Journal of Business & Economic Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ubes20

Estimating Spatial Autocorrelation with Sampled Network Data

Jing Zhou[a], Yundong Tu[a], Yuxin Chen[b] & Hansheng Wang[a]

[a] Peking University

[b] New York University Shanghai

Accepted author version posted online: 17 Jul 2015.

To cite this article: Jing Zhou, Yundong Tu, Yuxin Chen & Hansheng Wang (2015): Estimating Spatial Autocorrelation with Sampled Network Data, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2015.1061437

To link to this article: http://dx.doi.org/10.1080/07350015.2015.1061437

Disclaimer: This is a version of an unedited manuscript that has been accepted for publication. As a service to authors and researchers we are providing this version of the accepted manuscript (AM). Copyediting, typesetting, and review of the resulting proof will be undertaken on this manuscript before final publication of the Version of Record (VoR). During production and pre-press, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal relate to this version also.

Estimating Spatial Autocorrelation with Sampled Network Data

Jing Zhou[1], Yundong Tu[1], Yuxin Chen[2], and Hansheng Wang[1]

[1]*Peking University and* [2]*New York University Shanghai*

This version: June 27, 2015

Abstract

Spatial autocorrelation is a parameter of importance for network data analysis. To estimate spatial autocorrelation, maximum likelihood has been popularly used. However, its rigorous implementation requires the whole network to be observed. This is practically infeasible if network size is huge (e.g., Facebook, Twitter, Weibo, WeChat, etc). In that case, one has to rely on sampled network data to infer about spatial autocorrelation. By doing so, network relationships (i.e., edges) involving unsampled nodes are overlooked. This leads to distorted network structure and underestimated spatial autocorrelation. To solve the problem, we propose here a novel solution. By temporarily assuming that the spatial autocorrelation is small, we are able to approximate the likelihood function by its first order Taylor's expansion. This leads to the method of approximate maximum likelihood estimator (AMLE), which further inspires the development of paired maximum likelihood estimator (PMLE). Compared with AMLE, PMLE is computationally superior and thus is particularly useful for large scale network data analysis. Under appropriate regularity conditions (without assuming a small spatial autocorrelation), we show theoretically that PMLE is consistent and asymptotically normal. Numerical studies based on both simulated and real datasets are presented for illustration purpose.

KEY WORDS: Approximate Maximum Likelihood Estimator; Network Data Analysis; Paired Maximum Likelihood Estimator; Spatial Autocorrelation;

[†]Both Zhou and Wang's research are partly supported in part by National Natural Science Foundation of China (NSFC, 11131002, 11271032), the Business Intelligence Research Center at Peking University, and the Center for Statistical Science at Peking University. Tu's research is partly supported by Guanghua School of Management and the Center for Statistical Science at Peking University, and partly by National Natural Science Foundation of China (NSFC, 71301004 and 71472007).

1. INTRODUCTION

In the past few decades, there has been a surge of interest in analysis of network data. This is witnessed by a number of published book volumes including, for example, Scott (1992), Wasserman and Faust (1994), Cohendet et al. (1998), LeSage and Pace (2009), and research papers such as Case (1991), Brock and Durlauf (2001), Calvó-Armengol et al. (2009), Lee et al. (2013) among others. Social network analysis has produced a set of methods to analyze social structure. Practitioners are particularly interested in spatial autocorrelation, which plays an important role in characterizing spatial correlation between different nodes. Once spatial autocorrelation is estimated and the network structure is fixed, one can predict a node's behavior by inferring about its friends. This allows practitioners to (for example) evaluate an applicant's credibility by the credit history of its connected network friends. This makes fast, accurate, and large scale online credit scoring practically feasible. See also Lee et al. (2013) and Bronnenberg and Mahajan (2001) for some other interesting economics and marketing applications. To estimate spatial autocorrelation, a spatial autoregression model has been proposed and the method of maximum likelihood has been popularly used (Ord, 1975; Anselin, 1980; Lee et al., 2013).

Despite its popularity, the practical implementation of the spatial autoregression model and the corresponding maximum likelihood estimation are problematic. The main problem is that the popularly used spatial autoregression model is assumed for the network of the entire population, while statistical analysis is conducted based on sampled data. Inevitably, social interactions generated between the sampled and unsampled units are ignored. To fix the problem, one might want to assume that the whole network data is available. Unfortunately, this is seldom true in real world. Consider for example, the Facebook contains more than 700 million active users. Except the Facebook itself, nobody else can depict the entire network structure easily. Even for the Facebook, computing the whole network data for every research project is not wise, because the cost is to be significant. As a popular remedy, one might want to collect a sample with a practical size.

Subsequently, it is assumed that the intended network model holds for the sampled data. This is also problematic, because the autocorrelation between the sampled and unsampled units are overlooked. As a consequence, the true spatial autocorrelation would be underestimated, if the method of maximum likelihood is incorrectly applied (Chen et al., 2013). Then, how to conduct a correct maximum likelihood estimation for spatial autocorrelation based on sampled network data becomes a problem of interest.

Following Chen et al. (2013) and Lee et al. (2013), we assume a normal disturbance. This enables us to rigorously spell out the marginal likelihood function for the sampled data. Unfortunately, the resulting likelihood function involves both the observed and unobserved social network structure, which cannot be practically optimized. To solve the problem, we propose a novel solution. To fix the idea, we temporarily assume that the spatial autocorrelation is small. As a result, we are able to approximate the actual log likelihood function by its first order Taylor's expansion with respect to spatial autocorrelation. Surprisingly, we find that the resulting approximation involves mainly the observed network structure and the degree (i.e., the number of followers or followees) of each unit. Fortunately, degree numbers are summarized for each unit by most popular social network websites. See for example Facebook, Twitter, Sina Weibo, and others. Thus, they can be easily obtained, and the approximated log likelihood function can be practically optimized. This leads to an approximate maximum likelihood estimator (AMLE). Unreported numerical studies demonstrate that the AMLE is consistent and asymptotically normal when spatial autocorrelation is reasonably small.

Despite its theoretical attractiveness, AMLE is not cheap computationally. Let n be the sample size. The computation of AMLE involves a $n \times n$ matrix, whose determinant needs to be evaluated and thus results in expensive computation. As a result, AMLE cannot be our final solution for large scale network data analysis. Instead, it can only serve as an intermediate step. However, this intermediate step inspires the following novel solution. Specifically, for a total of n sampled units, we form them into different pairs and each pair contains two different nodes, denoted by i and j. This leads to a total of $n(n-1)/2$ pairs. By treating $\{i, j\}$ as a small sample and following the idea of AMLE, the first order approximation of their log likelihood function can be obtained. Interestingly, we find that the resulting objective function is free of spatial autocorrelation, unless the two samples i and j are connected with each other, by either one or two edges. This suggests that the disconnected pairs should carry little information about spatial autocorrelation and can be ignored for parameter estimation. The consequence is that a tremendous amount of computation can be saved, because the dominating portion of the paired samples are disconnected. We sum the approximated log likelihood functions over all the connected pairs and then maximize the summation with respect to spatial autocorrelation. The resulting estimator enjoys an elegant analytical solution and is referred to as a paired maximum likelihood estimator (PMLE).

Even though PMLE is inspired by the idea of AMLE under the temporary assumption that the spatial autocorrelation is small, the consistency and asymptotic normality of PMLE can be rigorously established without such a stringent assumption. Instead, we make use of the fact that most large scale social networks are extremely sparse, so that two sampled nodes can hardly be indirectly connected through unobserved social networks. Under this assumption, we show theoretically that PMLE is $\sqrt{n}$-consistent and asymptotically normal. Compared with AMLE, PMLE is computationally superior. Specifically, the computational complexity of PMLE is linear in the number of observed edges. This makes PMLE particularly attractive for big data analysis and thus can serve as our final solution.

To summarize, we provide in this work the following important contributions to the literature. Chen et al. (2013) documented solid numerical evidence, which shows that spatial autocorrelation can be seriously underestimated, if the method of maximum likelihood is incorrectly applied on sampled network data. However, how to conduct a correct maximum likelihood estimation is less well understood. We then fill the theoretical gap by the method of PMLE with well developed asymptotic theories. This is our first important contribution. Second, do big data have to call for big computation? We argue that this is not always necessary. We believe that big data call for smart computation! This is because for most big data applications the sample size is huge. However, the computational resources available to most researchers and practitioners are limited. It is then of great interest to develop novel method, which is efficient not only statistically but also computationally. This is the spirit of smart computation for big data analysis, and leads to the development of PMLE for network data analysis, which is probably one of the most typical and important types of big data analysis. Then, the spirit of "big data but smart computation" is our second important contribution.

The rest of the paper is organized as follows. Section 2 presents the model setup and the approximate likelihood theory. This leads to the method of AMLE, which further inspires the PMLE method, whose asymptotic theory is rigorously established. To demonstrate its finite sample performances, numerical studies based on both simulated and real datasets are conducted in Section 3. Lastly, the article is concluded with a short discussion in Section 4. All technical details are left to the Appendix.

2. THE METHODOLOGY

2.1. Model Setup

We consider a large network with N nodes. Its structure is captured by a network adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, where $a_{ij} = 1$ if the node i follows the node j and $a_{ij} = 0$ otherwise. For each i, we observe a continuous response Y_i. Due to the existence of spatial autocorrelation, the responses of those connected nodes are expected to be correlated with each other. To model such an interactive dependence structure, the following spatial autoregression model has been popularly used (Ord, 1975; Anselin, 1980; Bronnenberg and Mahajan, 2001; Lee et al., 2013).

$$\mathbb{Y} = \rho W \mathbb{Y} + \varepsilon, \tag{2.1}$$

where $\rho \in \mathbb{R}^1$ is referred to as spatial autoregression parameter (Banerjee et al., 2004), $\mathbb{Y} = (Y_1, \cdots, Y_N)^\top \in \mathbb{R}^N$ is the response vector, $W = (w_{ij}) \in \mathbb{R}^{N \times N}$ with $w_{ij} = a_{ij}/d_i$ and $d_i = \sum_{j=1}^N a_{ij}$ is the normalized adjacency matrix, and $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_N)^\top \in \mathbb{R}^N$ is the residual vector with mean 0 and covariance $\sigma^2 I \in \mathbb{R}^{N \times N}$. Here I stands for a $N \times N$ identity matrix.

By (2.1), we know that $\mathbb{Y} = (I - \rho W)^{-1}\varepsilon$, provided that $I - \rho W$ is invertible. According to Banerjee et al. (2004), we know that the largest singular value of W is 1. As a result, in order to ensure the invertibility of $(I - \rho W)$ for an arbitrary W matrix, we must require $|\rho| < 1$. Otherwise, there always exits a possibility to find a W matrix such that $(I - \rho W)$ is singular. We thus follow Banerjee et al. (2004) and assume $|\rho| < 1$ throughout the rest of this article. This implies that $\mathbb{Y}$ follows a normal distribution with mean 0 and covariance

$$\Sigma = (\sigma_{ij}) = \sigma^2 (I - \rho W)^{-1} (I - \rho W^\top)^{-1}. \tag{2.2}$$

Obtaining $\mathbb{Y}$ and W need to have the whole network observed, which is practically infeasible in most cases. Instead, one can draw a random sample of size n from $\mathcal{S}_F = \{1, 2, \cdots, N\}$. Without loss of generality, we assume that the first n nodes are randomly selected from $\mathcal{S}_F$ and are collected by $\mathcal{S} = \{1, 2, \cdots, n\}$. Accordingly, the observed response vector is $\mathbb{Y}_1 = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ and the observed network adjacency matrix is $A_{11} = (a_{ij} : 1 \leq i, j \leq n) \in \mathbb{R}^{n \times n}$. As we mentioned before, we also assume that the degree number of each node is observed, that is $D_1 = (d_i : 1 \leq i \leq n) \in \mathbb{R}^n$. Accordingly, the normalized adjacency matrix $W_{11} = (w_{ij} : 1 \leq i, j \leq n) \in \mathbb{R}^{n \times n}$ is observed. Define $\mathbb{Y}_2 = (Y_{n+1}, \cdots, Y_N)^\top \in \mathbb{R}^{N-n}$, which collects the responses from those unsampled nodes. Thus, $\mathbb{Y} = (\mathbb{Y}_1^\top, \mathbb{Y}_2^\top)^\top \in \mathbb{R}^N$. The matrix A is partitioned accordingly as $A = (A_{11}, A_{12}; A_{21}, A_{22})$. Similarly, W and I can be partitioned as $W = (W_{11}, W_{12}; W_{21}, W_{22})$ and $I = (I_{11}, O_{12}; O_{21}, I_{22})$.

Subsequently, we need to estimate the unknown parameter ρ based on observed response $\mathbb{Y}_1$ and observed network structure W_{11}. It is noted that $\mathbb{Y}_1$, A_{11}, and W_{11} are observed. In contrast, $\mathbb{Y}_2$, A_{12}, A_{21}, A_{22}, W_{12}, W_{21}, and W_{22} are not.

As pointed out by Wall (2004), the interpretation about ρ is not immediately straightforward. By (2.2) we know that the actual spatial covariance (i.e., σ_{ij}) depends on both ρ and W. As a result, ρ cannot be easily interpreted unless the network structure W is fixed. This immediately suggests that comparing ρ across different networks is not desirable. With a fixed W and assuming $|\rho| < 1$, the following Taylor's expansion can be justified

$$\Sigma = (\sigma_{ij}) = \sigma^2 \Big(\sum_{k=0}^{\infty} \rho^k W^k \Big) \Big\{ \sum_{k=0}^{\infty} \rho^k (W^\top)^k \Big\} = \sigma^2 \sum_{m=0}^{\infty} \rho^m \Big\{ \sum_{k_1+k_2=m}^{k_1,k_2 \geq 0} W^{k_1} (W^\top)^{k_2} \Big\}.$$

Note that all the components involved in W (and also $W^\top$) are nonnegative. This suggests that σ_{ij} (for two arbitrary nodes $i \neq j$) should be a monotonically increasing function in ρ, if the network structure W is fixed and ρ is nonnegative. Consequently, ρ can be more precisely interpreted if the following three conditions are satisfied simultaneously. They are, respectively, (1) nonnegative ρ value, (2) a fixed network structure W, and (3) a given node pair (i, j). Under these three conditions, larger size in ρ does lead to larger spatial covariance. Otherwise, the interpretation could be much more complicated. See for example Figure 5 on page 320 (Wall, 2004) for some counterintuitive but illuminating discussion.

2.2. Approximate Maximum Likelihood

By model (2.1) and (2.2), we can define $\Omega = \Sigma^{-1} = \sigma^{-2}(I - \rho W^\top)(I - \rho W) = \sigma^{-2}(\Omega_{11}, \Omega_{12}; \Omega_{21}, \Omega_{22})$, where

$$
\begin{aligned}
\Omega_{11} &= I_{11} - \rho(W_{11} + W_{11}^\top) + \rho^2(W_{11}W_{11}^\top + W_{21}W_{21}^\top),\\
\Omega_{12} &= -\rho(W_{21}^\top + W_{12}) + \rho^2(W_{12}W_{11}^\top + W_{22}W_{21}^\top),\\
\Omega_{21} &= -\rho(W_{21} + W_{12}^\top) + \rho^2(W_{11}W_{12}^\top + W_{21}W_{22}^\top),\\
\Omega_{22} &= I_{22} - \rho(W_{22} + W_{22}^\top) + \rho^2(W_{12}W_{12}^\top + W_{22}W_{22}^\top).
\end{aligned}
$$

Note that $\mathrm{cov}(\mathbb{Y}_1) = \Sigma_{11}$, which is the first $n \times n$ diagonal block matrix of Σ. We then have $\Sigma_{11}^{-1} = \sigma^{-2}(\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21})$. Unfortunately, Σ_{11} is not practically computable, because it involves Ω_{22}, which is an unobserved and huge sized matrix with dimension $N-n$. In the meanwhile, Σ_{11}^{-1} is a function of ρ. Under the model assumption $|\rho| < 1$, one can verify that Σ_{11}^{-1} has a Taylor's expansion as $\Sigma_{11}^{-1} = \sum_{k=0}^{\infty} \rho^k \Sigma_{11}^{(k)} \approx \sum_{k=0}^{K} \rho^k \Sigma_{11}^{(k)}$, where K is some pre-specified approximation order and $\Sigma_{11}^{(k)}$ is some matrix-valued derivative. Obviously, larger K leads to better approximation. However, it also calls for substantially increased sampling efforts. Thus, practically it is appealing to consider $K = 1$ as follows.

$$
\begin{aligned}
\sigma^2\Sigma_{11}^{-1} &= \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}\\
&= I_{11} - \rho(W_{11} + W_{11}^\top) + \rho^2(W_{11}W_{11}^\top + W_{21}W_{21}^\top)\\
&\quad -\rho^2\Big\{(W_{21}^\top + W_{12}) + \rho(W_{12}W_{11}^\top + W_{22}W_{21}^\top)\Big\}\Omega_{22}^{-1}\\
&\quad \Big\{(W_{21} + W_{12}^\top) + \rho(W_{11}W_{12}^\top + W_{21}W_{22}^\top)\Big\}\\
&= I_{11} - \rho(W_{11} + W_{11}^\top) + \rho^2(W_{11}W_{11}^\top + W_{21}W_{21}^\top)\\
&\quad -\rho^2(W_{21}^\top W_{21} + W_{21}^\top W_{12}^\top + W_{12}W_{21} + W_{12}W_{12}^\top) + \sum_{k>2}\rho^k\Sigma_{11}^{(k)},
\end{aligned}
$$

where the last equality is due to the fact that $\Omega_{22} \approx I_{22}$. Accordingly, we know that $\Sigma_{11}^{(1)} = (W_{11} +$

$W_{11}^\top$). This leads to the following first order approximation

$$\sigma^2\Sigma_{11}^{-1} \approx I_{11} - \rho(W_{11} + W_{11}^\top). \tag{2.3}$$

Note that the expression (2.3) is indeed the first order approximation of the sub-matrix of $\sigma^2\Sigma^{-1}$ corresponding to $\mathbb{Y}_1$. Surprisingly, we find that only W_{11} is involved in this approximation (2.3) while other network structures (e.g., W_{12}, W_{21}, and W_{22}) are not. This implies that the first order approximation of Σ_{11}^{-1} with respect to ρ is practically computable, even though Σ_{11}^{-1} itself is not. Accordingly, the corresponding approximation (after negative two-times log transformation) should be computable. It is given by

$$\log\left|I_{11} - \rho(W_{11} + W_{11}^\top)\right| - \sigma^{-2}\mathbb{Y}_1^\top\left\{I_{11} - \rho(W_{11} + W_{11}^\top)\right\}\mathbb{Y}_1 - n\log\sigma^2.$$

Fix ρ and optimize the above objective function with respect to σ^2. This leads to $\hat\sigma^2 = n^{-1}\mathbb{Y}_1^\top\{I_{11} - \rho(W_{11} + W_{11}^\top)\}\mathbb{Y}_1$. We then replace σ^2 in (2.3) by $\hat\sigma^2$. This gives a profiled objective function as

$$\begin{aligned} & \log\left|I_{11} - \rho(W_{11} + W_{11}^\top)\right| - n\log\left[n^{-1}\mathbb{Y}_1^\top\{I_{11} - \rho(W_{11} + W_{11}^\top)\}\mathbb{Y}_1\right] \\ = \; & \log\left|I_{11} - \rho(W_{11} + W_{11}^\top)\right| - n\log\left[1 - \rho n^{-1}\hat\sigma_Y^{-2}\mathbb{Y}_1^\top(W_{11} + W_{11}^\top)\mathbb{Y}_1\right] \\ \approx \; & \log\left|I_{11} - \rho(W_{11} + W_{11}^\top)\right| + \rho\hat\sigma_Y^{-2}\mathbb{Y}_1^\top(W_{11} + W_{11}^\top)\mathbb{Y}_1, \end{aligned}$$

where the constants independent of ρ are omitted, $\hat\sigma_Y^2 = n^{-1}\mathbb{Y}_1^\top\mathbb{Y}_1$, and the last approximation is due to Taylor's expansion and the temporary assumption that ρ is small. For a practical dataset, we can always assume that the data have been standardized so that $\hat\sigma_Y^2 = 1$. This leads to the following highly simplified objective function

$$\ell_a(\rho) = \log\left|I_{11} - \rho(W_{11} + W_{11}^\top)\right| + \rho\mathbb{Y}_1^\top(W_{11} + W_{11}^\top)\mathbb{Y}_1. \tag{2.4}$$

It is noted that $\ell_a(\rho)$ in (2.4) is not constructed based on the accurate likelihood function of $\mathbb{Y}_1$.

Instead, it is obtained by its first order approximation. We thus call the resulting estimator, denoted by $\hat{\rho}_a = \mathrm{argmax}\ell_a(\rho)$, as an approximate maximum likelihood estimator (AMLE).

2.3. Paired Maximum Likelihood

As one can note, optimizing AMLE is not cheap. It is mainly because $I_{11} - \rho(W_{11} + W_{11}^\top)$ is a $n \times n$ matrix, whose determinant needs to be computed. This is not a problem if the sample size n is small or moderate. However, it could be a serious burden if n is large. This motivates us to use AMLE as an intermediate step to further inspire an estimator, which is computationally superior.

Specifically, we consider an extreme situation with only two nodes (denoted by i and j). In that case, the approximation (2.3) is still valid with $W_{11} = (0, w_{ij}; w_{ji}, 0) \in \mathbb{R}^{2\times 2}$. Accordingly, the objective function (2.4) can be used. We follow the idea of composite likelihood (Shao, 2003) and sum together all the paired objective functions. This leads to

$$\sum_{i,j} \log\Big\{1 - \rho^2(a_{ij}/d_i + a_{ji}/d_j)^2\Big\} + 2\rho \sum_{i,j} Y_i Y_j (a_{ij}/d_i + a_{ji}/d_j).$$

It is noted that $a_{ij} = a_{ji} = 0$ for those disconnected pairs, and thus the corresponding quantity is free of the spatial autocorrelation ρ. As a result, those disconnected pairs can be ignored and the above quantity can be simplified as

$$\sum_{a_{ij}+a_{ji}>0} \log\Big\{1 - \rho^2(a_{ij}/d_i + a_{ji}/d_j)^2\Big\} + 2\rho \sum_{a_{ij}+a_{ji}>0} Y_i Y_j (a_{ij}/d_i + a_{ji}/d_j)$$

$$\approx -\rho^2 \sum_{a_{ij}+a_{ji}>0} (a_{ij}/d_i + a_{ji}/d_j)^2 + 2\rho \sum_{a_{ij}+a_{ji}>0} Y_i Y_j (a_{ij}/d_i + a_{ji}/d_j), \tag{2.5}$$

where the approximation is due to Taylor's expansion and the temporary assumption that ρ is small. Then, it is interesting to note that the quantity in (2.5) is a quadratic function in ρ, whose optimizer enjoys an analytical solution given by

$$\hat{\rho}_p = \Big\{ \sum_{a_{ij}+a_{ji}>0} (a_{ij}/d_i + a_{ji}/d_j)^2 \Big\}^{-1} \Big\{ \sum_{a_{ij}+a_{ji}>0} Y_i Y_j (a_{ij}/d_i + a_{ji}/d_j) \Big\}$$

$$= (n\omega_n)^{-1} \sum_{(i,j)\in\mathcal{D}} Y_i Y_j d_{ij},$$

where $\mathcal{D} = \{(i, j) : a_{ij} + a_{ji} > 0\}$ collects all the connected pairs, $d_{ij} = d_{ji} = a_{ij}/d_i + a_{ji}/d_j$, and $\omega_n = n^{-1}\sum_{(i,j)\in\mathcal{D}} d_{ij}^2$. Because $\hat{\rho}_p$ is an estimator obtained by optimizing the paired likelihood function, we refer to it as a paired maximum likelihood estimator (PMLE). Compared with the AMLE $\hat{\rho}_a$, the PMLE $\hat{\rho}_p$ is computationally much more efficient. This is because its computation only involves those connected pairs. This makes PMLE particularly attractive for large scale network data analysis.

2.4. The Asymptotic Properties

For a given N-dimensional square matrix $\mathcal{M} = (m_{i_1 i_2} : 1 \le i_1, i_2 \le N) \in \mathbb{R}^{N\times N}$, we define $\|\mathcal{M}\|_{(n)} = \sum_{i_1,i_2\le n} |m_{i_1 i_2}|$. Note that, even though PMLE was inspired under the assumption ρ is small, its consistency and asymptotic normality are free of such a stringent requirement. They can be rigorously justified under fairly reasonable conditions as given below.

(A1) Law of Large Number. There exists a constant $\omega > 0$ such that $tr(\mathbb{W}^2)/n = n^{-1}\sum_{ij}(a_{ij}/d_i + a_{ji}/d_j)^2 = \omega_n \to \omega$ as $n \to \infty$.

(A2) Network Sparsity. Write $\Delta_{\max} = \max_{k>1}\|W^k\|_{(n)} + \max_{k_1,k_2\ge 1}\|W^{k_1,k_2}\|_{(n)} + \max_{k_1,k_2,k_3,k_4\ge 1}\|W^{k_1,k_2,k_3,k_4}\|_{(n)}$, where $W^{k_1,k_2} = W^{k_1}(W^{k_2})^\top$ and $W^{k_1,k_2,k_3,k_4} = W^{k_2}(W^{k_3})^\top W^{k_4}(W^{k_1})^\top$. As $n \to \infty$, we require that $\Delta_{\max} = o(n^{1/2})$.

We argue that all those conditions are quite intuitive and reasonable. The detailed explanations are given below.

Condition (A1) is a Law of Large Number type condition. It requires that the sampled network structure to maintain a reasonable density level. For example, every node should be involved in at least one edge. Otherwise, the network structure could too sparse (e.g., a network with no edge). In that case, we should have $\omega = 0$. Obviously, a network without enough edges (i.e., too sparse) cannot provide sufficient information about spatial autocorrelation. As a result, condition (A1) warrants the number of edges observed in the network also diverges to infinity as $n \to \infty$.

Condition (A2) basically requires that the network structure W to be sparse. To see this, consider two arbitrary nodes (e.g., i and j). If the network structure is sparse, then the likelihood for them to be indirectly connected is small. For example, a typical indirect connection with length 2 could be $i \to k \to j$ for some $1 \le k \le N$. In this case, we should expect $\sum_{k=1}^{N} a_{ik}a_{kj}$ to be small on average for all possible (i, j) pairs. This suggests that $\|W^2\|_{(n)}$ should be small as compared with $\sqrt{n}$. This is typically true if the sampling fraction n/N is very small. Same argument applies to paths with higher order lengths. Thus, we should expect $\max_k \|W^k\|_{(n)}$ to be well controlled. Another typical indirect connection with length 2 could be $i \to k$ and $j \to k$ for some $1 \le k \le N$. In this case, we should have $\sum_{k=1}^{N} a_{ik}a_{jk}$ to be small on average for all possible (i, j) pairs. This suggests that $\|W^{1,1}\|_{(n)}$ should be small. Same argument applies to paths with higher order lengths. This leads to well bounded $\max_{k_1,k_2} \|W^{k_1,k_2}\|_{(n)}$ and $\max_{k_1,k_2,k_3,k_4} \|W^{k_1,k_2,k_3,k_4}\|_{(n)}$. This explains why condition (A2) controls the network sparsity level. Intuitively, if the network is sufficiently sparse, then the sampled edges should be mostly important for explaining the spatial autocorrelation. As a result, if the sampled edges are used appropriately, spatial autocorrelation should be estimated consistently.

Theorem 1. *Assume (A1) and (A2), then* $\sqrt{n}(\hat{\rho}_p - \rho) \xrightarrow{d} N(0, 2/\omega)$ *as* $n \to \infty$.

By Theorem 1, we find that the asymptotic variance of PMLE is analytically extremely simple and elegant. It is noted that the asymptotic variance ω can be easily estimated by ω_n, which is just a function of the observed network structure W_{11}. This makes the practical inference simple.

3. NUMERICAL STUDIES

3.1. Simulating Network Data

To evaluate the finite sample performance of the proposed methods, we present here a number of simulation studies. For a fixed N, the network adjacency matrix $A = (a_{ij})$ is simulated as follows. First, generate N independent and identically distributed random variables according to an exponential distribution with mean 10. Denote these variables by E_i with $1 \leq i \leq N$. For each node i, we randomly select a sample size of $[E_i]$ from $\mathcal{S}_F = \{1, 2, \cdots, n\}$ without replacement, where $[E_i]$ stands for the smallest integer no less than E_i. Denote the sample by $\mathcal{S}_i$. Define $a_{ij} = 1$ if $j \in \mathcal{S}_i$ and $a_{ij} = 0$ otherwise. In the third step, we force $a_{ij} = a_{ji}$ for every $i < j$. In the fourth step, we re-define $a_{ij} = d_{ij}a_{ij}$, where d_{ij}s are independent binary random variables with $P(d_{ij} = 1) = 0.5$. Lastly, let $a_{ii} = 0$ for every $1 \leq i \leq N$. This leads to the final adjacency matrix A. Subsequently, W can be computed by normalizing each row of A. Thereafter, W is fixed throughout the rest of the simulation study. For a reliable evaluation, each experiment is randomly replicated $M = 1000$ times. For each random replication, the response is generated according to $\mathbb{Y} = (I - \rho W)^{-1}\varepsilon$, where $\varepsilon \in \mathbb{R}^N$ is simulated from a N-dimensional standard normal random vector. This leads to the whole network data W and $\mathbb{Y}$.

3.2. PMLE

In this study, we fix $\rho = 0$ or 0.2. Various combinations of N and n are considered. For each combination, we fix sampling proportion of n/N to be equal to 10%. Once W and $\mathbb{Y}$ are simulated, a random sample size of n is obtained. Based on the sampled data, PMLE is computed. Its estimated standard error (SE) is also obtained as $\widehat{\mathrm{SE}} = \sqrt{2}\omega_n^{-1/2}n^{-1/2}$. Denote the estimator obtained in the mth simulation replication ($1 \leq m \leq M$.) by $\hat{\rho}^{(m)}$, and the corresponding SE estimate by $\widehat{\mathrm{SE}}^{(m)}$. Then the bias is evaluated as $\flat = \rho - \bar{\rho}$ with $\bar{\rho} = M^{-1}\sum_{m=1}^{M}\hat{\rho}$, and the true SE as $\mathrm{SE} = \{M^{-1}\sum_{m=1}^{M}(\hat{\rho}^{(m)} - \bar{\rho})^2\}^{1/2}$. We also compute the averaged SE estimate (i.e., $\widehat{\mathrm{SE}}$) as $M^{-1}\sum_{m=1}^{M}\widehat{\mathrm{SE}}^{(m)}$. With the estimated SE, the statistical significance of the spatial autocorrelation can be tested. Specifically, for each simulation iteration, a Z-type test statistic is constructed as $Z^{(m)} = \hat{\rho}^{(m)}/\widehat{\mathrm{SE}}^{(m)}$. For a given significance level $\alpha = 5\%$, we reject the null hypothesis of $H_0 : \rho = 0$ if $|Z^{(m)}| > z_{1-\alpha/2}$, where z_α stands for the αth quantile of a standard normal distribution. Accordingly, we summarize the empirical rejection probability (ERP) as $\mathrm{ERP} = M^{-1}\sum I(|Z^{(m)}| > z_{1-\alpha/2})$. Theoretically, ERP corresponds to the empirical size if $\rho = 0$ and power if $\rho \neq 0$.

Detailed results are summarized in Table 1, from where we can draw the following two conclusions. First, PMLE is consistent, with both bias and SE decreasing towards 0 as $N \to \infty$ and $n \to \infty$, regardless of ρ. Additionally, the estimated SE (i.e., $\widehat{\mathrm{SE}}$) approximates the true SE quite well, because their average values are very close to each other. Secondly, the reported ERP values are fairly close to their nominal level $\alpha = 5\%$ for $\rho = 0$. This suggests that the implemented Z-type test can control Type I error well. On the other side, the reported ERP values steadily increases towards 100% as $N \to \infty$ and $n \to \infty$ if $\rho = 0.2$. This confirms that the proposed Z-type test has a reasonable power.

3.3. Sampling Method

In this study, we compare different sampling methods and then evaluate their impact on PMLE accuracy. Data are generated in the same way as the previous subsection but with a fix $N = 100,000$. By Theorem 1, the asymptotic efficiency of PMLE is fully determined by the network structure through the quantity $\omega \approx \omega_n = n^{-1}\sum(a_{ij}/d_i + a_{ji}/d_j)^2$. A larger ω_n value implies better estimation accuracy. Thus, a good sampling method should maximize the number of observed edges (i.e., a_{ij} and a_{ji}). Obviously, the method of simple random sampling (SRS) without replacement (Thompson, 2012), is unlikely to be the optimal choice. Instead, a snowball type sampling method might be a good alternative. Here we investigate one particular type of snowball sampling method. It is an iterative method. In each iterative step, one seed node (e.g., i) is randomly selected and all its connected friends (i.e., every j satisfying $a_{ij} = 1$) are collected. Both the sampled seed node and its connected friends are accumulated. If the current accumulated sample size is still below the target n, the above iterative sampling process should be repeated. Otherwise, some sampled nodes are randomly dropped so that the final sample size is exactly n. For convenience, we refer to this sampling method as SNOW.

With a slight abuse of notation, we use $\hat{\rho}^{(m)}$ to denote PMLE obtained in the mth iteration corresponding to one particular sampling method (i.e., SRS and SNOW). We then evaluate its estimation accuracy by mean squared error MSE $= M^{-1}\sum_{m=1}^{M}(\hat{\rho}^{(m)}-\rho)^2$. Various choices of n are considered. The resulting MSE are plotted in Figure 1 in log-scale, from where we can draw the following observations. First, regardless of the sampling method, PMLE is consistent, because the log MSE value monotonically decreases as the sample size increases. Second, compared with SRS, SNOW offers a significant improvement in terms of estimation accuracy. The difference exhibited by SRS and SNOW in terms of log MSE can be as large as 1.2 approximately for example $n = 10,000$. This suggests that sampling method do play an important role for spatial autocorrelation estimation. SNOW should be a useful method for network sampling.

3.4. Sina Weibo Network Analysis

As our last numerical study, we present here a real network example about Sina Weibo (*www.weibo.com*), which can be viewed as a Twitter-type social media in Chinese community. The objective of this study is to understand how the users of Sina Weibo interact with each other in terms of their posting activity. For illustration purpose, we start with the Weibo accounts of four major online travel agencies in mainland China. They are respectively: CTRIP (*www.ctrip.com*), ELONG (*www.elong.com*), MangoCity (*wwww.mangocity.co*), and QUNAR (*www.qunar.com*). For each travel agency, we randomly select 5,000 nodes from their followers. Subsequently, those followers' followers are also collected. Because condition (A2) is better satisfied by sparse network, this motivates us to keep only those active users with relatively small degree numbers. This gives the final network size $N = 557,818$. Their follower-followee relationships (i.e., A) are recorded. This is then treated as our whole network with a total of $\sum a_{ij}$ =1,496,399 edges and $\sum_{i<j} a_{ij}a_{ji} = 535,408$ mutually connected pairs. For each node, we define the response as the number of its posted messages in log-scale. The responses are standardized so that its mean is 0 and variance is 1. With such a large network size, obtaining the maximum likelihood estimator or its approximate (i.e., AMLE) is extremely difficult. However, the PMLE can be readily computed by using a personal computer without much difficulty. It gives $\hat{\rho}_p = 0.154$ with $\widehat{\text{SE}} = 1.55 \times 10^{-3}$. We thus conclude that the estimated spatial autocorrelation is statistically significant at 5% level. This implies that a Sina Weibo user's posting activity does correlated with each other in a nontrivial way.

Given the whole network data, we next conducted a real data based simulation study to check the effect of sampling on the subsequent inferences. The study is implemented in a similar manner as in Section 3.2. However, the difference is that the whole network data (i.e., W and $\mathbb{Y}$) are not generated by simulation. Instead, they are directly derived from Sina Weibo. As a result, the response values are fixed for each node across different simulation iteration. The method of SRS is used. It is noted that the true spatial autocorrelation coefficient of this real data is unknown. We then treat the PMLE computed based on the whole network data (given in the previous paragraph) as if it were the true parameter. This gives us $\rho = 0.154$. We are able to do this because the sample size considered subsequently is considerably smaller than the network size $N = 557,818$. The detailed results are given in Table 2. However, it should be noted that the interpretation of SE in Table 2 should be slightly different from that of Table 1 in Section 3.2. The reason is the following. For this real data based simulation study, the response values are fixed for each node across different simulation iterations. The randomness due to response value regeneration was not involved. Consequently, the randomness of PMLE is fully due to sampling. As a result, the SE values reported in Table 2 should be interpreted as a standard error measure for $\hat{\rho}$ when only sampling randomness is involved. Then, the ERP values in Table 2 should be interpreted similarly.

By Table 2, we find that, in order to have about 95% power (i.e., ERP$\approx$ 95%), only $n = 20,000$ nodes need to be sampled. It accounts for about $n/N = 3.59\%$ of the entire network size.

4. CONCLUSION

We investigate here the problem of spatial autocorrelation estimation based on sampled network data. To capture spatial autocorrelation, the classical spatial autoregression model is considered. We find that the exact maximum likelihood estimator for the sampled data is practically infeasible when the network size is large. To fix the problem, a novel approximation method was proposed, leading to the method of AMLE, which further inspires the development of the PMLE method. These findings are also confirmed by numerical studies. We further illustrate our methods by a real dataset about Sina Weibo. Significant spatial autocorrelation in terms of posting activity is detected.

To conclude the article, we discuss here a number of interesting topics for future study. First, the spatial autoregression model (2.1) considers only those directly connected nodes for autoregression. Instead, empirical research provides evidence that those indirectly connected ones might also have impact on each other. Naturally, this calls for spatial autoregression models with higher order neighbors. Approximation of the likelihood in this general setup is a nontrivial extension of our proposed method and deserves a separate study. Second, by our theoretical analysis, the quantity ω_n determines the asymptotic efficiency and its value can be much improved by SNOW. However, how much SNOW can be further improved is not clear. Third, as in the literature of spatial statistics and econometrics, the current model assumes that the adjacency matrix (therefore the weight matrix) is pre-determined before the response is generated. However, in many cases the adjacency matrix is endogeneously determined, because nodes sharing common features are more likely to be connected. Then, how to model this endogeneous phenomenon is another important topic deserved further investigation.

APPENDIX

Appendix A. A Useful Lemma

To establish the asymptotic normality for PMLE, we need a Central Limit Theorem for normal quadrature. We thus first state and prove a useful lemma in this regards.

Lemma 1. *Let $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top \in \mathbb{R}^n$ be a n-dimensional standard normal random vector. Let $Q = (q_{ij}) \in \mathbb{R}^{n\times n}$ be a symmetric matrix satisfying that $\lambda_{\max}(Q) \le c_{\max}$ and $n^{-1}tr(Q^2) > c_{\min}$ as $n \to \infty$, for some positive constants $c_{\max} > 0$ and $c_{\min} > 0$. Define $\mathcal{Q}_n = \epsilon^\top Q\epsilon$, then $\{\mathcal{Q}_n - tr(Q)\}/tr^{1/2}(2Q^2) \to_d N(0,1)$ as $n \to \infty$.*

Proof. Because $Q \in \mathbb{R}^{n\times n}$ is a symmetric matrix. There exists an orthonormal matrix $U \in \mathbb{R}^{n\times n}$ (i.e., $U^\top U = UU^\top = I$), such that $Q = U^\top DU$ with $D = \mathrm{diag}\{\lambda_1, \cdots, \lambda_n\}$ being a diagonal matrix. We then have $\mathcal{Q}_n = \epsilon^\top Q\epsilon = Z^\top DZ = \sum \lambda_i Z_i^2$, where $Z = (Z_i) = U\epsilon \in \mathbb{R}^n$ and Z_is are independent standard normal random variables. We know immediately $E(\mathcal{Q}_n) = \sum \lambda_i = tr(Q)$ and $\mathrm{var}(\mathcal{Q}_n) = 2\sum \lambda_i^2 = 2tr(Q^2) \ge 2nc_{\min}$. Furthermore, we know that $\sum E(\lambda_i Z_i^2 - \lambda_i)^4 = (\sum \lambda_i^4)E(Z_i^2 - 1)^4 = E(Z_i^2 - 1)^4 tr(Q^4) \le E(Z_i^2 - 1)^4 (nc_{max}^4)$. Consequently, we know that $\sum E(\lambda_i Z_i^2 - \lambda_i)^4/\{\mathrm{var}(\mathcal{Q}_n)\}^2 \to 0$ as $n \to \infty$. This verifies the Lyapunov condition. As a result, the Central Limit Theorem can be established, that is $\{\mathcal{Q}_n - tr(Q)\}/tr^{1/2}(2Q^2) \to_d N(0,1)$ as $n \to \infty$. This proves the lemma conclusion.

Appendi B. Proof of Theorem 1

We are going to establish the theorem in three steps. In the 1st step, we show that $\hat{\rho}_p$ is asymptotically unbiased. In the 2nd step, we obtain its asymptotic variance. Lastly, we establish the asymptotic normality in Step 3.

Step 1 (Bias). Recall $\mathbb{Y} = (I - \rho W)^{-1}\varepsilon$. In the meanwhile, by Banerjee et al. (2004) we know that $\lambda_{\max}(W) \leq 1$. Furthermore, due to constraint $|\rho| < 1$, we can obtain $\mathbb{Y} = (I + \sum_{k=1}^{\infty} \rho^k W^k)\varepsilon$. Write $W^k = (w_{ij}^{(k)})$. We then have

$$w_{ij}^{(k)} = \sum_{s_1 s_2 \cdots s_{k-1}} \left(\frac{a_{i_1 s_1}}{d_{i_1}}\right)\left(\frac{a_{s_1 s_2}}{d_{s_1}}\right)\cdots\left(\frac{a_{s_{k-2} s_{k-1}}}{d_{s_{k-2}}}\right)\left(\frac{a_{s_{k-1} j}}{d_{s_{k-1}}}\right).$$

Then, for any sampled node $1 \leq i \leq n$, we should have $Y_i = \varepsilon_i + \sum_{k=1}^{\infty} \rho^k \sum_{j=1}^{N} w_{ij}^{(k)} \varepsilon_j = \varepsilon_i + \sum_{j=1}^{N} \varepsilon_j (\sum_{k=1}^{\infty} \rho^k w_{ij}^{(k)})$. Consider two arbitrarily sampled nodes i and j, then

$$\sigma^{-2} E(Y_i Y_j) = \sum_{k=1}^{\infty} \rho^k \Big(w_{ij}^{(k)} + w_{ji}^{(k)}\Big) + \sum_{s} \Big(\sum_{k \geq 1} \rho^k w_{is}^{(k)}\Big)\Big(\sum_{k \geq 1} \rho^k w_{js}^{(k)}\Big)$$

$$= \rho d_{ij} + \sum_{k=2}^{\infty} \rho^k \Big(w_{ij}^{(k)} + w_{ji}^{(k)}\Big) + \sum_{k_1, k_2 \geq 1} \rho^{k_1 + k_2} \Big(\sum_{s} w_{is}^{(k_1)} w_{js}^{(k_2)}\Big).$$

Therefore, we have $\sum_{ij} E(Y_i Y_j) d_{ij} = \rho \sum_{i,j \leq n} d_{ij}^2 + \mathcal{O}_1$. Here $\mathcal{O}_1 \geq 0$ is a nonnegative quantity given by

$$\mathcal{O}_1 = \sum_{i,j \leq n} d_{ij} \sum_{k=2}^{\infty} \rho^k \Big(w_{ij}^{(k)} + w_{ji}^{(k)}\Big) + \sum_{i,j \leq n} d_{ij} \sum_{k_1, k_2 \geq 1} \rho^{k_1 + k_2} \Big(\sum_{s} w_{is}^{(k_1)} w_{js}^{(k_2)}\Big)$$

$$\leq 2 \sum_{i,j \leq n} \sum_{k=2}^{\infty} \rho^k \Big(w_{ij}^{(k)} + w_{ji}^{(k)}\Big) + 2 \sum_{i,j \leq n} \sum_{k_1, k_2 \geq 1} \rho^{k_1 + k_2} \Big(\sum_{s} w_{is}^{(k_1)} w_{js}^{(k_2)}\Big),$$

because $d_{ij} \leq 2$ and all other involved quantities are nonnegative. We can then further write the right hand side of the above inequality as

$$= 2 \sum_{k=2}^{\infty} \rho^k \sum_{i,j \leq n} \Big(w_{ij}^{(k)} + w_{ji}^{(k)}\Big) + 2 \sum_{k_1, k_2 \geq 1} \rho^{k_1 + k_2} \sum_{i,j \leq n} \Big(\sum_{s} w_{is}^{(k_1)} w_{js}^{(k_2)}\Big)$$

$$= 4 \sum_{k \geq 2} \rho^k \|W^k\|_{(n)} + 2 \sum_{k_1, k_2 \geq 1} \rho^{k_1 + k_2} \|W^{k_1, k_2}\|_{(n)}.$$

By (A2), the right hand side of the above equality can be further bounded by

$$\leq 4\Delta_{\max}\sum_{k\geq 2}\rho^k + 2\Delta_{\max}\sum_{k_1,k_2\geq 1}\rho^{k_1+k_2} = 4\Delta_{\max}\left(\frac{\rho^2}{1-\rho}\right) + 2\Delta_{\max}\Big(\sum_{k\geq 1}\rho^k\Big)^2$$

$$= 4\Delta_{\max}\left(\frac{\rho^2}{1-\rho}\right) + 2\Delta_{\max}\left(\frac{\rho}{1-\rho}\right)^2 = \left(\Delta_{\max}\rho^2\right)\cdot O(1).$$

Recall $\hat{\rho}_p = (\sum_{1\leq i,j\leq n} d_{ij}Y_iY_j)/(\sum_{i,j} d_{i,j}^2)$. We thus have $E(\hat{\rho}_p) = \rho + o(n^{-1/2})$ by Conditions (A1) and (A2).

Step 2 (Variance). Write $\mathbb{W} = (W_{11} + W_{11}^\top)$. One can then verify that $\hat{\rho}_p = \mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1/tr(\mathbb{W}^2)$. Consequently, we know that $\text{var}(\hat{\rho}_p) = \text{var}(\mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1)/tr^2(\mathbb{W}^2)$. As a result, the key is to obtain an analytically tractable formula for $\text{var}(\mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1)$. To this end, recall that $\mathbb{Y} = (I-\rho W)^{-1}\varepsilon = \sum_{k\geq 0}\rho^k W^k\varepsilon$. Consequently, $\mathbb{Y}_1 = \sum_{k\geq 0}\rho^k(W^k)_{(n)}\varepsilon$, where $(W^k)_{(n)} = (w_{ij}^{(k)} : 1\leq i\leq n, 1\leq j\leq N) \in \mathbb{R}^{n\times N}$. We can then write $\mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1 = \varepsilon^\top\mathbb{M}\varepsilon$, where

$$\mathbb{M} = \Big\{\sum_{k\geq 0}\rho^k(W^k)_{(n)}^\top\Big\}\mathbb{W}\Big\{\sum_{k\geq 0}\rho^k(W^k)_{(n)}\Big\}.$$

Because $\sigma^{-1}\varepsilon$ follows a multivariate standard normal distribution, $\mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1$ is distributed as a weighted chi-square and its variance is given by $2\sigma^4 tr(\mathbb{M}^2)$. We next compute $tr(\mathbb{M}^2)$. Direct algebra reveals that

$$tr(\mathbb{M}^2) = \sum_{k_1,k_2,k_3,k_4\geq 0}\rho^{k_1+k_2+k_3+k_4}tr\Big\{(W^{k_1})_{(n)}^\top\mathbb{W}(W^{k_2})_{(n)}(W^{k_3})_{(n)}^\top\mathbb{W}(W^{k_4})_{(n)}\Big\}$$

$$= tr(\mathbb{W}^2) + \sum_{k_1+k_2+k_3+k_4\geq 1}\rho^{k_1+k_2+k_3+k_4}tr\Big\{(W^{k_1})_{(n)}^\top\mathbb{W}(W^{k_2})_{(n)}(W^{k_3})_{(n)}^\top\mathbb{W}(W^{k_4})_{(n)}\Big\}.$$

As a reuslt, we can write $tr(\mathbb{M}^2) = tr(\mathbb{W}^2) + O_2$, where

$$O_2 = \sum_{k_1+k_2+k_3+k_4\geq 1}\rho^{k_1+k_2+k_3+k_4}tr\Big\{(W^{k_1})_{(n)}^\top\mathbb{W}(W^{k_2})_{(n)}(W^{k_3})_{(n)}^\top\mathbb{W}(W^{k_4})_{(n)}\Big\}$$

Note that $\mathbb{W} = (d_{ij}) = (a_{ij}/n_i + a_{ji}/n_j) \in \mathbb{R}^{n\times n}$ with $0 \le d_{ij} \le 2$ for any $1 \le i, j \le n$. On the other side, all the components involved in W^k are positive for any $k > 0$. This implies that the right hand side of the above quality can be bounded by

$$\le 4 \sum_{k_1+k_2+k_3+k_4\ge 1} \rho^{k_1+k_2+k_3+k_4} tr\Big\{(W^{k_1})^\top_{(n)}(W^{k_2})_{(n)}(W^{k_3})^\top_{(n)}(W^{k_4})_{(n)}\Big\}$$

$$= 4 \sum_{k_1+k_2+k_3+k_4\ge 1} \rho^{k_1+k_2+k_3+k_4} tr\Big\{(W^{k_2})_{(n)}(W^{k_3})^\top_{(n)}(W^{k_4})_{(n)}(W^{k_1})^\top_{(n)}\Big\}. \tag{A.1}$$

Because the components involved in W are all nonnegative, we have

$$tr\Big\{(W^{k_2})_{(n)}(W^{k_3})^\top_{(n)}(W^{k_4})_{(n)}(W^{k_1})^\top_{(n)} \in \mathbb{R}^{n\times n}\Big\}$$

$$\le tr\Big\{(W^{k_2})_{(n)}(W^{k_3})^\top(W^{k_4})(W^{k_1})^\top_{(n)} \in \mathbb{R}^{n\times n}\Big\}. \tag{A.2}$$

Note that $(W^{k_2})_{(n)}(W^{k_3})^\top(W^{k_4})(W^{k_1})^\top_{(n)} \in \mathbb{R}^{n\times n}$ happens to be the corresponding sub-matrix of $W^{k_1,k_2,k_3,k_4} = W^{k_2}(W^{k_3})^\top W^{k_4}(W^{k_1})^\top \in \mathbb{R}^{N\times N}$. Once again, because the components of W are all nonnegative, we have the trace in (A.2) is bounded by $\|W^{k_1,k_2,k_3,k_4}\|_{(n)}$. This suggests that the right hand side of (A.1) can be bounded by

$$\le 4 \sum_{k_1+k_2+k_3+k_4\ge 1} \rho^{k_1+k_2+k_3+k_4}\|W^{k_1,k_2,k_3,k_4}\|_{(n)}$$

$$\le 4\Delta_{\max} \sum_{k_1+k_2+k_3+k_4\ge 1} \rho^{k_1+k_2+k_3+k_4} \le 4\Delta_{\max}\left(\frac{1}{1-\rho}\right)^4 = o(n^{1/2}),$$

by technical condition (A2). On the other hand, by (A1) we know that $tr(\mathbb{W}^2)$ is of the order n. Consequently, we know that $tr(\mathbb{M}^2) = tr(\mathbb{W}^2) + o(n^{1/2}) = tr(\mathbb{W}^2)\{1 + o(n^{1/2})tr^{-1}(\mathbb{W}^2)\} = tr(\mathbb{W}^2)\{1 + o(n^{-1/2})\} = tr(\mathbb{W}^2)\{1 + o(1)\}$. This further implies that $\text{var}(\hat{\rho}_p) = tr^{-1}(\mathbb{W}^2)\{1 + o(1)\}$.

Step 3 (Asymptotic Normality). Recall $\hat{\rho}_p = \mathbb{Y}_1^\top \mathbb{W}\mathbb{Y}_1/tr(\mathbb{W}^2)$. As a result, its asymptotic normality is fully determined by $\mathbb{Y}_1^\top\mathbb{W}\mathbb{Y}_1$. One can easily verify that $\lambda_{\max}(\mathbb{W}) \le 2\lambda_{\max}(W_{11}) \le$

$2\lambda_{\max}(W) \leq 2$. By assumption (A1), we have $n^{-1}tr(\mathbb{W}^2) = n^{-1}\sum d_{ij}^2 = \omega_n \to \omega > 0$. As a result, we have $\{\hat{\rho}_p - E(\hat{\rho}_p)\}\text{var}^{-1/2}(\hat{\rho}_p) \xrightarrow{d} N(0,1)$ by Lemma 1. This completes the entire theorem proof.

REFERENCES

Anselin , L. (1980) “Estimation Methods for Spatial Autoregressive Structures,” *Regional Science Dissertation and Monograph Series 8*, Cornell University, Ithaca, NY.

Banerjee, S., Carlin, B.P. and Gelfand, A.E. (2004), *Hierarchical Modeling and Analysis for spatial Data*, Champman & Hall/CRC.

Brock, W. A. and Durlauf, S. N. (2001), “Discrete choice with social interaction,” *Review of Economics Studies*, 68, 235–260.

Bronnenberg, B. J. and V. Mahajan (2001), “Unobserved retailer behavior in multimarket data: Joint spatial dependence in Marketing Shares and Promotion Variables,” *Marketing Science*, 20, 284-299.

Calvó-Armengol, A., Patacchini, E., and Zenou, Y. (2009), “Peer Effects and Social Networks in Education,” *Review of Economic Studies*, 76, 1239–1267.

Case, A. C. (1991), “Spatial patterns in house demand,” *Econometrica*, 59, 953–965.

Chen, X., Chen, Y., and Xiao, P. (2013), “The impact of sampling and network topology on the estimation of social intercorrelation,” *Journal of Marketing Research*, L, 95–110.

Cohendet, P., Llerena, P., Stahn, H., and Umbhauer, G. (1998), *The Economics of Networks: Interactions and Behaviors*, New York: Springer.

Lee, L. F., Li, J., and Lin, X. (2013), “Specification and estimation of social interaction models with network structure,” *Review of Economics and Statistics*, To appear.

LeSage, J. and Pace, R. K. (2009), *Introduction to Spatial Econometrics*, New York: Chapman & Hall.

Ord, J. (1975), “Estimation methods for models of spatial interaction,” *Journal of the American Statistical Association* 70, 120C26.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice and visualization*, New York: John Wiely.

Shao, J. (2003), *Mathematical Statistics*, New York: John Wiely.

Thompson, S. K. (2012), *Sampling*, New York: John Wiely.

Wall, M. M. (2004) “A close look at the spatial structure implied by the CAR and SAR models,” *Journal of Statistical planning and inference*, 121,311-324.

Wasserman, S. and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge University Press.

Table 1: Simulation results for PMLE with $n/N = 10\%$

		$\rho = 0.2$				$\rho = 0$			
N	n	$\flat$	SE	$\widehat{\text{SE}}$	ERP	$\flat$	SE	$\widehat{\text{SE}}$	ERP
1,000	100	0.0218	0.5391	0.5471	5.60%	0.0165	0.5357	0.5471	4.30%
5,000	500	0.0035	0.2452	0.2387	13.40%	0.0041	0.2390	0.2387	4.20%
10,000	1000	0.0012	0.1647	0.1639	21.60%	0.0011	0.1603	0.1639	4.20%
100,000	10,000	0.0003	0.0537	0.0521	95.90%	0.0002	0.0532	0.0521	5.40%
500,000	50,000	0.0001	0.0237	0.0232	100.0%	0.0001	0.0232	0.0232	4.40%

Table 2: The real data simulation results for PMLE with $N = 557,818$ and $\rho = 0.154$

n	$\flat$	SE	$\widehat{\text{SE}}$	ERP
2,000	0.0274	0.5139	0.4728	7.5%
5,000	0.0093	0.1743	0.1727	15.8%
10,000	0.0037	0.0907	0.0853	47.6%
20,000	0.0026	0.0451	0.0426	94.1%
50,000	0.0009	0.0189	0.0170	100.0%

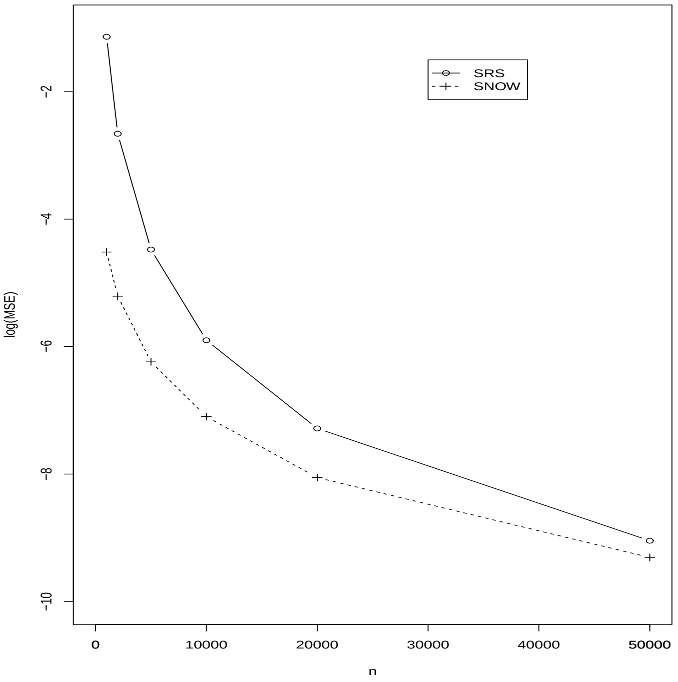

Figure 1: Comparing different sampling methods by MSE in log-scale