rdata: a list including 101 directed network adjacency matrix kgpcleanmat_undirt.Rdata: a list including 101 undirected network adjacency matrix kgpcleanblock.Rdata: a block directed adjacency matrix kgpcleanblock_undirt.Rdata: a block undirected adjacency matrix or matrix G in the model		
`PeerEffect_model1.R`	Code for model 1	Need load the above Rdatas and `PeerEffect_fun.R`
`PeerEffect_model2mle.R`	Code for model 2(MLE)	Need load the above Rdatas and `PeerEffect_fun.R`
`PeerEffect_model2sls.R`	Code for model 2(2SLS)	Need load the above Rdatas and `PeerEffect_fun.R`
`PeerEffect_model3.R`	Code for model 3	Need load the above Rdatas and `PeerEffect_fun.R`
`PeerEffect_simulation.R`	Code for simulation on 2SLS	Need load the above Rdatas and `PeerEffect_fun.R` It will save results simresults.Rdata simdensity.Rdata
`PeerEffect_plot.R`	Company code for plotting the figures	Need run with different models, in has same figure indexes with this paper

References

Albert, R. and A.-L. Barabási (2002): "Statistical mechanics of complex networks," *Reviews of modern physics*, 74, 47.

Ballester, C., A. Calvó-Armengol, and Y. Zenou (2006): "Who's who in networks. Wanted: The key player," *Econometrica*, 74, 1403–1417.

Barabási, A.-L. et al. (2016): *Network science*, Cambridge university press.

Bonacich, P. (1987): "Power and centrality: A family of measures," *American journal of sociology*, 92, 1170–1182.

Bramoullé, Y., H. Djebbari, and B. Fortin (2009): "Identification of peer effects through social networks," *Journal of econometrics*, 150, 41–55.

Calvó-Armengol, A., E. Patacchini, and Y. Zenou (2009): "Peer effects and social networks in education," *The Review of Economic Studies*, 76, 1239–1267.

Hsieh, C.-S. and L. F. Lee (2016): "A social interactions model with endogenous friendship formation and selectivity," *Journal of Applied Econometrics*, 31, 301–319.

Hummell, Hans J.; Klein, M. W.-M. M. Z. R. (1970): "Structure Analysis of the School (Schoolchildren Survey)," Data file, Version 1.0.0, `http://dx.doi.org/10.4232/1.0600`.

Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris (2008): "ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks," *Journal of Statistical Software*, 24, 1–29.

Jackson, M. O. (2010): *Social and economic networks*, Princeton university press.

Katz, L. (1953): "A new status index derived from sociometric analysis," *Psychometrika*, 18, 39–43.

Manski, C. F. (1993): "Identification of endogenous social effects: The reflection problem," *The review of economic studies*, 60, 531–542.

Sacerdote, B. (2001): "Peer effects with random assignment: Results for Dartmouth roommates," *The Quarterly journal of economics*, 116, 681–704.

Startz, R. and A. Wood-Doughty (2017): "Improved Estimation of Peer Effects using Network data," Unpublished, Available at `http://econ.ucsb.edu/~startz/Improved%20Estimation%20of%20Peer%20Effects%20Using%20Network%20Data.pdf`.