

Effect of observation time on source identification of diffusion in complex networks

Chaoyi Shi(史朝义), Qi Zhang(张琦), and Tianguang Chu(楚天广)

Citation: Chin. Phys. B, 2022, 31 (**7**): 070203. DOI: 10.1088/1674-1056/ac5985
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Dynamics and intermittent stochastic stabilization of a rumor spreading model with guidance mechanism in heterogeneous network

Xiaojing Zhong(钟晓静), Yukun Yang(杨宇琨), Runqing Miao(苗润青), Yuqing Peng(彭雨晴), and Guiyun Liu(刘贵云)

Chin. Phys. B, 2022, 31 (**4**): 040205. DOI: 10.1088/1674-1056/ac587d

Effect of degree correlation on edge controllability of real networks

Shu-Lin Liu(刘树林) and Shao-Peng Pang(庞少鹏)†

Chin. Phys. B, 2020, 29 (**10**): 100202. DOI: 10.1088/1674-1056/ab99ab

Ranking important nodes in complex networks by simulated annealing

Yu Sun(孙昱), Pei-Yang Yao(姚佩阳), Lu-Jun Wan(万路军), Jian Shen(申健), Yun Zhong(钟赟)

Chin. Phys. B, 2017, 26 (**2**): 020201. DOI: 10.1088/1674-1056/26/2/020201

Co-evolution of the brand effect and competitiveness in evolving networks

Guo Jin-Li

Chin. Phys. B, 2014, 23 (**7**): 070206. DOI: 10.1088/1674-1056/23/7/070206

Exponential synchronization of complex dynamical networks with Markovian jumping parameters using sampled-data and mode-dependent probabilistic time-varying delays

R. Rakkiyappan, N. Sakthivel, S. Lakshmanan

Chin. Phys. B, 2014, 23 (**2**): 020205. DOI: 10.1088/1674-1056/23/2/020205

Effect of observation time on source identification of diffusion in complex networks

Chaoyi Shi(史朝义)[1], Qi Zhang(张琦)[2], and Tianguang Chu(楚天广)[1,†]

[1] *College of Engineering, Peking University, Beijing* 100871, *China*

[2] *School of Information Technology and Management, University of International Business and Economics, Beijing* 100105, *China*

(Received 29 December 2021; revised manuscript received 30 January 2022; accepted manuscript online 2 March 2022)

This paper examines the effect of the observation time on source identification of a discrete-time susceptible-infected-recovered diffusion process in a network with snapshot of partial nodes. We formulate the source identification problem as a maximum likelihood (ML) estimator and develop a statistical inference method based on Monte Carlo simulation (MCS) to estimate the source location and the initial time of diffusion. Experimental results in synthetic networks and real-world networks demonstrate evident impact of the observation time as well as the fraction of the observers on the concerned problem.

Keywords: complex network, source identification, statistical inference, partial observation

PACS: 02.70.Uu, 87.23.Ge, 89.75.–k **DOI:** 10.1088/1674-1056/ac5985

1. Introduction

Recent years, researchers have witnessed tremendous growth in studies concerning the diffusion processes in complex networks, such as modeling epidemic spreading,[1–4] identifying influential spreaders,[5–9] and analyzing diffusion dynamics.[10–12] Among a plethora of research thrusts, the problem of identifying the diffusion sources in complex networks based on observed information has gathered increasing attention due to its broad applications in such problems as identifying the origin of a rumor,[13,14] locating the source of epidemic,[15–18] and so on. Clearly, studies in this direction are crucial in preventing or controlling spread of adverse or detrimental influences in a network system consisting of large amount of interconnected individuals or nodes.

Basically, source identification problem aims at estimating the location of source node that initiates the diffusion in a network by using certain observation information of the diffusion. So far, several approaches have been proposed to solve the source identification problem with observations that involve the states of observed nodes,[13,15–18] or the (infection) time at which the diffusion arrives at the observed nodes.[14,19–22] Early studies often assume complete observation condition, i.e., the state of each node is observed, and make use of the measures of certain centricity to evaluate the source detection probability of the maximum likelihood (ML) estimator.[13,15] However, it is usually hard in practice to obtain complete observations of a network, since the costs for observing all nodes can be expensive and the states of some nodes may be hidden.[3,14] In view of these, considerable efforts have been devoted to the study of source identification problems under partial observation conditions, i.e., the states or infection time of only a fraction of nodes can be available. For instance, in Refs. [16–18] the authors considered a single snapshot of partial observations at a given time and proposed different propagation algorithms to find the source of diffusion in networks. There have been also works addressing source localization by exploiting observed information of propagation delays from the source, under various assumptions on delay distributions, e.g., Gaussian and uniform distributions.[19–21] Another effective approach to locating diffusion source via incomplete observations in a network makes use of statistical inference based on Monte Carlo simulations.[22–24] The basic idea of this approach is to infer the diffusion source by evaluating the similarity between the simulation data and the observed data. Essentially, the approach applies to arbitrary network structures and therefore has wide applications. Notice that most existing studies assume that the initial time of diffusion is known in estimating source location, some of them are also concerned with the impact of the size of observation set on the results.[13,15–18]

In this paper, we intend to discuss the effect of the observation time on source identification in a diffusive network. So far in literature, little has been available concerning such an issue. Generally speaking, identification of diffusion source in a network essentially relies on the diffusion model, the observation condition, and the underlying network structure. Here we consider the case that the diffusion process is described by a discrete-time SIR model, with the initial condition of all nodes in susceptible states except one in infected state as the source. In literature, this type of models is often used to describe the spread of epidemics[16,17] and computer virus.[15,19] We consider the problem of estimating both the source location and the initial time of diffusion with snapshot of partial nodes, and formulate it as an ML estimator in terms of the marginal proba-

†Corresponding author. E-mail: chutg@pku.edu.cn

© 2022 Chinese Physical Society and IOP Publishing Ltd

bilities of a node in different states. A method based on Monte Carlo simulation (MCS) is developed to evaluate the marginal probabilities of a candidate source node, whereby to examine the effect of the observation time as well as the fraction of the observers on solutions of the concerned problem with numerical experiments on typical synthetic and real-world networks.

2. Problem formulation

We consider the case that the diffusion outbreak occurs at a single node and the states of only a fraction of nodes can be observed with a snapshot. The task of source identification is to determine the location of source and the initial time of diffusion in a network based on the partial observation. In the following, we will introduce the diffusion model and the ML estimator for the source identification problem.

2.1. The diffusion model

To describe the discrete-time SIR diffusion process, we model the underlying contact network as an undirected graph $G(V,E)$ with N nodes, where V is the set of nodes, $E \in V \times V$ is the set of edges. At time t, each node $i \in V$ has three possible states represented by a variable X_i^t: susceptible, $X_i^t = S$; infected, $X_i^t = I$; or recovered, $X_i^t = R$. In each time step, an infected node i infects its susceptible neighbor j with probability λ_{ij}, or recovers with probability μ_i, and no longer get infected. Thus, a susceptible node i becomes infected with probability,

$$P\left(X_i^t = I | X_i^{t-1} = S\right) = 1 - \prod_{\left\{j \in \partial i:\, X_j^{t-1} = I\right\}} (1 - \lambda_{ji}),$$

where ∂i is the set of neighbors of node i. As mentioned above, we assume that at an unknown initial time t_0, all nodes are in state S except only one infected node i_0, the source, that initiates diffusion process in the network.

2.2. Snapshot observation

The condition we will use to infer the diffusion source is a snapshot of K nodes at a certain time $T > t_0$, $0 < K \leq N$. Let O^T denote the set of K observed nodes, called observers. According to their susceptible, infected, or recovered states at time T, the observers can be divided into three subsets O_S^T, O_I^T, and O_R^T, respectively. Clearly, we should locate the source node $i_0 \in V \backslash O_S^T$, the set V excluding O_S^T.

For the issue of estimating the initial time t_0, we notice that by Ref. [25] it would become difficult to locate the source beyond $t_0 + t_{\max}$ time steps, in which $t_{\max}$ denotes the maximum duration of diffusion and can be determined by the condition

$$\sum_{i \in V} P(X_i^t = I) + \sum_{i \in V} P(X_i^t = R) \sim N,\ t = t_0 + t_{\max},$$

where $P(X_i^t = I)$ and $P(X_i^t = R)$ denote the probabilities of node i to be infected or recovered at time T, respectively. From this, we can assume $t_0 \in \mathbb{Z} \cap [T - t_{\max}, T-1]$ for estimation of the initial time of diffusion, with $\mathbb{Z}$ being the set of nonnegative integers.

Figure 1 shows an example of the diffusion process and snapshot observation on an arbitrary graph G.

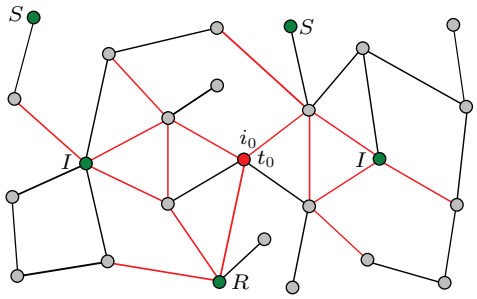


Fig. 1. Visualization of the diffusion process and observed data on graph G. At unknown time t_0, the source node i_0 (red) initiates the diffusion. The red edges are the links on which the spread occurs. In this example, a snapshot of the observers (green) is taken at time T, which has three types of nodes with the state S, I, or R for each. The goal is to identify the source location and initial time of diffusion from the snapshot observation.

2.3. Maximum likelihood estimator

To determine the location of source and the initial time of diffusion in a network with the snapshot observation, we formulate the problem as an ML estimator. Specifically, let $P\left(i_0, t_0 | O_S^T, O_I^T, O_R^T\right)$ denote the posterior probability that i_0 is source and t_0 is initial time given the observation sets O_S^T, O_I^T, and O_R^T. Suppose that there is no prior knowledge of the source location and initial time of diffusion, the posterior probability is thus proportional to the joint probability of obtaining the observations O_S^T, O_I^T and O_R^T given the source node i_0 and initial time t_0,

$$P\left(i_0, t_0 | O_S^T, O_I^T, O_R^T\right) \propto P\left(O_S^T, O_I^T, O_R^T | i_0, t_0\right).$$

Therefore, we can formulate the source identification problem as the following ML estimator:

$$\{\hat{i}_0, \hat{t}_0\} = \arg \max_{\substack{i_0 \in V \backslash O_S^T \\ t_0 \in \mathbb{Z} \cap [T - t_{\max}, T-1]}} P\left(O_S^T, O_I^T, O_R^T | i_0, t_0\right), \quad (1)$$

the solutions $\hat{i}_0$ and $\hat{t}_0$ give estimations of the source node and initial time. The maximizer in the estimator is determined by the SIR model and the underlying network structure.

As there is usually no tractable way to compute exactly the joint probability in the estimator (1), we will approach it approximately as follows. For the source node i_0 and initial time t_0, let $P(X_i^T = S | i_0, t_0)$, $P(X_i^T = I | i_0, t_0)$, and $P(X_i^T = R | i_0, t_0)$ denote the marginal probabilities of node i to be susceptible, infected or recovered at time T, respectively. Then,

$$P\left(O_S^T, O_I^T, O_R^T | i_0, t_0\right) \propto \sum_{i \in O_S^T} \ln\left(P(X_i^T = S | i_0, t_0)\right) + \sum_{j \in O_I^T} \ln\left(P(X_j^T = I | i_0, t_0)\right) + \sum_{k \in O_R^T} \ln\left(P(X_k^T = R | i_0, t_0)\right), \quad (2)$$

and clearly,

$$P\left(X_i^T = S|i_0,t_0\right) + P\left(X_i^T = I|i_0,t_0\right) + P\left(X_i^T = R|i_0,t_0\right) = 1.$$

We will make use of the marginal probabilities instead of the joint probability in the estimator (1) to infer the source node and initial time of diffusion.

3. Monte Carlo simulation

Now we present an MCS approach to estimate the marginal probabilities of a node. The basic idea is that for a candidate source node $i_0 \in V\backslash O_S^T$, we perform simulations of m times for the diffusion process following the SIR model on time interval $[0,t_{\max}]$, and sample the state of every observer $i \in O$ at time $t \in \mathbb{Z}\cap[0,t_{\max}]$. Then, by examining the similarity of the samples and the snapshot to determine the source and initial time of the diffusion.

To be specific, let $Y_i^t(i_0,l)$ be the state of node i at time t for the l-th simulation with the candidate source node i_0, it is clear that

$$Y_i^t(i_0,l) \in \{S,I,R\},\ i_0 \in V\backslash O_S^T,$$

and for $t = 0$,

$$Y_i^0(i_0,l) = \begin{cases} I, & i = i_0, \\ S, & i \neq i_0, \end{cases}$$

for $l = 1,\ldots,m$. In simulations, a node in the network updates its state by the following rule.

For a susceptible node i at time $t-1$, each infected neighbor j will execute an independent infection attempt with success probability λ_{ji}, triggered by a Bernoulli trial (flipping an independent coin) that generates a random number r_0 from the uniform distribution between 0 and 1 as threshold. Namely, the state of node i in the next step t is given by

$$Y_i^t(i_0,l) = \begin{cases} I,\ \exists j \in \partial i : X_j^{t-1} = I, \lambda_{ji} \geq r_0 \\ S,\ \forall j \in \partial i : X_j^{t-1} = I, \lambda_{ji} < r_0 \end{cases}, \text{ if } Y_i^{t-1}(i_0,l) = S.$$

Similarly, for a node i in the state I at time step $t-1$, it will execute a recovery attempt with success probability μ_i according to a Bernoulli trial number r_0 as before, and updates its state at time step t as follows:

$$Y_i^t(i_0,l) = \begin{cases} R,\ \mu_i \geq r_0 \\ I,\ \mu_i < r_0 \end{cases}, \quad \text{if } Y_i^{t-1}(i_0,l) = I.$$

According to the specification of the SIR model described before, all recovered nodes will not be infected anymore and remain the recovered state during the simulation process, i.e.,

$$Y_i^t(i_0,l) = R, \quad \text{if} \quad Y_i^{t-1}(i_0,l) = R.$$

After m simulations with a candidate source node $i_0 \in V\backslash O_S^T$, the marginal probability of node i in state S at time T, for a given initial time $t_0 \in \mathbb{Z}\cap[0,t_{\max}]$, is given by

$$P\left(X_i^T = S|i_0,t_0\right) = \frac{1}{m}\sum_{l=1}^{m}\delta\left(S, Y_i^{T-t_0}(i_0,l)\right), \tag{3}$$

where $\delta(\cdot,\cdot)$ denotes the Kronecker delta function. Similarly, the marginal probabilities of the node i being in states I and R at time T are given respectively by

$$P\left(X_i^T = I|i_0,t_0\right) = \frac{1}{m}\sum_{l=1}^{m}\delta\left(I, Y_i^{T-t_0}(i_0,l)\right), \tag{4}$$

$$P\left(X_i^T = R|i_0,t_0\right) = \frac{1}{m}\sum_{l=1}^{m}\delta\left(R, Y_i^{T-t_0}(i_0,l)\right). \tag{5}$$

This along with Eq. (2) will be used to solve the estimator (1) later in numerical experiments. To be specific, for each node $i_0 \in V\backslash O_S^T$, a number m of simulations for the SIR diffusion processes are performed with i_0 as a candidate source of diffusion so as to sample the state $Y_i^t(i_0,l)$ of each observer $i \in O$ on time interval $t \in \mathbb{Z}\cap[0,t_{\max}]$. Then, for a candidate initial time t_0, the marginal probabilities of the observers in different states at the snapshot time can be evaluated by Eqs. (3)–(5). Finally, we evaluate the joint probability $P\left(O_S^T, O_I^T, O_R^T|i_0,t_0\right)$ by Eq. (2) and solve the ML estimator (1). See Algorithm 1 for the procedures.

Algorithm 1 MSC-based source identification algorithm

Input: $G(V,E)$; λ_{ij}; μ_i; O_S^T; O_I^T; O_R^T, $t_{\max}$
Output: the source location $\hat{i}_0$; the initial time of diffusion $\hat{t}_0$
for a candidate source node $i_0 \in V\backslash O_S^T$ **do**
  for $l = 1$ to m **do**
    Sample the state $Y_i^t(i_0,l)$ of each observer $i \in O$ at time $t \in \mathbb{Z}\cap[0,t_{\max}]$
  end for
  for a candidate initial time $t_0 \in \mathbb{Z}\cap[T-t_{\max},T-1]$ **do**
    Evaluate the marginal probabilities $P\left(X_i^T = S|i_0,t_0\right)$, $P\left(X_i^T = I|i_0,t_0\right)$, $P\left(X_i^T = R|i_0,t_0\right)$, $i \in O$
    Evaluate the joint probability $P\left(O_S^T, O_I^T, O_R^T|i_0,t_0\right)$ by Eq. (2)
  end for
end for
$\{\hat{i}_0,\hat{t}_0\} = \arg \max_{\substack{i_0 \in V\backslash O_S^T \\ t_0 \in \mathbb{Z}\cap[T-t_{\max},T-1]}} P\left(O_S^T, O_I^T, O_R^T|i_0,t_0\right)$

We remark that the algorithmic complexity for a single SIR simulation is $\mathcal{O}(N\langle k\rangle t_{\max})$, where $\langle k\rangle$ is the average degree of the graph G, thus running m simulations for a candidate source node i_0 requires $\mathcal{O}(mN\langle k\rangle t_{\max})$ time. In experiments we found that it is enough to take $m = 200$ for obtaining a satisfactory identification result, further increasing the simulation times does not improve the accuracy of identification significantly.

Compared with some related studies, our method can be implemented directly by Monte Carlo simulations, without the need of complicated calculations or additional assumptions. For instance, the methods proposed in Refs. [21,22] require to calculate the multiple integrations of joint probability for partial observations. In Ref. [24], the source likelihood was estimated by using the Gaussian weighting function.

4. Experiments and discussion

We give numerical experiments to examine the effect of the observation time as well as the fraction of the observers on source identification for synthetic networks and real-world networks. In all the cases a fraction $\gamma = K/N$ of nodes is available in observation at time T for the source identification problem, and the source i_0 and the initial time t_0 are estimated by Algorithm 1. The performance of the method is assessed in terms of the success rate of locating source node for different fraction γ and snapshot time T.

4.1. Synthetic networks

We consider three typical synthetic networks, i.e., Erdös–Rényi (ER) random network, Barabási–Albert (BA) scale-free network, and Watts–Strogtz (WS) small world network,[1] all containing $N = 1000$ nodes with average degree $\langle k\rangle \approx 6$. For convenience in simulation and without loss of generality,[13,17,24] we consider uniform infection and recovery probabilities by letting $\lambda_{ij} = \lambda$ and $\mu_i = \mu$. In each experiment, a node is randomly selected to initiate the diffusion at time $t = 0$ and we run SIR process for $t_{\max}$ steps, which can be approximated by

$$t_{\max} = \frac{d}{\lambda(1-\mu)} + \frac{1}{\mu},$$

where d is the diameter of a network, the first term on the left side is the maximum infection time of a node, and the second term is the average recovery time of a node. Then, the state of a fraction γ of nodes is observed at time T, where $\gamma \in \{0.1, 0.2, \ldots, 0.9, 1.0\}$ and the snapshot time $T \in \{1, \ldots, t_{\max}\}$. We repeated the experiment 4000 times for each case to get the average success rate of located source.

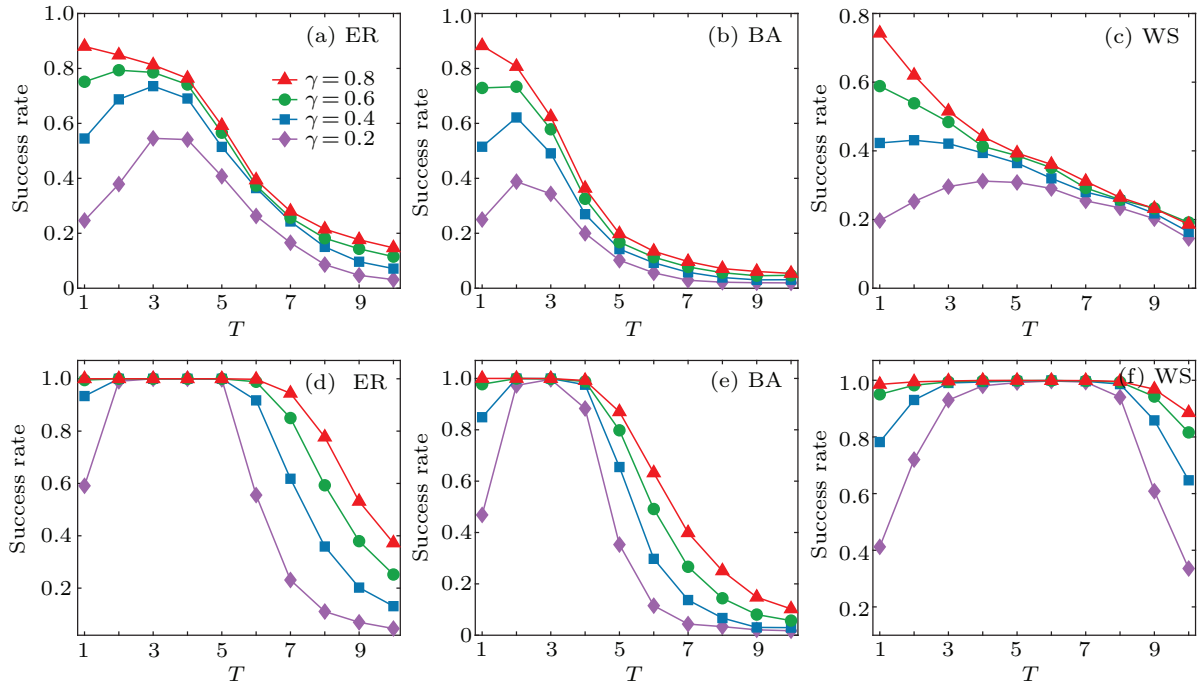


Fig. 2. Effects of the snapshot time T and the fraction γ of observers on source identification results: (a)–(c) $\lambda = 0.4$, $\mu = 0.2$, (d)–(f) $\lambda = 0.8$, $\mu = 0.2$. The success rate of locating source on ER, BA, and WS networks, with $N = 1000$ nodes and average degree, $\langle k\rangle \approx 6$. The diffusion starts at $t = 0$, a fraction $\gamma \in \{0.2, 0.4, 0.6, 0.8\}$ of nodes is observed at time $1 \le T \le 10$.

We examine the effect of the snapshot time T and the fraction γ of observers on source identification results. Existing studies mainly considered the effect of γ on source localization accuracy, regardless of considering the snapshot time T. Our results as illustrated in Fig. 2 reveal an involved influence of the two factors on source identification problem. We can see that, in general, the source identification problem will be hard to attack for observations with large time delay; and the

greater the number of observed nodes, the more accurately our methods can infer the true source. Indeed, a large value of T allows for a full development of diffusion (hence infection) in a network, this would reduce effectiveness of the snapshot information. The cases for different infection probability λ and recovery probability μ are similar and hence omitted.

In more detail, Fig. 2 indicates different trends of the success rate of source localization with respect to T and γ. For very large values of γ (e.g., not less than 0.8), the success rate decreases monotonically with the increase of T; whereas for the major portion of values of γ (e.g., $(0,0.8)$), the success rate first rises to a highest level and then drops down with the increase of T. This implies that one could try to choose an optimal time T for observation to get more reliable results. As shown in Fig. 2, it is better in general to take a snapshot at early stage of the diffusion process, and the range for such a time T also depends on SIR model parameters λ and μ. However, for very small γ (e.g., 0.2) the accuracy of source localization is low in early stages of diffusion as little information about infected nodes is available. Thus, it may be difficult to infer the source at early stages of diffusion from a very small observation set.

Figure 3 illustrates the impact of the snapshot time T and the infection probability λ on the success rate of locating the source, where the value of λ varies from 0.1 to 1.0. The results are different, depending on the network structures. For the ER network, the observation time T has little effect on the source localization when the infection probability λ is small (e.g., (0.1,0.3)). For the other two types of networks, this is true for relatively large values of λ (e.g., (0.1,0.4) for BA, (0.1,0.7) for WS). On the other hand, for a large value of λ, a relatively earlier observation is in favor of source localization in ER and BA networks. For the WS network, the effect of different observation time T is less evident in the case of larger values of λ.

Figure 4 depicts the average snapshot time T of successfully locating the source for different observation fraction γ. It shows that in general T decreases monotonically with the increase of γ for all networks. This implies that a snapshot of a larger observation set facilitates the inference of the source at early stages of diffusion. This is consistent with the results as shown in Fig. 2. In particular, given an observation fraction γ, one could try to choose an appropriate time T for observation by Fig. 4 to get more reliable results.

Figure 5 shows the source localization and initial time estimate results for different fraction γ of observers and snapshot time T. From Fig. 5(a) we can see that the success rate of locating source increases rapidly as γ increasing from a small

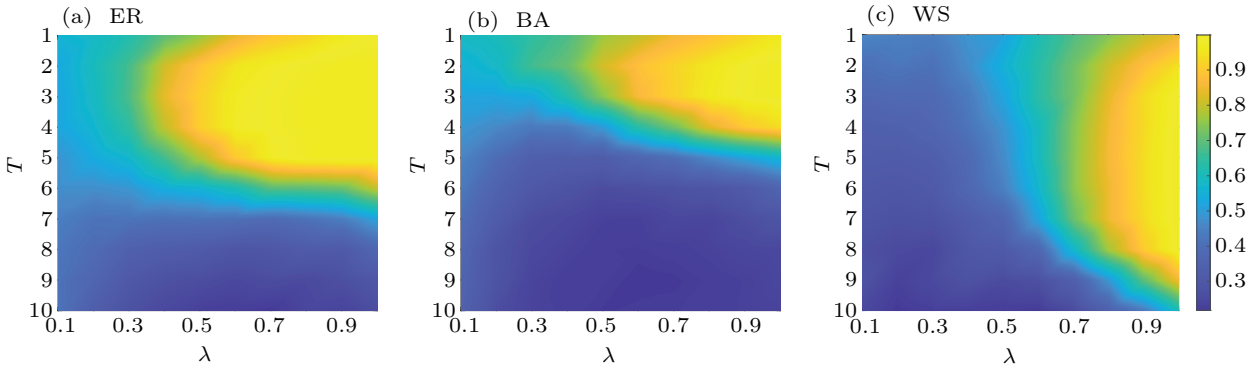


Fig. 3. Effects of the snapshot time T and infection probability λ on success rate of locating source for ER, BA, and WS networks, with $N = 1000$ nodes and average degree $\langle k\rangle \approx 6$. The diffusion starts at $t = 0$ with $\lambda = 0.1,\ldots,1.0$ and $\mu = 0.1$, a fraction $\gamma = 0.4$ of nodes is observed at time $1 \le T \le 10$.

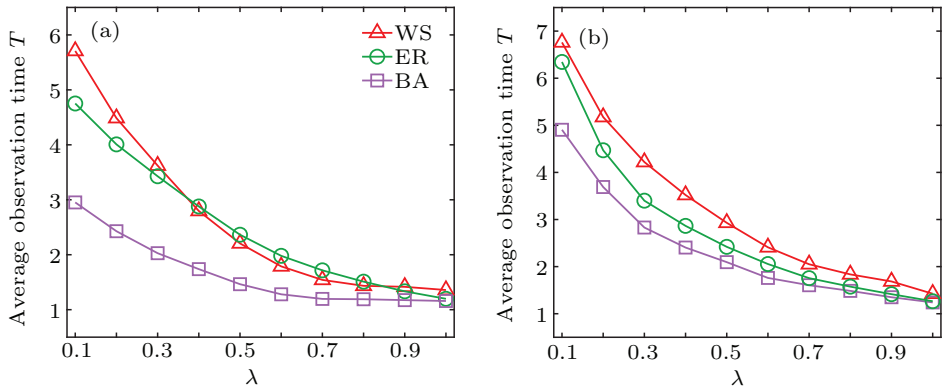


Fig. 4. Observation time T of successfully locating the source for different fraction γ on ER, BA, and WS networks, all containing $N = 1000$ nodes with average degree $\langle k\rangle \approx 6$: (a) $\lambda = 0.4$, $\mu = 0.2$, (b) $\lambda = 0.3$, $\mu = 0.1$. The diffusion process starts at $t = 0$ from a randomly selected node, and a snapshot of γN nodes is taken for $\gamma \in \{0.1, 0.2, \ldots, 0.9, 1\}$.

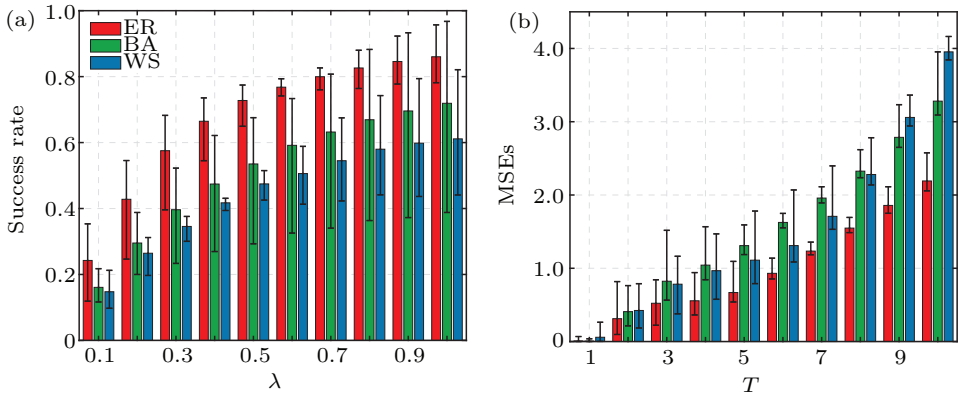


Fig. 5. Source localization and initial time estimate results on ER, BA, and WS networks, all containing $N = 1000$ nodes with average degree $\langle k \rangle \approx 6$. The diffusion process starts at $t = 0$ from a randomly selected node with $\lambda = 0.4$ and $\mu = 0.2$, and a snapshot of γN nodes is taken at time T for $\gamma \in \{0.1, 0.2, \ldots, 0.9, 1\}$ and $1 \leq T \leq 10$. (a) Success rates of locating source node for different fraction γ. The lower and upper values of the success rate are calculated by using snapshot observation at different time T. (b) MSEs of the initial time estimates for different snapshot time T. The lower and upper values of each MSE are computed with different γ.

value till to certain value (e.g., 0.6), after that, adding more observer nodes does not improve the identification accuracy significantly. Figure 5(b) gives the mean square error (MSE) of the initial time estimates for different snapshot time T, it indicates that the snapshots taken at early stage of diffusion result in small MSE values for different graph structures.

4.2. Real-world networks

Furthermore, we consider two real-world network examples of email[26] and US West-Coast power grid,[27] as detailed in Table 1, which are often cited to test source identification methods in the literature.[17,18] In our experiments, the diffusion process starts at $t = 0$ from a randomly selected node, with $\lambda \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$ and $\mu = 0.2$ respectively in each running, and a snapshot of γN nodes is taken at time T for $\gamma \in \{0.05, 0.1, \ldots, 0.45, 0.5\}$ and $1 \leq T \leq 10$ respectively in each time.

Table 1. The network parameters in the experiments, N and $|E|$ denote the network size and the number of edges, respectively.

Data sets	N	$\lvert E \rvert$	Description
Email[26]	1133	5451	email communication network
Power grid[27]	4941	6594	electrical power grid in western US

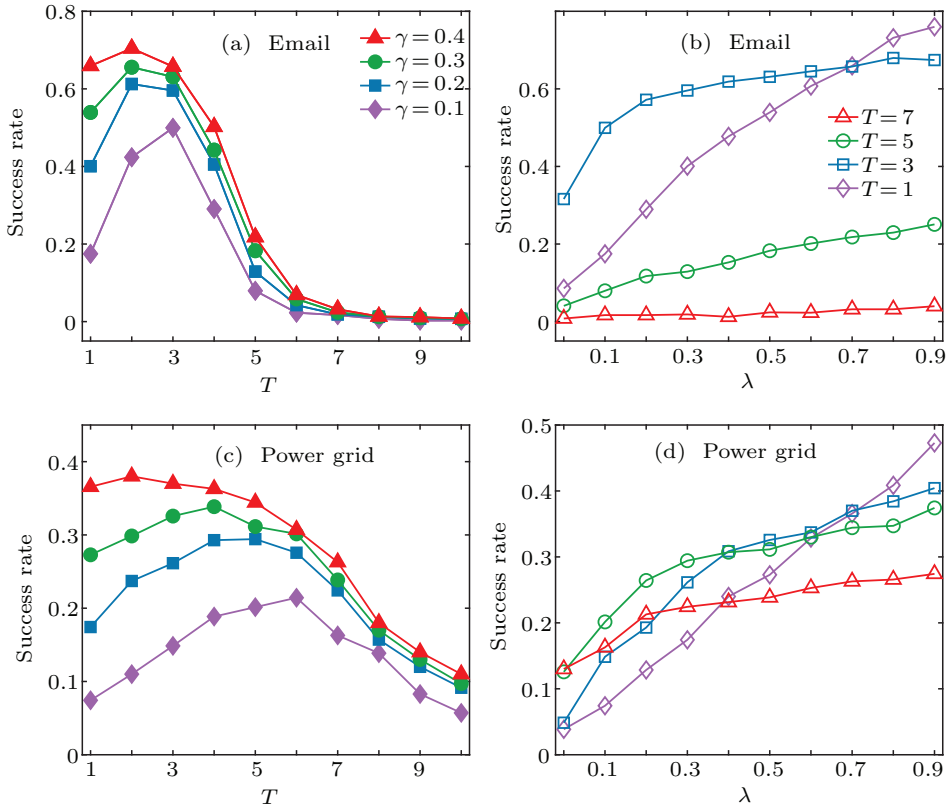


Fig. 6. Effects of the snapshot time T and the fraction γ of observers on source identification results for email and US West-Coast power grid. The diffusion process starts at $t = 0$ with $\lambda \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$ and $\mu = 0.2$, and a snapshot of γN nodes is taken at time T for $\gamma \in \{0.05, 0.1, \ldots, 0.45, 0.5\}$ and $1 \leq T \leq 10$. (a)–(d) Success rate of locating source node for different T and γ.

Figure 6 depicts the average results over 1000 experiments. It shows an evident impact of the observation time as well as the fraction of the observers on the source identification problem, which is similar to the case of the synthetic networks. Particularly, Figs. 6(a) and 6(c) indicate that the success rate of locating source is small at early stage of the diffusion process, and it will be hard to locate the source for large observation time. Figures 6(b) and 6(d) show that, in general, the success rate of locating source increases with the number of observer for different snapshot time T. For a small value of T (e.g., 1), the success rate of locating source increases rapidly as the fraction γ of observers increasing, whereas for a big value of T (e.g., 7), adding more observer nodes does not improve the source localization accuracy significantly.

5. Summary

We have discussed the effect of the observation time as well as the size of the observation set on source identification of diffusion in a network with discrete-time SIR spreading process, under incomplete observation conditions. The method makes use of the Monte Carlo simulation in evaluation of the marginal probabilities of a node that are involved in ML estimator of the source. Numerical experiments in synthetic networks and real-world networks reveal interesting yet involved nonlinear effects of the observation time and the fraction of observers on source identification. In general, a large value of snapshot time will not be in favor of the source identification for most values of the fraction of observers, since it allows for a full development of diffusion (hence infection) in a network, reducing effectiveness of the snapshot information. In particular, for very large values of the fraction of observers, the success rate of source localization decreases monotonically with the increase of the observation time. On the other hand, however, for a very small observation set, it is usually difficult to infer the source at early stages of diffusion because little information about infected nodes is available. More complicated and interesting cases occur for neither too small nor too large values of the fraction of observers, where the success rate of source localization first rises to a highest level and then drops down with the increase of the snapshot time, indicating possible choice of an optimal observation time to get more reliable results.

Acknowledgements

Project supported by the National Natural Science Foundation of China (Grant Nos. 61673027 and 62106047), the Beijing Social Science Foundation (Grant No. 21GLC042), and the Humanity and Social Science Youth foundation of Ministry of Education, China (Grant No. 20YJCZH228).

References

[1] Newman M E J 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
[2] Ji S G, Lü L Y, Yeung C H and Hu Y Q 2017 *New J. Phys.* **19** 073020
[3] Brockmann D and Helbing D 2013 *Science* **342** 6164
[4] Feng Y, Ding L, Huang Y H and Guan Z H 2016 *Chin. Phys. B* **25** 128903
[5] Hao X and Li X 2021 *Europhys. Lett.* **134** 58001
[6] Yan X L, Cui Y P and Ni S J 2020 *Chin. Phys. B* **29** 048902
[7] Ruan Y R, Lao S Y, Xiao Y D, Wang J D and Bai L 2016 *Chin. Phys. Lett.* **33** 028901
[8] Yang Y Z, Hu M and Huang T Y 2020 *Chin. Phys. B* **29** 088903
[9] Huang L Y, Tang P C, Huo Y L, Zheng Y and Cheng X F 2019 *Acta Phys. Sin.* **68** 128901 (in Chinese)
[10] Pan L M, Wang W, Cai S M and Zhou T 2019 *Phys. Rev. E* **100** 022316
[11] Ding Z J, Liu T, Lou X X, Shen Z W, Zhu K J, Jiang R, Wang B H and Chen B K 2019 *Physica A* **516** 6684
[12] Liu S L, Pang S P 2020 *Chin. Phys. B* **29** 100202
[13] Shah D and Zaman T 2015 arXiv:1110.6230[math.PR]
[14] Wang H S, Wu J, Pan S R, Zhang P and Chen L 2017 *Comput. Netw.* **114** 154
[15] Comin C H and Costa L D 2011 *Phys. Rev. E* **84** 056105
[16] Altarelli F, Braunstein A, Dall'Asta L, Lage-Castellanos A and Zecchina R 2014 *Phys. Rev. Lett.* **112** 118701
[17] Lokhov A Y, Mézard M, Ohta H and Zdeborová L 2014 *Phys. Rev. E* **90** 012801
[18] Zhu K and Ying L 2014 arXiv:1510.01814[cs.SI]
[19] Shen Z S, Cao S N, Wang W X, Di Z R and Stanley H E 2016 *Phys. Rev. E* **93** 032301
[20] Wang H J, Zhang F F and Sun K J 2021 *Phys. Lett. A* **393** 127184
[21] Pinto P C, Thiran P and Vetterli M 2012 *Phys. Rev. Lett.* **109** 068702
[22] Farajtabar M, Rodriguez M G, Zamani M, Du N, Zha H Y and Song L 2015 arXiv:1501.06582v1[cs.SI]
[23] Zhai X M, Wu W L and Xu W 2015 *Comput. Soc. Netw.* **2** 17
[24] Antulov-Fantulin N, Lančić A, Šmuc T, Štefančić H and Šikić M 2015 *Phys. Rev. Lett.* **114** 248701
[25] Shah C, Dehmamy N, Perra N, Chinazzi M, Barabási A L, Vespignani A and Yu R 2020 arXiv:2006.11913
[26] Guimera R, Danon L, Diaz-Guilera A, Giralt F and Arenas A 2003 *Phys. Rev. E* **68** 065103
[27] Watts D J and Strogatz S H 1998 *Nature* **393** 440