Statistica Sinica **30** (2020), 653-672
doi:https://doi.org/10.5705/ss.202017.0135

A NAIVE LEAST SQUARES METHOD FOR SPATIAL AUTOREGRESSION WITH COVARIATES

Yingying Ma, Rui Pan, Tao Zou and Hansheng Wang

Beihang University, Central University of Finance and Economics, Australian National University and Peking University

Abstract: The rapid development of social networks has resulted in an increase in the use of the spatial autoregression model with covariates. However, traditional estimation methods, such as the maximum likelihood estimation, are practically infeasible if the network size n is very large. Here, we propose a novel estimation approach, that reduces the computational complexity from $O(n^3)$ to $O(n)$. This approach is developed by ignoring the endogeneity issue induced by network dependence. We show that the resulting estimator is consistent and asymptotically normal under certain conditions. Extensive simulation studies are presented to demonstrate its finite-sample performance, and a real social network data set is analyzed for illustration purposes.

Key words and phrases: Maximum likelihood estimator, naive least squares estimator, social network analysis, spatial autoregression model.

1. Introduction

A social network is a collection of nodes (i.e., actors) and the associated social relationships (Snijders (2011)). These relationships can be based on, for example, friendship, kinship, common interests, or influence. Social network data and modeling are important for a number of reasons. First, these has been a proliferation of social networks, such as Facebook (`www.facebook.com`) and Sina Weibo (`www.weibo.com`), during the past decade. As a result, social network data are becoming increasingly available, and their associated commercial value is significant. Second, social network modeling is important because social network data are typically associated with complex dependence relationships, which pose serious challenges for traditional modeling techniques. Thus, several approaches have been developed to model such data. These include, but are not limited to, the exponential random graph model (Holland and Leinhardt (1981)), stochastic block model (Wang and Wong (1987); Nowicki and Snijders (2001)), and latent space model (Hoff, Raftery and Handcock (2002); Sewell and Chen (2015)).

In this study, we explore a specific type of statistical model used in network data analyses. The model can be viewed as a natural combination of two components. The first is the traditional linear regression model (Draper and Smith (1998)), and the second is the popular spatial autoregression (SAR) model (Ord (1975)). The linear regression component handles traditional covariate information, which is typically specific to each node, and is referred to as nodal information. If each node represents a user, then the nodal information (i.e., the covariates) could be, for example, the user's gender, age, educational background, or other demographic characteristics. The SAR component handles network dependence, and relates the response of one node to the responses of its connected neighbors. Thus, this model can handle both traditional covariates and network dependence in a convenient manner (Anselin (1988); Lee (2004)). SAR models can also be applied in fields other than social networks, including real estate (Dubin, Pace and Thibodeau (1999); Osland (2010)), crime incidents (Kakamu, Polasek and Wago (2008)), and geospatial data (Chawla et al. (2001)).

The maximum likelihood estimator (MLE) can be used to estimate a SAR model with covariates (Smirnov and Anselin (2001); Lee (2004)). However, to compute the MLE, the determinant of an ultrahigh-dimensional matrix also needs to be computed. The matrix is of the same size as the network, which renders the computational cost extremely high for large-scale social networks. LeSage and Pace (2009) developed approximate algorithms to improve the computational speed. The resulting estimator performs well if the network size is not too large (e.g., no more than 500). However, its performance deteriorates considerably as the network size increases (e.g., more than 5,000). Thus, a more computational efficient estimation procedure for the SAR model with covariates is required.

To this end, we propose a novel method. Our method is motivated by empirical studies, in which the estimated network autocorrelation is practically small (Chen, Chen and Xiao (2013); Zhou et al. (2017)). Thus, the endogeneity issue induced by network dependence can probably be ignored. This assumption leads to a naive least squares estimator (NLSE). We show that the NLSE is consistent and asymptotically normal under certain conditions. Extensive simulation studies are presented to demonstrate its finite-sample performance and a real Sina Weibo data set is analyzed for illustration purposes.

The remainder of the article is organized as follows. Section 2 introduces the naive least squares method and delineates the asymptotic properties of the corresponding NLSE. The numerical studies, including the simulation studies and real-data analysis, are presented in Section 3. Section 4 concludes the paper.

All technical details are relegated to the Supplementary Material.

2. The Methodology

2.1. Model setup

Assume a network with n nodes, indexed by $1 \leq i \leq n$. Define $a_{ij} = 1$ if node i follows node j, and 0 otherwise. We do not allow any node to follow itself directly. Thus, we require $a_{ii} = 0$ for any $1 \leq i \leq n$, which leads to the $n \times n$ adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. Its row-normalized version is given by $W = (w_{ij}) \in \mathbb{R}^{n \times n}$, with $w_{ij} = a_{ij}/d_i^{\text{out}}$, where $d_i^{\text{out}} = \sum_j a_{ij}$ is the nodal out-degree of node i. Correspondingly, we denote $d_i^{\text{in}} = \sum_j a_{ji}$ as the nodal in-degree of node i (Snijders (2011)). Consider, for example, a Twitter-type social network. In this case, the nodal in-degree measures the focal node's popularity, whereas the nodal out-degree reflects how active it is.

For each node, we assume a continuous response $Y_i \in \mathbb{R}^1$ and an associated p-dimensional predictor $X_i = (X_{i1}, \ldots, X_{ip})^\top \in \mathbb{R}^p$, where p is a fixed number. Let $\mathbb{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the response vector and $\mathbb{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times p}$ be the design matrix. Then, we consider the following SAR model with covariates:

$$\mathbb{Y} = \rho W \mathbb{Y} + \mathbb{X}\beta + \mathcal{E}, \tag{2.1}$$

where $\rho \in \mathbb{R}^1$ is the so-called *network autocorrelation coefficient* and $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ is the regression coefficient vector. This model is commonly used in the empirical literature; see, for example, Dubin, Pace and Thibodeau (1999), Kakamu, Polasek and Wago (2008), and Osland (2010). Now, let $\theta = (\rho, \beta^\top)^\top \in \mathbb{R}^{p+1}$, with the true value given by $\theta_0 = (\theta_{01}, \ldots, \theta_{0p+1})^\top = (\rho_0, \beta_0^\top)^\top$ and $\beta_0 = (\beta_{01}, \ldots, \beta_{0p})^\top$. Furthermore, denote $\mathcal{E} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$ as the error vector. We can further write $Y_i = \rho \sum_{j=1}^n w_{ij} Y_j + X_i\beta + \varepsilon_i$, which intuitively implies that the response of the ith subject is linearly related to its neighbors and covariates. Thus, ρ can be intuitively interpreted as a measure of the strength of the network dependence. We assume that ε_i are independent and i.i.d. as $N(0, \sigma^2)$. As a result, $\text{cov}(\mathcal{E}) = \sigma^2 I_n$, where $I_n \in \mathbb{R}^{n \times n}$ is an identity matrix. We thus have

$$\mathbb{Y} = (I_n - \rho W)^{-1}(\mathbb{X}\beta + \mathcal{E}). \tag{2.2}$$

To render the foregoing equality valid, it is critical that $I_n - \rho W$ be invertible for a general W matrix. Because W is a row-normalized adjacency matrix, it has been proved that its largest singular value is 1 (Banerjee et al. (2004)). As a result, a necessary and sufficient condition for the invertibility of $I_n - \rho W$, for an arbitrary W matrix, is $|\rho| < 1$. Thus, for the rest of this paper, we assume

that $|\rho| < 1$.

2.2. Maximum likelihood estimation

By (2.2), we know that $\mathbb{Y}|\mathbb{X}$ is jointly normal with mean $G\mathbb{X}\beta$ and covariance $\sigma^2(GG^\top)$, where $G = (I_n - \rho W)^{-1}$. This leads to the following negative two times log-likelihood function:

$$\ell(\theta, \sigma^2) = \sigma^{-2}(\mathbb{Y} - G\mathbb{X}\beta)^\top (GG^\top)^{-1}(\mathbb{Y} - G\mathbb{X}\beta) + 2\log|G| + n\log\sigma^2,$$

where several constants are ignored. Now, fix θ and optimize $\ell(\theta, \sigma^2)$ with respect to σ^2, which leads to $\tilde{\sigma}^2 = (\mathbb{Y} - G\mathbb{X}\beta)^\top (GG^\top)^{-1}(\mathbb{Y} - G\mathbb{X}\beta)/n$. Apply this back to $\ell(\theta, \sigma^2)$, and we obtain a profiled objective function $\ell(\theta) = \ell(\theta, \tilde{\sigma}^2)$. Then, the MLE of θ can be derived as $\hat{\theta}_{\mathrm{MLE}} = \mathrm{argmax}_\theta \ell(\theta)$, the asymptotic properties of which have been rigorously investigated by Anselin (1988), Anselin and Bera (1998), Lee (2004), and many others.

Under the assumption that the ε_i are independent and identically normal, the resulting $\hat{\theta}_{\mathrm{MLE}}$ performs quite well. However, its computational cost is high, because the log-likelihood function involves $\log|I_n - \rho W|$. The associated computational complexity is $O(n^3)$; see Smirnov and Anselin (2001). A number of suggestions have been made for dealing with the determinant $|I_n - \rho W|$ in a computationally efficient manner; see, for example, Ord (1975), Barry and Pace (1999), and Smirnov and Anselin (2001). However, a very restrictive structure needs to be imposed on W. For instance, Barry and Pace (1999) assumed that all eigenvalues of weighting matrix W are real, which may not be satisfied when W is asymmetric. These computational difficulties motivate us to explore alternative estimation methods that are more computationally feasible.

2.3. Naive least squares estimation

To develop a new estimation method with better computational capability, we here propose the following novel least squares estimation method. By (2.1), we know that the response $\mathbb{Y}$ is linear in $W\mathbb{Y}$, $\mathbb{X}$, and $\mathcal{E}$. It is then natural to consider the following least squares objective function:

$$Q(\theta) = Q(\rho, \beta) = \|\mathbb{Y} - \rho W\mathbb{Y} - \mathbb{X}\beta\|^2. \tag{2.3}$$

Intuitively, equation (2.3) can be viewed as a least squares problem, with $\mathbb{Y}$ as the response and $(W\mathbb{Y}, \mathbb{X})$ as the covariates. This leads to a least squares-type estimator because $\hat{\theta} = \mathrm{argmin}_\theta Q(\theta)$. More specifically, we denote $\hat{\theta} = (\hat{\rho}, \hat{\beta}^\top)^\top$. Define $\widetilde{\mathbb{X}} = (W\mathbb{Y}, \mathbb{X}) \in \mathbb{R}^{n\times(p+1)}$. It can be easily verified that $\hat{\theta} = (n^{-1}\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}})^{-1}(n^{-1}\widetilde{\mathbb{X}}^\top\mathbb{Y})$. We refer to $\hat{\theta}$ as a naive least squares estimator (NLSE),

because it ignores the potential endogeneity issue induced by $W\mathbb{Y}$ and $\mathcal{E}$. One might expect the consistency of such an estimator to be unlikely under a general condition. However, to our surprise, we are able to prove that $\hat{\theta}$ is indeed consistent under certain conditions (see below). Although these are not the most flexible conditions, they are reasonable in many situations. The specific conditions are as follows.

(C1) (Covariate Assumption) Assume that the X_i are i.i.d. with mean 0, covariance Σ_X, and a finite fourth-order moment.

(C2) (Network Structure) Define $\widetilde{W} = W^\top W$ and $\widehat{W} = W^\top + W$. Assume that there exist two positive constants, $c_{\max}$ and $c_{\min}$, such that $\lambda_{\max}(\widetilde{W}) < c_{\max} < \infty$, $\lambda_{\max}(\widehat{W}) < c_{\max} < \infty$, and $\lambda_{\min}(\widetilde{W}) > c_{\min}$, where $\lambda_{\max}(B)$ and $\lambda_{\min}(B)$ are the largest and smallest singular values, respectively, for an arbitrary matrix B. Furthermore, we assume that there exist two positive constants, C_1 and C_2, such that $\lim_{n\to\infty} n^{-1}tr(\widetilde{W}) \to C_1$ and $\lim_{n\to\infty} n^{-1}tr(\widehat{W}^2) \to C_2$.

(C3) (Identification Conditions) Assume that $\rho \to 0$ as $n \to \infty$.

Condition (C1) ensures that the sample covariance matrix computed from X_i converges in probability toward a positive-definite limit. It can be replaced by milder conditions (e.g., mixing conditions for X_i). Condition (C2) imposes constraints on the network structure, and is typically satisfied when the network structure is sufficiently sparse. To see this, consider, for example, a cycle-type network structure with $a_{i(i+1)} = 1$ for $1 \le i < n$, $a_{n1} = 1$, and $a_{ij} = 0$ otherwise. In this case, we can verify that $\widetilde{W} = (\widetilde{w}_{ij})$, which is an identity matrix. In addition, $\widehat{W} = (\widehat{w}_{ij})$, where $\widehat{w}_{ij} = 1$ if $|i - j| = 1$, and $\widehat{w}_{ij} = 0$ otherwise. Accordingly, we have $\lambda_{\max}(\widetilde{W}) = 1$ and $\lambda_{\max}(\widehat{W}) \le 2$. Thus, condition (C2) is satisfied. Intuitively, condition (C3) implies that the network autocorrelation coefficient ρ is practically small (Chen, Chen and Xiao (2013); Zhou et al. (2017)).

Theorem 1. *Assume model* (2.1) *and conditions* (C1)-(C3)*, as* $n \to \infty$*, we have* $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, \Sigma_1^{-1})$*. Specifically,* $\Sigma_1 = (\sigma_{11}, 0; 0, \sigma_{22})$*, where* $\sigma_{11} = C_1^2(\beta^\top \Sigma_X \beta + \sigma^2)^2/\{\sigma^2(C_1\beta^\top \Sigma_X \beta + \sigma^2 C_2/2)\}$ *and* $\sigma_{22} = \sigma^{-2}\Sigma_X$.

By Theorem 1, we know that $\hat{\theta}$, the proposed NLSE of θ, is $\sqrt{n}$-consistent and asymptotically normal. Moreover, $\hat{\rho}$ and $\hat{\beta}$ are asymptotically independent of each other. Furthermore, Σ_1 can be estimated by $\hat{\Sigma}_1 = (\hat{\sigma}_{11}, 0; 0, \hat{\sigma}_{22})$. Note that $\hat{C}_1 = n^{-1}tr(\widetilde{W})$, $\hat{C}_2 = n^{-1}tr(\widehat{W}^2)$, $\hat{\Sigma}_X = n^{-1}\mathbb{X}^\top\mathbb{X}$, and $\hat{\sigma}^2 = (\mathbb{Y} - \hat{\rho}W\mathbb{Y} - \mathbb{X}\hat{\beta})^\top(\mathbb{Y} - \hat{\rho}W\mathbb{Y} - \mathbb{X}\hat{\beta})/n$. Therefore,

$$\hat{\sigma}_{11} = \hat{C}_1^2 \frac{(\hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} + \hat{\sigma}^2)^2}{\{\hat{\sigma}^2(\hat{C}_1 \hat{\beta}^\top \hat{\Sigma}_X \hat{\beta} + \hat{\sigma}^2 \hat{C}_2/2)\}} \quad \text{and} \quad \hat{\sigma}_{22} = \hat{\sigma}^{-2}\hat{\Sigma}_X. \tag{2.4}$$

3. Numerical Studies

3.1. Preliminaries

To evaluate the finite-sample performance of our proposed method, we present a number of simulation models. For each simulation model and parameter setup, the experiment is randomly replicated $M = 5{,}000$ times. Write $\hat{\theta}^{(m)} = (\hat{\theta}_1^{(m)}, \ldots, \hat{\theta}_{p+1}^{(m)})^\top = (\hat{\rho}^{(m)}, \hat{\beta}_1^{(m)}, \ldots, \hat{\beta}_p^{(m)})^\top \in \mathbb{R}^{p+1}$ as the estimate obtained in the mth simulation replication. For each parameter $1 \le j \le p+1$, the bias (BIAS) is calculated as $\text{BIAS}_j = M^{-1}\sum_{m=1}^{M}(\hat{\theta}_j^{(m)} - \theta_{0j})$ and the root mean square error (RMSE) calculated as $\text{RMSE}_j = \sqrt{M^{-1}\sum_{m=1}^{M}(\hat{\theta}_j^{(m)} - \theta_{0j})^2}$. We next compute its standard error estimate ($\widehat{\text{SE}}$) as $\widehat{\text{SE}}_j^{(m)}$, where $\widehat{\text{SE}}_j^{(m)}$ is the square root of the jth diagonal element in $\hat{\Sigma}_1$. This can be computed according to (2.4), but from the mth simulation replication. This leads to the final average $\widehat{\text{SE}}_j = M^{-1}\sum_{m=1}^{M}\widehat{\text{SE}}_j^{(m)}$. The true standard error (SE) is estimated by $\text{SE}_j = \{M^{-1}\sum_{m=1}^{M}(\hat{\theta}_j^{(m)} - \bar{\theta}_j)^2\}^{1/2}$, where $\bar{\theta}_j = M^{-1}\sum_{m=1}^{M}\hat{\theta}_j^{(m)}$. A 95% confidence interval is constructed for θ_j in the mth replication as $\text{CI}_j^{(m)} = (\hat{\theta}_j^{(m)} - z_{0.975}\widehat{\text{SE}}_j^{(m)}, \hat{\theta}_j^{(m)} + z_{0.975}\widehat{\text{SE}}_j^{(m)})$, where z_α is the αth lower quantile of a standard normal distribution. Then, the empirical coverage probability (CP) is derived as $\text{CP}_j = M^{-1}\sum_{m=1}^{M} I(\theta_j \in \text{CI}_j^{(m)})$, where $I(\cdot)$ represents the indicator function. The original values of BIAS, RMSE, SE, $\widehat{\text{SE}}$, and CP are all small. For case of comparison, we report these values in percentage form (i.e., multiplied by 100) throughout the paper. They are thus represented as BIAS (%), RMSE (%), SE (%), $\widehat{\text{SE}}$ (%), and CP (%), respectively.

For comparison purposes, we also consider the MLE, which we computing using the algorithm developed by LeSage and Pace (2009). The corresponding MATLAB toolbox can be found at `http://www.spatial-econometrics.com`. We could follow the strict Newton Raphson-type algorithm to compute the MLE, but doing so would render the computational cost of the MLE extremely high for a large network size. The CPU times of the two methods (i.e., the NLSE and MLE) are compared using a PC, Windows 7, CPU E5-1620 V2 @ 3.70GHZ. The network autocorrelation coefficient ρ is set as $1/\log n$ to satisfy condition (C3). Moreover, we compare the independence model $\mathbb{Y} = \mathbb{X}\beta + \mathcal{E}$ with the

corresponding estimator as the ordinary least squares estimator (OLSE).

To summarize the network structure, we also compute the network density (i.e., ND $= \sum_{i,j} a_{ij}/\{n(n-1)\} \times 100\%$) and total number of edges (i.e., TNE $= \sum_{i,j} a_{ij}$) for each simulation replication. The network density is the percentage of observed edges from the number of all possible edges. The total number of edges can be viewed as the number of observed connections in the network. Therefore, the network density falls between zero and one. Intuitively, if a network has a low density, the nodes should be loosely connected with one another. Similarly, the nodes in the network are closely related if the network density is high. These performance measures are averaged across the $M = 5{,}000$ simulation replications. We next introduce four different models of network generation. The corresponding simulation results are summarized in Tables 1–4.

3.2. An ER model

We start with the simplest ER model (Erdös and Rényi (1959)), where edges are assumed to be independent. Assume that $P(a_{ij} = 1) = n^{-0.5}$, leading to adjacency matrix A and weighting matrix W. For an intuitive understanding, we generate a network with 50 nodes and plot the corresponding network structure in the left panel of Figure 1. As expected, the pattern is quite random. Furthermore, the histogram of the nodal in-degree in the right panel shows its distribution to be approximately normal.

Next, we follow Tibshiranit (1996) and fix the predictor dimension to $p = 7$. Here, $\beta_0 = (3, 1.5, 0, 0, 2, 0, 0)^\top$ and ε_i is generated independently from a standard normal distribution. Accordingly, predictor $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is simulated from a multivariate normal distribution with mean 0 and covariance $\Sigma_X = (\widetilde{\sigma}_{j_1 j_2})$, where $\widetilde{\sigma}_{j_1 j_2} = 0.5^{|j_1 - j_2|}$. The error term, ε_i, is generated independently from $N(0, 1)$, and response $\mathbb{Y}$ is then generated according to (2.2). Various network sizes (i.e., $n = 500$, 2,000, and 5,000) are considered. As the network size increases from $n = 500$ to $n = 5{,}000$, the network density drops from 4.5% to 1.4% and the total number of edges increases from 11,160.1 to 353,487.4, on average. The simulation results are summarized in Table 1.

In Table 1, the following patterns can be detected for the NLSE. First, the RMSE (%), SE (%), and $\widehat{\text{SE}}$ (%) values all decrease toward zero as the network size n increases. Second, the BIAS (%) values are much smaller than the RMSE (%) values, and appear to be practically negligible. Third, the estimated $\widehat{\text{SE}}$ (%) approximates the true SE (%) quite well. As a consequence, the CP (%) values (reported coverage probabilities) are fairly close to their nominal level of 95%.

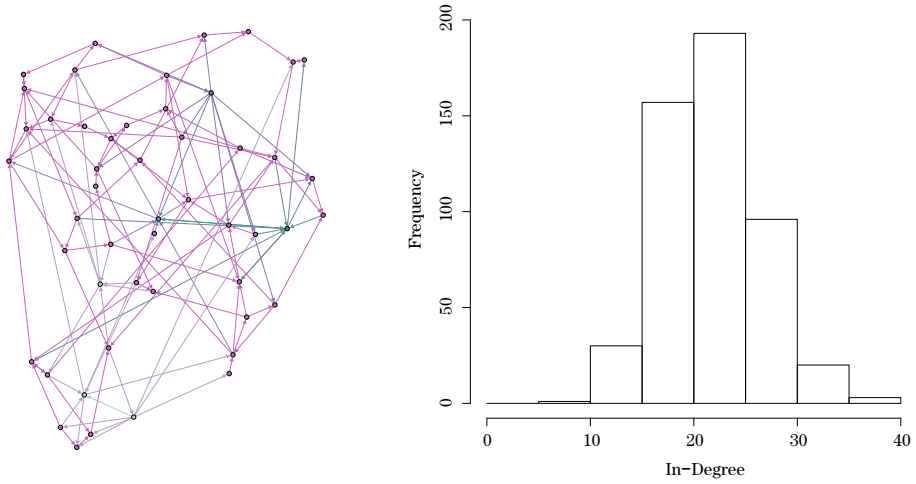

Figure 1. A randomly simulated ER network structure with 50 nodes. The network structure is plotted in the left panel. The right panel is the histogram of the nodal in-degree for the 50 nodes.

These results corroborate the asymptotic theory given in Theorem 1 quite well.

Comparatively speaking, the difference between the NLSE and MSE in terms of RMSE (%) is small, which implies that the estimation efficiency of the NLSE is almost as good as that of the MLE. Furthermore, as the network size n increases, the CP (%) values (reported coverage probabilities) of our estimator (NLSE) are stable at the nominal level of 95%. However, when the network size n increases, it becomes far more difficult for the MLE to obtain a reliable $\hat{\rho}$. For example, when $n = 5{,}000$, the reported coverage probability for the MLE is as low as 17.16%. We suspect that this problem stems primarily from the fast approximation algorithm. The strict Newton Raphson-type algorithm can be used to compute the MLE, but the required computational time would be much greater. In contrast, the proposed NLSE method is computationally efficient, as demonstrated by the reported CPU time in Table 1. Moreover, the performance of the OLSE is worse than that of the NLSE and MLE; therefore we omit it here to save space.

3.3. Dyad independence model

The ER model is a highly simplified network structure, and cannot mimic the reciprocity phenomenon in real network data. Intuitively, if node j follows i, the likelihood of node i following back should be large. This suggests that $P(a_{ij} = 1|a_{ji} = 1)$ should be well bounded above zero, even if $P(a_{ji} = 1)$ could

Table 1. Detailed simulation results for ER networks with $\rho = 1/\log n$ and predictor dimension $p = 7$. In this table, BIAS (%) = bias, RMSE (%) = root mean square error, SE(%) = estimated true standard error, $\widehat{\text{SE}}$ (%) = average standard error estimate, and CP (%) = coverage probability. Computational time (TIME) in seconds is also reported.

		BIAS(%)			RMSE(%)			SE (%)			$\widehat{\text{SE}}$(%)			CP (%)			TIME	
n	Parameter	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	MLE
500	ρ	0.12	-	0.11	4.46	-	4.45	4.45	-	4.45	4.41	-	4.34	94.58	-	94.20	0.009	0.202
	β_1	0.12	0.13	0.12	5.17	5.22	5.17	5.17	5.21	5.17	5.19	5.27	5.16	95.06	95.44	94.92		
	β_2	-0.06	-0.05	-0.06	5.92	6.00	5.92	5.92	6.00	5.92	5.81	5.90	5.78	94.90	94.94	94.86		
	β_3	0.11	0.12	0.11	5.81	5.88	5.81	5.81	5.88	5.81	5.81	5.90	5.77	95.06	94.80	94.94		
	β_4	-0.09	-0.07	-0.09	5.81	5.90	5.81	5.81	5.90	5.81	5.81	5.90	5.78	94.32	94.52	94.22		
	β_5	0.07	0.08	0.07	5.85	5.93	5.85	5.85	5.93	5.85	5.81	5.90	5.78	94.56	94.78	94.32		
	β_6	0.03	0.02	0.03	5.76	5.84	5.76	5.76	5.84	5.76	5.81	5.90	5.78	94.98	95.18	94.90		
	β_7	-0.08	-0.08	-0.08	5.27	5.33	5.27	5.27	5.33	5.27	5.19	5.27	5.16	94.86	95.04	94.82		
2,000	ρ	0.04	-	0.04	3.14	-	3.13	3.14	-	3.13	3.14	-	3.17	95.20	-	92.08	0.232	2.929
	β_1	0.01	0.01	0.01	2.59	2.60	2.59	2.59	2.60	2.59	2.59	2.60	2.58	94.82	94.88	94.78		
	β_2	0.01	0.01	0.01	2.90	2.91	2.90	2.90	2.91	2.90	2.89	2.90	2.89	94.86	94.78	94.84		
	β_3	0.03	0.03	0.03	2.90	2.92	2.90	2.90	2.92	2.90	2.89	2.91	2.89	94.86	94.94	94.84		
	β_4	-0.02	-0.02	-0.02	2.84	2.85	2.84	2.84	2.85	2.84	2.89	2.90	2.89	95.80	95.58	95.78		
	β_5	-0.06	-0.06	-0.06	2.88	2.89	2.88	2.88	2.89	2.88	2.89	2.90	2.89	95.02	95.26	94.94		
	β_6	0.02	0.02	0.02	2.85	2.86	2.85	2.85	2.86	2.85	2.89	2.91	2.89	95.30	95.32	95.28		
	β_7	0.05	0.06	0.05	2.59	2.60	2.59	2.59	2.60	2.59	2.59	2.60	2.58	94.68	94.84	94.66		
5,000	ρ	-0.03	-	-0.51	2.46	-	2.39	2.46	-	2.33	2.50	-	0.27	95.42	-	17.16	2.077	27.073
	β_1	-0.01	-0.01	-0.01	1.62	1.62	1.62	1.62	1.62	1.62	1.63	1.64	1.63	95.26	95.30	95.26		
	β_2	-0.02	-0.02	-0.02	1.79	1.80	1.79	1.79	1.80	1.79	1.83	1.83	1.82	95.40	95.48	95.42		
	β_3	0.00	0.00	0.00	1.81	1.81	1.81	1.81	1.81	1.81	1.83	1.83	1.82	95.30	95.36	95.30		
	β_4	-0.01	0.00	-0.01	1.83	1.84	1.83	1.83	1.84	1.83	1.83	1.83	1.82	95.26	95.34	95.22		
	β_5	0.01	0.01	0.01	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	94.50	94.44	94.50		
	β_6	0.00	0.00	0.00	1.79	1.79	1.79	1.79	1.79	1.79	1.83	1.83	1.83	95.32	95.44	95.32		
	β_7	0.01	0.00	0.01	1.64	1.64	1.64	1.64	1.64	1.64	1.63	1.64	1.63	94.94	95.00	94.90		

be close to zero. Therefore, we are motivated to simulate a network structure tht demonstrates the reciprocity property. To this end, we consider the p_1 model proposed by Holland and Leinhardt (1981).

In this model, edges form dyads. For two arbitrary nodes $i < j$, define a dyad as $D_{ij} = (a_{ij}, a_{ji})$. Different from the ER model, the p_1 model assumes that the different dyads are independent. However, within a given dyad D_{ij}, the two edges (i.e., a_{ij} and a_{ji}) can be dependent on each other in an arbitrary manner. In this simulation example, we generate the dyads according to $P(D_{ij} = (1,0)) = P(D_{ij} = (0,1)) = 0.5n^{-0.6}$, $P(D_{ij} = (1,1)) = n^{-0.6}$, and $P(D_{ij} = (0,0)) = 1 - 2n^{-0.6}$. It can be verified that $P(a_{ij} = 1|a_{ji} = 1) = P(D_{ij} = (1,1))/P(a_{ji} = 1) = 2/3$, which is a constant that does not converge toward 0 as $n \to \infty$, therefore protecting the reciprocity property. For an intuitive understanding, we generate a network of 50 nodes and plot the corresponding network structure in the left panel of Figure 2. A large number of reciprocated pairs (i.e., $a_{ij} = a_{ji} = 1$) are evident in the figure. A histogram of the nodal in-degree, which is approximately normal, is plotted in the right panel of Figure 2.

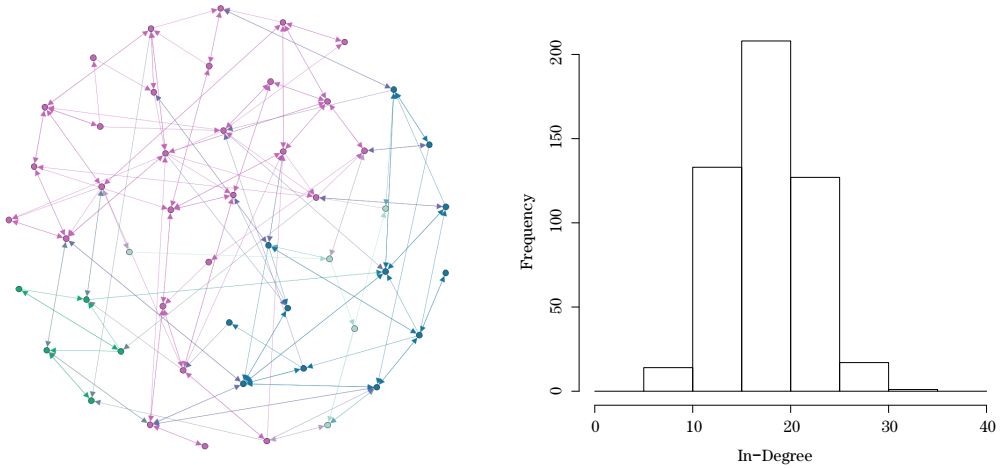

Figure 2. A randomly simulated dyad independence model network structure with 50 nodes. The network structure is plotted in the left panel. The right panel is a histogram of the nodal in-degree.

Next, we follow an example from Fan and Lv (2008) and fix the predictor dimension to $p = 3$. Here, $\beta_{0j} = 5$, for all $1 \leq j \leq p$, and ε_i is generated independently from $N(0, 1)$. The predictor $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is generated from a multivariate normal distribution with mean zero and covariance $\Sigma_X = (\widetilde{\sigma}_{j_1 j_2})$, where $\widetilde{\sigma}_{j_1 j_2} = 0.5$ for any $j_1 \neq j_2$ and $\widetilde{\sigma}_{jj} = 1$. The error term ε_i is generated independently from $N(0, 1)$ and the response $\mathbb{Y}$ is then generated according to (2.2). Various network sizes are considered (i.e., $n = 500$, 750, and 2,000). As the network size increases from 500 to 2,000, the network density drops from 3.6% to 1.6%, whereas the total number of edges increases from 8,991.5 to 62,711.5, on average. The simulation results are summarized in Table 2, and are qualitatively similar to the results in Table 1. In particular, Table 2 shows that when the network size $n > 500$, it becomes difficult for the MLE to obtain a reliable $\hat{\rho}$. Specifically, when $n = 750$, the reported coverage probability for the MLE is 62.04%, which is far from the nominal level, whereas that for the NLSE (i.e., 94.90%) remains close to the nominal level of 95%.

3.4. Stochastic block model

In addition to the reciprocity property, many real network structures exhibit a strong clustering property. As a result, the entire network can be classified into different groups. Nodes within the same group are much more likely to

Table 2. Detailed simulation results for dyad independence networks with $\rho = 1/\log n$ and $p = 3$. In this table, BIAS (%) = bias, RMSE (%) = root mean square error, SE(%) = estimated true standard error, $\widehat{\text{SE}}$ (%) = average standard error estimate, and CP (%) = coverage probability. Computational time (TIME) in seconds is also reported.

		BIAS(%)			RMSE(%)			SE (%)			$\widehat{\text{SE}}$(%)			CP (%)			TIME	
n	Parameter	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	MLE
500	ρ	0.06	-	0.03	1.48	-	1.47	1.48	-	1.47	1.51	-	1.49	95.14	-	94.74	0.011	0.269
	β_1	-0.05	0.43	-0.05	5.45	6.11	5.45	5.45	6.09	5.45	5.49	6.11	5.48	95.28	95.30	95.24		
	β_2	-0.05	0.39	-0.05	5.49	6.13	5.49	5.49	6.12	5.49	5.49	6.11	5.48	94.86	94.66	94.86		
	β_3	0.08	0.57	0.08	5.49	6.20	5.49	5.49	6.17	5.49	5.48	6.10	5.47	94.78	94.70	94.66		
750	ρ	0.05	-	0.02	1.34	-	1.33	1.33	-	1.33	1.34	-	0.59	94.90	-	62.04	0.022	0.389
	β_1	0.01	0.36	0.01	4.47	4.85	4.47	4.47	4.84	4.47	4.48	4.86	4.47	94.66	95.36	94.62		
	β_2	-0.03	0.35	-0.03	4.44	4.84	4.44	4.44	4.83	4.44	4.48	4.86	4.47	95.06	94.96	95.00		
	β_3	-0.04	0.36	-0.04	4.48	4.91	4.48	4.48	4.90	4.48	4.48	4.86	4.47	95.08	94.72	95.00		
2,000	ρ	0.04	-	0.02	1.00	-	0.99	1.00	-	0.99	1.00	-	0.40	94.88	-	57.72	0.216	2.457
	β_1	0.07	0.25	0.07	2.74	2.85	2.74	2.74	2.84	2.74	2.74	2.86	2.74	94.78	95.08	94.76		
	β_2	-0.01	0.16	-0.01	2.73	2.88	2.73	2.73	2.88	2.73	2.74	2.86	2.74	95.08	95.16	95.08		
	β_3	-0.02	0.16	-0.02	2.67	2.79	2.67	2.67	2.79	2.67	2.74	2.86	2.74	95.34	95.56	95.34		

be connected than are from different groups. In the literature, such groups are also referred to as communities (Girvan and Newman (2002)). To mimic this type of network structure, we follow Nowicki and Snijders (2001) and simulate a stochastic block structure, as follows. Assume a total of $K = 5$ blocks. We randomly assign node i to one of the five blocks, with equal probability. We then independently generate a_{ij} according to: (1) $P(a_{ij} = 1) = n^{-0.4}$ if i and j are from the same block; and (2) $P(a_{ij} = 1) = n^{-0.8}$ otherwise. Lastly, set $a_{ii} = 0$ for every $1 \le i \le n$. This leads to the network adjacency matrix A. We then generate a network with 300 nodes, the network structure of which is plotted in the left panel of Figure 3. A structure of five communities is shown. The histogram of the nodal in-degree plotted in the right panel shows an approximate normal pattern. In fact, the stochastic model considered here cannot capture a highly skewed in-degree distribution, primarily because the number of blocks is limited (only five). Within each block, the edges are generated in a random manner (i.i.d.), which renders the resulting degree distribution fairly normal, rather than skewed.

We now simulate the regression model following Zou and Hastie (2005). Specifically, fix the predictor dimension to $p = 7$ and set $\beta_{0j} = 0.85$, for all $1 \le j \le p$. The predictor $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is generated from a multivariate normal distribution with mean zero and covariance $\Sigma_X = (\widetilde{\sigma}_{j_1 j_2})$, where $\widetilde{\sigma}_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ for any j_1 and j_2. The error term ε_i is generated independently from $N(0, 1)$ and the response $\mathbb{Y}$ is then generated according to (2.2). Various network sizes are considered (i.e., $n = 500$, 1,000, and 5,000). As the network

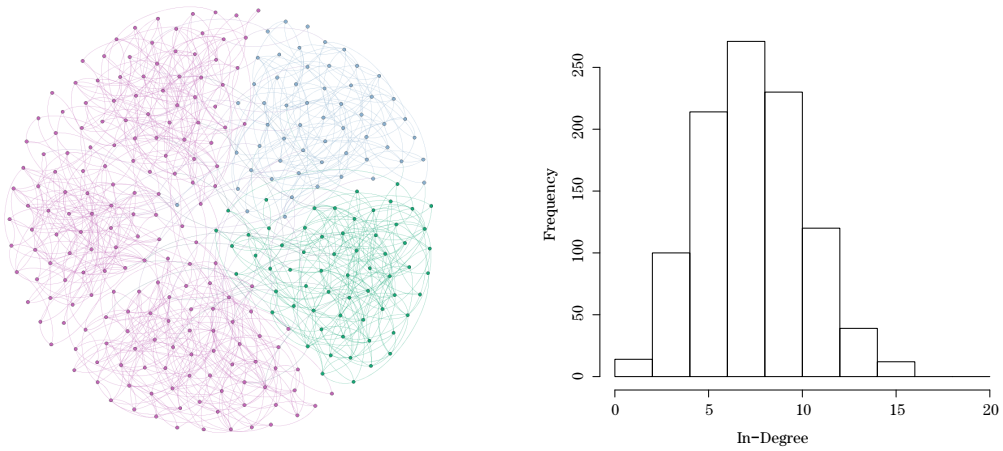

Figure 3. A randomly simulated stochastic block model network structure with 300 nodes and $K = 5$. The left panel presents the network structure and the right panel is a histogram of the nodal in-degree.

size increases from 500 to 5,000, the network density declines from 6.7% to 2.7%, whereas the total number of edges rises from 16,717.7 to 664,400.6, on average. The simulation results are summarized in Table 3, and are qualitatively similar to those in Table 1. We find the proposed NLSE to be consistent and asymptotically normal. The performance measures presented in Table 3 all corroborate our theory quite well.

3.5. Power-law distribution

The three simulation examples above suffer from one common limitation; that is, their in-degree distributions are approximately normal. As a result, we are unlikely to observe nodes with very large in-degree values. However, in practice, most Twitter-type social media platforms (e.g., Twitter and Weibo) contain nodes with extremely large in-degree values, typically representing celebrities or entities of importance. Accordingly, the in-degree distribution can be highly skewed, and should thus be of the power-law-type (Adamic and Huberman (2000)). Hence, we are motivated to simulate a power-law-type network structure.

Specifically, for each node $1 \leq i \leq n$, we generate n i.i.d. random variables according to a power-law distribution with a probability density function given by $ck^{-\alpha}$, where c is a normalizing constant, α is the exponent parameter, and $\alpha = 2.5$ is fixed. We denote these variables by E_i. Then, for the ith node, we randomly

Table 3. Detailed simulation results for stochastic block networks with $\rho = 1/\log n$ and $p = 7$. In this table, BIAS (%) = bias, RMSE (%) = root mean square error, SE(%) = estimated true standard error, $\widehat{\text{SE}}$ (%) = average standard error estimate, and CP (%) = coverage probability. Computational time (TIME) in seconds is also reported.

		BIAS(%)			RMSE(%)			SE (%)			$\widehat{\text{SE}}$(%)			CP (%)			TIME	
n	Parameter	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	MLE
500	ρ	0.13	-	0.04	7.02	-	6.96	7.02	-	6.97	7.05	-	6.90	94.96	-	94.82	0.009	0.235
	β_1	0.13	0.12	0.13	5.16	5.19	5.16	5.16	5.19	5.16	5.19	5.23	5.16	95.14	95.22	94.92		
	β_2	-0.07	-0.06	-0.07	5.92	5.94	5.92	5.92	5.94	5.92	5.81	5.85	5.78	95.00	95.02	94.90		
	β_3	0.11	0.13	0.11	5.81	5.83	5.81	5.81	5.83	5.80	5.81	5.85	5.78	94.92	95.38	94.80		
	β_4	-0.09	-0.09	-0.09	5.82	5.86	5.82	5.82	5.86	5.82	5.81	5.85	5.78	94.46	94.52	94.32		
	β_5	0.08	0.07	0.08	5.85	5.88	5.85	5.85	5.88	5.85	5.81	5.85	5.78	94.52	94.64	94.30		
	β_6	0.03	0.05	0.03	5.76	5.80	5.76	5.76	5.80	5.76	5.81	5.85	5.78	94.98	95.16	94.74		
	β_7	-0.08	-0.07	-0.08	5.27	5.29	5.27	5.27	5.29	5.27	5.19	5.23	5.16	94.86	94.98	94.68		
1,000	ρ	0.04	-	-0.03	6.09	-	6.06	6.09	-	6.06	6.15	-	2.78	95.18	-	63.48	0.041	0.703
	β_1	0.04	0.04	0.04	3.71	3.72	3.71	3.71	3.72	3.71	3.66	3.67	3.65	95.00	94.90	94.94		
	β_2	-0.14	-0.14	-0.14	4.07	4.07	4.07	4.07	4.07	4.07	4.09	4.11	4.08	94.72	94.78	94.66		
	β_3	0.06	0.06	0.06	4.12	4.13	4.12	4.12	4.13	4.12	4.09	4.11	4.08	95.14	95.14	94.96		
	β_4	-0.07	-0.06	-0.07	4.04	4.05	4.04	4.04	4.05	4.04	4.09	4.11	4.08	95.62	95.48	95.56		
	β_5	0.00	-0.01	0.00	4.05	4.06	4.05	4.05	4.06	4.05	4.09	4.11	4.08	95.26	95.10	95.16		
	β_6	-0.05	-0.04	-0.05	4.16	4.16	4.16	4.16	4.16	4.16	4.09	4.11	4.08	94.82	94.72	94.68		
	β_7	0.06	0.06	0.06	3.67	3.69	3.67	3.67	3.69	3.67	3.66	3.67	3.65	95.18	95.12	95.08		
5,000	ρ	-0.01	-	-0.78	4.52	-	4.29	4.52	-	4.22	4.46	-	0.40	94.68	-	13.84	2.546	29.085
	β_1	-0.01	-0.01	-0.01	1.62	1.62	1.62	1.62	1.62	1.62	1.63	1.64	1.63	95.28	95.44	95.28		
	β_2	-0.02	-0.02	-0.02	1.79	1.79	1.79	1.79	1.79	1.79	1.83	1.83	1.82	95.36	95.36	95.36		
	β_3	0.00	-0.01	0.00	1.81	1.81	1.81	1.81	1.81	1.81	1.83	1.83	1.82	95.36	95.20	95.34		
	β_4	0.00	0.00	0.00	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.82	95.22	95.16	95.20		
	β_5	0.01	0.01	0.01	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	1.83	94.54	94.46	94.48		
	β_6	0.00	0.00	0.00	1.79	1.79	1.79	1.79	1.79	1.79	1.83	1.83	1.83	95.38	95.48	95.32		
	β_7	0.01	0.01	0.01	1.64	1.64	1.64	1.64	1.64	1.64	1.63	1.64	1.63	94.96	95.16	94.98		

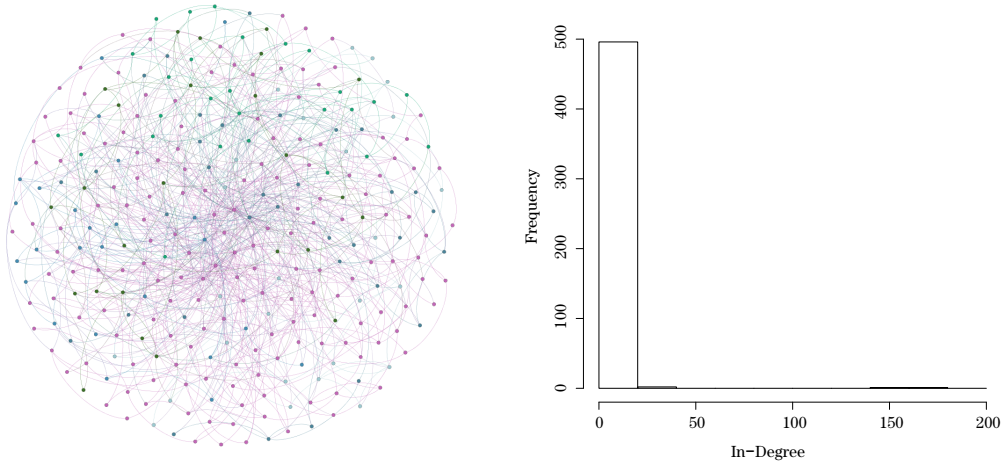

Figure 4. A randomly simulated power-law-type network structure with 300 nodes. The left panel is the network structure. The right panel is the histogram of the nodal in-degree.

select a sample of size $[E_i]$ from $\mathcal{S}_{\mathcal{F}} = \{1, 2, \ldots, n\}$, without replacement, where $[E_i]$ is the largest integer no greater than $\min\{E_i, n\}$. Denote the sample by $\mathcal{S}_i$. We next define $a_{ij} = 1$ if $j \in \mathcal{S}_i$, and $a_{ij} = 0$ otherwise. We force $a_{ii} = 0$ for every $1 \leq i \leq n$, which leads to our final adjacency matrix A. A randomly simulated network with 300 nodes is generated and the network structure is plotted in the left panel of Figure 4. Nodes with a large number of in-degree values are shown. The nodal in-degree histogram is plotted in the right panel. It is highly skewed and closely matches a power-law-type distribution.

Now, we borrow an example from Fan and Lv (2008) and fix the predictor dimension to $p = 4$. Here, $\beta_{0j} = 5$, for $1 \leq j \leq p-1$, and $\beta_{0p} = -15\sqrt{0.5}$. The error term ε_i is generated independently from $N(0, 1)$. The predictor $X_i = (X_{i1}, \ldots, X_{ip})^\top$ is generated from a multivariate normal distribution with mean 0 and covariance $\Sigma_X = (\widetilde{\sigma}_{j_1 j_2})$, where $\widetilde{\sigma}_{j_1 j_2} = 0.5$ for $1 \leq j_1 \neq j_2 \leq p-1$. Let $\widetilde{\sigma}_{j_1 j_2} = \sqrt{0.5}$ for either $j_1 = p$ or $j_2 = p$, and $\widetilde{\sigma}_{jj} = 1$ for $1 \leq j \leq p$. Thus, X_{ip} has a correlation of $\sqrt{0.5}$ with the other predictors, but is uncorrelated with the response Y_i. The response $\mathbb{Y}$ is then generated according to (2.2). Various network sizes are considered (i.e., $n = 500$, 1,000, and 2,000).

As the network size increases from 500 to 2,000, the network density decreases form 0.6% to 0.1%, and the total number of edges increases from 1,455.1 to 5,925.4, on average. We thus have a sparse network structure. The detailed simulation results are summarized in Table 4. Note that little difference can be observed between the MLE and the NLSE, except for the computational time. The NLSE is still much more computationally efficient than the MLE. In other words, the amount of CPU time consumed by the latter is substantially greater than that consumed by the NLSE; see the last two columns of Table 4.

3.6. Robustness check

It is clear that the technical condition (C3) requires that $\rho \to 0$ as the network size $n \to \infty$. We thus perform a robustness check for this condition. More specifically, we examine the finite-sample performance of the proposed estimator (NLSE) when condition (C3) is violated (e.g., a fixed ρ). To that end, we conduct a number of robustness studies. We replicate all of the simulation examples in the previous subsections, except here we fix $\rho = 0.5$ instead of allowing $\rho = 1/\log(n) \to 0$. The results are given in Tables S1–S4 in the Supplementary Material. The empirical results are qualitatively similar to those of the previous simulation studies, and the performance of the proposed NLSE remains relatively good.

Table 4. Detailed simulation results for power-law-type networks with $\rho = 1/\log n$ and $p = 4$. In this table, BIAS (%) = bias, RMSE (%) = root mean square error, SE(%) = estimated true standard error, $\widehat{\text{SE}}$ (%) = average standard error estimate, and CP (%) = coverage probability. Computational time (TIME) in seconds is also reported.

		BIAS(%)			RMSE(%)			SE (%)			$\widehat{\text{SE}}$(%)			CP (%)			TIME	
n	Parameter	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	OLSE	MLE	NLSE	MLE
500	ρ	0.01	-	0.01	0.91	-	0.91	0.91	-	0.91	0.90	-	0.89	94.90	-	94.44	0.009	0.206
	β_1	0.04	0.12	0.04	6.39	8.05	6.39	6.39	8.05	6.39	6.33	8.16	6.32	94.94	95.28	94.82		
	β_2	0.05	0.21	0.05	6.28	8.12	6.28	6.28	8.12	6.28	6.34	8.16	6.32	95.16	95.16	95.08		
	β_3	-0.01	0.11	-0.01	6.30	8.17	6.30	6.30	8.17	6.30	6.34	8.17	6.32	94.72	94.90	94.68		
	β_4	-0.14	-0.38	-0.14	8.96	11.44	8.96	8.96	11.43	8.96	8.96	11.54	8.94	94.54	95.06	94.50		
1,000	ρ	0.01	-	0.01	0.64	-	0.64	0.64	-	0.64	0.64	-	0.63	94.96	-	94.72	0.043	0.617
	β_1	0.08	0.08	0.08	4.50	5.49	4.50	4.50	5.49	4.50	4.47	5.54	4.47	94.92	94.86	94.90		
	β_2	0.00	0.05	0.00	4.54	5.62	4.54	4.54	5.62	4.54	4.48	5.54	4.47	94.76	94.44	94.74		
	β_3	0.02	0.00	0.02	4.48	5.53	4.48	4.48	5.53	4.48	4.48	5.54	4.47	95.06	95.16	94.96		
	β_4	-0.03	-0.01	-0.03	6.41	7.83	6.41	6.41	7.83	6.41	6.33	7.83	6.32	94.88	94.80	94.80		
2,000	ρ	0.01	-	0.01	0.44	-	0.44	0.44	-	0.44	0.45	-	0.45	95.38	-	95.20	0.232	3.100
	β_1	-0.01	-0.03	-0.01	3.17	3.78	3.17	3.17	3.78	3.17	3.17	3.79	3.16	95.30	95.32	95.28		
	β_2	-0.01	-0.01	-0.01	3.14	3.76	3.14	3.14	3.76	3.14	3.17	3.80	3.17	95.22	95.30	95.22		
	β_3	-0.06	-0.05	-0.06	3.17	3.80	3.17	3.17	3.80	3.17	3.17	3.80	3.17	95.04	94.68	94.98		
	β_4	0.02	0.05	0.02	4.43	5.37	4.43	4.43	5.37	4.43	4.48	5.37	4.47	95.28	94.92	95.22		

3.7. A real example

We now illustrate our proposed method using a real example with data collected from Sina Weibo (`www.weibo.com`), the most popular Twitter-type social media platform in China. Our data set contains $n = 6{,}916$ nodes, all of which are followers of an official Weibo account of a leading business school in China. The nodes are observed for a total of $T = 116$ consecutive days, allowing us to summarize the total number of tweets posted for each node during that period. Response Y_i is defined as the $\log(1 + x)$-transformed number of tweets. A histogram of the response is presented in the left panel of Figure 5, showing it to be approximately normal. Because the observation period is relatively short, the follower-followee relationship remains relatively stable. Thus, the social relationships on any observation day can be used to generate network adjacency matrix A. The analytical results are nearly identical; thus we use the network structure for the last observation day, for simplicity, leading to a total of $\sum_{i,j} a_{ij} = 547{,}856$ edges and $\sum_{i<j} a_{ij}a_{ji} = 99{,}678$ mutually connected pairs. The network density is roughly 1.0%. The nodal in-degree histogram plotted in the right panel of Figure 5 is highly skewed and exhibits a power-law-type distribution.

We next construct a number of predictors, as follows. (1) Define X_{i1} as a given node's number of followers. (2) X_{i2} is defined as tenure, measured in days, starting from the day the user created his/her Sina Weibo account. (3) Define $X_{i3} = 1$ if the ith user is an officially verified Sina Weibo member, and $X_{i3} = 0$

Table 5. Detailed estimation results for the Sina Weibo data set.

Regression Coefficient	NLSE	OLSE	MLE	SE_{NLSE}	SE_{OLSE}	SE_{MLE}
Network Autocorrelation (ρ)	0.269	-	0.224	0.031	-	0.021
Number of Followers (β_1)	0.188	0.159	0.183	0.017	0.016	0.016
Tenure (β_2)	−0.135	−0.106	−0.130	0.016	0.016	0.016
Verified Account (β_3)	0.110	0.125	0.112	0.016	0.016	0.016
Active Level(β_4)	0.069	0.070	0.069	0.014	0.015	0.014
Official Account (β_5)	0.166	0.175	0.167	0.014	0.014	0.014
Average Absolute Prediction Error	0.757	0.768	0.768	-	-	-

otherwise. (4) X_{i4} is a positive integer reflecting the activity level of the ith user, and is provided by Sina Weibo. (5) Finally, let $X_{i5} = 1$ if the ith node is an official account owned by a company or organization (as opposed to a natural person). These predictors help us to determine the kinds of nodes that are most likely to post in Sina Weibo.

We now compare the performance of NLSE with that of the MLE (LeSage and Pace (2009)) and OLSE (the model assuming independence). For evaluation purposes, the data set is split randomly into two subsets of equal size. One subset serves as the training sample, and the other is used to test the average absolute prediction errors. The experiment is randomly replicated 500 times, and the results are summarized in Table 5. First, all estimates are statistically significant at the 1% level, and the variance inflation factor (VIF) values are far below 10. Thus, it is acceptable to ignore collinearity (Belsley et al. (1980)). Second, the NLSE has the smallest average absolute prediction error of the three estimators. In addition, the $\hat{\rho}$ values for the NLSE and MLE are quite close and are all positive, suggesting that the posting activities of connected nodes are positively correlated. Third, the signs of the estimated coefficients for the three methods are consistent. More specifically, $\hat{\beta}_1$ is positive, which implies that users with a large number of followers are more likely to post in Sina Weibo, whereas $\hat{\beta}_2$ is negative, indicating that users lose interest in Sina Weibo over time. Further, $\hat{\beta}_3$ is positive, indicating that verified users are more likely to post, and $\hat{\beta}_4$ and $\hat{\beta}_5$ are both positive, suggesting that active users and official accounts are more likely to post on the social network.

4. Concluding Remarks

In this study, we employ the classical SAR model to capture the network dependency. However, the traditional maximum likelihood estimation method

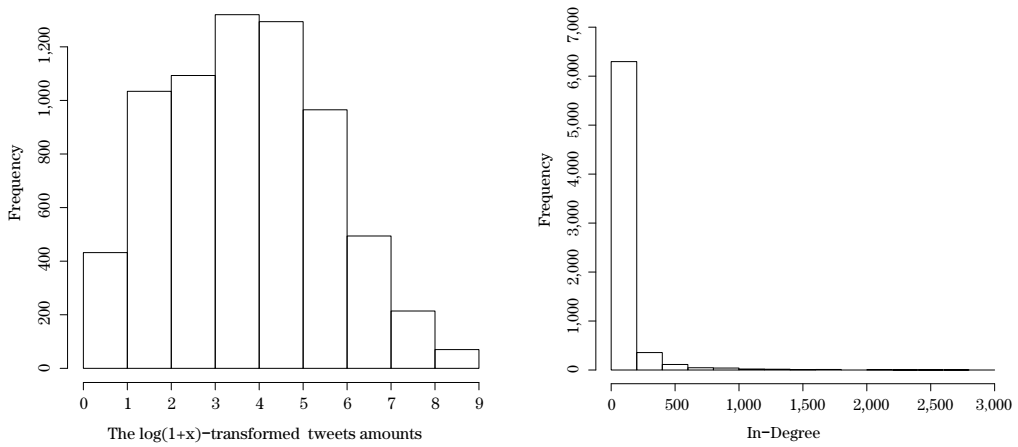

Figure 5. Real-data analysis with network size $n = 6{,}916$. The left panel is a histogram of responses and the right panel is a histogram of the nodal in-degree.

cannot be applied directly because the computational complexity is $O(n^3)$. To address this problem, we propose a novel estimation method, called the naive least squares estimation (NLSE), for the SAR model with covariates. Our simulation results suggest that the NLSE performs well across a large range of ρ-values ($|\rho| \leq 0.5$), which seems to cover most practical cases. Moreover, the real example results indicate that the NLSE exhibits superior out-of-sample performance compared with the OLSE and MLE.

We conclude the article by proposing three directions for future research. First, a richer and more general SAR model, combined with spatial correlated disturbances can be considered (i.e., $\mathbb{Y} = \rho_1 W_1 \mathbb{Y} + \mathbb{X}\beta + \mathcal{E}_1, \mathcal{E}_1 = \rho_2 W_2 \mathcal{E}_1 + \mathcal{E}_2$). Here, W_1, W_2 represent two network structures, and $\mathcal{E}_1$, $\mathcal{E}_2$ are the disturbances. The computational complexity of the MLE is $O(n^3)$. Thus, determining how to compute this model may require further effort. Second, our extensive empirical studies revealed that indirectly connected nodes may also expect to affect one another. Thus, an SAR model with higher-order neighbors would be a natural extension of the SAR model with covariates considered here. Third, a community structure may exist within the network, and different communities may correspond to different network autocorrelation coefficients. Hence, finding a way to extend the SAR model to a group setting that allows for a group-specific network autocorrelation coefficient is another interesting topic deserving of further investigation. Finally, note that our limited experience suggests that $\hat{\theta}$ may be consistent under a fairly general condition. A small ρ should be a sufficient

condition, but it may not be necessary.

Supplementary Material

The Supplementary Material available at Statistica Sinica online contains the simulation results for the robustness check, as well as the theoretical proofs and lemmas.

Acknowledgments

The authors sincerely thank the Co-Editor, Associate Editor, and two referees for their constructive suggestions and comments. This research was supported by the National Natural Science Foundation of China (NSFC, 11801022, 11525101, 71532001, 11831008, 11971504, 11601539, 11631003, 71771224, 71420107025), Humanities and Social Science Fund of Ministry of Education of China 17YJC910006, China's National Key Research Special Program Grant 2016YFC0207704, the Fundamental Research Funds for the Central Universities (QL18010), the Program for Innovation Research in Central University of Finance and Economics, and ANU College of Business and Economics Early Career Researcher Grant.

References

Adamic, L. A. and Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science* **287**, 2115–2115.

Anselin, L. (1988). Spatial econometrics: Methods and models. *Journal of the American Statistical Association* **85**, 905–907.

Anselin, L., and A. Bera (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In *Handbook of Applied Economic Statistics* (Edited by A. Ullah and D.E.A. Giles). Marcel Dekker: New York.

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

Barry, R. P. and Pace, R. K. (1999). Monte carlo estimates of the log determinant of large sparse matrices. *Linear Algebra and its Applications* **289**, 41–54.

Belsley, D. A., E. Kuh, and R. E. Welsch (1980). *Regression Diagnostics.* John Wiley & Sons, NJ: Hoboken.

Chawla, S., Shekhar, S., Wu, W. and Ozesmi, U. (2001). Modeling spatial dependencies for mining geospatial data. In: *Proceedings of the 2001 SIAM International Conference on Data Mining. SIAM*, pp. 1–17.

Chen, X., Chen, Y. and Xiao, P. (2013). The impact of sampling and network topology on the estimation of social intercorrelations. *Journal of Marketing Research* **50**, 95–110.

Coburn, T. C. (2007). Hierarchical modeling and analysis for spatial data. *Mathematical Geo-*

sciences **39**, 261–262.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. **326**. John Wiley & Sons.

Dubin, R. A., Pace, R. K. and Thibodeau, T. G. (1999). Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature* **7**, 79.

Erdös, P. and Rényi, A. (1959). On random graphs. I.. *Publicationes Mathematicae (Debrecen)* 6, 290–297.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7821–7826.

Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97**, 1090–1098.

Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association* **76**, 33–50.

Kakamu, K., Polasek, W. and Wago, H. (2008). Spatial interaction of crime incidents in japan. *Mathematics and Computers in Simulation* **78**, 276–282.

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

LeSage, J. and Pace, R. K. (2009). *Introduction to Spatial Econometrics*. Chapman and Hall/CRC.

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

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* **70**, 120–126.

Osland, L. (2010). An application of spatial econometrics in relation to hedonic house price modelling. *Journal of Real Estate Research* **32**, 289–320.

Sewell, D. K. and Chen, Y. (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association* **110**, 1646–1657.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: a characteristic polynomial approach. *Computational Statistics & Data Analysis* **35**, 301–319.

Snijders, T. A. B. (2011). Statistical models for social networks. *Annual Review of Sociology* **37**, 131–153.

Tibshiranit, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288.

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

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

School of Economics and Management, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, China, 100191.

E-mail: mayingying_11@163.com

School of Statistics and Mathematics, Central University of Finance and Economics, 39 South College Road, Haidian District, Beijing, China, 100081.

E-mail: panrui_cufe@126.com

Research School of Finance, Actuarial Studies and Statistics, Level 4, CBE Building, 26C Kingsley St, Acton, ACT2601, Australia.

E-mail: tao.zou@anu.edu.au

Guanghua School of Management, Peking University, No. 5 Yiheyuan Road, Haidian District, Beijing, China, 100871.

E-mail: hansheng@pku.edu.cn

(Received April 2016; accepted May 2018)