Advances in Intelligent Systems Research, volume 153
International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Localization of Diffusion Source in Networks With the Noisy Data

Tiantian Li and Lu Niu*
School of Mathematics and System Science, Beihang University, China
*Corresponding author

***Abstract*—Locating diffusion source in network is an important issue in network data analysis. Many methods have been developed. However, noiseless assumption used in the literature is restrictive and the methods are not robust enough. In this paper, we consider the problem of locating diffusion source in networks with the noisy presented. Since the sample size is much smaller than the dimension of unknown parameters, the Lasso method is used to identify the locating diffusion source in networks. Simulation results confirm the effectiveness of our method.**

Keywords-localization; spreading nodes; lasso method; noisy data

I. INTRODUCTION

Dynamical process is commonly taken place in complex network such as epidemic spreading in the human society [1], [2]. Identifying the spreading source is a very important issue in network data analysis. Many works have been developed on this issue, such as maximum-likelihood estimation [3], belief propagation [4], the phenomena of hidden geometry of contagion [5] and inverse spreading [6]. Satterlee and Penman [7] studied some simple models of disease transmission on small-world networks. Balcan et al. [8] gave a computational model of infectious diseases. These researches found the important sources in the networks. So how to locate the source exactly in network is a challenging problem. Yuan et al. [9] considered the exact controllability of the network. Shen et al. [10] considered a general locatability framework for sources localization in complex networks by time-reversal backward spreading. Moreover, vital nodes identification [11], [12], [13] attracted great attentions in different complex networks. Hu et al. [14] considered the localization of diffusion sources in complex network, where a theoretically optimal algorithm has been proposed to identify the source node. However, this algorithm is actually solved a nonconvex optimization problem and was computational heavily when the number of the nodes was large. Then the authors proposed a compress sensing CS method to identify the source node, which could be solved efficiently.

We briefly review this CS based approach. Let N be the number of the nodes in the network. Variable $x_i(t)$ denotes the state of node i at time t, which describes the concentration of water or air pollutant, etc. Denote β the diffusion coefficient and W_{ij} the weight of the directed link from node j to node i. Specifically $w_{ij} = w_{ji}$ for undirected networks. Let $W = (w_{ij}) \in R^{N\times N}$ be the weighted adjacent matrix and $D = \mathrm{diag}(d_1, \cdots, d_N) \in R^{N\times N}$ be the diagonal matrix with $d_i = \sum_{j\in N_i} w_{ij}$, where N_i denote the neighborhood of node i. Denote $L = W - D$ the Laplacian matrix. Define

$$\begin{cases} \mathbf{x}(t+1) = (I + \beta L)\mathbf{x}(t) \\ \mathbf{y}(t+1) = C\mathbf{x}(t) \end{cases} \tag{1}$$

where $\mathbf{x}(t) = (x_1(t), \cdots, x_N(t))$. The vector $\mathbf{y}(t) \in R^q$ is the output at time t and $C \in R^{q\times N}$ is the output matrix. The matrix C satisfies the observability rank condition [9] and can be determined by the algorithm of Hu et al [14].

Denote t_0 the initial time of the spreading. Then

$$\mathbf{x}(t) = (I + \beta L)^{t-t_0}\mathbf{x}(t_0)$$

and consequently

$$\mathbf{y}(t) = C(I + \beta L)^{t-t_0}\mathbf{x}(t_0)$$

Generally, to obtain a unique solution, no fewer than N snapshots of measurement are needed. Assume that uninterrupted time series from t_0 to $t_0 + N + 1$ are available. Then one has the equation

$$\mathrm{Y} = O\mathbf{x}(t_0) \tag{2}$$

where

$$\mathrm{Y} = \begin{pmatrix} \mathbf{y}(t_0) \\ \mathbf{y}(t_0+1) \\ \vdots \\ \mathbf{y}(t_0+N-1) \end{pmatrix} \in R^{qN}, \quad O = \begin{pmatrix} C \\ C[I+\beta L] \\ \vdots \\ C[I+\beta L]^{N-1} \end{pmatrix} \in R^{qN\times N}$$

Given t_0, CS method can be applied to recover the sparse signal $\mathbf{x}(t_0)$, based on the equation (2).

In practice, the observations often contain many noise or measurement error. Recall that $\mathbf{x}(t)$ denotes the true state of nodes at time t. However, in practice, it is inevitable that the observation $\mathbf{y}(t)$ contains some noisy or measurement error. Therefore, it is more reasonable to consider the model

$$\mathbf{y}(t) = C\mathbf{x}(t) + \epsilon(t)$$

where $\epsilon(t) \in R^q$ stands for the independent noisy with mean zero and unknown variance.

The classical CS methods aim to recover a sparse signal for the case of no noise and consequently are well suited for the model (2). When there is measurement error or noise presented, CS method is not applicable any more. In this paper, we propose a Lasso method to solve the problem.

II. IDENTIFY THE SOURCE DIFFUSION BY LASSO WHEN MEASUREMENT ERROR IS PRESENTED

A. *Review of Lasso*

Lasso proposed by Tibshirani [15] is a statistical method that can be used to recover the sparse signal for a linear model

$$y_i = x_i{}^T \beta + \epsilon_i, \quad 1 \le i \le n$$

where $\{(y_i, x_i), 1 \le i \le n\}$ are n observations with $x_i \in R^p$ and $y_i \in R^1$ and $\beta = (\beta_1, \cdots, \beta_p)^T \in R^p$ is the unknown sparse vector, while ϵ_i denotes the random error. Here p can be much large than n. Write the model in matrix form, we have

$$\mathbb{Y} = \mathbb{X}\beta + \epsilon$$

where $\mathbb{Y} = (Y_1, \cdots, Y_n)^T \in R^n, \mathbb{X} = (x_1, \cdots, x_n)^T \in R^{n \times p}$ and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \in R^n$. The goal is to estimate the sparse coefficient β. Let $Q(\beta) = n^{-1}\sum_{i=1}^n (y_i - x_i{}^T \beta)^2$ and $\|\beta\|_1 = \sum_{j=1}^p |\beta_i|$.

Lasso estimates β by minimizing the following objective function

$$\hat{\beta} = \arg\min_{\beta \in R^p} Q(\beta) + \lambda\|\beta\|_1.$$

Here λ is the tuning parameter and can be calculated by a modification of the LARS algorithm (Tibshirani (2004)). In addition, the estimator $\hat{\beta}$ is generally sparse.

B. *Localization of Diffusion Sources by Lasso*

When the noises are presented, similar to (2), we have the model

$$\begin{cases} \mathrm{x}(t+1) = (I + \beta L)\mathrm{x}(t) \\ \mathrm{y}(t+1) = C\mathrm{x}(t) + \epsilon(t) \end{cases} \tag{3}$$

Then $\mathrm{x}(t) = (I + \beta L)^{t-t_0}\mathrm{x}(t_0)$ and consequently $\mathrm{y}(t) = C(I + \beta L)^{t-t_0}\mathrm{x}(t_0) + \epsilon(t)$. Since Lasso can handle the case where the sample size is much small than the dimension of unknown parameter. Therefore, the number of observations can be significantly reduced. Therefore, snapshot of the network can be much less than N and requirement on C can also be relaxed. In this paper, we assume that t_0 and β are known. Given the observation at time point $t_0, t_0 + 1, \cdots, t_0 + n - 1$, where n can be much small than N, we have the model

$$\widetilde{\mathbb{Y}}_{t_0,n} = \widetilde{\mathbb{X}}_n \cdot \mathrm{x}(t_0) + \varepsilon_{t_0,n} \tag{4}$$

where

$$\widetilde{\mathbb{Y}}_{t_0,n} = \begin{pmatrix} \mathrm{y}(t_0) \\ \mathrm{y}(t_0+1) \\ \vdots \\ \mathrm{y}(t_0+n-1) \end{pmatrix} \in R^{qn}, \quad \widetilde{\mathbb{X}}_n = \begin{pmatrix} C \\ C[I+\beta L] \\ \vdots \\ C[I+\beta L]^{n-1} \end{pmatrix} \in R^{qn \times N}$$

and $\varepsilon_{t_0,n} = (\epsilon(t_0), \cdots, \epsilon(t_0 + n - 1))^T$.

Then we can apply Lasso method to estimate the sparse vector $\mathrm{x}(t_0)$, minimizing the objective function

$$\hat{\mathrm{x}}(t) = \arg\min_{\gamma \in R^N} \tilde{Q}_{t_0,n}(\gamma) + \lambda\|\gamma\|_1 \tag{5}$$

where $\tilde{Q}_{t_0,n}(\gamma) = n^{-1}\left\|\widetilde{\mathbb{Y}}_{t_0,n} - \widetilde{\mathbb{X}}_n \gamma\right\|^2$.

III. SIMULATION

To illustrate the performance of our diffusion source localization in networks with the noisy data, we consider different kinds of unweighted networks. Because of the lack of controllable noise in the real networks, we set parameters and build ER networks. We use a standard index, receiver operating characteristic curve (ROC curve), to test the performance. The larger the area under the curve(AUC), the better the performance. AUC = 1 means that the initial messenger nodes can be found entirely.

We generate the data from the model (4) with $\mathrm{x}(t_0) = (1, \cdots, 1, 0, \cdots, 0)$, where the number of 1 depends on the value of M.

There are 5 parameters.

p: With probability p, we connect each pair of nodes;

N: number of nodes;

M: number of source spreading nodes;

n: number of observed nodes;

δ: parameter controls the diffusion coefficient β.

$$\beta = \frac{1}{\left(\delta + \max\left(\sum_{i=1, i \ne i}^N W_{ij}\right)\right)}, \qquad i \in 1, 2, \cdots, N$$

In the following figures, for each setting, we repeat T = 200 times to compute the average TPR and FPR and plot the ROC curve.

A. *Effect of* p

We set parameter p to control the sparsity of the networks. As shown in Figure I, for p = 0.1,0.2,0.3,0.4 and 0.5, there are slightly differences of the ROC and the AUC remained above 0.9. In the case of sparse network, the sparse degree does not affect the effectiveness of our method.

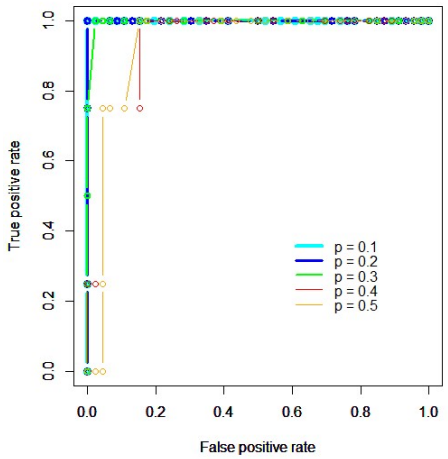

FIGURE I. EFFECT OF p. WE CHANGE p = 0.1,0.2,0.3,0.4 AND 0.5. AND SET N = 50, M = 4, n = 10, δ = 1. THE RESULTS ARE OBTAINED BY OVER 200 SIMULATIONS.

B. *Effect of the Number of Nodes N*

We build network with different number of nodes, e.g., 50,100,150,200. Figure II shows that smaller values of N results in higher values of AUC. With the same amount of the initial messenger nodes, larger network is sparser than the smaller one.

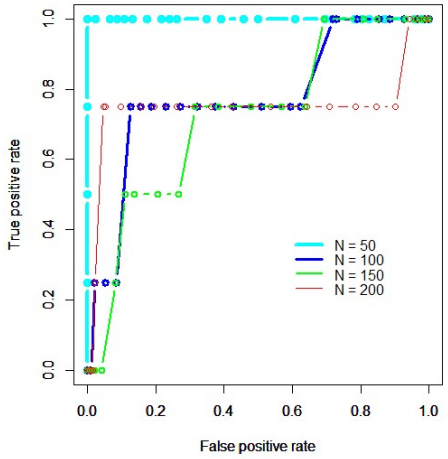

FIGURE II. EFFECT OF N. WE CHANGE N = 50, 100, 150 AND 200. AND SET p = 0.4. THE OTHER PARAMETERS ARE THE SAME AS IN FIGURE I. THE RESULTS ARE OBTAINED BY OVER 200 SIMULATIONS.

C. *Effect of the Source Spreading Nodes M*

We set the amount of the source spreading nodes M = 1,2,5,10. In Figure III, it shows the value of AUC changes slightly if M is equal to 1,2,5 or 10. And the value of AUC always stays at a higher level.

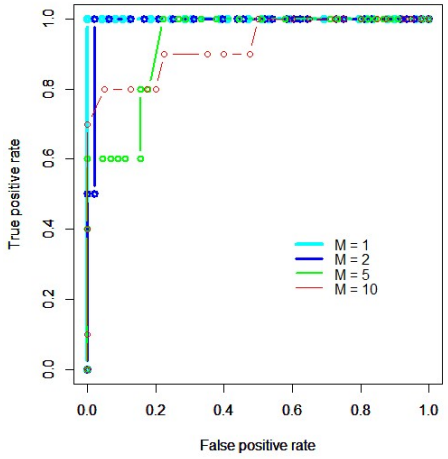

FIGURE III. EFFECT OF M. WE CHANGE M = 1, 2, 5 AND 10. AND SET p = 0.4. THE OTHER PARAMETERS ARE THE SAME AS IN FIGURE I. THE RESULTS ARE OBTAINED BY OVER 200 SIMULATIONS.

D. *Effect of the Observed Nodes n*

As we can see in Figure IV, the value of the AUC increases when the number of the observed nodes increases. With twenty percent of the nodes in network are observed, the accuracy of the diffusion source localization is close to 1.

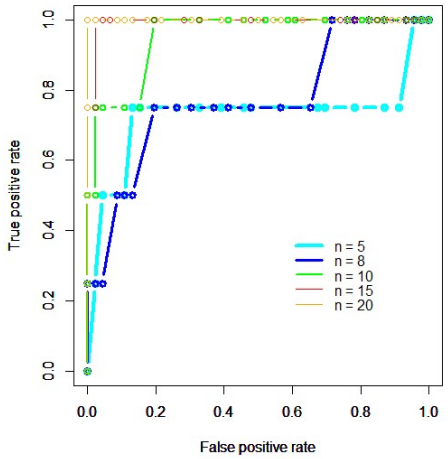

FIGURE IV. EFFECT OF n. WE CHANGE n = 5, 8, 10, 15 AND 20. AND SET p = 0.4. THE OTHER PARAMETERS ARE THE SAME AS IN FIGURE I. THE RESULTS ARE OBTAINED BY OVER 200 SIMULATIONS.

E. *Effect of the Diffusion Coefficient*

In Figure V, we test the influence of the diffusion coefficient δ with controlling the parameter β. We find that the value of AUC is always close to 1. So we think the diffusion coefficient does not affect the source localization much in network.

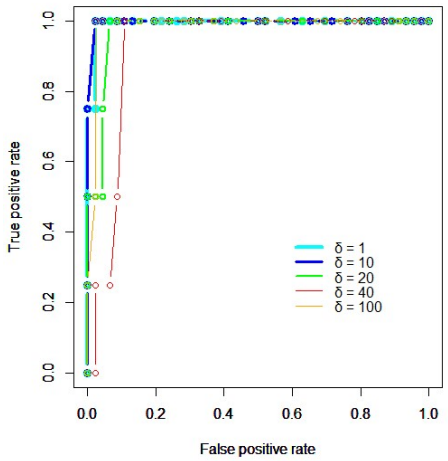

FIGURE V. EFFECT OF δ. WE CHANGE δ = 1, 10, 20, 40 AND 100. AND SET p = 0.4. THE OTHER PARAMETERS ARE THE SAME AS IN FIGURE I. THE RESULTS ARE OBTAINED BY OVER 200 SIMULATIONS.

IV. DISCUSSION

We propose a model for locating diffusion source in networks with noisy data, which is commonly encountered in application. And the Lasso method is used to identify the diffusion resource. Simulation results show that the proposed method works well under different settings. However, here we only consider the case of t_0 and β are known. When t_0 and β are unknown, how to estimate them is problem for future study.

ACKNOWLEDGMENT

We thank Prof. Junlong Zhao for helpful discussion and suggestions.

REFERENCES

[1] Neumann G, Noda T, Kawaoka Y. Emergence and pandemic potential of swine-origin H1N1 influenza virus[J]. Nature, 2009, 459(7249):931-9.

[2] Hvistendahl M, Cohen J. Influenza. Despite large research effort, H7N9 continues to baffle[J]. Science, 2013, 340(6131):414-5.

[3] Pinto P C, Thiran P, Vetterli M. Locating the source of diffusion in large-scale networks[J]. Physical Review Letters, 2012, 109(6):068702.

[4] Altarelli F, Braunstein A, Dall'Asta L, et al. Bayesian inference of epidemics on networks via belief propagation[J]. Physical Review Letters, 2014, 112(11):118701.

[5] Brockmann D, Helbing D. The hidden geometry of complex, network-driven contagion phenomena[J]. Science, 2013, 342(6164):1337-1342.

[6] Zhu K, Ying L. Information Source Detection in the SIR Model: A Sample-Path-Based Approach[J]. IEEE/ACM Transactions on Networking, 2012, 24(1):408-421.

[7] Satterlee C, Penman D. Method and apparatus for locating source of error in high-speed synchronous systems: U.S. Patent 5,383,201[P]. 1995-1-17.

[8] Balcan D, Colizza V, Gonçalves B, et al. Multiscale mobility networks and the spatial spreading of infectious diseases[J]. Proceedings of the National Academy of Sciences, 2009, 106(51): 21484-21489.

[9] Yuan Z, Zhao C, Di Z, et al. Exact controllability of complex networks[J]. Nature Communications, 2013, 4.

[10] Shen Z, Cao S, Wang W X, et al. Locating the source of diffusion in complex networks by time-reversal backward spreading[J]. Physical Review E, 2015, 93(3-1):032301.

[11] Kitsak M, Gallos L K, Havlin S, et al. Identification of influential spreaders in complex networks[J]. Nature Physics, 2010, 6(11):888-893.

[12] Pei S, Makse H A. Spreading dynamics in complex networks[J]. 2013(12):131-136.

[13] Lü L, Chen D, Ren X L, et al. Vital nodes identification in complex networks[J]. Physics Reports, 2016, 650:1-63.

[14] Hu Z L, Han X, Lai Y C, et al. Optimal localization of diffusion sources in complex networks[J]. Royal Society Open Science, 2017, 4(4):170091.

[15] Tibshirani R. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society, 2011, 73(3):273-282.