Portal Nodes Screening for Large Scale Social Networks

Xuening Zhu[1], Xiangyu Chang[2], Runze Li[3] and Hansheng Wang[4]

[1]*School of Data Science, Fudan University, Shanghai, P.R. China*

[2]*Center of Data Science and Information Quality, School of Management, Xi'an Jiaotong University, Xi'an, P.R. China*

[1,3]*Department of Statistics and the Methodology Center, The Pennsylvania State University, University Park, PA 16802-2111, USA;*

[4]*Guanghua School of Management, Peking University, Beijing, P.R. China*

Abstract

Network autoregression model (NAM), as a powerful tool to study user social behaviors on large scale social networks, has drawn great attention in recent years. In this paper, we are interested in identifying the influential users (i.e., portal nodes) in a social network under the framework of NAM. Especially, we consider the autoregression model that allows to have a heterogenous and sparse network effect coefficients. Therefore, the portal nodes take influential powers which are corresponding to the nonzero network effect coefficients. A screening procedure is designed to screen out the portal nodes and the strong screening consistency is established theoretically. A quasi maximum likelihood method is applied to estimate the influential powers. The asymptotic normality of the resulting estimator is established. Further selection procedure is given by taking advantage of the local linear approximation algorithm. Extensive numerical studies are conducted by using a Sina Weibo dataset for illustration purpose.

KEY WORDS: Network Autoregression; Social Network; Portal Nodes Screening; Heterogenous Network Effects.

*The research of Zhu and Li was supported by NIDA, NIH grants P50 DA039838, National Library of Medicine, T32 LM012415, National Institute of Allergy and Infectious Diseases, U19AI089672 and National Science Foundation grant, DMS 1820702, and National Nature Science Foundation of China (NNSFC) 11690015. This work was also partially supported by National Natural Science Foundation of China grant 11690015. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NSF, the NIDA, the NIH, or the NNSFC. Chang's research was supported in part by NNSFC 11771012, 61502342 and 71472023. Wang's research was supported by NNSFC 11525101 and 71532001 and Center for Statistical Science at Peking University. The authors are very grateful to the Editor, the Associate Editor and two reviewers for their constructive comments, which leads to a significant improvement of this work.

1. INTRODUCTION

The online social media (e.g. Facebook, Twitter, Wechat) is becoming an increasingly important resource for collecting large-scale social network data. By analyzing the social network data, online social media companies can provide more convenient and punctual services, such as accurately recommending friends or production (Tang et al., 2013) and finding target communities (Girvan and Newman, 2002) for social media users, which improve the user stickiness and activities. Recently, detecting opinion leaders or celebrities, as one of the significant applications of network data, has been drawn great attention (Bodendorf and Kaiser, 2009). Because, the online celebrities are influential users who can help online social media companies to spread news, release products, and launch promotion campaigns (Aral and Walker, 2011; Stephen and Galak, 2012; Gong et al., 2016).

There is an emerging stream of literature using network data for statistical analysis (Goldenberg et al., 2010; Kolaczyk, 2009). Particularly, to quantitatively model the social influence effect, network autoregression models (NAMs) are intensively studied by the researchers (Chen et al., 2013; Zhou et al., 2017; Zhu et al., 2017, 2018b; Huang et al., 2017; Cohen-Cole et al., 2018). The models basically assume that the behaviours of the network users are closely related to their connected friends. To name a few, Chen et al. (2013) and Zhou et al. (2017) study a Twitter-type online social network, where they discover a positive network effect among the user posting behaviors. Zhu et al. (2017) propose a network vector autoregression model, which focus on the dynamic social behaviours. Subsequently, Huang et al. (2017) consider to apply the network autoregression model to networks with repeated measurements. Cohen-Cole et al. (2018) find that there are non-trivial within- and cross-choice peer effects for the students' social activities. Other than the social behaviour study, the network analysis framework is also widely applied to complex financial systems, see Hautsch et al. (2014); Zou et al. (2017); Härdle et al. (2016); Zhu et al. (2018b) for

further discussions.

Despite the great usefulness of network analysis, however, to our best knowledge, most the aforementioned network autoregression models only allow for homogeneous network effect. In practice, one could observe the phenomenon of great unbalance in social influential powers. For example, there are typically only a small amount of users, as celebrities or opinion leaders on Twitter, whose opinions can resonate with their followers significantly. The influential users, which are referred to as *portal nodes* in this article, are of critical importance for social network marketing (Hinz et al., 2011; Iyengar et al., 2011; Katona et al., 2011; Aral and Walker, 2012). By locating the portal nodes, the marketers can target the vast potential customers and launch a successful marketing campaign. Hence, how to identify the portal nodes is a significant and valuable problem.

To identify the portal nodes, the network topology information has been widely used. As a straightforward implication, the influential power of nodes can be characterized by the nodal degrees (Carrington et al., 2005; Scott, 2012). As an alternative, the centrality measurements are also widely accepted indexes to rank nodes' importance, which essentially requires a favorable connectivity with other nodes in the network (Kolaczyk, 2009; Newman, 2010). However, it is not always true that the nodes with large nodal degrees (or centralities) will have greater influential powers. As revealed by the real data example in Section 4, some social network users might be identified as portal nodes since they have a lot of followers (e.g., actor or actress), however, they might not have great influence on the followers' daily activities because they are not closely related to the users' backgrounds. As an alternative, a direct measure of nodal influences on its followers' behaviors should be considered.

Different from the above topology-based methods, we aim at a model-based approach under the framework of NAM, which allows us to quantify the nodal influential powers. We propose a portal nodes identification procedure to identify the portal n-

odes with greater influential powers. The model considers a sparse structure of the network effects in NAM, which allows for the potential heterogeneity of network nodes. Particularly, the portal nodes are specified to take nonzero network effects, while others with zero network effects are information receivers from the portal nodes. For the portal nodes identification, we take advantage of the variable screening and selection techniques (Fan and Li, 2001; Fan and Lv, 2008; Fan et al., 2014). The theoretical screening consistency is established and the selection procedure is designed by taking advantage of the local linear approximation (LLA) algorithm. Lastly, the proposed method is applied to an online social network dataset for illustration purpose.

The rest of the article is organized as follows. Section 2 introduces the model setting and portal nodes screening procedure. Section 3 discusses the post screening estimation and presents a portal nodes selection procedure. Numerical studies are given in Section 4. The article is concluded with a brief discussion in Section 5. All technical details are delegated to the Appendix in a separate supplementary material.

2. PORTAL NODES SCREENING

2.1. Model and Notations

Consider a large scale network with N nodes, which are indexed by $i = 1, \cdots, N$. To describe their following relationship, we employ an adjacency matrix $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, where $a_{ij} = 1$ if the node i follows the node j, otherwise $a_{ij} = 0$. Particularly, we do not allow self-following relations, i.e., $a_{ii} = 0$. In addition, define $W = (w_{ij}) \in \mathbb{R}^{N\times N}$ to be the row-normalized adjacency matrix, where $w_{ij} = n_i^{-1}a_{ij}$ and $n_i = \sum_j a_{ij}$, namely the out-degree of node i. Let $Y_t = (Y_{1t}, \cdots, Y_{Nt})^\top \in \mathbb{R}^N$ be the continuous response (e.g., tweet length) collected at time point t for $1 \le t \le T$. In addition, assume for each node, a p-dimensional covariate is collected as $Z_{it} = (Z_{i1t}, \cdots, Z_{ipt})^\top \in$

$\mathbb{R}^p$. We consider the following network autoregression model (NAM),

$$Y_{it} = \sum_{j=1}^{N} d_j w_{ij} Y_{jt} + Z_{it}^\top \gamma + \varepsilon_{it}, \tag{2.1}$$

where $\gamma = (\gamma_1, \cdots, \gamma_p)^\top \in \mathbb{R}^p$ is the nodal coefficient, and ε_{it} is the noise term. Note the parameter d_j associated with node j reflects the social influence of node j on other nodes. Therefore, we refer $d = (d_1, d_2, \cdots, d_N)^\top$ as *influential powers* of nodes.

The model is motivated from the famous spatio-temporal model in literature (Yu et al., 2008; Dou et al., 2016), where the applications are mostly in spatial data. Recently, there are researches extending the spatial data analysis to the network data context and establish the NAM models (Zhou et al., 2017; Liu et al., 2017; Huang et al., 2017; Zhu et al., 2018b,a), where they assume the same network influence parameters for all network nodes. However, in practice, one could observe the phenomenon that there are typically a small amount of users, as opinion leaders on social networks, whose opinions can resonate with their followers significantly. This observation motivates us to assume a sparse structure of d.

By sparsity, we mean only a limited number of d_js are nonzero, while for the rest $d_j = 0$. The nodes corresponding to the nonzero social influential powers are collected as $\mathcal{M} = \{j : d_j \neq 0, 1 \leq j \leq N\}$, which are then referred to as *portal nodes*. By this specification, we could conclude immediately from (2.1) that the portal nodes could have influence on each other, while the other nodes could only be influenced by the portal nodes. Namely, the word "portal" is used to illustrate the role of the influential nodes as information portals. Consequently, the information is diffused to the whole network mainly through the portal nodes.

Let $D = \mathrm{diag}(d_1, \cdots, d_N) \in \mathbb{R}^{N \times N}$ and $Z_t = (Z_{1t}, \cdots, Z_{Nt})^\top \in \mathbb{R}^{N \times p}$. One could rewrite the model (2.1) as

$$Y_t = WDY_t + Z_t\gamma + \mathcal{E}_t, \tag{2.2}$$

where $\mathcal{E}_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Nt})^\top \in \mathbb{R}^N$ is the noise vector. The response Y_t can be represented by $Y_t = (I - WD)^{-1}(Z_t\gamma + \mathcal{E}_t)$. We assume that $(I - WD)$ could be inverted to make the representation valid. Note that here we allow D to have heterogeneous diagonal elements, which is consequently more generalized than the autoregressive models (Lee, 2004; Lee and Yu, 2009; Zhu et al., 2018b). In addition, instead of dealing with the cross-sectional data, here we require that one should have replications of Y_t along the time for $t = 1, \cdots, T$. As one could see in the theoretical development, the time replications of Y_t essentially facilitate the theoretical analysis and recovering of D.

Remark 1. The model (2.2) is flexible to extend to the case with dynamic dependence. Specifically, the lag terms of Y_t (i.e., Y_{t-k} with $k \geq 1$) can be added and included in Z_t. Consequently, the screening technique discussed in this article can also be employed. For the sake of simplicity, we keep the parsimonious model form (2.2).

Remark 2. The model form (2.2) is similar to the spatio-temporal model discussed by Dou et al. (2016). They consider the following type of modelling,

$$Y_t = D_1 W Y_t + D_2 W Y_{t-1} + \mathcal{E}_t, \tag{2.3}$$

where $D_k = \text{diag}(d_{k1}, \cdots, d_{kN}) \in \mathbb{R}^{N \times N}$ $(k = 1, 2)$ are diagonal matrices. In their setting, d_{ki} is interpreted as how much the node i is influenced by the other nodes. Although they are capable of capturing the heterogeneous influences of the connected friends on the focal node, the framework could not directly quantify the influential powers of the nodes. Our model, on the other side, allows us to identify the portal nodes and directly estimate out their influential powers. One should also note that by rewriting (2.3) as $Y_t = WD_1Y_t + WD_2Y_{t-1} + \mathcal{E}_t$, the model could then be transferred into our framework. The screening and model selection technique proposed in this work could be readily applied.

2.2. Portal Nodes Screening

Let $W_{\cdot i}$ be the ith column of W and $X_t = W\mathrm{diag}(Y_t) = (W_{\cdot 1}Y_{1t}, W_{\cdot 2}Y_{2t}, \cdots, W_{\cdot N}Y_{Nt}) \in \mathbb{R}^{N\times N}$. Then the model (2.2) could be rewritten as $Y_t = X_t d + Z_t\gamma + \mathcal{E}_t$. In addition, define $\mathbb{X} = (X_1^\top, \cdots, X_T^\top)^\top \in \mathbb{R}^{(NT)\times N}$, $\mathbb{Y} = (Y_1^\top, \cdots, Y_T^\top)^\top \in \mathbb{R}^{NT}$, $\mathbb{Z} = (Z_1^\top, \cdots, Z_T^\top)^\top = (\mathbb{Z}_1, \cdots, \mathbb{Z}_p) \in \mathbb{R}^{(NT)\times p}$, and $\mathcal{E} = (\mathcal{E}_1^\top, \mathcal{E}_2^\top, \cdots, \mathcal{E}_T^\top) \in \mathbb{R}^{NT}$. The model (2.2) could spell as

$$\mathbb{Y} = \mathbb{X}d + \mathbb{Z}\gamma + \mathcal{E}. \tag{2.4}$$

Consequently, each node in the network can be treated as a covariate and the NAM can be transformed to a linear regression form. However, it should be noted that screening on d might depend on the unknown parameter γ. To take this into consideration, we project each column of $\mathbb{X} = (\mathbb{X}_1, \cdots, \mathbb{X}_N)$ on the column space of $\widetilde{\mathbb{Y}} = (\mathbb{Y}, \mathbb{Z})$. That leads to a linear regression with $\mathbb{X}_j$ being response and $\widetilde{\mathbb{Y}}$ being predictors. Consequently, a regression R^2 with respect to each $\mathbb{X}_j$ can be approximated as

$$\widehat{R}_j^2 = \frac{\mathbb{X}_j^\top \left\{\widetilde{\mathbb{Y}}(\widetilde{\mathbb{Y}}^\top\widetilde{\mathbb{Y}})^{-1}\widetilde{\mathbb{Y}}^\top\right\}\mathbb{X}_j}{\mathbb{X}_j^\top\mathbb{X}_j}. \tag{2.5}$$

For a given proper thresholding value c, we could estimate the portal nodes set by

$$\widehat{\mathcal{M}} = \left\{1 \le j \le N : \widehat{R}_j^2 > c\right\}. \tag{2.6}$$

Remark 3. To get more insights of the screening measure $\widehat{R}_j^2$, we consider the special case without exogenous variables $\mathbb{Z}$. Therefore $\widetilde{\mathbb{Y}} = \mathbb{Y}$ and we could rewrite $\widehat{R}_j^2$ as

$$\widehat{R}_j^2 = \frac{(\mathbb{Y}^\top\mathbb{X}_j)^2}{\|\mathbb{Y}\|^2\|\mathbb{X}_j\|^2} = \frac{\{\sum_{t=1}^T (W_{\cdot j}^\top Y_t)Y_{jt}\}^2}{(\sum_{t=1}^T Y_t^\top Y_t)(\sum_{t=1}^T W_{\cdot j}^\top W_{\cdot j}Y_{jt}^2)},$$

which is exactly the square of sample correlation between $(Y_{jt} : 1 \le t \le T)^\top$ and

$(W_{\cdot j}^\top Y_t : 1 \le t \le T)^\top$. In other words, it measures the dependency of the activeness for the focal node j with respect to its followers' activeness. Consequently, the correlation based measure $\widehat{R}_j^2$ could filter out the nodes, who have weak correlation with their followers.

Remark 4. Note that the proposed portal nodes screening measure (2.5) is in spirit similar to the conditional sure independence screening (CSIS) method (Barut et al., 2016). Namely, our approach is to project $\mathbb{X}_j$ on the column space of $\widetilde{\mathbb{Y}} = (\mathbb{Y}, \mathbb{Z})$, while Barut et al. (2016) utilizes the projection of $\mathbb{Y}$ on the column space of $(\mathbb{X}_j, \mathbb{Z})$. By further assuming $\mathbb{X}_j$ is standardized, the two approaches could be unified by ranking the following two measures,

$$\begin{aligned}
\mathrm{SSE}_j^{(1)} &= \min_{\alpha_j,\gamma} E\|\mathbb{Y} - \alpha_j \mathbb{X}_j - \mathbb{Z}\gamma\|^2 = E\|\mathbb{Y} - \alpha_j^{(1)}\mathbb{X}_j - \mathbb{Z}\gamma^{(1)}\|^2, \\
\mathrm{SSE}_j^{(2)} &= \min_{\alpha_j,\gamma} E\|\mathbb{X}_j - \alpha_j \mathbb{Y} - \mathbb{Z}\gamma\|^2 = E\|\mathbb{X}_j - \alpha_j^{(2)}\mathbb{Y} - \mathbb{Z}\gamma^{(2)}\|^2
\end{aligned}$$

where the CSIS method ranks the nodes from low to high using $\mathrm{SSE}_j^{(1)}$ and our approach using $\mathrm{SSE}_j^{(2)}$. The following relationship holds for $\mathrm{SSE}_j^{(1)}$ and $\mathrm{SSE}_j^{(2)}$,

$$\begin{aligned}
\mathrm{SSE}_j^{(1)} &= \frac{1}{\alpha_j^{(1)2}} E\|\mathbb{X}_j + \mathbb{Z}\gamma^{(2)}/\alpha_j^{(1)} - \mathbb{Y}/\alpha_j^{(1)}\|^2 \ge \mathrm{SSE}_j^{(2)}/\alpha_j^{(1)2}, \\
\mathrm{SSE}_j^{(2)} &= \frac{1}{\alpha_j^{(2)2}} E\|\mathbb{Y} + \mathbb{Z}\gamma^{(2)}/\alpha_j^{(2)} - \mathbb{X}_j/\alpha_j^{(2)}\|^2 \ge \mathrm{SSE}_j^{(1)}/\alpha_j^{(2)2}.
\end{aligned}$$

This implies that $\mathrm{SSE}_j^{(1)} \cong \mathrm{SSE}_j^{(2)}$ in the sense that $\mathrm{SSE}_j^{(1)} \to 0$ implies $\mathrm{SSE}_j^{(2)} \to 0$ and vice versa. This indicates when the signal is strong enough, both measures are able to detect it. As a result, their performances are fairly comparable.

It is noteworthy that although the model (2.4) can be written in a linear regression form, it cannot be directly estimated by a ordinary least squares estimation. This is because the response information $\mathbb{Y}$ is included in both sides of (2.4). As an alternative, the screening measure (2.5) allows us to first marginally calculate the dependence of

$\mathbb{X}_j$ with respect to $\widetilde{\mathbb{Y}} = (\mathbb{Y}, \mathbb{Z})$, and then conduct the estimation.

2.3. Screening Consistency Property

Let $\Sigma_Y = \text{cov}(Y_t) = (\sigma_{Y,ij}) \in \mathbb{R}^{N\times N}$. Assume Z_{it} is independently distributed for $1 \le i \le N$ and $1 \le t \le T$ with $\text{cov}(Z_{it}) = \Sigma_Z \in \mathbb{R}^{p\times p}$. In addition, let $\mathcal{E}_t$ be independent over t with $\text{cov}(\mathcal{E}_t) = \sigma_e^2 I_N$. Define $S = I - WD$, it can be easily verified $\Sigma_Y = c_{\gamma e} S^{-1}(S^{-1})^\top$, where $c_{\gamma e} = \gamma^\top \Sigma_Z \gamma + \sigma_e^2$. Before we go into details about the theoretical properties of the screening procedure, we first clarify the notations. First define

$$R_j^2 = \frac{1}{\kappa_{1j}\sigma_{Y,jj}}\Big(\frac{\kappa_{2j}^2}{c_y} + \frac{c_z\kappa_{3j}^2}{N} + \frac{\kappa_{2j}^2 c_s^2 c_z}{c_y^2 N} - \frac{2\kappa_{2j}\kappa_{3j}c_s c_z}{c_y N}\Big), \tag{2.7}$$

where

$$\kappa_{1j} = W_{\cdot j}^\top W_{\cdot j},\ \kappa_{2j} = e_j^\top \Sigma_Y W_{\cdot j},\ \kappa_{3j} = e_j^\top S^{-1} W_{\cdot j},$$
$$c_z = \gamma^\top \Sigma_Z \widetilde{\Sigma}_{zy}^{-1} \Sigma_Z \gamma,\ c_y = \text{tr}(\Sigma_Y),\quad c_s = \text{tr}(S^{-1}),$$

$\widetilde{\Sigma}_{zy} = \Sigma_Z - N^{-1}c_y^{-1}c_s^2 \Sigma_Z \gamma\gamma^\top \Sigma_Z$ and $e_j \in \mathbb{R}^N$ is a vector with the jth element being 1 and others being 0.

Although R_j^2 defined in (2.7) takes a complex form, but as it will be shown in Proposition 1, $\widehat{R}_j^2$ performs as a good approximation for R_j^2. Namely, $\max_j |\widehat{R}_j^2 - R_j^2| \to_p 0$ as $N \to \infty$. Therefore, intuitively, R_j^2 could be comprehended as the population version of $\widehat{R}_j^2$. Technically, the following conditions are required.

(C1) (Sub-Gaussian Distribution) The random errors ε_{it} $(1 \le i \le N, 1 \le t \le T)$ are *i.i.d* sub-Gaussian random variables with mean zero and scale factor $0 < \sigma_e < \infty$, i.e., $E\{\exp(\delta\varepsilon_{it})\} \le \exp(\sigma_e^2\delta^2/2)$ for any $\delta \in \mathbb{R}$. Similarly, let the nodal covariates Z_{it} $(1 \le i \le N, 1 \le t \le T)$ be *i.i.d* sub-Gaussian random vectors.

(C2) (Minimum Signal) Let $c_{\min} \ge c$ as $N \to \infty$ and $T \to \infty$, where $c_{\min} =$

$\min_{j\in\mathcal{M}} R_j^2$ and $c = O(N^{\zeta-1})$ defined in (2.6) for $0 < \zeta \leq 1$.

(C3) (REGULARITY) Let $0 < \tau_{\min} \leq \lambda_{\min}(\Sigma_Y) \leq \lambda_{\max}(\Sigma_Y) \leq \tau_{\max}$, where and $\tau_{\min}$ is a finite constant, $\tau_{\max} = O(N^{\tau})$ with $\tau < \min\{\zeta, 1/2\}$ and ζ is defined in condition (C2).

Condition (C1) is imposed on the distribution of the noise term. It should be noted that the sub-Gaussian assumption is relaxed than the normality assumption (Wang et al., 2013). The assumption can be further relaxed to allow weak dependence among the nodal covariates. Next, Condition (C2) restricts that the signal strength R_j^2 of the portal nodes should not be too weak to be detected. Lastly, Condition (C3) is assumed on the covariance matrix Σ_Y, which is a uniformity condition (Zhu et al., 2018b) imposed on the network nodes. Given all the technical conditions, we give the following proposition.

Proposition 1. *Under Conditions (C1)–(C3), further assume* $T = O\{(N^{2-2\zeta}\log N)^{\xi}\}$ *for* $\xi > 1$*, then we have* $P(\max_j |\widehat{R}_j^2 - R_j^2| > \delta_1) \to 0$ *as* $N \to \infty$*, where* $\delta_1 = O(N^{\zeta-1})$.

The proof of Proposition 1 is given in Appendix A.2 in the supplementary material. By Proposition 1, we have $\max_j |\widehat{R}_j^2 - R_j^2| \to_p 0$ as $N \to \infty$. Particularly, it requires that the total time periods should diverge along with $N \to \infty$. Subsequently, we show the screening consistency property of $\widehat{R}_j^2$.

Theorem 1. *Assume* $T = O\{(N^{2-2\zeta}\log N)^{\xi}\}$ *for* $\xi > 1$*. Under conditions (C1)–(C3), there exists constant* c *and* $m_{\max} = O(N^{1+\tau-\zeta})$*, such that,*

$$P\big(\mathcal{M} \subset \widehat{\mathcal{M}}\big) \to 1, \tag{2.8}$$

$$P\big(|\widehat{\mathcal{M}}| \leq m_{\max}\big) \to 1. \tag{2.9}$$

as $N \to \infty$.

The proof of Theorem 1 is given in Appendix A.3 in the supplementary material. Note that to ensure the screening consistency property, we require the diverging rate of T should be slightly faster than $N^{2-2\zeta}\log N$. In addition, by (2.8) and (2.9), we are implicitly requiring the true model size $|\mathcal{M}| \leq |\widehat{\mathcal{M}}| \leq m_{\max}$. In addition, with respect to the upper model size $m_{\max}$, we could have $m_{\max} = O(1)$ if we set $\zeta = 1$ and $\tau = 0$, which implies a stronger signal in (C2) and a tighter regularity condition in (C3)

Subsequently, we focus on the signal R_j^2 in condition (C2). In the following proposition, we connect the assumption to the network structure conditions under the scenario $\min_{i\in\mathcal{M}} d_i > 0$. This leads to more insights about the topology features of portal nodes.

Proposition 2. *Assume* $d_{\min} \overset{\text{def}}{=} \min_{i\in\mathcal{M}} d_i > 0$. *In addition, assume there exists constants* c_1, c_2, c_3, c_4 *such that*

$$c_1 N^{\zeta} \leq \min_{j_1,j_2\in\mathcal{M}} \left(W_{\cdot j_1}^{\top} W_{\cdot j_2}\right) \leq \max_{j_1,j_2\in\mathcal{M}} \left(W_{\cdot j_1}^{\top} W_{\cdot j_2}\right) \leq c_2 N^{\zeta}, \tag{2.10}$$

$$c_3 N^{-1} tr(\Sigma_Y) \leq \min_{j\in\mathcal{M}} \sigma_{Y,jj} \leq \max_{j\in\mathcal{M}} \sigma_{Y,jj} \leq c_4 N^{-1} tr(\Sigma_Y) \tag{2.11}$$

for some $0 < \zeta \leq 1$. *In addition, assume*

$$\min_{j_1\in\mathcal{M}} \left\{ \max_{j_2\in\mathcal{M}} (W_{j_1 j_2}) \right\} \geq c_w, \tag{2.12}$$

where $c_w > 0$ *is a positive constant. Then condition (C2) holds.*

The proof of Proposition 2 is given in Appendix A.4 in the supplementary material. We next illustrate the conditions (2.10)–(2.12) with more insights. First, it can be noted $W_{\cdot j_1}^{\top} W_{\cdot j_2} = (n_{j_1} n_{j_2})^{-1} \sum_{i=1}^{N} a_{ij_1} a_{ij_2}$ measures the amount of weighted common followers of the nodes j_1 and j_2. Consequently, Condition (2.10) essentially requires the order of weighted common followers for arbitrary two portal nodes should be in the order of N^{ζ}. Next, Condition (2.11) states a regularity condition that the rate of $\sigma_{Y,jj}$ for the portal nodes should be both upper and lower bounded by the average level, i.e.,

$N^{-1}\mathrm{tr}(\Sigma_Y)$. Lastly, Condition (2.12) states that certain extent of connectivity should exist among the portal nodes. Specifically, it requires that for any portal node j_1, at least it should follow one another portal node j_2.

3. POST SCREENING ESTIMATION

3.1. Quasi Log-likelihood Estimation

In this section, we discuss the estimation procedure after portal nodes screening. We treat the post screening set to be $\mathcal{M}$ to save notations and assume $|\mathcal{M}| = m$. Specifically, let $d_{\mathcal{M}} \in \mathbb{R}^m$ collect the coefficients d_j in d for $j \in \mathcal{M}$. In addition, we restrict $d_j = 0$ for $j \neq \mathcal{M}$. Define the parameter of interest as $\theta_{\mathcal{M}} = (d_{\mathcal{M}}^\top, \gamma^\top)^\top \in \mathbb{R}^{m+p}$. We could write down the quasi log-likelihood as

$$\ell(\theta_{\mathcal{M}}) = T \log |S| - (NT/2) \log \sigma_e^2 - \frac{1}{2\sigma_e^2} \Big\{ \sum_{t=1}^{T} (SY_t - Z_t\gamma)^\top (SY_t - Z_t\gamma) \Big\} \tag{3.1}$$

with some constants ignored. The quasi maximum likelihood estimator (QMLE) can be obtained as $\widehat{\theta}_{\mathcal{M}} = \arg\max_{\theta_{\mathcal{M}}} \ell(\theta_{\mathcal{M}})$. In the following we establish the asymptotic properties of $\widehat{\theta}_{\mathcal{M}}$ when $\mathcal{M}$ covers the true model. We first state the conditions as follows.

(C4) (Noise Term) Assume $E(\varepsilon_{it}^3) = 0$ for all $1 \leq i \leq N$ and $1 \leq t \leq T$.

(C5) (Regularity) Denote $\mathbb{W} = W^\top W/N$. In addition, let $\mathbb{W}_{\mathcal{M}} = (\mathbb{W}_{j_1 j_2} : j_1 \in \mathcal{M}, j_2 \in \mathcal{M}) \in \mathbb{R}^{m \times m}$ and $\Sigma_{Y,\mathcal{M}} = (\Sigma_{Y,j_1 j_2} : j_1 \in \mathcal{M}, j_2 \in \mathcal{M}) \in \mathbb{R}^{m \times m}$. For an arbitrary matrix $M = (m_{ij}) \in \mathbb{R}^{m_1 \times m_2}$, denote $|M|_e = (|m_{ij}|) \in \mathbb{R}^{m_1 \times m_2}$. Assume as $N \to \infty$ there exists positive constants $0 < \tau_1 < \tau_2$ that

$$\tau_1 \leq \min\{\lambda_{\min}(\mathbb{W}_{\mathcal{M}}), \lambda_{\min}(\Sigma_{Y,\mathcal{M}})\} \leq \max\{\lambda_{\max}(\mathbb{W}_{\mathcal{M}}), \lambda_{\max}(|\Sigma_{Y,\mathcal{M}}|_e)\} \leq \tau_2. \tag{3.2}$$

Condition (C4) gives moment conditions on the noise terms. Technically, this condition is assumed to facilitate the theoretical discussion and could be further relaxed. Condition (C5) is a regularity condition on the network and covariance structure related to the nodes in $\mathcal{M}$. Particularly, by requiring $\lambda_{\min}(\mathbb{W}_{\mathcal{M}}) \geq \tau_1$, we have that $N^{-2}W_{\cdot j}^{\top}W_{\cdot j}$ is lower bounded by a constant as $N \to \infty$ for $j \in \mathcal{M}$. Then we establish the following theorem.

Theorem 2. *Assume conditions (C1), (C4), (C5). Furthermore, assume* $m = o(T^{\delta_1})$ *with* $0 \leq \delta_1 < 1/4$. *Then we have (a)* $\|\widehat{\theta}_{\mathcal{M}} - \theta_{\mathcal{M}}\| = O_p(\sqrt{(NT)^{-1}m})$; *and (b)* $\sqrt{NT}(\widehat{d}_j - d_j) \to_d N(0, \sigma_j^2)$ *and* $\sqrt{NT}(\widehat{\gamma} - \gamma) \to_d N(0, \sigma_e^2\Sigma_Z^{-1})$, *where* σ_j^2 *is the* j*th diagonal element of* $\Sigma_2^{-1}\Sigma_1\Sigma_2^{-1}$. *The forms of* Σ_1 *and* Σ_2 *are given in Appendix A.5.*

The proof of Theorem 2 is given in Appendix A.6 in the supplementary material. The Theorem 2 establishes the consistency and asymptotic normality result for the QMLE under the scenario that $\mathcal{M}$ covers the true model. In practice, the portal screening procedure tends to over select nodes and include many false positives. To control the false positive rate, a popular approach is to conduct penalization after the screening procedure. We introduce the procedure and give a numerical algorithm for implementation in the next section.

3.2. Penalized Quasi Log-likelihood Estimation

In this section, we provide a penalized quasi log-likelihood estimation procedure after portal set screening to precisely identify the true portal nodes. Given a regularization parameter λ, the following penalized quasi log-likelihood is considered

$$Q(\theta_{\mathcal{M}}) = -\ell\big(\theta_{\mathcal{M}}\big) + \sum_{j\in\mathcal{M}} p_\lambda\big(|d_j|\big), \tag{3.3}$$

where $p_\lambda(|\cdot|)$ is the penalty function related to a regularization parameter λ. The widely used penalty forms include: Lasso (Tibshirani, 1996) with L_1 penalty $p_\lambda(|\theta|) = \lambda|\theta|$,

bridge regression (Fu, 1998) with L_q penalty $p_\lambda(|\theta|) = \lambda|\theta|^q, q \geq 1$, hard thresholding penalty $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda^2)I(|\theta| < \lambda)$, and many others. As pointed by Fan and Li (2001), three mathematical properties should be examined for a penalty function, which are *unbiasedness, sparsity, continuity*. They conclude that the L_q and hard thresholding penalties cannot satisfy the three mathematical conditions simultaneously. As an alternative, they proposed the smoothly clipped absolute deviation (SCAD) penalty, which could lead to a consistent estimator and enjoy the oracle property. Subsequently, the minimax concave penalty (MCP) is proposed by Zhang et al. (2010), which equally has the desirable oracle property.

To optimize the non-convex objective functions in SCAD and MCP, several algorithms are designed, for example, the local quadratic/linear approximation (LQA/LLA) algorithm (Fan and Li, 2001; Zou and Li, 2008), PLUS algorithm, coordinate descent algorithm (Breheny and Huang, 2011), and many others. Further to deal with the potential issue of multiple local minimizers, both Wang et al. (2013) and Fan et al. (2014) have proposed possible solutions to revise the LLA algorithm. Theoretically, Fan et al. (2014) provides a guarantee for the LLA algorithm to obtain the oracle estimator in the folded concave penalized problem, when it is initialized by an appropriate initial estimator. Define $\ell^{(j)}(x) = \ell(x, \theta_{\mathcal{M}}^{(-j)})$ to be a function of $\ell(\theta)$ at $d_j = x$ given the other parameters $\theta_{\mathcal{M}}^{(-j)}$ fixed. Accordingly, let $\dot{\ell}^{(j)}(\cdot)$ and $\ddot{\ell}^{(j)}(\cdot)$ denote the first and second derivative function of $\ell^{(j)}(\cdot)$. In this work, we use the framework of the LLA algorithm to solve (3.3) and state the algorithm in Algorithm 1. The detailed development of the algorithm is given in Appendix B.1 in the supplementary material.

Given a solution path, a critical problem is to select the tuning parameter λ. Here we follow Wang et al. (2013) to employ the HBIC criterion, which spells as

$$\text{HBIC}(\lambda) = \ell(\widehat{\theta}_{\mathcal{M}}) + |\mathcal{M}_\lambda| \frac{C_{n^*} \log(q)}{n^*}, \tag{3.6}$$

Algorithm 1 Local Linear Approximation (LLA) Algorithm

1. Solve the following Lasso problem to obtain the initial estimator $\widehat{\theta}_{\mathcal{M}}^{(0)}$ as

$$\widehat{\theta}_{\mathcal{M}}^{(0)} = \arg\min_{\theta_{\mathcal{M}}} \Big\{ -\ell(\theta_{\mathcal{M}}) + \sum_{j\in\mathcal{M}} \lambda^{(0)}\delta_j^{(0)}|d_j| \Big\}, \tag{3.4}$$

where $\lambda^{(0)} = \lambda\eta$ with $\eta = 1/\log(NT)$ is the initial regularization parameter, and $\delta_j^{(0)} = \ddot{\ell}^{(j)}(\hat{d}_j^{(0)})$.

2. For $m = 0, 1, 2, \cdots$, repeat the LLA iteration till convergence

(2.a) Update the adaptive weights as $w_j^{(m)} = p'_\lambda(|\hat{d}_j^{(m-1)}|)$ for $j \in \mathcal{M}$.

(2.b) Obtain $\widehat{\theta}_{\mathcal{M}}^{(m+1)}$ by solving the following optimization problem

$$\widehat{\theta}_{\mathcal{M}}^{(m+1)} = \arg\min_{\theta_{\mathcal{M}}} \Big\{ -\ell(\theta_{\mathcal{M}}) + \sum_{j\in\mathcal{M}} w_j^{(m)}\delta_j^{(m)}|d_j| \Big\}. \tag{3.5}$$

The series $\delta_j^{(0)}, \delta_j^{(1)}, \cdots$ are scaling parameters given by $\delta_j^{(m)} = \ddot{\ell}^{(j)}(\hat{d}_j^{(m)})$.

where $\ell(\widehat{\theta}_{\mathcal{M}})$ is the log-likelihood defined in (3.1), $q = |\mathcal{M}|$ is the number of nodes after screening, $\mathcal{M}_\lambda = \{j : \hat{d}_j(\lambda) \neq 0\}$ is the selected set, n^* is the effective sample size, and $C_{n^*} = \log\{\log(n^*)\}$ is slowly diverging with n^*. To obtain the effective sample size n^*, we focus on the second order derivative matrix of $\ell(\theta_{\mathcal{M}})$ with respect to $d_{\mathcal{M}}$, i.e., $H(\theta_{\mathcal{M}}) = \partial^2\ell(\theta_{\mathcal{M}})/\partial d_{\mathcal{M}}\partial d_{\mathcal{M}}^\top$. One could verify that the $\partial^2\ell(\widehat{\theta}_{\mathcal{M}})/\partial d_j^2$ is in the order of $T(W_{\cdot j}^\top W_{\cdot j})$ for $j \in \mathcal{M}$ in the diagonal of $H(\theta_{\mathcal{M}})$. Therefore, we take $n^* = T\max_{j\in\mathcal{M}_\lambda}(W_{\cdot j}^\top W_{\cdot j})$ to be the effective sample size in (3.6).

Remark 5. If we let $\mathcal{M}$ to be the full nodes set, the LLA algorithm could be applied to identify the portal nodes set directly. Although it is feasible, it might be computationally inefficient and instable numerically (Fan et al., 2009). This is because optimizing (3.3) is computationally expensive at each iteration since the determinant of a high dimensional matrix is involved. Moreover, it could take more iterations to optimize many parameters at the same time and hard to converge. As an alternative, conducting the screening method (2.5)–(2.6) in the first step is efficient and could screen out a great portion of unimportant nodes.

4. NUMERICAL STUDIES

4.1. Simulation Models

To verify the theoretical properties, we present here three simulation examples. The main difference lies in the generating mechanism of network structures for non-portal nodes. For the portal nodes, we consider here two typical types. The first type, which is also the most common type, is the portal nodes with large number of followers, while the second type only have limited number of followers. We first specify the first $s = 10$ nodes as portal nodes. The corresponding coefficients are set to be $(0.3+0.05i : 0 \leq i \leq 9)^\top \in \mathbb{R}^{10}$. Specifically, we set the 5th and 6th portal nodes to be the second type, and the others to be the first type with nodal in-degrees $d_i = N^\delta$. The number of followers for the second type is set to be median number of the non-portal nodes' in-degrees. In addition, we further randomly select 10 non-portal nodes with nodal in-degrees $d_i = N^\delta$, who have large number of followers but with zero influential powers. Two cases with $\delta = 1/2$ and $\delta = 1/4$ are considered respectively.

Subsequently, for each node i at time t, we sample the covariates Z_{it} as follows. First, sample a multivariate normal random variable $\widetilde{Z}_{it} = (\widetilde{Z}_{it,1}, \cdots, \widetilde{Z}_{it,5})^\top \in \mathbb{R}^5$ independently with $E(\widetilde{Z}_{it}) = \mathbf{0}$ and $\mathrm{cov}(\widetilde{Z}_{it}) = \Sigma_z = (\sigma_{j_1 j_2})$, where $\sigma_{j_1 j_2} = 0.5^{|j_1 - j_2|}$. Then, the covariates are constructed as $Z_{it} = (1, \widetilde{Z}_{it}^\top, Y_{i(t-1)}, Y_{i(t-2)})^\top \in \mathbb{R}^8$, where the last two elements are lag-2 autoregression terms. Accordingly, the coefficients are set to be $(1, \widetilde{\gamma}^\top, 0.5, 0.3)^\top$ respectively, where $\widetilde{\gamma} = (1, 2, 3, 4, 5)^\top \in \mathbb{R}^5$. Lastly, the noise term ε_{it} is generated from $N(0, 1)$ independently and the responses Y_{it} are generated by model (2.1) accordingly. We then introduce the following three examples of different network structures among the remaining non-portal nodes.

Example 1. (Dyad Independence Network) A dyad is defined as $\mathbb{A}_{ij} = (a_{ij}, a_{ji})$ where $1 \leq i < j \leq N$, a_{ij} and a_{ji} are the entries of A. Holland and Leinhardt (1981) proposed a generative model for networks based on the independent assumption of

dyads. Following the model, we set $P(\mathbb{A}_{ij} = (1,1)) = 4N^{-1}$ to reflect network sparsity. Therefore, the expected number of the mutually connected dyads (i.e., $\mathbb{A}_{ij} = (1,1)$) is $O(N)$. Next, we allow the expected degree of each node to be slowly diverging in the order of $O(N^{0.2})$ by setting $P(\mathbb{A}_{ij} = (1,0)) = P(D_{ij} = (0,1)) = 0.5N^{-0.8}$. As a result, the probability of forming a null dyad should be $P(\mathbb{A}_{i,j} = (0,0)) = 1 - 4N^{-1} - N^{-0.8}$, which is close to 1 as the network size N is large.

Example 2. (Stochastic Block Network) One of the most essential structures of networks is the community structure. Typically, it is assumed that communities are tightly knit groups with denser connections within the communities, and relatively sparser connections between the communities (Newman and Girvan, 2004). To model the community structure, stochastic block model is proposed by Wang and Wong (1987). In this example, we follow Nowicki and Snijders (2001) to randomly assign a block label for each node ($k = 1, \cdots, K$), where $K \in \{5, 10\}$ is the total number of blocks. Then, let $P(a_{ij} = 1) = 0.9N^{-1}$ if i and j belong to the same block, and $P(a_{ij} = 1) = 0.3N^{-1}$ otherwise. As a result, nodes within the same community are more likely to be connected, when compared with nodes from different communities.

Example 3. (Power-Law Distribution Network) In a social network, it is commonly observed that there exists a small portion of nodes have a large amount of followers, which are usually referred to as "hubs" (Barabási and Albert, 1999). This phenomenon leads to the power-law distribution of network degrees (Clauset et al., 2009). To mimic this phenomenon, we simulate the adjacency matrix A according to Clauset et al. (2009) as follows. First, we generate the in-degree $n_i = \sum_j a_{ji}$ for node i by the discrete power-law distribution, i.e., $P(n_i = k) = ck^{-\alpha}$ with a normalizing constant c and exponent parameter $\alpha = 2.5$. Here larger α indicates heavier tail of the distribution. Then, for the ith node, we randomly select n_i nodes to be its followers with $a_{ji} = 1$. Note using the Power-Law distribution network for the non-portal nodes, we could generate some nodes with large in-degrees but with zero influential powers.

We then evaluate the screening and selection accuracy as well as the estimation properties. For each example, we consider (a) MEDIAN NETWORK with $(N, T) = (100, 50)$, $(200, 100)$, and (b) LARGE NETWORK with $(N, T) = (2000, 200)$, $(5000, 500)$ respectively. The experiment is repeated for 100 times. First, we compare the proposed screening measure with the topology-based screening method (i.e., ranking nodes by in-degrees). Next, with regards to portal nodes selection, we implement the following algorithms to obtain the resulting estimator. They are, the Lasso estimator (Tibshirani, 1996), the adaptive Lasso (ALasso) estimator (Zou, 2006; Bühlmann and Van De Geer, 2011), the SCAD (with $a = 3.7$) and MCP (with $a = 1.5$) estimators implemented by Algorithm 1 respectively. To choose the tuning parameter, the HBIC criterion (3.6) is calculated and applied. For the LARGE NETWORK, we only implement the screening method for evaluating the performances. Lastly, the experiment for the QMLE estimation and inference are given in Appendix B.2.

For each replication, $n = [N/\log(N)]$ nodes with highest $\widehat{R}_j^2$ are selected as portal nodes, which is denoted as $\widehat{\mathcal{M}}$. We first evaluate the coverage properties for the screening procedure. The coverage number for the true portal nodes set in the screening step is computed as $\text{TP}_s = \sum_{r=1}^{100} |\widehat{\mathcal{M}}^{(r)} \cap \mathcal{M}|/100$, where $\widehat{\mathcal{M}}^{(r)}$ is the screened set in the rth replication and $\mathcal{M} = \{j : d_j > 0\}$ is the true portal nodes set. To compare the screening efficiency, we rank the nodes by the screening measurements from high to low, then we calculate the largest rank of all the portal nodes, i.e., $\mathfrak{R}_r = \max_{i\in\mathcal{M}} \sum_j I(\widehat{R}_j^2 > \widehat{R}_i^2)$, and $\mathfrak{R}_d = \max_{i\in\mathcal{M}} \sum_j I(d_j > d_i)$. Next, for the estimation step, we evaluate the sparsity recovery and estimation accuracy properties. Let $\hat{d}^{(r)}_{\widehat{\mathcal{M}}^{(r)}} = (\hat{d}_j^{(r)} : j \in \widehat{\mathcal{M}}^{(r)})$ be estimator given $\widehat{\mathcal{M}}^{(r)}$. Further define $\widehat{\mathcal{M}}_1^{(r)} = \{j \in \widehat{\mathcal{M}}^{(r)} : \hat{d}_j \neq 0\}$. First, the true positive (TP) is defined as the average number of nonzero coefficients correctly estimated to be nonzero, i.e., $\text{TP}_e = \sum_{r=1}^{100} |\widehat{\mathcal{M}}_1^{(r)} \cap \mathcal{M}|/100$. The false positive (FP) is defined as the number of zero coefficients incorrectly estimated to be nonzero, i.e.,

$\text{FP}_e = \sum_{r=1}^{100} |\widehat{\mathcal{M}}_1^{(r)} \cap \mathcal{M}_0|/100$, where $\mathcal{M}_0$ is defined to be true non-portal nodes set, i.e., $\mathcal{M}_0 = \{j : d_j = 0\}$. Lastly, we define TM_e to be the proportion of all replications that the true portal nodes set being exactly identified. To evaluate the estimation accuracy, we calculate the root mean square error (RMSE) for d and γ as $\text{RMSE}_d = \{n^{-1} \sum_{j \in \widehat{\mathcal{M}}^{(r)}} (\hat{d}_j^{(r)} - d_j)^2\}^{1/2}$ and $\text{RMSE}_\gamma = \{p^{-1} \|\hat{\gamma}^{(r)} - \gamma\|^2\}^{1/2}$. The median of RMSEs for all replications is reported for both estimators. Lastly, the ratio of average in-degrees for portal nodes versus the non-portal nodes is reported.

The simulation results are given in Table 1–3. First, the TP_s for both the screening and estimation increases as N and T increase. For example, the TP_s increases from 9.10 to 9.64 as (N, T) increases from $(100, 50)$ to $(200, 100)$ for dyad independence model with $\delta = 1/4$. This corroborates the screening consistency properties. Moreover, the proposed portal nodes screening method shows higher screening accuracy with lower $\mathfrak{R}_r$ compared with $\mathfrak{R}_d$, where the performance is consistent for both MEDIAN NETWORK and LARGE NETWORK.

For the estimation part, as the sample size is increased, all the methods have a higher probability to identify the true model. With regards to the estimation accuracy, one could observe that the methods ALasso, MCP, and SCAD perform better than Lasso with lower RMSE levels. For example, the RMSE_d ($\times 10^2$) for ALasso, MCP, and SCAD are 1.07, 0.64, 1.48, which is much lower than Lasso estimator (i.e., 3.79) with $(N, T) = (200, 100)$ for the stochastic block network with $\delta = 1/4$.

4.3. A Sina Weibo Dataset

We next illustrate the portal nodes screening and estimation method using a Sina Weibo dataset. The data are collected from Sina Weibo (*www.weibo.com*), which is the largest Twitter-type social media in mainland China. From the perspective of managing such online social media platform, identifying the influential users is of particular interest. It could help to attract active users, launch marketing campaigns,

increase advertising profits and so on. As a result, it will lead to growing more profits for the social media platform.

Specifically, a total number of $N = 842$ active followers of an MBA program account are recorded for continuous $T = 30$ days. The network structure is constructed by the following relationship of network users, where the network density is 4.5%. The histograms of the in- and out-degrees are given in Figure 1. It can be observed that the distribution of in-degrees is much skewer than the out-degrees, which indicates the existence of influential network users.

The response Y_{it} is defined as the $\log(1+x)$-transformed Weibo post length for the ith user at the tth day. To account for the user posting behaviors, several covariates Z_{it} are taken into consideration. The first one is the lag-1 term $Y_{i(t-1)}$, which accounts for the autoregressive effect of Y_{it}. Next, the nodal covariates are taken into consideration to reflect the user characteristics. They are, the gender (male = 1, female = 0), tenure (i.e., the time length since the registration with Sina Weibo), number of personal labels, and the user description length. Particularly, the labels and descriptions are created by the user themselves to describe their life styles, characteristics, and so on.

We first conduct the screening procedure to estimate out the portal node set. The top 10 user accounts with highest $\widehat{R}_j^2$ are given by Figure 2. Most of them are celebrities and online social medias, which has backgrounds of finance, economics, and business. In addition, we compare the top 20 users selected by the portal nodes screening and the topology-based method (i.e., ranking nodes using in-degrees). Among them 12 are selected by both screening methods, while the topology-based screening method also include the "super stars" in more general backgrounds. For example, a famous Chinese actress (named Yao Chen) is included in the top 20 list of the degree screening method but not in ours. In addition, the top 20 portal nodes identified by our approach are not necessarily with large in-degrees. The smallest in-degree of the top 20 portal nodes of our approach is 109, which is quite small comparing to last portal node identified

by the topology-based method (with in-degree 206). That suggests the portal nodes selected by our method is more related to the user's specific backgrounds.

Next, the top $n = [N/\log(N)]$ users are kept after screening for further estimation. Specifically, the estimation methods, i.e., Lasso, ALasso, MCP ($a = 1.5$), and SCAD ($a = 3.7$), are applied. Subsequently, the HBIC criterion is employed to select the tuning parameter λ. Table 4 gives the QMLE estimation and inference results, where the listed 15 users are identified as portal nodes by at least one method. Among them there are nine famous online social media accounts, who release latest news about business, economics and finance. The other six users are famous celebrities also in these related fields. The estimated lag-1 autoregressive coefficient is around 0.46, which illustrates a positive momentum effect for user behaviours. For the nodal covariates, it is found the length of description of the users is positively related to the user's activeness level, and is significant under the 5% significance level. However, with respect to the other covariates (i.e., GENDER, TENURE, and LABEL), we cannot find any significant evidence showing that they are significantly related to the response.

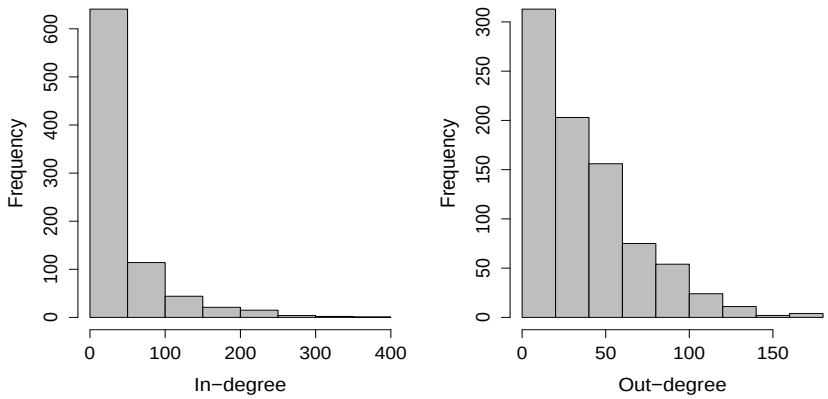

Figure 1: The Sina Weibo data analysis. The left panel: histogram of in-degrees for $N = 842$ nodes. The highly right skewed shape indicates the existence of "super stars" in the network; The right panel: similar histogram but for out-degrees.

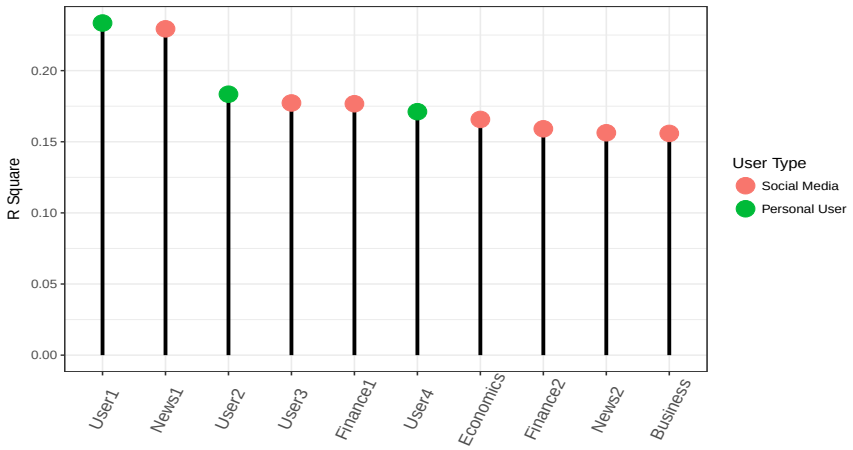

Figure 2: Users with top 10 $\widehat{R}_j^2$ for Sina Weibo dataset. Most accounts are famous online social medias. For privacy reasons, only the type of user accounts are shown here in different colors.

5. CONCLUDING REMARKS

This paper considers a new type network autoregression model. The major contributions of this work are summarized as follows.

- We quantify influential powers in the proposed NAM to represent the heterogeneous and sparse nodal influences. This is an essential generalization of traditional NAM (Chen et al., 2013; Zhu et al., 2017; Huang et al., 2017; Cohen-Cole et al., 2018) with homogenous influential powers.

- We provide a NAM-based screening method, that is a model-based method, to identify portal nodes. Comparing with the traditional topology-based methods (Carrington et al., 2005; Newman, 2010; Scott, 2012), the superiority of NAM-based screening method is that employs the dynamic response information Y_{it} to detect portal nodes.

- We prove that the NAM-based screening method has the screening consistency property. And the asymptotic normality of QMLE is established. These are

the advantages of NAM-based screening method comparing with the traditional topology-based methods without any theoretical guarantee.

- We further consider a portal node selection procedure after screening by giving an LLA algorithm. The finite sample performance is further demonstrated by a number of numerical studies.

To conclude the article, we discuss several potential future research topics.

First, as we have mentioned, the CSIS based method (Barut et al., 2016) could be also utilized as a screening measure to detect portal nodes. Then an important extension is to establish its screening consistency property for the network data. Next, although the theoretical properties of the portal nodes screening are established, the post selection and estimation properties remain largely unknown. In addition, the computation could be further reduced by utilizing the sparsity feature of the network structure (Zhu et al., 2018a). Since it might be beyond the scope of this work, we add this direction as an important future research topic.

Next, note that the influential powers defined in this work are associated with the response types. However, different types of responses might lead to different quantifications of the influential powers. In this work, we use posted tweet length as our response to quantify the users' influences, which is also widely used in literature (Chen et al., 2013; Zhou et al., 2017; Zhu et al., 2017). Other types of responses (e.g., online time length) could be used in the future works and a unified framework to measure the users' influences should be proposed.

Thirdly, the context of network analysis could be more general. The analytical procedures should also be developed and applied to broaden types of data, for example, financial time series, gene expression, geographical information, and so on. Moreover, the NAM model setting in this work could be further generalized. For example, when the time replications are involved, one could further consider to add fixed effects (e.g.,

Yu et al. (2008)) as

$$Y_{it} = \alpha_i + \sum_{j=1}^{N} d_j w_{ij} Y_{jt} + Z_{it}^{\top}\gamma + \varepsilon_{it},$$

where α_i denotes the fixed effect. In this case, one could consider to follow the techniques of Yu et al. (2008) to concentrate out the fixed effects for each node i. Next the screening and selection methods proposed in this work could still be applied. It could be interesting to study the theoretical properties under the new model settings.

References

Aral, S. and Walker, D. (2011), "Creating social contagion through viral product design: a randomized trial of peer influence in networks," *Management Science*, 57, 1623–1639.

— (2012), "Identifying influential and susceptible members of social networks," *Science*, 337, 337–341.

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

Barut, E., Fan, J., and Verhasselt, A. (2016), "Conditional sure independence screening," *Journal of the American Statistical Association*, 111, 1266–1277.

Bodendorf, F. and Kaiser, C. (2009), "Detecting opinion leaders and trends in online social networks," in *Proceedings of the 2nd ACM workshop on Social web search and mining*, ACM, pp. 65–68.

Breheny, P. and Huang, J. (2011), "Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection," *The Annals of Applied Statistics*, 5, 232–253.

Bühlmann, P. and Van De Geer, S. (2011), *Statistics for High-dimensional Data: Methods, Theory and Applications*, Springer Science & Business Media.

Carrington, P. J., Scott, J., and Stanley Wasserman, S. (2005), *Models and Methods in Social Network Analysis*, Cambridge University Press.

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.

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

Cohen-Cole, E., Liu, X., and Zenou, Y. (2018), "Multivariate choices and identification of social interactions," *Journal of Applied Econometrics*, 33, 165–178.

Dou, B., Parrella, M. L., and Yao, Q. (2016), "Generalized Yule–Walker estimation for spatio-temporal models with unknown diagonal coefficients," *Journal of Econometrics*, 194, 369–382.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultra-high dimensional feature space (with discussion)," *Journal of the Royal Statistical Society, Series B*, 70, 849–911.

Fan, J., Samworth, R., and Wu, Y. (2009), "Ultrahigh dimensional feature selection: beyond the linear model," *Journal of machine learning research*, 10, 2013–2038.

Fan, J., Xue, L., and Zou, H. (2014), "Strong oracle optimality of folded concave penalized estimation," *Annals of Statistics*, 42, 819–849.

Fu, W. J. (1998), "Penalized regressions: the bridge versus the lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

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

Goldenberg, A., Zheng, A. X., Fienberg, S. E., Airoldi, E. M., et al. (2010), "A survey of statistical network models," *Foundations and Trends® in Machine Learning*, 2, 129–233.

Gong, S., Zhang, J., Zhao, P., and Jiang, X. (2016), "Tweets and Sales," *Working Paper*.

Härdle, W. K., Wang, W., and Yu, L. (2016), "TENET: Tail-event driven network risk," *Journal of Econometrics*, 192, 499–513.

Hautsch, N., Schaumburg, J., and Schienle, M. (2014), "Financial network systemic risk contributions," *Review of Finance*, 19, 685–738.

Hinz, O., Skiera, B., Barrot, C., and Becker, J. U. (2011), "Seeding strategies for viral marketing: an empirical comparison," *Journal of Marketing*, 75, 55–71.

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.

Huang, D., Chang, X., and Wang, H. (2017), "Spatial autoregression with repeated measurements for social networks," *Communications in Statistics-Theory and Methods*, 1–13.

Iyengar, R., Den Bulte, C. V., and Valente, T. W. (2011), "Opinion leadership and social contagion in new product diffusion," *Marketing Science*, 30, 195–212.

Katona, Z., Zubcsek, P. P., and Sarvary, M. (2011), "Network effects and personal influences: The diffusion of an online social network," *Journal of Marketing Research*, 48, 425–443.

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer.

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

Lee, L.-f. and Yu, J. (2009), "Spatial nonstationarity and spurious regression: the case with a row-normalized spatial weights matrix," *Spatial Economic Analysis*, 4, 301–327.

Liu, X., Patacchini, E., and Rainone, E. (2017), "Peer effects in bedtime decisions among adolescents: a social network model with sampled data," *The Econometrics Journal*, 20, S103–S125.

Newman, M. (2010), *Networks: An Introduction*, Oxford University Press.

Newman, M. E. and Girvan, M. (2004), "Finding and evaluating community structure in networks," *Physical Review E*, 69, 026113.

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

Scott, J. (2012), *Social Network Analysis*, Sage.

Stephen, A. T. and Galak, J. (2012), "The effects of traditional and social earned media on sales: a study of a microlending marketplace," *Journal of Marketing Research*, 49, 624–639.

Tang, J., Hu, X., and Liu, H. (2013), "Social recommendation: a review," *Social Network Analysis and Mining*, 3, 1113–1133.

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

Wang, L., Kim, Y., and Li, R. (2013), "Calibrating non-convex penalized regression in ultra-high dimension," *Annals of Statistics*, 41, 2505–2536.

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

Yu, J., De Jong, R., and Lee, L.-f. (2008), "Quasi-maximum likelihood estimators for spatial dynamic panel data with fixed effects when both n and T are large," *Journal of Econometrics*, 146, 118–134.

Zhang, C.-H. et al. (2010), "Nearly unbiased variable selection under minimax concave penalty," *Annals of Statistics*, 38, 894–942.

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.

Zhu, X., Huang, D., Pan, R., and Wang, H. (2018a), "Multivariate Spatial Autoregression for Large Scale Social Networks," *Journal of Econometrics*, To appear.

Zhu, X., Pan, R., Li, G., Liu, Y., and Wang, H. (2017), "Network vector autoregression," *Annals of Statistics*, 45, 1096–1123.

Zhu, X., Wang, W., Wang, H., and Härdle, W. K. (2018b), "Network quantile autoregression," *Journal of Econometrics*, To appear.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *Annals of Statistics*, 36, 1509–1533.

Zou, T., Lan, W., Wang, H., and Tsai, C.-L. (2017), "Covariance Regression Analysis," *Journal of the American Statistical Association*, 112, 266–281.

Table 1: Simulation results ($\mathrm{TP}_s, \mathfrak{R}_r, \mathfrak{R}_d, \mathrm{TP}_e, \mathrm{TM}_e, \mathrm{RMSE}(\times 10^2)$) for the MEDIAN NETWORK with 100 replications for three examples with $\delta = 1/2$ are reported. The last column is the ratio of average nodal in-degrees of the portal nodes versus the non-portal nodes.

(N,T)	TP_s	$\mathfrak{R}_r$	$\mathfrak{R}_d$	Est.	TP_e	FP_e	TM_e	RMSE_d	RMSE_γ	Ratio
Example 1 : Dyad Independence Network										
(100,50)	9.8	17	58	Lasso	9.32	0.02	0.83	9.52	1.31	1.5
				ALasso	9.78	0.05	0.95	2.77	0.94	
				MCP	9.63	0.01	0.82	2.36	0.92	
				SCAD	9.61	0.01	0.81	3.49	0.95	
(200,100)	10	12	96	Lasso	9.9	0	0.9	5.00	0.44	2.0
				ALasso	10	0.01	0.99	1.26	0.33	
				MCP	9.88	0	0.93	1.73	0.33	
				SCAD	9.8	0	0.8	2.41	0.34	
Example 2 : Stochastic Block Model										
(100,50)	9.67	24	43	Lasso	9.56	0.06	0.89	7.51	1.66	4.3
				ALasso	9.67	0.12	0.89	2.44	0.98	
				MCP	9.44	0	0.77	2.57	0.93	
				SCAD	9.53	0	0.86	3.14	0.98	
(200,100)	9.92	19	67	Lasso	9.92	0.25	0.77	4.19	0.63	7.3
				ALasso	9.92	0.07	0.93	1.48	0.36	
				MCP	9.92	0	1	0.77	0.33	
				SCAD	9.92	0.01	0.99	1.11	0.34	
Example 3 : Power-law Distribution Network										
(100,50)	9.71	21	54	Lasso	9.47	0.03	0.73	8.43	1.16	1.7
				ALasso	9.71	0.13	0.88	2.78	0.98	
				MCP	9.44	0.03	0.7	3.00	0.96	
				SCAD	9.36	0.03	0.62	4.69	0.98	
(200,100)	9.98	16	97	Lasso	9.78	0.01	0.79	6.16	0.43	2.4
				ALasso	9.98	0.06	0.94	1.71	0.34	
				MCP	9.88	0.01	0.89	1.51	0.33	
				SCAD	9.76	0.01	0.77	2.73	0.34	

Table 2: Simulation results ($\text{TP}_s, \mathfrak{R}_r, \mathfrak{R}_d, \text{TP}_e, \text{TM}_e, \text{RMSE}(\times 10^2)$) for the MEDIAN NETWORK with 100 replications for three examples with $\delta = 1/4$ are reported. The last column is the ratio of average nodal in-degrees of the portal nodes versus the non-portal nodes.

(N,T)	TP_s	$\mathfrak{R}_r$	$\mathfrak{R}_d$	Est.	TP_e	FP_e	TM_e	RMSE_d	RMSE_γ	Ratio
Example 1 : Dyad Independence Network										
(100,50)	9.1	35	84	Lasso	9.05	0.04	0.95	6.69	1.00	0.7
				ALasso	9.1	0.03	0.98	2.50	0.91	
				MCP	9.09	0	0.99	2.00	0.91	
				SCAD	8.96	0	0.87	3.67	0.91	
(200,100)	9.64	36	167	Lasso	9.62	0	0.99	4.25	0.35	0.6
				ALasso	9.64	0	1	1.69	0.34	
				MCP	9.63	0	0.99	1.33	0.33	
				SCAD	9.4	0	0.76	2.83	0.34	
Example 2 : Stochastic Block Model										
(100,50)	9.97	12	47	Lasso	9.94	0.06	0.92	5.94	1.10	2.2
				ALasso	9.97	0.03	0.97	2.03	0.93	
				MCP	9.97	0.02	0.98	1.45	0.92	
				SCAD	9.79	0.01	0.81	3.47	0.94	
(200,100)	10	10	75	Lasso	10	0	1	3.79	0.38	2.8
				ALasso	10	0	1	1.07	0.34	
				MCP	10	0	1	0.64	0.33	
				SCAD	9.93	0	0.93	1.48	0.34	
Example 3 : Power-law Distribution Network										
(100,50)	9.69	21	54	Lasso	9.68	0.09	0.92	6.73	0.93	0.7
				ALasso	9.68	0.1	0.89	3.01	0.90	
				MCP	9.66	0.06	0.93	2.54	0.90	
				SCAD	9.4	0.04	0.7	4.67	0.90	
(200,100)	9.97	17	95	Lasso	9.97	0.01	0.99	4.83	0.32	0.7
				ALasso	9.97	0.05	0.95	1.89	0.32	
				MCP	9.96	0	0.99	1.38	0.32	
				SCAD	9.49	0	0.52	3.98	0.32	

Table 3: Simulation results ($\mathrm{TP}_s, \mathfrak{R}_r$, $\mathfrak{R}_d, \mathrm{TP}_e$) for the LARGE NETWORK with 100 replications for three examples with $\delta = 1/2$ and $\delta = 1/4$ are reported. The Ratio is average nodal in-degrees of the portal nodes versus the non-portal nodes, and the ND is the network density.

	$\delta = 1/2$					$\delta = 1/4$				
(N, T)	TP_s	$\mathfrak{R}_r$	$\mathfrak{R}_d$	Ratio	ND (%)	TP_s	$\mathfrak{R}_r$	$\mathfrak{R}_d$	Ratio	ND (%)
Example 1 : Dyad Independence Network										
(2000, 200)	10	11	850	5.73	0.33	10	24	851	0.98	0.31
(5000, 500)	10	11	2271	8.82	0.13	10	10	2270	1.20	0.13
Example 2 : Stochastic Block Model										
(2000, 200)	10	14	451	32.82	0.13	10	10	458	5.97	0.09
(5000, 500)	10	15	1025	56.93	0.04	10	10	1031	7.59	0.04
Example 3 : Power-law Distribution Network										
(2000, 200)	10	29	978	7.42	0.25	10	24	986	1.14	0.23
(5000, 500)	10	39	2464	11.73	0.10	10	30	2477	1.46	0.09

Table 4: Estimation result for Sina Weibo dataset. The user account types for the portal nodes are reported with estimated influential powers. In addition, the estimated coefficients for nodal covarates are also given. For the last three covariates (i.e., user labels, tenure, and description), the coefficients are presented with $\times 10^3$ times. In addition, the p-values are also reported with p-value less than 0.05 marked by *.

User/Covariates	Est.	SE	p-value
Portal Nodes			
News	3.457	0.234	< 0.001*
News	0.963	0.122	0.006*
News	0.724	0.033	< 0.001*
Business	1.560	0.444	0.019*
Business	0.897	0.255	0.076
Business	0.839	0.085	0.004*
Economics	0.643	0.092	0.034*
Economics	0.529	0.120	0.128
Finance	0.697	0.054	0.003*
User	1.431	0.285	0.007*
User	1.280	0.300	0.020*
User	1.155	0.287	0.031*
User	0.800	0.229	0.095
User	0.658	0.162	0.102
User	0.191	0.192	0.663
Covariates Estimation			
Intercept	2.669	0.029	< 0.001*
Lag Response	0.462	0.000	< 0.001*
Gender	0.062	0.002	0.129
Labels	0.408	1.656	0.751
Tenure	-0.065	0.003	0.258
Description	6.574	0.511	< 0.001*