CHAPTER 3

SNOWBALL SAMPLING AND SAMPLE SELECTION IN A SOCIAL NETWORK

Julian TszKin Chan

ABSTRACT

This chapter studies a snowball sampling method for social networks with endogenous peer selection. Snowball sampling is a sampling design which preserves the dependence structure of the network. It sequentially collects the information of vertices linked to the vertices collected in the previous iteration. The snowball samples suffer from a sample selection problem because of the endogenous peer selection. The author proposes a new estimation method that uses the relationship between samples in different iterations to correct selection. The author uses the snowball samples collected from Facebook to estimate the proportion of users who support the Umbrella Movement *in Hong Kong.*

Keywords: Social network; snowball sampling; sample selection; generalized method of moment; six degrees of separation; Umbrella Movement

1 INTRODUCTION

Snowball sampling is a network sampling design that preserves the information of the network structure (Kolaczyk, 2009). It is an iterative procedure of collecting vertices' information that is linked with vertices collected in the previous iteration. It has several advantages over the simple random sampling. First, all the information of the network structure is preserved because

The Econometrics of Networks
Advances in Econometrics, Volume 42, 61–80

ISSN: 0731-9053/doi:10.1108/S0731-905320200000042008

it collects samples through the network connections from previous samples. This is essential in obtaining unbiased estimates for the network structure, such as the average number of friends (also known as the average degree of a vertex). Simple random sampling leads to measurement error and bias in estimates as network structure is not preserved (Chandrasekhar & Lewis, 2011). Second, snowball sampling is a cost-efficient method for collecting samples. Researchers can quickly obtain large amount of data by collecting information from the vertices linked with the samples in the previous iteration. For example, researchers can collect data from neighbors, friends, or family in the initial samples. In some cases, collecting data through snowball sampling would be easier than through traditional random sampling. For example, if researchers want to collect information from Facebook or Twitter, snowball sampling provides a natural way to collect data through the connections of the agents in the social network.

However, snowball samples are subject to sample selection problem because people are more likely to associate with others like themselves. This is a common phenomenon in social network called the homophily principle of social network (McPherson, Smith-Lovin, & Cook, 2001). For example, male students could be more likely to have male friends than female friends or people with similar political views are more likely to be friends. As a result, samples collected with snowball sampling are correlated with the previous samples and the initial samples determine the distribution of the rest of the samples. Therefore, if the initial samples do not come from a random sample or if the sample size is small, the sample selection problem could be severe and the estimates using snowball sampling could be biased. Although the bias would be reduced as the number of iterations of the snowball sampling increases, the number of iterations rarely exceeds 5. If the number of iterations is larger than 5, it is likely the researchers collect data from the whole population. This is known as the six-degree separation theory or the small world problem (Gurevitch, 1961; Milgram, 1967). The theory states that on average, the friendship distance between two individuals is about 6. The famous postcard experiment by Travers and Milgram (1969) showed that the average distance between two individuals is about 5.7. More recent studies (Backstrom, Boldi, Rosa, Ugander, & Vigna, 2012; Ugander, Karrer, Backstrom, & Marlow, 2011) indicated that the average distance between Facebook users in May 2011 was 4.7 and the average distance between individuals in the United States at the same time was 4.3.

Here is an example that illustrates the sample selection problem of snowball sampling. Suppose we are interested in the proportion of the types of agents. Let $Y_i \in \{0, 1\}$ be the type of agent i. The objective is to identify $P_1 \equiv \mathbb{P}(Y_i = 1)$. Let the proportion of agents with type 0 be 0.5; that is, $P_0 = 0.5$. Researchers have an initial sample of 10 agents, 7 of whom are type 0. Agents with the same type are more likely to be connected. Suppose, on average, 8 of 10 friends of an agent are the same type as the agent. When researchers start to collect more samples using the snowball sampling method, we expect the proportion of different types of agents to be correlated with the initial samples. In this example, the expected proportion of different types would be

$$\begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}\begin{pmatrix} 7 \\ 3 \end{pmatrix}=\begin{pmatrix} 6.2 \\ 3.8 \end{pmatrix}\neq\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}. \tag{1}$$

Therefore, if the initial samples were not drawn from random samples, then the estimates from the snowball samples suffer from the same sample selection problem as in the initial samples.

As the number of iterations of the snowball samples goes to infinity, the proportion of types will converge to $(0.5, 0.5)$[1]:

$$\lim_{n\to\infty}\begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}^{n}=\begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{pmatrix}. \tag{2}$$

As mentioned earlier, this result is not feasible in the data collection process because of the small world theory.

In this chapter, we propose a new estimation method that uses the relationship between the samples in different iterations to correct for the sample selection problem. Although the snowball samples are subject to sample selection bias, the information about the network connections is not. The average number of type 0 friends of an agent of type 1 can be consistently estimated despite the proportion of the types of agents selected in the initial samples. We could use this information to construct consistent estimates of the proportion of the types of agents. Further, we could determine a weight for the samples based on the estimates of the proportion of the types of agents to adjust for the sample selection problem for other statistical models such as regression.

The proposed method relies on two observations. First, the adjacency matrix of an undirected graph is symmetric. When we observe a type 1 friend of a type 0 agent, there must be a corresponding type 1 agent who is a friend of the same type 0 agent. Therefore, the total number of links from the type 0 to the type 1 agents and links between type 0 and type 1 agents must be the same in the population.

Second, we can consistently estimate the average number of friends of an agent given her type. This is possible because the snowball sampling method preserves the dependence structure of the network; hence, we can consistently estimate the average number of friends. This is an important feature of the snowball sampling method. Although the snowball samples are subject to sample selection, the sample selection does not play a role in the estimation of the number of friends given the types of an agent. In addition, we are considering the average number of friends of an agent given their types. The initial proportion of the types of agents would not affect this conditional expectation. Using these two observations, we derived the moment equations for the model, estimate the proportion of the types of agents by generalized method of moment and derive the asymptotic distribution of the estimator.

As an empirical application, we collect snowball samples from Facebook to estimate the proportion of users who support the *Umbrella Movement* in Hong Kong in 2014. The Facebook users changed their profile pictures to yellow ribbons to show their support for the movement and blue ribbons to show their

support for the government and the police. We find that the sample proportion is underestimated by 40% of the proportion of users who changed their profile pictures to blue ribbons.

We define the snowball sampling method and discuss the setup of our model in Section 2. Then we will discuss the estimation method and the statistical properties in Section 3. We conduct some simulation experiments to examine the statistical properties of the proposed estimator in Section 4. In Section 5, we describe the empirical application using the Facebook data we collected. Finally, Section 6 summarizes the findings.

2 SETUP

We follow the definition of snowball sampling method in Kolaczyk (2009) and consider only undirected graphs in this chapter. Let $G = (V, E)$ be the population graph, where V is the set of all vertices and E is the set of all edges. An edge is a pair of vertices $\{i, j\}$, where i and j are connected. Let V_0 be set of vertices in the initial samples and C_i be a set of vertices connected to i; that is, $C_i = \{j: \{i, j\} \in E\}$. Let N be the sample size of the initial sample. The initial sample may not draw from a simple random sample.

In the first iteration, we collect the information of all the vertices connected with the vertices in the initial sample that are not included in the initial samples. The set of vertices collected in the first iteration is

$$V_1 = \left(\cup_{i \in V_0} C_i\right) \setminus V_0. \tag{3}$$

The second iteration follows the same procedure but we are collecting all the vertices connected with vertices in V_1 except those in V_0 and V_1.

Each of the vertices $i \in \cup_{t=0,\ldots,T} V_t$ has a variable y that takes a discrete value ($y \in \{0, 1\}$), where T is the number of iterations of the snowball samples.

Our objective is to estimate $\mathbb{P}(y = 0)$ or construct a weight w_i such that $\sum_i w_i 1(y_i = 0)$ consistently estimates $\mathbb{P}(y = 0)$. The information of the edges is represented by an adjacency matrix A. The $\{i, j\}$ element of A is equal to 1 if i and j are connected, otherwise 0. Once we have an estimate, $\hat{P}_0$, it is easy obtain $w_i = \frac{\hat{P}_0}{\sum_j 1(y_j = 0)}(1 - y_i) + \frac{1 - \hat{P}_0}{\sum_j 1(y_j = 1)} y_i$.

In this setup, we assume T to be small and consider large N asymptotic. In fact, we can have a reasonable sample size even if N is small. The actual sample size is roughly Nd^T, where d is the average number of connections. The actual sample size increases exponentially by T. For example, suppose the average number of connections is 10. When $N = 5$ and $T = 1$, the actual sample size is 50. When $T = 2$, the actual sample size is 500. In our empirical example, the average number of connections of a Facebook user is over 250.

We do not assume the initial samples are random and start with a small initial sample that is subject to sample selection and collect data using the snowball sampling.

3 ESTIMATION

The estimation strategy is based on two observations on the adjacency matrix of an undirected graph and the snowball sampling.

First, the adjacency matrix of an undirected graph is symmetric. When we observe a type 1 friend from a type 0 agent, there must be a corresponding type 1 agent who is a friend of the same type 0 agent. Therefore, the total number of links between the type 0 and the type 1 agents and links from type 0 to type 1 agents must be the same in the population.

The second observation is that the characteristics related to the social network of an agent conditional on the type of the agent are not subject to the sample selection problem caused by the initial sample.

When the initial sample is subject to sample selection, the sample average of the types of the snowball samples is biased and inconsistent. However, the initial proportion of the types does not play a role in the characteristics related to social network of an agent given her type. We can consistently estimate the average number of friends of an agent given her type because the snowball sampling preserves the dependence structure of the network; hence, we can consistently estimate the average number of friends.

As long as we are estimating statistics that are conditional on the type, the proportion of types in the initial sample does not play a role. For example, the consistency of the number of connections of a type 0 or type 1 agent does not depend on the number of observations of type 0 and type 1 agents.[2] Therefore, we can consistently estimate the expected number of different types of friends conditional on the type of an agent.

The first observation provides the basis for the moment equations, and the second observation provides a means to estimate the moment equations. We will discuss the estimation method by introducing the moment equations of the population. For simplicity, we start with two types of agents.

3.1 Moment Equation

The number of type $a \in \{0, 1\}$ of agent i is defined as

$$d_{i,a} \equiv \sum_{j \in C_i} 1\big(y_j = a\big). \tag{4}$$

The average number of type a friends of a type b agent is

$$\begin{aligned} d_{a|b} &\equiv \mathbb{E}\big(d_{i,a} \mid y_i = b\big) \\ &= \mathbb{P}\big(y_i = b\big)^{-1} \mathbb{E}\big(d_{i,a} \times 1\big(y_i = b\big)\big). \end{aligned} \tag{5}$$

Because the graph is undirected, we have $j \in C_i$ if and only if $i \in C_j$ for all $i \neq j \in V$. For every type a friend of a type b agent, there must be a corresponding type b friend of a type a agent, that is,

$$\mathbb{E}\big(d_{i,b} \times 1\big(y_i = a\big)\big) = \mathbb{E}\big(d_{i,a} \times 1\big(y_i = b\big)\big). \tag{6}$$

Combining equations (5) and (6) results in the following moment equation:

$$\mathrm{d}_{b|a} \times \mathbb{P}(y_i = a) = d_{a|b} \times \mathbb{P}(y_i = b). \tag{7}$$

Rearranging the equation, the odd ratio of type a agents is

$$\frac{\mathbb{P}(y_i = a)}{\mathbb{P}(y_i = b)} = \frac{d_{a|b}}{\mathrm{d}_{b|a}} \tag{8}$$

or the proportion of type a agents is

$$\mathbb{P}(y_i = a) = \frac{d_{a|b}}{d_{a|b} + \mathrm{d}_{b|a}}. \tag{9}$$

Equation (9) indicates if we can consistently estimate the expected number of type a friends of the type b agents, then we can have a consistent estimate of the proportion of type a agents. This result is implied by the symmetricity of the adjacency matrix of an undirected graph.

3.2 Sample Moment Equation

Using the snowball sampling method, we collect the information of all the friends of the samples in the previous iteration. For example, if we have 10 agents in the initial iteration, then we will collect the information of all the friends of these 10 agents. Thus, we can estimate the value of $d_{b|a}$ by computing the average number of type b friends of a type a agents. The estimator of $d_{b|a}$ is defined as

$$\hat{d}_{b|a} = \sum_{i \in V_0} \sum_{j \in C_i, y_j = b} N_a^{-1} 1(y_i = a) 1(y_j = b), \tag{10}$$

where $N_a^{-1} = \sum_i^N 1(y_i = a)$. We can define the estimator of $\hat{d}_{a|b}$ in the same way. Notice that $\hat{d}_{b|a}$ and $\hat{d}_{a|b}$ are consistent even if the initial sample is subject to sample selection in terms of types. The initial proportion of the types of agents does not affect the consistency of the estimation of $d_{b|a}$ because it is conditional on the type of agents. The initial proportion of types of agents affects only the variance of $\hat{d}_{b|a}$ and $\hat{d}_{a|b}$ because it affects the number of observations of different types of agents.

Combining equations (9) and (10), the estimator of $\mathbb{P}(y_i = a)$ is

$$\hat{\mathbb{P}}(y_i = a) = \frac{\hat{d}_{a|b}}{\hat{d}_{a|b} + \hat{d}_{b|a}}. \tag{11}$$

3.3 Asymptotic Distribution

In this section, we discuss the assumptions and derive the asymptotic distribution of the proposed estimator.

Assumption 3.1. Conditional on the types of agents, the samples are independent and identically distributed and $Var(d_{i|a}|y_i = b) < \infty, \forall\, a \neq b$,

The sample selection problem studied in this chapter is focused on the selection of the types of the agents. The estimation does not rely on how the researchers select the proportion of the types. However, given the types of the agents, the samples collected are assumed to be independent and identically distributed.

Assumption (3.1) is not a necessary condition. If the number of friends of different types for different types of agents satisfied the Linderberg condition:

$$\sum_i \lim_{N\to\infty} s_{a,N}^{-2} \sum_i^N \Big(d_{i,b}1(y_i = a) - \mathbb{E}\big(d_{i,b}1(y_i = a)\big)1\Big(\big|d_{i,b}1(y_i = a)\big| > \varepsilon s_n\Big)\Big) = 0 \;\; \forall a \neq b, \quad (12)$$

where $s_{a,N} = \sum_i^N Var\big(d_{i,b}1(y_i = a)\big)$, then we can use the Linderberg central limit theorem instead of the Linderberg-Lévy Central Limit Theorem. For simplicity, we will stick with the independent and identically distributed assumption.

Assumption 3.2. $\mathbb{P}>(A_{i,j} = 1|y_i = a,\ y_j = b) > 0$ and $\sum_{i,j} A_{i,j}1(y_i = a)1(y_i = b) = Op(N)$

Assumption (3.2) requires some links between two types of agents. The identification of the model relies on the fact that the number of links between type b agents and type a agents is always the same as the number of links between type a agents and type b agents. The model is not identified if there are no links between different types of agents. It also requires the number of links increases at a slower rate the same sample size N. This restricts the type a agents connect to all type b agents.

Assumption 3.3. Let $N = N_a + N_b$, where N_a and N_b are the initial observations of type a and type b agents. As $N \to \infty, \dfrac{N_a}{N} \to sp_a \in (0,1)$.

The initial sample proportion of type a agents converge to a constant proportion $0 < sp_a < 1$.

Proposition 1. Under *Assumptions (3.1)* to *(3.3)*, the asymptotic distribution of the estimator is

$$\sqrt{N}\Big(\hat{\mathbb{P}}\big(y_i = a\big) - \mathbb{P}\big(y_i = a\big)\Big) \to N\big(0, \sigma_{pa}^2\big), \quad (13)$$

where

$$\sigma_{pa}^2 = \sigma^2\left(d_{b|a} + d_{b|a}\right)^{-4}\left(d_{b|a}sp_a^{-1} - d_{a|b}sp_b^{-1}\right)^2 \quad (14)$$

$$\sigma^2 = \mathrm{p}\lim_{N\to\infty} Var\left(N^{-1}\sum_{i,j\in V} A_{i,j}1\big(y_i = a\big)1\big(y_i = b\big)\right). \quad (15)$$

The equation for the variance of the proposed estimator indicates, when $\sigma^2_{b|a}$ or $\sigma^2_{a|b}$ increases, the variance of the proposed estimator increases. This is an obvious result. On the other hand, when $d_{b|a}$ or $d_{a|b}$ increases, the variance of the proposed estimator decreases because the identification strategy depends on the number of type a friends of an type b agent. When these numbers are small, we do not have much information about the network structure between two types of agents; hence, the variance of the proposed estimator would be large.

We can estimate σ^2 using the sample variance of $d_{a|b}$ and $d_{b|a}$, that is,

$$\hat{\sigma}^2 = \frac{sp_b^2 \hat{\sigma}^2_{d_{b|a}} + sp_a^2 \hat{\sigma}^2_{d_{a|b}}}{2}. \tag{16}$$

3.4 More Than Two Types

Suppose there are k types of agents and the types are labeled 1, 2, …, k. The moment equation can be written as

$$d_{a|b} \times \mathbb{P}(y_i = b) - d_{b|a} \times \mathbb{P}(y_i = a) = 0 \;\; \forall a \neq b \tag{17}$$

or the number of moment equations is $\frac{k(k-1)}{2}$. When $k = 2$, the model is just identified. When $k > 2$, we have more moment equations than parameters. We can estimate the model by generalized method of moment. The sample analog of the moment equations are

$$\hat{d}_{a|b} \times \mathbb{P}(y_i = b) - \hat{d}_{b|a} \times \mathbb{P}(y_i = a) = 0 \;\; \forall a \neq b \tag{18}$$

where $\hat{d}_{a|b} = N_b^{-1} \sum_{i, y_i = b} \sum_{j, y_j = a} A_{[i,j]}$ and $\hat{d}_{b|a} = N_a^{-1} \sum_{i, y_i = a} \sum_{j, y_j = b} A_{[i,j]}$

We can arrange the moment equation in the following matrix form:

$$\begin{pmatrix} -d_{2|1} & d_{1|2} & 0 & 0 & \dots \\ -d_{3|1} & 0 & d_{1|3} & 0 & \dots \\ 0 & -d_{3|2} & d_{2|3} & 0 & \dots \\ \vdots & & \dots & & \vdots \\ \vdots & 0 & \dots & 0 & \vdots \end{pmatrix} \begin{pmatrix} \mathbb{P}(y_i = 1) \\ \mathbb{P}(y_i = 2) \\ \mathbb{P}(y_i = 3) \\ \vdots \\ \mathbb{P}(y_i = k) \end{pmatrix} \equiv DP = 0. \tag{19}$$

Because $\mathbb{P}(y_i = k) = 1 - \mathbb{P}(y_i = 1) - \cdots - \mathbb{P}(y_i = k - 1)$, we can rewrite the moment equations as

$$\tilde{D} P_{-k} + D_k = 0 \tag{20}$$

where $\tilde{D} = \left(D_{-k} - D_k 1'_{k-1}\right)$, $D = [D_{-k}, D_k]$, $P = (P_{-k}, P_k)$, and 1_{k-1} is a $k - 1 \times 1$ vector of ones.

Similar to equation (10), we replace $d_{b|a}$ for any $a \neq b$ with their sample analog $\hat{d}_{b|a}$. The generalized method of moment estimator is the solution of

$$\hat{P} = \underset{v \in \mathbb{R}_{+}^{k-1}}{\arg\min} \left(\tilde{D}_v + D_k\right)' W \left(\tilde{D}_v + D_k\right) \tag{21}$$

where W is a $\frac{(k-2)(k-1)}{2}$ weighting matrix and $W \to_p W_0$ and W_0 is a $\frac{(k-2)(k-1)}{2}$ positive definite matrix.

The first-order condition of the above minimization problem is

$$\tilde{D}' W \left(\tilde{D}_v + D_k\right) = 0. \tag{22}$$

We can rewrite the moment equations in equation (18) for any $a \neq b$ as

$$N^{-1} \sum_i h_i = 0 \tag{23}$$

where $h_i = (h_{i,1,2}, h_{i,1,3}, \ldots, h_{i,k-1,k})'$ is a $k \times 1$ vector of all the moment equations of i and $h_{i,a,b} \equiv \frac{1(y_i = b)}{sp_b} \sum_{j, y_j = a} A_{[i,j]} \mathbb{P}(y_i = b) - \frac{1(y_i = a)}{sp_a} \sum_{j, y_j = b} A_{[i,j]} \mathbb{P}(y_i = a)$. Please see Section 7 for the derivation.

Let $S = \text{plim} N^{-1} \sum_i h_i h_i'$.[3] The diagonal elements of S are

$$\sigma_{a,b}^2 \left(p_b^2 sp_b^{-2} + p_a^2 sp_a^{-2} - 2P_a P_b sp_a^{-1} sp_b^{-1}\right) \tag{24}$$

where $P_k = 1 - ||P_{-k}||_1$ and $\sigma_{a,b}^2 = \text{p}\lim_{N \to \infty} Var\left(N^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right)$

There are two possible cases for the off-diagonal elements. Each of the moment equations links two types of agents. When we look at the covariance of a pair of moment equations, if all the types of agents are different in both moment equations, then the covariance of the moment equations is 0.

When there is one common type of agent, the covariance of the moment equation is

$$\sigma_{b,c|a} P_a^2 sp_a^{-1} \tag{25}$$

where $\sigma_{b,c|a} = Cov(d_{i|a}|y_i = b, d_{i|a}|y_i = c)$.

If there are two common types of agents between moment equations, this implies two moment equations are the same and the covariance would be the variance of the moment equation.

Proposition 2. Under *Assumptions (3.1)* to *(3.3)* for all $a \neq b$ and $a, b \in 1, \ldots, k$, the asymptotic distribution of the generalized method of moment estimator is

$$\sqrt{N}\left(\hat{P}_k - P_k\right) \to N(0, \Omega) \tag{26}$$

where $\Omega = \left(\tilde{D}'W\tilde{D}\right)^{-1}\left(\tilde{D}'WSW\tilde{D}\right)\left(\tilde{D}'W\tilde{D}\right)^{-1}$.

When $k = 1$, the asymptotic variance of P_a[4] is

$$\left(d_{a|b} + d_{b|a}\right)^{-4}\left(d_{a|b} + d_{b|a}\right)^{-2}\sigma^2_{a,b}\left(p_b^2 sp_b^{-2} + p_a^2 sp_a^{-2} - 2P_a P_b sp_a^{-1} sp_b^{-1}\right). \quad (27)$$

Since $\hat{P}_a = \dfrac{d_{b|a}}{d_{b|a} + d_{a|b}}$ and $\hat{P}_b = \dfrac{d_{b|a}}{d_{b|a} + d_{a|b}}$, Equation (27) reduces to

$$\sigma^2_{a,b}\left(d_{a|b} + d_{b|a}\right)^{-4}\left(sp_b^{-1}d_{a|b} - sp_a^{-1}d_{b|a}\right)^2 \quad (28)$$

which is the same as the result in the previous section.

3.5 Optimal Weighting

We can obtain a more efficient estimator by setting the weighting matrix, $W = \hat{S}^{-1}$, where $\hat{S}^{-1} \to_p S$. The element of S is defined in equations (24) and (25). The terms σ^2_a, sp_a, and $\sigma_{b,c|a}$ for all $a \neq b \neq c$ can be estimated from the samples, and P_a for all $a = 1, \ldots, k$ can be estimated using the procedure in the previous section by assuming the weighting matrix to be identity matrix. The limiting variance of the estimators would become $(D'S^{-1}D)^{-1}$.

The proportion of types of agents in the initial sample could affect the asymptotic variance of the proposed estimator. For example, when $k = 2$, the variance is

$$\sigma^2_{a,b}\left(d_{a|b} + d_{b|a}\right)^{-4}\left(d_{a|b}sp_b^{-1} - d_{b|a}sp_a^{-1}\right)^2 \quad (29)$$

4 SIMULATION

In this section, we conduct a simulation experiment to study the sample selection problem in the snowball sampling and the performance of the proposed method. We have two different setups for the simulation. In the first setup, we simulate the social network by the stochastic block model by Holland, Laskey, and Leinhardt (1983). In the second setup, we draw a subsample from the social network data collected in the first and second chapters. In both setups, we assume two types of agents.

4.1 Data Generating Process: Stochastic Block Model

The stochastic block model (Holland et al., 1983) is a model for network formation. It assumes k types of vertices and the probability of having a link between type a vertex and type b vertex is a constant parameter $P_{a,b}$. In our simulation, there are two types of vertices, type 0 and type 1. Let y_0 be a vector of binary variables; the ith element of $y_0 = 1$ if vertex i is of type 0, otherwise 0. Similarly, the ith element of $y_1 = 1$ if vertex i is of type 1.

The probability of the adjacency matrix is

$$\mathbb{P}(A) = \begin{pmatrix} y_1 & y_2 \end{pmatrix} \begin{pmatrix} P_{0,0} & P_{0,1} \\ P_{1,0} & P_{1,1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \tag{30}$$

where $P_{0,1} = P_{1,0}$ and $A_{i,j} = A_{j,i}$.

In the simulation, we assume the population size of the network is 500 and control the probability of the link by the expected connections between types of agents:

$$P_{0,0} = \frac{d_{0|0}}{N_0} \tag{31}$$

$$P_{1,1} = \frac{d_{1|1}}{N_1} \tag{32}$$

$$P_{1,0} = P_{0,1} = \frac{d_{1|0}}{N_0} = \frac{d_{0|1}}{N_1} \tag{33}$$

To study the sample selection problem and the performance of the proposed estimator, we change the proportion of type 0 vertices in the initial sample, the proportion of type 0 vertices in the underlying data generating process, and the probability of the connections between two types of agents. Table 1 shows the simulation results when vertices are homophilic, which means two vertices with the same type are more likely to have a connection than two vertices with different types. In this simulation, we assume the expected number of connections between type 0 and type 1 agents is 500. The values of $d_{0|1}$ and $d_{1|0}$ depend on the proportion of different types of vertices in the simulation, shown in Table 1. In addition, we assume $d_{0|0} = 5$ and $d_{1|1} = 7$. That is, on average, type 0 (type 1) vertices will have 5 (7) connections to other vertices is also of type 0 (type 1).

The results indicate the sample proportion of the snowball samples suffer from sample selection bias, while the proposed estimator does not show bias. In fact, we can approximate the sample selection bias of the sample proportion using the following equation.

$$\begin{pmatrix} d_{0|0} & d_{0|1} \\ d_{1|0} & d_{1|1} \end{pmatrix} \begin{pmatrix} \frac{N_0}{N} \\ \frac{N_1}{N} \end{pmatrix} \tag{34}$$

For example, when the proportion is (0.3, 0.7), the equation gives

$$\begin{pmatrix} 5 & 1.43 \\ 3.33 & 7 \end{pmatrix} \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} \tag{35}$$

Table 1. Simulation Results of the Estimators (Vertices Are Homophilic, $d_{0|0} = 5$ and $d_{1|1} = 7$).

		True	Initial	Sample Proportion			Proposed Estimator		
$d_{1\|0}$	$d_{0\|1}$	Proportion	Proportion	Bias	SD	RMSE	Bias	SD	RMSE
3.33	1.43	0.30	0.30	0.00	0.02	0.02	0.00	0.05	0.05
3.33	1.43	0.30	0.50	0.09	0.03	0.09	−0.00	0.05	0.05
3.33	1.43	0.30	0.70	0.17	0.03	0.17	−0.00	0.05	0.05
2.00	2.00	0.50	0.30	−0.15	0.03	0.15	0.01	0.06	0.06
2.00	2.00	0.50	0.50	−0.07	0.03	0.07	0.00	0.05	0.05
2.00	2.00	0.50	0.70	0.03	0.03	0.04	−0.01	0.06	0.06
1.43	3.33	0.70	0.30	−0.28	0.02	0.28	0.00	0.05	0.05
1.43	3.33	0.70	0.50	−0.21	0.03	0.21	−0.00	0.04	0.04
1.43	3.33	0.70	0.70	−0.12	0.03	0.12	−0.00	0.04	0.04

After normalization, the sample proportion is (0.298, 0.702) and the bias is close to 0. Another example, when the proportion is 0.5 and initial proportion is 0.3, the equation gives

$$\begin{pmatrix} 5 & 2 \\ 2 & 7 \end{pmatrix} \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix} \tag{36}$$

After normalization, the sample proportion is (0.345, 0.655), and the bias is –0.155. Both examples give numbers similar to the simulation results.

Table 2 shows the simulation results when vertices are heterophilic, which means two vertices with the same type are less likely to have a connection than two vertices with different types. In this simulation, we assume the expected number of connections between type 0 and type 1 agents is 1,000. The values of $d_{0|1}$ and $d_{1|0}$ depend on the proportion of different types of vertices in the simulation as shown in Table 6. In addition, we assume $d_{0|0} = 3$ and $d_{1|1} = 3$. That is, on average, type 0 (type 1) vertices will have 5 (7) connections to other vertices also of type 0 (type 1).

Table 2. Simulation Results of the Estimators (Vertices Are Heterophilic, $d_{0|0} = d_{1|1} = 3$).

		True	Initial	Sample Proportion			Proposed Estimator		
$d_{1\|0}$	$d_{0\|1}$	Proportion	Proportion	Bias	SD	RMSE	Bias	SD	RMSE
6.67	2.86	0.30	0.30	0.11	0.03	0.12	0.00	0.03	0.03
6.67	2.86	0.30	0.50	0.08	0.03	0.08	−0.00	0.03	0.03
6.67	2.86	0.30	0.70	0.05	0.03	0.05	−0.00	0.04	0.04
4.00	4.00	0.50	0.30	0.02	0.03	0.04	0.01	0.04	0.05
4.00	4.00	0.50	0.50	−0.01	0.03	0.03	−0.00	0.04	0.04
4.00	4.00	0.50	0.70	−0.04	0.03	0.04	−0.00	0.04	0.04
2.86	6.67	0.70	0.30	−0.06	0.02	0.06	0.00	0.04	0.04
2.86	6.67	0.70	0.50	−0.09	0.03	0.09	0.00	0.03	0.03
2.86	6.67	0.70	0.70	−0.12	0.03	0.13	−0.00	0.03	0.03

The simulation results also indicate the sample proportions are biased and the proposed estimators are unbiased.

4.2 Data Generating Process: Subsample of Social Network Data

In this setup, we draw a subsample from social network data discussed in the first and second chapters. We select all grade nine students with at least one friend in a school in the dataset used in the first and second chapter. There are 194 students with at least one friend, 73 of them were male, and 121 of them were female. The proportion of male students was 37.62%. Fig. 1 shows the visualization of the friendship network. Each vertex is a student, and the shape of the vertices represent gender of the students. If two students are friends, they are linked with a black edge. As indicated in Fig. 1, students are more likely to have friends of the same gender. This observation motivates the sample selection bias in the snowball sampling. Table 3 shows the average number of friends for each student by gender. The first row shows the average number of female friends by gender of female students, and the second row shows the same statistics for male students.

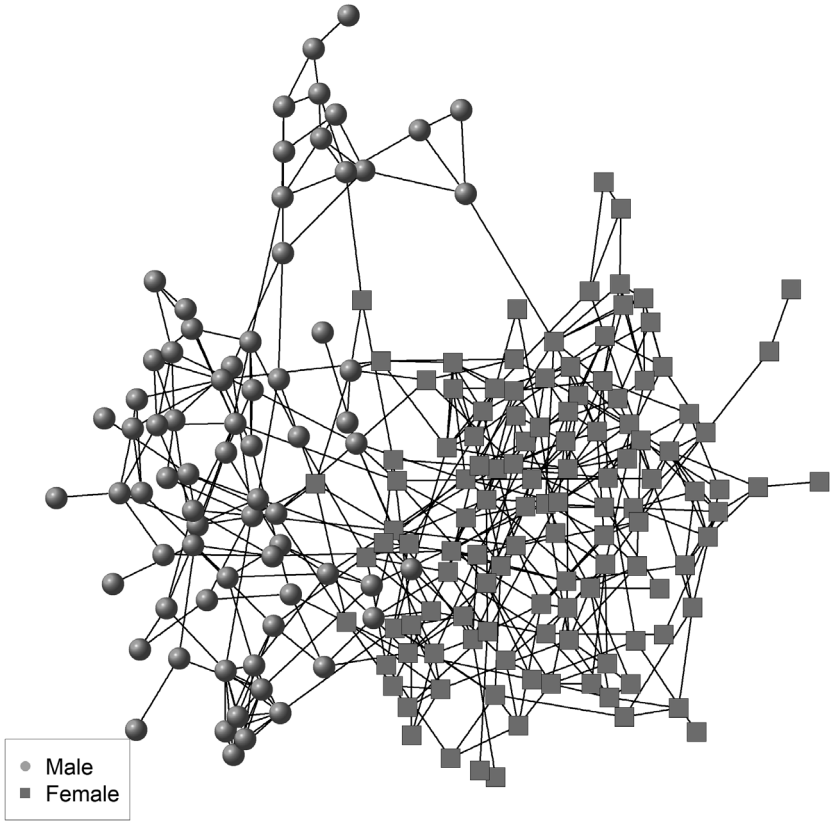

Fig. 1. Friendship Network of Students from the Same Grade and School.

Table 3. Average Numbers of Friends by Gender.

	Male Friends	Female Friends
Female students	0.264	4.959
Male students	3.836	0.438

In the simulations, we drew from the data $n \in \{20, 40, 60\}$ students, with the proportion of male students being equal to $\frac{N_{\text{male}}}{N} \in \{0.3, 0.40, 5, 0.6, 0.7\}$, with replacement and then collected the information of their friends with the snowball sampling.

Table 4 shows the simulation results. The proportion of male students was 37.62%. In general, the sample proportion was biased because of the sample selection problem of the snowball sampling. The bias of the proposed estimator is much smaller than the sample proportion. When the sample size is small ($N = 20$), we still observed a small bias in the estimator. The bias decreases as the sample size increases.

The standard error of the proposed estimator is much higher than it is for the sample proportion because the expected number of female friends of a male student and the expected number of male friends of a female student is very small. They are equal to 0.438 and 0.264, respectively. Since the identification strategy relies on these two quantities, if both of them are very small, the variance of the

Table 4. Simulations of the Estimators Using Social Network Data.

Initial Proportion of Male Students	Sample Proportion			Proposed Estimator		
	Bias	SD	RMSE	Bias	SD	RMSE
			$N = 20$			
0.350	−0.058	0.041	0.320	0.040	0.281	0.502
0.450	0.025	0.043	0.403	0.001	0.267	0.462
0.500	0.067	0.043	0.445	−0.014	0.258	0.445
0.550	0.111	0.044	0.490	−0.021	0.276	0.450
0.650	0.200	0.045	0.578	−0.056	0.273	0.421
			$N = 40$			
0.350	−0.060	0.029	0.317	0.017	0.190	0.436
0.450	0.026	0.029	0.403	−0.000	0.178	0.416
0.500	0.068	0.032	0.446	−0.006	0.184	0.413
0.550	0.112	0.031	0.489	−0.019	0.189	0.404
0.650	0.202	0.032	0.578	−0.025	0.201	0.405
			$N = 60$			
0.350	−0.059	0.024	0.318	0.013	0.154	0.418
0.450	0.026	0.024	0.403	0.005	0.148	0.408
0.500	0.066	0.026	0.443	−0.003	0.150	0.401
0.550	0.120	0.026	0.496	−0.003	0.157	0.405
0.650	0.202	0.026	0.579	−0.010	0.170	0.403

proposed estimator can be large. The equation of the variance of the proposed estimator also suggests the same result. In addition, the number of initial samples is small in our simulations. For this reason, the standard deviation of the proposed estimator in the simulations is large.

5 EMPIRICAL APPLICATION

On September 28, 2014, activists in Hong Kong protested outside the government headquarters and then occupied several major roads in the city. The name *Umbrella Movement* was suggested by Adam Cotton on Twitter because the umbrellas were used for defense against tear gas. The public quickly accepted the name. In the next few months, many young people showed their support for the movement by changing their Facebook profile pictures to yellow ribbons or yellow umbrellas on a black background. The supporters of the government and police changed their profile pictures to blue ribbons.

Telephone surveys are one way to collect opinions of citizens about the movement. However, conducting random sampling of the population to collect the data would be costly. It is relatively easy to collect data by snowball sampling through Facebook. Our objective is to estimate how many people switched their profile pictures to yellow ribbons or blue ribbons.

From an initial sample of 142 individuals from Facebook, we collected 45,785 individuals through snowball sampling. In total, we had 45,927 profile pictures. Each individual is classified as one of the following types: no change, yellow ribbons, and blue ribbons. These types corresponded to those who did not change their profile pictures, changed their profile pictures to yellow ribbons and changed their profile pictures to blue ribbons.

In the initial sample, 46 individuals changed their profile pictures to yellow ribbons, 11 changed their profile pictures to blue ribbons, and 85 did not change their profile pictures. Of the samples collected through the snowball sampling, 9,418 changed their profile pictures to yellow ribbons, 468 changed their profile pictures to blue ribbons, and 35,899 did not change their profile pictures. Table 5 shows the summary of the proportion of supporters of the movement and supporters of the government in the initial samples and snowball samples.

The initial samples were not randomly selected. We intentionally selected more initial samples with blue ribbons because the proportion of profiles with blue ribbons is relatively small. More importantly, the data collected through the

Table 5. Sample Proportion of the Types of Profile Pictures in the Samples.

	No Change	Yellow Ribbons	Blue Ribbons	Total
Initial samples	85	46	11	142
	59.86%	32.39%	7.75%	
Snowball samples	35,899	9,418	468	45,785
	78.40%	20.57%	1.02%	
Total	35,984	9,464	479	45,927
	78.36%	20.60%	1.04%	

Table 6. Average Number of Friends by Type for Each Type of Sample.

Types	Average Number of Friends by Type			
	No Change	Yellow Ribbons	Blue Ribbons	Total
No change	276.35	57.29	3.76	337.41
Yellow ribbon	218.37	91.41	1.41	312.20
Blue ribbon	214.91	31.18	7.55	255.64

snowball sampling method were subject to sampling bias because people have friends with similar political views. Table 6 shows the average number of friends by type for each type in the samples and Table 7 shows the average proportion of number of friends by type for each type in the samples.

The proportion of friends with yellow ribbons of a user with a yellow ribbon was 29.4%, which is higher than the proportion of friends with yellow ribbons for a user with a blue ribbon (12.3%). Similarly, the proportion of friends with blue ribbons of a user with blue ribbon was 3%, which is lower than the proportion of friends with blue ribbons for a user with a yellow ribbon (12.3%).

The number of friends also plays a role in the sampling bias. The average number of friends for those who did not change their profile pictures is 337.4. This is higher than the average number of friends for those who changed their profile picture to yellow or blue ribbons, which have 312.2 and 255.64 friends, respectively. Users with yellow ribbons had more friends than those with blue ribbons, and hence, the snowball samples would be biased toward the proportion of samples with yellow ribbons.

Using the proposed method, we solved in moment equation (37) to obtain the estimates of the proportion of each type. The estimates are shown in Table 8. The estimates of the proportion of users displaying blue ribbons increased to 1.4% using the proposed method, which is 40% higher than the sample average of the initial samples and samples collected through snowball sampling. This suggests the selection bias may greatly affect the estimated proportion.

Since most of the Facebook users in our samples (78%) did not change their profile pictures, the proportion of users with yellow and blue ribbons did not change much. Focusing on the users who changed their profile pictures made observing changes easier. Table 9 shows the proportion of users changed their profile pictures to yellow or blue ribbons. In the initial samples, the proportions were 80.7% and 19.3%. With the samples collected through snowball sampling, the proportions became 95.27% and 4.73%. Finally, using the proposed method,

Table 7. Average Proportion of Friends by Type for Each Type of Sample.

Types	Average Proportion of Friends by Type		
	No Change	Yellow Ribbon	Blue Ribbon
No change	0.819	0.170	0.011
Yellow ribbon	0.702	0.294	0.005
Blue ribbon	0.847	0.123	0.030

Table 8. Estimates of Sample Mean and the Proposed Method Using the Snowball Samples.

	No Change	Yellow Ribbon	Blue Ribbon
Initial samples	0.599 (0.041)	0.324 (0.039)	0.077 (0.022)
Initial samples and snowball samples	0.784 (0.002)	0.206 (0.002)	0.010 (0.001)
New method	0.781 (0.028)	0.205 (0.028)	0.014 (0.003)

Table 9. Estimates of Sample Mean and the Proposed Method Using the Snowball Samples.

	Yellow Ribbon	Blue Ribbon
Initial samples	0.8070 (0.06)	0.1930 (0.035)
Initial Samples and Snowball Samples	0.9527 (0.004)	0.0473 (0.002)
New method	0.9378 (0.06)	0.0622 (0.006)

the proportions were 93.78% and 6.22%. The estimates of the proportion of users changed their profile pictures to blue ribbons increased from 4.73% to 6.22%.

$$\begin{aligned} 57.29P_{\text{yellow}} - 218.36P_{\text{nochange}} &= 0 \\ 3.76P_{\text{nochange}} - 214.9P_{\text{blue}} &= 0 \\ 1.41P_{\text{yellow}} - 31.18P_{\text{blue}} &= 0 \end{aligned} \tag{37}$$

6 CONCLUSION

This chapter proposes a new estimation method that corrects for the sample selection problem in snowball sampling. The method relies on two important observations. First, we can consistently estimate the average number of friends of an agent given the type of agent. Although the snowball samples are subject to sample selection, it does not play a role in estimating the number of friends given the type of an agent. Second, the adjacency matrix of an undirected graph is symmetric. This implies the number of links from type 0 agent to type 1 agent is the same as the link from type 0 agent to type 1 agent.

Using these two observations, we derived the moment equations, estimated the proportion of the types of agents, and derived the asymptotic distribution of the estimator. We also investigated the finite sample properties with two simulation studies.

As an empirical application, we used samples collected from Facebook to estimate the proportion of Facebook users who supported the *Umbrella Movement* in Hong Kong in 2014. Facebook users changed their profile pictures to yellow ribbons to show their support for the movement and to blue ribbons to show their support for the government and the police. The results indicated that the simple average of the proportion in the snowball samples underestimated by 40% the proportion of Facebook users who changed their profile pictures to blue ribbons.

7 PROOF

Proof of Proposition 1

First, we have to find out the asymptotic distribution of $\hat{d}_{b|a}$ and $\hat{d}_{a|b}$. $\hat{d}_{b|a}$ and $\hat{d}_{a|b}$ can be written as,

$$d_{b|a} = N_a^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b) \tag{38}$$

and

$$d_{a|b} = N_b^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b) \tag{39}$$

The variance and covariance of $d_{b|a}$ and $d_{a|b}$ are

$$Var\left(d_{b|a}\right) = \left(\frac{N}{N_a}\right)^2 Var\left(N^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right) \tag{40}$$

$$Var\left(d_{a|b}\right) = \left(\frac{N}{N_b}\right)^2 Var\left(N^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right) \tag{41}$$

$$Cov\left(d_{a|b}, d_{b|a}\right) = \left(\frac{N^2}{N_b N_a}\right) Var\left(N^{-1} \sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right) \tag{42}$$

By *Assumption (3.2)*, $Var\left(N^{-1}\sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right) \xrightarrow{p} \sigma^2$. The asymptotic normality of $\hat{d}_{b|a}$ is implied by the Linderberg-Lévy Central Limit Theorem. By *Assumption (3.3)*, $\frac{N_a}{N} \to sp_a$ and $\frac{N_b}{N} \to sp_b$, the asymptotic distribution of $\hat{d}_{b|a}$ and $\hat{d}_{a|b}$ is

$$\begin{bmatrix} \sqrt{N}\left(\hat{d}_{b|a} - \hat{d}_{b|a}\right) \\ \sqrt{N}\left(\hat{d}_{a|b} - \hat{d}_{a|b}\right) \end{bmatrix} \to N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \begin{pmatrix} sp_b^{-2} & sp_a^{-1} sp_a^{-1} \\ sp_a^{-1} sp_a^{-1} & sp_b^{-2} \end{pmatrix}\right] \tag{43}$$

where $Var\left(N^{-1}\sum_{i,j \in V} A_{i,j} 1(y_i = a) 1(y_i = b)\right) \xrightarrow{p} \sigma^2$

Then we apply the delta method to obtain the asymptotic distribution of $\hat{\mathbb{P}}(y_i = a) = \frac{\hat{d}_{a|b}}{\hat{d}_{a|b} + \hat{d}_{b|a}}$, that is,

$$\sqrt{N}\left(\hat{\mathbb{P}}(y_i = a) - \mathbb{P}(y_i = a)\right) \to N\left(0, \sigma_{pa}^2\right) \tag{44}$$

Where

$$\sigma_{pa}^2 \equiv \begin{pmatrix} -\dfrac{d_{b|a}}{\left(d_{a|b}+d_{b|a}\right)^2} \\ \dfrac{d_{a|b}}{\left(d_{a|b}+d_{b|a}\right)^2} \end{pmatrix}' \begin{pmatrix} sp_a^{-2}\sigma^2 & sp_a^{-1}sp_b^{-1}\sigma^2 \\ sp_a^{-1}sp_b^{-1}\sigma^2 & sp_b^{-2}\sigma^2 \end{pmatrix} \begin{pmatrix} -\dfrac{d_{b|a}}{\left(d_{a|b}+d_{b|a}\right)^2} \\ \dfrac{d_{a|b}}{\left(d_{a|b}+d_{b|a}\right)^2} \end{pmatrix} \quad (45)$$

$$= \sigma^2\left(d_{b|a}+d_{b|a}\right)^{-4}\left(d_{b|a}sp_a^{-1}-d_{a|b}sp_b^{-1}\right)$$

Proof of **equation (23)**

$$\hat{d}_{a|b}\times\mathbb{P}\left(y_i=b\right)-\hat{d}_{b|a}\times\mathbb{P}\left(y_i=a\right)=0 \quad (46)$$

$$N_b^{-1}\sum_{j,y_j=b}\sum_{i,y_i=a}A_{[i,j]}\times\mathbb{P}\left(y_i=b\right)-N_a^{-1}\sum_{j,y_j=a}\sum_{i,y_i=b}A_{[i,j]}\times\mathbb{P}\left(y_i=a\right)=0 \quad (47)$$

$$N_b^{-1}\sum_i 1\left(y_i=b\right)\sum_{i,y_i=a}A_{[i,j]}\times\mathbb{P}\left(y_i=b\right)-N_a^{-1}\sum_i 1\left(y_i=a\right)\sum_{i,y_i=b}A_{[i,j]}\times\mathbb{P}\left(y_i=a\right)=0 \quad (48)$$

$$N^{-1}\sum_i\left(\frac{1\left(y_i=b\right)}{sp_b}\sum_{i,y_i=a}A_{[i,j]}\mathbb{P}\left(y_i=b\right)-\frac{1\left(y_i=a\right)}{sp_a}\sum_{i,y_i=b}A_{[i,j]}\mathbb{P}\left(y_i=a\right)\right)=0 \quad (49)$$

Finally, let

$$h_{i,a,b}\equiv\frac{1\left(y_i=b\right)}{sp_b}\sum\nolimits_{i,y_i=a}A_{[i,j]}\mathbb{P}\left(y_i=b\right)-\frac{1\left(y_i=a\right)}{sp_a}\sum\nolimits_{i,y_i=b}A_{[i,j]}\mathbb{P}\left(y_i=a\right)$$

NOTES

1. Because agents selected during the previous iteration of snowball sampling, adjustment in the transition matrix is needed. For simplicity, we assume the transition matrix is same for every iterations of snowball sampling. The stationary proportion of the types of agents is the eigenvectors of the transition matrix.
2. The variance of the estimator would depend on the number of observations.
3. $\sum_{i,j}h_ih_i^{'}=\sum_i h_ih_i^{'}$ because the samples are independent except for the proportion of the initial samples.
4. $P_b = 1 - P_a$ and $Var(P_a) = Var(P_b)$.

ACKNOWLEDGMENTS

I am grateful to my main advisor Iván Fernández-Val for his guidance and patience. I appreciate the helpful feedback from Pierre Perron, M. Daniele Paserman, Hiroaki Kaido, Ho-Po Crystal Wong, Tak-Yuen Wong, Vladimir Yankov, and seminar participants in Econometrics seminar in Boston University. I also thank the anonymous reviewers for their insightful comments and suggestions.

The views expressed here are solely the author's and do not represent the views of Bates White, or their other employees. All errors are mine.

REFERENCES

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four degrees of separation. In *Proceedings of the 4th annual ACM web science conference* (WebSci '12). Association for Computing Machinery, New York, NY, USA, 33–42. DOI:https://doi.org/10.1145/2380718.2380723.

Chandrasekhar, A., & Lewis, R. (2011). Econometrics of sampled networks. Unpublished manuscript, Massachusetts Institute of Technology.

Gurevitch, M. (1961). The social structure of acquaintanceship networks. Ph.D. thesis, Massachusetts Institute of Technology.

Holland, P. W., Laskey, K. B., & Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, *5*(2), 109–137.

Kolaczyk, E. D. (2009). *Statistical analysis of network data: Methods and models*. Springer Series in Statistics. New York, NY: Springer.

McPherson, M. Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1), 60–67.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 425–443.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the Facebook social graph. *arXiv preprint*, arXiv:1111.4503.