Electronic Journal of Statistics
Vol. 13 (2019) 1135–1165
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1549

Least squares estimation of spatial autoregressive models for large-scale social networks*

Danyang Huang

Renmin University of China

Wei Lan

Southwestern University of Finance and Economics
e-mail: lanwei@swufe.edu.cn

Hao Helen Zhang

University of Arizona

Hansheng Wang

Peking University

Abstract: Due to the rapid development of various social networks, the spatial autoregressive (SAR) model is becoming an important tool in social network analysis. However, major bottlenecks remain in analyzing large-scale networks (e.g., Facebook has over 700 million active users), including computational scalability, estimation consistency, and proper network sampling. To address these challenges, we propose a novel least squares estimator (LSE) for analyzing large sparse networks based on the SAR model. Computationally, the LSE is linear in the network size, making it scalable to analysis of huge networks. In theory, the LSE is $\sqrt{n}$-consistent and asymptotically normal under certain regularity conditions. A new LSE-based network sampling technique is further developed, which can automatically adjust autocorrelation between sampled and unsampled units and hence guarantee valid statistical inferences. Moreover, we generalize the LSE approach for the classical SAR model to more complex networks associated with multiple sources of social interaction effect. Numerical results for simulated and real data are presented to illustrate performance of the LSE.

*Danyang Huang's research is supported by National Natural Science Foundation of China (Grant No. 11701560), the Beijing Municipal Social Science Foundation(Grant No. 17GLC051) and the Center for Applied Statistics, School of Statistics, Renmin University of China. Wei Lan's research is partially supported by National Natural Science Foundation of China (No. 11401482, No. 71532001), and the Center of Statistical Research, Southwestern University of Finance and Economics. Hansheng Wang's research is partially supported by National Natural Science Foundation of China (No. 71532001, No. 11525101, No. 71332006). It is also supported in part by China's National Key Research Special Program (No. 2016YFC0207700). Zhang's research was supported in part by NSF DMS-1309507, DMS-1418172, and NSFC 11571009. We deeply appreciate Professor James P. LeSage for generously sharing with us the well written MLE codes; see *http://www.spatial-statistics.com/software\protect_index.htm*.

Keywords and phrases: Large-scale social networks, least squares estimation, network sampling, social interaction.

Received May 2018.

1. Introduction

We consider a network with n nodes. An adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ could be defined to describe the network structure. Let $Y = (Y_1, \cdots, Y_n)^\top \in \mathbb{R}^n$ be the continuous responses collected from the n nodes. In social network analysis, the spatial autoregressive (SAR) model has been popularly employed for modeling the social interaction within a network (Bronnenberg and Mahajan, 2001; Lee et al., 2010; Anselin, 2013), which is,

$$Y = \rho WY + \mathcal{E}, \tag{1.1}$$

where $W = (w_{ij}) \in \mathbb{R}^{n\times n}$ with $w_{ij} = a_{ij}/\sum_{j=1}^n a_{ij}$. And ρ is the autocorrelation parameter representing the social interaction (Lee et al., 2010). The random noises are collected in the vector $\mathcal{E} = (\varepsilon_1, \cdots, \varepsilon_n)^\top \in \mathbb{R}^n$, which is assumed to have mean $\mathbf{0}_n = (0, \cdots, 0) \in \mathbb{R}^n$ and covariance matrix $\sigma^2 I_n \in \mathbb{R}^{n\times n}$, where I_n is the identity matrix of dimension n. Due to the rapid development of online social network websites (e.g. Facebook, Twitter, Sina Weibo, Wechat), the usefulness of the SAR model has been increasingly recognized in recent years (Sampson et al., 1999; Leenders, 2002; Fujimoto et al., 2011; Robins, 2013).

Remark 1. The estimation of social interaction coefficient ρ is important in large networks. For example, in marketing research, a node i could be a user on a social network platform, and Y_i could be defined to represent a person's attitude about a brand. It could be influenced by his or her friends on the social network platform. Studying the influence is important (Lee et al., 2010). Statistically, this amounts to estimate the social interaction ρ. A positive and significant social interaction indicates potential profit in marketing strategy on the platform. This further suggests that estimation of ρ is important in large networks. See Chen et al. (2013) for more detailed discussions.

We assume $|\rho| < 1$ throughout the article. According to the proof in Banerjee et al. (2004), $I_n - \rho W$ is invertible in this case. Thus we have $Y = (I_n - \rho W)^{-1}\mathcal{E}$. This implies that $E(Y) = \mathbf{0}_n$, and $\mathrm{cov}(Y) = \Sigma = \sigma^2(I_n - \rho W)^{-1}(I_n - \rho W^\top)^{-1}$. If $\mathcal{E}$ is further assumed to follow a normal distribution, one can obtain maximum likelihood estimator (MLE) of ρ and σ^2 (Barry and Pace, 1999). The estimator's asymptotic distribution has been studied by Lee (2004) and Hillier and Martellosio (2014). Some higher order asymptotic results have been investigated recently by Robinson and Rossi (2014).

Despite many recent advances and successes in the SAR model applications, existing estimation methods do experience bottlenecks when analyzing large networks. The size of many popular social websites can be enormously large. For example, the Sina Weibo (*www.weibo.com*) network analyzed in this paper has $n = 557,818$ nodes. Thus the traditional methods are computationally

infeasible with usual computational resource. The computational complexity of evaluating the determinant $|I_n - \rho W|$ in the log-likelihood function is in general $O(n^3)$ (Trefethen and Bau, 1997; Barry and Pace, 1999). Huang et al. (2018) proposed the pseudo likelihood estimate for SAR with random effects. Because this is a likelihood-type method, complex matrix computation (e.g. log determinant) is needed. Thus the computational complexity is still $O(n^3)$. For a sparse matrix, more efficient algorithms have been proposed (Barry and Pace, 1999; Smirnov and Anselin, 2001; LeSage and Pace, 2007). However, these methods usually rely on some stringent assumptions on $I_n - \rho W$, which can hardly hold for real social network data. For example, one commonly used assumption is that all of the eigenvalues of the weight matrix W are real (Barry and Pace, 1999). This condition is not necessarily satisfied if W is asymmetric. Furthermore, the aforementioned methods need to evaluate the Jacobian term, which is computationally expensive for large n.

Other major challenges include network sampling and statistical inference. Most existing network analysis methods such as the MLE assume that the entire network are observed. However, it is often not possible to observe the entire network, and only sampled network data can be collected. Since the sampled data may omit relationships between sampled and unsampled nodes, it may lead to biased estimation and the resulting statistical inference could be misleading (Frank, 1979; Costenbader and Valente, 2003; Handcock and Gile, 2010; Shalizi and Rinaldo, 2013). Chen et al. (2013) pointed out that sampled network may lead to seriously biased estimation for the SAR model when the MLE is applied. Zhou et al. (2017) proposed the paired maximum likelihood estimator. However, the method is based on Taylor's expansion, which works well only when ρ is sufficiently small. See euation (2.5) and the relative discussion in the paper for more details. Thus better techniques for network sampling are needed to ensure consistent estimation of social interaction effect.

Motivated by these challenges, we propose a novel, fast and scalable estimation method for the SAR model. The new method is particularly designed for large social networks with tens of thousands of (or millions of) nodes. One main advantage of the new method is that, under appropriate sparsity assumptions, its computational complexity is linear in network size. Our basic idea is described as follows. For any node i, define $Y_{(-i)} = (Y_j, 1 \le j \le n, j \ne i)$. Under the normal assumption, we can show that $E\{Y_i|Y_{(-i)}\} = Y_{(-i)}^\top \gamma_i^*(\rho)$, where the coefficient vector $\gamma_i^*(\rho) \in \mathbb{R}^{(n-1)}$ depends on both i and ρ. Based on this, we propose to construct a least squares type objective function $Q(\rho) = \sum_i \{Y_i - Y_{(-i)}^\top \gamma_i^*(\rho)\}^2$ and obtain the least squares estimator (LSE) of ρ by minimizing $Q(\rho)$. We can show that under appropriate assumption, the computational complexity of minimizing $Q(\rho)$ is $O(n)$, i.e., linear in the network size.

In practice, the effect of social interaction on individuals may come from multiple sources (Leenders, 2002). First, the existence of an edge is the result of a combination of different properties of the network (Krivitsky et al., 2009), such as the *homophily* (nodes with similar characteristics are more likely to relate) or *degree heterogeneity* (super stars receive edges more than others).

Therefore different relationships between nodes, such as friends or fans, may have disparate social interaction effects. Second, researchers often collect data from multiple social network platforms, such as instant messaging, mobile phone communication, and online social networks. Networks obtained from different platforms may have separate impact on the response. This leads to multiple weighting matrices. In this paper, we will also generalize the LSE approach to estimate the SAR model with multiple weighting matrices, which allows to capture different types of social interaction from multiple sources. The resulting estimator is referred to as the mLSE.

The rest of the article is organized as follows. Section 2 introduces the proposed LSE approach, establishes its theoretical properties, and presents the generalization of the proposed approach. Asymptotic properties of the estimators are established as well. Simulation studies and a real data example are given in Section 3. Section 4 concludes the article with discussions. All the theoretical proofs are relegated to Appendix.

2. Least squares estimation

2.1. Motivation

Consider a network with n nodes. Let Y_i be the response for node i with $1 \leq i \leq n$, and $Y = (Y_1, Y_2, \cdots, Y_n)^\top \in \mathbb{R}^n$. Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be the network adjacency matrix, where $a_{ij} = 1$ if node i follows node j and $a_{ij} = 0$ otherwise. For completeness, we assume $a_{ii} = 0$ for $i = 1, \cdots, n$. To assess the social interaction effect of the network structure, we assume the model in (1.1).

In order to ensure $(I_n - \rho W)$ to be invertible, the eigenvalues of ρW should be all different from 1. Banerjee et al. (2004) showed that the largest eigenvalue of W is 1. These two facts imply that $|\rho| < 1$ is a sufficient condition to guarantee the invertibility of $(I_n - \rho W)$ for a general W. In fact, this is also a necessary condition. Otherwise, one can always find an appropriately defined matrix W so that $(I_n - \rho W)$ is not invertible; see Banerjee et al. (2004) for more detailed discussions. Consequently, we assume that $|\rho| < 1$.

The autocorrelation parameter ρ measures the effect of social interaction. By omitting some constants, the quasi log likelihood function can be written as,

$$\ell(\rho, \sigma) = \ln |I_n - \rho W| - (n/2) \ln \sigma^2 - (1/2)\sigma^{-2} Y^\top (I_n - \rho W)^\top (I_n - \rho W) Y. \quad (2.1)$$

The MLE could be obtained by maximizing (2.1). However, this classical MLE approach becomes computationally expensive when the network size n is large, mainly due to high cost of computing $|I_n - \rho W|$ (Trefethen and Bau, 1997; Barry and Pace, 1999; Smirnov and Anselin, 2001).

2.2. Least squares estimation

We propose to estimate the SAR model (1.1) based on the following intriguing observation. Recall that $Y_{(-i)} = (Y_1, \cdots, Y_{i-1}, Y_{i+1}, \cdots, Y_n)^\top \in \mathbb{R}^{(n-1)}$ for each i. Then, we have the following result, with the proof given in Appendix A.1.

Proposition 1. *Assuming that $\mathcal{E}$ is normally distributed, then*

$$E\{Y_i|Y_{(-i)}\} = Y_{(-i)}^\top \gamma_i^*(\rho) = \sum_{k\neq i} \frac{\rho(\omega_{ik}+\omega_{ki}) - \rho^2 \sum_j \omega_{ji}\omega_{jk}}{1+\rho^2\sum_j \omega_{ji}^2} Y_k.$$

By Proposition 1, the conditional expectation of Y_i given $Y_{(-i)}$ is linear in $Y_{(-i)}$ with the coefficient $\gamma_i^*(\rho)$. Correspondingly, we propose to construct the following least squares type objective function,

$$Q(\rho) = \sum_i \left\{Y_i - Y_{(-i)}^\top \gamma_i^*(\rho)\right\}^2. \tag{2.2}$$

Define $\Omega_\rho = (I_n - \rho W)^\top (I_n - \rho W)$, and $d_\rho = \text{diag}\{(1+\rho^2||W_{\cdot i}||^2)^{-1}, 1 \le i \le n\} \in \mathbb{R}^{n\times n}$, where $W_{\cdot,i} \in \mathbb{R}^{n\times 1}$ represents the ith column of W. As a result, we have

$$Q(\rho) = ||d_\rho \Omega_\rho Y||^2. \tag{2.3}$$

The detailed verification of (2.3) is given in Appendix A.2. This leads to the proposed LSE $\hat{\rho} = \text{argmin}_\rho Q(\rho)$.

Next, we point out that minimizing $Q(\rho)$ in (2.2) is computationally feasible even when n is large. This is based on the fact that, even though $\gamma_i^*(\rho) \in \mathbb{R}^{n-1}$ is high-dimensional, it is extremely sparse. The kth entry of $\gamma_i^*(\rho) = \{\gamma_{ik}^*(\rho), k \neq i\}$ is $\gamma_{ik}^*(\rho) = \{\rho(\omega_{ik}+\omega_{ki}) - \rho^2\sum_j \omega_{ji}\omega_{jk}\}(1+\rho^2\sum_j \omega_{ji}^2)^{-1}$. A necessary condition for $\gamma_{ik}^*(\rho) \neq 0$ is either $\omega_{ik}+\omega_{ki} > 0$ or $\sum_j \omega_{ji}\omega_{jk} > 0$, which is equivalent to either $a_{ik}+a_{ki} > 0$ or $\sum_j a_{ji}a_{jk} > 0$. For each node i, there are two types of nodes (denoted by ks) satisfying the aforementioned conditions: its "direct" friends (those ks satisfying $a_{ik}+a_{ki} > 0$), and its one particular type of "indirect" friends (those ks satisfying $\sum_j a_{ji}a_{jk} > 0$).

In an ideal situation, if the number of friends connected with each node is finite, then one can verify that the computational complexity of $\hat{\rho}$ is linear in n. The following proposition presents a formal result.

Proposition 2. *If we further make the following assumptions: (1) Assume the objective function (2.3) is optimized by the Newton-Raphson algorithm, which converges in a finite number of steps; (2) Assume that there exists a finite constant $d_{\max}$ such that $\max_i d_i = \max_i \sum_j a_{ji} \le d_{\max}$ as $n \to \infty$. Then, the computational complexity demanded for optimizing (2.3) is $O(n)$.*

The proof of Proposition 2 is given in Appendix A.3. We make some explanations about the assumptions in this proposition.

Remark 2. The assumption (2) in Proposition 2 is a constraint on network sparsity. By this assumption, we know that the degree d_i of each node is bounded by a finite constant. This means the density of the network is $O(n^{-1})$, which goes to 0 as n goes to infinity. This implies that the network structure is very sparse. In the meanwhile, if $d_{\max}$ diverges to infinity at a low speed (e.g., $\log(n)$), then the computational complexity becomes $O(n\log(n))$, which is slightly higher than $O(n)$.

2.3. Asymptotic properties

The following conditions are needed to establish asymptotic results for the LSE. Define $\lambda_j(B)$ to be the jth eigenvalue of an arbitrary matrix $B \in \mathbb{R}^{p\times p}$ such that $|\lambda_1(B)| \geq |\lambda_2(B)| \geq \cdots \geq |\lambda_p(B)|$. Further define $S = I_n - \rho W$, $M_1 = \sigma d_\rho S^\top$, $M_2 = \sigma(\dot{d}_\rho S^\top - d_\rho W^\top - d_\rho S^\top W S^{-1})$, and $M = M_1^\top M_2$. The following conditions are needed.

(C1) (Network Connectivity) The set $\{1, \cdots, n\}$ is defined including all nodes as the state space of a Markov chain. The transition probability is given by the weighting matrix W. We assume the Markov chain to be irreducible and aperiodic. Additionally, we define $\pi = (\pi_i)^\top \in \mathbb{R}^n$ to be the stationary distribution vector of the Markov chain, which satisfies $\pi_i \geq 0$, $\sum_i \pi_i = 1$, and $W^\top \pi = \pi$. We further assume $\sum_{i=1}^n \pi_i^2 = O(n^{-1/2-\delta})$, where $0 < \delta \leq 1/2$ is a positive constant.

(C2) (Network Uniformity) Define $\mathbb{W}_1 = W + W^\top$, which is a symmetric matrix. Assume $|\lambda_1(\mathbb{W}_1)| = O(\log n)$.

(C3) (Noise Term) Define $\widetilde{\varepsilon} = \sigma^{-1}\mathcal{E} = (\widetilde{\varepsilon}_1, \cdots, \widetilde{\varepsilon}_n)^\top \in \mathbb{R}^n$. Assume $E(\widetilde{\varepsilon}_i^4) = \kappa_4$, and $E(\widetilde{\varepsilon}_{i_1}\widetilde{\varepsilon}_{i_2}\widetilde{\varepsilon}_{i_3}) = 0$ for $1 \leq i_1, i_2, i_3, i \leq n$, where κ_4 is a finite constant. Further assume $c_\varepsilon = E(\widetilde{\varepsilon}_i^2 - 1)^4$ is a finite constant.

(C4) (Law of Large Numbers) Assume the limits of the following network features exist: $\sigma_1^2 = \lim_{n\to\infty} \sigma_{1n}^2$, and $\sigma_2^2 = \lim_{n\to\infty} \sigma_{2n}^2$, where

$$\sigma_{1n}^2 = n^{-1}\Big[\mathrm{tr}(MM^\top) + \mathrm{tr}(M^2) + (\kappa_4 - 3)\mathrm{tr}\big\{\mathrm{diag}^2(M)\big\}\Big],$$
$$\sigma_{2n}^2 = 2n^{-1}\mathrm{tr}(M_2^\top M_2).$$

First, in the first two conditions, we impose assumptions on the network structure. Condition (C1) assumes certain connectivity for the network structure. Specifically, it could be verified that, if the network is fully connected after a finite number of steps, then both irreducibility and aperiodicity could be satisfied. If the famous six degrees of separation (Newman et al., 2006) holds, then the condition is satisfied. According to Meyn and Tweedie (2012), if condition (C1) is satisfied, then it holds that $\lim_{m\to\infty} W^m = \mathbf{1}_n \pi^\top$, where $\mathbf{1}_n$ is an n-dimensional vector with all elements to be 1. In condition (C2), we assume certain uniformity on the network structure. Classical SAR models (Lee, 2004; Yang and Lee, 2017) require the row and column sums of W to be bounded. While (C2) allows $\lambda(\mathbb{W}_1)$ to diverge with the rate of $O(\log n)$. It is remarkable that conditions (C1) and (C2) are not related to the assumption (2) in Proposition 2. This means the asymptotic property of LSE in the following theorem does not rely on sparsity of the network. Second, condition (C3) is a regularity assumption on the noise terms. It could be verified that the normal distribution with mean 0 satisfies this condition. Third, condition (C4) is a law of large number type condition. We consider two special cases to help understand the condition. In this cases, the existence of limits could be theoretically verified.

Case 1. (Circle Network) In this network, we assume $a_{i,i+1} = 1$ for $1 \leq i \leq n-1$, and $a_{n1} = 1$; otherwise, $a_{ij} = 0$ $(1 \leq i, j \leq n)$. In this case, the nodes are connected as a circle. We could show the detailed expression of σ_{1n}^2

and σ_{2n}^2, which are $\sigma_{1n}^2 = 2(1+\rho)^2 \sum_{k=0}^{N-3}(-\rho)^k(1+\rho^2)^{-3}(\sum_{k=0}^{N-1}\rho^k)^{-1}$, and $\sigma_{2n}^2 = 4(1-\rho^2)(1+\rho^2)^{-3}$. Thus we have $\sigma_1^2 = \lim_{n\to\infty}\sigma_{1n}^2 = 2(1-\rho^2)(1+\rho^2)^{-3}$, and $\sigma_2^2 = 4(1-\rho^2)(1+\rho^2)^{-3}$. In this case, the limits of σ_{1n}^2 and σ_{2n}^2 exist.

Case 2. (Fully Connected Block) In this network, for a fixed positive integer k, define $A_k = (a_{ij}^{*k})$, where $a_{ij}^{*k} = 1$ $(1 \le i, j \le k$, and $i \ne j)$, and $a_{ii}^{*k} = 0$ $(1 \le i \le k)$. Thus the network adjacency matrix could be generated as $A = I_c \otimes A_k \in \mathbb{R}^{n\times n}$ for sample size n, where n is assumed to be $n = ck$ for a positive integer c. In this case, the network density is $\sum a_{ij}/\{n(n-1)\} = (n-1)^{-1}$. As a result, it could be calculated that, $\sigma_1^2 = 8(k-1)^{-3}(1+\rho^2)^{-6}\big\{(k-1)^2\rho^4 + 2(k-1)(k-2)\rho^3 + (k^3-6k^2+8k-2)\rho^2 - 2(k-1)(k-2)\rho + (k-1)^2\big\}$, and $\sigma_2^2 = 8(k-1)^{-1}(1+\rho^2)^{-4}\big\{(k-1)\rho^2+1\big\}$. Thus the limits exist. One could verify that if k is diverging, then the limits still exist.

The above assumptions are to facilitate the asymptotic analysis based on the central limit theorem. Next, we establish asymptotic properties of the LSE $\hat{\rho}$ in the following theorem, which provides theoretical justifications for the new estimator.

Theorem 1. *Assume that the conditions (C1)– (C4) hold, then we have*

$$\sqrt{n}(\hat{\rho}-\rho) \to_d N\Big(0, \sigma_2^{-4}\sigma_1^2\Big), \quad as \quad n \to \infty.$$

The proof of Theorem 1 is left to Appendix C.1. By Theorem 1, we know that the LSE $\hat{\rho}$ is asymptotically normally distributed and $\sqrt{n}$-consistent. This is the same convergence rate of the classical SAR models. The LSE's asymptotic variance is $\sigma_2^{-4}\sigma_1^2$. Thus consistent estimators for σ_2^2 and σ_1^2 are to be derived to make valid inferences. See Appendix C.2 for discussion of computationally feasible estimators for σ_1^2 and σ_2^2.

By assuming $\mathcal{E}$ is normally distributed, MLE of ρ could be obtained. Then it is of interest to theoretically compare the relative estimation efficiency between MLE and LSE. It is worth noting that in real practice, the network effect is typically small (Chen et al., 2013). We are then motivated to conduct Taylor's expansion of ρ to approximate the asymptotic covariance of MLE and LSE by their leading terms. For simplicity, we assume that $\sigma^2 = 1$. Define $\sigma_L^2 = \sigma_2^{-4}\sigma_1^2$ to be the asymptotic variance of LSE, and σ_M^2 to be that of MLE. Thus, under appropriate assumptions for the MLE, the following theorem could be established.

Theorem 2. *Assume* $\pi_A = \lim_{n\to\infty} n^{-1}\big\{tr(W^2) + tr(W^\top W)\big\}$ *exists. Thus we have,* $\sigma_L^2 = \pi_A^{-1} + o(1)$ *and* $\sigma_M^2 = \pi_A^{-1} + o(1)$.

The proof of Theorem 2 is given in Appendix C.3. See the appendix for more detailed discussions. By Theorem 2, the conclusion could be drawn that, similar estimation efficiency can be obtained for MLE and LSE if ρ is small.

2.4. New LSE-based scheme for sampling networks

In this section, we develop a novel sampling scheme to cope with the LSE approach, and further show that the sampled data can lead to a consistent

estimation for the SAR model. For each node i, define a sampling indicator s_i, which equals to 1 if node i is sampled as the response node and equals to 0 otherwise. Define $\mathcal{S}_y = \{i : s_i = 1\}$, which is the collection of all the sampled response nodes. In order to evaluate $\gamma_i^*(\rho)$ correctly for each sampled node i, we need to sample its "direct" and appropriately defined "indirect" friends. Altogether, the related nodes are denoted by,

$$\mathcal{S}_x = \{k : a_{ik} + a_{ki} > 0 \ \text{ or } \ \sum_j a_{ji}a_{jk} > 0, \ \text{ for some } \ i \in \mathcal{S}_y\}.$$

Without involving all the nodes in networks, the new sampling scheme can be implemented in the following three steps:

- Step 1: Obtain the response set $\mathcal{S}_y$ via some convenient sampling method.
- Step 2: Collect all the nodes js, which are directly connected with some node $i \in \mathcal{S}_y$ in different directions ("direct" friends). Denote the two types of nodes as $\mathcal{S}_{m_1} = \{j : a_{ji} = 1 \text{ and } i \in \mathcal{S}_y\}$ and $\mathcal{S}_{m_2} = \{j : a_{ij} = 1 \text{ and } i \in \mathcal{S}_y\}$, respectively.
- Step 3: Obtain the "indirectly" connected nodes by searching for ks which are connected with node $j \in \mathcal{S}_{m_1}$, in the direction $a_{jk} = 1$. Denote $\mathcal{S}_{m_3} = \{k : a_{jk} = 1 \text{ and } j \in \mathcal{S}_{m_1}\}$. Putting all together, we get $\mathcal{S}_x = \mathcal{S}_{m_1} \bigcup \mathcal{S}_{m_2} \bigcup \mathcal{S}_{m_3}$.

Using $\mathcal{S}_x$ and $\mathcal{S}_y$, one can construct the following least squares objective function

$$Q_s(\rho) = \sum_{i \in \mathcal{S}_y} \Big\{ Y_i - Y_{(-i)}^\top \gamma_i^*(\rho) \Big\}^2.$$

Define the resulting sampling-based LSE as $\hat{\rho}_s = \text{argmin}_\rho Q_s(\rho)$, which is referred to as the Sample-LSE. Define $\mathbb{S} = \text{diag}(s_1, \cdots, s_n) \in \mathbb{R}^{n \times n}$, where $s_i = 1$ if $i \in \mathcal{S}_y$ and 0 otherwise. Define n_s to be the number of nodes collected in $\mathcal{S}_y$. $\dot{Q}_s(\rho)$ and $\ddot{Q}_s(\rho)$ are similarly defined as $\dot{Q}(\rho)$ and $\ddot{Q}(\rho)$. Further define $M_{1s} = \sigma d_\rho \mathbb{S} S^\top$, $M_{2s} = \sigma(\dot{d}_\rho \mathbb{S} S^\top - d_\rho \mathbb{S} W^\top - d_\rho \mathbb{S} S^\top W S^{-1})$ and $M_s = M_{1s}^\top M_{2s}$. The following condition is needed to establish the asymptotic property of Sample-LSE.

(C5) (Law of Large Numbers) Assume the limits of the following network features exist: $\sigma_{1s}^2 = \lim_{n\to\infty} \sigma_{1ns}^2$, and $\sigma_{2s}^2 = \lim_{n\to\infty} \sigma_{2ns}^2$, where

$$\sigma_{1ns}^2 = n_s^{-1}\Big[\text{tr}(M_s M_s^\top) + \text{tr}(M_s^2) + (\kappa_4 - 3)\text{tr}\big\{\text{diag}^2(M_s)\big\}\Big],$$
$$\sigma_{2ns}^2 = 2n_s^{-1}\text{tr}(M_{2s}^\top M_{2s}).$$

Then the following theorem could be established.

Theorem 3. *Assume that the conditions (C1)– (C3) and (C5) hold, then we have*

$$\sqrt{n_s}(\hat{\rho}_s - \rho) \to_d N\Big(0, \sigma_{2s}^{-4}\sigma_{1s}^2\Big), \quad as \ \ n_s \to \infty.$$

Proof of Theorem 3 is similar to that of Theorem 1 and is hence omitted. By Theorem 3, the Sample-LSE $\hat{\rho}_s$ is $\sqrt{n_s}$-consistent and asymptotically normal. The discussion of computationally feasible estimators for σ_{1s}^2 and σ_{2s}^2 is similar to that in the previous section and thus omitted. It is worthy mentioning that although $\mathbb{S}$ is an $n \times n$ diagonal matrix, it only has n_s nonzero diagonal elements. Therefore, only the sampled nodes in $\mathcal{S}_x$ and $\mathcal{S}_y$ are involved in the estimation. Both the performances of simple random sampling and snowball sampling are demonstrated in the simulation studies.

2.5. mLSE: generalization to multiple weighting matrices

Now we generalize the LSE approach to models with multiple weighting matrices. It allows us to consider separate social interaction effects from multiple sources. Multiple weighting matrices are sometimes used in spatial statistics; see for example LeSage and Pace (2009). Consider the SAR model with L weighting matrices. For simplicity, we assume $L = 2$ in this paper, but it is straightforward to extend the proposed estimation procedure and theoretical results to the case of $L > 2$. The model is defined as,

$$Y = \rho_1 W_1 Y + \rho_2 W_2 Y + \mathcal{E}, \tag{2.4}$$

where $w_{l,ij} = a_{l,ij}/d_{l,i}$, $d_{l,i} = \sum_{j=1}^{n} a_{l,ij}$ for $l \in \{1, 2\}$. $\mathcal{E} \in \mathbb{R}^n$ has mean $\mathbf{0}_n$ and covariance matrix $\sigma^2 I_n \in \mathbb{R}^{n \times n}$. In order to insure the invertibility of the matrix $I_n - \rho_1 W_1 - \rho_2 W_2$, one could verify that a sufficient condition is $|\rho_1| + |\rho_2| < 1$. Thus we assume $|\rho_1| + |\rho_2| < 1$ throughout the rest of this article. Then we have $\Sigma = \mathrm{var}(Y) = \sigma^2 (I_n - \rho_1 W_1 - \rho_2 W_2)^{-1} (I_n - \rho_1 W_1^\top - \rho_2 W_2^\top)^{-1}$. Each W_l, for $1 \le l \le L$, represents the weight of the adjacency matrix A_l. Correspondingly, the autocorrelation parameters ρ_ls represent influence effects of different types of social interactions.

Remark 3. For model defined in (1.1) (Anselin, 2013), the parameter estimation becomes more and more inaccurate as $\rho \approx 1$. Similarly, for model defined in (2.4), when $|\rho_1| + |\rho_2| \approx 1$, model estimation becomes inaccurate. However, in practice, the autocorrelation could be small (Chen et al., 2013). We have also demonstrated this fact through the Weibo example in real data analysis.

To employ the LSE approach to (2.4), we define $\gamma_i^*(\theta)$, where $\theta = (\rho_1, \rho_2)^\top$. Then, the conditional expectation of Y_i given $Y_{(-i)}$ is linear in $Y_{(-i)}$ for $i = 1, \ldots, n$. The result is given in the following proposition.

Proposition 3. *Assume $\mathcal{E}$ is normally distributed, then*

$$\begin{aligned} &E\{Y_i | Y_{(-i)}\} = Y_{(-i)}^\top \gamma_i^*(\theta) \\ &= \sum_{k \ne i} \frac{\sum_{l=1}^{2} \left\{ \rho_l (\omega_{l,ik} + \omega_{l,ki}) - \rho_l \sum_j \omega_{l,ji} (\rho_1 \omega_{1,jk} + \rho_2 \omega_{2,jk}) \right\}}{1 + \sum_j (\rho_1 \omega_{1,ji} + \rho_2 \omega_{2,ji})^2} Y_k. \end{aligned}$$

The coefficient vector $\gamma_i^*(\theta)$ is sparse if W_ls are sparse. The proof is similar to that of Proposition 1 and given in Appendix A.1. Next, we propose the objective function,

$$Q(\theta) = \sum_i \left\{ Y_i - Y_{(-i)}^\top \gamma_i^*(\theta) \right\}^2 = ||d_\theta \Omega_\theta Y||^2, \tag{2.5}$$

where $\Omega_\theta = (I_n - \rho_1 W_1 - \rho_2 W_2)^\top (I_n - \rho_1 W_1 - \rho_2 W_2)$, $d_\theta = \text{diag}\{(1 + ||\rho_1 W_{1,\cdot i} + \rho_2 W_{2,\cdot i}||^2)^{-1}, 1 \le i \le n\} \in \mathbb{R}^{n\times n}$, and $W_{l,\cdot i}$ is the ith column of W_l ($l \in \{1,2\}$). The second equality in (2.5) is verified in Appendix A.2. The LSE for model (2.4) is denoted by $\hat\theta = \min_\theta Q(\theta)$. We refer to the LSE for multiple weighting matrices as the mLSE.

Computationally, it is worth pointing out that, minimizing $Q(\theta)$ in (2.5) is still feasible for a very large n. A necessary condition for the kth entry of $\gamma_i^*(\theta) = \{\gamma_{ik}^*(\theta), k \ne i\}$ is nonzero, is either $a_{l,ik} + a_{l,ki} > 0$ or $\sum_j a_{l_1,ji} a_{l_2,jk} > 0$ for $l, l_1, l_2 \in \{1,2\}$. Typically, A_ls are extremely sparse, which implies that the total number of nodes involved in $\gamma_i^*(\theta)$ is finite for any i. This would dramatically reduce the computational cost of $\hat\theta$.

Before establishment of the property of θ, the following conditions are needed. Define $S_\theta = I_n - \rho_1 W_1 - \rho_2 W_2$, $M_{1\theta} = \sigma d_\theta S_\theta^\top$, $M_{2\theta,l} = \sigma(\dot d_{\theta l} S_\theta^\top - d_\theta W_l^\top - d_\theta S_\theta^\top W_l S_\theta^{-1})$ and $M_{\theta,l} = M_{1\theta}^\top M_{2\theta,l}$. Further define $\rho_a = |\rho_1| + |\rho_2|$, and $W_a = \rho_a^{-1}|\rho_1| W_1 + \rho_a^{-1}|\rho_2| W_2$.

(C6) (Network Connectivity) Define $\{1, \cdots, n\}$ to be the set including all nodes as the state space of a Markov chain. Weighting matrix W_a is the transition probability matrix. The Markov chain is assumed to be irreducible and aperiodic. Further define $\pi_a = (\pi_{a,i})^\top \in \mathbb{R}^n$ to be the stationary distribution vector of the Markov chain, satisfying $\pi_{a,i} \ge 0$, $\sum_i \pi_{a,i} = 1$, and $W_a^\top \pi_a = \pi_a$. Assume $\sum_{i=1}^n \pi_{a,i}^2 = O(n^{-1/2-\delta_a})$, where $0 < \delta_a \le 1/2$ is a positive constant.

(C7) (Network Uniformity) Define $\mathbb{W}_{a1} = W_a + W_a^\top$, which is a symmetric matrix. Further assume $|\lambda_1(\mathbb{W}_{a1})| = O(\log n)$.

(C8) (Law of Large Numbers) Assume the limits of the following network features exist: for $1 \le l_1 \le l_2 \le 2$, $\pi_{1,l_1 l_2} = \lim_{n\to\infty} n^{-1}\big[\text{tr}(M_{\theta,l_1} M_{\theta,l_2}^\top) + \text{tr}(M_{\theta,l_1} M_{\theta,l_2}) + (\kappa_4 - 3)\text{tr}\big\{\text{diag}(M_{\theta,l_1})\text{diag}(M_{\theta,l_2})\big\}\big]$ and $\pi_{2,l_1 l_2} = \lim_{n\to\infty} 2n^{-1}\text{tr}(M_{2\theta,l_1} M_{2\theta,l_2}^\top)$. Define $\Pi_1 = (\pi_{1,11}, \pi_{1,12}; \pi_{1,12}, \pi_{1,22}) \in \mathbb{R}^{2\times 2}$, and $\Pi_2 = (\pi_{2,11}, \pi_{2,12}; \pi_{2,12}, \pi_{2,22}) \in \mathbb{R}^{2\times 2}$ to be positive definite matrices.

The conditions are similar to (C1), (C2), and (C4). Conditions (C6)-(C7) are assumptions on the network structure. Condition (C8) is a law of large numbers type condition. Then, we could establish the asymptotic property of $\hat\theta$.

Theorem 4. *Assume that the conditions (C3), and (C6)–(C8) hold. We then have* $\sqrt{n}(\hat\theta - \theta) \to_d N(\mathbf{0}_2, \Pi_2^{-1}\Pi_1\Pi_2^{-1})$ *as* $n \to \infty$.

The proof of Theorem 4 is similar to that of Theorem 1 and thus omited. By Theorem 4, we could see that the mLSE $\hat\theta$ is asymptotically normally distributed

and $\sqrt{n}$-consistent. The computationally feasible estimators for Π_1 and Π_2 can be obtained similarly with those for σ_2^2 and σ_1^2. Similar to the previous result, when the number of nodes directly connected with any arbitrary node is finite in each W_l, only a finite number of nonzero coefficients will be involved in $\{Y_i - Y_{(-i)}^\top \gamma_i^*(\theta)\}^2$.

3. Numerical studies

To evaluate finite sample performance of the proposed LSE methods, we carry out a number of simulation studies. Specifically, we report the performance of the LSE and compare it with the classical MLE, in terms of both estimation efficiency and computational complexity. We also demonstrate the performance of the Sample-LSE and the mLSE. Finally, the LSE methods are applied in the network of the Sina Weibo.

3.1. Performance of the LSE

In the following, we evaluate the finite sample performance of the proposed methodology for three typically encountered network models, and a special case of stochastic block model. Given n, we first generate the adjacency matrix $A = (a_{ij})$, and then set $a_{ii} = 0$ for each $1 \le i \le n$. Note that A is not necessarily symmetric. Subsequently, W can be computed by normalizing each row of A.

Example 1. (Dyad Independence Model) By Holland and Leinhardt (1981), define a dyad as $D_{ij} = (a_{ij}, a_{ji})$ for any $1 \le i < j \le n$. Assume D_{ij}s are mutually independent of each other. To allow for sparsity of the network, we define $P\{D_{ij} = (1,1)\} = 0.5n^{-1}$ and $P\{D_{ij} = (1,0)\} = P\{D_{ij} = (0,1)\} = 5n^{-1}$. Then we have $P\{D_{ij} = (0,0)\} = 1 - 5.5n^{-1}$, which is very close to 1 for a large n.

Example 2. (Stochastic Block Model) Consider the network structure generated from the stochastic block model (Wang and Wong, 1987; Nowicki and Snijders, 2001). Let $K = 20$ be the total number of blocks. First, we follow Nowicki and Snijders (2001), and randomly assign each node a block label $(k = 1, \cdots, K)$ with equal probability $1/K$. Next, let $P(a_{ij} = 1) = 20n^{-1}$ if i and j are in the same block; $P(a_{ij} = 1) = 2n^{-1}$ otherwise. In this way, nodes within the same block are more likely to be connected with each other, when compared with nodes from different blocks.

Example 3. (Power-Law Distribution) According to Barabási and Albert (1999), the majority of nodes have few edges but a small amount have huge number of edges. Following Clauset et al. (2009), we simulate A as follows. First, for node i, generate its in-degree $d_i = \sum_j a_{ji}$ according to the discrete power-law distribution with $P(d_i = k) = ck^{-\alpha}$, where c is a normalizing constant and the parameter α is set to be 2. Next, for each i, randomly select d_i nodes as its potential followers.

Example 4. (Densely Connected Communities) Although the social networks are often sparse, they are expected to have densely connected communities. To

TABLE 1
Summary of LSE simulation results for the 3 examples with 1000 replications.

	ρ=0					ρ=0.2				
n	Density	$\flat$	SE	SE*	ERP	Density	$\flat$	SE	SE*	ERP
EXAMPLE 1										
2000	0.0015	0.001	0.032	0.031	5.4%	0.0015	0.001	0.032	0.032	100.0%
5000	0.0006	0.001	0.020	0.020	4.7%	0.0006	0.001	0.020	0.020	100.0%
10000	0.0003	0.000	0.014	0.014	4.9%	0.0003	0.000	0.014	0.014	100.0%
20000	0.0002	-0.000	0.010	0.010	4.8%	0.0002	0.000	0.010	0.010	100.0%
EXAMPLE 2										
2000	0.0015	-0.001	0.033	0.032	4.9%	0.0015	0.001	0.032	0.033	100.0%
5000	0.0006	0.000	0.021	0.020	5.3%	0.0006	0.001	0.022	0.021	100.0%
10000	0.0003	0.000	0.014	0.014	4.1%	0.0003	0.000	0.014	0.015	100.0%
20000	0.0001	0.001	0.010	0.010	4.6%	0.0001	0.000	0.011	0.010	100.0%
EXAMPLE 3										
2000	0.0025	-0.001	0.045	0.047	4.1%	0.0025	0.000	0.046	0.046	98.5%
5000	0.0011	-0.001	0.031	0.032	5.4%	0.0011	0.001	0.031	0.031	100.0%
10000	0.0006	-0.000	0.023	0.023	5.1%	0.0006	0.000	0.023	0.023	100.0%
20000	0.0003	0.000	0.017	0.017	4.7%	0.0003	0.001	0.017	0.017	100.0%
EXAMPLE 4										
2000	0.0060	-0.002	0.055	0.055	4.6%	0.0060	0.004	0.047	0.047	98.6%
5000	0.0024	-0.002	0.035	0.035	5.1%	0.0024	0.002	0.031	0.030	100.0%
10000	0.0012	-0.000	0.024	0.024	4.3%	0.0012	0.001	0.021	0.021	100.0%
20000	0.0006	0.000	0.018	0.017	5.5%	0.0006	0.001	0.015	0.015	100.0%

further examine the robustness to the sparsity assumption for the LSE method, we consider a fully connected block defined in Case 2 in Section 2.3 with $k = 10$. This means $d_i = 10$ for $1 \le i \le n$. To simulate A, for each node i, randomly select d_i nodes as its potential followers.

In each experiment, ρ is set to be 0 or 0.2 and σ^2 is fixed at 1. We consider various network sizes (n=2000, 5,000, 10,000, 20,000). For a reliable evaluation, each experiment is randomly replicated $M = 1,000$ times. For the mth replication, write $\hat{\rho}^{(m)}$ as the estimate of ρ. Then the bias is evaluated as $\flat = \rho - \bar{\rho}$, where $\bar{\rho} = M^{-1}\sum_m \hat{\rho}^{(m)}$. We estimate the standard error $\widehat{\mathrm{SE}}^{(m)}$ by $\hat{\sigma}_2^{-2}\hat{\sigma}_1^*$ based on Theorem 1 and report the average $\mathrm{SE} = M^{-1}\sum_m \widehat{\mathrm{SE}}^{(m)}$. We compare SE with the Monte Carlo standard deviation of $\hat{\rho}$, which is calculated by $\mathrm{SE}^* = \{M^{-1}\sum_m(\hat{\rho}^{(m)} - \bar{\rho})^2\}^{1/2}$. For each coefficient estimate, we compute its Z-type test statistic as $Z^{(m)} = \hat{\rho}^{(m)}/\widehat{\mathrm{SE}}^{(m)}$. The null hypothesis of $H_0 : \rho = 0$ is rejected if $|Z^{(m)}| > z_{\alpha/2}$. $\alpha = 0.05$ is used throughout the rest of this article. The empirical rejection probability (ERP) is computed as $M^{-1}\sum_m I(|Z^{(m)}| > z_{\alpha/2})$. According to whether the true parameter ρ is 0 or not, the ERP might correspond to either empirical size or power.

We summarize numerical results of the LSE in Table 1. It is observed that, for all of the three examples, the ERP results of nonzero ρ (i.e. $\rho = 0.2$) are always larger than 95% when n is not smaller than 2000. This suggests that the proposed Z-test is consistent in power. On the other hand, the ERP results of

TABLE 2
Comparison of the MLE and the LSE (in terms of estimation efficiency and computation time) with 500 replications.

Method	n	Density	$\flat$	SE	SE*	ERP	Time
MLE-NR	245	0.0225	0.0072	0.0978	0.0957	54.8%	0.018
	490	0.0112	-0.0007	0.0690	0.0675	83.2%	0.120
	2450	0.0022	0.0012	0.0299	0.0303	100.0%	6.222
	4900	0.0011	0.0009	0.0209	0.0214	100.0%	45.767
MLE-BP	245	0.0225	0.0128	0.0957	0.1340	25.2%	0.026
	490	0.0112	0.0054	0.0673	0.0946	57.4%	0.032
	2450	0.0022	0.0070	0.0291	0.0424	100.0%	0.312
	4900	0.0011	0.0063	0.0203	0.0300	100.0%	1.213
LSE	245	0.0225	0.0049	0.1006	0.0993	54.4%	0.003
	490	0.0112	-0.0024	0.0712	0.0695	82.8%	0.005
	2450	0.0022	0.0010	0.0306	0.0308	100.0%	0.033
	4900	0.0011	0.0006	0.0215	0.0218	100.0%	0.053

zero ρ (i.e., $\rho = 0$) are always close to the nominal level 5%, suggesting that the proposed Z-test can control Type I error very well. This is not surprising since the difference between SE and SE* is very small, showing that the true standard error can be well approximated by the estimators derived in Section 2.3, in accordance with Theorem 1.

3.2. Comparison of the LSE vs. the MLE

We compare the LSE and the MLE in terms of estimation efficiency and computational efficiency. In this study, we fix $\rho = 0.2$ and $\sigma^2 = 1$. In order to implement the MLE, two algorithms are considered. The first one is the traditional Newton-Raphson algorithm, which is expected to be accurate but slow. For fast computation, we consider the toolbox in *http://www.spatial-statistics.com/software_index.htm*. The function *far* is used and the log-determinant algorithm of Barry and Pace (1999) is implemented. This algorithm provides an approximated solution to the MLE of (1.1); it is expected to be faster than the Newton-Raphson algorithm.

Following Lee and Liu (2010), we generate W as follows. Let W_A be the weighting matrix for the study of crimes across 49 districts in Columbus, Ohio (Anselin, 2013). This is a contiguity matrix constructed based on the latitude and longitude coordinates of the districts. Fix the sample size to be $n = 49m$, where $m = 5$, 10, 50, or 100. Therefore, the network size varies from 245 to 4,900. Accordingly, the spatial weighting matrix W is generated as $I_m \bigotimes W_A$, where $\bigotimes$ denotes the Kronecker product operator. The response variable Y is generated according to (1.1). We compare three different estimators of ρ, including the MLE computed by the Newton-Raphson algorithm (denoted as the MLE-NR), the MLE computed by the log-determinant algorithm of Barry and Pace (1999) with the Matlab function *far* (denoted as the MLE-BP), and the new LSE. For each parameter setup, the experiment is randomly replicated for $M = 500$ times. To compare different methods' computational efficiency, their average CPU times (Time) for each experiment are also reported. The detailed results are summarized in Table 2.

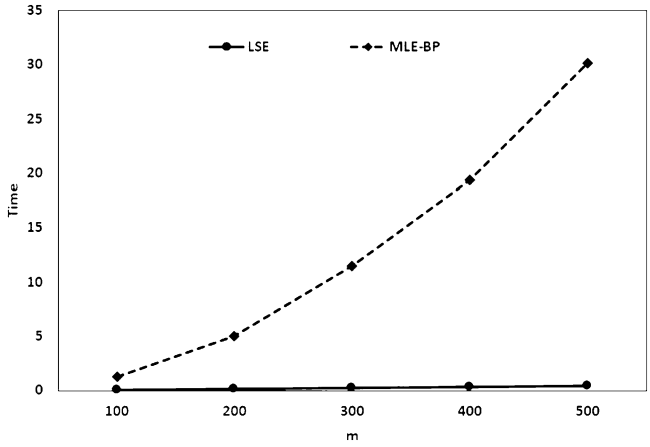

FIG 1. *Average Computation Time. m in horizontal axis controls the sample size as* $n = 49m$. *The vertical axis represents the average computation time based on 100 simulation replications. The network size varies from n =4,900 to 24,500. The solid line for the LSE and the dash line for the MLE-BP.*

Based on Table 2, we make the following interesting observations. First, the biases of all the estimation approaches are close to 0. Second, all the three estimators are fairly comparable in terms of SE. However, the SE* (i.e., the SE estimate) results of the MLE-BP are seriously biased. This makes the corresponding statistical inference problematic. Lastly, in terms of computational efficiency, the MLE-NR is the worst; the MLE-BP is substantially better in terms of the average computation time (i.e., Time); and the LSE is the best. This is particularly true for a large sample size (i.e., $n = 4,900$). See Figure 1 for further numerical evidence. Based on the above, the LSE method is the only method (among the three in Table 2) which can deliver highly accurate estimation results and statistically sound inference, yet with a minimal computational cost.

3.3. Performance of the sample-LSE

Here we demonstrate finite sample performance of the Sample-LSE method. The data are simulated in the same manner as in Section 3.1 based on Examples 1–3. We fix the whole network size to be $n = 20,000$ and the autocorrelation parameter $\rho \in \{0, 0.2\}$. Various sizes of the network sample are considered: $n_s =$ 2,000, 5,000, and 10,000. The experiments are replicated in a similar manner as before with $M =$1,000.

Two different sampling mechanisms are considered: (a) Simple Random Sampling (SNS); (b) Snowball Sampling (SNOW). In the SNOW, the following steps are conducted: (1) We start with $s_s = 10$ selected nodes and define the set to be $\mathcal{S}_{y_0}$ with $|\mathcal{S}_{y_0}| = s_s = 10$; (2) In the kth $(1 \leq k \leq K)$ step , all of the n_k connected friends of nodes in $\mathcal{S}_{y_{k-1}}$ are selected and added to the sample,

TABLE 3
Summary of Sample-LSE simulation results of SRS for the 3 examples with 1000 replications.

	ρ=0				ρ=0.2			
n	$\flat$	SE	SE*	ERP	$\flat$	SE	SE*	ERP
EXAMPLE 1								
2000	0.000	0.024	0.023	5.6%	0.001	0.025	0.024	100.0%
5000	0.000	0.016	0.016	5.6%	0.001	0.017	0.016	100.0%
10000	0.000	0.012	0.012	4.7%	0.000	0.012	0.012	100.0%
EXAMPLE 2								
2000	0.001	0.025	0.024	5.2%	0.001	0.026	0.025	100.0%
5000	0.000	0.016	0.016	4.0%	0.000	0.017	0.017	100.0%
10000	0.000	0.013	0.013	5.5%	0.000	0.013	0.013	100.0%
EXAMPLE 3								
2000	-0.000	0.023	0.023	5.1%	0.000	0.046	0.046	98.5%
5000	0.000	0.016	0.016	6.4%	0.001	0.031	0.031	100.0%
10000	0.001	0.012	0.012	5.4%	0.000	0.023	0.023	100.0%

TABLE 4
Summary of Sample-LSE simulation results of SNOW for the 3 examples with 1000 replications.

	ρ=0				ρ=0.2			
n	$\flat$	SE	SE*	ERP	$\flat$	SE	SE*	ERP
EXAMPLE 1								
2000	0.001	0.027	0.025	6.6%	0.001	0.028	0.026	100.0%
5000	0.000	0.017	0.016	5.0%	0.000	0.017	0.017	100.0%
10000	0.000	0.013	0.012	5.8%	0.000	0.013	0.012	100.0%
EXAMPLE 2								
2000	0.001	0.025	0.026	4.8%	0.001	0.027	0.027	100.0%
5000	0.001	0.017	0.017	4.4%	0.001	0.018	0.017	100.0%
10000	0.001	0.012	0.013	4.2%	0.001	0.013	0.013	100.0%
EXAMPLE 3								
2000	0.001	0.041	0.043	6.0%	0.000	0.044	0.044	99.0%
5000	0.000	0.026	0.027	5.0%	0.000	0.029	0.029	100.0%
10000	0.000	0.020	0.020	4.4%	0.000	0.021	0.022	100.0%

which makes $\mathcal{S}_{y_k}$. (3) If in the kth step, there is no more connected node that could be selected, then a random node $i \in \mathcal{S} \setminus \mathcal{S}_{y_{k-1}}$ is added to the sample set, which leads to $\mathcal{S}_{y_k}$. (4) Repeat (2) and (3) until at least n_s nodes are selected in the Kth step, which means $|\mathcal{S}_{yK}| \geq n_s$. (5) Randomly select n_K^* nodes from the sampled n_K nodes in the Kth step, which makes $|\mathcal{S}_{y_{K-1}}| + n_K^* = n_s$. Thus $|\mathcal{S}_y| = |\mathcal{S}_{yK}| = n_s$.

The detailed results are presented in Tables 3 and 4. We observe that: (1) the Sample-LSE is consistent with ignorable bias and decreasing SE as $n \to \infty$; (2) the SE estimator developed for the Sample-LSE in Section 2.4 works quite well, because the difference between SE and SE* values reported in Tables 3 and 4

TABLE 5
Comparison of the MLE and the Sample-LSE (in terms of estimation efficiency and computation time) with 500 replications.

Method	n_s	$\flat$	SE	SE*	ERP	Time
MLE-NR	1000	0.1519	0.0354	0.0364	25.4%	0.075
	2000	0.1414	0.0233	0.0222	74.0%	1.296
	5000	0.0997	0.0160	0.0155	100.0%	47.871
MLE-BP	1000	0.1541	0.0341	0.0507	7.4%	0.027
	2000	0.1458	0.0220	0.0306	41.2%	0.067
	5000	0.1083	0.0153	0.0212	100.0%	0.755
Sample-LSE	1000	0.0016	0.0376	0.0382	100.0%	0.135
	2000	-0.0012	0.0264	0.0279	100.0%	0.153
	5000	0.0005	0.0191	0.0192	100.0%	0.183

is very small; and (3) as a consequence, the corresponding Z-test can control Type I error quite well (since the ERP values associated with $\rho = 0$ are fairly close to their nominal level 5%) and is consistent in power (since the ERP values associated with $\rho = 0.2$ are all close to 100%).

3.4. Comparison of the sample-LSE vs. the MLE

In this subsection, we compare three methods, based on the sampled data: the MLE-NR, the MLE-BP, and the Sample-LSE. Data are simulated in the same manner as in Section 3.2. For the purpose of illustration, we fix $n = 9,800$, $\rho = 0.2$, and $n_s \in \{1000, 2000, 5000\}$. In order to implement the two MLE methods, which are the MLE-NR and the MLE-BP, we treat the sampled network structure as if they were the whole network. This leads to another row-normalized W matrix based on the sampled data only (Chen et al., 2013). Then the isolated nodes are eliminated from the data accordingly. We replicate the experiment M =500 times and report the summary in Table 5.

One immediate observation is that, both the MLE-NR and the MLE-BP methods are inconsistent with the sampled data, because their estimation biases are clearly above 0. Such a finding is not surprising and is essentially consistent with that of Chen et al. (2013). In contrast, the Sample-LSE remains to be consistent and statistically valid.

3.5. Performance of the mLSE

In this subsection, we demonstrate finite sample performance of the mLSE. To allow for different features of adjacency matrices, we generate W_1 and W_2 in the same manner as in Examples 3 and Example 2, separately. The parameters are set as $\rho_1 \in \{0, 0.1\}$ and $\rho_2 = 0.2$. Various network sizes are considered: $n =$ 2,000, 5,000, 10,000 and 20,000. The experiments are replicated in the same way as before with M =1,000.

Numerical results of the mLSE are summarized in Table 6. For all the three examples, the ERP results of nonzero ρ_ls ($l \in \{1, 2\}$) are always larger than

TABLE 6
Summary of mLSE simulation results with 1000 replications.

n	$\flat_{\rho_1}$	SE_{ρ_1}	$\text{SE}^*_{\rho_1}$	ERP_{ρ_1}	$\flat_{\rho_2}$	SE_{ρ_2}	$\text{SE}^*_{\rho_2}$	ERP_{ρ_2}
	ρ_1=0				ρ_2=0.2			
2000	0.001	0.031	0.032	5.5%	0.006	0.034	0.032	100.0%
5000	0.000	0.020	0.020	6.0%	0.005	0.022	0.020	100.0%
10000	-0.000	0.014	0.014	4.8%	0.005	0.016	0.014	100.0%
20000	0.000	0.010	0.010	5.1%	0.005	0.011	0.010	100.0%
	ρ_1=0.1				ρ_2=0.2			
2000	0.003	0.031	0.030	91.2%	0.008	0.035	0.032	100.0%
5000	0.001	0.020	0.019	99.7%	0.006	0.023	0.020	100.0%
10000	0.001	0.014	0.013	100.0%	0.006	0.017	0.014	100.0%
20000	0.001	0.010	0.009	100.0%	0.006	0.012	0.010	100.0%

95%. This means the proposed Z-test is consistent in power. By contrast, the ERP results of zero ρ_1 are always close to the nominal level 5%, which indicates that the proposed Z-test controls Type I error very well. This is because the difference between SE and SE* is very small, suggesting that the true standard error can be well approximated by the estimators derived in Section 2.5, in accordance with Theorem 4.

3.6. Sina Weibo network analysis

We apply the LSE methods to a real social network collected from Sina Weibo (*www.weibo.com*), the largest Twitter-type social media in China. The goal of this study is to understand how the Sina Weibo users interact with each other in terms of their posting activity. To collect the data, we start with the Sina Weibo accounts of four major online travel agencies in China. We randomly select 5,000 nodes from each travel agency's followers and collect the followers' friends. To better mimic a sparse network, only active users is with out-degree ($d_i = \sum_j a_{ij}$) no more than 20 are kept. The final network has $n = 557,818$ nodes. Their follower-followee relationships are recorded by the adjacency matrix A. In total, the network includes $\sum a_{ij}$ =1,496,399 edges and $\sum_{i<j} a_{ij}a_{ji} = 535,408$ mutually connected pairs. The density of the network is 4.8×10^{-6}, which is extremely sparse.

For each node, the response is defined to be total amount of its posted messages in log-scale. The responses are standardized to be mean 0 and variance 1. For such a large network size, both the MLE-NR and the MLE-BP are computationally too expensive to be implemented. However, the LSE can be easily computed using a personal computer within 58 seconds. It gives the estimate $\hat{\rho} = 0.125$ with $\widehat{\text{SE}} = 1.4 \times 10^{-3}$, implying that the social interaction is statistically significant at 5% level. This suggests that on average, a Weibo user's posting activity does positively correlate with his/her connected friends.

To further demonstrate usefulness of the Sample-LSE method, we conduct an interesting simulation study as follows. Specifically, we treat the above sampled nodes and edges as if they were the whole social network. We then treat the "whole network" LSE estimator $\hat{\rho} = 0.125$ as if it were the true parameter. This

TABLE 7
Real example: estimation results for the Sina Weibo network.

n	$\flat$	SE	SE*	ERP
2,000	0.000	0.019	0.018	99.9%
5,000	0.001	0.011	0.012	100.0%
10,000	0.000	0.008	0.008	100.0%
20,000	0.000	0.006	0.006	100.0%

allows us to conduct the simulation experiments in a similar manner as we have stated in Sections 3.3 and 3.4. The results are presented in Table 7. By Table 7, $n_s = 2,000$ is large enough to detect a positive and statistically significant ρ. The corresponding ERP (i.e., power) is as large as 99.9%. However, $n_s = 2,000$ only accounts for about $n_s/n = 0.36\%$ of the entire network size. Consequently, the saving in both sampling and computational costs is significant.

Last, we demonstrate the usefulness of the proposed mLSE method. According to Holland and Leinhardt (1981), and Huang et al. (2016), for different nodes i and j ($i \neq j$), mutual relationship ($a_{ij} = a_{ji} = 1$) and asymmetric relationship ($a_{ij} + a_{ij} = 1$) are different types of relationship, i.e., friends and fans. To test which type of relationship has a stronger impact on one's posting activity in Weibo network, we construct A_1 and A_2 based on the above mentioned $n = 557,818$ nodes. Define $A_1 = (a_{ij}) \in \mathbb{R}^{n\times n}$ to record all the relationships with $a_{ij} = a_{ji} = 1$, and $A_2 = (a_{ij}) \in \mathbb{R}^{n\times n}$ to record those with $a_{ij} + a_{ij} = 1$. Thus W_1 and W_2 could be obtained. Correspondingly, ρ_1 measures the social interaction impact from mutually connected friends, and ρ_2 measures that from asymmetrically connected friends. Thus we obtain the mLSE for ρ_1 is 0.115 with $\widehat{\text{SE}}_{\rho_1} = 1.3\times10^{-3}$, and that for ρ_2 is 0.061 with $\widehat{\text{SE}}_{\rho_2} = 1.2\times10^{-3}$. Both the coefficients are statistically significant at 5% level. This suggests that on average, (1) a Weibo user's posting activity positively correlates with his/her mutually connected or asymmetrically connected friends; (2) mutual relationship has a stronger impact on one's posting activity than asymmetric relationship.

4. Conclusion

To conclude this article, we discuss four interesting topics for future study. First, the proposed LSE method requires the adjacency matrices to be sparse. It would be intriguing to study the problem without the network sparsity assumption. Second, the proposed approach is developed for the SAR model with no covariates. A natural question would be how to extend the approach by taking covariate information into consideration. Third, we have assumed that the adjacency A is pre-determined and the response Y is generated base on the weighting matrix W. However in practice, network structure could be influenced by individual features, implying a possible two-way relationship (Robins et al., 2001). How to model this bilateral relationship is another interesting topic for future study. Lastly, we have only empirically verified the performance of simple random sampling and snow ball sampling. We find both methods work fairly

well and comparable. However, what would be the effect of a general sampling method is not clear. Further study along this direction is needed.

Appendix A

A.1. Proof of Proposition 1 and Proposition 3

For both model (1.1) and model (2.4), without the loss of generality, define $\Sigma_{i.}$ and $\Sigma_{.i}$ to be the ith row and column of Σ separately. Further define $\Sigma_{i(-i)} \in \mathbb{R}^{1\times(n-1)}$ to be the ith row of Σ with out the i-th element. Thus $\Sigma_{(-i)i} \in \mathbb{R}^{(n-1)\times 1}$ can be defined in the similar manner. $\Sigma_{(-i)(-i)} \in \mathbb{R}^{(n-1)\times(n-1)}$ is defined to be Σ without its ith row and ith column. Further let $\Sigma = (\sigma_{ij})$.

According to the distribution of Y, we know that the joint distribution of $\{Y_i, Y_{(-i)}^\top\}^\top$ is $N(0, \Sigma_i)$ with $\Sigma_i = \{\sigma_{ii}, \Sigma_{i(-i)}; \Sigma_{(-i)i}, \Sigma_{(-i)(-i)}\}$. Therefore, by the normality assumption of $\mathcal{E}$ in both of the models, we know that the conditional distribution of Y_i given $Y_{(-i)}$ is $N\Big(\Sigma_{i(-i)}\Sigma_{(-i)(-i)}^{-1}Y_{(-i)}, \sigma_{ii} - \Sigma_{i(-i)}\Sigma_{(-i)(-i)}^{-1}\Sigma_{(-i)i}\Big)$. In accordance with Σ_i, we define Ω_i and thus we have, $\Omega_i = \Sigma_i^{-1} = \big(\Omega_{i,11}, \Omega_{i,12}; \Omega_{i,21}, \Omega_{i,22}\big)$, $\Sigma_{(-i)(-i)}^{-1} = \Omega_{i,12} - \Omega_{i,21}\Omega_{i,11}^{-1}\Omega_{i,12}$, and $\Sigma_{i(-i)} = -\Omega_{i,11}^{-1}\Omega_{i,12}(\Omega_{i,22} - \Omega_{i,21}\Omega_{i,11}^{-1}\Omega_{i,12})^{-1}$. As a result, $E\Big\{Y_i|Y_{(-i)}\Big\} = \Sigma_{i(-i)}\Sigma_{(-i)(-i)}^{-1}Y_{(-i)} = -\Omega_{i,11}^{-1}\Omega_{i,12}Y_{(-i)}$.

To prove Proposition 1, recall that $W_{.i} \in \mathbb{R}^{n\times 1}$ and $W_{i.} \in \mathbb{R}^{n\times 1}$ are the i-th column and row of W, respectively. For convenience, further define $W_{i(-i)} \in \mathbb{R}^{1\times(n-1)}$ to be the i-th row of W without the i-th element, $W_{(-i)i} \in \mathbb{R}^{(n-1)\times 1}$ to be the i-th column of W without the i-th element, and $W_{(-i)(-i)} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the weighting matrix W but without the i-th column and ith row. Thus we have, $\Omega_i = \sigma^{-2}\big[I_n - \rho\{w_{ii}, W_{i(-i)}; W_{(-i)i}, W_{(-i)(-i)}\}\big]^\top\big[I_n - \rho\{w_{ii}, W_{i(-i)}; W_{(-i)i}, W_{(-i)(-i)}\}\big]$. Simple calculations imply that,

$$\begin{aligned}
\sigma^2\Omega_{i,11} &= (1-\rho w_{ii})^2 + \rho^2 W_{(-i)i}^\top W_{(-i)i} = 1 + \rho^2 W_{.i}^\top W_{.i},\\
\sigma^2\Omega_{i,12} &= -(1-\rho w_{ii})\rho W_{i(-i)} - \rho W_{(-i)i}^\top\{I_n - \rho W_{(-i)(-i)}\}\\
&= -\rho\{W_{i(-i)} + W_{(-i)i}^\top\} + \rho^2 W_{(-i)i}^\top W_{(-i)(-i)}.
\end{aligned}$$

Note that $w_{ii} = 0$. Define $\gamma_i^*(\rho) = -\Omega_{i,12}^\top(\Omega_{i,11}^{-1})^\top \in \mathbb{R}^{n-1}$. By combing the results above, we obtain that

$$E\big\{Y_i|Y_{(-i)}\big\} = Y_{(-i)}^\top\gamma_i^*(\rho) = \sum_{k\neq i}\frac{\rho(w_{ik}+w_{ki}) - \rho^2\sum_j w_{ji}w_{jk}}{1+\rho^2\sum_j w_{ji}^2}Y_k,$$

which completes the entire calculations of Proposition 1.

To prove Proposition 3, $W_{l,i.}$, $W_{l,.i}$, $W_{l,(-i)i}$, $W_{l,i(-i)}$ and $W_{l,(-i)(-i)}$ is defined for $l \in \{1,2\}$ and $1 \le i \le n$ similarly. We could define Ω_i to be $\Omega_i = \sigma^{-2}\big[I_n -$

$\sum_l \rho_l\{w_{l,ii}, W_{l,i(-i)}; W_{l,(-i)i}, W_{l,(-i)(-i)}\}\big]^\top \big[I_n - \sum_l \rho_l\{w_{l,ii}, W_{l,i(-i)}; W_{l,(-i)i}, W_{l,(-i)(-i)}\}\big]$ for model (2.4). This leads to,

$$\begin{aligned}
\sigma^2\Omega_{i,11} &= 1 + (\rho_1 W_{1,.i} + \rho_2 W_{2,.i})^\top(\rho_1 W_{1,.i} + \rho_2 W_{2,.i}),\\
\sigma^2\Omega_{i,12} &= -\rho_1 W_{1,i(-i)} - \rho_2 W_{2,i(-i)}\\
&\quad - (\rho_1 W_{1,(-i)i}^\top + \rho_2 W_{2,(-i)i}^\top)\Big\{I_{n-1} - \rho_1 W_{1,(-i)(-i)} - \rho_2 W_{2,(-i)(-i)}\Big\}\\
&= -\rho_1\Big\{W_{1,i(-i)} + W_{1,(-i)i}^\top\Big\} - \rho_2\Big\{W_{2,i(-i)} + W_{2,(-i)i}^\top\Big\}\\
&\quad + \Big\{\rho_1 W_{1,(-i)i}^\top + \rho_2 W_{2,(-i)i}^\top\Big\}\Big\{\rho_1 W_{1,(-i)(-i)} + \rho_2 W_{2,(-i)(-i)}\Big\}.
\end{aligned}$$

If we further define $\gamma^*(\theta) = -\Omega_{i,12}^\top(\Omega_{i,11}^{-1})^\top \in \mathbb{R}^{n-1}$, thus the conclusion in Proposition 3 could be obtained.

A.2. Notations and the verification of (2.3) *and* (2.5)

Before the proof of the theorems, we firstly verify (2.3), (2.5) and define some useful notations for the first and second order derivatives of the objective functions .

Verification of (2.3). First, we consider model (1.1). Let $W_{ii}^0 \in \mathbb{R}^{n\times n}$ to have the same elements as W except that the i-th column and the i-th row equal to 0 for notation convenience for $1 \le i \le n$. Recalling that $W_{.i} \in \mathbb{R}^n$ is defined to be the i-th column of W, $W_{i.} \in \mathbb{R}^n$ is defined to be the ith row of W, respectively. Then by Proposition 1, we could verify that, for $1 \le i \le n$,

$$\begin{aligned}
E\{Y_i|Y_{(-i)}\} &= (1+\rho^2 W_{.i}^\top W_{.i})^{-1}\{\rho W_{i.} + \rho W_{.i}^\top - \rho^2 W_{.i}^\top W_{ii}^0\}Y\\
&= (1+\rho^2||W_{.i}||^2)^{-1}\Big[\{\rho W_{i.} + \rho W_{.i}^\top - \rho^2 W_{.i}^\top W\}Y\\
&\quad +\rho^2 W_{.i}^\top W_{.i}Y_i + Y_i - Y_i\Big]\\
&= (1+\rho^2||W_{.i}||^2)^{-1}\Big[\{\rho W_{i.} + \rho W_{.i}^\top - \rho^2 W_{.i}^\top W\}Y - Y_i\Big] + Y_i.
\end{aligned}$$

Therefore, the objective function can be also written in the form of (2.3).

Next, we derive the expressions for $\dot{Q}(\rho)$ and $\ddot{Q}(\rho)$. It could also be shown that $\dot{d}(\rho)$, $\ddot{d}(\rho)$ can be expressed as follows,

$$\dot{d}_\rho = -2\rho d_\rho^2 \mathrm{diag}(W^\top W), \tag{A.1}$$

$$\ddot{d}_\rho = -2d_\rho^2\mathrm{diag}(W^\top W) + 8\rho^2 d_\rho^3 \mathrm{diag}^2(W^\top W) \tag{A.2}$$

Define $F = d_\rho\Omega_\rho Y$. Thus $Q(\rho) = F^\top F$. Define $\dot{F}$ and $\ddot{F}$ to be the first and second order derivatives of F with respect to ρ, respectively. Then, it could be obtained that, $\dot{Q}(\rho) = 2F^\top\dot{F} = 2\tilde{\varepsilon}M_1^\top M_2\tilde{\varepsilon}$, where $\tilde{\varepsilon} = \sigma^{-1}\mathcal{E}$. and $\ddot{Q}(\rho) = 2\dot{F}^\top\dot{F} + 2F^\top\ddot{F} = 2\tilde{\varepsilon}^\top M_2^\top M_2\tilde{\varepsilon} + 2F^\top\ddot{F}$.

VERIFICATION OF (2.5). Second, we consider model (2.4). By Proposition 3, $W^*_{\rho,.i} = \rho_1 W_{1,.i} + \rho_2 W_{2,.i}$ $(1 \le i \le n)$ is defined for notation convenience. Similarly define $W_{1,.i}$, $W_{2,.i}$, $W_{1,i.}$, $W_{2,i.}$, $W^0_{1,ii}$ and $W^0_{2,ii}$ as in the VERIFICATION OF (2.3). Thus it could be calculated that,

$$\begin{aligned}
E\{Y_i|Y_{(-i)}\} &= \big(1 + W^{*\top}_{\rho,.i} W^*_{\rho,.i}\big)^{-1} \Big\{ \sum_l \rho_l (W_{l,i.} + W^\top_{l,.i}) - \sum_l \rho_l^2 W^\top_{l,.i} W^0_{l,ii} \\
&\quad - \rho_1\rho_2 (W^\top_{1,.i} W^0_{2,ii} + W^\top_{2,.i} W^0_{1,ii}) \Big\} Y \\
&= \big(1 + W^{*\top}_{\rho,.i} W^*_{\rho,.i}\big)^{-1} \Bigg[\Big\{ \sum_l \rho_l (W_{l,i.} + W^\top_{l,.i}) \Big\} Y - \big(\sum_l \rho_l^2 W^\top_{l,.i} W_l \big) Y \\
&\quad - \rho_1\rho_2 (W^\top_{1,.i} W_2 + W^\top_{2,.i} W_1) Y + W^{*\top}_{\rho,.i} W^*_{\rho,.i} Y_i + Y_i - Y_i \Bigg] \\
&= Y_i + \big(1 + W^{*\top}_{\rho,.i} W^*_{\rho,.i}\big)^{-1} \Bigg[\Big\{ \sum_l \rho_l (W_{l,i.} + W^\top_{l,.i}) - \sum_l \rho_l^2 W^\top_{l,.i} W_l \\
&\quad - \rho_1\rho_2 (W^\top_{1,.i} W_2 + W^\top_{2,.i} W_1) \Big\} Y - Y_i \Bigg],
\end{aligned}$$

for $l \in \{1, 2\}$ and $1 \le i \le n$. After this, simple calculation written in vector form will show that $Q(\theta) = ||d_\theta \Omega_\theta Y||^2$.

A.3. Proof of Proposition 2

The conclusion in Proposition 2 could be obtained as follows. Under the assumption (2) in Proposition 2, $\sum_j \omega^2_{ji}$ and $\sum_j \omega_{ji}\omega_{jk}$ are both summations of finite terms for any node i $(1 \le i \le n)$. Thus the calculation complexity of $Q(\rho)$ and its derivatives is linear in the network size n. See Appendix A.2 for the detailed expressions for the forms of its derivatives. If the Newton Raphson iteration converges in K steps, where K is finite, then complexity of estimation is also $O(n)$. This completes the proof.

Appendix B

To investigate the theoretical property of the LSE, several useful lemmas are proved before the establishment of the theorems.

Lemma 1. *Define $B_1 \preccurlyeq B_2$ if $b^{(1)}_{ij} \le b^{(2)}_{ij}$, where $B_1 = (b^{(1)}_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ and $B_2 = (b^{(2)}_{ij}) \in \mathbb{R}^{n_1 \times n_2}$ are two arbitrary matrices. Define $|B|_e = (|b_{ij}|)$ for any arbitrary matrix B. Then, we have the following results.*

(1) Assume the condition (C1) is satisfied. Then, we can find an integer sufficiently large such that for $c_K \ge K$, $W^{c_K} \preccurlyeq c_w \mathbf{1}_n \pi^\top$, where c_w is a positive constant. Define $\mathcal{W}_0 = \sum_{m=0}^K W^m + \mathbf{1}_n \pi^\top$, and $\mathcal{W}_q = W^q \mathcal{W}_0$. For positive

integer q, further define $\Delta_n = (\log n)^{2(K+q)}$ if $\delta = 1/2$ and $\Delta_n = n^{1/2-\delta}$ if $0 < \delta < 1/2$, and δ here is the positive constant, which is defined in condition (C1). Then, we have, for $0 \le q_1, \cdots, q_4 \le 1$, and finite positive integers r_1, r_2, r_3 and q, as $n \to \infty$,

$$\lambda_{\max}(W^\top W) = O\{(\log n)^2\}, \quad \lambda_{\max}(\mathcal{W}_q^\top \mathcal{W}_q) = O(\Delta_n), \tag{B.1}$$

$$n^{-2} tr\Big\{(W^{r_1} W^{\top r_2})^{r_3}\Big\} \to 0, \tag{B.2}$$

$$n^{-2} tr\Big\{(W^\top W)^{q_1} (\mathcal{W}_q^\top \mathcal{W}_q)^{q_2} (W^\top W)^{q_3} (\mathcal{W}_q^\top \mathcal{W}_q)^{q_4}\Big\} \to 0. \tag{B.3}$$

(2) Recall that $S = I_n - \rho W$, then we have

$$|S^{-1}|_e \preccurlyeq c_s \mathcal{W}_0, \quad |W^q S^{-1}|_e \preccurlyeq c_s \mathcal{W}_q,$$

where $c_s = \max\{1, c_w c_\rho\}$, and $c_\rho = (1-\rho)^{-1}\rho^{K+1}$.
(3) For the notations of matrices defined in Section 2 and Appendix A.2, we have the following conclusions:

$$|d_\rho|_e \preccurlyeq c_{0m} I_n, \quad |\dot{d}_\rho|_e \preccurlyeq c_{1m} W^\top W, \tag{B.4}$$

$$|\ddot{d}_\rho|_e \preccurlyeq c_{2m}\{W^\top W + (W^\top W)^2\}, \tag{B.5}$$

$$|S|_e \preccurlyeq I_n + c_{1s} W, \tag{B.6}$$

$$|M_1|_e \preccurlyeq c_{1M}(I_n + W^\top), \quad |M_2|_e \preccurlyeq c_{2M}\widetilde{\mathcal{W}} \tag{B.7}$$

$$|M|_e \preccurlyeq c_{3M}(\widetilde{\mathcal{W}} + W\widetilde{\mathcal{W}}), \tag{B.8}$$

$$\lambda_{\max}(\widetilde{\mathcal{W}}^\top \widetilde{\mathcal{W}}) = O\{(\log n)^6 n^{1/2-\delta}\}. \tag{B.9}$$

where $\widetilde{\mathcal{W}} = W^\top + W^\top W + W^\top W W^\top + W^\top \mathcal{W}_1$.

Proof. In this section, we prove (1)–(3) in Lemma 1 subsequently.

Proof of (1). We prove each of the results in the following parts.

Proof of (B.1). Recall that we have defined $\mathbb{W}_1 = W + W^\top$. As a result, we have $\lambda_{\max}(W^\top W) = \max_{\|u\|=1} u^\top (W^\top W) u$. Thus, for any $u \in \mathbb{R}^n$, $u^\top W^\top W u \le |u|_e^\top \mathbb{W}_1^\top \mathbb{W}_1 |u|_e \le \lambda_{\max}^2(\mathbb{W}_1)\|u\|^2 = \lambda_{\max}^2(\mathbb{W}_1)$. By condition (C2), we could draw the conclusion that $\lambda_{\max}(W^\top W) = O\{(\log n)^2\}$. Furthermore, we could verify that $\mathcal{W}_q = \sum_{m=0}^K W^{m+q} + \mathbf{1}_n \pi^\top$. Thus, by the Cauchy-Schwarz inequality, we could calculate that

$$\lambda_{\max}(\mathcal{W}_q^\top \mathcal{W}_q) \le c_{wq}\Big[\sum_{m=0}^K \lambda_{\max}\{W^{m+q}(W^{m+q})^\top\} + n\lambda_{\max}(\pi\pi^\top)\Big],$$

where c_{wq} is a finite constant. In addition, it could be proved $\lambda_{\max}\{W^{m+q}(W^{m+q})^\top\} \le \lambda_{\max}\{\mathbb{W}_1^{2(m+q)}\} = O\{(\log n)^{2(m+q)}\}$ by similar techniques. Next, by condition (C1), one could calculate that $n\lambda_{\max}(\pi\pi^\top) =$

$n(\pi^\top\pi) = n^{1/2-\delta}$. As a result, if $\delta = 1/2$, $(\log N)^{2(K+q)}$ will dominate the diverging speed; otherwise, it will diverge with the speed in the order of $n^{1/2-\delta}$.

PROOF OF (B.2). By $W \preccurlyeq \mathbb{W}_1$ and $W^\top \preccurlyeq \mathbb{W}_1$, we could be verify the conclusion that $n^{-2}\mathrm{tr}\{(W^{r_1}W^{\top r_2})^{r_3}\} \le n^{-2}\mathrm{tr}(W^{*(r_1+r_2+r_3)}) \le n^{-1}\lambda_{\max}^{r_1+r_2+r_3}(\mathbb{W}_1)$. As $\lambda_{\max}(\mathbb{W}_1) = O(\log n)$, then it can be calculated that $n^{-1}\lambda_{\max}^{r_1+r_2+r_3}(\mathbb{W}_1) \to 0$ as $n \to \infty$. Thus, (B.2) can be obtained.

PROOF OF (B.3). We assume $0 < \delta < 1/2$. Thus, it can be obtained that

$$\begin{aligned} &n^{-2}\mathrm{tr}\{(W^\top W)^{q_1}(\mathcal{W}_q^\top\mathcal{W}_q)^{q_2}(W^\top W)^{q_3}(\mathcal{W}_q^\top\mathcal{W}_q)^{q_4}\} \\ &\quad\le n^{-1}\lambda_{\max}^{q_1+q_3}(W^\top W)\lambda_{\max}^{q_2+q_4}(\mathcal{W}_q^\top\mathcal{W}_q). \end{aligned}$$

By (B.1), as $n \to \infty$, $n^{-1}(\log n)^{2(q_1+q_3)}n^{(1/2-\delta)(q_2+q_4)} \le n^{-2\delta}(\log n)^4 \to 0$. Then, (B.3) holds.

Proof of (2). By condition (C1), one can obtain the conclusion that there exists an integer $K > 0$, such that for $n > K$, $W^n \preccurlyeq c_w\mathbf{1}_N\pi^\top$. Thus,

$$\begin{aligned} \sum_{m=0}^{\infty}\rho^m W^m &= \sum_{m\le K}\rho^m W^m + \sum_{m>K}\rho^m W^m \\ &\preccurlyeq \sum_{m=0}^{K}\rho^m W^m + c_w\mathbf{1}_N\pi^\top\Big(\sum_{m>K}\rho^m\Big) \\ &\preccurlyeq \sum_{m=0}^{K} W^m + c_\rho c_w\mathbf{1}_N\pi^\top \preccurlyeq c_s\mathcal{W}_0. \end{aligned}$$

As a result, $|S^{-1}|_e \preccurlyeq c_s\mathcal{W}_0$ can be obtained. Additionally, $|W^qS^{-1}|_e \preccurlyeq c_s\mathcal{W}_q$ can be subsequently verified.

Proof of (3). The proofs of (B.4)–(B.8) are similar, and we only calculate the upper bound for $|M_2|$ as an example. We have $|M_2|_e \preccurlyeq \sigma\{c_{1m}W^\top W(I_n + c_{1s}W^\top) + c_{0m}W^\top + c_{0m}(I_n + c_{1s}W^\top)W(\sum_{m=0}^{K}W^m + c_\rho c\mathbf{1}_n\pi^\top)\} \preccurlyeq c_{2M}(W^\top W + W^\top WW^\top + W^\top + W\mathcal{W}_0 + W^\top W\mathcal{W}_0) = c_{2M}\widetilde{\mathcal{W}}$.

Next, we prove (B.9). One can verify $\lambda_{\max}(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}) \le= c_{\widetilde{\mathcal{W}}}[\lambda_{\max}(\mathcal{W}_1^\top WW^\top\mathcal{W}_1) + \lambda_{\max}\{(W^\top W)^2\} + \lambda_{\max}\{(W^\top W)^3\} + \lambda_{\max}(W^\top W) + \lambda_{\max}(\mathcal{W}_1^\top\mathcal{W}_1)] = c_{\widetilde{\mathcal{W}}}\{\lambda_{\max}^2(W^\top W) + \lambda_{\max}^3(W^\top W) + \lambda_{\max}(W^\top W) + \lambda_{\max}(\mathcal{W}_1^\top\mathcal{W}_1) + \lambda_{\max}(\mathcal{W}_1^\top\mathcal{W}_1)\lambda_{\max}(WW^\top)\}$. Thus, the order is $O\{(\log n)^6 n^{1/2-\delta}\}$. □

Lemma 2. *Let $\tilde{\varepsilon}_i \in \mathbb{R}^1$ $(1 \le i \le n)$ be identically distributed variables. Assume the following conditions are satisfied:*
(1) $E(\tilde{\varepsilon}_i) = 0$ for $1 \le i \le n$;
(2) $E(\tilde{\varepsilon}_i, \tilde{\varepsilon}_j) = 0$ for any $i \ne j$;
(3) $E(\tilde{\varepsilon}_i\tilde{\varepsilon}_j\tilde{\varepsilon}_k) = 0$ for any $1 \le i, j, k \le n$;
(4) $E(\tilde{\varepsilon}_i^2) = 1$ and $E(\tilde{\varepsilon}_i^4) = \kappa_4$, where κ_4 is a finite positive constant.

Let $\tilde{\varepsilon} = (\tilde{\varepsilon}_1, \tilde{\varepsilon}_2, \cdots, \tilde{\varepsilon}_n)^\top \in \mathbb{R}^n$, $Q_1 = \tilde{\varepsilon}^\top \mathcal{M}_1 \tilde{\varepsilon}$, *and* $Q_2 = \tilde{\varepsilon}^\top \mathcal{M}_2 \tilde{\varepsilon}$, *where* $\mathcal{M}_1$ *and* $\mathcal{M}_2$ *are* $n \times n$ *dimensional matrices. Then, we have,*

$$cov(Q_1, Q_2) = tr(\mathcal{M}_1 \mathcal{M}_2^\top) + tr(\mathcal{M}_1 \mathcal{M}_2) + (\kappa_4 - 3) tr\big\{ diag(\mathcal{M}_1) diag(\mathcal{M}_2) \big\}. \tag{B.10}$$

Proof. Define $\mathcal{M}_1 = (m_{1,ij}) \in \mathbb{R}^{n \times n}$ and $\mathcal{M}_2 = (m_{2,ij}) \in \mathbb{R}^{n \times n}$. Then we have, $E(\tilde{\varepsilon}^\top \mathcal{M}_1 \tilde{\varepsilon}) = \mathrm{tr}(\mathcal{M}_1)$. Next, it can be calculated that

$$\begin{aligned}
&E\Big\{ (\tilde{\varepsilon}^\top \mathcal{M}_1 \tilde{\varepsilon})(\tilde{\varepsilon}^\top \mathcal{M}_2 \tilde{\varepsilon}) \Big\} = \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \sum_{m=1}^n m_{1,ij} m_{2,lm} E(\tilde{\varepsilon}_i \tilde{\varepsilon}_j \tilde{\varepsilon}_l \tilde{\varepsilon}_m) \\
&= \sum_{i \neq j} m_{1,ii} m_{2,jj} + \sum_{i \neq j} m_{1,ij} m_{2,ij} + \sum_{i \neq j} m_{1,ij} m_{2,ji} + \sum_i m_{1,ii} m_{2,ii} \kappa_4 \\
&= \mathrm{tr}(\mathcal{M}_1)\mathrm{tr}(\mathcal{M}_2) + \mathrm{tr}(\mathcal{M}_1 \mathcal{M}_2^\top) + \mathrm{tr}(\mathcal{M}_1 \mathcal{M}_2) \\
&\quad + \mathrm{tr}\Big\{ \mathrm{diag}(\mathcal{M}_1) \mathrm{diag}(\mathcal{M}_2) \Big\} (\kappa_4 - 3).
\end{aligned}$$

Therefore we have $E\{(\tilde{\varepsilon}^\top \mathcal{M}_1 \tilde{\varepsilon})(\tilde{\varepsilon}^\top \mathcal{M}_2 \tilde{\varepsilon})\} - E(\tilde{\varepsilon}^\top \mathcal{M}_1 \tilde{\varepsilon}) E(\tilde{\varepsilon}^\top \mathcal{M}_2 \tilde{\varepsilon}) = \mathrm{tr}(\mathcal{M}_1 \mathcal{M}_2^\top) + \mathrm{tr}(\mathcal{M}_1 \mathcal{M}_2) + \mathrm{tr}\{\mathrm{diag}(\mathcal{M}_1)\mathrm{diag}(\mathcal{M}_2)\}(\kappa_4 - 3)$. This completes the proof. □

Lemma 3. *Suppose* $\varepsilon = (\varepsilon_1, \cdots, \varepsilon_n) \in \mathbb{R}^n$. *Moreover, assume* $\max_i E(\varepsilon_i^4) \leq \kappa_4$, *and* $\max_i E(\varepsilon_i^2 - \sigma^2)^4 \leq c_\varepsilon$, *where* κ_4 *and* c_ε *are finite positive constants. Let*

$$Q = \varepsilon^\top \mathcal{M} \varepsilon - \sigma^2 tr(\mathcal{M}),$$

where $\mathcal{M} \in \mathbb{R}^{n \times n}$. *Let* $\mathbb{M} = |\mathcal{M}|_e$. *Define* $\sigma_{1q}^2 = \lim_{n \to \infty} n^{-1} var(Q)$. *Then, we have* $n^{-1/2} Q \to_d N(0, \sigma_{1q}^2)$ *as* $n \to \infty$ *if*

$$n^{-2} tr\Big\{ \mathbb{M}\mathbb{M}^\top \mathbb{M}\mathbb{M}^\top \Big\} \to 0, \tag{B.11}$$

Proof. We apply the martingale difference theorem to prove the asymptotic normality of Q. First, we construct a martingale difference array Q_i which satisfies $Q = \sum_{i=1}^n Q_i$. Second, we apply the martingale difference theorem to prove $n^{-1/2} Q \to_d N(0, \sigma_{1q}^2)$ as $n \to \infty$.

Define $\mathcal{M} = (M_{ij}) \in \mathbb{R}^{n \times n}$. And define $\mathcal{F}_i$ to be the σ-field generated by

$$Q_i = M_{ii}(\varepsilon_i^2 - \sigma^2) + \sum_{j=1}^{i-1} M_{ij} \varepsilon_i \varepsilon_j + \sum_{j=1}^{i-1} M_{ji} \varepsilon_i \varepsilon_j,$$

where $e_i \in \mathbb{R}^n$ is a zero vector with only the ith element being 1. Then, we have $Q = \sum_i Q_i$ and $E(Q_i | \mathcal{F}_{i-1}) = 0$, where $\mathcal{F}_i$ is defined to be the σ-field generated by $\{\varepsilon_j : 1 \leq j \leq i\}$. Thus, we only need to prove the following two results,

$$n^{-2} \sum_{i=1}^n E(Q_i^4) \to 0, \tag{B.12}$$

$$n^{-1}\sum_{i=1}^{n} E(Q_i^2|\mathcal{F}_{i-1}) \to_p \sigma_{1q}^2. \tag{B.13}$$

Next, we prove (B.12) and (B.13) separately.

STEP I: PROOF OF (B.12). Define $Q_{i_1} = M_{ii}(\varepsilon_i^2 - \sigma^2)$, $Q_{i_2} = \sum_{j=1}^{i-1} M_{ij}\varepsilon_i\varepsilon_j$, and $Q_{i3} = \sum_{j=1}^{i-1} M_{ji}\varepsilon_i\varepsilon_j$. By the Cauchy-Schwarz inequality, we only need to prove that $n^{-2}\sum_{j=1} E(Q_{id}^4) \to 0$ holds for all $d = 1, 2, 3$.

For $d = 1$, we have $EQ_{i_1}^4 = EM_{ii}^4(\varepsilon_i^2 - \sigma^2)^4$. As $\max_i E(\varepsilon_i^2 - \sigma^2)^4 \le c_\varepsilon$, it can be calculated that $n^{-2}\sum EQ_{i_1}^4 \le n^{-1}c_\varepsilon \to 0$.

For $d = 2$, we have $Q_{i2}^4 = \sum_{j_1,j_2\le i-1} M_{ij_1}M_{ij_2}\varepsilon_i^2\varepsilon_{j_1}\varepsilon_{j_2}$. Thus, we have,

$$\begin{aligned} EQ_{i2}^4 &= \sum_{j_1\ne j_2, j_1 j_2<i} M_{ij_1}^2 M_{ij_2}^2 \kappa_4 E\varepsilon_{j_1}^2 E\varepsilon_{j_2}^2 + \sum_{j<i} M_{ij}^4\kappa_4^2 \\ &= \sum_{j_1\ne j_2, j_1 j_2<i} M_{ij_1}^2 M_{ij_2}^2 \kappa_4\sigma^4 + \sum_{j<i} M_{ij}^4\kappa_4^2. \end{aligned}$$

Define $c_q = \max\{\kappa\sigma^4, (\kappa_4)^2\}$. Then, we have

$$\sum_i EQ_{i2}^4 \le c_q \sum_i \sum_{j_1 j_2\le i, j_1\ne j_2} (M_{ij_1}^2 M_{ij_2}^2) \le c_q \mathrm{tr}(\mathbb{M}\mathbb{M}^\top\mathbb{M}\mathbb{M}^\top).$$

Then, it can be concluded that $n^{-2}\sum_i E(Q_{i2}^4) \to 0$ by (B.11).

For $d = 3$, the proof is similar to that of $d = 2$ and thus omitted.

STEP II: PROOF OF (B.13). First, it can be derived $n^{-1}\sum_{i=1}^n E\{E(Q_i^2|\mathcal{F}_{i-1})\} = n^{-1}\sum_{i=1}^n E(Q_i^2) = n^{-1}E(Q^2) \to \sigma_{1q}^2$. We next verify that, as $n \to \infty$,

$$n^{-2}\mathrm{var}\Big\{\sum_{i=1}^{n} E(Q_i^2|\mathcal{F}_{i-1})\Big\} \to 0.$$

By the Cauchy-Schwarz inequality, we only need to show $n^{-2}\sum_{i=1}^n \mathrm{var}\{E(Q_{id}^2|\mathcal{F}_{i-1})\} \to 0$ for $d = 1, \cdots, 3$, as $n \to \infty$. It can be easily verified $\mathrm{var}\{E(Q_{id}^2|\mathcal{F}_{i-1})\} = 0$ for $d = 1$. Next, we prove the case for $d = 2$. The proof of the case when $d = 3$ is similar and omitted here.

We can calculate that

$$\begin{aligned} \sum_{i=1}^{n} E(Q_{i2}^2|\mathcal{F}_{i-1}) &= \sum_i \sum_{j_1,j_2<i} (M_{ij_1}M_{ij_2})\sigma^2\varepsilon_{j_1}\varepsilon_{j_2} \\ &= \sum_{i=1}^{n} \varepsilon^\top \mathbb{I}_{i-1} M_{i\cdot} M_{i\cdot}^\top \mathbb{I}_{i-1}\varepsilon, \end{aligned}$$

where $M_{i\cdot} \in \mathbb{R}^n$ is the ith row vector of $\mathcal{M}$, $\mathbb{I}_i = \sum_{j=1}^i e_j e_j^\top$, recalling that e_j is a zero vector with only the jth element being 1. Thus, we only need

to prove, $n^{-2}\mathrm{var}(\varepsilon^\top \mathcal{M}^*\varepsilon) \to 0$ where $\mathcal{M}^* = \sum_{i=1}^N \mathbb{I}_{i-1}M_{i\cdot}M_{i\cdot}^\top \mathbb{I}_{i-1}$. Owing to $|\mathbb{I}_{i-1}M_{i\cdot}M_{i\cdot}^\top \mathbb{I}_{i-1}| \preccurlyeq |M_{i\cdot}|_e|M_{i\cdot}|_e^\top$, we only need to prove $N^{-2}\mathrm{tr}(\mathcal{M}^*\mathcal{M}^{*\top}) \to 0$. It can be calculated that

$$\mathrm{tr}(\mathcal{M}^*\mathcal{M}^{*\top}) \le \sum_{i_1,i_2}(|M_{i_1,\cdot}|_e^\top |M_{i_2,\cdot}|_e)(|M_{i_2,\cdot}|_e^\top |M_{i_1,\cdot}|_e) \le \mathrm{tr}(\mathbb{M}\mathbb{M}^\top \mathbb{M}\mathbb{M}^\top).$$

By (B.11), the result can be obtained. □

Lemma 4. *Define $\ddot{Q}(\rho)$ to be the second-order derivative of $Q(\rho)$ with respect to ρ. Assume all the conditions satisfied in Theorem 1; then, we have $-n^{-1}\ddot{Q}(\rho) \to_p \sigma_2^2$.*

Proof. We have already calculated in Appendix A.2 that $\ddot{Q}(\rho) = 2\dot{F}^\top\dot{F} + 2F^\top\ddot{F}$. Then it suffices to show that,

$$2n^{-1}\dot{F}^\top\dot{F} \to_p \sigma_2^2 \tag{B.14}$$

$$2n^{-1}F^\top\ddot{F} \to_p 0 \tag{B.15}$$

It could be easily verified that $2n^{-1}E\dot{F}^\top\dot{F} = \sigma_2^2$. Next, we verify $EF^\top\ddot{F} = 0$. It could be derived that $\ddot{F} = \ddot{d}_\rho S^\top\mathcal{E} - 2\dot{d}_\rho(W^\top\mathcal{E} + S^\top WY) + 2d_\rho W^\top WY$. Thus $F^\top\ddot{F} = S_1 + S_2 + S_3 = \mathcal{E}^\top S d_\rho \ddot{d}_\rho S^\top\mathcal{E} - 2\mathcal{E}^\top S d_\rho \dot{d}_\rho(W^\top\mathcal{E} + S^\top WY) + 2\mathcal{E}^\top S d_\rho^2 W^\top WY$. Thus, $E(S_1) = \sigma^2\mathrm{tr}(Sd_\rho\ddot{d}_\rho S^\top) = \sigma^2\mathrm{tr}(\ddot{d}_\rho)$, $E(S_2) = -2\sigma^2\mathrm{tr}(d_\rho\dot{d}_\rho W^\top S) - 2\sigma^2\mathrm{tr}(d_\rho\dot{d}_\rho S^\top W)$, and $E(S_3) = 2\sigma^2\mathrm{tr}(d_\rho^2 W^\top W)$. As a result, by (A.1) and (A.2), it could be calculated that $E(S_1) + E(S_3) = -E(S_2) = 8\sigma^2\rho^2\mathrm{tr}\{d_\rho^3\mathrm{diag}^2(W^\top W)\}$. Therefore, $EF^\top\ddot{F} = 0$.

Therefore, it suffices to verify $n^{-2}\mathrm{var}(\dot{F}^\top\dot{F}) \to 0$ and $n^{-2}\mathrm{var}(F^\top\ddot{F}) \to 0$. Since the verifications are similar, we only prove $n^{-2}\mathrm{var}(\dot{F}^\top\dot{F}) \to 0$ for example. Since $\dot{F}^\top\dot{F} = \tilde{\varepsilon}^\top M_2^\top M_2\tilde{\varepsilon}$, by Lemma 2, we only need to prove $n^{-2}\mathrm{tr}(M_2^\top M_2 M_2^\top M_2) \to 0$. By (B.7), it can be derived $n^{-2}\mathrm{var}(\dot{F}^\top\dot{F}) \le n^{-2}c_{sm}\mathrm{tr}(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}})$, where c_{sm} is a finite constant. It can be further derived $n^{-2}\mathrm{tr}(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}) \le n^{-1}\lambda_{\max}^2(\widetilde{\mathcal{W}}^\top\widetilde{\mathcal{W}}) \to 0$ by Lemma 1. Thus $n^{-2}\mathrm{var}(\dot{F}^\top\dot{F}) \to 0$.

Thus (B.14) and (B.15) could be obtained. This completes the proof. □

Appendix C

C.1. Proof of Theorem 1

The proof will be accomplished in the following two steps accordingly. In the first step, we establish the $\sqrt{n}$-consistency of $\hat{\rho}$. In the second step, we show that $\hat{\rho}$ is asymptotically normal.

Step 1. One can verify that the objective function $Q(\rho)$ is a strict convex function in ρ under condition (C2). As a result, to complete this step, it suffices

to follow the technique in Fan and Li (2001) to show that: for any $\epsilon > 0$, there exists a finite constant $C > 0$ such that

$$\liminf_n P\Big\{ \inf_{|u|=C} Q(\rho + n^{-1/2}u) > Q(\rho) \Big\} \geq 1 - \epsilon. \tag{C.1}$$

To this end, we apply Taylor's expansion and obtain

$$\inf_{|u|=C} \Big\{ Q(\rho + n^{-1/2}u) - Q(\rho) \Big\} = n^{-1/2}\dot{Q}(\rho)u + (2n)^{-1}C^2\ddot{Q}(\rho) + o_p(1)$$

$$\geq (2n)^{-1}C^2\ddot{Q}(\rho) - n^{-1/2}|\dot{Q}(\rho)|C + o_p(1). \tag{C.2}$$

We next consider the two terms of $\ddot{Q}(\rho)$ and $\dot{Q}(\rho)$ separately in (C.2). First, by definition, we have that,

$$\dot{Q}(\rho) = -2\sum_{i=1}^{n} Y_{(-i)}^{\top}\dot{\gamma}_i^*(\rho)\big\{Y_i - Y_{(-i)}^{\top}\gamma_i^*(\rho)\big\}, \tag{C.3}$$

where $\dot{\gamma}_i^*(\rho)$ is defined to be the first order derivative of $\gamma_i^*(\rho)$. Moreover, according to Proposition 1, we know that $E\{Y_i|Y_{(-i)}\} = Y_{(-i)}^{\top}\gamma_i^*(\rho)$. Consequently, by (C.3), $E\big\{\dot{Q}(\rho)\big\} = E\big[E\big\{\dot{Q}(\rho)|Y_{(-i)}\big\}\big] = 0$. Next, it can be concluded $\lim_{n\to\infty} n^{-1}\mathrm{var}\{\dot{Q}(\rho)\} = \sigma_1^2$ by Lemma 2. This suggests that the coefficient for the linear term of C in (C.2), $n^{-1/2}\dot{Q}(\rho)$ is $O_p(1)$. In addition, by Lemma 4, we know that $n^{-1}\ddot{Q}(\rho) > 0$ asymptotically. As a result, the coefficient for the quadratic term of C in (C.2) is a positive constant asymptotically. Consequently, as long as C is sufficiently large, the quadratic term in (C.2) would dominate its linear term. Therefore, $Q(\rho + n^{-1/2}u) - Q(\rho) > 0$ with probability tending to 1 as $n \to \infty$. This proves the desired conclusion in (C.1). As a result, it completes the first part proof of the theorem.

Step 2. By the first step of proof, we know that $\hat{\rho}$ is $\sqrt{n}$-consistent. This enables us to apply the Taylor's expansion technique to obtain the following asymptotic approximation,

$$\sqrt{n}(\hat{\rho} - \rho) = \{n^{-1}\ddot{Q}(\rho^*)\}^{-1}\{n^{-1/2}\dot{Q}(\rho)\},$$

where ρ^* lies between ρ and $\hat{\rho}$. By the proof of Lemma 4, we have already known that $n^{-1}\ddot{Q}(\rho^*) \to_p \sigma_2^2$.

We next show that $n^{-1/2}\dot{Q}(\rho) \to_d N(0, \sigma_1^2)$. By Lemma 3, it suffices to show $n^{-2}\mathrm{tr}(|M|_e|M|_e^{\top}|M|_e|M|_e^{\top}) \to 0$. By (B.8), the desired result could be obtained. This completes the proof.

C.2. Computationally feasible estimators for σ_1^2 and σ_2^2

Consider σ_2^2 first. By the Lemma 4 in Appendix B, we can obtain $\hat{\sigma}_2^2$ by replacing the unknown parameter ρ with the LSE $\hat{\rho}$ and replacing σ^2 with $\hat{\sigma}^2 = n^{-1}(Y -$

$\hat{\rho}WY)^\top(Y-\hat{\rho}WY)$ in $n^{-1}\ddot{Q}(\rho)$. For σ_1^2, intuitively, one could also estimate it based on $\hat{\rho}$ and $\hat{\sigma}^2$. It could be verified that,

$$\begin{aligned}\hat{\sigma}_1^2 &= \frac{\hat{\sigma}^4}{n}\mathrm{tr}\Big[8(\Omega_{\hat{\rho}}d_{\hat{\rho}}\ddot{d}_{\hat{\rho}})^2+4\big\{(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\big\}^2+16\Omega_{\hat{\rho}}d_{\hat{\rho}}\ddot{d}_{\hat{\rho}}(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\Big]\\ &\quad+\frac{\hat{\sigma}^2}{n}\mathrm{tr}\Big\{4(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\Omega_{\hat{\rho}}d_{\hat{\rho}}^2(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)\hat{\Sigma}\Big\},\end{aligned}\tag{C.4}$$

where $\mathbb{W}_2=W^\top W$, $\Omega_{\hat{\rho}}$ is defined with the true parameter ρ replaced by $\hat{\rho}$ in Ω_ρ; $d_{\hat{\rho}}$, $\dot{d}_{\hat{\rho}}$, $\ddot{d}_{\hat{\rho}}$ and $\hat{\Sigma}$ are correspondingly defined. Numerical experience suggests that the estimator (C.4) works well. However, its computational cost is high for large n by computing the inverse of a high-dimensional matrix $(I-\hat{\rho}W^\top)(I-\hat{\rho}W)$. To provide a consistent estimator which is computationally feasible for large networks, we next propose an alternative estimator. Note that the last term of σ_1^2 is $4\sigma^2n^{-1}\mathrm{tr}\big\{(2\rho\mathbb{W}_2-\mathbb{W}_1)d_\rho^2\Omega_\rho d_\rho^2(2\rho\mathbb{W}_2-\mathbb{W}_1)\Sigma\big\}$. Moreover, $\Sigma=\mathrm{var}(Y)$ and $n^{-1}E\big(Y^\top M_3Y\big)=n^{-1}\mathrm{tr}(M_3\Sigma)$, where $M_3=4(2\rho\mathbb{W}_2-\mathbb{W}_1)d_\rho^2\Omega_\rho d_\rho^2(2\rho\mathbb{W}_2-\mathbb{W}_1)$. This suggests that $n^{-1}Y^\top M_3Y$ is an unbiased estimator of $n^{-1}\mathrm{tr}(M_3\Sigma)$.

By the similar proof technique of Lemma 4, it could be verified that $n^{-1}Y^\top M_3Y$ is indeed a consistent estimator of $n^{-1}\mathrm{tr}(M_3\Sigma)$. Thus a computationally feasible estimator of σ_1^2 is given by,

$$\begin{aligned}\hat{\sigma}_1^{*2} &= \frac{\hat{\sigma}^4}{n}\mathrm{tr}\Big[8(\Omega_{\hat{\rho}}d_{\hat{\rho}}\ddot{d}_{\hat{\rho}})^2+4\big\{(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\big\}^2+16\Omega_{\hat{\rho}}d_{\hat{\rho}}\ddot{d}_{\hat{\rho}}(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\Big]\\ &\quad+\frac{4\hat{\sigma}^2}{n}Y^\top(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)d_{\hat{\rho}}^2\Omega_{\hat{\rho}}d_{\hat{\rho}}^2(2\hat{\rho}\mathbb{W}_2-\mathbb{W}_1)Y.\end{aligned}\tag{C.5}$$

Accordingly, for a given confidence level $1-\alpha$, a confidence interval of ρ can be constructed as $(\hat{\rho}-Z_{\alpha/2}n^{-1/2}\hat{\sigma}_2^{-2}\hat{\sigma}_1^*,\hat{\rho}+Z_{\alpha/2}n^{-1/2}\hat{\sigma}_2^{-2}\hat{\sigma}_1^*)$, where $Z_{\alpha/2}$ represents for the upper $\alpha/2$-th quantile of a standard normal distribution.

C.3. Proof of Theorem 2

We derive the approximated form for MLE and LSE respectively in the following.

Part I. Approximation for MLE. Assume appropriate conditions for MLE satisfied. To get the approximation for the asymptotic covariance for MLE, we first get approximations for the basic matrices. Define $\gamma_M=(\rho,\sigma^2)$. Assuming the limits of $n^{-1}\lim_{n\to\infty}\big\{\mathrm{tr}(G_\rho^2)+\mathrm{tr}(G_\rho^\top G_\rho)\big\}$ and $n^{-1}\lim_{n\to\infty}\mathrm{tr}(G_\rho)$ exist. Thus by (2.1), it could be derived that, $\sqrt{n}\gamma_M\to_d N(\gamma_M,\Sigma_M^{-2})$, where $\Sigma_M=[A,B;B,C]$, and $A=n^{-1}\lim_{n\to\infty}\big\{\mathrm{tr}(G_\rho^2)+\mathrm{tr}(G_\rho^\top G_\rho)\big\}$, $B=(n\sigma^2)^{-1}\lim_{n\to\infty}\mathrm{tr}(G_\rho)$, $C=(2\sigma^4)^{-1}$. Then we have $\sigma_M^2=A^{-1}+A^{-2}B^2(C-B^2A^{-1})$. Since $G_\rho=W+o(1)$, one could verify that $B=o(1)$. Furthermore, we could obtain that $A=n^{-1}\big\{\mathrm{tr}(W^2)+\mathrm{tr}(W^\top W)\big\}+o(1)$. As a result, $\sigma_M^2=\pi_A^{-1}+o(1)$.

Part II. Approximation for LSE. It could be derived that $d_\rho=I_n+o(1)$, $S=I_n+o(1)$, $S^{-1}=I_n+o(1)$, and $\dot{d}_\rho=o(1)$. Thus we have $M_1=I_n+o(1)$,

and $M_2 = -(W^\top + W) + o(1)$. As as result,

$$\sigma_1^2 = \lim_{n\to\infty} \frac{16}{N}\Big\{\mathrm{tr}(W^2) + \mathrm{tr}(WW^\top)\Big\} + o(1), \tag{C.6}$$

$$\sigma_2^2 = \lim_{n\to\infty} \frac{4}{N}\Big\{\mathrm{tr}(W^2) + \mathrm{tr}(WW^\top)\Big\} + o(1). \tag{C.7}$$

By (C.6) and (C.7), we have $\sigma_L^2 = \pi_A^{-1} + o(1)$.

This completes the proof.

References

Anselin, L. (2013), *Spatial Econometrics: Methods and Models*, Springer Science & Business Media.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Champman & Hall/CRC. MR3362184

Barabási, A.-L. and Albert, R. (1999), "Emergence of scaling in random networks," *Science*, 286, 509–512. MR2091634

Barry, P. and Pace, K. (1999), "Monte Carlo estimates of the log determinant of large sparse matrices," *Linear Algebra Application*, 289, 41–54. MR1670972

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

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*, 50, 95–110.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009), "Power-law distributions in empirical data," *SIAM review*, 51(4), 661–703. MR2563829

Costenbader, E. and Valente, T. W. (2003), "The stability of centrality measures when networks are sampled," *Social Networks*, 25(4), 283–307.

Frank, O. (1979), "Sampling and estimation in large social networks," *Social Networks*, 1(1), 91–101. MR0506605

Fujimoto, K., Chou, C. P., and Valente, T. W. (2011), "The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter," *Social Networks*, 33(3), 231–243. MR2793441

Handcock, M. S. and Gile, K. J. (2010), "Modeling social networks from sampled data," *The Annals of Applied Statistics*. MR2758082

Hillier, G. and Martellosio, F. (2014), "Properties of the maximum likelihood estimator in spacial autoregressive models," *Working Paper*.

Holland, P. W. and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–5. MR0608176

Huang, D., Chang, X., and Wang, H. (2018), "Spatial autoregression with repeated measurements for social networks," *Communications in Statistics - Theory and Methods*, 47, 3715–3727. MR3803429

Huang, D., Yin, J., Shi, T., and Wang, H. (2016), "A statistical model for social network labeling," *Journal of Business & Economic Statistics*, 34, 368–374. MR3523781

Krivitsky, P. N.and Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009), "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, 31.

Lee, L. (2004), "Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models," *Econometrica*, 72, 1899–1925. MR2095537

Lee, L., Li, J., and Lin, X. (2010), "Specification and estimation of social interaction models with network structure," *The Econometrics Journal*, 13, 145–176. MR2722880

Lee, L. F. and Liu, X. (2010), "Efficient GMM estimation of high order spatial autoregressive models with autoregressive disturbances," *Econometric Theory*, 26, 187–230. MR2587107

Leenders, R. T. A. (2002), "Modeling social influence through network autocorrelation: constructing the weight matrix," *Social Networks*, 24, 21–47.

LeSage, J. and Pace, R. (2007), "A Matrix Exponential Spatial Specification," *Journal of Econometrics*, 140:1, 190–214. MR2395921

LeSage, J. and Pace, R. K. (2009), *Introduction to Spatial Econometrics*, New York: Chapman & Hall. MR2730939

Meyn, S. P. and Tweedie, R. L. (2012), *Markov Chains and Stochastic Stability*, Springer Science & Business Media. MR1287609

Newman, M., Barabasi, A.-L., and Watts, D. J. (2006), *The Structure and Dynamics of Networks*, Princeton University Press. MR2352222

Nowicki, K. and Snijders, T. A. B. (2001), "Estimation and prediction for stochastic block structures," *Journal of the American Statistical Association*, 96, 1077–1087. MR1947255

Robins, G. (2013), "A tutorial on methods for the modeling and analysis of social network data," *Journal of Mathematical Psychology*, 57(6), 261–274. MR3137880

Robins, G., Elliott, P., and Pattison, P. (2001), "Network models for social selection processes," *Social Networks*, 23(1), 1–30.

Robinson, P. and Rossi, F. (2014), "Improved Lagrange multiplier tests in spatial autoregressions," *Econometrics Journal*, 17, 139–164. MR3171215

Sampson, R. J., Morenoff, J. D., and Earls, F. (1999), "Beyond social capital: Spatial dynamics of collective efficacy for children," *American Sociological Review*.

Shalizi, C. R. and Rinaldo, A. (2013), "Consistency under sampling of exponential random graph models," *The Annals of Statistics*, 41(2), 508–535. MR3099112

Smirnov, O. and Anselin, L. (2001), "Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach," *Computational Statistics and Data Analysis*, 35, 301–319. MR1819042

Trefethen, L. N. and Bau, D. (1997), *Numerical Linear Algebra*, vol. Vol.50,

Siam. MR1444820

Wang, Y. J. and Wong, G. Y. (1987), "Stochastic blockmodels for directed graphs," *Journal of the American Statistical Association*, 82, 8–19. MR0883333

Yang, K. and Lee, L.-f. (2017), "Identification and QML estimation of multivariate and simultaneous equations spatial autoregressive models," *Journal of Econometrics*, 196, 196–214. MR3572822

Zhou, J., Tu, Y., Chen, Y., and Wang, H. (2017), "Estimating spatial autocorrelation with sampled network data," *Journal of Business & Economic Statistics*, 35(1), 130–138. MR3591541