IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 4, 2021, accepted February 9, 2021, date of publication February 23, 2021, date of current version March 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3061599

Source Identification of Asymptomatic Spread on Networks

HUAN LIU[1], QING BAO[1], HONGJUN QIU[1], MING XU[1], AND BENYUN SHI[2]
[1]School of Cyberspace, Hangzhou Dianzi University, Hangzhou 310018, China
[2]School of Computer Science and Technology, Nanjing Tech University, Nanjing 211800, China

Corresponding authors: Qing Bao (qbao@hdu.edu.cn) and Benyun Shi (benyunshi@outlook.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61806061, Grant 61572165, Grant 61702150, and Grant 61803135; in part by the Key Research and Development Plan Project of Zhejiang Province under Grant 2017C01065; and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LQ19F030011.

ABSTRACT Identifying the sources of spreading dynamics on networks has drawn extensive attention in recent years, where a variety of epidemic models have been adopted to simulate the propagation dynamics among network nodes, such as the Susceptible-Infectious (SI) model. The objective is to identify the most possible source of infection based on the observed state transition events (e.g., from susceptible to infectious) on a small set of monitored nodes. Most existing studies assumed that once a monitored node is infected, it will be immediately observed. However, in reality, it is likely that for many infectious diseases, a newly infected person becomes infectious without showing any disease symptoms. In this case, the state transition cannot be observed until symptoms arise after a time period (i.e., the incubation period). Accordingly, in this paper, we focus on investigating the source identification problem of asymptomatic spread on complex networks by monitoring only a small number of network nodes. Specifically, we adopt a continuous-time SI model with a contagious incubation period to simulate the asymptomatic spread on networks, where the length of the incubation period is assumed to be an independent and identically distributed exponential distribution. In doing so, we formulate the source identification problem as a likelihood maximization problem and solve it using the Monte Carlo approximation and importance sampling. Finally, we validate the performance of our method on both synthetic and real-world networks with different experimental settings. The results show that our method can achieve higher identification accuracy than several benchmark methods.

INDEX TERMS Source identification, asymptomatic spread, incubation period, susceptible-infectious model, Monte Carlo approximation.

I. INTRODUCTION

Understanding and control the spreading dynamics on networks has attracted a considerable number of studies in the field of complex networks for many years [1]–[4]. A great deal of research has been focusing on investigating the properties of various spreading dynamics on networks, such as the propagation of worms on computer networks [5], the spread of rumors on social networks [6]–[9], and the epidemic of infectious diseases on human contact networks [10]–[14]. Evidence has shown that the spread of malicious information or diseases on networks can cause immeasurable losses. To prevent and control a harmful spread on networks, one of the most effective ways is to identify the possible sources of the propagation (i.e., the origin of a spread). If the origin could be identified in the early stage, it will have far-reaching applications in developing intervention and control strategies to curtailing the spread, and reducing the potential losses incurred. For example, finding the rumor-monger may reduce disinformation in a social network [15]–[17]; identifying patient zeros may help control an epidemic [18], [19]; locating the source of a computer worm can help improve the security of computer networks [20]. In this paper, we aim to address the problem of estimating the epidemic origin on networks.

In recent years, extensive methods have been proposed to solve the problem of source identification on networks [17]–[25]. Given a network and the observed configurations, the object is to determine the most probable source of propagation on the network. Early studies focused

The associate editor coordinating the review of this manuscript and approving it for publication was Dimitrios Katsaros.

on identifying the propagation sources on tree-like regular networks [22]. Then, the problem was generalized to estimate the propagation sources in complex networks [19], [24]. Along this line, by analyzing the properties of network structure, a set of graph-centrality measures have been proposed, such as the distance centrality or the spectral centrality [22]. However, relying on structural information alone cannot reflect the characteristics of spreading dynamics on networks, let alone estimate the source of spreading dynamics.

Another challenge of source identification lies in the stochastic nature of spreading dynamics on networks, where different initial conditions can lead to the same observed configuration. Previous work used discrete-time sequential propagation models such as the independent cascade model [26]. However, it is difficult to reflect real-world situations. Later on, various epidemic models were used to simulate the spreading dynamics, such as the Susceptible-Infectious (SI) model, the Susceptible-Infectious-Recovered (SIR) model, and the Susceptible-Infectious-Susceptible (SIS) model [15], [20]–[22], most of which assumed that a newly infected person will immediately become infectious. However, in social networks, users may hesitate to believe a confusing message after receiving it, and will not forward it immediately [8], [9]. Accordingly, the Susceptible-Exposed-Infectious-Recovered (SEIR) model was introduced to involve the latent period, that is, the period from the time of infection to the time of becoming infectious [27].

In reality, for many infectious diseases, infected individuals do not necessarily have symptoms when they are infectious (namely, asymptomatic spread). For example, a person infected by the coronavirus disease 2019 (COVID-19) may be infectious before they develop symptoms [28]–[31]. In other words, it is difficult to determine when the person was infected due to the uncertainty of the incubation period, which is defined as the time period between when an individual gets infected and when the symptoms start. Concerning the epidemic source identification, when a node in the network has symptoms, it may have been infected for a long time and may have led to new infections in the network. This will greatly increase the difficulty of the problem. In this paper, we focus on investigating the source identification problem of asymptomatic spread on networks with partially observed propagation traces, which record the time when symptoms appeared on a small set of monitored nodes.

Specifically, we use the continuous-time SI model with contagious incubation periods to simulate the asymptomatic spread on networks. Then, we propose a novel source identification method to estimate the source nodes based on partially observed state changes of the censored node. The main contributions of this paper are summarized as follows:

1) We present a source identification problem for the asymptomatic spread on networks, taking into consideration the uncertainty of the contagious incubation period. So far as we know, this is the first work to identify the epidemic origin of asymptomatic spread.
2) We adopt the continuous-time SI model with varying incubation periods, the length of which depends on the personal physical condition and follows an exponential distribution. Moreover, the propagation sources are estimated based only on the observations about the onset of symptoms of a small number of nodes.
3) We present a Source Identification of Asymptomatic Spread (SIAS) method that casts the source identification problem as a maximum likelihood estimation problem, which maximizes the likelihood of observed propagation traces under the propagation model. Specifically, we propose an efficient importance sampling method to approximate the objective function.
4) We conduct a series of simulations on both synthetic and real-world networks with different settings to evaluate the properties and performance of the proposed SIAS method.

The remainder of this paper is organized as follows. In Section II, we introduce the related work about source identification on networks. In Section III, we formulate the source identification problem based on the Susceptible-Infectious model with asymptomatic spread. In Section IV, we present the SIAS method in detail. Then, we carry out experiments to evaluate the performance of the proposed SIAS method in Section V. Finally, we conclude this work in Section VI.

II. RELATED WORK

In the past decades, a variety of methods have been proposed to identify the propagation sources by maximizing the likelihood of the observed traces [17], [21], [24], [32], [33]. Most early studies focus on spreading dynamics on tree-like networks, which is simulated using the traditional epidemic models [15], [20], [22]. For example, Shah and Zaman proposed a rumor centrality metric and proved that the node with the largest rumor centrality can maximize the likelihood of the observed data [20], [22]. Accordingly, Dong *et al.* further proposed a local rumor center method to identify propagation sources, which reduced the seeking scale for the origin of spread [15]. Zhu and Ying proposed a novel Jordan center method for the Susceptible-Infectious-Recovered (SIR) model and proved that the source node is more likely to be at the Jordan center of the network [34].

Later on, researchers extended the source identification problems from tree-like networks to general networks [18], [19], [24], and temporal networks [35]–[37]. Early studies tried to estimate the propagation origin based on a complete steady-state snapshot at a given time. Taking the estimation of an epidemic origin as an example, we can only observe *which* nodes got infected rather than *when* they did so. In doing so, the information about the nodes to which the spread did not reach cannot be fully explored [18]. In this case, many researchers tried to identify the sources by injecting a set of sensors in networks in advance such that the specific state changes of the sensor nodes and the corresponding time of infection can be observed [23], [38]–[40].

The source identification method based on partial observations of sensor nodes was first proposed by Pinto *et al.* [38]. In their study, the authors have assumed the propagation time for each edge follows a Gaussian distribution. Then, the propagation origin can be identified based on the differences in time of infection among those sensor nodes. Recently, Yang *et al.* extended the Gaussian estimation method to a direction-induced search-based Gaussian estimator, which can identify propagation sources in general networks with high accuracy [25]. Along this line, many other researchers have tried to estimate the origin based on propagation traces among network nodes [41]–[43]. Usually, only a small fraction of network nodes are monitored and, if infected, their infection time can be observed [44]–[46]. The challenge lies in that we need to unroll the incomplete traces into the past to pinpoint the source.

The spreading dynamics on networks directly determine the difficulty of the source identification problem. Many existing studies used epidemic models to simulate the spreading dynamics on networks, such as the Susceptible-Infectious-Recovered (SIR) model and the Susceptible-Infectious-Susceptible (SIS) model [18], [47]. Further, the latent period was considered by the epidemic models, i.e., the time interval between when an individual is infected and when he or she becomes infectious [48], [49]. While for some diseases, it takes a while for an individual to develop symptoms after infection [28]–[31]. In this case, the contagious incubation period was considered to model the epidemic dynamics [10], [12], [50]. For example, Dhar *et al.* studied the role of the contagious incubation period based on the SIS model with infectious incubated state [50]. Yu *et al.* also extended the SIS model and proposed a corresponding epidemic model with the contagious incubation period on directed and heterogeneous networks [12]. Further, Zhu *et al.* proposed a generalized epidemic model on complex heterogeneous networks based on the SIR model with the incubation period, and studied how the heterogeneous connectivity patterns and the underlying network structures affect the disease propagation [51]. In this paper, we aim to tackle the source identification problem for the asymptomatic spread on networks, using the SI model with a varying incubation period.

III. PROBLEM STATEMENT

In this section, we first introduce a continuous-time SI model with the incubation period to characterize the asymptomatic spreading on networks. Then, we formulate the source identification problem to be a likelihood maximization problem based on partially observed propagation traces on network nodes.

A. ASYMPTOMATIC SPREAD ON NETWORKS

Let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be a directed network, where $\mathcal{V}$ is defined as a set of nodes and $\mathcal{E} = \{(i, j)|i, j \in \mathcal{V}\}$ is defined as a set of edges between nodes i and j. In this paper, we focus on the asymptomatic spread on $\mathcal{G}$ using the Susceptible-Infectious (SI) model, where a newly infected node may become infectious without showing any symptoms. Therefore, each node in $\mathcal{G}$ should be in any of the following three states: (i) Susceptible (denoted as S), nodes that have not been infected; (ii) Infectious but asymptomatic (denoted as A), nodes that are infectious but asymptomatic; and (iii) Infectious and symptomatic (denoted as I), nodes that are contagious and symptomatic.

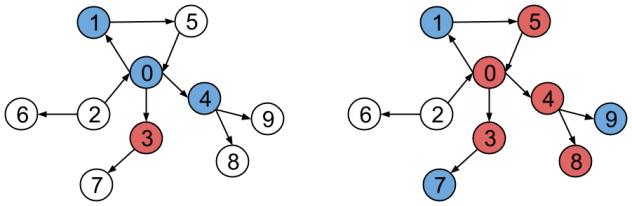


FIGURE 1. An example of the infection process on a directed graph under the Susceptible-Infectious model with asymptomatic spread. Two snapshots are taken recording the states of nodes during the epidemic spreading. Nodes in the *S*-state, *A*-state, and *I*-state are colored with white, blue, and red, respectively. Initially, the node 0 is infectious but asymptomatic (*A*-state), and all other nodes are in *S*-state. (a) After a period of transmission, nodes 1, 3 and 4 are infected by the source node 0. All of them become infectious while nodes 1 and 4 are asymptomatic (*A*-state). (b) After the incubation period, node 0 and node 4 change to be *I*-state. Meanwhile, nodes 7 and 9 are infected and change to *A*-state; Nodes 5 and 8 have passed a short incubation period and become *I*-state.

Figure 1 demonstrates the process of asymptomatic spreading on a network. Initially, all nodes except for the source node s are in S-state. Starting from the source node s, an epidemic propagates along its out-going edges to its direct neighbors and further spread to other nodes in $\mathcal{G}$. When a node i is infected, its state will change from S to A. Then, it will take a time period φ_i (i.e., the incubation period) to develop disease symptoms. After the incubation period φ_i, the node enters I-state. We cannot determine whether a node is infected or not before it enters I-state. Note that an infectious node can infect multiple neighboring nodes, but it must be infected by one of its neighboring nodes.

The force of infection along each edge (i, j) is described by τ_{ij}, which represents the time of transmission from i to j. In this paper, we assume that τ_{ij} is randomly drawn from a probability density function $f(\tau_{ij}; \alpha_{ij})$ parameterized by a transmission rate α_{ij}. Moreover, because a node cannot be infected by another node infected later in time, we assume that the force of infection τ_{ij} is independent and non-negative, that is, if $\tau_{ij} < 0$, $f(\tau_{ij}; \alpha_{ij}) = 0$. With respect to the incubation period, we assume that φ_i is randomly drawn from a probability density function $g(\varphi_i; \beta_i)$ parameterized by a transition rate β_i. Similarly, we assume the $\{\varphi_i\}$ is independent and non-negative, that is, if $\varphi_i < 0$, $g(\varphi_i; \beta_i) = 0$.

A temporal trace $\boldsymbol{t^I} = (t_1^I, \cdots, t_N^I)$ will be observed after the propagation process, which records when the symptoms appear on each node. Here, $t_i^I \in [0, \infty]$ and ∞ denotes the corresponding node is not infected during the observation window. However, in many real-world scenarios, due to the large network size, we cannot record the states of all network

nodes. In this paper, we assume that only a limited number of sensors can be placed to observe when the nodes develop symptoms. Accordingly, we represent the observable set of nodes in the networks as $\mathcal{O}$, and the other nodes as $\mathcal{H}$. Meanwhile, the time of infection of all nodes is denoted by an N-dimensional vector $\boldsymbol{t^A} = (t_1^A, \cdots, t_N^A)$, which is unobservable as well. The goal of this paper is to estimate the source of an epidemic spreading on $\mathcal{G}$ from the partially observable traces of nodes in $\mathcal{O}$.

B. SOURCE IDENTIFICATION OF ASYMPTOMATIC SPREAD

Given the times of the subset of nodes $\mathcal{O}$ in the network when the symptoms of the nodes become observable, *i.e.*, $\{t_i^I\}_{i\in\mathcal{O}}$, our goal is to find the source node s and its infection time t_s^A, which maximize the likelihood of the observed data. Thus, we aim to solve

$$s^* = \underset{s\in V}{\operatorname{argmax}} \max_{t_s^A \in (-\infty, \min_{i\in\mathcal{O}} t_i^I)} p\left(\left\{t_i^I\right\}_{i\in\mathcal{O}} | t_s^A\right). \quad (1)$$

where we assume $t_s^A < \min_{i\in\mathcal{O}} t_i^I$ and $p\left(\left\{t_i^I\right\}_{i\in\mathcal{O}} | t_s^A\right)$ is defined as below.

According to the conditional independence relation proposed in the continuous-time model, the complete likelihood of the infection process for both observed and hidden times can be given as:

$$p(\boldsymbol{t^I}, \boldsymbol{t^A} | t_s^A) = \prod_{i\in\mathcal{O}\cup\mathcal{H}} p(t_i^I | t_i^A)\, p\,(t_i^A | \{t_j^A\}_{j\in\pi_i}), \quad (2)$$

where π_i is the set of parents of node i in the directed graph and the likelihood of the incubation period $\varphi_i = t_i^I - t_i^A$, *i.e.*, $p(t_i^I | t_i^A)$, is defined with a probability density function $g(\varphi_i; \beta_i)$. Furthermore, the likelihood of the infection times can be written as below, as is given in [41]:

$$p\left(t_i^A | \left\{t_j^A\right\}_{j\in\pi_i}\right) = \prod_{j\in\pi_i} S\left(t_i^A - t_j^A; \alpha_{ij}\right) \times \sum_{l\in\pi_i} H\left(t_i^A - t_l^A; \alpha_{il}\right),$$

where $S\left(\tau_{ij}; \alpha_{ij}\right) = 1 - F\left(\tau_{ij}; \alpha_{ij}\right)$ is the survival function which represents the probability that node i has not been infected by node j at time t_i^A. $F\left(\tau_{ij}; \alpha_{ij}\right) = \int_0^{\tau_{ij}} f\left(\tau_{ij}; \alpha_{ij}\right) dt$ is the cumulative distribution function computed with the probability density function $f(\tau_{ij}; \alpha_{ij})$, and $H\left(\tau_{ij}; \alpha_{ij}\right) = \frac{f(\tau_{ij};\alpha_{ij})}{S(\tau_{ij};\alpha_{ij})}$ is the hazard function, represented as the instantaneous infection rate.

For the probability density functions, to make our method more suitable for real world propagation data [41], [52], we here focus on the exponential distributions for modeling the delays $f(\tau_{ij}; \alpha_{ij})$ and $f(\varphi_i; \beta_i)$. In this case,

$$f\left(\tau_{ij}; \alpha_{ij}\right) = \frac{1}{\alpha_{ij}} e^{-\frac{\tau_{ij}}{\alpha_{ij}}} = \frac{1}{\alpha_{ij}} e^{-\frac{t_i^A - t_j^A}{\alpha_{ij}}},$$

$$f\left(\varphi_i; \beta_i\right) = \frac{1}{\beta_i} e^{-\frac{\varphi_i}{\beta_i}} = \frac{1}{\beta_i} e^{-\frac{t_i^I - t_i^A}{\beta_i}}.$$

where the transmission rates $\{\alpha_{ij}\}$ and transition rates $\{\beta_i\}$ control the length of propagation delays and conversion delays respectively.

Unfortunately, to use Eq. 2, the state transition times of all the nodes in the network need to be fully observed. As we can only observe the times of the subset of nodes in the network when the symptoms of the nodes become observable, the likelihood of incomplete propagation data is computed as:

$$p(\{t_i^I\}_{i\in\mathcal{O}} | t_s^A) = \int_\Omega \prod_{i\in\mathcal{O}\cup\mathcal{H}} p(t_i^I | t_i^A)\, p\,(t_i^A | \{t_j^A\}_{j\in\pi_i}) \times \prod_{j\in\mathcal{H}} dt_j^I \prod_{i\in O\cup\mathcal{H}} dt_i^A, \quad (3)$$

which marginalizes out the times of all hidden nodes $\mathcal{H}$ and all hidden infection times over a product space $\Omega := [t_s^A, \infty)^{|\mathcal{H}|}$. However, the computation of the incomplete likelihood involves high dimensional integration. In the next section, we will address this difficult problem with a novel importance sampling method.

IV. METHOD

As stated above, to solve the problem of source identification, there are two remaining difficulties. First, how to approximate the high dimensional integration in the objective function. We address this problem by an approximation algorithm based on importance sampling to simplify the calculation. Second, the maximization over the infection time of the source node in Eq. 1 is a non-convex problem. To solve this problem, we utilize the piece-wise structure of the objective function and optimize the objective function with efficient algorithms.

A. APPROXIMATION OF MULTIPLE INTEGRALS

With incomplete observation, it is difficult to directly calculate the likelihood function in Eq. 3 due to multiple integrals. Thus, we consider Monte Carlo approximation methods to sample from the posterior distribution of latent variables given the source time t_s^A and the observed times $\{t_i\}_{i\in\mathcal{O}}$. However, it is still difficult to sample directly from this distribution, so we use importance sampling to approximate the likelihood value [44].

More specifically, we first introduce a set of auxiliary random variables $\{\eta_i^I\}_{i\in\mathcal{O}}$, which corresponds to the observed times, and with its arbitrary joint probability distribution $q(\{\eta_i^I\}_{i\in\mathcal{O}})$. The auxiliary variables will be used in the importance sampling. Given the auxiliary distribution, the objective function becomes:

$$p(\{t_i^I\}_{i\in\mathcal{O}} | t_s^A) = \int_\Omega \prod_{i\in\mathcal{O}\cup\mathcal{H}} p(t_i^I | t_i^A) p(t_i^A | \{t_j^A\}_{j\in\pi_i}) \times q(\{\eta_i^I\}_{i\in\mathcal{O}}) \prod_{j\in\mathcal{H}} dt_j^I \prod_{i\in O\cup\mathcal{H}} dt_i^A \prod_{i\in\mathcal{O}} d\eta_i^I. \quad (4)$$

Then, we use importance sampling on the auxiliary and hidden variables by introducing the proposal distribution

$\widetilde{q}(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}})$, which gives:

$$\begin{aligned}
&p(\{t_i^I\}_{i\in\mathcal{O}}|t_s^A)\\
&= \int_\Omega \frac{\prod_{i\in\mathcal{O}\cup\mathcal{H}} p(t_i^I|t_i^A)p(t_i^A|\{t_j^A\}_{j\in\pi_i})}{\widetilde{q}(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}})} q(\{\eta_i^I\}_{i\in\mathcal{O}})\\
&\quad\times\widetilde{q}(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}})\\
&\quad\times\prod_{j\in\mathcal{H}} dt_j^I \prod_{i\in\mathcal{O}\cup\mathcal{H}} dt_i^A \prod_{i\in\mathcal{O}} d\eta_i^I\\
&\approx \frac{1}{L}\sum_{l=1}^{L} \frac{p(\{t_i^I\}_{i\in O}, \{(t_i^I)^l\}_{i\in H}|\{(t_i^A)^l\}_{i\in O\cup H})}{\widetilde{q}(\{(\eta_i^I)^l\}_{i\in\mathcal{O}}, \{(t_i^A)^l\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{(t_i^I)^l\}_{i\in\mathcal{H}})}\\
&\quad p(\{(t_i^A)^l\}_{i\in O\cup H}|t_s^A) q(\{(\eta_i^I)^l\}_{i\in\mathcal{O}})\\
&\triangleq \phi_L(t_s^A),
\end{aligned} \tag{5}$$

where we draw L samples from the proposal distribution $\widetilde{q}(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}})$ to approximate the multiple integrals. Here we define the proposal distribution with the generative process of the epidemic spreading. In particular, the proposal distribution $\widetilde{q}(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}})$ will be $p(\{\eta_i^I\}_{i\in\mathcal{O}}, \{t_i^A\}_{i\in\mathcal{O}\cup\mathcal{H}}, \{t_i^I\}_{i\in\mathcal{H}}|t_s^A)$, the distribution of the auxiliary and hidden variables under the epidemic spreading process with s as the source node. In addition, the auxiliary distribution $q(\{\eta_i^I\}_{i\in\mathcal{O}})$ is chosen to be equal to $p(\{\eta_i^I\}_{i\in\mathcal{O}}|\{t_i^A\}_{i\in\mathcal{O}})$, which greatly simplifies the objective function in Eq. 5.

Finally, with the proposal distribution and auxiliary distributions proposed above, we can further simplify the objective function as follows:

$$\begin{aligned}
\phi_L(t_s^A) &= \frac{1}{L}\sum_{l=1}^{L} \frac{p(\{t_i^I\}_{i\in O}, \{(t_i^I)^l\}_{i\in H}|\{(t_i^A)^l\}_{i\in O\cup H})}{p(\{(\eta_i^I)^l\}_{i\in\mathcal{O}}, \{(t_i^I)^l\}_{i\in\mathcal{H}})|\{(t_i^A)^l\}_{i\in O\cup H}}\\
&\quad\times p(\{(\eta_i^I)^l\}_{i\in\mathcal{O}}|\{(t_i^A)^l\}_{i\in O})\\
&= \frac{1}{L}\sum_{l=1}^{L}\prod_{i\in O} p\left(t_i^I|(t_i^A)^l\right).
\end{aligned} \tag{6}$$

It is worth mentioning that as the proposal distribution is defined with the generative process of the epidemic spreading, we sample L sets of infection times in an efficient way, which is independent of the actual value of t_s^A and only depends on the real source node s. More specifically, to sample the infection times $\{(t_i^A)^l\}_{i\in\mathcal{O}\cup\mathcal{H}}$, we first sample transmission times $\{(\tau_{ij})^l\}_{(i,j)\in\mathcal{E}}$ for each edge in the networks, which is independent of the choice of source nodes. Then, for each potential source node s, we utilize the shortest-path first property [53] under this model, which means that the infection propagate through the shortest path, to effectively calculate $\{(t_i^A)^l\}_{i\in\mathcal{O}\cup\mathcal{H}}$ with different source node infection times t_s^A. Let $\mathcal{Q}_i(s)$ be the set of directed paths from source node s to a node i, where each path $q \in \mathcal{Q}_i(s)$ contains a sequence of directed edges (j, n), and we assume the infection time of the source node to be t_s^A, thus we can calculate $(t_i^A)^l$ as follows:

$$\begin{aligned}
(t_i^A)^l &= g_i(\{(\tau_{jn})^l\}_{(j,n)\in E}|s) + t_s^A\\
&= \min_{q\in\mathcal{Q}_i(s)} \sum_{(j,n)\in q} (\tau_{jn})^l + t_s^A,
\end{aligned} \tag{7}$$

where $g_i(\cdot)$ is the value of shortest-path from source node s to node i.

B. MAXIMIZING THE OBJECTIVE FUNCTION

The objective function, given by Eq. 1, consists of two layers of maximization. In the outer layer, we only need to rank the maximum likelihood values corresponding to all candidate source nodes. The node with the largest value is estimated as the source node s^*. The outer layer maximization is straightforward, however, it involves the inner layer maximization, which requires calculating the likelihood value of each candidate source node at its optimal starting time t_s^A. To this end, we utilize the Monte Carlo approximation and aim to find the optimal value t_s^A to maximize $\phi_L(t_s^A)$. That is,

$$\max_{t_s^A\in(-\infty, \min_{i\in\mathcal{O}} t_i^I)} \phi_L(t_s^A), \tag{8}$$

Although the inner maximization is a one-dimensional problem, defined in Eq. 6, the objective function is piecewise continuous with respect to t_s^A. The reason is that as the value of the start time t_s^A increases, the value of the asymptomatic infection time $(t_i^A)^l$ sampled by an observation node i will exceed its observed infection time t_i^I, which is inconsistent with the concept of the non-negative incubation times mentioned above. But by using the characteristics of the objective function, we can effectively find the maximum value in each of its continuous pieces.

To solve the inner layer maximization problem, we first need to find all the endpoint $t_{s_i}^A$ of each continuous piece in the approximated likelihood function $\phi_L(t_s^A)$. With the starting time t_s^A of the candidate source node increasing, when the infection time $(t_i^A)^l$ of any observation node i is equal to its observed infection time t_i^I, the starting time t_s^A corresponding to the current source node is the change point. More specifically, it can be seen from Eq. 6 that the objective function $\phi_L(t_s^A)$ can be mainly composed of one part: $p(t_i^I|(t_i^A)^l)$. The change of t_s^A will affect the difference between the asymptomatic infection time $(t_i^A)^l$ and the observation time t_i^I of the observed node i. When the asymptomatic infection time $(t_i^A)^l$ of the observed node i obtained by sampling is equal to t_i^I, the current t_s^A is the change point. We set $t_s^A = 0$ at the beginning and computing the asymptomatic infection time $(t_i^A)^l$ for each observed node $i \in \mathcal{O}$ and realization l through sampling and the shortest path characteristics. Then we can find the change points by computing the time difference $t_i^I - (t_i^A)^l$, $i \in \mathcal{O}$, $l = 1, \ldots, L$ with the starting time t_s^A increasing. If the time difference is equal to zero, we record the current change point $t_{s_i}^A$. Once the time difference of an observation node is negative, the node will be ignored in the subsequent process and the related time difference will no longer be calculated, since the time difference will never be

greater than zero. After that, the inner maximization problem is transformed into finding the maximum values for the continuous pieces. Moreover, within each continuous piece, the objective function is a monotonic increasing function with respect to t_s^A, as $\phi_L(t_s^A)' > 0$ in each piece. Thus, we can get the maximum value of each continuous piece at the right endpoint. Finally, we can obtain the maximum value of the objective function by comparing the maximum values calculated for each piece.

We summarize the above algorithm in Algorithm 1.

Algorithm 1 SIAS: Source Identification of Asymptomatic Spread

Require: $G, \boldsymbol{t^I}, \{\alpha_{ij}\}_{(i,j)\in E}, \{\beta_i\}_{i\in V}, L$
Ensure: $s^*, t_{s^*}^A$
Sample L sets of transmission times $\{\tau_{ij}\}_{(i,j)\in E}$
Compute estimated infection times $(\hat{t}_{i\in\mathcal{V}}^A)^l$, $l = 1, \ldots, L$ assuming $t_s^A = 0$ for each candidate source node s using Eq. 7
Compute change points: $t_i^I - (\hat{t}_i^A)^l$, $i \in \mathcal{O} \cap t_i^I \geq (\hat{t}_i^A)^l$, $l = 1, \ldots, L$, for each candidate source node s
for $i \in V$ **do**
  $t_i^* = argmax_{t_i^A}\phi_L(t_i^A)$ (the maximum value of each continuous piece is at the right end change point)
end for
$s^* = argmax_{i\in\mathcal{V}}\phi_L(t_i^*)$

C. COMPUTATIONAL COMPLEXITY

As is summarized in Algorithm 1, implementing the learning algorithm involves two main steps: 1) sample transmission times, compute estimated infection times and change points; 2) optimize the inner layer of maximization for each candidate source node and then obtain the optimal source node.

For the first step, the cost for sampling L sets of the transmission times $\{\tau_{ij}\}_{(i,j)\in\mathcal{E}}$ is $O(L \times |\mathcal{E}|)$. To obtain the L sets of estimated infection times $(\hat{t}_{i\in\mathcal{V}}^A)^l$ when $t_s^A = 0$ for each candidate source node s using Eq. 7, we need to apply the Dijkstra algorithm for each source node $s \in \mathcal{V}$ to calculate the shortest path. As the complexity of Dijkstra algorithm is $O(|\mathcal{V}|^2)$ in general, the complexity to obtain the estimated infection times is thus $O(L \times |\mathcal{V}|^3)$. In addition, to compute the change points for each candidate source node whenever a sample of an observed node i reaches the condition that $t_i^I = (t_i^A)^l$, the complexity is $O(|\mathcal{V}| \times L \times |\mathcal{O}|)$. For the second step, to calculate $t_i^* = argmax_{t_i^A}\phi_L(t_i^A)$ corresponding to each candidate source node i, we need to calculate the value of $\phi_L(t_i^A)$ at each change point. According to Eq. 6, the complexity for calculating $\phi_L(t_i^A)$ each time is $O(L \times |\mathcal{O}|)$, and we need to compare the maximum values for each piece, thus the complexity for each candidate node becomes $O(L^2 \times |\mathcal{O}|^2 + |\mathcal{O}| \times L)$. As there are $|\mathcal{V}|$ candidate nodes i, the complexity becomes $O(|\mathcal{V}| \times (L^2 \times |\mathcal{O}|^2 + |\mathcal{O}| \times L))$. At last, we need to compare the values $\phi_L(t_i^*)$ for all candidate nodes i and obtain the optimal source node, and the corresponding complexity is $O(|\mathcal{V}|)$. As the set of observed nodes $\mathcal{O}$ is a subset of V, the worst-case complexity for our algorithm is thus $O(L^2 \times |\mathcal{V}|^3)$.

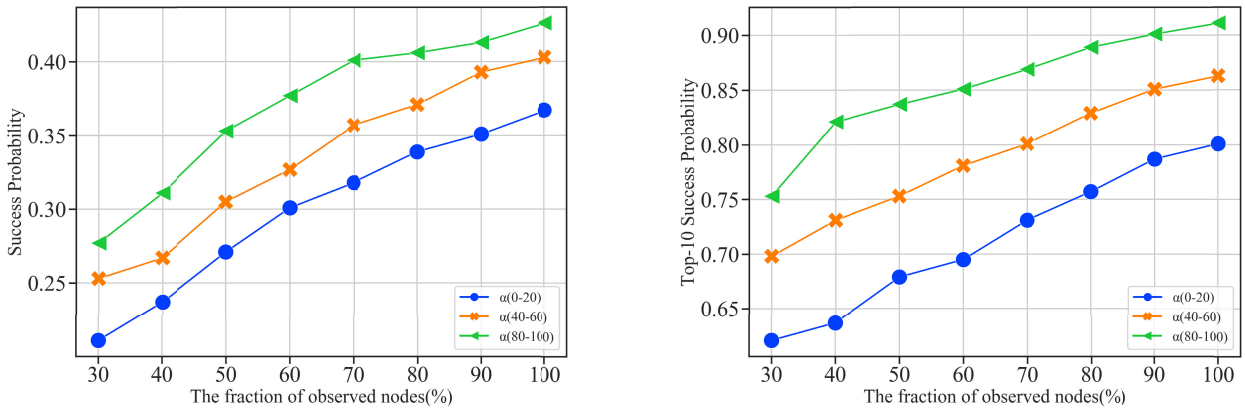
V. EXPERIMENTS

In this section, we conducted a series of simulation experiments to evaluate the performance of our algorithm and studied the influence of different parameters involved in the source node identification process on the experimental results. First, we investigated the accuracy of our method under infection processes with different transmission rates and transition rates. Then, we studied the impact of observer selection strategies for locating the source using our method. Finally, we evaluated the performance for source identification using both synthetic and real network data sets, and compared it with other benchmark methods. Except when it is explicitly mentioned, the transmission rates and transition rates are set by drawing samples from uniform distribution $\alpha \sim U(0, 100)$ and $\beta \sim U(0, 100)$ [41]. And in the experiments about studying the characteristics of the algorithm, we first randomly generated 10 scale-free networks of 1024 nodes and 2000 edges using the SNAP platform [54] as real networks are mostly scale-free. Then, in each setting, the simulations of the infection process were carried out from ten different random sources for each network respectively. In addition, we only considered the source nodes that triggered at least 200 nodes to be infected as we are interested in detecting vital source nodes. In all the following experiments, we set the sample size to be 500, and evaluated the method with observed infected nodes in different proportions as only partial observations are available in real situations. The results showed that our approach can discover the true source of the spread with high accuracy.

A. IMPACT OF TRANSMISSION RATES

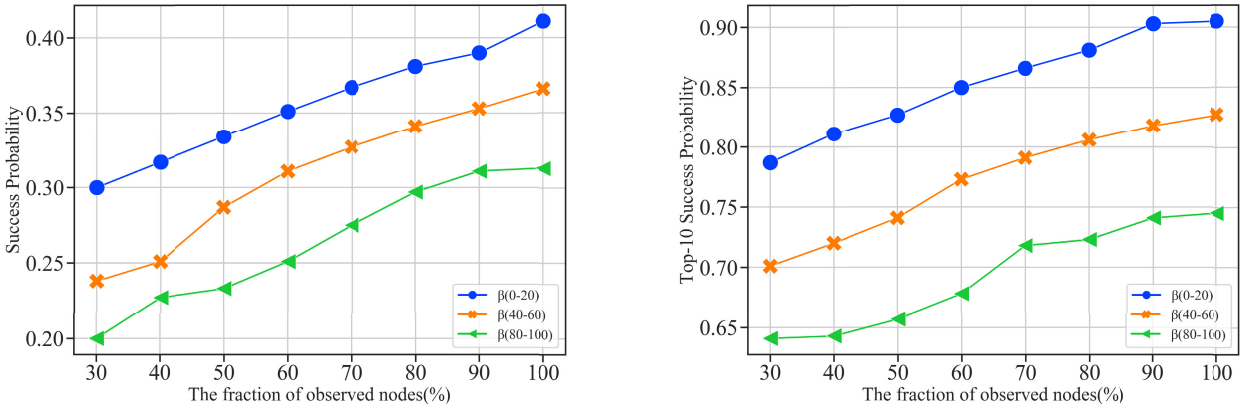
We compared the accuracy of our source identification method under different transmission rates α when setting the value range of transition rates as $\beta \sim U(0, 100)$ to study the influence of the transmission rate on the experimental results. Here we set the transmission rates α of different edges by drawing samples from $\alpha \sim U(0, 20)$, $\alpha \sim U(40, 60)$, and $\alpha \sim U(80, 100)$ respectively.

Figure 2 showed the impact of transmission rates α on the accuracy of our source identification method. We used the success probability (SP) and top-10 success probability (Top-10) to measure accuracy. We defined the success probability as the probability of finding the true source in all cases, and the top-10 success probability as the probability that the true source is among the top-10 nodes ranked according to the likelihood estimation. It can be observed that our method achieves high accuracy for source identification and the accuracy increases with the transmission rates α. As the propagation delays $\{\tau_{ij}\}$ along edges increase with the transmission rates $\{\alpha_{ij}\}$, the infection spreads more slowly with larger values of transmission rates, which increases the



(a) Success Probability (b) Top-10 Success Probability

FIGURE 2. The impact of transmission rates α on the performance of our method in terms of (a) the Success Probability (SP) and (b) the TOP-10 Success Probability (Top-10 SP).



(a) Success Probability (b) Top-10 Success Probability

FIGURE 3. The impact of transition rates β on the performance of our method in terms of (a) the Success Probability (SP) and (b) the TOP-10 Success Probability (Top-10 SP).

gap between the infection time of different nodes, reducing the difficulty of identifying the source node.

B. IMPACT OF TRANSITION RATES

To study the impact of transition rates β on our method, we compared the experimental results of different transition rates when the transmission rates α is not limited. We set the transition rates β of different nodes in the networks by drawing samples from $\beta \sim U(0, 20)$, $\beta \sim U(40, 60)$, and $\beta \sim U(80, 100)$ respectively. It can be observed from Figure 3 that our method can achieve higher accuracy with smaller values of transition rates β. The reason is that the length of the incubation period is related to transition rates β. As the transition rates $\{\beta_i\}$ increase, the incubation periods $\{\varphi_i\}$ of the nodes in the network become longer, which makes it more difficult to accurately determine the relationship between the observed infection time t_i^I and the true infection time t_i^A of the sensor nodes in the networks, and there are more possible candidate source nodes. When there are more candidate source nodes, the source of propagation cannot be accurately identified.

C. IMPACT OF OBSERVER SELECTION STRATEGIES

We further evaluated the performance of our method when the observed nodes are chosen with different strategies to investigate the impact of observer selection strategies on our algorithm. In real situations, as the only limited number of sensors can be placed to get the time when the nodes show symptoms, we need to select observation nodes to make better use of the sensors. Here we adopted three widely used metrics to measure node importance: 1) degree centrality, the degree of nodes [55] in the networks; 2) PageRank score, calculated with PageRank algorithm [56]; 3) closeness centrality, the inverse of the average distance to all other nodes in the networks [55]. Accordingly, to manifest the effect of considering node importance, we compared the two strategies for each metric: 1) the nodes with highest centrality values as observers, and 2) the nodes with lowest centrality values as observers. Generally, selecting nodes with high centrality

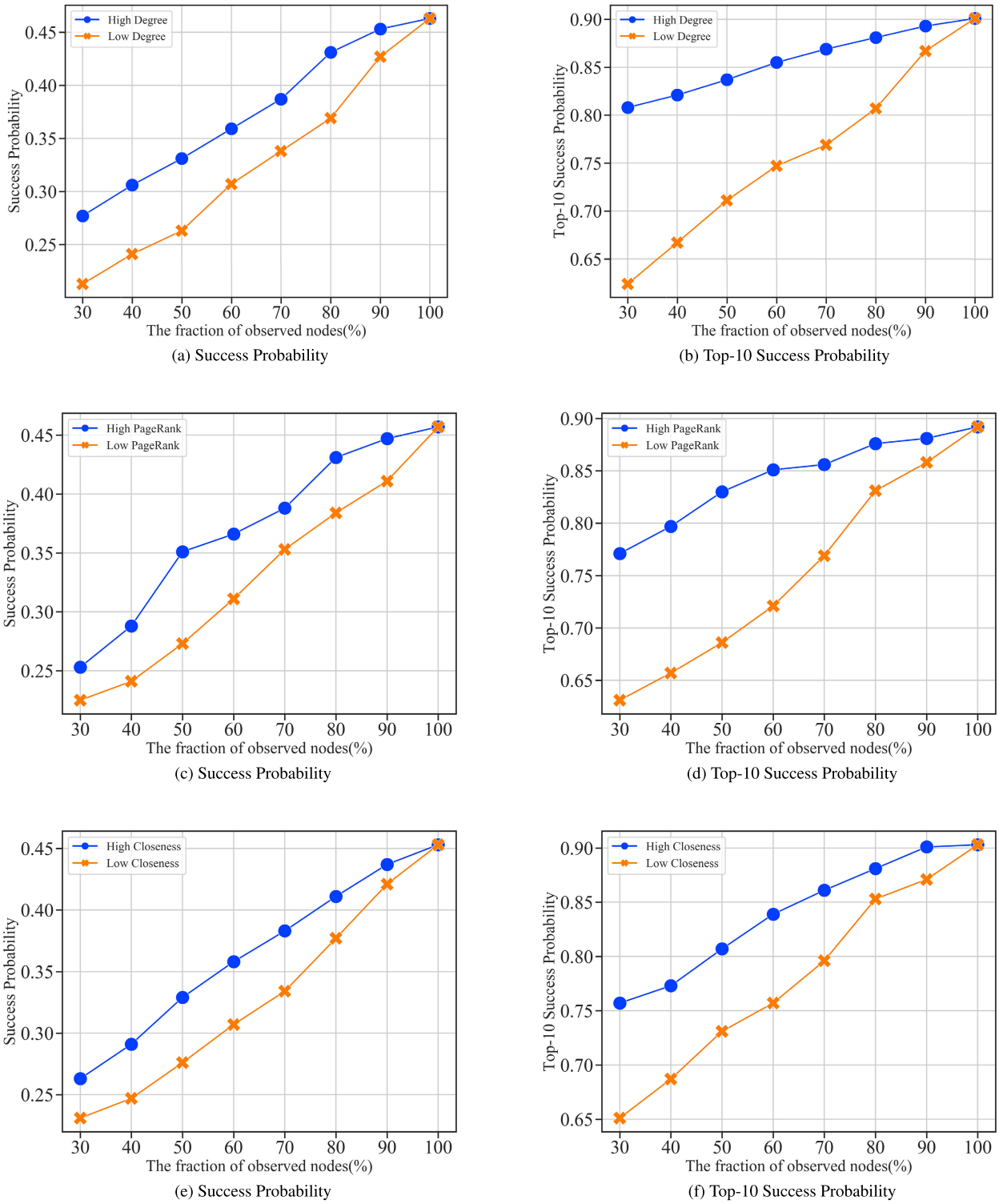


FIGURE 4. **The impact of different observer selection strategies on the performance of our method in terms of three metrics for measure node importance: degree centrility, pagerank score and closeness centrality.**

values as sensor nodes may contain more infection information. However, the nodes with high centrality value may be connected to each other, which will induce some redundant observers to decrease the accuracy of the algorithm. For this reason, we not only selected nodes with high centrality values as observation nodes for experiments, but also tested the performance of the method when the observation nodes have low centrality values.

Figure 4 showed that the accuracy of our approach under different node selection strategies for the three metrics.

The results identified that our algorithm performs better when the nodes with the highest centrality values are chosen as observers than the ones with the lowest centrality values. Therefore, considering nodes with high centrality values when selecting observer nodes can achieve better performance. As the number of observers increases, the overlap of the node sets involved in the two different strategies is also increasing, which makes the difference between the two strategies gradually decrease, and then disappears when all the nodes in the network are selected as observers.

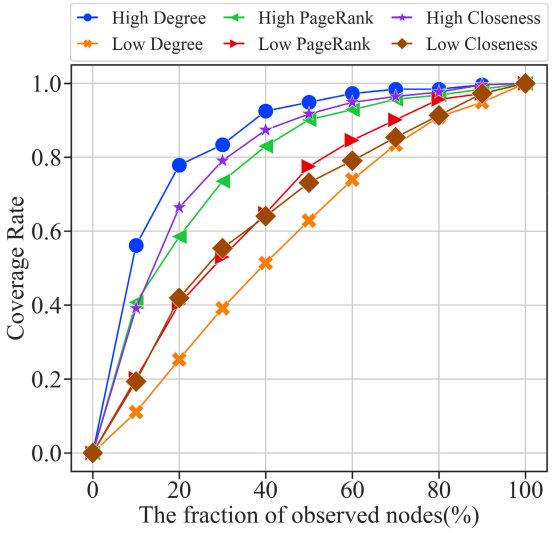


FIGURE 5. Coverage rates of different observer selection strategies with different metrics. The coverage rate is defined as the proportion of all observers and their neighbors in the entire network.

The reasons behind this can be explained as follows. When the observers are close to the true source, the uncertainty of the accumulated delay along the propagation path will be smaller, and the accuracy of identifying the true source will be higher. In other words, selecting the observers close to the true source can achieve higher accuracy. As there is no prior information of the true source, if the observers cover more nodes in a network or there are more neighbors around the observers, the more likely that the distance between the true source and its nearest observer is shorter, which will improve the accuracy of source identification. Therefore, we used the coverage rate to measure the coverage states of observers in a network, which is defined as the proportion of all observers and their neighbors in the entire network. Figure 5 showed the coverage rates of different observer selection strategies with different metrics. As anticipated, we observed that the coverage rates are higher when the nodes with the highest degree centrality, PageRank score, and closeness centrality are chosen as observers than the ones with the lowest degree centrality, PageRank score, and closeness centrality.

D. EXPERIMENTS ON DIFFERENT NETWORKS

1) SYNTHETIC NETWORKS

We generated three typical types of synthetic networks with Kronecker graph approach [57]: (i) core-periphery networks (parameter matrix: [0.9 0.5; 0.5 0.3]), which mimic the real world networks [42], (ii) random networks ([0.5 0.5; 0.5 0.5]), typically used in studies of physics and graph theory [58], and (iii) hierarchical networks ([0.9 0.1; 0.1 0.9]) [59]. In order to verify the effectiveness of the algorithm, we generated networks of 1024 nodes and 2000 edges for each type of Kronecker network.

Figure 6 showed the performance of our method on synthetic networks. It can be observed that our algorithm achieves high accuracy for source identification and performs obviously better in hierarchical networks than the other two types of networks. We investigated the temporal traces left by the infection processes and found that in most cases the sets of infected nodes caused by different source nodes did not overlap much in hierarchical networks, as compared with the other two types of networks. This is due to the structure of hierarchical networks, where infection cannot propagate far from current branches. As a result, the likelihood values of many nodes equal zero, which narrows down the range of candidate source nodes. Besides, the performance on random networks is better than that on core-periphery networks. This is also due to the difference in the network structure. The core-periphery networks entail densely-connected core nodes and sparsely-connected periphery nodes, while the edges in the random networks are uniformly generated. Therefore, the network core has more available propagation paths, which brings difficulties for the identification of source nodes in core-periphery networks.

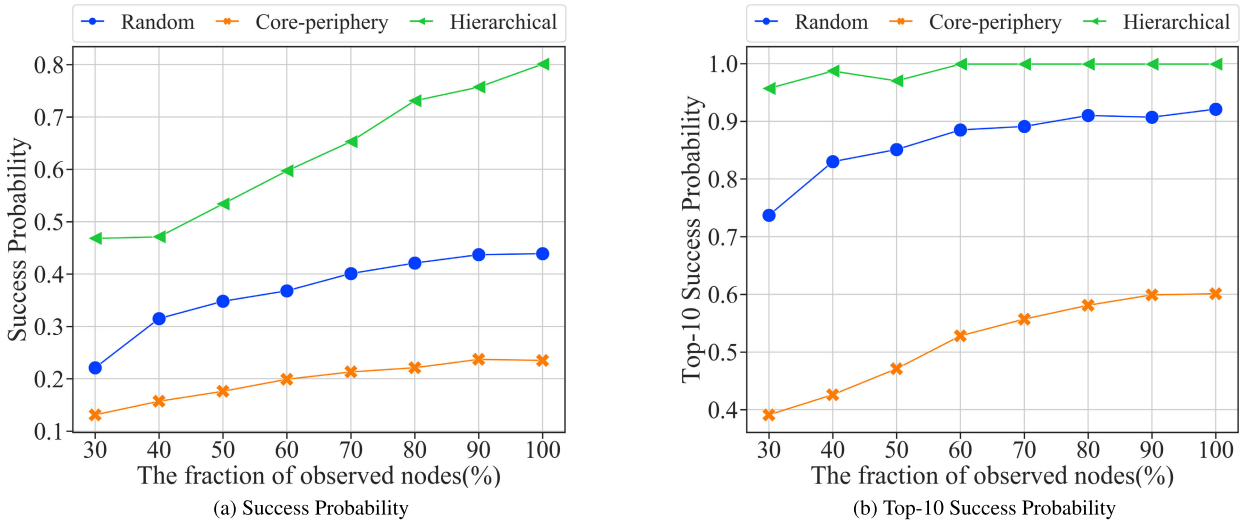
2) REAL NETWORKS

Specifically, the experiments are carried out in the following six real-world networks, which have been widely used for the research of source identification in complex networks:

- *Enron email network* [60]: the email communication network described the email connection of employees in the Enron Corporation.
- *Doctor friendship network* [61]: the network captured innovation spread among physicians in Illinois, Peoria, Bloomington, Quincy, and Galesburg.
- *USAir network* [62]: the network detailed the US air transportation system.
- *Food web network* [63]: the network was formed by the food web in Florida Bay during the dry season.
- *Email network* [64]: the data of this e-mail network was collected from the relationship between members of the University Rovira i Virgili (Tarragona).
- *Facebook-like social network* [65]: the network originated from an online community for students at the University of California, Irvine.

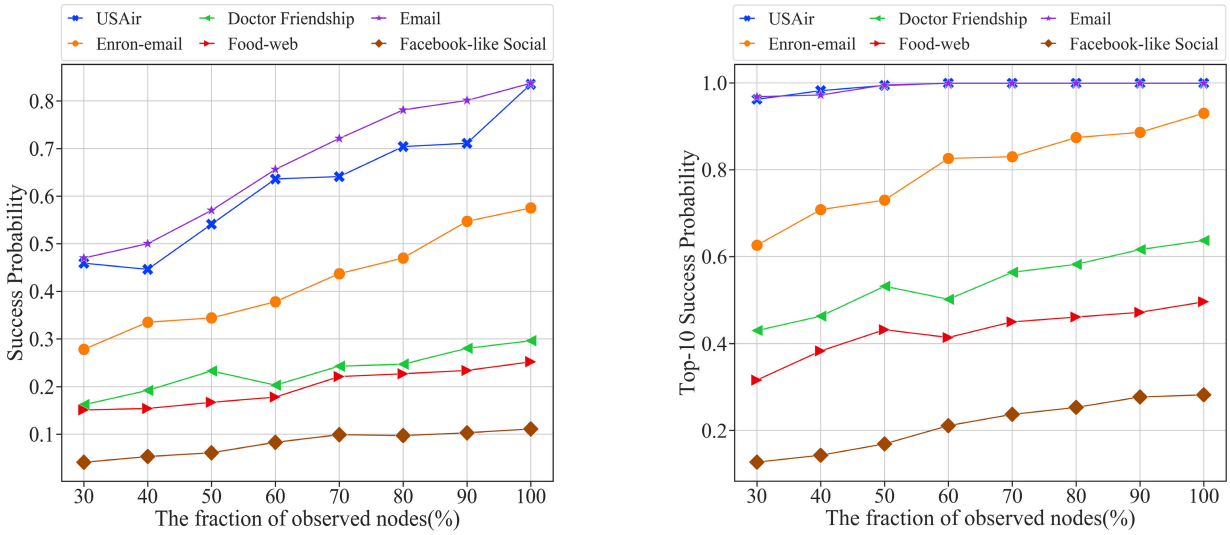
The basic topological features of the six real networks are shown in Table 1.

Figure 7 showed the performance of our method on real networks. We observed that the performance is apparently better in the Email and USAir network than in other networks. The Email network is divided into multiple different communities, and the final set of infected nodes is different for



(a) Success Probability (b) Top-10 Success Probability

FIGURE 6. The performance of our method on synthetic networks in terms of (a) the Success Probability (SP) and (b) the TOP-10 Success Probability (Top-10 SP).



(a) Success Probability (b) Top-10 Success Probability

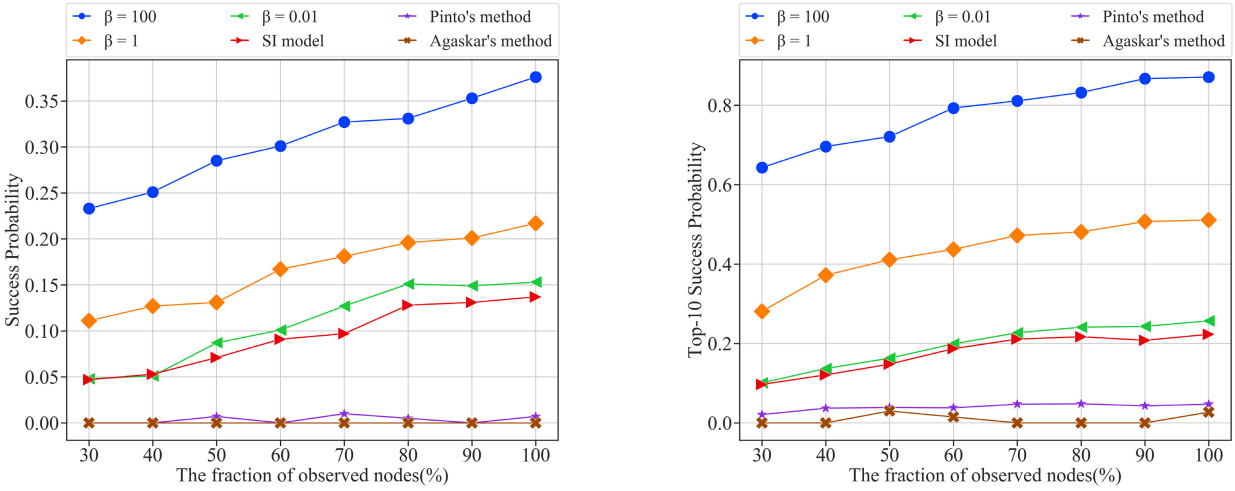
FIGURE 7. The performance of our method on real networks in terms of (a) the Success Probability (SP) and (b) the TOP-10 Success Probability (Top-10 SP).

different propagation source nodes. Therefore, except for the true source node, the maximum likelihood estimates of other candidate source nodes are all equal to zero, which makes it possible to identify the propagation source more accurately. The situation on the USAir network is similar, in an air transportation system, major airports in different areas are interconnected to facilitate long-distance travel. Meanwhile, small airports within an area are interconnected locally and they are also connected to their nearby major airport. Due to the structure of the USAir network, the infection cannot easily propagate to remote areas, and in most cases, the sets of infected nodes caused by different source nodes cannot overlap much. Thus, the likelihood values of many nodes are equal to zero, and the true source can be easily located through the final infection networks. For the other four real networks, we calculated the respective network characteristics as shown in Table 1, and found that the networks with larger average degree achieved lower accuracy when the network scale is comparable. The reason behind this is that, as the average degree of nodes increases, there will be more available propagation paths, which brings difficulties for source node identification.

Besides, the performance suggested the law of diminishing returns, that is, as the proportion of observation nodes increases, when it exceeds a certain value, there are progressively smaller rises in the accuracy. In fact, with the number of

TABLE 1. The basic topological features of real networks. | V | is the number of nodes, | E | is the number directed edges, $\langle k \rangle$ is the average degree, and $\langle d \rangle$ is the average shortest path length.

Network	$\|V\|$	$\|E\|$	$\|E\|/\|V\|$	$\langle k \rangle$	$\langle d \rangle$
Enron email network	151	266	1.76	3.5	3.91
Doctor friendship network	241	1098	4.56	9.13	2.37
USAir network	332	2126	16.6	12.81	2.74
Food web network	128	2196	17.16	30.71	1.69
Email network	1133	5451	4.81	4.81	3.61
Facebook-like social network	1899	20296	10.69	10.68	3.06



(a) Success Probability

(b) Top-10 Success Probability

FIGURE 8. The performance comparison of our methods with different transition rates and the existing method for traditional Susceptible-Infectious (SI) model in terms of (a) the Success Probability (SP) and (b) the TOP-10 Success Probability (Top-10 SP). The simulations are carried out with transition rates $\beta = 100$.

the observers increasing, there will be more redundant information provided by nearby observers. Thus, our algorithm can guarantee the accuracy of the results with only a part of nodes observed, given a limited budget for sensor placement.

E. PERFORMANCE COMPARISON

The key contribution of our paper is to consider the contagious incubation period in the Susceptible-Infectious infection processes with the asymptomatic spread for source identification. For these situations, directly adopting the existing method based on the traditional Susceptible-Infectious propagation model would give a bad performance as the infection times are not accurate. To demonstrate the difference, we set the same transition rates $\beta = 100$ for all nodes in the networks for the simulations of the infection process to study the duration of incubation periods as below. We evaluated the accuracy of our method given correct values of transition rates and compared it with two states of the art methods, Agaskar's method [23] and Pinto's method [38], and two baseline methods: 1) existing method for the traditional Susceptible-Infectious propagation model without consideration of contagious incubation period [44], and 2) our method given smaller values of transition rates ($\beta = 1, 0.01$). The first baseline can be regarded as a special case of our method when β approaches zero since the length of the incubation period approaches zero. Agaskar and Lu [23] proposed a fast Monte Carlo method for source identification. They assumed propagation follows the Susceptible-Infected propagation model based on geodesic distances on a randomly-weighted version of the graph and used the Monte Carlo method to approximate the gap between the observed infection time and the sampled infection time of sensor nodes. The node which can minimize the gap was considered as the propagation source node. When sampling the observed infection time, the random weight variable w_{ij} of the edge and the infection probability λ_{ij} are involved. We set the infection probability λ_{ij} of all edges to 0.4, which is the same as their work. Pinto's method [38] similarly calculated the gap between the observed delays and the deterministic delays of sensors. The node, which can minimize the distance of sensor nodes, was considered as the propagation origin. They assumed that all propagation delays in the network follow the same Gaussian distribution. In order to ensure that the propagation delay is non-negative, the mean must be greater than the standard deviation. We followed their experimental settings, set $\mu/\sigma = 4$, and took μ as the average of the delays of all edges.

Referring to Figure 8, our method given correct values of transition rates ($\beta = 100$) achieves best performance and the baseline method with smaller transition rates $\beta = 1$ gives lower accuracy. When the transition rate β equals 0.01, the performance is apparently worse

and very close to that of the method for the traditional Susceptible-Infectious infection process which gives the lowest accuracy. The results are consistent with the fact that the traditional Susceptible-Infectious propagation model is a special case of the SI model with a contagious incubation period when the duration of the incubation period approaches zero. The low performance of the state-of-the-art methods may be explained as follows. First of all, the propagation model and influence parameters considered in the methods are too simple. Second, in both cases, the network structure used to verify the algorithm is special and small in scale, such as square networks [23] and spatial (geographical) networks [38], which makes the source identification problem easier. Therefore, our method is more general than the existing method for the traditional SI model and can achieve much better performance in situations with longer incubation periods, such as COVID-19 and SARS with days of the incubation period.

VI. CONCLUSION

In this paper, we proposed to solve the single source identification problem based on the Susceptible-Infectious model with a contagious incubation period. To the best of our knowledge, we are the first group to investigate the source detection problem for epidemic dynamics with a contagious incubation period. In particular, we use the continuous-time Susceptible-Infectious model with a contagious incubation period to simulate the epidemic dynamics and assume only partial infection information can be observed, which often occurs in real situations. Accordingly, we have presented a source identification method that maximizes the likelihood of observed infection information under the propagation model. Then, we derive an efficient importance sampling approach to approximate the likelihood function and an effective optimal process. Simulations on several synthetic and real-world networks indicate that our method can identify sources of propagation with high accuracy in different situations, and outperform benchmark methods without considering contagious incubation periods for disease propagation with the asymptomatic spread.

Meanwhile, there are still some limitations and our work opens some interesting future work. For instance, it would be useful to extend our method to consider the presence of multiple propagation sources in the network. Also, the source identification problem based on other epidemic models with contagious incubation period such as the Susceptible-Infectious-Recovered model and the Susceptible-Infectious-Susceptible model can be investigated. Moreover, the network is static in this paper, our method can be extended to support temporal networks. Finally, other issues such as scalability may be further explored.

ACKNOWLEDGMENT

(Huan Liu and Qing Bao contributed equally to this work.)

REFERENCES

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol. 45, no. 2, pp. 167–256, 2003.
[2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Phys. Rep.*, vol. 424, nos. 4–5, pp. 175–308, 2006.
[3] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási, "Controllability of complex networks," *Nature*, vol. 473, pp. 167–173, May 2011.
[4] W. Wang, Q.-H. Liu, J. Liang, Y. Hu, and T. Zhou, "Coevolution spreading in complex networks," *Phys. Rep.*, vol. 820, pp. 1–51, Aug. 2019.
[5] Y. Wang, S. Wen, Y. Xiang, and W. Zhou, "Modeling the propagation of worms in networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 942–960, 2nd Quart., 2014.
[6] B. Doer, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, p. 70, 2012.
[7] M. Li, X. Wang, K. Gao, and S. Zhang, "A survey on information diffusion in online social networks: Models and methods," *Information*, vol. 8, no. 4, p. 118, Sep. 2017.
[8] L.-L. Xia, G.-P. Jiang, B. Song, and Y.-R. Song, "Rumor spreading model considering hesitating mechanism in complex social networks," *Phys. A, Stat. Mech. Appl.*, vol. 437, pp. 295–303, Nov. 2015.
[9] Q. Liu, T. Li, and M. Sun, "The analysis of an SEIR rumor propagation model on heterogeneous network," *Phys. A, Stat. Mech. Appl.*, vol. 469, pp. 372–380, Mar. 2017.
[10] Z. Jin, J. Zhang, L.-P. Song, G.-Q. Sun, J. Kan, and H. Zhu, "Modelling and analysis of influenza a (H1N1) on networks," *BMC Public Health*, vol. 11, no. S1, pp. 1–9, Dec. 2011.
[11] B. Shi, G. Liu, H. Qiu, Z. Wang, Y. Ren, and D. Chen, "Exploring voluntary vaccination with bounded rationality through reinforcement learning," *Phys. A, Stat. Mech. Appl.*, vol. 515, pp. 171–182, Feb. 2019.
[12] Y. Yu, L. Ding, L. Lin, P. Hu, and X. An, "Stability of the SNIS epidemic spreading model with contagious incubation period over heterogeneous networks," *Phys. A, Stat. Mech. Appl.*, vol. 526, Jul. 2019, Art. no. 120878.
[13] A. Koher, H. H. K. Lentz, J. P. Gleeson, and P. Hövel, "Contact-based model for epidemic spreading on temporal networks," *Phys. Rev. X*, vol. 9, no. 3, Aug. 2019, Art. no. 031017.
[14] L. Feng, Q. Zhao, and C. Zhou, "Epidemic spreading in heterogeneous networks with recurrent mobility patterns," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 102, no. 2, Aug. 2020, Art. no. 022306.
[15] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul. 2013, pp. 2671–2675.
[16] S. Shelke and V. Attar, "Origin identification of a rumor in social network," in *Cognition and Machine Learning Applications*. Singapore: Springer, 2020, pp. 89–96.
[17] M. Dong, B. Zheng, N. Quoc Viet Hung, H. Su, and G. Li, "Multiple rumor source detection with graph convolutional networks," in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 569–578.
[18] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 1, Jul. 2014, Art. no. 012801.
[19] S. J. Kazemitabar and A. A. Amini, "Approximate identification of the optimal epidemic source in complex networks," in *Proc. Int. Conf. Netw. Sci.*, 2020, pp. 107–125.
[20] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Modeling Comput. Syst. - SIGMETRICS*, 2010, pp. 203–214.
[21] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "Identifying propagation sources in networks: State-of-the-Art and comparative studies," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 465–481, 1st Quart., 2017.
[22] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?" *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5163–5181, Aug. 2011.
[23] A. Agaskar and Y. M. Lu, "A fast Monte Carlo algorithm for source localization on graphs," *Proc. SPIE*, vol. 8858, Sep. 2013, Art. no. 88581N.
[24] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "K-center: An approach on the multi-source identification of information diffusion," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 12, pp. 2616–2626, Dec. 2015.
[25] F. Yang, S. Yang, Y. Peng, Y. Yao, Z. Wang, H. Li, J. Liu, R. Zhang, and C. Li, "Locating the propagation source in complex networks with a direction-induced search based Gaussian estimator," *Knowl.-Based Syst.*, vol. 195, May 2020, Art. no. 105674.
[26] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2003, pp. 137–146.

[27] Y. Zhou, C. Wu, Q. Zhu, Y. Xiang, and S. W. Loke, "Rumor source detection in networks based on the SEIR model," *IEEE Access*, vol. 7, pp. 45240–45258, 2019.
[28] M. K. Slifka and L. Gao, "Is presymptomatic spread a major contributor to COVID-19 transmission?" *Nature Med.*, vol. 26, no. 10, pp. 1531–1533, Oct. 2020.
[29] Z. Wu and J. M. McGoogan, "Asymptomatic and pre-symptomatic COVID-19 in China," *Infectious Diseases Poverty*, vol. 9, no. 1, p. 72, Dec. 2020.
[30] B. Tang, X. Wang, Q. Li, N. L. Bragazzi, S. Tang, Y. Xiao, and J. Wu, "Estimation of the transmission risk of the 2019-nCoV and its implication for public health interventions," *J. Clin. Med.*, vol. 9, no. 2, p. 462, Feb. 2020.
[31] Y. Liu, Z. Gu, S. Xia, B. Shi, X.-N. Zhou, Y. Shi, and J. Liu, "What are the underlying transmission patterns of COVID-19 outbreak? An age-specific social contact characterization," *EClinicalMedicine*, vol. 22, May 2020, Art. no. 100354.
[32] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Trans. Signal Process.*, vol. 61, no. 11, pp. 2850–2865, Jun. 2013.
[33] W. Zang, P. Zhang, C. Zhou, and L. Guo, "Locating multiple sources in social networks under the SIR model: A divide-and-conquer approach," *J. Comput. Sci.*, vol. 10, pp. 278–287, Sep. 2015.
[34] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample-path-based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2016.
[35] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, "Rumor source identification in social networks with time-varying topology," *IEEE Trans. Depend. Sec. Comput.*, vol. 15, no. 1, pp. 166–179, Jan. 2018.
[36] Q. Huang, C. Zhao, X. Zhang, and D. Yi, "Locating the source of spreading in temporal networks," *Phys. A, Stat. Mech. Appl.*, vol. 468, pp. 434–444, Feb. 2017.
[37] Z.-L. Hu, Z. Shen, S. Cao, B. Podobnik, H. Yang, W.-X. Wang, and Y.-C. Lai, "Locating multiple diffusion sources in time varying networks from sparse observations," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, Dec. 2018.
[38] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.*, vol. 109, no. 6, Aug. 2012, Art. no. 068702.
[39] T. Zhao and A. Nehorai, "Distributed sequential Bayesian estimation of a diffusive source in wireless sensor networks," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1511–1524, Apr. 2007.
[40] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, "Bayesian inference of epidemics on networks via belief propagation," *Phys. Rev. Lett.*, vol. 112, no. 11, Mar. 2014, Art. no. 118701.
[41] M. G. Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. 28th Int. Conf. Mach. Learn.*, Bellevue, WA, USA, 2011, pp. 561–568.
[42] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 4, pp. 21:1–21:37, Feb. 2012.
[43] B. Abrahao, F. Chierichetti, R. Kleinberg, and A. Panconesi, "Trace complexity of network inference," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2013, pp. 491–499.
[44] M. Farajtabar, M. Gomez-Rodriguez, N. Du, M. Zamani, H. Zha, and L. Song, "Back to the past: Source identification in diffusion networks from partially observed cascades," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, San Diego, CA, USA, 2015, pp. 232–240.
[45] Z. Zhang, W. Xu, W. Wu, and D.-Z. Du, "A novel approach for detecting multiple rumor sources in networks with partial observations," *J. Combinat. Optim.*, vol. 33, no. 1, pp. 132–146, Jan. 2017.
[46] R. I. Alexandru and P. L. Dragotti, "Diffusion source detection in a network using partial observations," *Proc. SPIE*, vol. 11138, Sep. 2019, Art. no. 111380L.
[47] W. Luo and W. P. Tay, "Finding an infection source under the SIS model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 2930–2934.
[48] Y. Wang, J. Cao, A. Alsaedi, and B. Ahmad, "Edge-based SEIR dynamics with or without infectious force in latent period on random networks," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 45, pp. 35–54, Apr. 2017.
[49] G. Li and J. Zhen, "Global stability of an SEI epidemic model with general contact rate," *Chaos, Solitons Fractals*, vol. 23, no. 3, pp. 997–1004, Feb. 2005.
[50] J. Dhar and A. K. Sharma, "The role of the incubation period in a disease model," *Appl. Math. E Notes*, vol. 9, no. 9, pp. 146–153, 2009.
[51] G. Zhu, X. Fu, and G. Chen, "Spreading dynamics and global stability of a generalized epidemic model on complex heterogeneous networks," *Appl. Math. Model.*, vol. 36, no. 12, pp. 5808–5817, Dec. 2012.
[52] N. Du, L. Song, H. Woo, and H. Zha, "Uncover topic-sensitive information diffusion networks," in *Proc. 16th Int. Conf. Artif. Intell. Statist.*, Scottsdale, AZ, USA, Apr. 2013, pp. 229–237.
[53] N. Du, L. Song, M. G. Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," *Adv. Neural Inf. Process. Syst.*, vol. 26, no. 3, p. 3147, 2013.
[54] J. Leskovec. *SNAP: Stanford Network Analysis Platform*. Accessed: Feb. 24, 2021. [Online]. Available: http://snap.stanford.edu/snap/index.html
[55] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, M. Granovetter, Ed. Cambridge, U.K.: Cambridge Univ. Press, 1994.
[56] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the Web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, 1999.
[57] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, "Kronecker graphs: An approach to modeling networks," *J. Mach. Learn. Res.*, vol. 11, no. 3, pp. 985–1042, 2010.
[58] C. Arney, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
[59] A. Clauset, C. Moore, and M. E. J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, no. 7191, pp. 98–101, May 2008.
[60] B. Klimt and Y. Yang, "The Enron corpus: A new dataset for email classification research," in *Proc. Eur. Conf. Mach. Learn.*, Pisa, Italy, 2004, pp. 217–226.
[61] J. S. Coleman, E. Katz, and H. Menzel, "The diffusion of an innovation among physicians," *Sociometry*, vol. 20, no. 4, pp. 253–269, 1957.
[62] V. Batagelj and A. Mrvar, "Pajek-analysis and visualization of large networks," in *Graph Drawing Software*. Berlin, Germany: Springer, 2004, pp. 77–103.
[63] C. J. Melián and J. Bascompte, "FOOD Web COHESION," *Ecology*, vol. 85, no. 2, pp. 352–358, Feb. 2004.
[64] R. Guimera, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 68, no. 6, Dec. 2003, Art. no. 065103.
[65] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Netw.*, vol. 31, no. 2, pp. 155–163, May 2009.

HUAN LIU received the B.Eng. degree in network engineering from the Taiyuan University of Technology, Taiyuan, China, in 2018. He is currently pursuing the master's degree with the School of Cyberspace, Hangzhou Dianzi University, Hangzhou, China. His research interests include data-driven modeling and analytics, machine learning, and complex systems/networks.

QING BAO received the B.Sc. degree from the Department of Computer Science and Technology, East China Normal University, Shanghai, China, in 2011, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2016. She is currently an Associate Professor with the School of Cyberspace, Hangzhou Dianzi University, China. Before that, she was a Postdoctoral Research Fellow with Hong Kong Baptist University. Her research interests include graph data mining, and social networks analysis and health informatics. She was a recipient of the Best Student Paper Award at the 2013 IEEE/WIC/ACM International Conference on Web Intelligence. She is a reviewer for various journals and a program committee member of several conferences.

HONGJUN QIU received the B.Sc. degree in computer science and technology from Beijing Forestry University, Beijing, China, in 2003, and the Ph.D. degree in computer application from the Beijing University of Technology, Beijing, in 2010. She is currently a Lecturer with the School of Cyberspace, Hangzhou Dianzi University, China. Her research interests include complex systems/networks, autonomy-oriented computing, and health informatics.

MING XU received the M.S. and Ph.D. degrees from Zhejiang University, Hangzhou, China, in 2000 and 2004, respectively. He is currently a Full Professor with Hangzhou Dianzi University, China. His research interests include network security and digital forensics.

BENYUN SHI received the B.Sc. degree in mathematics from Hohai University, Nanjing, China, in 2003, and the M.Phil. and Ph.D. degrees in computer science from Hong Kong Baptist University, Hong Kong, in 2008 and 2012, respectively. He is currently a Professor with the School of Computer Science and Technology, Nanjing Tech University, China. Before that, he worked as a Full Professor with Hangzhou Dianzi University, China, and a Research Assistant Professor with Hong Kong Baptist University. His research interests include data-driven modeling and analytics, machine learning, complex systems/networks, multi-agent autonomy-oriented computing, computational epidemiology, and health informatics.