Statistics and Its Interface Volume 14 (2021) 115–129

Information diffusion with network structures

Xuening Zhu*, Rui Pan†,‡, Yuxuan Zhang, Yu Chen,
Wenquan Mi, and Hansheng Wang§

Information diffusion refers to the process about passing certain information from one subject to another. It is a typical and critical phenomenon observed in large scale social networks. To statistically model such a phenomenon, a network diffusion model is proposed and studied. The diffusion process is then investigated under the modeling framework from both the short term and long term perspectives. To estimate the model, a maximum likelihood estimator and a moment estimator are proposed, whose asymptotic properties are further established. The resulting estimators are manifested to have a reliable finite sample performance through a number of numerical studies. Lastly, the diffusion of earthquake news on Sina Weibo is analyzed to illustrate the practical usefulness.

Keywords and phrases: Diffusion process, Maximum likelihood estimator, Moment estimator, Social network.

1. INTRODUCTION

Diffusion refers to the dynamics concerning passing certain information (or action) from one subject to another (Bass, 1969). Particularly, network diffusion refers to the diffusion occurred on a social network. According to data published by GlobalWebIndex (www.globalwebindex.net), internet users spend 135 minutes every day on social networking. It is conceivable that internet users are now under great exposure to messages posted by others. On social network platform like Twitter, users are influenced by those messages from day to day, and they might re-publish them to others. Empirical study conducted by Zhou et al. (2010) shows that the number of tweets concerning a specific event increases gradually and dies down as the time goes by. The network diffusion process, which is induced by network diffusion, can be mathematically represented as follows.

First, define a network adjacency matrix $A = (a_{i_1 i_2}) \in \mathbb{R}^{N \times N}$ to describe a static network structure with N nodes, where $a_{i_1 i_2} = 1$ if node i_1 follows i_2, and $a_{i_1 i_2} = 0$ otherwise. Following the convention in Zhu et al. (2017), we do not allow self-connected edges, i.e., $a_{ii} = 0$ for $1 \le i \le N$. Next, at a given time point t ($1 \le t \le T$), a binary variable $Y_{it} \in \{0, 1\}$ is recorded for each node i. Define $Y_{it} = 1$ if node i responds to certain information stimulus at time point t, and $Y_{it} = 0$ otherwise. Take Twitter-type social network as an example: node i might be exposed to a tweet (i.e., a information stimulus) from its followees. Subsequently, the node should decide to re-tweet (i.e., $Y_{it} = 1$) or not (i.e., $Y_{it} = 0$). Define $\mathbb{Y}_t = (Y_{1t}, \cdots, Y_{Nt})^\top \in \mathbb{R}^N$ to be the response vector at time t. This leads to the network diffusion process $\{\mathbb{Y}_t : 1 \le t \le T\}$. The objective of this work is to develop a statistical model to describe the underlying dynamics for $\mathbb{Y}_t$.

It is obvious that diffusion is a classical problem of fundamental importance. Accordingly, various diffusion models have been developed. In the seminal work of Bass (1969), potential customers are classified into innovators and imitators. This leads to an elegant differential equation with an analytical solution. In addition, various estimation methods have been discussed for Bass model (Schmittlein and Mahajan, 1982; Srinivasan and Mason, 1986; Kou et al., 2012). The Bass model and its extensions have a large amount of applications in marketing (Goswami, 2001; Van den Bulte and Stremersch, 2004; Rogers, 2010). Despite its popularity, Bass model and its extensions are based on aggregate data. As an alternative, Jun and Park (1999) and Niu (2002) proposed diffusion models at individual level to better understand diffusion process. Later, Yang and Leskovec (2010) and Du et al. (2014) studied the influence function based model where influence function is estimated for each node. Gao et al. (2017) used embedding model for information diffusion prediction.

Nevertheless, the models mentioned above either ignore or bypass network structure. How to include network structure in diffusion model is still open and widely studied. Independent cascade model (Goldenberg, Libai and Muller, 2001) is one of the earliest models incorporating explicit network structure. It assumes each activated node has a chance to activate its neighbors in each step.

*Xuening Zhu is supported by the National Natural Science Foundation of China (NSFC, 11901105, 71991472, U1811461), the Shanghai Sailing Program for Youth Science and Technology Excellence (19YF1402700), and the Fudan-Xinzailing Joint Research Centre for Big Data, School of Data Science, Fudan University.
†Corresponding author.
‡The research of Rui Pan is supported by National Natural Science Foundation of China (NSFC, 11971504, 11631003, 71771224), the Fundamental Research Funds for the Central Universities (QL18010), the Youth Talent Development Support Program (QYP1911) and the Program for Innovation Research in Central University of Finance and Economics.
§Hansheng Wang's research is partially supported by National Natural Science Foundation of China (NSFC, 11831008, 11525101, 71532001). It is also supported in part by China's National Key Research Special Program (No. 2016YFC0207704).

Katona, Zubcsek and Sarvary (2011) used complementary log-log link function to formulate adoption probability. Wang, Wang and Xu (2012) modeled temporal and spatial characteristics of information by diffusive logistic equation. In recent works, various relationships among users and information are introduced into diffusion models. For example, Li et al. (2017) studied rational agents assumption in information diffusion. Zhang et al. (2018) incorporated interactions among users and contagions into a unified framework.

However, most previous models assume that a node will stay active once it is activated ($Y_{it} = 1$ if $Y_{i(t-1)} = 1$ in discrete case). Although this assumption is reasonable in the field of marketing, it is too restrictive to study the dynamic of an event. This motivates us to propose an interesting *network diffusion model*, which enables us to understand the diffusion process with network structure from both aggregate and individual perspectives.

Furthermore, the process is also investigated both at the very beginning (i.e., short term analysis) and at the end (i.e., long term analysis) of the diffusion. Moreover, because the model is specified at individual level, parameter estimation can be conducted at a cross-sectional level. This enables us to estimate the model at very early stage of the diffusion process. A maximum likelihood estimator (MLE) and a moment estimator (ME) are then proposed. The asymptotic properties are studied and their finite sample performances are compared by extensive numerical studies.

We would like to summarize our contributions to the literature in the following three regards. First, we study an individualized network diffusion model, which takes simple form and embeds the known network structure information. Second, two estimation approaches are investigated and compared both in theory and computational side. Third, we provide abundant theoretical justifications for the proposed network diffusion model and study the corresponding asymptotic properties. Although the proposed network diffusion model framework is parsimonious, we would like to emphasize that the contribution of this work is mainly on the theoretical side and abundant extensions could be investigated.

The rest of this article is organized as follows. Section 2 introduces the network diffusion model, where the short term and long term analysis are conducted subsequently. In Section 3 we propose two types of estimators and their corresponding asymptotic properties are investigated. Extensive numerical studies and a Sina Weibo data analysis are given in Section 4. The article is concluded with a brief discussion in Section 5. All technical details are left in the Appendix.

2. NETWORK DIFFUSION MODEL

2.1 Model and notations

Recall that N is the network size and Y_{it} is the binary response collected from the ith subject at time point t. Define $\mathcal{F}_t = \{\mathbb{Y}_t, \cdots, \mathbb{Y}_0\}$ be a set of the historical information up to time point t. We further assume that the response variable Y_{it} given $\mathcal{F}_{t-1}$ is fully determined by $\mathbb{Y}_{t-1}$ as

$$P(Y_{it} = 1|\mathcal{F}_{t-1}) = p_{it} = \rho n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)}, \tag{1}$$

where $n_i = \sum_{j \neq i} a_{ij}$ is the total number of nodes that i follows, and it is referred to as nodal out-degree (Wasserman et al., 1994). For a fixed t, different Y_{it}s are assumed to be independent given $\mathcal{F}_{t-1}$. As one can see, the quantity $n_i^{-1} \sum_{j=1}^{N} a_{ij} Y_{j(t-1)}$ is the average stimulus received by i from the nodes it follows at $t-1$. Its impact on Y_{it} is quantified by ρ, which is referred to as *network diffusion effect*. In order to ensure (1) to be a valid probability measure, a sufficient condition is $0 \leq \rho \leq 1$. Practically, the estimated network diffusion effect is usually very small; see Section 4.3 for the empirical evidence. Therefore, this assumption is easy to satisfy. To sum up, model (1) specifies the network diffusion model. It explicitly takes the network structure into consideration. By model (1), we know that the information stimulus, received at time point t, is induced by $\mathbb{Y}_{t-1}$. This information stimulus passes through the network, and eventually affects whether $Y_{it} = 1$ or 0 for $1 \leq i \leq N$.

For convenience, define $\mathbb{P}_t = (p_{1t}, \cdots, p_{Nt})^\top \in \mathbb{R}^N$. Given $\mathcal{F}_{t-1}$, $\mathbb{Y}_t$ follows multivariate Bernoulli distribution with conditional probability $\mathbb{P}_t$, which is defined as

$$\mathbb{P}_t = \rho W \mathbb{Y}_{t-1}, \tag{2}$$

where $W = (w_{ij}) = \text{diag}\{n_1^{-1}, \cdots, n_N^{-1}\}A$ is the row-normalized adjacency matrix. Therefore, a weighted adjacency matrix is involved here in our network diffusion model, where the weights are related to the out-degrees of the network nodes (Zhu et al., 2017). As a result, compared to a node following ten friends, a node follows ten thousand friends will receive less influence from its following friends. In this work, we treat the weighting matrix to be static. While in practice, the weighting matrix could be in a more general form, e.g., in dynamic forms. We would like to leave it as an important future study direction to the proposed model.

The network diffusion model (1) is given at an individual level. However, on some occasions it is of particular interest to analyze the network diffusion at an aggregate level. For a given time point t, the total amount of responses to the information stimulus is given by $\mathbf{1}^\top \mathbb{Y}_t$, where $\mathbf{1} = (1, \cdots, 1)^\top \in \mathbb{R}^N$. For convenience, we refer to $\text{IDS}(t) = \mathbf{1}^\top \mathbb{Y}_t$ as the *incremental diffusion size* (IDS). It is then of interest to theoretically investigate the behavior of $\text{IDS}(t)$ in both short and long terms, which is discussed in the following two subsections respectively.

2.2 Short term analysis of IDS

In this subsection, we consider short term analysis of IDS. Assume $\mathcal{F}_t$ is given, we then investigate the probabilistic

characteristics of both IDS$(t+1)$ and IDS$(t+2)$. Define $E^*(\cdot) = E(\cdot|\mathcal{F}_t)$ and $\text{var}^*(\cdot) = \text{var}(\cdot|\mathcal{F}_t)$. We then analyze the IDS in one step and two steps ahead respectively as follows.

2.2.1 One step ahead analysis

Note that $p_{i(t+1)} = E^*(Y_{i(t+1)}) = \rho w_i^\top \mathbb{Y}_t$, where $w_i = (w_{i1}, \cdots, w_{iN})^\top \in \mathbb{R}^N$ is the ith row vector of W. We then have

$$E^*\Big\{\text{IDS}\,(t+1)\Big\} = \sum_{i=1}^{N} \rho w_i^\top \mathbb{Y}_t = \rho \mathbf{1}^\top W \mathbb{Y}_t = \rho \sum_{j=1}^{N} Y_{jt}\Big(\sum_{i=1}^{N} w_{ij}\Big). \tag{3}$$

By (3), we know that the expected value of IDS$(t+1)$ depends on two important factors. First, it depends on Y_{jt}s, i.e., the response of node j at time t for $1 \le j \le N$. The more nodes respond to the information stimulus (i.e., $Y_{jt} = 1$), the more likely to observe a large expected value for IDS$(t+1)$. Second, IDS$(t+1)$ depends on the quantity $\sum_{i=1}^N w_{ij} = \sum_{i=1}^N n_i^{-1} a_{ij}$. Given a response node j (i.e., $Y_{jt} = 1$), the more followers j has (i.e., $\sum_{i=1}^N a_{ij}$), it is more likely to have a large value for $\sum_i w_{ij}$. The value could be even larger if the followers are *loyal* with small nodal out-degree (i.e., node i satisfying $a_{ij} = 1$ with small n_i). We can define the incremental value of $E^*\{\text{IDS}(t+1)\}$ with respect to IDS(t) as

$$\Delta(t+1) = E^*\Big\{\text{IDS}(t+1)\Big\} - \text{IDS}(t) = \sum_{j=1}^{N} Y_{jt}\Big(\rho \sum_{i=1}^{N} w_{ij} - 1\Big). \tag{4}$$

By (4), we know that IDS$(t+1)$ is more likely to be larger than IDS(t) if $\sum_i w_{ij}$ is large, for many js with $Y_{jt} = 1$. This implies that there are many famous users (i.e., users with large amount of loyal followers) who respond at time t. In this case, the diffusion process is likely to expand. Otherwise, it is expected to shrink.

We next consider the conditional variance of IDS$(t+1)$. Conditional on $\mathcal{F}_t$, $Y_{i(t+1)}$s are independent. Accordingly, we have

$$\text{var}^*\Big\{\text{IDS}(t+1)\Big\} = \sum_{i=1}^{N} V_{i(t+1)}, \tag{5}$$

where $V_{i(t+1)} = p_{i(t+1)}(1 - p_{i(t+1)})$. In practice, the conditional probability that node i will respond (i.e., $p_{i(t+1)}$) is typically small and close to 0. Furthermore, $\text{var}^*\{\text{IDS}(t+1)\}$ could be expanded as

$$\text{var}^*\Big\{\text{IDS}(t+1)\Big\} = \sum_{i=1}^{N} p_{i(t+1)} - \sum_{i=1}^{N} p_{i(t+1)}^2 = E^*\Big\{\text{IDS}(t+1)\Big\} - \sum_{i=1}^{N} p_{i(t+1)}^2. \tag{6}$$

This suggests that the magnitude of the conditional variance of IDS$(t+1)$ is mainly determined by its conditional mean. Based on the above results, we then conduct the the two steps analysis of IDS subsequently.

2.2.2 Two steps ahead analysis

Similarly, we can evaluate the conditional mean and variance for IDS$(t+2)$. This leads to

$$E^*\Big\{\text{IDS}(t+2)\Big\} = \sum_{i=1}^{N} p_{it}^{(2)} \tag{7}$$

and $\text{var}^*\{\text{IDS}(t+2)\} = \sum_{i=1}^N V_{i(t+2)}$, where $p_{it}^{(2)} = \sum_{j=1}^N \rho^2 w_{ij} w_j^\top \mathbb{Y}_t = \rho^2 w_i^\top W \mathbb{Y}_t$, $V_{i(t+2)} = p_{it}^{(2)}(1 - p_{it}^{(2)}) + \rho^2 \sum_{1 \le j \ne k \le N} w_{ji} w_{ki} V_{i(t+1)}$. We then have

$$\text{var}^*\Big\{\text{IDS}(t+2)\Big\} = E^*\Big\{\text{IDS}(t+2)\Big\} - \sum_{i=1}^{N} \Big(p_{it}^{(2)}\Big)^2 + \rho^2 \sum_{i=1}^{N} \sum_{1 \le j \ne k \le N} w_{ji} w_{ki} p_{i(t+1)}\Big(1 - p_{i(t+1)}\Big). \tag{8}$$

The proof of (8) is given in Appendix A.1 Comparing (8) with (6), we find that the conditional variance of IDS$(t+2)$ could be much larger than its conditional expectation. While in the previous analysis, the conditional variance of IDS$(t+1)$ is mainly determined by its conditional expectation. This is because of the quantity $\sum_i \sum_{j \ne k} w_{ji} w_{ki} p_{i(t+1)}(1 - p_{i(t+1)})$, which could be large if there exists one or multiple nodes i such that $\sum_{j \ne k} w_{ji} w_{ki}$ are extremely large. This might happen if i is a node with a huge amount of followers (e.g., a celebrity). To summarize, the comparison between (6) and (8) suggests that an accurate prediction of IDS$(t+2)$ is substantially more difficult than that of IDS$(t+1)$. This implies that a reliable long-term inference for IDS(t) is theoretically difficult.

2.3 Long term analysis of IDS

Even though it is difficult to access an accurate long-term forecasting, it does not rule out the possibility to make some reliable judgment for its primary trend. The primary trend mainly concerns about whether the diffusion process should: (1) gradually shrink and eventually disappear, or (2) maintain at a relatively stable level. To this end, we mainly concern the mean of IDS(t) and study its relationship with t.

To facilitate the discussion, we treat $\mathbb{Y}_0$ as fixed in the following. Define EIDS(t) to be the expectation of IDS(t). We then have EIDS(t) $= E(\mathbf{1}^\top \mathbb{Y}_t) = \rho^t \mathbf{1}^\top W^t \mathbb{Y}_0$. Note that W can be viewed 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., $\{1, \cdots, N\}$. We further assume that the Markov chain is irreducible, which implies that two arbitrary nodes in the network can always be connected with a path of finite length. In real life, irreducibility holds commonly according to six degrees of separation theory (Watts and Strogatz, 1998). If irreducibility is not satisfied, this indicates that the network could be divided into several isolated parts, and models could be built separately on each part. Under the assumption of irreducibility, the classical Markov chain theory states that there should exist a stationary distribution $\pi = (\pi_1, \cdots, \pi_N)^\top \in \mathbb{R}^N$ such that $\pi^\top W = \pi^\top$ and $W^t \to \mathbf{1}\pi^\top$ as $t \to \infty$. We then study the limit and convergence rate of EIDS(t) respectively as follows. First, we investigate the limit of EIDS(t) as $t \to \infty$. The main results are given in the following theorem.

Theorem 1. *Assume the Markov chain associated with W is irreducible and aperiodic. Then as $t \to \infty$, (1) EIDS(t) $\to_p 0$, if $0 \le \rho < 1$; and (2) EIDS(t) $\to_p N\pi^\top \mathbb{Y}_0$, if $\rho = 1$.*

The proof of Theorem 1 is given in Appendix A.2. By Theorem 1, we know that the long-term primary trend of a network diffusion process is mainly determined by its network diffusion effect (i.e., ρ). EIDS(t) shrinks towards 0 (i.e. the diffusion process eventually disappear) if $0 \le \rho < 1$, while it reaches a quantity if $\rho = 1$.

Next, we investigate the rate that EIDS(t) converges towards its limit EIDS$^* = \lim_{t\to\infty}$ EIDS(t). It helps us to understand how long the diffusion process will reach its stable status or disappear. This is important for a network platform operator, since the survival time of information determines the activeness of the network. To this end, write the distance between EIDS(t) and EIDS* as $d(t) = |\text{EIDS}(t) - \text{EIDS}^*|$. Accordingly, we can define the *ϵ-convergence time* as $t(\epsilon) = \min\{t : d(t) \le \epsilon\}$. It can be seen $t(\epsilon)$ characterizes the minimum time which is required to make the distance between EIDS(t) and EIDS* to be less than ϵ. In the long term analysis we focus on a symmetric network structure (i.e., $A = A^\top$) to facilitate theoretical discussions as follows.

Theorem 2. *Assume the adjacency matrix A is symmetric and let $W = (w_{ij}) = \text{diag}\{n_1^{-1}, \cdots, n_N^{-1}\}A$ be the row-normalized adjacency matrix. Define $\lambda^* = \max\{|\lambda_i| : |\lambda_i| \ne 1\}$, where λ_i is the ith eigenvalue of W. Then we have*

$$t(\epsilon) \le \frac{1}{1 - \lambda^* - \log\rho} \log \frac{2N}{\epsilon \gamma_{\min}}, \tag{9}$$

where $\gamma_{\min} = \min\{n_i / \sum_j n_j : 1 \le i \le N\}$.

The proof of Theorem 2 is given in Appendix A.3. Although the symmetric assumption on A seems to be restrictive, in practice, a lot of famous social networks (e.g., Facebook, WeChat) satisfy this assumption. For instance, on social network platforms like Facebook, relationship is built only when two nodes are mutually followed with each other, which yields a symmetric network relationship.

From Theorem 2, it can be concluded that the upper bound of ϵ-convergence time is mainly determined by the following four factors. They are, respectively, the network diffusion effect ρ, the network size N, the spectral gap of W (i.e., $1 - \lambda^*$), and $\gamma_{\min}$. Firstly, by (9), the upper bound increases as ρ increases. Thus, smaller network diffusion effect leads to faster convergence speed for EIDS(t). Secondly, the upper bound is smaller when network size N is small. This implies that the convergence speed is faster for small networks. Thirdly, as demonstrated by Banerjee, Carlin and Gelfand (2014), it holds that $\max_i |\lambda_i| = 1$. Therefore one must have the spectral gap $1 - \lambda^* > 0$. As suggested by (9), it will result in faster convergence of EIDS(t) to its limit. Lastly, as $\gamma_{\min}$ gets larger, a faster convergence speed could be achieved.

Remark. In practice, typically we have $0 \le \rho < 1$, which implies that the network diffusion process will disappear eventually. Theorem 2 provides some insights that how fast the diffusion process will disappear. In practice, it provides a rough estimation about the disappearing time of the diffusion process, which helps the practitioners to make quick decisions in the early stage when the diffusion occurs. Furthermore, We note that the disappearing speed of the diffusion process is related to several factors as we have mentioned above. As a result, if a practitioner would like to intervene the diffusion process, he/she may need to pay attention to the above factors. For instance, in the context of new product releasing, the practitioners may pursue to lengthen the diffusion process as much as possible. Therefore, he is suggested to (1) release the product in a large social network platform; (2) try to make more users to repost relevant news and information; (3) increase the connectivity of the network structure (i.e., lower spectral gaps) and try to introduce more super stars (i.e., make $\gamma_{\min}$ to be lower) in the network. On the contrary, in the context of epidemiology, the practitioners may want to shorter the diffusion process. Therefore, he is advised to intervene the diffusion process in an opposite way.

3. PARAMETER ESTIMATION

In this section, we discuss the parameter estimation methods for the network diffusion model (1). Two types of estimators are studied. The first one is the MLE and the second one is the ME. We would like to remark that in this section we do not assume the adjacency matrix A to be symmetric.

We first discuss the maximum likelihood estimation method. To this end, we first write the log-likelihood function $\ell(\rho)$ as

$$\ell(\rho) = \sum_{t=1}^{T}\sum_{i=1}^{N} I(w_i^\top \mathbb{Y}_{t-1} > 0)\Big\{Y_{it}\log(\rho w_i^\top \mathbb{Y}_{t-1}) + (1-Y_{it})\log(1-\rho w_i^\top \mathbb{Y}_{t-1})\Big\}, \tag{10}$$

where $I(\cdot)$ is the indicator function. Then, the MLE could be obtained as $\hat{\rho}_{mle} = \text{argmax}_\rho \ell(\rho)$.

We then investigate the asymptotic properties for MLE. First, we assume that $0 \le \rho < 1$, since this is the most realistic situation where the diffusion process disappear eventually. In this case, we have $\text{EIDS}(t) \to 0$. This suggests that the number of observed responses is limited as $t \to \infty$. As a consequence, the asymptotic theory should be established with a bounded T. In such a situation, as long as the expected number of potential responses (i.e., $\sum_{t=1}^{T} E(\mathbb{Y}_t) = \sum_{t=1}^{T} \rho^{t-1}\mathbf{1}^\top W^t \mathbb{Y}_0$) is large, the MLE is expected to be consistent. This motivates us to define $N^* = \sum_{t=1}^{T} \rho^{t-1}\mathbf{1}^\top W^t \mathbb{Y}_0$ as the effective sample size and assume $N^* \to \infty$ as $N \to \infty$. We first derive the asymptotic properties for $\hat{\rho}_{mle}$ as follows.

Theorem 3. *Let* $Z_i = \sum_{t=0}^{T-1}\Big\{\rho(1-\rho w_i^\top \mathbb{Y}_t)\Big\}^{-1} w_i^\top \mathbb{Y}_t$ *and assume (1)* $N^* \to \infty$ *as* $N \to \infty$*; (2)* $N^{*-1}\sum_i Z_i \to_p \sigma_\rho^2$*. We then have* $\sqrt{N^*}(\hat{\rho}_{mle} - \rho) \to_d N(0, \sigma_\rho^{-2})$ *as* $N^* \to \infty$*.*

The proof of Theorem 3 is given in Appendix A.4. Assumption (1) in Theorem 3 guarantees that the effective sample size N^* diverges to infinity as the network size N diverges. Assumption (2) in Theorem 3 is a law of large number type assumption imposed on Z_i, which assumes certain type of uniformity among the network nodes. Although MLE has higher estimation efficiency compared to ME, it is hard to obtain an analytical form of $\hat{\rho}_{mle}$. As an alternative, the ME $\hat{\rho}_{me}$ is proposed with analytical form as

$$\hat{\rho}_{me} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} w_i^\top \mathbb{Y}_{t-1}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}. \tag{11}$$

To derive the theoretical results of $\hat{\rho}_{me}$, we first write $W^t = (w_{ij}^{(t)})$, where $w_{ij}^{(t)}$ is the (i,j)th element of W^t. For convenience, we further define $\nu = (\nu_1, \cdots, \nu_N)^\top = \sum_{t=1}^{T} E(\mathbb{Y}_t) = \sum_{t=1}^{T} \rho^t W^t \mathbb{Y}_0$, where ν_i is the expected number of responses for node i during $1 \le t \le T$. Let $\tilde{\nu} = (\tilde{\nu}_1, \cdots, \tilde{\nu}_N)^\top \in \mathbb{R}^N$, where $\tilde{\nu}_i = I(\nu_i > 0)$ indicates whether node i has a positive probability to response during the whole time period T. The asymptotic property of $\hat{\rho}_{me}$ is then given by Theorem 4 as follows.

Theorem 4. *Assume (1)* $N^* \to \infty$ *as* $N \to \infty$*; (2)* $(N^*)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\Big\{p_{it}(1-p_{it})\Big\} \to_p \sigma_1^2$*; (3)* $(N^*)^{-2}\sum_{t=1}^{T}\sum_{j\in J}(\sum_i w_{ij}^{(t)})^2 \to 0$*, where* $J = \{j : \tilde{\nu}_j = 1\}$*, as* $N^* \to \infty$*. Then we have* $\sqrt{N^*}(\hat{\rho}_{me} - \rho) \to_d N(0, \sigma_1^2)$ *as* $N^* \to \infty$*.*

The proof of Theorem 4 is given in Appendix A.5. We illustrate the conditions respectively as follows. The first condition guarantees that the effective sample size N^* should diverge to infinity as the network size N diverges. The second condition is a law of large number type condition, which restricts the heterogeneity levels of the nodes. For the last condition, note that $w_{ij}^{(t)} \neq 0$ implies that the ith node could be connected with the jth node within t steps. Therefore, it imposes a sparsity condition for the network structure. It is remarkable that $\hat{\rho}_{me}$ is an $\sqrt{N^*}$-consistent estimator. In practice, the network is extremely sparse thus this condition will be easily satisfied.

4. NUMERICAL STUDIES

4.1 Simulation models

To demonstrate the finite sample performance of the proposed methodology, we present three examples in this subsection. The main difference is the generating mechanism of the adjacency matrix A. We first generate Y_{i0} independently according to Bernoulli distribution with $p = 0.3$ for $1 \le i \le N$. Then, $\mathbb{Y}_t$s $(t = 1, \cdots, T)$ are randomly generated according to model (1). Lastly, for each example, we set the network diffusion effect $\rho \in \{0.1, 0.3, 0.5\}$.

Example 1 (Dyad Independence Model)**.** We follow Holland and Leinhardt (1981) to define a dyad as $D_{ij} = (a_{ij}, a_{ji})$ $(1 \le i < j \le N)$, which is assumed to be independent with each other. To ensure the network sparsity, we set the probability of symmetric dyad (i.e., $D_{ij} = (1,1)$) as $P(D_{ij} = (1,1)) = 20N^{-1}$. Next, we set $P(D_{ij} = (1,0)) = P(D_{ij} = (1,0)) = 0.5N^{-0.8}$. This implies the expected nodal in-degree and out-degree is $O(N^{0.2})$, which diverges to infinity at a slow rate. Lastly, we have $P(D_{ij} = (0,0)) = 1 - 20N^{-1} - N^{-0.8}$, which implies null dyads (i.e., $D_{ij} = (0,0)$) are the most frequently observed in real social networks.

Example 2 (Stochastic Block Model)**.** Next, we consider the stochastic block model (Wang and Wong, 1987; Nowicki and Snijders, 2001), which is of particular interest for community detection (Zhao et al., 2012). Specifically, we follow Nowicki and Snijders (2001), and randomly generate a block label $(k = 1, \cdots, K)$ for each node with equal probability. The number of blocks is set to be $K = N/100$. Next, we 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. Consequently, nodes have higher probability to connect if they belong the same block.

Example 3 (Power-Law Distribution Model)**.** According to Clauset, Shalizi and Newman (2009), the nodal in-degrees

Table 1. Simulation Results with 1000 Replications for dyad independence network. The Root Mean Square Error ($\times 10^{-2}$), Coverage Probability (%), CPU time (in second), and the Network Density (%) are reported

ρ	Sample Size		MLE			ME			
	N	N^*	RMSE	CP(%)	CPU	RMSE	CP(%)	CPU	ND(%)
0.1	500	183.5	2.27	92.9	0.2	2.27	92.8	0.1	4.69
	1000	331.9	1.71	94.6	0.7	1.71	94.5	0.2	2.40
	2000	688.9	1.18	93.7	3.0	1.18	93.6	1.0	1.23
	10000	3300.0	0.53	94.8	70.0	0.53	94.8	23.4	0.26
0.3	500	218.8	3.57	94.3	0.2	3.57	94.4	0.1	4.69
	1000	395.3	2.62	95.2	0.8	2.63	95.0	0.2	2.40
	2000	887.7	1.73	95.5	3.4	1.73	95.5	1.0	1.23
	10000	4299.6	0.82	94.8	72.1	0.82	94.5	23.2	0.26
0.5	500	270.9	4.28	93.6	0.2	4.29	93.7	0.1	4.69
	1000	562.8	2.86	95.2	0.8	2.88	94.9	0.2	2.40
	2000	1140.7	2.01	95.5	3.5	2.01	95.6	1.0	1.23
	10000	5534.1	0.93	94.7	74.4	0.93	94.8	23.0	0.26

approximately follow a power-law distribution. This reflects the fact that in real social networks the majority of nodes have few followers but a small amount have a huge number of followers. To mimic this phenomenon, we simulate A as follows. First, generate nodal in-degree ($d_i = \sum_j a_{ji}$) from the discrete power-law distribution, i.e, $P(d_i = k) = ck^{-\alpha}$, where α is set to be 2.5 according to Clauset, Shalizi and Newman (2009). Lastly, for the ith node, we randomly select d_i nodes to be its followers.

4.2 Performance measurements and simulation results

Different network sizes are considered for each simulation example as $N = 500, 1000, 2000, 10000$. To obtain a reliable result, the experiment is replicated for $R = 1000$ times. In order to verify the conditions in Theorem 3 and Theorem 4, we have conducted some works on the three simulations and the verification results are shown in Appendix A.6 For each example, the average effective sample size (i.e., N^*) is reported. Let $\hat{\rho}^{(r)}$ be the estimator in the rth replication. In order to obtain $\hat{\rho}_{mle}$, Newton-Raphson method is adopted to maximize $\ell(\rho)$. The following measurements are considered to evaluate the finite sample performance. First, the root mean square error (RMSE) is calculated by $\text{RMSE} = \{R^{-1}\sum_{r=1}^{R}(\hat{\rho}^{(r)} - \rho)^2\}^{1/2}$ to gauge the estimation accuracy. Next, a 95% confidence interval is constructed for ρ as $\text{CI}^{(r)} = (\rho^{(r)} - z_{0.975}N^{*-1/2}\hat{\sigma}^{(r)}, \rho^{(r)} + z_{0.975}N^{*-1/2}\hat{\sigma}^{(r)})$, where z_α is the αth upper quantile of the standard normal distribution and $\hat{\sigma}^{2(r)}$ is the estimated asymptotic variance in the rth replication. The corresponding asymptotic variances of MLE and ME are given by conclusion of Theorem 3 and Theorem 4 respectively. Then, we compute the coverage probability as $\text{CP} = R^{-1}\sum_{r=1}^{R} I(\hat{\rho}^{(r)} \in \text{CI}^{(r)})$, where $I(\cdot)$ is the indicator function. In addition, the average computing time (CPU) measured in second and network density (ND, defined as $\sum_{i,j=1}^{N} a_{ij}/(N^2 - N)$) are reported.

The simulation results are summarized in Table 1 to Table 3. The patterns are quite similar across different networks structures, which indicate a robust performance of the proposed methods. Take the dyad independence network (i.e., Table 1) for example. It can be seen the RMSE is decreased for both MLE and ME as the effective sample size N^* increases, which corroborates the consistency results in Theorem 3 and Theorem 4. Particularly, although MLE is slightly more efficient than ME, one could see that the difference is sufficiently small (i.e., the RMSEs of MLE and ME are almost equivalent across all settings). This implies that the ME is almost as optimal as MLE. With regards to the computational time, ME is found to be faster, where the CPU time for ME is only one third of MLE. Lastly, one could see the CPs of both estimators are stable around 95%. This is consistent with the theoretical results in Theorem 3 and Theorem 4.

4.3 Short term and long term analysis

In this section, we conduct the short term and long term analysis in the simulation study. For a reliable evaluation, we repeat the procedure for $R = 1000$ times. In the rth round, given each type of network structure, we generate the responses $Y_{it}^{(r)}$ by the model (1). Next, we could calculate $\text{IDS}^{(r)}(t)$ by $\text{IDS}^{(r)}(t) = \sum_i Y_{it}$, the one step ahead prediction $E^*\{\text{IDS}^{(r)}(t)\}$ by (3), and two steps ahead prediction by (7) respectively. Then we could obtain $\overline{\text{IDS}}(t) = R^{-1}\sum_r \text{IDS}^{(r)}(t)$ and its average short term prediction as $R^{-1}\sum_r \text{EIDS}^{(r)}(t)$. To evaluate the prediction accuracy, we calculate the mean square error (MSE) of one step and two steps prediction. The results are shown in Figure 1, where the one step prediction has lower MSE than two step prediction. This corroborates with the theoretical analysis the short term analysis.

Next, we conduct a long term analysis by a simulation study. In this study, we generate the adjacency matrix A

Table 2. Simulation Results with 1000 Replications for stochastic block network. The Root Mean Square Error ($\times 10^{-2}$), Coverage Probability (%), CPU time (in second), and the Network Density (%) are reported

ρ	Sample Size		MLE			ME			
	N	N^*	RMSE	CP(%)	CPU	RMSE	CP(%)	CPU	ND(%)
0.1	500	158.7	2.44	91.1	0.2	2.44	91.1	0.1	0.98
	1000	350.3	1.61	94.0	0.7	1.61	93.9	0.3	0.38
	2000	687.4	1.15	94.8	3.0	1.15	94.6	0.9	0.16
	10000	3399.7	0.50	95.6	72.0	0.50	95.4	23.3	0.02
0.3	500	185.6	3.85	94.3	0.2	3.87	94.4	0.1	1.05
	1000	415.1	2.59	94.2	0.8	2.60	94.5	0.3	0.42
	2000	837.8	1.65	95.5	3.3	1.66	95.5	0.9	0.18
	10000	4081.6	0.77	95.6	81.7	0.77	95.6	23.4	0.03
0.5	500	287.3	3.65	95.1	0.2	3.67	95.6	0.1	0.98
	1000	609.1	2.63	94.5	0.8	2.66	94.0	0.2	0.46
	2000	1073.3	1.88	95.4	3.4	1.91	95.5	0.9	0.18
	10000	5532.8	0.76	96.1	85.4	0.78	95.4	24.2	0.02

Table 3. Simulation Results with 1000 Replications for power-law distribution network. The Root Mean Square Error ($\times 10^{-2}$), Coverage Probability (%), CPU time (in second), and the Network Density (%) are reported

ρ	Sample Size		MLE			ME			
	N	N^*	RMSE	CP(%)	CPU	RMSE	CP(%)	CPU	ND(%)
0.1	500	180.4	2.35	93.8	0.2	2.35	93.9	0.0	0.49
	1000	480.1	1.35	95.0	0.8	1.35	95.1	0.2	0.26
	2000	617.2	1.21	93.4	3.1	1.21	93.3	0.9	0.12
	10000	3169.3	0.53	94.9	71.6	0.53	94.8	23.5	0.02
0.3	500	182.8	3.67	94.6	0.2	3.69	94.4	0.1	0.50
	1000	449.0	2.27	94.6	0.8	2.30	94.5	0.2	0.23
	2000	939.6	1.57	95.0	3.4	1.58	94.8	0.9	0.12
	10000	4710.3	0.72	95.0	79.9	0.72	95.2	23.4	0.02
0.5	500	217.8	4.18	93.4	0.2	4.26	93.6	0.0	0.46
	1000	623.2	2.31	94.7	0.9	2.38	95.4	0.2	0.23
	2000	1510.0	1.49	95.3	3.4	1.51	95.5	0.9	0.12
	10000	5846.2	0.79	94.5	81.2	0.81	94.1	23.4	0.02

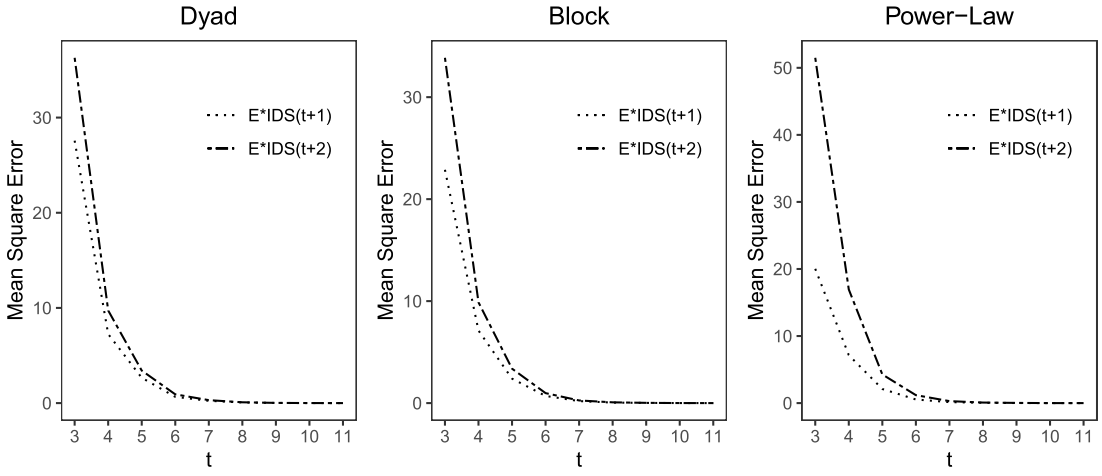

Figure 1. The MSE of one step and two step predictions for three network models, i.e., Dyad Independence Model, Stochastic Block Model, and Power-law Distribution Model.

by a symmetric stochastic block model. Then we verify the upper bound of the ϵ-*convergence time* given by Theorem 2. Set $\epsilon = 10^{-i}$, $i = 1, 2, \ldots, 100$, we calculate $t(\epsilon)$ and its upper bound in (9) for $R = 100$ repeated experiments. Then we report corresponding average values respectively in Figure 2. As shown by the Figure 2, the upper bound given by Theorem 2 is pretty tight for $t(\epsilon)$.

4.4 Diffusion of earthquake news on Sina Weibo

In this section, we apply the newly proposed network diffusion model (1) to a Sina Weibo dataset. Specifically, $N = 9{,}830$ users are collected from Sina Weibo's open API (open.weibo.com), and Sina Weibo is the largest Twitter type social network platform in China. Particularly, their corresponding Weibo posts are recorded for one week after the 7 magnitude of earthquake in Ya'an (a city in Sichuan province of China) on April 20th, 2013. At first, 100 users are randomly selected from overall users and their followees are added into the poll subsequently. After repeating tracking followees of newly added users four times, users who did not publish message containing Ya'an during certain time period are excluded. Finally, 9830 users are randomly selected to keep a reasonable size of the sample. A diffusion process among the network users is triggered immediately after an official account announced the earthquake news. In this section, we aim to estimate the network diffusion effect among users with regards to the diffusion of earthquake news.

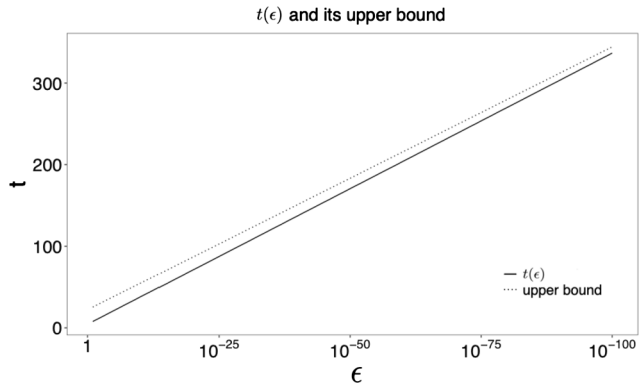

Figure 2. $t(\epsilon)$ *and its upper bound for Symmetric Stochastic Block Model.*

First, the adjacency matrix A is defined to be the following-followee relationship between users. The adjacency matrix A is asymmetric in this context since Sina Weibo does not restrict the users to mutually follow each other, i.e., it is allowed $a_{ij} \neq a_{ji}$. The resulting network density is 0.146%, which indicates a sufficiently sparse network. The histograms of nodal in-degrees and out-degrees are visualized in Figure 3, where the distribution of in-degrees are more skewed than out-degrees. The response variable Y_{it} is defined to be whether the ith user has re-tweeted the earthquake news on the tth day. For instance, user i could re-tweet news about the earthquake on the first and second day, but retweeted nothing on the third day. In this case, $Y_{i1} = Y_{i2} = 1$, while $Y_{i3} = 0$. To conduct a preliminary data clean procedure, it should be noted that the whole network structure is not available, therefore one might repost the weibo from some node outside our scope. To alleviate this problem, we add a "virtual" node to represent the outside nodes beyond the dataset, which is followed by all the nodes but has no followees. We then let the responses of this virtual node to be 1 over all $1 \leq t \leq T$. In this way, each node in our collected dataset can be equipped with a positive probability to repost news from the outside network.

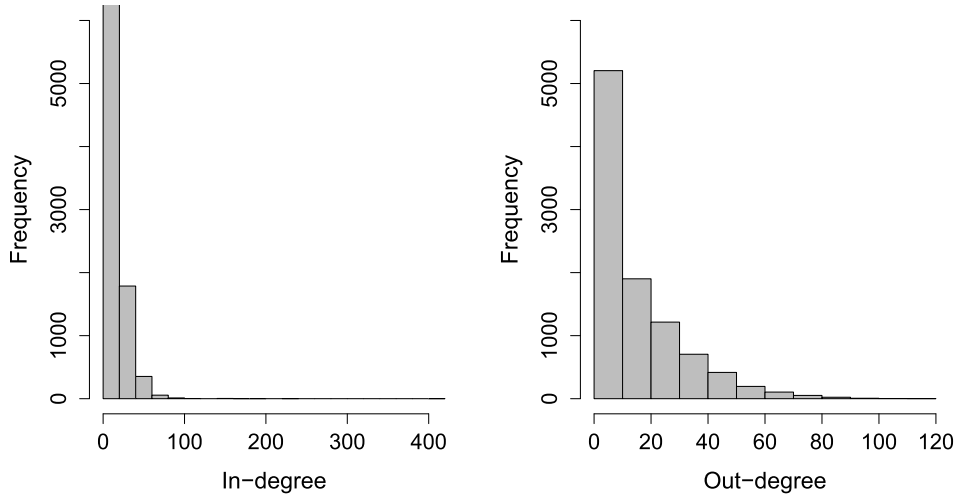

Figure 3. The left panel: histogram of nodal in-degree for $N = 9831$ *nodes. The highly right skewed shape can be detected; The right panel: histogram of nodal out-degree for* $N = 9831$ *nodes. Right skewed phenomenon can also be detected.*

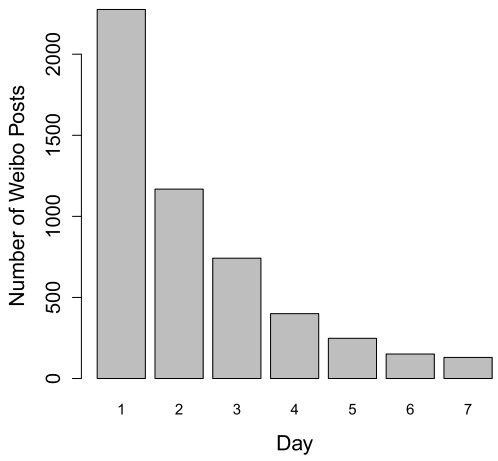

Figure 4. Number of earthquake related Weibo posts for 7 days after the occurrence of the earthquake. A sharply decreasing pattern can be captured.

To visualize the overall diffusion pattern, the total number of earthquake related Weibo posts are summarized in each day (i.e., $\sum_i Y_{it}$) in Figure 4. A sharply decreasing trend can be captured for the diffusion process at aggregate level. We then conduct estimation on the proposed network diffusion model. The estimated network diffusion effects (i.e., ρ) of MLE and ME are 0.093 and 0.096 respectively, which are almost identical to each other. Moreover, the effective sample size is estimated to be 4824.5, and both estimates are statistically significant under the 5% significance level.

5. CONCLUDING REMARKS

In this article, we study the diffusion process in large scale social networks. Particularly, a network diffusion model is proposed and investigated in both individual and aggregate level. Both the short term and the long term analysis of the network diffusion size are discussed. To estimate the parameter of the network diffusion effect, two estimators (i.e., MLE and ME) are proposed. In addition, their asymptotic properties are theoretically investigated and verified through extensive numerical studies. It is found the ME enjoys the same convergence rate as the MLE, and can be as almost optimal as MLE by the simulation results.

To conclude this work, we have the following discussions for potential future research topics. First, it is noteworthy that the network diffusion model (1) is in a linear form of the previous responses. However, in practice, various nonlinear modeling frameworks can be taken into consideration. Then, the analysis of the diffusion size and the parameter estimation method should be re-examined under the new framework. Next, although only the response information (i.e., Y_{it}) is involved in the network diffusion model, other exogenous variables (e.g., node specific covariates) can be employed to improve the model accuracy. Third, it can be restricted to assume the network diffusion parameter ρ to be the same for all the nodes in the network. As an alternative, it is more flexible to allow the parameter ρ to vary with different nodes, i.e., ρ_i $(1 \le i \le N)$, which could better capture the nodal heterogeneity across the network. In addition, we cannot numerically guarantee that the estimated $\widehat{\rho}_{mle}$ is within $(0, 1)$, especially when the true value of ρ is close to 1 (e.g., $\rho = 0.99$) and the sample size is not large. In such a case, one may use a thresholding technique to modify the estimator as $\widetilde{\rho}_{mle} = \min\{\widehat{\rho}_{mle}, \tau\}$, where τ could be set as $\tau = 0.99$. However, it might be problematic and we take this case as a future research study topic. Lastly, even though the structure of a network will not change significantly in a short time, it would be more precise and interesting to include this change by introducing A_t into the model.

APPENDIX A

A.1 Proof of (8)

The two steps ahead conditional variance can be calculated as follows:

$$
\begin{aligned}
&\mathrm{var}^*\Big\{\mathrm{IDS}(t+2)\Big\} = \mathrm{var}(\mathbf{1}^\top \mathbb{Y}_{t+2}|\mathcal{F}_t) \\
=&E\Big\{\mathrm{var}(\mathbf{1}^\top \mathbb{Y}_{t+2}|\mathcal{F}_{t+1})|\mathcal{F}_t\Big\} \\
&+ \mathrm{var}\Big\{E(\mathbf{1}^\top \mathbb{Y}_{t+2}|\mathcal{F}_{t+1})|\mathcal{F}_t\Big\} \\
=&E\Big\{\sum_{i=1}^{N} \rho w_i^\top \mathbb{Y}_{t+1}(1-\rho w_i^\top \mathbb{Y}_{t+1})|\mathcal{F}_t\Big\} \\
&+ \mathrm{var}(\mathbf{1}^\top \rho W \mathbb{Y}_{t+1}|\mathcal{F}_t) \\
=&E\Big\{\sum_{i=1}^{N} \rho w_i^\top \mathbb{Y}_{t+1}|\mathcal{F}_t\Big\} \\
&- E\Big\{\sum_{i=1}^{N} \rho^2 w_i^\top \mathbb{Y}_{t+1} w_i^\top \mathbb{Y}_{t+1})|\mathcal{F}_t\Big\} \\
&+ \rho^2 \mathbf{1}^\top W^\top \mathrm{var}(\mathbb{Y}_{t+1}|\mathcal{F}_t) \\
=&E^*\Big\{\mathrm{IDS}(t+2)\Big\} - \sum_{i=1}^{N} (p_{it}^{(2)})^2 \\
&+ \rho^2 \sum_{i=1}^{N} \sum_{1\le j\ne k\le N} w_{ji} w_{ki} p_{i(t+1)}(1-p_{i(t+1)}).
\end{aligned}
\tag{12}
$$

This completes the proof.

A.2 Proof of Theorem 1

By the assumption of irreducibility and aperiodicity, it could be concluded that there exists a stationary distribution π for the Markov chain by Levin and Peres (2017). As a result, we have $W^t \to \mathbf{1}\pi^\top$ as $t \to \infty$. Accordingly, it could be verified that $\mathrm{EIDS}(t) = \rho^t \mathbf{1}^\top \mathbf{1}\pi^\top \mathbb{Y}_0 = \rho^t N \pi^\top \mathbb{Y}_0$. Then, we obtain that (1) $\mathrm{EIDS}(t) \to 0$ if $0 \le \rho < 1$; (2)

EIDS$(t) = \mathbf{1}^\top W^t \mathbb{Y}_0 \to N\pi^\top \mathbb{Y}_0$ as $t \to \infty$. This completes the proof.

A.3 Proof of Theorem 2

Recall that the ϵ-convergence time is $t(\epsilon) = \min\{t : d(t) \le \epsilon\}$. In this part, we intend to derive an upper bound for $t(\epsilon)$. To this end, we first verify that there exists a stationary distribution $\pi = (\pi_1, \cdots, \pi_i)^\top \in \mathbb{R}^N$ for the Markov chain with transition probability matrix W. As a result, $W^t \to \mathbf{1}\pi^\top$. Second, we find an upper bound $d^*(t)$ for $d(t)$ that $d(t) \le d^*(t)$. Then the upper bound for $t(\epsilon)$ could be constructed by deriving the upper bound for $t^*(\epsilon) = \sup\{t : d^*(t) \le \epsilon\}$. Since $t(\epsilon) < t^*(\epsilon)$, the upper bound for $t(\epsilon)$ could be obtained.

First, to verify the stationary distribution of the Markov chain (with transition probability matrix W) exists, it suffices to show the Markov chain is reversible by Levin and Peres (2017) under the irreducibility assumption in Theorem 2. By setting $\pi_i = n_i/(\sum_{i=1}^N n_i)$, we have $\pi_i w_{ij} = \pi_j w_{ji} = a_{ij}/(\sum_{i=1}^N n_i)$ for any $i, j \in \{1, \cdots, N\}$. As a result, the Markov chain is reversible (Meyn and Tweedie, 2012). Further more, it could be verified $\pi = (\pi_1, \cdots, \pi_i)^\top$ is the stationary distribution vector.

Second, we construct an upper bound $d^*(t)$ for $d(t)$. To do this, we first write $W^t = (w_{ij}^{(t)}) \in \mathbb{R}^{N\times N}$. Since we have $0 \le Y_{0i} \le 1$, it can be obtained that $d(t) \le \rho^t \mathbf{1}^\top \Delta_W^{(t)} \mathbf{1} = \rho^t \sum_{i,j} \delta_{ij}^{(t)}$, where $\Delta_W^{(t)} = (\delta_{ij}^{(t)}) \in \mathbb{R}^{N\times N}$ with $\delta_{ij}^{(t)} = |w_{ij}^{(t)} - \pi_j|$. Furthermore, due to that $\sum_{1\le i,j\le N} \delta_{ij}^{(t)} \le N \max_i\{\sum_{j=1}^N \delta_{ij}^{(t)}\}$, we have $d(t) \le d^*(t) = \rho^t N \max_i\{\sum_{j=1}^N \delta_{ij}^{(t)}\}$. It could be easily verified that $d^*(t) \le \epsilon$ is equal to $\tilde{d}(t) = \max_i\{\sum_{j=1}^N \delta_{ij}^{(t)}\} \le \rho^{-t} N^{-1}\epsilon$. As a result, we have $t^*(\epsilon) = \sup\{t : d^*(t) \le \epsilon\} = \sup\{t : \tilde{d}(t) \le \rho^{-t} N^{-1}\epsilon\}$. Subsequently, by Levin and Peres (2017) (Proposition 4.2 and Theorem 12.3), we have

$$t^*(\epsilon_0) \le \log\Big(\frac{2}{\epsilon_0 \gamma_{\min}}\Big) \frac{1}{1-\lambda^*},$$

where $\epsilon_0 = \rho^{-t^*(\epsilon)} N^{-1}\epsilon$. It then could be easily calculated that

$$t^*(\epsilon_0) \le \frac{1}{1-\lambda^* - \log\rho} \log\frac{2N}{\epsilon\gamma_{\min}}.$$

Further note that $t(\epsilon) \le t^*(\epsilon)$, then (9) can be obtained. This completes the proof.

A.4 Proof of Theorem 3

Lemma 1. *Assume the conditions in Theorem 3. Then we have* $N^{*-1}\sum_{t=1}^T \sum_{i=1}^N \rho^{-2}(1-p_{it})^{-2} I(w_i^\top \mathbb{Y}_{t-1} > 0)(Y_{it} - p_{it})^2 \to_p \sigma_\rho^2$.

Proof. By the condition (2) in Theorem 3, we have $N^{*-1}\sum_{i=1}^N Z_i = N^{*-1}\sum_{t=1}^T\sum_{i=1}^N \rho^{-2}\{1 - p_{it}\}^{-2} I(w_i^\top \mathbb{Y}_{t-1} > 0) p_{it}(1-p_{it}) \to \sigma_\rho^2$. Then it suffices to show

$$N^{*-1}\sum_{i=1}^N\sum_{t=1}^T \Delta_{it} = o_p(1), \tag{13}$$

where $\Delta_{it} = \rho^{-2}(1-p_{it})^{-2} I(w_i^\top \mathbb{Y}_{t-1} > 0)\{(Y_{it} - p_{it})^2 - p_{it}(1-p_{it})\}$. It could be easily verified that $E(\Delta_{it}) = E[\rho^{-2}(1-p_{it})^{-2} E\{(Y_{it} - p_{it})^2 - p_{it}(1-p_{it})|\mathbb{Y}_{t-1}\}] = 0$ due to that $E\Big\{(Y_{it} - p_{it})^2 - p_{it}(1-p_{it})|\mathbb{Y}_{t-1}\Big\} = 0$.

Next, we are going to show that $N^{*-2}\text{var}(\sum_{i=1}^N \sum_{t=1}^T \Delta_i) \to 0$ as $N^* \to \infty$. For any $t_1 > t_2$ or $i_1 \ne i_2$, it could be verified that $\text{cov}(\Delta_{i_1 t_1}, \Delta_{i_2 t_2}) = E(\Delta_{i_1 t_1}\Delta_{i_2 t_2}) = E\{E(\Delta_{i_1 t_1}|\Delta_{i_2 t_2})\Delta_{i_2 t_2}\} = 0$. Then we have $\text{var}(\sum_{i=1}^N \sum_{t=1}^T \Delta_i) = \sum_{i=1}^N \sum_{t=1}^T \text{var}(\Delta_{it})$. By noting that $\text{var}(\Delta_{it}) = \text{var}\{E(\Delta_{it}|\mathbb{Y}_{t-1})\} + E\{\text{var}(\Delta_{it}|\mathbb{Y}_{t-1})\} = E\{\text{var}(\Delta_{it}|\mathbb{Y}_{t-1})\}$ due to that $E(\Delta_{it}|\mathbb{Y}_{t-1}) = 0$. Furthermore, it could be derived that $\text{var}(\Delta_{it}|\mathbb{Y}_{t-1}) \le p_{it}^{-4}(1-p_{it})^{-4} E(Y_{it} - p_{it}|\mathbb{Y}_{t-1})^4 = \rho^{-4}(1-\rho)^{-4}\Big\{(1-p_{it})^4 p_{it} + p_{it}^4(1-p_{it})\Big\} \le 2\rho^{-4}(1-\rho)^{-4} p_{it}$. As a result, we have $\sum_{i=1}^N \sum_{t=1}^T \text{var}(\Delta_{it}) \le 2\rho^{-4}(1-\rho)^{-4}\sum_{i=1}^N\sum_{t=1}^T E(p_{it}) = CN^*$, where $C = 2\rho^{-3}(1-\rho)^{-4}$ is a constant. Consequently, we have $N^{*-2}\sum_{i=1}^N\sum_{t=1}^T \text{var}(\Delta_{it}) \le CN^{*-1} \to 0$. Therefore, (13) could be obtained. □

Proof of Theorem 3. The asymptotic properties can be established in 2 steps. in the 1st step, we are going to show that $\hat{\rho}_{mle}$ is a $\sqrt{N^*}$-consistent local maximizer. In the 2nd step, we prove the asymptotic normality of the MLE estimator.

Step 1. Let $a_n = N^{*-1/2}$. We follow Fan and Li (2001) to show that, for any $\epsilon > 0$, there exists a constant $C > 0$ such that

$$P\Big\{\sup_{|u|=1} \ell(\rho + a_n u C) < \ell(\rho)\Big\} \ge 1 - \epsilon. \tag{14}$$

This implies that with probability at least $1-\epsilon$, there exists a local optimizer $\hat{\rho}$ in the ball $\{\rho + a_n u C : |u| \le 1\}$. As a result, we have $|\hat{\rho} - \rho| = O_p(a_n)$. To this end, we apply Taylor's expansion and obtain

$$\begin{aligned}&\sup_{|u|=1}\Big\{\ell(\rho + a_n u C) - \ell(\rho)\Big\}\\ &= \sup_{|u|=1}\Big\{C a_n \dot{\ell}(\rho) u - 2^{-1} C^2 a_n^2 u^2 \ddot{\ell}(\rho) + o_p(1)\Big\}\end{aligned}$$

$$\ge C|a_n\dot{\ell}(\rho)| - 2^{-1}C^2\{-a_n^2\ddot{\ell}(\rho)\} + o_p(1) \tag{15}$$

which is a quadratic function in C asymptotically. Next, note that $\dot{\ell}(\rho) = \sum_{i=1}^N \sum_{t=1}^T$

$I(w_i^\top \mathbb{Y}_{t-1} > 0)\rho^{-1}(1-p_{it})^{-1}(Y_{it}-p_{it})$. Then we have

$$E\{a_n\dot{\ell}(\rho)\} = a_n \sum_{i=1}^{N}\sum_{t=1}^{T} E\Big\{I(w_i^\top \mathbb{Y}_{t-1} > 0)\rho^{-1} (1-p_{it})^{-1}E(Y_{it}-p_{it}|\mathbb{Y}_{t-1})\Big\} = 0. \tag{16}$$

By the result of Lemma 1, it could be shown that $a_n\dot{\ell}(\rho)$ is $O_p(1)$. Similarly, by Lemma 1 we have $a_n^2\ddot{\ell}(\rho) = -a_n^2\sum_{i=1}^{N}\sum_{t=1}^{T}\rho^{-2}(1-p_{it})^{-2}I(w_i^\top \mathbb{Y}_{t-1}>0)(Y_{it}-p_{it})^2 = \sigma_\rho^{-2}\{1+o_p(1)\}$. This implies that the coefficient for the quadratic term in (15) is a positive constant asymptotically. As a result, the quadratic term should dominate the linear term as long as C is sufficiently large (Fan and Li, 2001). This proves the result of (14).

Step 2. By the first step, we know that $\hat{\rho}_{mle}$ is $\sqrt{N^*}$-consistent. This enables us to obtain the following asymptotic approximation

$$\sqrt{N^*}(\hat{\rho}_{\text{mle}} - \rho) = \frac{\Big\{N^{*-1/2}\dot{\ell}(\rho)\Big\}\Big\{1+o_p(1)\Big\}}{\Big\{N^{*-1}\ddot{\ell}(\rho)\Big\}}. \tag{17}$$

By Lemma 1, we know that $-N^{*-1}\ddot{\ell}(\rho) \to \sigma_\rho^2$. We next show that

$$N^{*-1/2}\dot{\ell}(\rho) \to_d N(0,\sigma_\rho^2). \tag{18}$$

Let $\xi_{it} = N^{*-1/2}I(w_i^\top \mathbb{Y}_{t-1}>0)\rho^{-1}(1-p_{it})^{-1}(Y_{it}-p_{it})$. Then we have $N^{*-1/2}\dot{\ell}(\rho) = \sum_{i=1}^{N}\sum_{t=1}^{T}\xi_{it}$. Define the sequence $\{Y_j^* : Y^*_{i+(t-1)N} = Y_{it}, 1\le j\le NT\}$ and $\{\xi_j^* : \xi^*_{i+(t-1)N} = \xi_{it}, 1\le j\le NT\}$. In addition, let $\mathcal{F}_j^*$ be a set generated by $\{Y_k^*, 1\le k\le j\}$. Then it could be easily verified that $E(\xi_j^*|\mathcal{F}_{j-1}^*) = 0$. Therefore, $\{\xi_j^*, \mathcal{F}_j^*, 1\le j\le NT\}$ is a martingale difference array (MDA). We then employ the central limit theorem for MDA (Hall and Heyde, 2014, Corollary 3.1) to prove (18).

To this end, it could be first easily verified $E(\xi_{it}^2) \le 4(N^*)^{-1}\rho^{-2}(1-\rho)^{-2} < \infty$ due to that $(1-p_{it})^{-1}\le(1-\rho)^{-1}$ and $|I(w_i^\top \mathbb{Y}_{t-1}>0)(Y_{it}-\rho w_i^\top \mathbb{Y}_{t-1})|\le 2$. Second, it could be calculated that

$$\begin{aligned}
&\sum_{j=1}^{NT} E\Big\{(\xi_j^*)^2|\mathcal{F}_{j-1}^*\Big\}\\
=&\sum_{i=1}^{N}\sum_{t=1}^{T} E\Big\{(\xi_{i+(t-1)N})^2|\mathcal{F}^*_{i+(t-1)N-1}\Big\}\\
=&(N^*)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\rho^{-1}(1-p_{it})^{-1}w_i^\top \mathbb{Y}_{t-1} \to_p \sigma_\rho^2
\end{aligned}$$

according to condition (2) in Theorem 3. Third, for any $\epsilon>0$, as $N^*\to\infty$

$$\begin{aligned}
&\sum_{j=1}^{NT} E\Big\{\xi_j^{*2}I(|\xi_j^*|>\epsilon)|\mathcal{F}_{j-1}^*\Big\} \le \epsilon^{-2}\sum_j E\Big\{\xi_j^{*4}|\mathcal{F}_{j-1}^*\Big\}\\
=&\epsilon^{-2}N^{*-2}\rho^{-4}(1-\rho)^{-4}\sum_{i=1}^{N}\sum_{t=1}^{T}E\Big\{(Y_{it}-p_{it})^4|\mathbb{Y}_{t-1}\Big\}\\
=&2\epsilon^{-2}N^{*-2}\rho^{-4}(1-\rho)^{-4}\sum_{i=1}^{N}\sum_{t=1}^{T}p_{it}\\
=&2\epsilon^{-2}\rho^{-3}(1-\rho)^{-4}N^{*-1}\to_p 0
\end{aligned}$$

due to that $0\le\{p_{it}^3+(1-p_{it})^3\}\le 2$ as $N^*\to\infty$. Thus, by the central limit theorem for MDA (Hall and Heyde, 2014, Corollary 3.1), (18) holds. This completes the proof. □

A.5 Proof of Theorem 4

By (11), $\hat{\rho}_{me}$ can be written as $\hat{\rho}_{me} = \rho + S_1^{-1}S_2$, where $S_1 = N^{*-1}\sum_{i=1}^{N}\sum_{t=1}^{T}w_i^\top \mathbb{Y}_{t-1}$ and $S_2 = N^{*-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-p_{it})$. As a result, the conclusion of Theorem 3 holds if

$$S_1 \to_p 1, \tag{19}$$

$$\sqrt{N^*}S_2 \to_d N(0,\sigma_1^2), \tag{20}$$

as $N^*\to\infty$. Subsequently, we prove (19) and (20) in Step 1 and Step 2 respectively.

Step 1. Proof of (19). In this step, it suffices to show that $E(S_i)/N^*\to 1$, and $\text{var}(S_1)/N^{*2}\to 0$. We first rewrite S_1 as $S_1 = N^{*-1}\sum_{i=1}^{N}\sum_{t=1}^{T}w_i^\top \mathbb{Y}_{t-1} = \sum_{t=1}^{T}\mathbf{1}^\top W\mathbb{Y}_{t-1}$. Then it could be calculated that $E(S_1) = \sum_{t=1}^{T}\rho^{t-1}\mathbf{1}^\top W^t\mathbb{Y}_0 = N^*$. As a result, we have $E(S_i)/N^* = 1$. Second, to prove $\text{var}(S_1)/N^{*2}\to 0$, it suffices to show $\text{var}(\mathbf{1}^\top W\mathbb{Y}_t)/N^{*2}\to 0$ for $t = 1,\cdots,T-1$ by Cauchy inequality. First note that $\text{var}(\mathbf{1}^\top W\mathbb{Y}_t) = \mathbf{1}^\top W\text{cov}(\mathbb{Y}_t)W^\top\mathbf{1}$. Next, for matrices $M_1 = (m_{ij}^{(1)})\in\mathbb{R}^{k_1\times k_2}$ and $M_2 = (m_{ij}^{(2)})\in\mathbb{R}^{k_1\times k_2}$, define $M_1\preccurlyeq M_2$ as $m_{ij}^{(1)}\le m_{ij}^{(2)}$ for $1\le i\le k_1$ and $1\le j\le k_2$. Then it can be verified that $\text{cov}(\mathbb{Y}_t) = \text{cov}\Big\{E(\mathbb{Y}_t|\mathbb{Y}_{t-1})\Big\} + E\Big\{\text{cov}(\mathbb{Y}_t|\mathbb{Y}_{t-1})\Big\}$

$$\begin{aligned}
&=\rho^2 W\text{cov}(\mathbb{Y}_{t-1})W^\top\\
&\quad + E\Big[\text{diag}\{\rho W\mathbb{Y}_{t-1}\circ(\mathbf{1}-\rho W\mathbb{Y}_{t-1})\Big\}\Big]\\
&\preccurlyeq\rho^2 W\text{cov}(\mathbb{Y}_{t-1})W^\top + \text{diag}\Big\{E(\rho W\mathbb{Y}_{t-1})\Big\}\\
&=\rho^2 W\text{cov}(\mathbb{Y}_{t-1})W^\top + \rho^t\,\text{diag}(W^t\mathbb{Y}_0),
\end{aligned} \tag{21}$$

where $\circ$ is the Hadamard product and the inequality is due to that W is elementwisely non-negative. Apply (21) iteratively then it could be obtained that $\text{cov}(\mathbb{Y}_t)\preccurlyeq\rho^t\,\text{diag}(W^t\mathbb{Y}_0)+\rho^{t+1}W\,\text{diag}(W^{t-1}\mathbb{Y}_0)W^\top+\cdots+\rho^{2t-1}W^{t-1}\,\text{diag}(W\mathbb{Y}_0)(W^\top)^{t-1}$. Recall $\nu = \sum_{t=1}^{T}\rho^t W^t\mathbb{Y}_0$

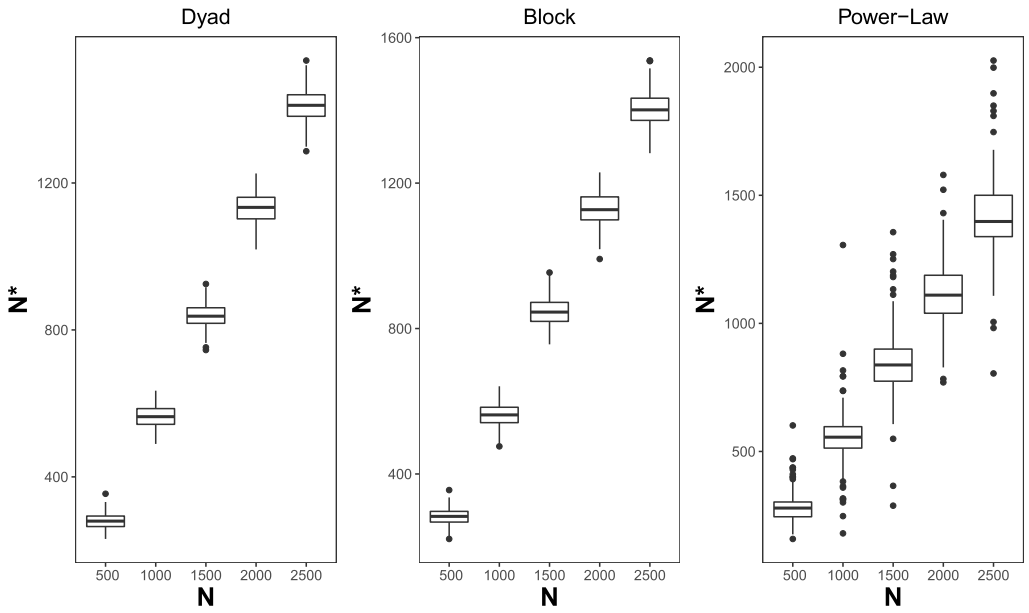

Figure 5. N^* *versus* N *in three models. The left panel for Dyad Independence Model. The middle panel for Stochastic Block Model. The right panel for Power-Law Model.*

and $\tilde{\nu} = (I(\nu_i > 0))$. Then we have $W^t\mathbb{Y}_0 \preccurlyeq \tilde{\nu}$. By condition (3) in Theorem 3, it could be derived that $(N^*)^{-2}\mathbf{1}^\top W^k \operatorname{diag}(W^{t-k}\mathbb{Y}_0)(W^\top)^k\mathbf{1} \leq (N^*)^{-2}\mathbf{1}^\top W^k \operatorname{diag}(\tilde{\nu})(W^\top)^k\mathbf{1} \to 0$ as $N^* \to \infty$.

Step 2. Proof of (20). Define $\zeta_{it} = (N^*)^{-1/2}(Y_{it} - \rho w_i^\top \mathbb{Y}_{t-1})$ and $\{\zeta_j^* : \zeta^*_{i+(t-1)N} = \zeta_{it}, 1 \leq j \leq NT\}$. Then we have $S_2 = \sum_{i=1}^N \sum_{t=1}^T \zeta_{it} = \sum_j \zeta_j^*$. Recall the sequence $\{Y_j^* : Y^*_{i+(t-1)N} = Y_{it}, 1 \leq j \leq NT\}$ and its corresponding set defined in Appendix A.3. Similar to the proof of Step 2 in Appendix A.3, it could be verified $E(\zeta_j^*|\mathcal{F}^*_{j-1}) = 0$. As a result, $\{\zeta_j^*, \mathcal{F}_j^*, 1 \leq j \leq NT\}$ is a martingale difference array (MDA) and the central limit theorem for MDA could be applied. To this end, first, it could be verified $E(\zeta_{it}^2) \leq 4(N^*)^{-1} < \infty$ due to that $|Y_{it} - \rho w_i^\top \mathbb{Y}_{t-1}| \leq 2$. Thus, $E(\zeta_j^*)^2 \leq 4(N^*)^{-1} < \infty$. Second, it could be calculated that $\sum_j E\left\{{\zeta_j^*}^2|\mathcal{F}^*_{j-1}\right\} = \sum_{i=1}^N \sum_{t=1}^T E\left\{{\zeta^*_{i+(t-1)N}}^2|\mathcal{F}^*_{i+(t-1)N-1}\right\} = N^{*-1}\sum_{i=1}^N \sum_{t=1}^T p_{it}(1-p_{it}) \to_p \sigma_1^2$ according to condition (2) in Theorem 3. Third, as $N^* \to \infty$, it holds that $0 \leq \{p_{it}^3 + (1-p_{it})^3\} \leq 2$ as $N^* \to \infty$. Then, for any $\epsilon > 0$, we have that $\sum_{j=1}^{NT} E\{{\zeta_j^*}^2 I(|\zeta_j^*| > \epsilon)|\mathcal{F}^*_{j-1}\} \leq$

$$\epsilon^{-2}\sum_j E\left\{{\zeta_j^*}^4|\mathcal{F}^*_{j-1}\right\}$$
$$= \epsilon^{-2}(N^*)^{-2}\sum_{i=1}^N\sum_{t=1}^T E\left\{I(w_i^\top\mathbb{Y}_{t-1} > 0)(Y_{it} - p_{it})^4|\mathbb{Y}_{t-1}\right\}$$
$$= \epsilon^{-2}(N^*)^{-2}\sum_{i=1}^N\sum_{t=1}^T p_{it}(1-p_{it})\left\{p_{it}^3 + (1-p_{it})^3\right\}$$
$$\leq 2\epsilon^{-2}(N^*)^{-2}\sum_{i=1}^N\sum_{t=1}^T p_{it}(1-p_{it}),$$

which converges to 0 in probability as $N^* \to \infty$. Thus, by the central limit theorem for MDA (Hall and Heyde, 2014, Corollary 3.1), (20) holds. This completes the proof.

A.6 Verifications of the conditions in Theorem 3 and Theorem 4

We devote this section to verify the assumptions in Theorem 3 and Theorem 4 for three simulation studies. We set different sample size $N = 500, 1000, 1500, 2000, 2500$ and generate Y_{i0} independently according to Bernoulli distribution with $p = 0.3$ for $1 \leq i \leq N$. Then, $\mathbb{Y}_t$s $(t = 1, \cdots, T)$ are randomly generated according to model (1). Lastly, for each example, we set the network diffusion effect $\rho = 0.5$. The experiment is randomly replicated for 200 times to obtain a reliable result.

A.6.1 Verification of conditions in Theorem 3

In Theorem 3, we assume that $N^* \to \infty$ as $N \to \infty$ and $N^{*-1}\sum_i Z_i \to_p \sigma_\rho^2$. To verify the condition, we calculate the corresponding N^* and $N^{*-1}\sum_i Z_i$. The detailed information is given in Figure 5 and Figure 6. A clear pattern for $N^* \to \infty$ as $N \to \infty$ and a convergence pattern for $N^{*-1}\sum_i Z_i \to_p \sigma_\rho^2$ could be observed.

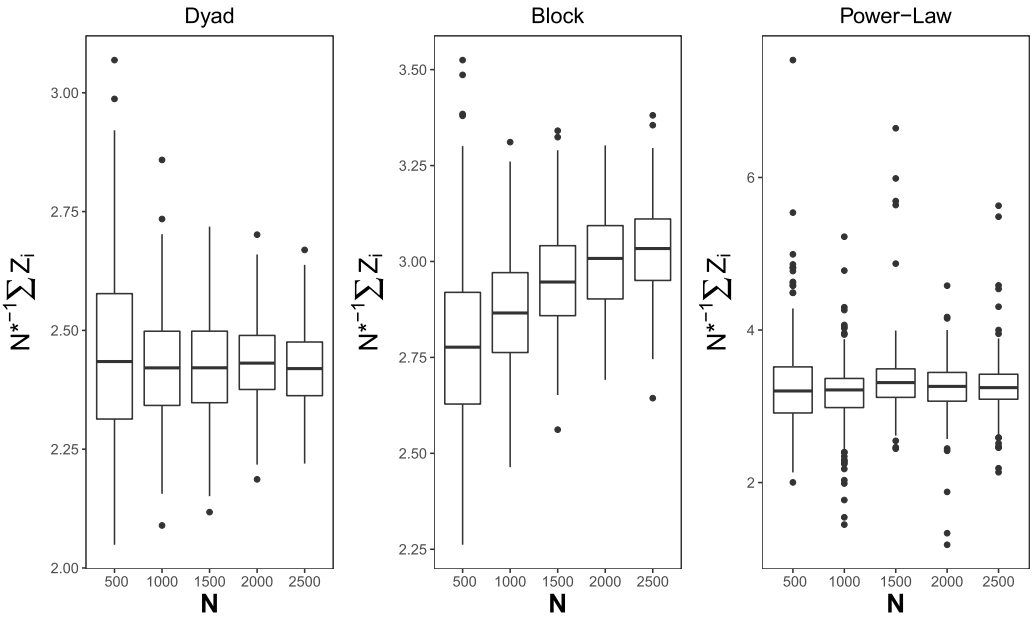

Figure 6. $N^{*-1}\sum_i Z_i$ *versus* N *in three models. The left panel for Dyad Independence Model. The middle panel for Stochastic Block Model. The right panel for Power-Law Model.*

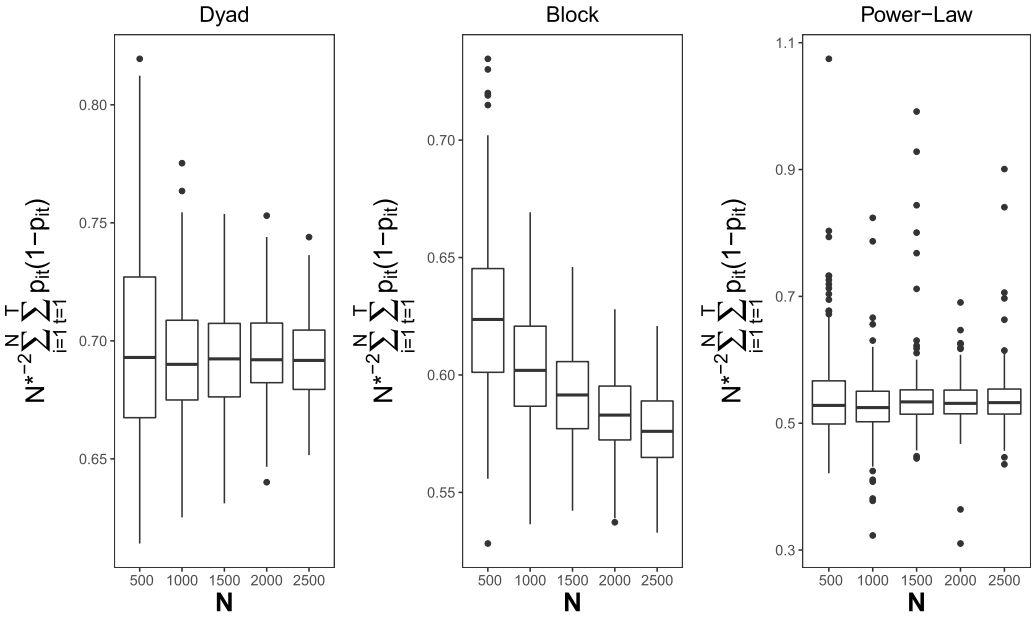

Figure 7. $(N^*)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{p_{it}(1-p_{it})\right\}$ *versus* N *in three models. The left panel for Dyad Independence Model. The middle panel for Stochastic Block Model. The right panel for Power-Law Model.*

A.6.2 Verification of conditions in Theorem 4

In Theorem 4, the verification of Assumption (1) is the same as the Assumption (1) in Theorem 3. Then, we verify the Assumption (2) and (3) in Theorem 4 respectively. For Assumption (2), the convergence pattern of $(N^*)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left\{p_{it}(1-p_{it})\right\}$ is shown in Figure 7. Next, we verify the Assumption (3) and draw the result in Figure 8. As one could observe, the sequence $(N^*)^{-2}\sum_{t=1}^{T}\sum_{j\in J}(\sum_i w_{ij}^{(t)})^2$ shows a clear convergence pattern to zero.

Received 1 July 2019

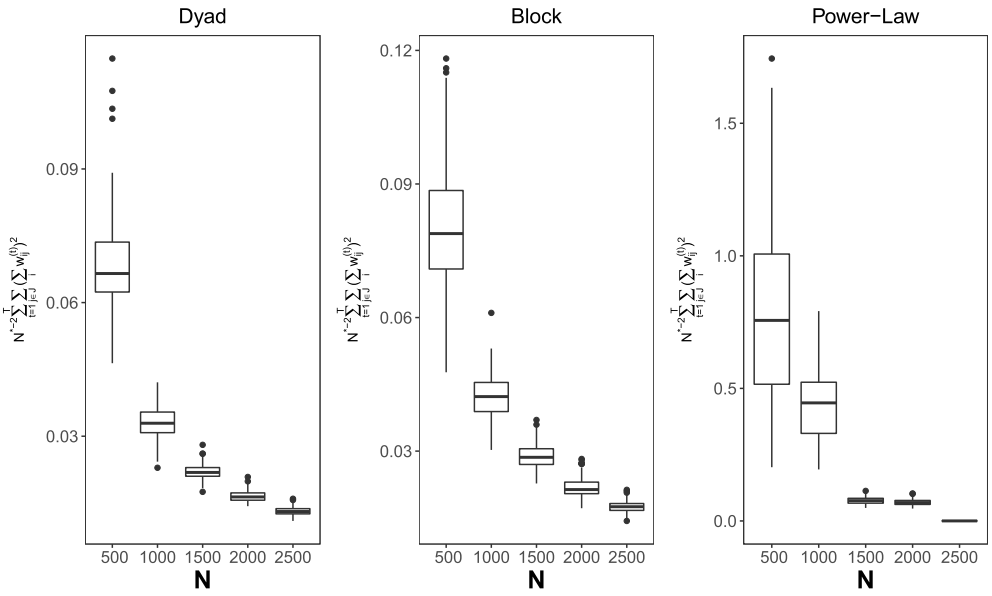

Figure 8. $(N^*)^{-2}\sum_{t=1}^{T}\sum_{j\in J}(\sum_i w_{ij}^{(t)})^2$ *versus* N *in three models. The left panel for Dyad Independence Model. The middle panel for Stochastic Block Model. The right panel for Power-Law Model.*

REFERENCES

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. CRC Press. MR3362184

Bass, F. M. (1969). A new product growth for model consumer durables. *Management Science* **15** 215–227.

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

Du, N., Liang, Y., Balcan, M. and Song, L. (2014). Influence function learning in information diffusion networks. In *International Conference on Machine Learning* 2016–2024.

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. MR1946581

Gao, S., Pang, H., Gallinari, P., Guo, J. and Kato, N. (2017). A novel embedding method for information diffusion prediction in social network big data. *IEEE Transactions on Industrial Informatics* **13** 2097–2105.

Goldenberg, J., Libai, B. and Muller, E. (2001). Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* **12** 211–223.

Goswami, D. (2001). Stochastic evolution of innovation diffusion in heterogeneous groups: Study of life cycle patterns. *IMA Journal of Management Mathematics* **12** 107–126.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press. MR0624435

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. MR0608176

Jun, D. B. and Park, Y. S. (1999). A choice-based diffusion model for multiple generations of products. *Technological Forecasting and Social Change* **61** 45–58.

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.

Kou, S., Olding, B. P., Lysy, M. and Liu, J. S. (2012). A multiresolution method for parameter estimation of diffusion processes. *Journal of the American Statistical Association* **107** 1558–1574. MR3036416

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times* **107**. American Mathematical Soc. MR3726904

Li, D., Zhang, S., Sun, X., Zhou, H., Li, S. and Li, X. (2017). Modeling information diffusion over social networks for temporal dynamic prediction. *IEEE Transactions on Knowledge and Data Engineering* **29** 1985–1997.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media. MR2509253

Niu, S.-C. (2002). A stochastic formulation of the Bass model of new-product diffusion. *Mathematical Problems in Engineering* **8** 249–263. MR1945003

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

Rogers, E. M. (2010). *Diffusion of innovations*. Simon and Schuster.

Schmittlein, D. C. and Mahajan, V. (1982). Maximum likelihood estimation for an innovation diffusion model of new product acceptance. *Marketing Science* **1** 57–78.

Srinivasan, V. and Mason, C. H. (1986). Nonlinear least squares estimation of new product diffusion models. *Marketing Science* **5** 169–178.

Van den Bulte, C. and Stremersch, S. (2004). Social contagion and income heterogeneity in new product diffusion: A meta-analytic test. *Marketing Science* **23** 530–544.

Wang, F., Wang, H. and Xu, K. (2012). Diffusive logistic model towards predicting information diffusion in online social networks. In *2012 32nd International Conference on Distributed Computing Systems Workshops* 133–139. IEEE.

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

Wasserman, S., Faust, K. et al. (1994). *Social network analysis: Methods and applications* **8**. Cambridge University Press.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393** 440.

Yang, J. and Leskovec, J. (2010). Modeling information diffusion in implicit networks. In *2010 IEEE International Conference on Data Mining* 599–608. IEEE.

Zhang, X., Su, Y., Qu, S., Xie, S., Fang, B. and Philip, S. Y. (2018). IAD: Interaction-aware diffusion framework in social networks. *IEEE Transactions on Knowledge and Data Engineering* **31** 1341–1354.
Zhao, Y., Levina, E., Zhu, J. et al. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40** 2266–2292. MR3059083
Zhou, Z., Bandari, R., Kong, J., Qian, H. and Roychowdhury, V. (2010). Information resonance on Twitter: watching Iran. In *Proceedings of the First Workshop on Social Media Analytics* 123–131.
Zhu, X., Pan, R., Li, G., Liu, Y., Wang, H. et al. (2017). Network vector autoregression. *The Annals of Statistics* **45** 1096–1123. MR3662449

Xuening Zhu
School of Data Science
Fudan-Xinzailing Joint Research Centre for Big Data
Fudan University
Shanghai
China
E-mail address: xueningzhu@fudan.edu.cn

Rui Pan
School of Statistics and Mathematics
Central University of Finance and Economics
Beijing
China
E-mail address: panrui_cufe@126.com

Yuxuan Zhang
Department of Statistics
University of California, Davis
Davis, CA
USA
E-mail address: yuxzh@ucdavis.edu

Yu Chen
Guanghua School of Management
Peking University
Beijing
China
E-mail address: yu.chen@pku.edu.cn

Wenquan Mi
School of Data Science
Fudan-Xinzailing Joint Research Centre for Big Data
Fudan University
Shanghai
China
E-mail address: venturerice@163.com

Hansheng Wang
Guanghua School of Management
Peking University
Beijing
China
E-mail address: hansheng@pku.edu.cn