Physics Letters A ••• (••••) •••–•••

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Localization of diffusion sources in complex networks with sparse observations

Zhao-Long Hu [a], Zhesi Shen [b], Chang-Bing Tang [a], Bin-Bin Xie [c], Jian-Feng Lu [a]

[a] College of Mathematics, Physics and Information Engineering, Zhejiang Normal University, Jinhua 321004, Zhejiang, PR China
[b] National Science Library, Chinese Academy of Sciences, Beijing 100190, PR China
[c] Hangzhou Institute of Advanced Studies, Zhejiang Normal University, Hangzhou 311231, Zhejiang, PR China

A R T I C L E I N F O

Article history:
Received 18 December 2017
Received in revised form 27 January 2018
Accepted 29 January 2018
Available online xxxx
Communicated by C.R. Doering

Keywords:
Complex networks
Source localization
Sparse observations
Diffusion process
Integer programming

A B S T R A C T

Locating sources in a large network is of paramount importance to reduce the spreading of disruptive behavior. Based on the backward diffusion-based method and integer programming, we propose an efficient approach to locate sources in complex networks with limited observers. The results on model networks and empirical networks demonstrate that, for a certain fraction of observers, the accuracy of our method for source localization will improve as the increase of network size. Besides, compared with the previous method (the maximum–minimum method), the performance of our method is much better with a small fraction of observers, especially in heterogeneous networks. Furthermore, our method is more robust against noise environments and strategies of choosing observers.

1. Introduction

Diffusion is a common phenomenon involved in many complex systems [1]. Some diffusion processes might incur huge losses to our society. Such examples include spread of rumors or infectious diseases [2–5], virus invasion in computer networks [6,7], and diffusion of air or water pollution [8,9]. Under these circumstances, being able to swiftly and accurately identify the origins of the diffusion processes is critically important and proper control methods must be devised to contain such harmful events.

Despite some recent developments in the field, locating diffusion sources has remained a difficult task. The challenge arises from the fact that the information obtained about the diffusion processes is either incomplete or "contaminated". For example, both the number of sources and the exact time when diffusion first occurs could be unknown. Based on various propagation models, several algorithms have been proposed, which include the maximum likelihood estimation [10–14], dynamic message passing [15], belief propagation [16], hidden geometry of contagion [17] and inverse diffusion [18,19]. These algorithms make the assumption of single-source localization. However, in complex systems, diffusion is oftentimes stemming from multiple sources [20–23]: Several individuals, for example, may be responsible for spreading the same malicious information. Therefore, developing a method to locate multiple sources is of great practical value. Although some methods have been proposed for the purpose [24–26], they require the specification of a snapshot of the states of all nodes involved, which is impractical for real large networks. In addition, the computational complexity would grow exponentially with the increase in the number of sources.

With a limited fraction of observers, Altarelli et al. has proposed an efficient belief propagation algorithm to perform source localization with the unknown the initial spreading time [16]. However, the performance can't be guaranteed if given the presence of numerous cyclic structures in the graph. Wang et al. put forward an iterative algorithm that makes no assumptions about the underlying propagation model. One problem with this method is that sometimes the number of sources is difficult to infer [27]. In a recent study [28], with the help of observability theory and compressive sensing, the problem of multi-source localization with minimum messengers was addressed. This general method, however, can only be applied to linear diffusion processes. A novel approach, i.e., the maximum–minimum method (MMM), was proposed to locate multi-sources with partial observations [29]. The accuracy of the algorithm, however, is adversely affected by extreme values, particularly in the case of a small fraction of observers. Inspired by MMM, our study proposes a method for identifying multi-sources with a small fraction of observers based on backward diffusion and integer programming (IP).

The outline of this letter is as follows. We will give an introduction to the propagation model and MMM before presenting

E-mail address: huzhaolong@zjnu.edu.cn (Z.-L. Hu).

https://doi.org/10.1016/j.physleta.2018.01.037
0375-9601/© 2018 Published by Elsevier B.V.

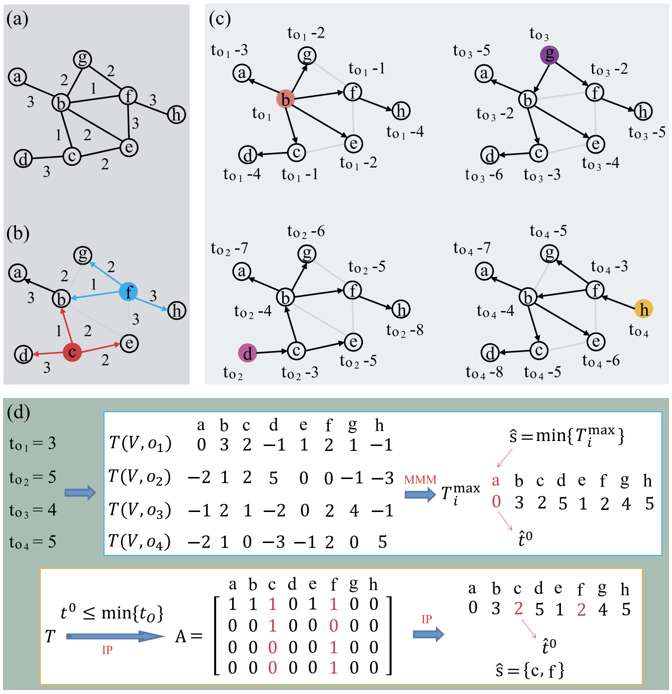

Fig. 1. (Color online.) An example of diffusion process with multi-source. (a) A simple undirected network, and the value on the corresponding edge is the delay time of the diffusion process. (b) The diffusion process with two sources $S=\{s_1,s_2\}$. The diffusion paths are highlighted by blue arrows and red arrows, respectively. (c) Illustration of the backward diffusion-based with four observers b, d, g, and h. The diffusion paths are highlighted by solid lines with arrows. The corresponding informed time t_O for these observers are t_{o_1}, t_{o_2}, t_{o_3} and t_{o_4}. The inferred initial diffusion times are presented around each node via backward diffusion-based method. (d) An example of the process of locating sources by MMM and IP.

our own method source localization with IP. We will then compare model performances using MMM versus IP using different structures of synthetic and empirical networks. Finally, we will investigate the robustness of these methods for source localization.

2. Model and methods

Propagation model. An undirected graph $G=\{V,E\}$ is consist of node set V and edge set E. Connections among nodes in the networks represent dependency relations through which information can spread. It is assumed that the topology of the network G is known and the topology is invariant. Define $S=\{s_m\}_{m=1}^{N_s}$ as the diffusion source set that triggers the diffusion process. Here $S\subset V$ and N_s represents the number of sources. Similar to the propagation model of [29], the details of the diffusion model are as follows (see Fig. 1):

i) At the initial diffusion time t^0, only sources are informed.

ii) The informed nodes will transmit the information to their neighbors along the shortest paths according to the corresponding delay time. The time of node i first receiving the information denoted as t_i and node i will transmit this information to all its neighbors $\Gamma(i)$. Then the time for each uninformed neighbor $j\in\Gamma(i)$ receiving the information is $t_i+\theta_{ij}$. Here θ_{ij} is the transmission delay time of the edge (i,j), which follows a known joint distribution. For any node $i\in V$, the informed time t_i can be expressed as

$$t_i=t^0+\min\{\Delta_{s_1,i},\Delta_{s_2,i},\cdots,\Delta_{s_{N_s},i}\}, \tag{1}$$

where $\Delta_{s_m,i}$ represents the shortest transmission time between s_m and i. The sign $\min\{\cdot\}$ denotes the minimum in the set $\{\cdot\}$.

iii) The process is terminated when all of nodes in V have been informed.

Source localization. Define $O=\{o_k\}_{k=1}^{K}$ as the set of K observers, and t_O as the corresponding informed time set, which can be recorded by these observers. Combining the network topology with delay time and using the backward propagation of each observer in O, we can infer the initial diffusion time of other nodes, see Fig. 1(c).

For each node i, if the transmission time $\Delta_{s_m,i}$ is the minimum of $\{\Delta_{s_1,i},\Delta_{s_2,i},\cdots,\Delta_{s_{N_s},i}\}$, then node i is informed by source s_m. Let's define $i\in\Pi_{s_m}$, where Π_{s_m} is the diffusion set of source s_m, shown by blue arrows and red arrows in Fig. 1(b). So the relationship between the informed time and the diffusion set can be obtained. According to Eq. (1), for any node $i\in\Pi_{s_m}$, the informed time must be

$$t_i=t^0+\Delta_{s_m,i}. \tag{2}$$

Thus for any observer o_k, the inferred initial diffusion time of node s from the observer node should satisfy

$$T(s,o_k)=t_{o_k}-\Delta_{s,o_k}\le t^0. \tag{3}$$

If the observer o_k is informed by s, then Δ_{s,o_k} must be the minimum in Δ_{S,o_k} set. Hence, by virtue of Eq. (3), the true initial diffusion time t^0 is the maximum of $T(s,o_k)$, defined as $T_s^{\max}$. In order to identify all sources, we should infer the initial diffusion time of all nodes from backward diffusion-based method, namely $T(i\subset V,o_k\subset O)$. According to $T(i\subset V,o_k\subset O)$, we can readily obtain $T_i^{\max}$ from all observers O. The true initial diffusion time t^0 should be the minimum in all the inferred time $T_i^{\max}$, then we choose those nodes with the minimum of $T_i^{\max}$ as sources [29]. Here we call it as the maximum–minimum method (MMM).

We will use a concrete example to illustrate the above method. Without loss of generality, we treat the edge (i,j) as an integer chosen from a uniform distribution $U(1,5)$, see Fig. 1(a). For instance, set nodes c and f as sources, see Fig. 1(b). Let's assume $t_b=3$, $t_d=5$, $t_g=4$ and $t_h=5$ which are collected from the four observers. Based on backward diffusion from observers b, d, g, and h, respectively, we can compute the estimated initial diffusion time $T(V,O)=\begin{bmatrix} & a & b & c & d & e & f & g & h \\ b & 0 & 3 & 2 & -1 & 1 & 2 & 1 & -1 \\ d & -2 & 1 & 2 & 5 & 0 & 0 & -1 & -3 \\ g & -1 & 2 & 1 & -2 & 0 & 2 & 4 & -1 \\ h & -2 & 1 & 0 & -3 & -1 & 2 & 0 & 5 \end{bmatrix}$. Thus $T_i^{\max}$ for each node $i\in V$ from the observers is $\begin{bmatrix} a & b & c & d & e & f & g & h \\ 0 & 3 & 2 & 5 & 1 & 2 & 4 & 5 \end{bmatrix}$. By using MMM, the sources are the nodes with the minimum from $T_i^{\max}$, so a is the source and the inferred initial time is $\hat{t}^0=0$, which doesn't coincide with the true sources c and f. Fig. 1(d) shows the process of locating sources by MMM.

Now we develop a method to locate sources by 0–1 integer programming. As t^0 can't be larger than the minimum of the set of observation time t_O, then the potential sources are those nodes must satisfy $T_i^{\max}\le\min\{t_O\}$, say Ω. So we find $\{a,b,c,e,f\}\subset\Omega$. Setting all the $T_{i\in\Omega}^{\max}$ as 1, while others are 0. The matrix $T(V,O)$ is translated into $A=\begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$.

Because only the true sources can account for the observers, then the sources are those nodes with the least number of columns to cover all the rows, i.e. all the observers, which can be described as

Please cite this article in press as: Z.-L. Hu et al., Localization of diffusion sources in complex networks with sparse observations, Phys. Lett. A (2018), https://doi.org/10.1016/j.physleta.2018.01.037