

## NETWORK VECTOR AUTOREGRESSION

BY XUENING ZHU<sup>\*,1</sup>, RUI PAN<sup>†,2</sup>, GUODONG LI<sup>‡,3</sup>,  
YUEWEN LIU<sup>§,4</sup> AND HANSHENG WANG<sup>\*,1</sup>

*Peking University*,<sup>\*</sup> *Central University of Finance and Economics*,<sup>†</sup>  
*University of Hong Kong*<sup>‡</sup> and *Xi'an Jiaotong University*<sup>§</sup>

We consider here a large-scale social network with a continuous response observed for each node at equally spaced time points. The responses from different nodes constitute an ultra-high dimensional vector, whose time series dynamic is to be investigated. In addition, the network structure is also taken into consideration, for which we propose a network vector autoregressive (NAR) model. The NAR model assumes each node's response at a given time point as a linear combination of (a) its previous value, (b) the average of its connected neighbors, (c) a set of node-specific covariates and (d) an independent noise. The corresponding coefficients are referred to as the momentum effect, the network effect and the nodal effect, respectively. Conditions for strict stationarity of the NAR models are obtained. In order to estimate the NAR model, an ordinary least squares type estimator is developed, and its asymptotic properties are investigated. We further illustrate the usefulness of the NAR model through a number of interesting potential applications. Simulation studies and an empirical example are presented.

**1. Introduction.** Consider a large-scale social network (e.g., Facebook or Twitter) with  $N$  nodes (i.e., users) indexed by  $1 \leq i \leq N$ . Throughout the rest of this article, we refer to  $N$  as the network size. To describe the network structure, define an adjacency matrix  $A = (a_{i_1 i_2}) \in \mathbb{R}^{N \times N}$ , where  $a_{i_1 i_2} = 1$  if there exists a social relationship (i.e., a directed edge) from  $i_1$  to  $i_2$  (e.g., user  $i_1$  follows  $i_2$  on Twitter), and  $a_{i_1 i_2} = 0$  otherwise [32]. Theoretically, we assume that  $A$  is nonrandom throughout the rest of this article. It can be both directed (i.e.,  $A^\top \neq A$ ) or undirected (i.e.,  $A^\top = A$ ). We follow the convention and do not allow

---

Received August 2015; revised April 2016.

<sup>1</sup>Supported by National Natural Science Foundation of China (NSFC, 11131002, 11271031, 11525101, 71532001), the Business Intelligence Research Center at Peking University, the Center for Statistical Science at Peking University.

<sup>2</sup>Supported by National Natural Science Foundation of China (NSFC, 11601539, 11631003), Foundation for Development of Young Teachers at Central University of Finance and Economics (QJJ1519), the Program for Innovation Research in Central University of Finance and Economics.

<sup>3</sup>Supported by the Hong Kong RGC Grant (HKU702613P).

<sup>4</sup>Supported by National Natural Science Foundation of China (NSFC, 71301128, 71331005).

*MSC2010 subject classifications.* Primary 62M10; secondary 62J05.

*Key words and phrases.* Multivariate time series, ordinary least squares, social network, vector autoregression.

any node to be self-related, so that  $a_{ii} = 0$  for any  $1 \leq i \leq N$  (e.g., any Twitter user cannot follow itself). Let  $Y_{it} \in \mathbb{R}^1$  be the continuous response obtained from node  $i$  at time point  $t$  (e.g., log-transformed total tweet length). Accordingly,  $\mathbb{Y}_t = (Y_{1t}, \dots, Y_{Nt})^\top \in \mathbb{R}^N$  constitutes an ultra-high dimensional vector with a very large  $N$ , and its time series dynamic needs to be statistically modeled and theoretically investigated.

Note that  $\mathbb{Y}_t \in \mathbb{R}^N$  with  $0 \leq t \leq T$  is an ultra-high dimensional time series. Consequently, it has a close relationship with multivariate time series literature [4, 11, 15, 23, 30]. For a usual multivariate time series, the following common wisdoms exist. The first one is to model each individual time series (i.e.,  $Y_{it}$  with  $0 \leq t \leq T$  but for a fixed  $i$ ) separately [5, 8, 25]. This approach is simple in both theory and computation. However, the relationship across different time series is lost. As an alternative, one can model  $\mathbb{Y}_t$  by a vector autoregressive (VAR) model [3, 22]. Accordingly, the information of all time series is fully considered. However, the number of parameters need to be estimated is of  $O(N^2)$ , which could be much larger than  $T$ , if  $N$  is sufficiently large. Therefore, many efforts have been taken to reduce parameter dimension by either sparse estimation [16, 18] or dimension reduction [10, 24, 29]. In particular, dimension reduction by factor modeling has been proved very useful; see, for example, [13, 27, 28] and [19] for some interesting discussions. To our best knowledge, none of these methods have taken the observed social relationships (i.e., network structure) into consideration. This is the key contribution we intend to make in this work.

Under a network framework,  $Y_{it}$  might be affected by four different factors. First,  $Y_{it}$  might be affected by itself but from the previous time point, which is  $Y_{i(t-1)}$ . Second,  $Y_{it}$  might be affected by its followees, which are collected by  $\{j : a_{ij} = 1\}$ . Third,  $Y_{it}$  might also be affected by a set of node-specific covariates (denoted by a  $p$ -dimensional vector  $Z_i \in \mathbb{R}^p$ ). Lastly, the unexplained variation should be attributed to an independent random noise. As a result, we propose a network vector autoregressive (NAR) model, which assumes that  $Y_{it}$  is a linear combination of: (a)  $Y_{i(t-1)}$ , (b)  $n_i^{-1} \sum_j a_{ij} Y_{j(t-1)}$  with  $n_i = \sum_j a_{ij}$ , (c) node-specific covariates  $Z_i$ , and (d) an independent noise. The associated coefficients are then referred to as the *momentum effect*, the *network effect* and the *nodal effect*, respectively.

Compared with a usual VAR model, where the total number of parameters diverges with  $N$ , the total number of unknown parameters in a NAR model is fixed. Consequently, the NAR model can be easily estimated for large-scale social networks. In particular, an ordinary least squares type estimator is proposed, and its asymptotic properties are investigated. Furthermore, the NAR model's strictly stationary solutions are obtained. This leads to a number of interesting potential applications.

The rest of the article is organized as follows. Section 2 introduces the NAR model with both asymptotic theory and a  $p$ -lag extension. Section 3 describes two potential applications. Extensive numerical studies are given in Section 4. The

article is concluded with a brief discussion in Section 5. All technical details are left to the [Appendix](#).

## 2. Network vector autoregression.

2.1. *Model and notation.* Recall that  $N$  is the network size and  $Y_{it}$  is the response collected from the  $i$ th subject at time point  $t$ . In addition, for each node  $i$ , assume a  $p$ -dimensional node-specific random vector  $Z_i = (Z_{i1}, \dots, Z_{ip})^\top \in \mathbb{R}^p$  can be observed. To model  $Y_{it}$ , we propose the following NAR model:

$$(2.1) \quad Y_{it} = \beta_0 + Z_i^\top \gamma + \beta_1 n_i^{-1} \sum_{j=1}^N a_{ij} Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + \varepsilon_{it},$$

where  $n_i = \sum_{j \neq i} a_{ij}$  is the total number of nodes that  $i$  follows, and it is called out-degree [32]. This model implicitly assumes that the focal node  $i$  is unlikely to be affected by another node  $j$ , unless  $i$  follows  $j$ . This is true in practice since the activities of nodes with  $a_{ij} = 0$  are invisible to  $i$ . Specifically, the term  $\beta_0 + Z_i^\top \gamma$  characterizes the *nodal impact* of the  $i$ th node, where  $\beta_0 \in \mathbb{R}^1$  is the intercept and  $\gamma = (\gamma_1, \dots, \gamma_p)^\top \in \mathbb{R}^p$  is the associated coefficient (i.e., nodal effect). For convenience, we write  $\beta_{0i} = \beta_0 + Z_i^\top \gamma$ . As one can see, the nodal impact (i.e.,  $\beta_{0i}$ ) is invariant over time  $t$ . The quantity  $n_i^{-1} \sum_{j=1}^N a_{ij} Y_{j(t-1)}$  is the average impact from  $i$ th neighbors [1, 12, 20, 21]. Its associated parameter  $\beta_1$  is referred to as the network effect. The term  $Y_{i(t-1)}$  is the standard autoregressive impact, and  $\beta_2$  is referred to as the momentum effect. Moreover,  $\varepsilon_{it}$  is the error term independent of  $Z_i$ 's, which follows normal distribution with  $E(\varepsilon_{it}) = 0$  and  $\text{var}(\varepsilon_{it}) = \sigma^2$ . Although it is more flexible to allow the error terms (i.e.,  $\varepsilon_{it}$ 's) to be correlated, however, this might lead to large amount of unknown parameters. To reduce the dimensionality, a diagonal covariance structure is assumed in this work.

For convenience, define  $\mathbb{Z} = (Z_1, \dots, Z_N)^\top \in \mathbb{R}^{N \times p}$  and  $\mathcal{B}_0 = (\beta_{01}, \dots, \beta_{0N})^\top = \beta_0 \mathbf{1} + \mathbb{Z} \gamma \in \mathbb{R}^N$ , where  $\mathbf{1} = (1, \dots, 1)^\top$  is a vector with compatible dimension. Recall  $\mathbb{Y}_t = (Y_{1t}, \dots, Y_{Nt})^\top \in \mathbb{R}^N$ . Then we can rewrite model (2.1) in a vector form as

$$(2.2) \quad \mathbb{Y}_t = \mathcal{B}_0 + G \mathbb{Y}_{t-1} + \mathcal{E}_t,$$

where  $G = \beta_1 W + \beta_2 I$ ,  $W = \text{diag}\{n_1^{-1}, \dots, n_N^{-1}\} A$  is the row-normalized adjacency matrix,  $I$  is an identity matrix with compatible dimension, and  $\mathcal{E}_t = (\varepsilon_{1t}, \dots, \varepsilon_{Nt})^\top \in \mathbb{R}^N$  is the innovation vector. Since  $A$  is assumed to be nonrandom in this work, both  $G$  and  $W$  are nonstochastic. However,  $\mathcal{B}_0$  is random.

2.2. *Strict stationarity: Type I.* Since  $\mathbb{Y}_t$  is a time series, it is of interest to study its stationary distribution. According to whether  $N$  is fixed or  $N \rightarrow \infty$ , we can define two different types of stationarities. They are referred to as, respectively, Type I ( $N$  is fixed) and Type II ( $N \rightarrow \infty$ ). We start with Type I stationarity, because its discussion is similar to that of the classical time series model.

**THEOREM 1.** *Suppose that  $E\|Z_i\| < \infty$  and  $N$  is fixed. If  $|\beta_1| + |\beta_2| < 1$ , then there exists a unique strictly stationary solution with a finite first-order moment to the NAR model (2.2). The solution has the form of*

$$(2.3) \quad \mathbb{Y}_t = (I - G)^{-1}B_0 + \sum_{j=0}^{\infty} G^j \mathcal{E}_{t-j}.$$

The proof of Theorem 1 is given in a separate technical Appendix containing supplementary materials. Based on the strict stationary solution (2.3), it is interesting to obtain its conditional distribution given the nodal information (i.e.,  $\mathbb{Z}$ ). For convenience, define  $E^*(\cdot) = E(\cdot|\mathbb{Z})$  and  $\text{cov}^*(\cdot) = \text{cov}(\cdot|\mathbb{Z})$ . For any integer  $h$ , further denote the conditional auto-covariance function as  $\Gamma(h) = \text{cov}^*(\mathbb{Y}_t, \mathbb{Y}_{t-h})$ . One could easily verify that,  $\Gamma(h) = G^h\Gamma(0)$  for  $h > 0$ , and  $\Gamma(h) = \Gamma(0)(G^\top)^{-h}$  for  $h < 0$ . By calculation similar to [22] (pages 28–29), the conditional mean and covariance of  $\mathbb{Y}_t$  can be obtained in the following proposition.

**PROPOSITION 1.** *Assume the same conditions as in Theorem 1. Then given  $\mathbb{Z}$ , the strictly stationary solution in (2.3) follows a normal distribution with mean and covariance given by*

$$(2.4) \quad \mu = (I - G)^{-1}B_0 = (I - \beta_1 W - \beta_2 I)^{-1}B_0,$$

$$(2.5) \quad \text{vec}\{\Gamma(0)\} = \sigma^2(I - G \otimes G)^{-1} \text{vec}(I).$$

Here,  $\text{vec}(\cdot)$  is the usual operator stacking the columns of a given matrix, and  $\otimes$  is the Kronecker product.

By (2.4), the conditional mean of  $\mathbb{Y}_t$  is determined by four factors. They are the nodal impact  $B_0$ , the network effect  $\beta_1$ , the momentum effect  $\beta_2$ , and the network structure  $W$  (i.e.,  $A$ ). To better interpret Proposition 1, we consider some special and interesting cases for discussions.

*Case 1.* Assume  $\beta_{0i} = \beta_0$  for every  $1 \leq i \leq N$ . In this case, we implicitly assume that  $\gamma_j = 0$  for every  $1 \leq j \leq p$  so that all the nodes have the same nodal impact. Without loss of generality, we assume that  $\beta_0 > 0$ . It can be verified that  $\mathbf{1}$  is an eigenvector of  $(I - \beta_1 W - \beta_2 I)^{-1}$  with eigenvalue  $(1 - \beta_1 - \beta_2)^{-1}$ , and recall that  $\mathbf{1} = (1, \dots, 1)^\top$  is a vector with compatible dimension. Accordingly, we have  $(I - \beta_1 W - \beta_2 I)^{-1}\mathbf{1} = (1 - \beta_1 - \beta_2)^{-1}\mathbf{1}$ . Consequently, the conditional mean of response at each node has the same analytical form, which is referred to as *nodal mean* and is given by  $E^*(Y_{it}) = \beta_0(1 - \beta_1 - \beta_2)^{-1}$ . In this case, the nodal mean is no longer influenced by the network structure. Further, note that  $E^*(Y_{it}) \rightarrow \infty$ , as  $\beta_1 + \beta_2 \rightarrow 1$ . This indicates that the larger network effect (i.e.,  $\beta_1$ ) and/or larger momentum effect (i.e.,  $\beta_2$ ) all lead to a higher nodal mean.

*Case 2. Assume  $a_{ij} = 1$  for any  $i \neq j$ .* In this case, all the nodes follow each other, and thus the network is fully connected and very dense. This would never happen in real practice but is still discussed here for the sake of theoretical completeness. Under this assumption, one can verify that  $(I - \beta_1 W - \beta_2 I)^{-1} = \tau_1 I + \tau_2 \mathbf{1}\mathbf{1}^\top$ , where  $\tau_1 = \{1 + (N - 1)^{-1} \beta_1 - \beta_2\}^{-1}$ , and  $\tau_2 = \tau_1 \beta_1 (N - 1)^{-1} (1 - \beta_1 - \beta_2)^{-1}$ . Then, according to (2.4), the nodal mean can be derived as  $E^*(Y_{it}) = \tau_1 \beta_{0i} + \tau_2 \mathbf{1}^\top \mathcal{B}_0$ , where the difference lies in  $\beta_{0i}$  for different nodes. Under the stationary condition  $|\beta_1| + |\beta_2| < 1$ , one can easily verify that  $\tau_1 > 0$ . Accordingly, the node with larger nodal impact (i.e.,  $\beta_{0i}$ ) should have larger nodal mean.

*Case 3. First-order Taylor's expansion.* With a general network structure, it is difficult to interpret formula (2.4) and (2.5). As a compromise, we consider to approximate both  $E^*(\mathbb{Y}_t)$  and  $\text{cov}^*(\mathbb{Y}_t)$  by their first order Taylor's expansion with respect to  $\beta_1$ . As one can expect, such an approximation could be poor (good), if  $|\beta_1|$  is large (small). However, it does give us a quick opportunity to explore theoretical insights. More specifically, we have verified, in a separate technical appendix containing supplementary materials, that the first-order Taylor's expansion is given by

$$(2.6) \quad E^*(\mathbb{Y}_t) \approx \frac{\mathcal{B}_0}{1 - \beta_2} + \frac{\beta_1}{(1 - \beta_2)^2} W \mathcal{B}_0,$$

$$(2.7) \quad \text{cov}^*(\mathbb{Y}_t) \approx \frac{\sigma^2}{1 - \beta_2^2} I + \frac{\sigma^2 \beta_1 \beta_2}{(1 - \beta_2^2)^2} (W + W^\top).$$

By (2.6), we have that  $E^*(Y_{it}) \approx (1 - \beta_2)^{-1} \beta_{0i} + (1 - \beta_2)^{-2} \beta_1 (\sum_j a_{ij} \beta_{0j}) / n_i$ . Note that  $(\sum_j a_{ij} \beta_{0j}) / n_i$  is the average nodal impact from  $i$ 's neighbors. We refer to it as the *local impact* for node  $i$  and denote it as  $l_i$ . As a result, (2.6) suggests that larger nodal impact and local impact both lead to a higher nodal mean. By (2.7), we find that the variance of an arbitrary node is  $(1 - \beta_2^2)^{-1} \sigma^2$  approximately, which is only determined by the momentum effect  $\beta_2$  and the variance of  $\varepsilon_{it}$ . In addition, the covariance of nodes  $i_1$  and  $i_2$  is given by  $\sigma^2 (1 - \beta_2^2)^{-2} \beta_1 \beta_2 (a_{i_1 i_2} / n_{i_1} + a_{i_2 i_1} / n_{i_2})$ . This indicates that nodes following each other (i.e.,  $a_{i_1 i_2} = a_{i_2 i_1} = 1$ ) are more likely to be correlated than those disconnected ones (i.e.,  $a_{i_1 i_2} = a_{i_2 i_1} = 0$ ). The correlation could be even stronger if both nodes are *loyal* to each other, in the sense that they do not follow many other nodes (i.e., small  $n_{i_1}$  and  $n_{i_2}$ ).

**2.3. Strict stationarity: Type II.** Next, we investigate the Type II strict stationarity with  $N \rightarrow \infty$ . It is remarkable that the dimension of  $\mathbb{Y}_t$  (i.e.,  $N$ ) is diverging, so how to define stationarity is challenging. To our best knowledge, there exists no widely accepted general definition. As one possible solution, we attempt to give one definition as follows.

DEFINITION 1. Let  $\{\mathbb{Y}_t \in \mathbb{R}^N\}$  be a  $N$ -dimensional time series with  $N \rightarrow \infty$ . Define  $\mathcal{W} = \{\omega \in \mathbb{R}^\infty : \sum |\omega_i| < \infty\}$ , where  $\omega = (\omega_i \in \mathbb{R}^1 : 1 \leq i \leq \infty)^\top \in \mathbb{R}^\infty$ . For each  $\omega \in \mathcal{W}$ , let  $\mathbf{w}_N = (\omega_1, \dots, \omega_N)^\top \in \mathbb{R}^N$  be the truncated  $N$ -dimensional vector.  $\{\mathbb{Y}_t\}$  is then said to be strictly stationary, if it satisfies the following conditions: for any  $\omega \in \mathcal{W}$ , (1)  $Y_t^\omega = \lim_{N \rightarrow \infty} \mathbf{w}_N^\top \mathbb{Y}_t$  exists in the almost sure sense; and (2)  $\{Y_t^\omega\}$  is strictly stationary. Moreover,  $\{\mathbb{Y}_t\}$  is said to have finite  $m$ th order moment if  $\max_{1 \leq i < \infty} E|Y_{it}|^m < \infty$ .

In the usual situation with a fixed  $N$ , one can verify that  $\{\mathbb{Y}_t\}$  is strictly stationary if and only if  $\{\mathbf{w}_N^\top \mathbb{Y}_t\}$  is strictly stationary for any  $\mathbf{w}_N \in \mathbb{R}^N$ . As a result, Definition 1 can be viewed as an extension from the usual stationarity with fixed  $N$  to the diverging case. We then have the following theorem for the NAR model. The proof is given in Appendix A.1.

THEOREM 2. Assume the same conditions as in Theorem 1 with  $N \rightarrow \infty$ . Then the solution defined in (2.3) is a unique strictly stationary solution (in the sense of Definition 1) to the NAR model with finite first-order moment.

2.4. Parameter estimation. Let  $\beta = (\beta_0, \beta_1, \beta_2)^\top \in \mathbb{R}^3$  and  $\theta = (\theta_j)^\top = (\beta^\top, \gamma^\top)^\top \in \mathbb{R}^{p+3}$ . In order to estimate the unknown parameter  $\theta$ , we rewrite the NAR model in (2.1) as

$$(2.8) \quad Y_{it} = \beta_0 + \beta_1 w_i^\top \mathbb{Y}_{t-1} + \beta_2 Y_{i(t-1)} + Z_i^\top \gamma + \varepsilon_{it} = X_{i(t-1)}^\top \theta + \varepsilon_{it},$$

where  $X_{i(t-1)} = (1, w_i^\top \mathbb{Y}_{t-1}, Y_{i(t-1)}, Z_i^\top)^\top \in \mathbb{R}^{p+3}$ , and  $w_i = (a_{ij}/n_i : 1 \leq j \leq N)^\top \in \mathbb{R}^N$  is the  $i$ th row vector of  $W$ . Further, denote  $\mathbb{X}_t = (X_{1t}, X_{2t}, \dots, X_{Nt})^\top \in \mathbb{R}^{N \times (p+3)}$ . Then model (2.8) can be rewritten in vector form as  $\mathbb{Y}_t = \mathbb{X}_{t-1} \theta + \mathcal{E}_t$ . As a result, an ordinary least squares type estimator can be obtained as

$$(2.9) \quad \hat{\theta} = \left( \sum_{t=1}^T \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1} \right)^{-1} \sum_{t=1}^T \mathbb{X}_{t-1}^\top \mathbb{Y}_t,$$

whose asymptotic properties are to be investigated subsequently. To this end, we need the following technical conditions:

- (C1) *Nodal Assumption*) Assume that  $Z_i$ 's are independent and identically distributed random vectors, with mean 0 and covariance  $\Sigma_z \in \mathbb{R}^{p \times p}$ . Furthermore, its fourth-order moment is finite. The same assumption is also needed for  $\varepsilon_{it}$  across both  $1 \leq i \leq N$  and  $0 \leq t \leq T$ . Moreover, we need  $\{Z_i\}$  and  $\{\varepsilon_{it}\}$  to be mutually independent.
- (C2) *Network Structure*) Assume  $W$  is a sequence of matrices indexed by  $N$ . They are assumed to be nonstochastic.

- (C2.1) (*Connectivity*) Treat  $W$  as a transition probability matrix of a Markov chain, whose state space is defined as the set of all the nodes in the network (i.e.,  $\{1, \dots, N\}$ ). We assume the Markov chain is irreducible and aperiodic. Further, define  $\pi = (\pi_1, \dots, \pi_N)^\top \in \mathbb{R}^N$  as the stationary distribution of the Markov chain, such that (a)  $\pi_i \geq 0$  and  $\sum_{i=1}^N \pi_i = 1$ , (b)  $\pi = W^\top \pi$ . Furthermore,  $\sum_{i=1}^N \pi_i^2$  is assumed to converge to 0 as  $N \rightarrow \infty$ .
- (C2.2) (*Uniformity*) Define  $W^* = W + W^\top$  as a symmetric matrix. Assume  $\lambda_{\max}(W^*) = O(\log N)$ , where  $\lambda_{\max}(M)$  stands for the largest absolute eigenvalue of an arbitrary symmetric matrix  $M$ .
- (C3) (*Law of Large Numbers*) Define  $Q = (I - G)^{-1}(I - G^\top)^{-1}$ , and recall  $G = \beta_1 W + \beta_2 I$ . Assume the following limits exist. They are, respectively,  $\kappa_1 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{\Gamma(0)\}$ ,  $\kappa_2 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{W\Gamma(0)\}$ ,  $\kappa_3 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{(I - G)^{-1}\}$  and  $\kappa_4 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(Q)$ . Here,  $\kappa_1, \kappa_2, \kappa_3$  and  $\kappa_4$  are fixed constants.

Condition (C1) provides some basic assumptions for nodal variables  $Z_i$  and  $\varepsilon_{it}$ , so that the standard law of large numbers and central limit theorem can be applied. In fact, this assumption can be moderately relaxed so that different  $Z_i$ s and  $\varepsilon_{it}$ 's are weakly dependent, as long as the law of large numbers and central limit theorem hold.

Condition (C2) is about the network structure (i.e.,  $A$  or  $W$ ). It can be further divided into two sub-conditions. First, (C2.1) assumes that all the nodes are reachable to each other (i.e., irreducibility). Specifically, for two arbitrary nodes  $i$  and  $j$ , there should be a path of finite length connecting  $i$  to  $j$ . Otherwise, one can assume that there exist two sets of nodes, and the nodes from different sets are completely disconnected with each other. Then those two sets of nodes can be modeled separately. A simple sufficient condition for both irreducibility and aperiodicity is that the network is always fully connected after a finite number of steps. That is, there exists an  $n^*$  such that, for any  $n \geq n^*$ , each component in  $W^n$  is always positive. Accordingly, the network structure discussed in *Case 2* in Section 2.2 satisfies this sufficient condition. Second, (C2.2) requires that the network structure should admit certain uniformity property so that the diverging rate of  $\lambda_{\max}(W^*)$  should be sufficiently slow.

Lastly, condition (C3) is a law of large numbers type assumption. To see this consider, for example, the first condition in (C3), that is,  $\kappa_1 = \lim_{N \rightarrow \infty} N^{-1} \times \text{tr}\{\Gamma(0)\} = N^{-1} \sum_{i=1}^N \text{var}^*(Y_{it})$ . Heuristically, if we treat  $\sigma_i^2 = \text{var}^*(Y_{it})$  as if they were independent and identically generated random variables, then this is a law of large numbers type assumption. With the help of (C3), we can verify that the following limits also exist. They are  $\kappa_5 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(WQW^\top)$ ,  $\kappa_6 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{W\Gamma(0)W^\top\}$ ,  $\kappa_7 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(WQ)$  and  $\kappa_8 = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{W(I - G)^{-1}\}$ . In fact, we have  $\kappa_5 = (a + b)^2 \kappa_4 - 2(a + b)b\kappa_3 + b^2$ ,  $\kappa_6 = (a^2 + 2ab\beta_2 + b^2)\kappa_1 + 2a\kappa_2 - b^2\sigma^2$ ,  $\kappa_7 = (a + b)\kappa_4 - b\kappa_3$  and  $\kappa_8 =$

$(a + b)\kappa_3 - b$ . Here,  $W = aI + bG$ , where  $a = -\beta_1^{-1}\beta_2$ , and  $b = \beta_1^{-1}$ . In addition, one can easily verify that all the limits (i.e.,  $\{\kappa_1, \dots, \kappa_8\}$ ) exist even when  $\beta_1 = 0$ , as long as the limit of  $N^{-1} \text{tr}\{W W^\top\}$  exists. We then have the following theorem.

**THEOREM 3.** *Assume the stationary condition  $|\beta_1| + |\beta_2| < 1$  and technical conditions (C1)–(C3) hold, we then have  $\sqrt{NT}(\hat{\theta} - \theta) \rightarrow_d N(0, \sigma^2 \Sigma^{-1})$  as  $\min\{N, T\} \rightarrow \infty$ , where*

$$(2.10) \quad \Sigma = \begin{pmatrix} 1 & c_\beta & c_\beta & \mathbf{0}^\top \\ c_\beta & \Sigma_1 & \Sigma_2 & \kappa_8 \gamma^\top \Sigma_z \\ c_\beta & \Sigma_2 & \Sigma_3 & \kappa_3 \gamma^\top \Sigma_z \\ \mathbf{0} & \kappa_8 \Sigma_z \gamma & \kappa_3 \Sigma_z \gamma & \Sigma_z \end{pmatrix},$$

$c_\beta = \beta_0(1 - \beta_1 - \beta_2)^{-1}$ ,  $\Sigma_1 = c_\beta^2 + \kappa_5 \gamma^\top \Sigma_z \gamma + \kappa_6$ ,  $\Sigma_2 = c_\beta^2 + \kappa_7 \gamma^\top \Sigma_z \gamma + \kappa_2$ ,  $\Sigma_3 = c_\beta^2 + \kappa_4 \gamma^\top \Sigma_z \gamma + \kappa_1$ , and  $\mathbf{0} = (0, \dots, 0)^\top$  is a vector with compatible dimension.

The proof of Theorem 3 is given in Appendix A.2. By Theorem 3, we know that  $\hat{\theta}$  is  $\sqrt{NT}$ -consistent with asymptotic variance  $\sigma^2 \Sigma^{-1}$ .

It is remarkable that Theorem 3 requires both  $N \rightarrow \infty$  and  $T \rightarrow \infty$ . For the sake of completeness, we might also want to investigate the other two types of asymptotics. One is that  $N$  is fixed but  $T \rightarrow \infty$ , and the other one is that  $T$  is fixed but  $N \rightarrow \infty$ . However, the nodal effect  $\gamma$  is only associated with nodal information  $Z_i$ . Therefore, to estimate  $\gamma$  consistently, we must have ample amount of information about  $Z_i$ . As a result,  $N$  is required to be diverging, and no consistency result can be established for  $\hat{\theta}$  with fixed  $N$ . In contrast, our theoretical analysis shows that both the consistency and asymptotic normality can be established for  $\hat{\theta}$  with fixed  $T$  but  $N \rightarrow \infty$ . The main results are given in the following proposition.

**PROPOSITION 2.** *Assume  $T$  is fixed and conditions in Theorem 3 hold. Then we have  $\sqrt{N}(\hat{\theta} - \theta) \rightarrow_d N(0, \sigma^2 T^{-1} \Sigma^{-1})$  as  $N \rightarrow \infty$ .*

The proof of Proposition 2 is given in a separate technical Appendix containing supplementary materials. By Proposition 2, the estimated parameter  $\hat{\theta}$  is  $\sqrt{N}$ -consistent with fixed  $T$ . This enables us to obtain a consistent estimator even with limited time periods.

**2.5. General NAR( $p$ ) model.** Note that the proposed NAR model (2.1) considers only one lag. For simplicity, we refer to it as a NAR(1) model. As a flexible extension, one could consider the NAR( $p$ ) model as follows:

$$(2.11) \quad Y_{it} = \beta_0 + Z_i^\top \gamma + \sum_{m=1}^p \alpha_m n_i^{-1} \sum_{j=1}^N a_{ij} Y_{j(t-m)} + \sum_{m=1}^p \beta_m Y_{i(t-m)} + \varepsilon_{it}.$$



Let  $\mathbb{Y}_t^* = (\mathbb{Y}_t^\top, \mathbb{Y}_{t-1}^\top, \dots, \mathbb{Y}_{t-p+1}^\top)^\top \in \mathbb{R}^{Np}$ . Then the NAR( $p$ ) model (2.11) can be expressed in vector form as

$$(2.12) \quad \mathbb{Y}_t^* = \mathcal{B}_0^* + G^* \mathbb{Y}_{t-1}^* + \mathcal{E}_t^*,$$

with  $\mathcal{B}_0^* = (\mathcal{B}_0^\top, \mathbf{0}_{N(p-1)}^\top)^\top \in \mathbb{R}^{Np}$ ,  $\mathcal{E}_t^* = (\mathcal{E}_t^\top, \mathbf{0}_{N(p-1)}^\top)^\top \in \mathbb{R}^{Np}$ , and

$$G^* = \begin{pmatrix} \mathfrak{A} & \alpha_p W + \beta_p I_N \\ I_{N(p-1)} & O_{N(p-1), N} \end{pmatrix},$$

where  $\mathfrak{A} = (\alpha_1 W + \beta_1 I_N, \dots, \alpha_{p-1} W + \beta_{p-1} I_N) \in \mathbb{R}^{N \times N(p-1)}$ ,  $\mathbf{0}_n$  is the  $n$ -dimensional zero vector,  $O_{n_1, n_2}$  is the  $n_1 \times n_2$  dimensional zero matrix, and  $I_n$  is the  $n \times n$  dimensional identity matrix. Similarly, we give the strictly stationary solution of NAR( $p$ ) model in the next theorem.

**THEOREM 4.** *Assume  $E\|Z_i\| \leq \infty$ . If  $\sum_{m=1}^p (|\alpha_m| + |\beta_m|) < 1$ , then there exists a unique strictly stationary solution with finite first-order moment to the NAR( $p$ ) model (2.11). The solution takes the form  $\mathbb{Y}_t = \mathfrak{J} \mathbb{Y}_t^*$ , where  $\mathbb{Y}_t^* = (I - G^*)^{-1} \mathcal{B}_0^* + \sum_{j=0}^\infty G^{*j} \mathcal{E}_{t-j}^*$ ,  $\mathfrak{J}(I - G^*)^{-1} \mathcal{B}_0^* = (I - \tilde{G})^{-1} \mathcal{B}_0$ ,  $\tilde{G} = \sum_{m=1}^p (\alpha_m W + \beta_m I_N)$ , and  $\mathfrak{J} = \{I_N, O_{N, N(p-1)}\}$ .*

The proof of Theorem 4 is given in a separate technical Appendix containing supplementary materials. It is remarkable that the solution, given in Theorem 4, is the unique strictly stationary solution of NAR( $p$ ) model, regardless of whether  $N$  is fixed or  $N \rightarrow \infty$ .

We next consider the estimation of NAR( $p$ ) model (2.11). Assume the dimension of  $Z_i$  is  $q$  throughout this subsection. Then write  $X_{i(t-1)}^* = \{1, w_i^\top \mathbb{Y}_{t-1}, \dots, w_i^\top \mathbb{Y}_{t-p}, Y_{i(t-1)}, \dots, Y_{i(t-p)}, Z_i^\top\}^\top \in \mathbb{R}^{2p+q+1}$  and  $\mathbb{X}_{t-1}^* = (X_{1(t-1)}^*, \dots, X_{N(t-1)}^*)^\top \in \mathbb{R}^{N \times (2p+q+1)}$ . Further, write  $\theta^* = (\beta_0, \alpha^\top, \beta^\top, \gamma^\top)^\top \in \mathbb{R}^{2p+q+1}$ , where  $\alpha = (\alpha_1, \dots, \alpha_p)^\top$  and  $\beta = (\beta_1, \dots, \beta_p)^\top$ . Then (2.11) can be written as  $\mathbb{Y}_t = \mathbb{X}_{t-1}^* \theta^* + \mathcal{E}_t$ . The corresponding ordinary least squares type estimator can be derived as

$$(2.13) \quad \hat{\theta}^* = \left( \sum_{t=p+1}^T \mathbb{X}_{t-1}^{*\top} \mathbb{X}_{t-1}^* \right)^{-1} \sum_{t=p+1}^T \mathbb{X}_{t-1}^{*\top} \mathbb{Y}_t.$$

We next investigate the asymptotic properties of  $\hat{\theta}^*$ . To this end, define  $\Gamma^*(h) = \text{cov}^*(\mathbb{Y}_t, \mathbb{Y}_{t-h})$  to be the conditional auto-covariance function for the NAR( $p$ ) model under strict stationarity. The following condition is required:

(C4) Assume the following limits exist as,  $v_1(h) = \lim_{N \rightarrow \infty} N^{-1} W \Gamma^*(h) W^\top$ ,  $v_2(h) = \lim_{N \rightarrow \infty} N^{-1} W \Gamma^*(h)$ ,  $v_3(h) = \lim_{N \rightarrow \infty} N^{-1} \Gamma^*(h)$  for  $0 \leq h \leq p - 1$ , and  $\kappa_3^* = \lim_{N \rightarrow \infty} N^{-1} \text{tr}\{(I - \tilde{G})^{-1}\}$ ,  $\kappa_4^* = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(\tilde{Q})$ ,  $\kappa_5^* = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(W \tilde{Q} W^\top)$ ,  $\kappa_7^* = \lim_{N \rightarrow \infty} N^{-1} \text{tr}(W \tilde{Q})$ ,  $\kappa_8^* = \lim_{N \rightarrow \infty} N^{-1} \times \text{tr}\{W(I - \tilde{G})^{-1}\}$ , where  $\tilde{Q} = (I - \tilde{G})^{-1}(I - \tilde{G}^\top)^{-1}$ .

We then have the following theorem.

**THEOREM 5.** Assume  $\sum_{m=1}^p (|\alpha_m| + |\beta_m|) < 1$  and the technical conditions (C1), (C2), (C4) hold. We then have  $\sqrt{NT}(\hat{\theta}^* - \theta^*) \rightarrow_d N(0, \sigma^2 \Sigma^{*-1})$  as  $\min\{N, T\} \rightarrow \infty$ , where

$$\Sigma^* = \begin{pmatrix} 1 & c_\beta^* \mathbf{1}_p^\top & c_\beta^* \mathbf{1}_p^\top & \mathbf{0}_q^\top \\ c_\beta^* \mathbf{1}_p & \Sigma_1^* & \Sigma_2^* & \mathbf{1}_p (\kappa_8^* \gamma^\top \Sigma_z) \\ c_\beta^* \mathbf{1}_p & \Sigma_2^{*\top} & \Sigma_3^* & \mathbf{1}_p (\kappa_3^* \gamma^\top \Sigma_z) \\ \mathbf{0}_q & (\kappa_8^* \Sigma_z \gamma) \mathbf{1}_p^\top & (\kappa_3^* \Sigma_z \gamma) \mathbf{1}_p^\top & \Sigma_z \end{pmatrix},$$

$c_\beta^* = \{1 - \sum_{m=1}^p (\alpha_m + \beta_m)\}^{-1} \beta_0$ ,  $\Sigma_1^* = \{\Sigma_{1,(m_1,m_2)}^*\} \in \mathbb{R}^{p \times p}$  with  $\Sigma_{1,(m_1,m_2)}^* = c_\beta^{*2} + \kappa_5^* \gamma^\top \Sigma_z \gamma + v_1(m_1 - m_2)$ ,  $\Sigma_2^* = \{\Sigma_{2,(m_1,m_2)}^*\} \in \mathbb{R}^{p \times p}$  with  $\Sigma_{2,(m_1,m_2)}^* = c_\beta^{*2} + \kappa_7^* \gamma^\top \Sigma_z \gamma + v_2(m_1 - m_2)$ ,  $\Sigma_3^* = \{\Sigma_{3,(m_1,m_2)}^*\} \in \mathbb{R}^{p \times p}$  with  $\Sigma_{3,(m_1,m_2)}^* = c_\beta^{*2} + \kappa_4^* \gamma^\top \Sigma_z \gamma + v_3(m_1 - m_2)$ , and  $\mathbf{1}_p = (1, \dots, 1)^\top$  is a  $p$ -dimensional vector with all elements to be 1.

The proof of Theorem 5 is given in a separate technical Appendix containing supplementary materials.

### 3. Potential applications.

**3.1. Network structure analysis.** To demonstrate the usefulness of the proposed NAR model, we next consider a number of interesting potential applications. We start with network structure analysis. By Proposition 1, the conditional mean is given by  $E^*(\mathbb{Y}_t) = (I - \beta_2 I - \beta_1 W)^{-1} \beta_0$ . For a more intuitive discussion, we assume that  $Y_{it}$  is the measure for the  $i$ th node’s activeness at time  $t$ . Then the expected average network activeness (EANA) is given by  $\mathbf{1}^\top E^*(\mathbb{Y}_t)/N$ . Next, how to maximize this quantity becomes an important problem. Previous discussion shows that identical nodal impact (i.e.,  $\beta_{i0} = \beta_0$ ) leads to the same nodal mean with EANA given by  $\beta_0(1 - \beta_1 - \beta_2)^{-1}$ . In this case, EANA is free of the network structure, and cannot be maximized by optimizing network structure. To make our discussion meaningful, we assume nonidentical nodal impact throughout the rest of this subsection. Similar to Case 3 in Section 2.2, we also employ the Taylor’s expansion technique, which greatly simplifies the theoretical results. Consequently, we can focus on the leading terms, which leads to fruitful insights and interpretations.

We start with the simplest network structure change. Consider two different nodes  $i$  and  $j$  where  $a_{ij} = 0$ . Then, by (2.6), the nodal mean for  $i$  is given by  $\mu_{i0} \approx (1 - \beta_2)^{-1} \beta_{0i} + (1 - \beta_2)^{-2} \beta_1 \sum_{k \neq j} a_{ik} \beta_{0k} / n_i$ . Now assume  $i$  starts to follow  $j$ , that is,  $a_{ij}$  changes from 0 to 1. Then the nodal mean of  $i$  changes from  $\mu_{i0}$  to  $\mu_{i1} \approx (1 - \beta_2)^{-1} \beta_{0i} + (1 - \beta_2)^{-2} \beta_1 (\sum_{k \neq j} a_{ik} \beta_{0k} + \beta_{0j}) / (n_i + 1) + o_p(\beta_1)$ . Hence, the change in  $a_{ij}$  (from 0 to 1) leads to a change in the nodal mean for node

$i$  as

$$(3.1) \quad \mu_{i1} - \mu_{i0} \approx (1 - \beta_2)^{-2} \beta_1 \left( \beta_{0j} - \sum_{k \neq j} a_{ik} \beta_{0k} / n_i \right) / (n_i + 1).$$

Practically, the network effect  $\beta_1$  is typically expected to be positive [6]. Then the sign of the change is mainly determined by  $(\beta_{0j} - \sum_{k \neq j} a_{ik} \beta_{0k} / n_i)$ , which involves the  $j$ th nodal impact (i.e.,  $\beta_{0j}$ ) and the  $i$ th local impact (i.e.,  $\sum_{k \neq j} a_{ik} \beta_{0k} / n_i$ ). If the nodal impact of  $j$  is larger than the local impact of  $i$ , then the  $i$ th nodal mean is expected to increase. Otherwise, it should decrease.

The above discussion leads to the following interesting findings, that is, the nodal mean of an arbitrary node is likely to be increased by: either (1) following nodes with nodal impact larger than its local impact, or (2) unfollowing nodes with nodal impact smaller than its local impact. One can repeat this process for each node till the network converges to an optimal structure. To gain some quick insights, we assume that  $\beta_{0i}$  are all different and the nodes are appropriately sorted so that  $\beta_{01} > \beta_{02} > \beta_{0j}$  for any  $j > 2$ . In this case, for any node  $i > 1$  with  $a_{i1} = 0$ , its nodal mean can always be increased by setting  $a_{i1} = 1$  (i.e., following node  $i = 1$ ). Accordingly, we should have  $a_{i1} = 1$  for any  $i > 1$ . Otherwise, the network structure cannot be optimal. Next, for any node  $i$  with  $n_i > 1$ , its nodal mean can always be increased by unfollowing its neighbor with the smallest  $\beta_{0j}$  value. As a result, we should have  $n_i = 1$  and  $a_{i1} = 1$  for every  $i > 1$ . For node  $i = 1$ , similar argument leads to  $n_1 = 1$  and  $a_{12} = 1$ . In this case, one can verify that the value of EANA could be approximated by

$$(3.2) \quad (1 - \beta_2)^{-1} N^{-1} \sum_i \beta_{0i} + (1 - \beta_2)^{-2} N^{-1} (N - 1) \beta_1 \beta_{01}.$$

For a large-scale network, we have  $(N - 1)/N \approx 1$ . Hence, the above quantity can be further approximated by  $(1 - \beta_2)^{-1} N^{-1} \sum_i \beta_{0i} + (1 - \beta_2)^{-2} \beta_1 \beta_{01}$ . It is mainly determined by the average nodal impact (i.e.,  $N^{-1} \sum \beta_{0i}$ ) and the largest nodal impact  $\beta_{01}$ .

**3.2. Intervention analysis.** We study in this subsection one type of intervention analysis, which refers to some stimulus imposed on a set of nodes. The stimulus aims at enhancing the network activeness as much as possible. If there were no intervention, the network activeness would be fully determined by the NAR model (2.1), which leads to a sequence of response vectors as  $\{\mathbb{Y}_t\}$  with  $t \geq 0$ . Without loss of generality, we assume that the intervention is given at time  $t = 0$  to each node  $i$  so that its initial activeness level can be increased from  $Y_{i0}$  to  $Y_{i0} + \delta_i$ , for some  $\delta_i \geq 0$ . Equivalently,  $\mathbb{Y}_0$  is changed to  $\mathbb{Y}_0 + \Delta$ , where  $\Delta = (\delta_1, \dots, \delta_N)^\top$ . This leads to a stimulated new time series sequence, given by  $\{\mathbb{Y}_t^*\}$  with  $t \geq 0$ . Both  $\{\mathbb{Y}_t\}$  and  $\{\mathbb{Y}_t^*\}$  follow the same NAR model as given in (2.1), sharing the same set of innovation process  $\mathcal{E}_t$ . The only difference is that  $\{\mathbb{Y}_t\}$  starts with  $\mathbb{Y}_0$  while  $\{\mathbb{Y}_t^*\}$  starts with  $\mathbb{Y}_0 + \Delta$ . We can then evaluate the intervention effect by

comparing the difference between  $\mathbb{Y}_t$  and  $\mathbb{Y}_t^*$ . This leads to the total intervention effect (TIE) given by  $\text{TIE}(\Delta) = \mathbf{1}^\top \sum_{t>0} (\mathbb{Y}_t^* - \mathbb{Y}_t)$ . Practically, the resource (i.e.,  $\sum \delta_i = \mathbf{1}^\top \Delta$ ) that can be used is limited, and we assume that  $\sum \delta_i \leq 1$ . It is then of interest to maximize  $\text{TIE}(\Delta)$  under the constraint  $\mathbf{1}^\top \Delta \leq 1$ .

It can be easily verified that  $\mathbb{Y}_t^* = \mathbb{Y}_t + G^t \Delta$ . As a result,  $\text{TIE}(\Delta) = \mathbf{1}^\top \times \sum_{t>0} (\mathbb{Y}_t^* - \mathbb{Y}_t) = \mathbf{1}^\top (\Delta + G \Delta + \dots) = \mathbf{1}^\top (I - G)^{-1} \Delta$ , under the stationary condition of Theorem 1. Then the optimization problem becomes

$$(3.3) \quad \begin{aligned} \max_{\Delta \in \mathbb{R}^N} \text{TIE}(\Delta) &= \mathbf{1}^\top (I - G)^{-1} \Delta, \\ \text{such that } \sum_{i=1}^N \delta_i &\leq 1, \quad \delta_i \geq 0, i = 1, \dots, N. \end{aligned}$$

Note that (3.3) is a linear function in  $\Delta$  with coefficients given by  $v = (v_1, \dots, v_N)^\top = (I - G^\top)^{-1} \mathbf{1}$ . Without loss of generality, we assume that the nodes are appropriately sorted so that  $v_1 > v_2 > \dots > v_N$ . One can verify that  $v_1 > 0$ . Accordingly, by standard linear programming theory [9], the solution to (3.3) is given by  $\Delta_{\max} = (1, 0, 0, \dots, 0)^\top \in \mathbb{R}^N$ . This suggests that all the resources should be given to the node with the largest  $v_i$ -values. For a practical interpretation, we refer to  $v_i$  as node  $i$ 's *influential power*.

It is then of interest to query which type of nodes is likely to have large  $v_i$ -values (i.e., influential powers). This amounts to computing  $v = (I - G^\top)^{-1} \mathbf{1}$  and studying its relationship with the network structure  $A$ . Once again, we follow Case 3 in Section 2.2 and use first-order Taylor's expansion for approximation. This enables us to approximate  $v$  by  $(1 - \beta_2)^{-1} \mathbf{1} + (1 - \beta_2)^{-2} \beta_1 W^\top \mathbf{1}$ . We then have  $v_i \approx 1/(1 - \beta_2) + (1 - \beta_2)^{-2} \beta_1 \sum_j n_j^{-1} a_{ji}$ . By assuming a positive network effect  $\beta_1 > 0$ , we then find that a node's influential power is mainly determined by  $\sum_j n_j^{-1} a_{ji}$ , which is referred to as *weighted degree*. Nodes with larger weighted degree are typically ones with a larger number of followers (i.e.,  $\sum_j a_{ji}$ ), and those followers are *loyal* in the sense that their out-degree  $n_j$ s are small. Consequently,  $\text{TIE}(\Delta_{\max}) = \max_i v_i \approx 1/(1 - \beta_2) + (1 - \beta_2)^{-2} \beta_1 \max_i \{\sum_j n_j^{-1} a_{ji}\}$ .

### 4. Numerical studies.

4.1. *Simulation models.* To demonstrate the finite sample performance of the proposed methodology, we present in this subsection three examples. The main difference is the generating mechanism of the adjacency matrix  $A$  and also the specification of  $\beta = (\beta_0, \beta_1, \beta_2)^\top \in \mathbb{R}^3$ . Other than that, they are fairly similar. Specifically, for each example, the random error  $\varepsilon_{it}$  is simulated from a standard normal distribution  $N(0, 1)$ , and the covariate  $Z_i = (Z_{i1}, \dots, Z_{i5})^\top \in \mathbb{R}^5$  is from a multivariate normal distribution with mean  $\mathbf{0}$  and covariance  $\Sigma_z = (\sigma_{j_1 j_2})$ , where  $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$ . For each example,  $\gamma$  is fixed to be  $\gamma = (-0.5, 0.3, 0.8, 0, 0)^\top$ . In order to generate  $\mathbb{Y}_t$ , an initial value  $\mathbb{Y}_0$  is randomly simulated according to the

stationary distribution as given in Proposition 1. Once  $\mathbb{Y}_0$ s are given,  $\mathbb{Y}_t$ s can be generated according to (2.8).

**EXAMPLE 1 (Dyad Independence Model).** By [17], a dyad is defined as  $D_{ij} = (a_{ij}, a_{ji})$  for any  $1 \leq i < j \leq N$ . Dyad independence assumes that different  $D_{ij}$ 's are independent. In order to ensure network sparsity, we set  $P(D_{ij} = (1, 1)) = 20N^{-1}$ . As a result, the expected number of mutually connected dyads [i.e.,  $D_{ij} = (1, 1)$ ] is of  $O(N)$ . Next, set  $P(D_{ij} = (1, 0)) = P(D_{ij} = (0, 1)) = 0.5N^{-0.8}$ . This makes the expected degree for each node  $O(N^{0.2})$ , which diverges toward infinity as the network size  $N$  increases, but at a slow rate. Otherwise, the network sparsity might be violated. Accordingly, we should have  $P(D_{ij} = (0, 0)) = 1 - 20N^{-1} - N^{-0.8}$ , which is very close to 1 for large  $N$ . Lastly, fix  $T = 10, 30, 100$ , and  $(\beta_0, \beta_1, \beta_2)^\top = (0.3, 0.0, 0.5)^\top$ .

**EXAMPLE 2 (Stochastic Block Model).** We next consider another popular network structure, which is the stochastic block model [26, 31]. This model is of particular interest for community detection [33]. Specifically, we follow [26], and randomly assign for each node a block label ( $k = 1, \dots, K$ ) with equal probability, where  $K \in \{5, 10, 20\}$  is the total number of blocks. Next, set  $P(a_{ij} = 1) = 0.3N^{-0.3}$  if  $i$  and  $j$  belong to the same block, and  $P(a_{ij} = 1) = 0.3N^{-1}$  otherwise. Accordingly, the nodes within the same block are more likely to be connected, as compared with nodes from different blocks. Lastly, fix  $T = 30$  and  $(\beta_0, \beta_1, \beta_2)^\top = (0.0, 0.1, -0.2)^\top$ .

**EXAMPLE 3 (Power-Law Distribution Model).** By [2], a power-law distribution reflects a popular network phenomenon, that is, the majority of nodes have very few edges but a small amount have a huge number of edges. To mimic this phenomenon, we follow [7] and simulate  $A$  as follows. First, generate for each node its in-degree  $d_i = \sum_j a_{ji}$  according to the discrete power-law distribution, that is,  $P(d_i = k) = ck^{-\alpha}$  for a normalizing constant  $c$  and the exponent parameter  $\alpha \in \{1.2, 2, 3\}$ . Smaller  $\alpha$  value implies a heavier distribution tail. Next, for the  $i$ th node, we randomly select  $d_i$  nodes to be its followers. Lastly, fix  $T = 30$  and  $(\beta_0, \beta_1, \beta_2)^\top = (0.3, -0.1, 0.5)^\top$ .

**4.2. Performance measurements and simulation results.** For each simulation example, different network sizes are considered (i.e.,  $N = 100, 500, 1000$ ), and the experiment is randomly replicated  $R = 1000$  times. Let  $\hat{\theta}^{(r)} = (\hat{\theta}_j^{(r)})^\top = (\hat{\beta}_0^{(r)}, \hat{\beta}_1^{(r)}, \hat{\beta}_2^{(r)}, \hat{\gamma}^{(r)\top})^\top$  be the estimator obtained in the  $r$ th replication. We then consider the following measures to gauge their performances. First, for a given parameter  $\theta_j$  with  $1 \leq j \leq p + 3$ , the root mean square error is evaluated by  $\text{RMSE}_j = \{R^{-1} \sum_{r=1}^R (\hat{\theta}_j^{(r)} - \theta_j)^2\}^{1/2}$ . Next, for each  $1 \leq j \leq p + 3$ , a 95% confidence interval is constructed for  $\theta_j$  as  $\text{CI}_j^{(r)} = (\hat{\theta}_j^{(r)} -$

$z_{0.975}\widehat{\text{SE}}_j^{(r)}, \hat{\theta}_j^{(r)} + z_{0.975}\widehat{\text{SE}}_j^{(r)}$ , where  $\widehat{\text{SE}}_j^{(r)}$  is root square of the  $j$ th diagonal element of  $(\sum_t \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1})^{-1} \hat{\sigma}^2$  with  $\hat{\sigma}^2 = (NT)^{-1} \sum_{i,t} (Y_{it} - X_{it}^\top \hat{\theta}^{(r)})^2$ , and  $z_\alpha$  is the  $\alpha$ th quantile of a standard normal distribution. Then the coverage probability (CP) is computed as  $\text{CP}_j = R^{-1} \sum_{r=1}^R I(\theta_j \in \text{CI}_j^{(r)})$ , where  $I(\cdot)$  is the indicator function. Lastly, the total number of observed edges (i.e.,  $\sum_{i_1, i_2} a_{i_1 i_2}$ ) and the network density [i.e.,  $\{N(N-1)\}^{-1} \sum_{i_1, i_2} a_{i_1 i_2}$ ] are also reported.

The detailed simulation results are summarized in Tables 1–3. For the first example (i.e., Table 1), we find that, if  $T$  is fixed, the RMSE values for all  $\hat{\theta}_j$ 's decrease toward 0 as  $N$  increases. Consider for example  $\hat{\beta}_1$  (i.e., the estimated network effect) with  $T = 30$ , the RMSE value drops from 4.5% to 1.3%, as  $N$  increases from 100 to 1000. In the meanwhile, the network density drops from 22.7% to 2.4%, which implies that the network structure is increasingly sparse. Moreover, the reported coverage probability values for each parameter (i.e.,  $\theta_j$ ) are all fairly close to their nominal level 95%, which suggests that the estimated standard error (i.e.,  $\widehat{\text{SE}}$ ) approximates the true SE very well. Quantitatively similar findings are also obtained for Examples 2 and 3 from Tables 2 and 3. All these results confirm that the proposed estimator  $\hat{\theta}$  is indeed consistent and asymptotically normal.

4.3. *Simulation for potential applications.* Recall the two potential applications discussed in Section 3, which are the network structure analysis and the intervention analysis. Both of them are demonstrated through the following simulation studies. For each study, three different network structures are simulated (i.e., dyad independence model, stochastic block model and power-law distribution model) in the same manner as in the previous subsection. For illustration purposes, we fix  $N = 100, T = 100, K = 5, \alpha = 2.5$ , and  $(\beta_0, \beta_1, \beta_2)^\top = (0.2, 0.1, 0.4)^\top$ .

STUDY 1 (Network Structure Analysis). We first study the effect of network structure change discussed in Section 3.1. The procedure is implemented as follows:

*Step (1) (Initializing)* Simulate an initial network structure  $A^{(0)}$  as in the previous simulation examples.

*Step (2) (Alternating)*

(2.1) (*Network Structure*) Once  $A^{(k-1)}$  is given, we then update it to  $A^{(k)}$  ( $1 \leq k \leq N$ ) as follows. Define two different node sets, which are  $\Omega_{k0} = \{j : a_{kj} = 1 \text{ and } \beta_{0j} < l_k\}$  and  $\Omega_{k1} = \{j : j \neq k, a_{kj} = 0 \text{ and } \beta_{0j} > l_k\}$ , where  $l_k = n_k^{-1} \sum_j a_{kj} \beta_{0j}$  is the local impact received by node  $k$ . If  $|\Omega_{k0}| \geq 1$  (i.e., the size of  $\Omega_{k0}$ ), we then randomly select one node from  $\Omega_{k0}$  (denoted by  $j_1$ ), and change  $a_{kj_1}$  from 1 to 0. Similarly, if  $|\Omega_{k1}| \geq 1$ , we then randomly select one node from  $\Omega_{k1}$  (denoted by  $j_2$ ), and change  $a_{kj_2}$  from 0 to 1. Do nothing if  $|\Omega_{k0}| = 0$  or  $|\Omega_{k1}| = 0$ .

TABLE 1

Simulation results for Example 1 with 1000 replications. The RMSE values ( $\times 10^2$ ) are reported for every  $\beta$  and  $\gamma$  estimates. The CP (in %) of every estimate is given in parentheses. Total number of observed edges (TNOE) and network density (ND) are also reported

	<b>T = 10</b>			<b>T = 30</b>			<b>T = 100</b>		
	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>
$\beta_0$	6.4 (93.8)	2.6 (94.4)	1.8 (93.9)	3.7 (93.5)	1.4 (94.7)	1.0 (94.9)	2.0 (94.7)	0.8 (94.8)	0.5 (95.5)
$\beta_1$	7.6 (95.1)	3.3 (93.9)	2.2 (94.8)	4.5 (94.3)	1.9 (93.9)	1.3 (93.7)	2.5 (94.4)	1.0 (94.4)	0.7 (95.8)
$\beta_2$	2.9 (94.3)	1.2 (95.0)	0.9 (95.0)	1.6 (93.4)	0.7 (94.5)	0.5 (94.8)	0.9 (94.1)	0.4 (94.9)	0.3 (93.9)
$\gamma_1$	4.9 (94.5)	2.0 (94.6)	1.5 (94.1)	2.7 (94.7)	1.2 (95.7)	0.8 (95.9)	1.5 (95.2)	0.6 (96.8)	0.4 (95.8)
$\gamma_2$	4.8 (94.0)	2.0 (94.2)	1.4 (94.2)	2.6 (94.7)	1.2 (95.2)	0.8 (94.7)	1.5 (93.9)	0.6 (94.3)	0.5 (94.4)
$\gamma_3$	6.0 (96.5)	2.6 (96.5)	1.9 (94.9)	3.6 (94.3)	1.5 (94.4)	1.1 (95.4)	2.0 (93.5)	0.8 (94.9)	0.6 (94.9)
$\gamma_4$	4.2 (95.2)	1.8 (95.1)	1.3 (95.0)	2.5 (95.4)	1.1 (94.2)	0.7 (94.9)	1.3 (95.6)	0.6 (95.1)	0.4 (94.3)
$\gamma_5$	3.8 (95.6)	1.7 (94.2)	1.2 (95.2)	2.2 (94.2)	1.0 (94.8)	0.7 (95.3)	1.2 (95.3)	0.5 (94.7)	0.4 (95.1)
TNOE	2251	11,732	23,981	2251	11,732	23,981	2251	11,732	23,981
ND (%)	22.7	4.7	2.4	22.7	4.7	2.4	22.7	4.7	2.4

TABLE 2

Simulation results for Example 2 with 1000 replications. The RMSE values ( $\times 10^2$ ) are reported for every  $\beta$  and  $\gamma$  estimates. The CP (in %) of every estimate is given in parentheses. Total number of observed edges (TNOE) and network density (ND) are also reported

	<b>K = 5</b>			<b>K = 10</b>			<b>K = 20</b>		
	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>	<b>N = 100</b>	<b>N = 500</b>	<b>N = 1000</b>
$\beta_0$	1.8 (95.1)	0.9 (92.6)	0.6 (95.4)	1.9 (95.2)	0.8 (95.3)	0.6 (95.4)	1.9 (94.2)	0.8 (95.2)	0.6 (94.3)
$\beta_1$	2.1 (95.9)	1.1 (95.1)	1.1 (95.4)	1.7 (95.4)	0.9 (95.4)	0.8 (94.1)	1.7 (95.3)	0.8 (94.7)	0.6 (95.7)
$\beta_2$	1.8 (95.2)	0.8 (95.0)	0.6 (94.9)	1.8 (93.9)	0.8 (94.8)	0.5 (95.9)	1.8 (94.9)	0.8 (94.8)	0.6 (93.6)
$\gamma_1$	2.3 (94.8)	1.0 (95.3)	0.7 (94.7)	2.4 (94.1)	1.0 (94.0)	0.7 (95.0)	2.3 (95.3)	1.0 (94.6)	0.7 (94.8)
$\gamma_2$	2.5 (95.3)	1.0 (95.7)	0.8 (95.0)	2.6 (95.0)	1.1 (93.7)	0.8 (94.1)	2.5 (95.8)	1.1 (94.1)	0.8 (93.7)
$\gamma_3$	2.8 (94.2)	1.2 (95.5)	0.8 (95.4)	2.7 (95.2)	1.2 (95.0)	0.8 (96.8)	2.7 (94.9)	1.2 (94.6)	0.8 (95.6)
$\gamma_4$	2.4 (94.9)	1.0 (96.2)	0.7 (95.8)	2.4 (96.0)	1.1 (94.6)	0.7 (96.0)	2.5 (93.2)	1.1 (95.4)	0.7 (95.4)
$\gamma_5$	2.2 (95.0)	1.0 (94.9)	0.7 (94.9)	2.2 (93.7)	1.0 (94.7)	0.7 (95.2)	2.2 (94.7)	1.0 (94.7)	0.7 (94.5)
TNOE	261	2339	7523	173	1239	3792	193	969	2777
ND (%)	2.6	0.9	0.8	1.7	0.5	0.4	1.9	0.4	0.3



TABLE 3

Simulation results for Example 3 with 1000 replications. The RMSE values ( $\times 10^2$ ) are reported for every  $\beta$  and  $\gamma$  estimates. The CP (in %) of every estimate is given in parentheses. Total number of observed edges (TNOE) and network density (ND) are also reported

	$\alpha = 1.2$			$\alpha = 2.5$			$\alpha = 5.0$		
	$N = 100$	$N = 500$	$N = 1000$	$N = 100$	$N = 500$	$N = 1000$	$N = 100$	$N = 500$	$N = 1000$
$\beta_0$	5.2 (95.2)	4.4 (94.5)	4.0 (94.7)	2.7 (95.8)	1.1 (95.3)	0.8 (94.4)	2.2 (95.0)	1.0 (94.5)	0.7 (95.0)
$\beta_1$	8.7 (94.8)	8.4 (95.1)	7.8 (94.7)	2.5 (94.3)	0.9 (92.8)	0.6 (93.8)	1.2 (93.7)	0.5 (95.2)	0.3 (94.5)
$\beta_2$	1.6 (94.6)	0.7 (94.9)	0.5 (95.4)	1.6 (95.4)	0.7 (96.5)	0.5 (95.6)	1.6 (94.1)	0.7 (95.3)	0.5 (94.3)
$\gamma_1$	2.8 (95.5)	1.2 (95.4)	0.8 (95.4)	2.8 (95.0)	1.2 (94.2)	0.9 (94.7)	2.7 (94.9)	1.2 (95.1)	0.8 (94.5)
$\gamma_2$	2.6 (94.7)	1.2 (95.2)	0.8 (94.7)	2.6 (95.1)	1.1 (96.2)	0.8 (94.1)	2.6 (95.0)	1.2 (95.1)	0.8 (95.1)
$\gamma_3$	3.7 (94.0)	1.6 (94.6)	1.1 (94.4)	3.5 (94.7)	1.5 (95.9)	1.1 (94.7)	3.5 (95.0)	1.5 (95.5)	1.1 (94.3)
$\gamma_4$	2.5 (94.3)	1.1 (94.5)	0.8 (93.9)	2.4 (95.8)	1.0 (95.4)	0.8 (94.4)	2.5 (93.5)	1.1 (95.1)	0.8 (94.7)
$\gamma_5$	2.3 (93.9)	0.9 (96.9)	0.7 (95.8)	2.2 (94.5)	1.0 (95.3)	0.7 (94.2)	2.2 (95.3)	1.0 (94.2)	0.7 (94.1)
TNOE	3535	77,498	268,983	561	2227	4563	220	1060	2132
ND (%)	35.7	31.1	26.9	5.7	0.9	0.5	2.2	0.4	0.2

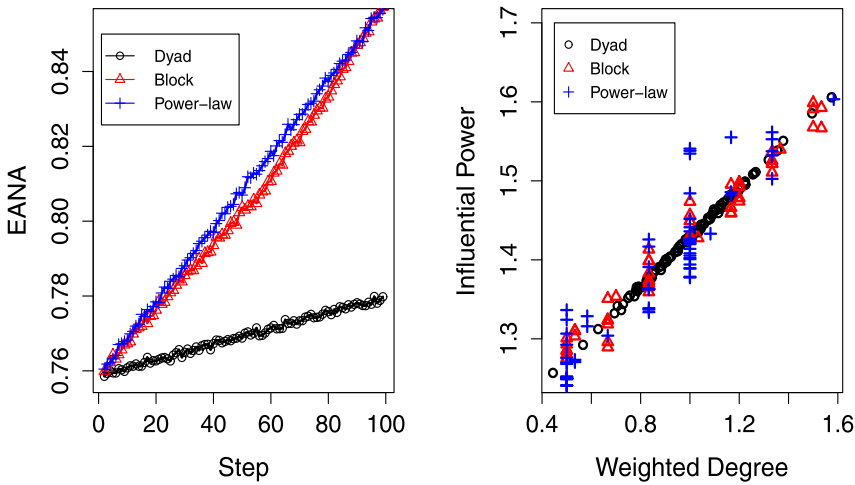


FIG. 1. Potential applications for three different network structures: black (o) for dyad independence model, red ( $\Delta$ ) for stochastic block model and blue (+) for power-law distribution model. The left panel: Network Structure Analysis; the right panel: Intervention Analysis.

(2.2) (Response Vector) With the updated network structure  $A^{(k)}$ , we simulate a series of response vectors as in the previous simulation examples and denote them by  $\mathbb{Y}_t^{(k)} = (Y_{it}^{(k)})^\top$  with  $1 \leq t \leq T$ .

Note that in Step (2.1), we distinguish  $k$ 's followers by comparing their nodal impacts (i.e.,  $\beta_{0j}$ ) with  $k$ 's local impact (i.e.,  $l_k$ ). By doing so, a node with nodal impact smaller than  $l_k$  is unfollowed, and a node with nodal impact larger than  $l_k$  is followed. Next, the expected average network activeness (i.e., EANA) can be estimated by  $\widehat{\text{EANA}}^{(k)} = (NT)^{-1} \sum_{i,t} Y_{it}^{(k)}$ . This leads to an EANA path  $\{\widehat{\text{EANA}}^{(k)} : 1 \leq k \leq N\}$ . For a reliable evaluation, we replicate the experiment 1000 times. This generates a total of 1000 EANA paths, which are then averaged and plotted in the left panel of Figure 1. A monotonically increasing pattern is detected, which corroborates our theoretical analysis in Section 3.1 quite well.

STUDY 2 (Intervention Analysis). In this study, we investigate the intervention effect discussed in Section 3.2. The network data are generated as in the previous study. According to the discussion in Section 3.2, a node with larger weighted degree is more likely to have larger influential power. To confirm this, we calculate for each node its weighted degree as  $\sum_j n_i^{-1} a_{ji}$ , and the corresponding influential power  $v_i$ , where  $v = (v_i)^\top = (I - G^\top)^{-1} \mathbf{1}$ . This leads to a scatter plot as given in the right panel of Figure 1. A clear monotonic relationship is detected, which indicates a strong positive correlation between weighted degree and influential power. It is remarkable that, for a large-scale network, an accurate computation of  $v$  is difficult, because an ultra-high dimensional matrix (i.e.,  $I - G^\top$ ) needs to be inverted.

However, the computation of weighted degree is much easier. This study confirms the practical usefulness of the weighted degree as a computationally feasible proxy for influential power.

4.4. *A Sina Weibo dataset.* We next illustrate our proposed method by a real example. The data are collected from Sina Weibo ([www.weibo.com](http://www.weibo.com)), which is the largest Twitter-type social media in China. Our dataset contains weekly observations for a total of  $N = 2982$  active followers of an official Weibo account. These Weibo users are observed for a total of  $T = 4$  consecutive weeks. The response (i.e.,  $Y_{it}$ ) considered here is the  $\log(1 + x)$ -transformed post length (i.e., the number of characters contained in the post) made by node  $i$  in week  $t$ . In addition, two time-invariant nodal covariates are recorded. They are the number of personal labels (created by the users to describe their life styles and career status) and the gender of each node (male = 1, female = 0). Lastly, the network structure  $A$  is naturally defined to be the followee–follower relationship. The resulting network density is around 2.2%.

We first provide the histogram of the in-degrees and out-degrees in Figure 2. It can be seen that the distribution of in-degrees is much more skewed than that of out-degrees. Their median values are 20 and 48, respectively. Next, the histogram of the responses is plotted in Figure 3. The response distribution is approximately normal with the mean value 6.26. To better motivate our method, we conduct some preliminary analysis. First, a simple linear regression is conducted for each node with  $Y_{it}$  as the response and  $Y_{i(t-1)}$  as the only covariate. As a consequence, R-squares can be computed for each node. This leads to a total of  $N = 2982$  R-square

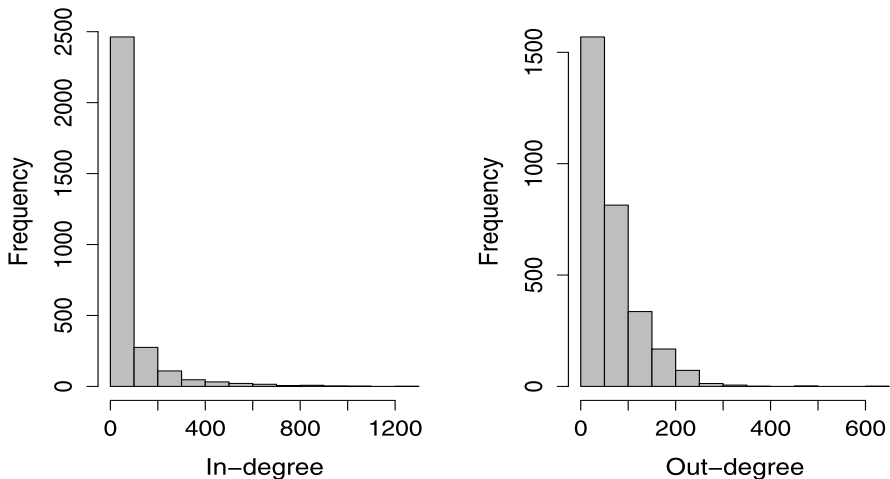


FIG. 2. *The Sina Weibo data analysis. The left panel: histogram of in-degrees for  $N = 2982$  nodes. The highly right skewed shape indicates the existence of “super stars” in the network; the right panel: similar histogram but for out-degrees.*

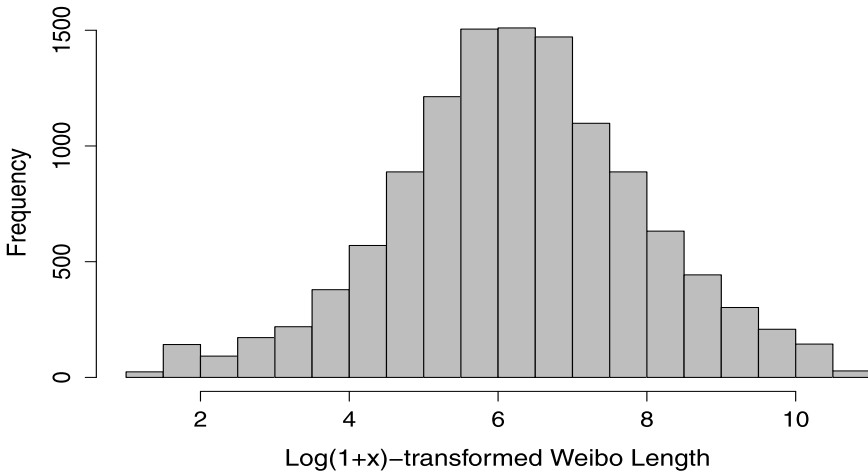


FIG. 3. *The Sina Weibo data analysis. Histogram of the response, which is the  $\log(1+x)$ -transformed Weibo length. The bell-shaped histogram suggests the response distribution is approximately normal.*

values, whose median level is around 31.8%. This suggests the existence of the momentum impact. Next, we compute residuals from this model for each node. These residuals are then treated as the responses and regressed against  $n_i^{-1} \sum_j a_{ij} Y_{j(t-1)}$  (i.e., the network impact). This leads to another  $N = 2982$  R-squares values, whose median is around 50.1%. This suggests that, even after controlling the momentum impact, the network effect exists.

Motivated by the preliminary analysis, we then fit the proposed NAR model to the Weibo dataset. The detailed estimation results are given in Table 4. As one can see, all estimates are at 5% level of significance. The estimated network effect (0.09) suggests that the activeness of a node is positively related to its connected neighbors. The estimated momentum effect (0.78) confirms that a node with higher (lower) activeness level in the past is likely to exhibit higher (lower) activeness in the future. The estimated nodal effects indicate that male users with more self-

TABLE 4  
*The detailed NAR analysis results for the Sina Weibo dataset*

Regression coefficient	Estimate	SE ( $\times 10^2$ )	p-value
Baseline Effect ( $\hat{\beta}_0$ )	0.53	13.13	<0.001
Network Effect ( $\hat{\beta}_1$ )	0.09	1.81	<0.001
Momentum Effect ( $\hat{\beta}_2$ )	0.78	0.68	<0.001
Number of Labels ( $\hat{\gamma}_1$ )	0.02	0.31	<0.001
Gender ( $\hat{\gamma}_2$ )	0.10	2.42	<0.001

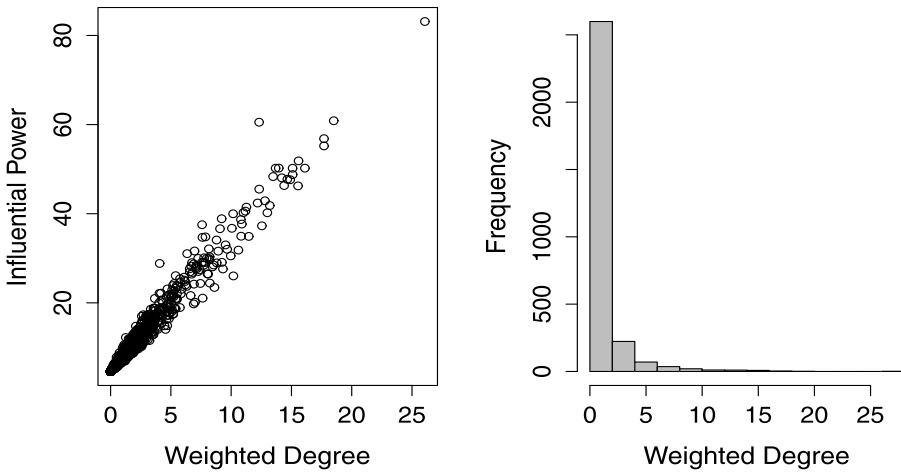


FIG. 4. *The Sina Weibo data analysis. The left panel: weighted degree versus influential power; the right panel: histogram for weighted degree.*

created labels tend to be more active. Next, in order to evaluate the out-sample prediction performance, we use the data from the first three weeks for estimation, and observations in the last week to evaluate its prediction accuracy. The resulting mean absolute prediction error (MAPE) is 0.78. As an alternative solution, one can also fit an AR(1) model to each individual node and then evaluate its prediction accuracy in a similar manner. The resulting MAPE is 3.34, which is substantially larger than that of the NAR model (i.e., 0.78). We do not consider a VAR model as a competitor for this particular dataset. This is because, with  $T = 4$  and  $N = 2982$ , the total number of the parameters demanded by a VAR model is too huge to be estimated.

Finally, we also calculate for each node its weighted degree and influential power, which are summarized in Figure 4. Once again, a positive relationship is observed; see the left panel in Figure 4. This further confirms that weighted degree should be a practically useful proxy for influential power. Meanwhile, the distribution of weighted degree is found to be heavily tailed; see the right panel in Figure 4. Further calculation reveals that only 26% of these top nodes carry about 80% of the total weighted degree. This means that a small fraction of the top nodes possesses the majority of the influential power in the whole network. Therefore, they should be the focus of the network intervention and marketing.

**5. Concluding remarks.** To conclude the article, we discuss here several interesting topics for future study. First, the NAR model proposed here requires the response to be continuous. However, discrete responses are commonly encountered in real practice. Then how to extend the NAR model for noncontinuous responses is the first topic for future study. Second, the network structure

discussed here is assumed to be static. However, in reality the network structure slowly changes with time. Then how to model such a network structure dynamic is another interesting problem worthwhile pursuing. Third, the NAR model assumes that, as long as  $a_{ij_1} \neq 0$  and  $a_{ij_2} \neq 0$ , the coefficients for  $Y_{j_1(t-1)}$  and  $Y_{j_2(t-1)}$  are the same. One may consider a more flexible model as  $Y_{it} = \beta_{0i} + \sum_{j=1}^N a_{ij} \eta_j Y_{j(t-1)} + \beta_2 Y_{i(t-1)} + \varepsilon_{it}$ . Then how to estimate those  $\eta_j$ 's is a problem of great importance. Lastly, the NAR model is established for the whole network. However, if the network is observed partially (e.g., sampled network), the resulting estimate for network effect (i.e.,  $\beta_1$ ) could be biased. Our finding corroborates that of [6] and [34] very well. Then how to correct this bias should be a challenging and interesting topic for future study.

APPENDIX

**A.1. Proof of Theorem 2.** Note that  $\{\mathbb{Y}_t\}$  defined as in (2.3) satisfies the NAR model (2.2) for any  $N$ . To prove the existence of a stationary solution, it is sufficient to show that  $\{\mathbb{Y}_t\}$  in (2.3) is strictly stationary according to Definition 1.

Define  $|M|_e$  as  $|M|_e = (|m_{ij}|) \in \mathbb{R}^{n \times p}$  for any arbitrary matrix  $M = (m_{ij}) \in \mathbb{R}^{n \times p}$ . Moreover, for matrices  $M_1 = (m_{ij}^{(1)}) \in \mathbb{R}^{n \times p}$  and  $M_2 = (m_{ij}^{(2)}) \in \mathbb{R}^{n \times p}$ , define  $M_1 \preceq M_2$  as  $m_{ij}^{(1)} \leq m_{ij}^{(2)}$  for  $1 \leq i \leq n$  and  $1 \leq j \leq p$ . Note that  $E|\mathcal{B}_0 + \mathcal{E}_{t-j}|_e \preceq (|\beta_0| + E|Z_i^\top \gamma| + E|\varepsilon_{it}|)\mathbf{1}$  and  $|G|_e^j \mathbf{1} = (|\beta_1|W + |\beta_2|I)^j \mathbf{1} = (|\beta_1| + |\beta_2|)^j \mathbf{1}$ . As a result,  $E|\mathbf{w}_N^\top \sum_{j=0}^\infty G^j(\mathcal{B}_0 + \mathcal{E}_{t-j})| \leq \sum_{i=1}^\infty |\omega_i| \sum_{j=0}^\infty (|\beta_1| + |\beta_2|)^j < \infty$ , which implies that  $\lim_{N \rightarrow \infty} \mathbf{w}_N^\top \mathbb{Y}_t = \lim_{N \rightarrow \infty} \mathbf{w}_N^\top \sum_{j=0}^\infty G^j(\mathcal{B}_0 + \mathcal{E}_{t-j})$  exists with probability one. Let  $Y_t^\omega = \lim_{N \rightarrow \infty} \mathbf{w}_N^\top \mathbb{Y}_t$ , and it is obvious that  $\{Y_t^\omega\}$  is strictly stationary. Hence,  $\{\mathbb{Y}_t\}$  is strictly stationary according to Definition 1.

Next, we verify the uniqueness of the strictly stationary solution. Assume that  $\{\tilde{\mathbb{Y}}_t\}$  is another strictly stationary solution to the NAR model with finite first order moment. Therefore,  $E|\tilde{\mathbb{Y}}_t|_e \preceq C_1 \mathbf{1}$  for some constant  $C_1$ . Then we have  $E|\mathbf{w}_N^\top \mathbb{Y}_t - \mathbf{w}_N^\top \tilde{\mathbb{Y}}_t| = E|\sum_{j=m}^\infty \mathbf{w}_N^\top G^j(\mathcal{B}_0 + \mathcal{E}_{t-j}) - \mathbf{w}_N^\top G^m \tilde{\mathbb{Y}}_{t-m}| \leq C_2 \sum_{j=m}^\infty \{(|\beta_1| + |\beta_2|)^j + C_1 (|\beta_1| + |\beta_2|)^m\} \sum_{i=1}^\infty |\omega_i|$  for any  $N$  and weight  $\omega$ , where  $C_2 = |\beta_0| + E|Z_i^\top \gamma| + E|\varepsilon_{it}|$ . Consequently, by the arbitrary specification of  $m$ , we have  $Y_t^\omega = \tilde{Y}_t^\omega$  with probability one. This completes the proof of Theorem 2.

**A.2. Proof of Theorem 3.** By (2.9),  $\hat{\theta}$  can be written as  $\hat{\theta} = \theta + \hat{\Sigma}^{-1} \hat{\Sigma}_{xe}$ , where  $\hat{\Sigma} = (NT)^{-1} \sum_{t=1}^T \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1}$  and  $\hat{\Sigma}_{xe} = (NT)^{-1} \sum_{t=1}^T \mathbb{X}_{t-1}^\top \mathcal{E}_t$ . As a result, the conclusion of Theorem 3 holds if

$$(A.1) \quad \hat{\Sigma} \xrightarrow{p} \Sigma,$$

$$(A.2) \quad \sqrt{NT} \hat{\Sigma}_{xe} \xrightarrow{d} N(0, \sigma^2 \Sigma),$$

as  $\min\{N, T\} \rightarrow \infty$ . Actually, (A.1) only requires  $N \rightarrow \infty$  in the following proofs. Subsequently, (A.1) will be proved in Step 1, and (A.2) will be proved in Step 2.

*Step 1. Proof of (A.1).* In this step, we attempt to prove that

$$\widehat{\Sigma} = \frac{1}{NT} \sum_{t=1}^T \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1} = \begin{pmatrix} 1 & \mathcal{S}_{12} & \mathcal{S}_{13} & \mathcal{S}_{14} \\ & \mathcal{S}_{22} & \mathcal{S}_{23} & \mathcal{S}_{24} \\ & & \mathcal{S}_{33} & \mathcal{S}_{34} \\ & & & \mathcal{S}_{44} \end{pmatrix} \xrightarrow{p} \begin{pmatrix} 1 & c_\beta & c_\beta & 0 \\ & \Sigma_1 & \Sigma_2 & \kappa_8 \gamma^\top \Sigma_z \\ & & \Sigma_3 & \kappa_3 \gamma^\top \Sigma_z \\ & & & \Sigma_z \end{pmatrix},$$

where

$$\mathcal{S}_{12} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1}, \quad \mathcal{S}_{13} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)}, \quad \mathcal{S}_{14} = \frac{1}{N} \sum_{i=1}^N Z_i^\top,$$

$$\mathcal{S}_{22} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N (w_i^\top \mathbb{Y}_{t-1})^2, \quad \mathcal{S}_{23} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1} Y_{i(t-1)},$$

$$\mathcal{S}_{24} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1} Z_i^\top, \quad \mathcal{S}_{33} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)}^2,$$

$\mathcal{S}_{34} = (NT)^{-1} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)} Z_i^\top$ ,  $\mathcal{S}_{44} = N^{-1} \sum_{i=1}^N Z_i Z_i^\top$ , and  $c_\beta = \beta_0(1 - \beta_1 - \beta_2)^{-1}$ ,  $\Sigma_1 = c_\beta^2 + \kappa_5 \gamma^\top \Sigma_z \gamma + \kappa_6$ ,  $\Sigma_2 = c_\beta^2 + \kappa_7 \gamma^\top \Sigma_z \gamma + \kappa_2$ , and  $\Sigma_3 = c_\beta^2 + \kappa_4 \gamma^\top \Sigma_z \gamma + \kappa_1$  are defined in Theorem 3. Note that, by equation (2.2),

$$(A.3) \quad \mathbb{Y}_t = c_\beta \mathbf{1} + (I - G)^{-1} \mathbb{Z} \gamma + \widetilde{\mathbb{Y}}_t$$

almost surely, where  $\widetilde{\mathbb{Y}}_t = \sum_{j=0}^\infty G^j \mathcal{E}_{t-j}$ ,  $\mathbb{Z} \gamma = (Z_1^\top \gamma, \dots, Z_N^\top \gamma)^\top$ , and  $\{Z_i^\top \gamma\}$  are i.i.d. random variables. By the law of large numbers, it is obvious that  $\mathcal{S}_{44} \rightarrow_p \Sigma_z$  and  $\mathcal{S}_{14} \rightarrow_p 0$  by (C1). We next show the convergence of the other entries in  $\widehat{\Sigma}$  one by one.

*Step 1.1. Convergence of  $\mathcal{S}_{12}$ .* Note that

$$\mathcal{S}_{12} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1} = \frac{1}{NT} \sum_{t=1}^T \mathbf{1}^\top W \mathbb{Y}_{t-1} = \frac{\beta_0}{1 - \beta_1 - \beta_2} + \mathcal{S}_{12a} + \mathcal{S}_{12b},$$

where  $\mathcal{S}_{12a} = N^{-1} \mathbf{1}^\top W (I - G)^{-1} \mathbb{Z} \gamma$  and  $\mathcal{S}_{12b} = (NT)^{-1} \sum_{t=1}^T \mathbf{1}^\top W \widetilde{\mathbb{Y}}_{t-1}$ . Then by (5.4) and (5.2) in Lemma 2, we have  $N^{-2} \mathbf{1}^\top W Q W^\top \mathbf{1} \rightarrow 0$  and  $N^{-1} \sum_{j=0}^\infty \{\mathbf{1}^\top W G^j (G^\top)^j W^\top \mathbf{1}\}^{1/2} \rightarrow 0$ , as  $N \rightarrow \infty$ . As a result, it is implied by Lemma 1(a) and (c) that  $\mathcal{S}_{12a} \rightarrow_p 0$  and  $\mathcal{S}_{12b} \rightarrow_p 0$ .

*Step 1.2. Convergence of  $\mathcal{S}_{13}$ .* Note that

$$\mathcal{S}_{13} = \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)} = \frac{1}{NT} \sum_{t=1}^T \mathbf{1}^\top \mathbb{Y}_{t-1} = \frac{\beta_0}{1 - \beta_1 - \beta_2} + \mathcal{S}_{13a} + \mathcal{S}_{13b},$$

where  $\mathcal{S}_{13a} = N^{-1}\mathbf{1}^\top(I - G)^{-1}\mathbb{Z}\gamma$  and  $\mathcal{S}_{13b} = (NT)^{-1}\sum_{t=1}^T\mathbf{1}^\top\tilde{\mathbb{Y}}_{t-1}$ . Similarly, we have  $N^{-2}\mathbf{1}^\top Q\mathbf{1} \rightarrow 0$  and  $N^{-1}\sum_{j=0}^\infty\{\mathbf{1}^\top G^j(G^\top)^j\mathbf{1}\}^{1/2} \rightarrow 0$ , as  $N \rightarrow \infty$ , by (5.4) and (5.2) of Lemma 2. As a result, by Lemma 1(a) and (c), we have  $\mathcal{S}_{13a} \rightarrow_p 0$ ,  $\mathcal{S}_{13b} \rightarrow_p 0$ , and  $\mathcal{S}_{13} \rightarrow_p c_\beta$ .

*Step 1.3. Convergence of  $\mathcal{S}_{22}$ .* Note that

$$\begin{aligned} \mathcal{S}_{22} &= \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N (w_i^\top \mathbb{Y}_{t-1})^2 = \frac{1}{NT} \sum_{t=1}^T \mathbb{Y}_{t-1}^\top W^\top W \mathbb{Y}_{t-1} \\ &= \frac{\beta_0^2}{(1 - \beta_1 - \beta_2)^2} + \mathcal{S}_{22a} + \mathcal{S}_{22b} + 2\mathcal{S}_{22c} + \frac{2\beta_0(\mathcal{S}_{12a} + \mathcal{S}_{12b})}{1 - \beta_1 - \beta_2}, \end{aligned}$$

where  $\mathcal{S}_{22a} = N^{-1}(\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}W^\top W(I - G)^{-1}\mathbb{Z}\gamma$ ,  $\mathcal{S}_{22b} = (NT)^{-1} \times \sum_{t=1}^T \tilde{\mathbb{Y}}_{t-1}^\top W^\top W \tilde{\mathbb{Y}}_{t-1}$  and  $\mathcal{S}_{22c} = (NT)^{-1} \sum_{t=1}^T (\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}W^\top W \tilde{\mathbb{Y}}_{t-1}$ . By (5.5), (5.3) and (5.6),  $N^{-1} \sum_{j=0}^\infty [\text{tr}\{WG^j(G^\top)^j W^\top W Q W^\top\}]^{1/2} \rightarrow 0$  and  $N^{-1} \sum_{i,j=0}^\infty [\text{tr}\{WG^i(G^\top)^i W^\top WG^j(G^\top)^j W^\top\}]^{1/2} \rightarrow 0$ . Furthermore, one can verify that  $N^{-2} \text{tr}\{(W Q W^\top)^2\} \rightarrow 0$ . Then by Lemma 1(b), (d) and (e), we have  $\mathcal{S}_{22a} \rightarrow_p \kappa_5 \gamma^\top \Sigma_z \gamma$ ,  $\mathcal{S}_{22b} \rightarrow_p \kappa_6$ ,  $\mathcal{S}_{22c} \rightarrow_p 0$ . Consequently, we have  $\mathcal{S}_{22} \rightarrow_p c_\beta^2 + \kappa_5 \gamma^\top \Sigma_z \gamma + \kappa_6$ , where  $\kappa_5$  and  $\kappa_6$  are defined in Theorem 3.

*Step 1.4. Convergence of  $\mathcal{S}_{23}$ .* Note that

$$\begin{aligned} \mathcal{S}_{23} &= \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1} Y_{i(t-1)} = \frac{1}{NT} \sum_{t=1}^T \mathbb{Y}_{t-1}^\top W \mathbb{Y}_{t-1} \\ &= \frac{\beta_0^2}{(1 - \beta_1 - \beta_2)^2} + \mathcal{S}_{23a} + \mathcal{S}_{23b} + 2\mathcal{S}_{23c} + \frac{2\beta_0(\mathcal{S}_{12a} + \mathcal{S}_{12b})}{1 - \beta_1 - \beta_2}, \end{aligned}$$

where  $\mathcal{S}_{23a} = N^{-1}(\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}W(I - G)^{-1}\mathbb{Z}\gamma$ ,  $\mathcal{S}_{23b} = NT^{-1} \sum_{t=1}^T \tilde{\mathbb{Y}}_{t-1}^\top \times W \tilde{\mathbb{Y}}_{t-1}$  and  $\mathcal{S}_{23c} = NT^{-1} \sum_{t=1}^T (\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}W \tilde{\mathbb{Y}}_{t-1}$ . Further, note that  $W Q W^\top = f_{1,1}(W, Q)$ ,  $Q = f_{0,1}(W, Q)$ ,  $G^j(G^\top)^j = g_{j,0,1}(G, W)$  and  $W G^j(G^\top)^j W^\top = g_{j,1,1}(G, W)$ . Then by (5.5), (5.3) and (5.6) of Lemma 2,  $N^{-2} \text{tr}\{(W Q W^\top) Q\} \rightarrow 0$ ,  $N^{-1} \sum_{i,j=0}^\infty [\text{tr}\{G^i(G^\top)^i W G^j(G^\top)^j W^\top\}]^{1/2} \rightarrow 0$ , and  $N^{-1} [\text{tr}\{W G^j(G^\top)^j W^\top Q\}]^{1/2} \rightarrow 0$ . Therefore, by Lemma 1(b), (d) and (e),  $\mathcal{S}_{23a} \rightarrow_p \kappa_7 \gamma^\top \Sigma_z \gamma$ ,  $\mathcal{S}_{23b} \rightarrow_p \kappa_2$ ,  $\mathcal{S}_{23c} \rightarrow_p 0$  and  $\mathcal{S}_{23} \rightarrow_p c_\beta^2 + \kappa_7 \gamma^\top \Sigma_z \gamma + \kappa_2$ , where  $\kappa_7$  and  $\kappa_2$  are defined in Theorem 3.

*Step 1.5. Convergence of  $\mathcal{S}_{33}$ .*

$$\begin{aligned} \mathcal{S}_{33} &= \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)}^2 = \frac{1}{NT} \sum_{t=1}^T \mathbb{Y}_{t-1}^\top \mathbb{Y}_{t-1} \\ &= \frac{\beta_0^2}{(1 - \beta_1 - \beta_2)^2} + \mathcal{S}_{33a} + \mathcal{S}_{33b} + 2\mathcal{S}_{33c} + \frac{2\beta_0(\mathcal{S}_{13a} + \mathcal{S}_{13b})}{1 - \beta_1 - \beta_2}, \end{aligned}$$



where  $\mathcal{S}_{33a} = N^{-1}(\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}(I - G)^{-1}\mathbb{Z}\gamma$ ,  $\mathcal{S}_{33b} = (NT)^{-1}\sum_{t=1}^T\tilde{\mathbb{Y}}_{t-1}^\top \times \tilde{\mathbb{Y}}_{t-1}$ , and  $\mathcal{S}_{33c} = (NT)^{-1}\sum_{t=1}^T(\mathbb{Z}\gamma)^\top(I - G^\top)^{-1}\tilde{\mathbb{Y}}_{t-1}$ . By (5.5), (5.3) and (5.6), we have  $N^{-2}\text{tr}(Q^2) \rightarrow 0$ ,  $N^{-1}\sum_{i,j=0}^\infty[\text{tr}\{G^i(G^\top)^i G^j(G^\top)^j\}]^{1/2} \rightarrow 0$ , and  $N^{-1}\sum_{j=0}^\infty[\text{tr}\{G^j(G^\top)^j Q\}]^{1/2} \rightarrow 0$ , as  $N \rightarrow \infty$ . Hence, by Lemma 1(b), (d) and (e),  $\mathcal{S}_{33a} \rightarrow_p \kappa_4\gamma^\top \Sigma_z \gamma$ ,  $\mathcal{S}_{33b} \rightarrow_p \kappa_1$ ,  $\mathcal{S}_{33c} \rightarrow_p 0$  and  $\mathcal{S}_{33} \rightarrow_p c_\beta^2 + \kappa_4\gamma^\top \Sigma_z \gamma + \kappa_1$ , where  $\kappa_4$  and  $\kappa_1$  are defined in Theorem 3.

Step 1.6. *Convergence of  $\mathcal{S}_{24}$  and  $\mathcal{S}_{34}$ .* Note that

$$\begin{aligned} \mathcal{S}_{24} &= \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N w_i^\top \mathbb{Y}_{t-1} Z_i^\top = \frac{1}{NT} \sum_{t=1}^T \mathbb{Y}_{t-1}^\top W^\top \mathbb{Z} = \mathcal{S}_{24c} + \mathcal{S}_{24a} + \mathcal{S}_{24b}, \\ \mathcal{S}_{34} &= \frac{1}{NT} \sum_{t=1}^T \sum_{i=1}^N Y_{i(t-1)} Z_i^\top = \frac{1}{NT} \sum_{t=1}^T \mathbb{Y}_{t-1}^\top \mathbb{Z} = \mathcal{S}_{24c} + \mathcal{S}_{34a} + \mathcal{S}_{34b}, \end{aligned}$$

where  $\mathcal{S}_{24a} = (NT)^{-1}\sum_{t=1}^T \gamma^\top \mathbb{Z}^\top (I - G^\top)^{-1} W^\top \mathbb{Z}$ ,  $\mathcal{S}_{24b} = (NT)^{-1} \times \sum_{t=1}^T \tilde{\mathbb{Y}}_{t-1}^\top W^\top \mathbb{Z}$ ,  $\mathcal{S}_{34a} = (NT)^{-1}\sum_{t=1}^T \gamma^\top \mathbb{Z}^\top (I - G^\top)^{-1} \mathbb{Z}$ ,  $\mathcal{S}_{34b} = (NT)^{-1} \times \sum_{t=1}^T \tilde{\mathbb{Y}}_{t-1}^\top \mathbb{Z}$ , and  $\mathcal{S}_{24c} = \beta_0(1 - \beta_1 - \beta_2)^{-1}N^{-1}\sum_{i=1}^N Z_i^\top$ . By (5.5) and (5.3) of Lemma 2, we have  $N^{-2}\text{tr}(WQW^\top) \rightarrow 0$ ,  $N^{-1}\sum_{j=0}^\infty[\text{tr}\{WG^j(G^\top)^j W^\top\}]^{1/2} \rightarrow 0$ ,  $N^{-2}\text{tr}(Q) \rightarrow 0$ , and  $N^{-1}\sum_{j=0}^\infty[\text{tr}\{G^j(G^\top)^j\}]^{1/2} \rightarrow 0$ , as  $N \rightarrow \infty$ . As a result, by the law of large numbers and Lemma 1(b) and (e), we have  $\mathcal{S}_{24c} \rightarrow_p 0$ ,  $\mathcal{S}_{24a} \rightarrow_p \kappa_8\gamma^\top \Sigma_z$ ,  $\mathcal{S}_{24b} \rightarrow_p 0$ ,  $\mathcal{S}_{34a} \rightarrow_p \kappa_3\gamma^\top \Sigma_z$  and  $\mathcal{S}_{34b} \rightarrow_p 0$ . Consequently, we have  $\mathcal{S}_{24} \rightarrow_p \kappa_8\gamma^\top \Sigma_z$  and  $\mathcal{S}_{34} \rightarrow_p \kappa_3\gamma^\top \Sigma_z$ , where  $\kappa_8$  and  $\kappa_3$  are defined in Theorem 3. This completes the proof of (A.1).

Step 2. *Proof of (A.2).* To prove (A.2), it suffices to show that  $(NT)^{1/2}\eta^\top \times \hat{\Sigma}_{xe} = (NT)^{-1/2}\sum_t \eta^\top \mathbb{X}_{t-1}^\top \mathcal{E}_t \rightarrow_d N(0, \eta^\top \Sigma \eta)$  for any  $\eta \in \mathbb{R}^{p+3}$ , where  $\sigma^2$  is set to be 1 in this step for simplicity. To this end, denote  $\xi_{Nt} = (NT_N)^{-1/2}\eta^\top \times \mathbb{X}_{t-1}^\top \mathcal{E}_t$ ,  $\mathbb{S}_{Nt} = \sum_{s=1}^t \xi_{Ns}$  and  $\mathcal{F}_{Nt} = \sigma\{\varepsilon_{is}, 1 \leq i \leq N, -\infty < s \leq t\}$ , where the number of observed time points  $T_N$  is assumed to depend on  $N$  with  $T_N \rightarrow \infty$  as  $N \rightarrow \infty$ . As a result, the double sequence  $\{\mathbb{S}_{Nt}, \mathcal{F}_{Nt}, -\infty < t \leq T_N, N \geq 1\}$  is a martingale array.

Similar to the proof of (A.1) in Step 1, we can show that  $N^{-2}E(\eta^\top \mathbb{X}_{t-1}^\top \times \mathbb{X}_{t-1}\eta)^2 < \infty$ , which implies that, for any  $\delta > 0$ , as  $N \rightarrow \infty$ ,

$$\begin{aligned} \sum_{t=1}^{T_N} E\{\xi_{Nt}^2 I(|\xi_{Nt}| > \delta) | \mathcal{F}_{N,t-1}\} &\leq \delta^{-2} \sum_{t=1}^{T_N} E(\xi_{Nt}^4 | \mathcal{F}_{N,t-1}) \\ \text{(A.4)} \qquad \qquad \qquad &\leq \frac{C\delta^{-2}}{(NT_N)^2} \sum_{t=1}^{T_N} (\eta^\top \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1}\eta)^2 \xrightarrow{p} 0, \end{aligned}$$

where  $C = 2\sigma^4 + |E(\varepsilon_{it}^4) - 3\sigma^4|$ . Moreover, by (A.1),

$$(A.5) \quad \sum_{t=1}^{T_N} E(\xi_{Nt}^2 | \mathcal{F}_{N,t-1}) = \frac{1}{NT_N} \sum_{t=1}^{T_N} \eta^\top \mathbb{X}_{t-1}^\top \mathbb{X}_{t-1} \eta \xrightarrow{p} \eta^\top \Sigma \eta$$

as  $N \rightarrow \infty$ . Therefore, by (A.4), (A.5) and the central limit theorem for martingale difference sequences [14], Corollary 3.1, we have that  $\mathbb{S}_{NT_N} = (NT)^{1/2} \eta^\top \hat{\Sigma}_{xe} \rightarrow_d N(0, \eta^\top \Sigma \eta)$ . This completes the proof.

## SUPPLEMENTARY MATERIAL

**Supplement to “Network vector autoregression”** (DOI: [10.1214/16-AOS1476SUPP](https://doi.org/10.1214/16-AOS1476SUPP); .pdf). The supplementary material [35] contains the verification of (2.6) and (2.7), proofs of Theorem 1, Theorem 4, Theorem 5, two useful lemmas and Proposition 2. The numerical verification of conditions (C2)–(C3) are also included.

## REFERENCES

- [1] ANSELIN, L. (1999). *Spatial Econometrics: Methods and Models*. Springer Science and Business Media, Dordrecht.
- [2] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. [MR2091634](#)
- [3] BEDRICK, E. J. and TSAI, C. L. (1994). Model selection for multivariate regression in small samples. *Biometrics* **50** 226–231.
- [4] BOX, G. E. P., JENKINS, G. M. and REINSEL, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd ed. Prentice Hall, Englewood Cliffs, NJ. [MR1312604](#)
- [5] BROWN, L. D., HAGERMAN, R. L., GRIFFIN, P. A. and ZMIJEWSKI, M. E. (1987). Security analyst superiority relative to univariate time-series models in forecasting quarterly earnings. *J. Account. Econ.* **9** 61–87.
- [6] CHEN, X., CHEN, Y. and XIAO, P. (2013). The impact of sampling and network topology on the estimation of social intercorrelations. *J. Mark. Res.* **50** 95–110.
- [7] CLAUSET, A., SHALIZI, C. R. and NEWMAN, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Rev.* **51** 661–703. [MR2563829](#)
- [8] CUARESMA, J. C., HLOUSKOVA, J., KOSSMEIER, S. and OBERSTEINER, M. (2004). Forecasting electricity spot-prices using linear univariate time-series models. *Appl. Energy* **77** 87–106.
- [9] DANTZIG, G. B. (1998). *Linear Programming and Extensions*, corrected ed. Princeton Univ. Press, Princeton, NJ. [MR1658673](#)
- [10] DE MOL, C., GIANNONE, D. and REICHLIN, L. (2008). Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components? *J. Econometrics* **146** 318–328. [MR2465176](#)
- [11] FAN, J. and YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York. [MR1964455](#)
- [12] FINGLETON, B. (1999). Spurious spatial regression: Some Monte Carlo results with a spatial unit root and spatial cointegration. *J. Reg. Sci.* **39** 1–19.
- [13] FORNI, M., HALLIN, M., LIPPI, M. and REICHLIN, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *J. Amer. Statist. Assoc.* **100** 830–840. [MR2201012](#)

- [14] HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York. [MR0624435](#)
- [15] HAMILTON, J. D. (1994). *Time Series Analysis*. Princeton Univ. Press, Princeton, NJ. [MR1278033](#)
- [16] HAN, F. and LIU, H. (2013). Transition matrix estimation in high dimensional time series. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* 172–180. ACM, Atlanta, GA.
- [17] HOLLAND, P. W. and LEINHARDT, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. [MR0608176](#)
- [18] HSU, N.-J., HUNG, H.-L. and CHANG, Y.-M. (2008). Subset selection for vector autoregressive processes using Lasso. *Comput. Statist. Data Anal.* **52** 3645–3657. [MR2427370](#)
- [19] LAM, C. and YAO, Q. (2012). Factor modeling for high-dimensional time series: Inference for the number of factors. *Ann. Statist.* **40** 694–726. [MR2933663](#)
- [20] LEE, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72** 1899–1925. [MR2095537](#)
- [21] LEE, L. F. and YU, J. (2009). Spatial nonstationarity and spurious regression: The case with a row-normalized spatial weights matrix. *Spatial Econ. Anal.* **4** 301–327.
- [22] LÜTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin. [MR2172368](#)
- [23] MCQUARRIE, A. D. R. and TSAI, C.-L. (1998). *Regression and Time Series Model Selection*. World Scientific, River Edge, NJ. [MR1641582](#)
- [24] NEGAHBAN, S. and WAINWRIGHT, M. J. (2011). Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Statist.* **39** 1069–1097. [MR2816348](#)
- [25] NEWBOLD, P. and GRANGER, C. W. J. (1974). Experience with forecasting univariate time series and the combination of forecasts. *J. Roy. Statist. Soc. Ser. A* **137** 131–164. [MR0451583](#)
- [26] NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *J. Amer. Statist. Assoc.* **96** 1077–1087. [MR1947255](#)
- [27] PAN, J. and YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika* **95** 365–379. [MR2521589](#)
- [28] PARK, B. U., MAMMEN, E., HÄRDLE, W. and BORAK, S. (2009). Time series modelling with semiparametric factor dynamics. *J. Amer. Statist. Assoc.* **104** 284–298. [MR2504378](#)
- [29] REINSEL, G. C. and VELU, R. P. (1998). *Multivariate Reduced-Rank Regression: Theory and Applications. Lecture Notes in Statistics* **136**. Springer, New York. [MR1719704](#)
- [30] SHUMWAY, R. H. and STOFFER, D. S. (2000). *Time Series Analysis and Its Applications*. Springer, New York. [MR1856572](#)
- [31] WANG, Y. J. and WONG, G. Y. (1987). Stochastic blockmodels for directed graphs. *J. Amer. Statist. Assoc.* **82** 8–19. [MR0883333](#)
- [32] WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis: Methods and Applications*. Cambridge Univ. Press, London.
- [33] ZHAO, Y., LEVINA, E. and ZHU, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *Ann. Statist.* **40** 2266–2292. [MR3059083](#)
- [34] ZHOU, J., TU, Y., CHEN, Y. and WANG, H. (2017). Estimating spatial autocorrelation with sampled network data. *J. Bus. Econ. Stat.* **35** 130–138. [MR3591541](#)
- [35] ZHU, X., PAN, R., LI, G., LIU, Y. and WANG, H. (2016). Supplement to “Network vector autoregression.” DOI:10.1214/16-AOS1476SUPP.

X. ZHU  
H. WANG  
GUANGHUA SCHOOL OF MANAGEMENT  
PEKING UNIVERSITY  
BEIJING, 100871  
P. R. CHINA  
E-MAIL: [xueningzhu@pku.edu.cn](mailto:xueningzhu@pku.edu.cn)  
[hansheng@gsm.pku.edu.cn](mailto:hansheng@gsm.pku.edu.cn)

G. LI  
DEPARTMENT OF STATISTICS  
AND ACTUARIAL SCIENCE  
UNIVERSITY OF HONG KONG  
HONG KONG  
P. R. CHINA  
E-MAIL: [gqli@hku.hk](mailto:gqli@hku.hk)

R. PAN  
SCHOOL OF STATISTICS  
AND MATHEMATICS  
CENTRAL UNIVERSITY OF FINANCE  
AND ECONOMICS  
BEIJING, 100081  
P. R. CHINA  
E-MAIL: [panrui\\_cufe@126.com](mailto:panrui_cufe@126.com)

Y. LIU  
SCHOOL OF MANAGEMENT  
XI'AN JIAOTONG UNIVERSITY  
XI'AN  
P. R. CHINA  
E-MAIL: [liuyuewen@mail.xjtu.edu.cn](mailto:liuyuewen@mail.xjtu.edu.cn)