Identifying Propagation Sources in Networks: State-of-the-Art and Comparative Studies

Jiaojiao Jiang, Sheng Wen, Shui Yu, Senior Member, IEEE,
Yang Xiang, Senior Member, IEEE, and Wanlei Zhou, Senior Member, IEEE

Abstract—It has long been a significant but difficult problem to identify propagation sources based on limited knowledge of network structures and the varying states of network nodes. In practice, real cases can be locating the sources of rumors in online social networks and finding origins of a rolling blackout in smart grids. This article reviews the state-of-the-art in source identification techniques, and discusses the pros and cons of current methods in this field. Furthermore, in order to gain a quantitative understanding of current methods, we provide a series of experiments and comparisons based on various environment settings. Especially, our observation reveals considerable differences in performance by employing different network topologies, various propagation schemes and diverse propagation probabilities. We therefore reach the following points for future work. First, current methods remain far from practice as their accuracy in terms of error distance (δ) is normally larger than three in most scenarios. Second, the majority of current methods are too time-consuming to quickly locate the origins of propagation. In addition, we list five open issues of current methods exposed by the analysis, from the perspectives of topology, number of sources, number of networks, temporal dynamics, and complexity and scalability. Solutions to these open issues are of great academic and practical significance.

Index Terms—Complex network, propagation, source identification, centrality measures.

I. INTRODUCTION

In the modern world, the ubiquity of networks has made us vulnerable to various network risks. For instance, rumors spread incredibly fast in online social networks, such as Facebook and Twitter [1]. Computer viruses propagate throughout the Internet and infect millions of computers [2]. In smart grids, isolated failures could lead to rolling blackouts in the networks [3]. Every year, tremendous damages caused by those risks have incurred massive losses to society in finance and labor [4].

Risks, in terms of rumors, computer viruses or smart grid failures, propagate on various networks. From both practical and technical aspects, it is of great significance to identify propagation sources. Practically, it is important to accurately identify the ‘culprit’ of the propagation for forensic purposes. Moreover, seeking the propagation origins as quickly as possible can find the causation of risks, and therefore, diminish the damages. Technically, the work in this field is aimed at identifying the sources of propagations based on limited knowledge of network structures and the states of a portion of nodes. In academia, traditional identification techniques, such as IP traceback [5] and stepping-stone detection [6], are not sufficient to seek the propagation origins of risks, as they only determine the true source of packets received by a destination. In the propagation of risks, the source of packets is almost never the origin of the propagation but just one of the many propagation participants [7]. Methods are needed to find propagation sources higher up in the application level and logic structures of networks, rather than in the IP level and packets.

In the past few years, researchers have proposed a series of methods to identify propagation sources. The initial methods are designed to work on tree-like networks and with propagation following the traditional susceptible-infected (SI) model [8]–[17]. Further, some other work are proposed to deal with tree-like networks but with different epidemic models, such as the susceptible-infected-recovery (SIR) model and the susceptible-infected-susceptible (SIS) model [18]–[22]. The constraints on tree-like topologies were then relaxed to generic network topologies in source identification techniques [23], [24]. [24]–[38]. In addition, researchers proposed methods to identify propagation sources by first injecting sensors into networks [39]–[46]. In many ways, source identification requires either high computational complexity to find near-optimal solutions, or simplified heuristics to achieve suboptimal performance. In order to summarize the state-of-the-art and to benefit future research, we are motivated to provide a survey about current work in this field. To the best of our knowledge, this is the first comprehensive survey that focuses on the techniques of seeking propagation origins in various networks.

This survey consists of three main parts. We list the contribution and usage of each part as follows. First, we review existing source identification methods and analyze their pros and cons. This part sheds light on the basic ideas of current work to readers. Second, comparative studies are provided according to various experiment settings and scenarios. The results provide readers a numerical understanding of existing methods. Third, we summarize the analysis and comparative studies of source identification methods, and further list currently unsolved problems in this field. The significance of addressing these problems is analyzed in this part.

This survey is structured as follows. In Section II, we introduce some basic knowledge used in this article. The analysis of existing methods is presented in Section III. Section IV shows comparative studies followed by Section V which provides extensive discussion on critical problems in this field. We finally conclude this survey in Section VI.

J. Jiang, S. Wen, S. Yu, Y. Xiang and W. Zhou are with the School of Information Technology, Deakin University, 3125, Melbourne, VIC, Australia. Email: {jiaojiao, wsheng, syu, yang, wanlei}@deakin.edu.au.
II. PRELIMINARIES

We introduce preliminary knowledge of source identification in this section. It consists of observation categories, epidemic models and centrality measures. For convenience, we borrow notions from the area of epidemics to represent the states of nodes in networks. A node being infected stands for a user believing rumors, viruses having compromised a computer, or a power station being out of operation. Reader can derive analogous meanings for a node being susceptible or recovered.

A. Categories of Observations

One of the major premises in source identification problems is the observation of node states during the propagation process. Diverse observations lead to a great variety of methods in this field. According to the literature, there are three main categories of observations:

**Complete Observation**: Given a time $t$ during the propagation, this type of observation presents the exact state for each node in the network at this moment. The state of a node stands for the node having been infected or recovered, or remaining susceptible. This type of observation provides comprehensive knowledge of a transient status of the network. Through this type of observation, source identification techniques are advised with sufficient knowledge. An example of the complete observation is shown in Fig. 1(A).

**Snapshot**: Snapshot provides partial knowledge of network status at a given time $t$. Partial knowledge is presented in four forms. First, nodes reveal if they have been infected with probability $\mu$. Second, we recognize all infected nodes, but cannot distinguish susceptible or recovered nodes. Third, only a set of nodes was observed at time $t$ when the snapshot was taken. Fourth, only the nodes infected at time $t$ were observed. We show an example of the snapshot in Fig. 1(B).

**Sensor Observation**: Sensors are first injected into networks, and then the propagation dynamics over these sensor nodes are collected, including their states, state transition time and infection directions. In fact, sensors also stand for users or computers in networks. The difference between sensors and normal nodes in networks is that they are usually monitored by network administrators in practice. Therefore, the sensors can record all details of the rumor propagation over themselves, and their life can be theoretically assumed to be everlasting during the propagation dynamics. This is different from the mobile sensor devices which may be out of work when their batteries run out. As an example, we show the sensor observation in Fig. 1(C).

An illustration of these three categories of observations is shown in Fig. 1. It is clear that the snapshot and sensor observation provide much less information for identifying propagation sources compared with the complete observation.

B. Epidemic Models

Epidemic models are employed to describe the infection and recovery processes of nodes in networks. As another foundation for this field, different models refer to different scenarios in seeking propagation origins. So far, researchers mainly employ three epidemic models:

**SI model**: In this model, nodes are initially susceptible and can be infected along with the propagation of risks. Once a node is infected, it remains infected forever. This model focuses on the infection process $S \rightarrow I$, regardless of the recovery process.

**SIR model**: Recovery processes are considered in this model. Similarly, nodes are initially susceptible and can be infected along with the propagation. Infected nodes can then be recovered, and never become susceptible again. This model deals with the infection and curing process $S \rightarrow I \rightarrow R$.

**SIS model**: In this model, infected nodes can become susceptible again after they are cured. This model stands for the infection and recovery process $S \rightarrow I \rightarrow S$.

There are also other epidemic models, such as SIRS [47], SEIR [48], MSIR [49], SEIRS [50]. As far as we know, these
models have not been applied in source identification methods. Future work may take these models into consideration. Readers could refer to the work of [51] and [2] for other epidemic models.

C. Centrality Measures

Centrality measures are utilized to describe the influence of nodes on propagation. Therefore, researchers employ various centrality measures to identify potential propagation sources. We list five commonly used centrality measures as follows:

Degree: The degree of a node in a network is the number of edges incident to the node. In the real world, popular users correspond to high-degree nodes in networks [52]. The theoretical bases of this measure are the scale-free and power-law properties of the Internet with a few highly-connected nodes playing a vital role in maintaining the network’s connectivity [53], [54]. We illustrate this centrality in Fig. 2 (A).

Betweenness: The betweenness of a node stands for the number of shortest paths passing through the node [55]. Researchers have found the nodes which do not have large degrees in networks also play a vital role in the information propagation [56], [57]. As shown in Fig. 2 (B), the degree of node E is smaller than node A, B, C and D. However, node E is noticeably more important to the spread of rumors as it is the connector of two large groups of users. To locate this kind of nodes in networks, researchers introduced the measure of betweenness.

Closeness: The closeness of a node is defined as the mean geodesic (i.e. shortest path) distance from this node to other reachable nodes [54], [55]. As shown in Fig. 2(C), this measure discloses the nodes that can rapidly disseminate information to all the other nodes. This measure concentrates more on the information propagation speed rather than the connectivity of a network [54].

Jordan centrality: The Jordan centrality of a node is defined as the maximum geodesic distance from this node to any other infected node in the network [58], [59]. Jordan centers stand for the nodes that have minimum Jordan centrality. Suppose all the nodes are infected in the graph in Fig. 2 (D), then node A, B, C are the Jordan centers of the graph with Jordan centrality equals 3. Equivalently, the set of Jordan centers is equal to the radius of a network [60].

Eigenvector centrality: Eigenvector Centrality is defined as the eigenvector of the adjacency matrix associated to the largest eigenvalue [61], [62]. The eigenvector centrality of a node is proportional to the sum of the centrality values of all its neighboring nodes. In the real world, an important node is characterized by its connectivity to other important nodes. A node with a high eigenvector centrality value is a well-connected node and has a dominant influence on the surrounding network. As shown in Fig. 2 (E), node V_1 and V_3 have the highest eigenvector centrality in the graph. Readers could refer to [61] for further computation methods.

III. Source Identification Techniques

In this section, we analyze different techniques for source identification and discuss their pros and cons. We classify the source identification methods into three categories in accordance with the three different types of observation in Section II-A. The taxonomy of current methods is shown in Fig. 3. We analyze each category of methods in the following subsections, respectively.
A. Source Identification Methods with Complete Observations

In this subsection, we summarize the methods of source identification developed for complete observations. There are two techniques in this category: rumor center and eigenvector center based methods.

1) Single Rumor Center: Shah and Zaman [8], [10] introduced rumor centrality for source identification. They assume that information spreads in tree-like networks and the information propagation follows SI model. They also assume each node receives information from only one of its neighbors. Since we consider the complete observations of networks, the source node must be in the infected nodes. This method is proposed for the propagation of risks originating from a single source.

Method: Assuming an infected node as the source, its rumor centrality is defined as the number of distinct propagation paths originating from the source. The node with the maximum rumor centrality is called the rumor center. For regular trees, the rumor center is considered as the propagation origin. For generic networks, researchers employ BFS trees to represent the original networks. Each BFS tree corresponds to a probability $p$ of a rumor that chooses this tree as the propagation path. In this case, the source node is revised as the one that holds the maximum product of rumor centrality and $p$.

Analysis: In essence, the method is to seek a node from which the propagation matches the complete observation the best. As proven in [10] and [8], the rumor center is equivalent to the closeness center for a tree-like network. However, for a generic network, the closeness center may not equal the rumor center. The effectiveness of the method is further examined by the work of [12]. The authors proved the rumor center method can still be guaranteed. Z. Wang et al. [63] extend the discussion of the single rumor center into a more complex scenario with multiple snapshots. Although snapshot only provides partial knowledge of rumor spreading, the authors prove that multiple independent snapshots can dramatically improve temporally sequential snapshots. The analysis in [63] suggests that the complete observation for rumor source can be approximated by multiple independent snapshots.

Discussion: There are several strong assumptions far from reality. First, it is considered on a very special class of networks: infinite trees. Generic networks will be reconstructed into BFS trees before seeking propagation origins. Second, risks are implicitly assumed to spread in a unicast way (an infectious node can only infect one of its neighbors at one time step). Third, the infection probability between neighboring nodes is equal to 1. In the real world, however, networks are far more complex than trees, with information often spreading in multicast or broadcast ways, and the infection probability between neighboring nodes differing from each other.

2) Local Rumor Center: Following the assumptions of the single rumor center method, Dong et al. [9] proposed a local rumor center method to identify propagation sources. This method designates a set of nodes as suspicious sources. Therefore, it reduces the scale of seeking origins.

Method: Dong et al. [9] utilized the approaches and results in [10] and [8] to identify the source of propagation in networks. Following the definition of the rumor center, they defined the local rumor center as the node with the highest rumor centrality compared to other suspicious infected nodes. The local rumor center is considered as the source node.

Analysis: For regular trees with node degree $d$, the authors analyze the accuracy $\gamma$ of the local rumor center method. To construct a regular tree, the degree $d$ of each node should be at least 2. However, W. Dong et al. derived that the accuracy of the local rumor center method follows $O(1/\sqrt{n})$. Therefore, when $n$ is sufficiently large, the accuracy is close to 0 when $d = 2$. As a result, $d$ starts from 3 to infinity in the analysis. First, when the suspicious set degenerates into
the entire network, $\gamma$ grows from 0.25 to 0.307 as $d$ increases from three to infinity. This means that the minimum accuracy $\gamma$ is 25\% and the maximum accuracy is 30.7\%. Second, when suspicious nodes form a connected subgraph of the network, $\gamma$ significantly exceeds $1/k$ when $d = 3$, where $k$ is the the number of suspicious nodes. Third, when there are only two suspect nodes, $\gamma$ is at least 0.75 if $d = 3$, and $\gamma$ increases with the distance between the two suspects. Fourth, when multiple suspicious nodes form a connected subgraph, the accuracy $\gamma$ is lower than when these nodes form several disconnected subgraphs.

Discussion: The local rumor center is actually the node with the highest rumor centrality in the priori set of suspects. The advantage of the local rumor center method is that it dramatically reduces the source-searching scale. However, it has the same drawbacks as the single rumor center method.

3) Multiple Rumor Centers: Luo et al. [13] extended the single rumor center method to identify multiple sources. In addition to the basic assumptions, researchers further assume the maximum number of sources is known for the method of identifying multiple rumor centers.

Method: Based on the definition of rumor centrality for a single node, Luo et al. [13] extended rumor centrality for a set of nodes, which is defined as the number of distinct propagation paths originating from the set. They propose a two-source estimator to compute the rumor centrality when there are only two sources. For multiple sources, they propose a two-step method. In the first step, they assume a set of infected nodes as sources. All infected nodes are divided into different partitions by using the Voronoi partition algorithm [64] on these sources. The single rumor center method is then employed to identify the source in each partition. In the second step, estimated sources are calibrated by the two-source estimator between any two neighboring partitions. These two steps are iterated until the estimated sources become steady.

Analysis: Luo et al. [13] are the first to employ the rumor center method to identify multiple sources. They further investigate the performance of the two-source estimator on geometric trees [10]. The accuracy approximates to one when these nodes form several disconnected subgraphs.

Discussion: According to the definition of rumor centrality for a set of nodes, we need to calculate the number of distinct propagation paths originating from the set. It is too computationally complex to obtain the result. Even though Luo et al. have proposed a two-step method to reduce the complexity, the two-step method still needs $O(k^k)$ computations, where $k$ is the number of source nodes. This method can hardly be used in the real world, especially on large-scale networks.

4) MDL: Prakash et al. [24], [66] proposed a minimum description length (MDL) method for source identification. This method is considered on generic networks. They assume propagation follows SI model.

Method: Given an arbitrary infected node as the source node, this corresponds to the probability of obtaining the infection graph. For generic networks, it is too computationally expensive to obtain the probability. Therefore, Prakash et al. [24] introduced an upper bound of the probability and sought the origin by maximizing the upper bound instead. They claimed that to maximize the upper bound is to find the smallest eigenvalue $\lambda_{\text{min}}$ and the corresponding eigenvector $u_{\text{min}}$ of the Laplacian matrix of the infection graph. The Laplacian matrix is widely used in the spectral graph theory and has many applications in various fields. This matrix is mathematically defined as $L = D - A$, where $D$ is the diagonal degree matrix and $A$ is the adjacency matrix. In Prakash et al.'s work [24], [66], the node with the largest score in the eigenvector $u_{\text{min}}$ of the Laplacian matrix refers to the propagation source.

Analysis: This method can also be used to seek multiple sources. The authors adopt the minimum description length (MDL) cost function [67]. This is used to evaluate the ‘goodness’ of a node being in the source set. To search the next source node, they first remove the previous source nodes from the infected set. Then, they replay the process of searching the single source in the remaining infection graph. These two steps are iterated until the MDL cost function stops decreasing.

Discussion: Due to the high complexity in computing matrix eigenvalues, generally $O(N^3)$, the DML method is not suitable for identifying sources in large-scale networks. Moreover, the number of true sources is unknown. Further to this, the gap between the upper bound and the real value of the probability has not been analyzed, and therefore, the accuracy of this method is not guaranteed.

5) Dynamic age: Fioriti et al. [23] introduced the dynamic age method for source identification in generic networks. The assumption for this method is the same as the DML method.

Method: Fioriti et al. took advantage of the correlation between the eigenvalue and the ‘age’ of a node. The ‘oldest’ nodes which are associated to those with largest eigenvalues will be considered as the sources of a propagation [68]. Meanwhile, they utilized the dynamical importance of node in [69]. It essentially calculates the reduction of the largest eigenvalue of the adjacency matrix after a node has been removed. A large reduction after removal of a node implies the node is relevant to the ‘aging’ of a propagation. By combing these two techniques, Fioriti et al. proposed the concept of dynamical age for an arbitrary node $i$ as follows,

$$DA_i = |\lambda_m - \lambda_i^m|/\lambda_m,$$  

where $\lambda_m$ is the maximum eigenvalue of the adjacency matrix, and $\lambda_i^m$ is the maximum eigenvalue of the adjacency matrix after node $i$ is removed. The nodes with the highest dynamical age are considered as the sources.

Analysis: This method is essentially different from the previous MDL method. The DML method is to find the smallest eigenvalues and the corresponding eigenvectors of Laplacian matrices, while the dynamic age method is to find the largest eigenvalues of the adjacency matrix.

Discussion: Similar to the MDL method, the dynamic age method is not suitable for identifying sources in large-scale networks.
networks. Moreover, since there is no threshold to determine the oldest nodes, the number of source nodes is uncertain.

B. Source Identification Methods with Snapshot

In the real world, a complete observation of an entire network is hardly possible, especially for large-scale networks. Snapshot is an observation close to reality. It only provides partial knowledge of propagation in networks. There are three techniques of source identification developed on snapshot: Jordan center, message passing and concentricity based methods.

1) Jordan Center: Zhu and Ying proposed a novel Jordan center method for source identification [18]. They assume information propagates in tree-like networks and the propagation follows SIR model. All infected nodes are known, but we cannot distinguish between susceptible nodes and recovered nodes. This method is proposed for single source propagation.

Method: Zhu and Ying [18] proposed a sample path based approach to identify the propagation source. An optimal sample path is the one which most likely leads to the observed snapshot of a network. The source associated with the optimal sample path is proven to be the Jordan center of the infection graph. Jordan center is considered as a propagation origin.

Analysis: Zhu and Ying further extended the sample path based approach to the heterogeneous SIR model [21]. Heterogeneous SIR model means the infection probabilities between any two neighboring nodes are different, and the recovery probabilities of infected nodes differ from each other. They prove that on infinite trees, the source node associated with the optimal sample path is also the Jordan center. Moreover, Luo et al. [19], [22] investigated the sample path based approach in SI and SIS models. They obtain the same conclusion as in the SIR model.

Discussion: Similar to rumor center based methods, the Jordan center method is considered on infinite tree-like networks, which are far different from real-world networks.

2) Dynamic Message Passing: In the dynamic message-passing (DMP) method [25], researchers suppose that propagation follows SIR model in generic networks. Only propagation time \( t \) and the states of a set of nodes at time \( t \) are known.

Method: The DMP method is based on the dynamic equations in [33]. Assuming an arbitrary node as the source node, it first estimates the probabilities of other nodes to be in different states at time \( t \). Then, it multiplies the probabilities of the observed set of nodes being in the observed states. The source node which can obtain the maximum product is considered the propagation origin.

Analysis: The DMP method takes into account the spreading dynamics of the propagation process. This is very different from the previous centrality based methods. Lokhov et al. [25] claim the DMP source identification method dramatically outperforms the previous centrality based methods.

Discussion: An important prerequisite of the DMP method is that we must know the propagation time \( t \). However, the propagation time \( t \) is generally unknown. Besides, the computational complexity of this method is \( O(tN^2d) \), where \( N \) is the number of nodes in a network and \( d \) is the average degree of the network. If the underlying network is strongly connected, it will be computationally expensive to use the DMP method to identify the propagation source.

3) Effective Distance Based Method: Assuming propagation follows SI model in weighted networks, Brockmann and Helbing proposed an effective distance based method for source identification [26]. This method is considered in another case of snapshot where we only know a spreading wavefront.

Method: Brockmann and Helbing [26] first proposed a new concept, the effective distance, to represent the propagation process. The effective distance from node \( n \) to neighboring node \( m \), \( d_{mn} \), is defined as

\[
d_{mn} = (1 - \log P_{mn}),
\]

where \( P_{mn} \) is the fraction of a propagation with destination \( m \) emanating from \( n \). From the perspective of a chosen source node \( v \), the set of shortest paths in terms of effective distance to all other nodes constitutes a shortest path tree \( \Psi_v \). They empirically obtain that the propagation process initiated from node \( v \) on the original network can be represented as wavefronts on the shortest path tree \( \Psi_v \). To illustrate this process, a simple example is shown in Fig. 4 (refers to [26]). According to the propagation process of wavefronts, the spreading concentricity can only be observed from the perspective of the true source. Then, the node, which has the minimum standard deviation and mean of effective distances to the nodes in the observed wavefront, is considered as the source node.

Analysis: The information propagation process in networks is complex and network-driven. The combined multiscale nature and intrinsic heterogeneity of real-world networks make...
it difficult to develop an intuitive understanding of these processes. Brockmann and Helbing [26] reduce the complex spatiotemporal patterns to a simple wavefront propagation process by using effective distance.

Discussion: To use the effective distance based method for source identification, we need to compute the shortest distances from any suspicious source to the observed infected nodes. This leads to high computational complexity, especially for large-scale networks.

C. Source Identification Methods with Sensor Observations

In the real world, a further strategy is used to identify propagation sources by injecting sensors into networks. The sensors report the direction in which information arrives to them, and the time at which the information arrives at the sensor. According to Fig. 3, there are two techniques developed in this category: statistics and greedy rules.

1) Gaussian Source Estimator: Assuming propagation follows SI model in tree-like networks, Pinto et al. proposed a Gaussian method for single source identification [40]. They also assume there is a deterministic propagation time for each edge, which are independent and identically distributed with Gaussian distribution.

Method: This method is divided into two steps. In the first step, they reduce the scale of seeking origins. According to the direction in which information arrived at the sensors, it uniquely determines a subtree $T_o$. The subtree $T_o$ is guaranteed to contain the propagation origin [40]. In the second step, they use the following Gaussian technique to seek the source in $T_o$. On the one hand, given a sensor node $o_1$, they calculate the ‘observed delay’ between $o_1$ and the other sensors. On the other hand, assuming an arbitrary node $s \in T_o$ as the source, they calculate the ‘deterministic delay’ for every sensor node relative to $o_1$ by using the deterministic propagation time of the edges. The node, which can minimize the distance between the ‘observed delays’ and the ‘deterministic delays’ of sensor nodes, is considered as the propagation origin.

Analysis: This method is considered on tree-like networks. For generic networks, Pinto et al. [40] assume that information spreads along a BFS tree, and then the origin is sought in the BFS trees. This method is improved by combining community recognition techniques in order to reduce the number of deployed sensors in networks. By choosing the nodes between communities and with high betweenness values for sensors, A. Louni et al. [70] reduce 3% fewer sensors than the original method [40].

Discussion: For generic networks, the Gaussian source estimator is of complexity $O(N^3)$. It is too computationally expensive to use this method for large-scale networks.

2) Monte Carlo Source Estimator: Agaskar and Lu [39] proposed a fast Monte Carlo method for source identification in generic networks. They assume propagation follows the heterogeneous SI model in which the infection probabilities between any two neighboring nodes are different. In addition, the observation of sensors is obtained in a fixed time window.

Method: This method consists of two steps. In the first step, assuming an arbitrary node as the source, they introduce an alternate representation for the infection process initiated from the source. The alternate representation is derived in terms of the infection time of each edge. Based on the alternate representation, they sample the infection time for each sensor. In the second step, they compute the gap between the observed infection time and the sampled infection time of sensors. They further use the Monte Carlo approach to approximate the gap. The node which can minimize the gap is considered as the propagation origin.

Analysis: The computational complexity of this method is $O(LN\log(N)/\varepsilon)$, where $L$ is the number of sensor nodes, and $\varepsilon$ is the assumed error. The complexity is less than other source identification methods, which are normally $O(N^2)$, or even $O(N^3)$.

Discussion: When sampling infection time for each edge, Agaskar and Lu [39] assume that information always spreads along the shortest paths to other nodes. However, in the real world, information generally reaches other nodes by random walk. Therefore, this method may not be suitable for other propagation schemes, such as random spreading or multicast spreading.

3) Bayesian Source Estimator: Distinguished from the DMP method which adopts the message-passing propagation model (see Section III-B2), F. Altarelli et al. proposed using the Bayesian belief propagation model to compute the probabilities of each node being in any state [27]. This method can work with different observations and in different propagation scenarios, however guaranteed accuracy is only obtained in tree-like networks.

Method: The propagation of risks are first presented by SI, SIR or other isomorphic models [2]. Second, given an observation on the infection of a network, either through a group of sensors or a snapshot at an unknown time, the belief propagation equations are derived for the posterior distribution of past states on all network nodes. By constructing a factor graph based on the original network, these equations provide the exact computation of posterior marginals in the models. Third, belief propagation equations are iterated with time until they converge. Nodes are then ranked according to the posterior probability of being the source.

Analysis: This method provides the exact identification of source in tree-like networks. This method is also effective for synthetic and real networks with cycles, both in a static and a dynamic context, and for more general networks, such as DTN [71]. This method relies on belief propagation model in order to be used with different observations and in various scenarios.

Discussion: The accuracy of this method can not be guaranteed other than in tree-like networks. Particularly for dynamically evolving networks [72], the average success rate is only $0.53 \pm 0.06$ and the average error reaches $0.76 \pm 0.23$.

4) Moon-walk Source Estimator: Xie et al. proposed a post-mortem technique on traffic logs to seek the origin of a worm (a kind of computer virus) [7]. There are four assumptions for this technique. First, it focuses on the scanning worm [73]. This kind of worm spreads on the Internet by making use of OS vulnerabilities. Victims will proceed to scan the whole IP space for vulnerable hosts. Famous examples of
this kind of worm includes Code Red [74] and Slammer [75].
Second, logs of infection from sensors cover the majority of
the propagation processes. Third, the worm propagation forms
a tree-like structure from its origin. Last, the attack flows of
a worm do not use sproofed source IP addresses.

**Methods:** Based on traffic logs, the network communication
between end-hosts are modelled by a directed host contact
graph. Propagation paths are then created by sampling edges
from the graph according to the time of corresponding logs.
The creation of each path stops when there is no contiguous
edge within $\Delta t$ seconds to continue the path. As the sampling
is performed, a count is kept of how many times each edge
from the contact graph is traversed. If the worm propagation
follows a tree-like structure, the edge with maximum count
will most likely be the top of the tree. The start of this directed
edge will be considered as the propagation source.

**Analysis:** There are several issues on this technique that
need to be further analyzed. First, it is reasonable to assume
worm do not use the IP sproof technique. In the real world,
the overwhelming majority of worm traffic involved in prop-
agation is initiated by victims instead of the original attacker.
Sproofed IP addresses would only decrease the number of
successful attacks without providing further anonymity to
the attacker. Second, IP traceback techniques [6] are related
to Moonwalk and other methods discussed in this article.
However, traceback on its own is not sufficient to track worms
to their origin, as traceback only determines the true source
of the IP packets received by a destination. In an epidemic
attack, the source of these packets is almost never the origin
of the attack, but just one of the many infected victims. The
methods introduced in this article are still needed to find the
hosts higher up in the propagation casual trees. Third, this
method relies only on traffic logs. This feature benefits itself
on its ability to work without any a priori knowledge about
the worm attack.

**Discussion:** Nowadays, the number of scanning worms has
largely decreased due to advances in OS development and
security techniques [76]. Therefore, the usage of Moonwalk,
which can only seek the propagation origin of the scanning
worm, is largely limited. Moreover, a full collection of in-
festation logs is hardly achieved in the real world. Finally,
current computer viruses are normally distributed by Botnet
[77]. Moonwalk, which can only seek single origin, may not
be helpful in this scenario.

5) **Four-metric Source Estimator:** Seo et al. [41] proposed
a four-metric source estimator to identify single source node
in directed networks. They assume propagation follows SI
model. The sensor nodes which transited from susceptible states
to infected states are regarded as positive sensors. Otherwise,
they are considered as negative sensors.

**Method:** Seo et al. use the intuition that the source node
must be close to the positive sensor nodes, but far away from
the negative sensor nodes. They proposed four metrics to
locate the source. First, they find out a set of nodes which
are reachable to all positive sensors. Second, they filter the
set of nodes by choosing the ones with the minimum sum
of distances to all positive sensor nodes. Third, they further
choose the nodes that are reachable to the minimum number of
negative sensor nodes. Finally, the node which satisfies all of
the above three metrics and has the maximum sum of distances
to all negative sensor nodes is considered as the source node.

**Analysis:** Seo et al. [41] studied and compared different
methods of choosing sensors, such as randomly choosing
(Random), choosing the nodes with high betweenness cen-
trality values (BC), choosing the nodes with a large number
of incoming edges (NI), and choosing the nodes which are
at least $d$ hops away from each other (Dist). Different sensor
selection methods produce different sets of sensor nodes, and
have different accuracies in source identification. They show
that the NI and BC sensor selection methods outperform the
others.

**Discussion:** For the four-metric source estimator, it needs to
compute the shortest paths from the sensors to any potential
source. Generally, the computational complexity is $O(N^3)$. It
is too computationally expensive to use this method.

**IV. COMPARATIVE STUDY**

In order to have a numerical understanding of the meth-
ods of source identification, we examine the methods under
different experiment environments. Furthermore, we analyze
potential impact factors on the accuracy of source identifica-
tion. We test the methods on both synthetic and real-world
networks. All the experiments were conducted on a desktop
computer running Microsoft Windows7 with 2 CPUs and 4G
memory. The implementation was done in Matlab2012.

For each category of observation, we examined one or two
typical source identification methods. In total, five methods
were examined. For complete observation, we tested the rumor
center method and the dynamic-age method. We also tested the
Jordan center method and the DMP method for snapshots of
networks. The Gaussian source estimator was examined for
sensor observation. In the experiments, we typically choose
infection probability ($q$) to be 0.75 and recovery probability
($p$) to be 0.5. We randomly choose a node as a source to initiate
a propagation, and then average the error distance $\delta$ between
the estimated sources and the true sources by 100 runs.

**A. Tests on Synthetic Networks**

In this subsection, we first compare the performance of
different source identification methods on synthetic networks.
Then, we study three potential impact factors on the accuracies of the methods.

1) Crosswise Comparison: We conducted experiments on two synthetic networks: a regular tree [8] and a small-world network [78]. Fig. 5 (A) and (B) show example topologies of a regular tree and a small-world network.

Fig. 6 shows the frequency of error distances $\delta$ of different methods on a 4-regular tree. We can see that, the sources estimated by the DMP method and the Jordan center method are the closest to the true sources, with an average of 1-2 hops away. The rumor center method and the dynamic age method estimate the sources with an average of 2-3 hops away from the true sources. The sources estimated using the dynamic age method were the farthest from the true sources. The dynamic age and Gaussian method have the worst performance.

2) The Impact of Network topologies: In Section III, we know that some existing methods of source identification are considered on tree-like networks. In the previous subsection, we have shown the results of methods implemented on regular trees and small-world networks. In order to analyze the impact of network topology on the methods, we introduce another two different network topologies: random trees and regular graphs. We further conduct performance evaluation on these two topologies.

Fig. 8 shows the experiment results of methods on a random tree. It is clear the Jordan center method has the best performance, with estimated sources around 1 hop away from the true sources. The DMP method also exposes good performances by showing estimated sources are an average of 1-2 hops away from the true sources. The dynamic age method and Gaussian have the worst performance.

Numerical Results: From the experiment results on the regular tree and small-world network, we can see that the DMP method and the Jordan center method have better performance than the other methods.

3) The Impact of Propagation Schemes: From Section III, we know that some existing methods of source identification are sensitive to network topology.

Fig. 9 shows the experiment results of methods on a regular graph. It shows that sources estimated by using the Jordan center method and the DMP method were the closest to the true sources. The sources estimated by the rumor center method were the farthest from true sources. The dynamic age method and the Gaussian method also show poor performance in this scenario.

Numerical Results: From the experiment results on the four different network topologies, we can see the source identification methods are sensitive to network topology.
are based on the assumption that information propagates along the BFS trees in networks. This means propagation follows the broadcast scheme. However, in the real world, propagation may follow various propagation schemes. We focus on three most common propagation schemes: snowball, random walk and contact process [28]. Their definitions are given below.

- **Random Walk**: A node can deliver a message randomly to one of its neighbors.
- **Contact Process**: A node can deliver a message to a group of its neighbors that have expressed interest in receiving the message.
- **Snowball Spreading**: A node can deliver a message to all of its neighbors.

An illustration of these three propagation schemes is shown in Fig. 10. We examine different propagation schemes on both regular trees and small-world networks.

Fig. 11 shows the experiment results of the methods with propagation following the random-walk scheme.

hops away from the true sources. The performances of the rumor center method, the dynamic age method and the Jordan center method are similar to each other, with estimated sources around 5 hops away from the true sources. The DMP method has the worst performance. Fig. 12 shows experiment results of the methods with propagation following the contact-process propagation scheme on a 4-regular tree. It is clear the results in Fig. 11 and Fig. 12 are similar to each other. This means the methods have similar performances on both the random-walk and contact-process propagation schemes. Fig. 13 shows the experiment results of the methods with propagation following the snowball propagation scheme on a 4-regular tree. The results show a big difference from the results of the previous two propagation schemes. The DMP method and the Jordan center method outperformed the others, with estimated sources around 1-2 hops away from the true sources. The rumor center method and the Gaussian method also showed good performances, with estimated sources around 2-3 hops away from the true sources. The dynamic age method had the worst
The experiment results of the methods with propagation following different propagation schemes on a small-world network are shown in Fig. 14, 15 and 16. The results are dramatically different from the results on regular trees. From Fig. 14 we can see the Gaussian source estimator obtains the best performance, followed by the DMP method. The rumor center method, the dynamic age method and the Jordan center method show identifying sources by randomly choosing. From Fig. 15, it is clear the Jordan center method, the DMP method and the Gaussian method show similar performances. These three methods outperform the others. From Fig. 16 we can see the Jordan center method outperforms the others, with estimated sources around 1 hop away from the true sources. The sources estimated using the DMP method are around 1-2 hops away from the true sources. The Gaussian source estimator has the worst performance.

**Numerical Results:** From the experiment results, we see the source identification methods are also sensitive to propagation schemes. The methods of source identification show better performance when propagation follows the snowball propagation scheme rather than the random-walk or contact-process propagation schemes.

4) The Impact of Infection Probabilities: In this subsection, we will analyze the impact of infection probability on the accuracy of source identification. We set the infection probability from 0.5 to 0.95.

The experiment results are shown in Fig. 17 and 18. From these figures, we can see that the rumor center method have similar performances when we change the infection proba-
In this subsection, we examine the methods of source identification on two real-world networks. The first one is an Enron email network [79]. This network has 143 nodes and 1,246 edges. On average, each node has 8.71 edges. Therefore, the Enron email network is a dense network. The second is a power grid network [80]. This network has 4,941 nodes and 6,594 edges. On average, each node has 1.33 edges. Therefore, the power grid network is a sparse network. Sample topologies of these two real-world networks are shown in Fig. 19.

Fig. 20 shows the frequency of error distance $\delta$ of different methods on the Enron email network. We can see the rumor center method, the Jordan center method and the dynamic age method outperform the others. The DMP method has the worst performance. The Enron email network is a small and dense network, complete observation of this network is reasonable and executable, and the identification accuracy is also acceptable. Fig. 21 shows the experiment results on the power grid network. It is clear the Jordan center method and the DMP method outperform the others, with estimated sources around 1-2 hops away from the true sources. The rumor center method and the Gaussian method show similar performance, with estimated sources around 2-4 hops away.
from the true sources. The dynamic age method has the worst performance.

**Numerical Results:** From the experiment results, we can see the accuracies of the methods are greatly different between these two real-world networks. For the Enron email network, the rumor center method and the dynamic age method outperform the other methods, while the DMP method has the worst performance. However, for the power grid network, the DMP method and the Jordan center have the best performance.

**V. SUMMARY AND OPEN ISSUES**

A. What We Learn from the-State-of-the-Art?

We summarize the source identification methods in this subsection. Based on the content in Section III, it is clear that current methods rely on either the topological centrality measures or the measures of the distance between the observations and mathematical estimations of the propagation.

In Table I, we collect seven features from the methods discussed in this article. A detailed summary on each feature is elaborated as follows:

1. **Topology:** As shown in Table I, a significant part the focus for current methods is tree-like topology. These methods can deal with generic network topologies by using the BFS technique to reconstruct generic networks into trees. According to comparative studies in Section IV, methods on different topologies show a great variety of accuracy in seeking origins.

2. **Observation:** Based on the analysis in Section III and IV, the category of observation is not a deterministic factor on the accuracy of source identification. The accuracy of each method varies according to the different conditions and scenarios. In the real world, complete observation is generally difficult to achieve. Snapshot and sensor observation are normally more realistic.

3. **Model:** The majority of methods employ SI model to present the propagation dynamics of risks. The SI model only considers the susceptible and infected states of nodes regardless of the recovery process. The extension to SIR/SIS will increase the complexity of source identification methods. Jordan center and Monte Carlo method is based on SIR/SIS model. In particular, the Bayesian source estimator can be used in scenarios with various propagation models as the belief propagation approach can estimate the probabilities of node states under various conditions.

4. **Source:** Most methods focus on single source identification. The multi-rumor center method and eigenvector center method can be used to identify multiple sources. However, these two methods are too computationally expensive to be implemented. In the real world, risks are normally distributed from multiple sources. For example, attackers generally employ a botnet which contains thousands of victims to help spread the computer virus [81], [82]. For source identification, these victims are the propagation origins.

5. **Probability:** For simplicity, earlier methods consider the infection probabilities to be identical among the edges in networks. Later, most methods are extended to varied infection probabilities among different edges. Noticably, this extension makes source identification methods more realistic.

6. **Time Delay:** Only the methods under sensor observations consider time delay for edges. Accurate time delay of risks is an important factor in the propagation [83]. It is important to consider the time delay in source identification techniques.

7. **Complexity:** Most current methods are too computationally expensive to quickly capture the sources of propagation. The complexity ranges from $O(N \log N / \varepsilon)$ to $O(N^k)$. In fact, the complexity of methods dominates the speed of seeking origins. Quickly identifying propagation sources in most cases is of great significance in the real world, such as capturing the culprits of rumors. Future work is needed to improve the identification speed.
C. Open Issues

Based on the summary of the-state-of-the-art and comparative studies in source identification, we extract five open issues. The solutions to these open issues will help provide more realistic results.

B. What We Learn from Comparative Studies?

A summary of the comparative studies in Section IV is shown in Table II. For the rumor center method, it is clear that the error distance \( \delta \) is normally from 3 to 4. The performance worsens when this method runs with the settings: regular graph, random-walk propagation scheme, contact-process propagation scheme or infection probability \( q = 0.5 \). Specifically, the performance of the dynamic age method is much worse than that of the rumor center method, as the error distance \( \delta \) is normally larger than 4. The Jordan center method and the DMP method normally outperform other methods in many settings, with error distance \( \delta \) between 1 and 3. The Gaussian method only runs well when propagation follows random-walk scheme or contact-process scheme on regular trees.

From the comparative studies, we can see that current methods are far from practice as their accuracy in terms of error distance \( \delta \) is normally larger than three in most scenarios. Although the sources estimated by the Jordan center method and the DMP method are close to the true sources under some settings, their performances are unstable and cannot meet our expectation with \( \delta > 4 \) under other settings in Table II.

C. Open Issues

Based on the summary of the-state-of-the-art and comparative studies in source identification, we extract five open issues. The solutions to these open issues will help provide more realistic results.

1) Tree-like Topology or Generic Topology: It is normal to have cycles in real-world networks [84]. It is essential to consider the propagation impact of topological cycles on source identification. Although current methods based on trees can identify sources on generic networks by using the BFS technique, its accuracy cannot be guaranteed as the impact of cycles is neglected in BFS trees. This is an inevitable drawback for tree-based methods working on generic networks. Therefore, we cannot directly use or extend tree-based methods for source identification on generic networks. On another hand, current methods which are considered on generic networks are quite sensitive to the topologies of networks (see details in Section IV-A2). We cannot obtain a guaranteed accuracy when the topology changes. We therefore propose an open issue of an accurate, steady and practical source identification method in generic networks.

2) Single Source or Multiple Sources: In the real world, the propagation of risks are often initiated from multiple sources. For example, culprits employ a botnet to spread rumors and computer viruses [81], [82]. However, few current methods are designed for multi-source identification. Technically, the methods of single source identification cannot be directly used for multiple source identification. This is because the spread initiated from multiple sources cannot be thought of as the superposition of multiple single-type propagation processes. Moreover, current multi-source identification methods are too computationally expensive to obtain results. The complexity is normally \( O(N^3) \) [23], [24]. Especially for the work [13], when the number of sources (\( k \)) increases, the complexity becomes
Our future work contains two parts. First, we will focus on novel methods that can identify multiple sources of propagation. The potential solution could be based on inverse techniques in mathematics. Second, the methods of source identification on interconnected networks are in development. These methods are more practical than current methods.

VI. CONCLUSION

In this article, we review state-of-the-art in source identification techniques. We first categorized current source identification techniques into three classes and analyze the pros and cons of each method. We further explored comparative studies on typical methods in order to provide a numerical understanding of current methods. We find current methods have a great variety of accuracy when the experiment environment changes. Open issues are finally proposed based on the analysis and comparison of the previous two parts. We believe this survey is timely and worthwhile.

Our future work contains two parts. First, we will focus on novel methods that can identify multiple sources of propagation. The potential solution could be based on inverse techniques in mathematics. Second, the methods of source identification on interconnected networks are in development. These methods are more practical than current methods.

REFERENCES


