Finding the Source in Networks: An Approach Based on Structural Entropy

CHONG ZHANG, QIANG GUO, LUOYI FU, JIAXIN DING, and XINDE CAO, Shanghai Jiao Tong University, China
FEI LONG, Xinhua News Agency, China
XINBING WANG, Shanghai Jiao Tong University, China
CHENGHU ZHOU, Chinese Academy of Sciences, China

The popularity of intelligent devices provides straightforward access to the Internet and online social networks. However, the quick and easy data updates from networks also benefit the risk spreading, such as rumor, malware, or computer viruses. To this end, this article studies the problem of source detection, which is to infer the source node out of an aftermath of a cascade, that is, the observed infected graph G_N of the network at some time. Prior arts have adopted various statistical quantities such as degree, distance, or infection size to reflect the structural centrality of the source. In this article, we propose a new metric that we call the infected tree entropy (ITE), to utilize richer underlying structural features for source detection. Our idea of ITE is inspired by the conception of structural entropy [21], which demonstrated that the minimization of average bits to encode the network structures with different partitions is the principle for detecting the natural or true structures in real-world networks. Accordingly, our proposed ITE based estimator for the source tries to minimize the coding of network partitions brought by the infected tree rooted at all the potential sources, thus minimizing the structural deviation between the cascades from the potential sources and the actual infection process included in G_N. On polynomially growing geometric trees, with increasing tree heterogeneity, the ITE estimator remarkably yields more reliable detection under only moderate infection sizes, and returns an asymptotically complete detection. In contrast, for regular expanding trees, we still observe guaranteed detection probability of ITE estimator even with an infinite infection size, thanks to the degree regularity property. We also algorithmically realize the ITE based detection that enjoys linear time complexity via a message-passing scheme, and further extend it to general graphs. Extensive experiments on synthetic and real datasets confirm the superiority of ITE to the baselines. For example, ITE returns an accuracy of 85%, ranking the source among the top 10%, far exceeding 55% of the classic algorithm on scale-free networks.

CCS Concepts: • **Networks** → **Network algorithms**; • **Mathematics of computing** → **Information theory**

Additional Key Words and Phrases: Source detection, structural entropy, graph theory, inference algorithms

This work was supported by NSF China (Grants No. 42050105, No. 62020106005, No. 62061146002, No. 61960206002), 100-Talents Program of Xinhua News Agency, and the Program of Shanghai Academic/Technology Research Leader under Grant No. 18XD1401800.

Authors' addresses: C. Zhang, Q. Guo, L. Fu, J. Ding, X. Cao, and X. Wang, Shanghai Jiao Tong University, Dongchuan Rd. 800, Shanghai, Shanghai 200240, China; emails: {zhangchong18, johnnyguo, yiluofu, jiaxinding, xdcao, xwang8}@sjtu.edu.cn; F. Long, Xinhua News Agency, Sanyuan St. 17, Beijing, Beijing 100077, China; email: longf05@tsinghua.org.cn; C. Zhou, Chinese Academy of Sciences, Datun Rd. 11, Beijing, China; email: zhouch@Ireis.ac.cn.

1533-5399/2023/03-ART17 $15.00
https://doi.org/10.1145/3568309

ACM Reference format:
Chong Zhang, Qiang Guo, Luoyi Fu, Jiaxin Ding, Xinde Cao, Fei Long, Xinbing Wang, and Chenghu Zhou. 2023. Finding the Source in Networks: An Approach Based on Structural Entropy. *ACM Trans. Internet Technol.* 23, 1, Article 17 (March 2023), 25 pages.
https://doi.org/10.1145/3568309

1 INTRODUCTION

The ubiquity of the Internet, online social networks, and many types of offline/physical networks has fundamentally changed the landscapes of information spreading, nowadays. Unfortunately, the same channels can be utilized to amplify isolated risks such as malicious rumors, computer viruses, malware, or an isolated failure in a power grid network that cause pernicious effects on the society. Therefore, inferring the initiator or the source of the malicious information is critical whether for forensic use or insights to prevent future epidemics.

Because of the wide range of applications, the source detection problem has gained a lot of attention during the past decade. The seminal work belongs to Shah and Zaman [32], which studies the problem under the natural epidemic model, that is, the **susceptible-infected** (**SI**) model. This was followed by numerous efforts [11, 13, 16, 17, 24–26, 36, 38, 40, 41], which investigate the problem in a common paradigm: given an observation O of the graph $\mathcal{G}$ at some time, the goal is to find the node $\hat{v}$ that maximizes the correct detection probability, given by $\mathbf{P}(O|\hat{v})$. Many of those prior arts try to utilize network topological features, and accordingly adopt various statistical quantities to describe the influence of nodes on propagation. Typical examples include (i) *degree* [11], where it is simply believed that the source node is the one surrounded by the most infected neighbors, (ii) *distance* [24, 25, 40, 41], that selects the potential source based on the minimum infection eccentricity, or (iii) *infection size* [13, 26, 32, 36, 38], where the estimators select the node that highly balances the infection size of each neighboring subtree. Despite those significant efforts, we notice that there may still remain some potential room of topology utilization for source detection. In addition, the side information such as infection timestamps, propagation directions, or queries to culprits [1, 7–9, 27, 35, 39] are often hard to obtain in reality due either to the privacy concern or to the unreliability of the truth. Hence, it is a natural way to exploit the structural features available inside the graph as much as possible to enhance the detection performance.

To this end, we present a new metric to seek for richer topological features mentioned above for source detection. The same as in [24, 26, 32, 33, 36, 38], we also assume that the infection spreading process follows an SI model, where a node that is infected with the information retains it forever. Our design of the new metric is mainly inspired by the structural entropy [21], where a principle for detecting the natural or true structures in real-world networks is proposed. The key point of structural entropy is to partition a given graph into different modules, where an exogenous process is launched to continuously collect the message delivery (named a *call* [22]) between nodes uniformly at random. In this manner, the structural entropy provably [21] captures the average number of bits needed in two-dimensional code to encode the receivers of the calls in a lossless way, which fully characterizes the corresponding structural information. (A more detailed introduction can be referred to in Section 3.3.) Accordingly, in our problem, given a snapshot of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, including the knowledge of the infected graph $G_N = (V_I, E_I)$ and beyond, where V_I is the infected nodes set, E_I not only includes all the edges between infected nodes but also those edges on the boundary between infected and uninfected nodes, the question becomes how can our proposed metric, by virtue of structural entropy, leverage more available topology to detect the true infection structure in G_N, thus inferring the source node more accurately?

To answer this question, we first note that on a tree network, for the infected tree rooted at any potential source, the information will eventually spread to the adjacent branches of the root. This provides an intrinsic structure of the infected tree, which is distinguished only by the location of the potential source in the infected tree. Since the structural entropy provides the minimum encoding principle to find the true structure inside a graph, analogously we partition the structure of the infected tree into modules in terms of different propagation branches for any potential source. Considering the message calls both between intra-module nodes and inter-module ones, we encode this two-dimensional structure based on the probability distribution of all infected nodes as the receivers of the calls, thus characterizing the extent to which the constructed structure deviates from the actual infection process. We name this proposed metric as **Infected Tree Entropy** (**ITE**). As can be seen, our ITE based estimator for the source is indeed able to capture more structural features in the following aspects. (i) The natural substructures in the infected graph; we call them different modules. (ii) The mutual connections between the nodes inside a module. (iii) The inter-connections between modules. (iv) The connections to uninfected nodes on the boundary. These features integrally lead to the complete form of the spreading cascade, and, as we will provably demonstrate in later sections (Sections 5 and 6), bring about improved source detection performance. Then, we extend this framework in general graphs by a **breadth-first search** (BFS) heuristic. To the best of our knowledge, we are the first to apply structural entropy in this problem.

Our main contributions are highlighted as follows:

— We propose a new structural entropy based approach for source detection, called the ITE estimator, which utilizes more underlying structural features. In a tree graph, the estimator can be efficiently solved via a message-passing algorithm, whose complexity scales linearly with the infection size. In general graphs, a BFS heuristic is incorporated to approximate the ITE estimator.
— We derive the performance of the ITE estimator on different networks. For geometric trees, with increasing heterogeneity of the subtrees, our estimator remarkably yields more reliable detection under only moderate infection sizes, which effectively prevents the isolated risks spreading to a wide range. Meanwhile, it returns an asymptotic detection probability of 1, which is as good as the best estimator. In contrast, for regular expanding trees, the ITE estimator can still guarantee a non-trivial detection even when the infection size goes to infinity.
— Besides theoretical guarantees, our ITE estimator also exhibits favorable performance in experiments. Specifically, we conduct extensive experiments on both synthetic networks and real-world networks with different scales (the maximum order of magnitude is almost millions) to evaluate the proposed algorithm. The results show that the ITE estimator not only achieves much lower error distance, but also higher accuracy of ranking the source than other source estimators.

Organization. The rest of this article is organized as follows. Section 2 presents the literature review. We introduce the ITE estimator in Section 3. For tree-type networks, we propose an efficient algorithm for its evaluation in Section 4. Section 5 summarizes the main theoretical results. The simulation based performance evaluations will be presented in Section 6, and all the proofs are provided in Section 7. Finally, we conclude the article in Section 8.

2 RELATED WORKS

In this section, we give a brief overview on related works in both source detection and graph entropy fields.

2.1 Source Detection in Networks

It is known that the source detection problem is highly challenging. In the seminal work [32], Shah and Zaman studied the single source inference problem, and proposed rumor centrality, a newly defined centrality quantity, which was proved to be the maximum likelihood estimator on regular trees under the SI model. They claimed that the node with the maximum rumor centrality is the rumor source, which is called the rumor center. They proved that rumor centrality is equivalent to closeness centrality on tree networks, but performs better on general graphs. This work was extended in [33] for general random trees, where the detection probability was quantified.

Later, the rumor centrality method has been further studied by many other researchers, which can identify rumor sources under different propagation models and assumptions. Luo et al. [26] extended the rumor centrality method in a single source scenario to multiple sources. Wang et al. [36] analyzed the performance of rumor centrality for tree networks when there are multiple observations of both sequential and independent diffusion processes from the same source. Dong et al. [13] further proposed a local rumor center method, and studied the detection probability under the assumption that the prior knowledge of suspect nodes is available. Therefore, they reduced the scale of the searching area. By considering the boundary effect in source detection, Yu et al. [38] extended the rumor centrality method to finite tree networks.

Besides the rumor centrality, several other algorithms based on a single snapshot of the network have been proposed. Zhu and Ying [40] proposed a sample path based approach to detect the single source under the **susceptible-infected-recovered** (**SIR**) model, while a message-passing algorithm was proposed under the same scenario by Lokhov et al. [23]. In [19], Lappas et al. analyzed the detection problem under the **independent cascade** (**IC**) model [15] by minimizing the distance between the expected states and the observed states of the nodes. Prakash et al. [28] proposed a minimum description length based algorithm called NETSLEUTH, which used an eigenvector based metric to rank nodes under SI model. Similarly, Fioriti and Chinnici [14] utilized the correlation between the eigenvalue and the age of a node, and introduced the dynamic age algorithm for the single source detection. In addition, there exist several other algorithms that utilized side information for the source detection problem, such as timestamps of the infected nodes [1, 7, 27, 35, 39], or directions from which a node gets infected [8, 9]. All these methods are unable to exploit the structural characteristics as much as possible.

2.2 Measures of Graph Entropy

As an important issue in graph analysis, graph entropy aims to measure the complexity of graphs, which refers to the level of organization of the structural features such as the scaling behavior of degree distribution, community structure, graph spectra, and so forth. In order to capture the inherent structural complexity of graphs, there is a flurry of existing works that contribute to the measures of graph entropy [2, 6, 10, 30, 31]. Most of them are specific forms of the Shannon entropy [34] for different types of distributions extracted from graph structures. Braunstein et al. [6] proposed the notion of von Neumann entropy, defined as the Shannon entropy of the spectrum of the trace rescaled Laplacian matrix of a graph. Based on the network ensembles, Bianconi [4] introduced the Gibbs entropy to determine the code of the network constructed by the ensemble. Rashevsky [29] proposed the entropy measure based on the distribution of the number of topologically equivalent vertices. Following the similar idea, Bonchev and Trinajstić [5] focused on the distribution of the distances between any two vertices. Raychaudhury et al. [30] proposed the first local measure of graph entropy, which is interpreted as a kind of vertex complexity. Dehmer [12] defined an entropy measure based on several parametric information functions, which characterize metrical properties of various graphs.

Table 1. Notation and Definition

Notation	Definition
v^*	original source node
$\hat{v}$	estimated source node
$\mathcal{G}(\mathcal{V}, \mathcal{E})$	underlying network
$G_N(V_I, E_I)$	infected graph
N	the number of infected nodes
d_v	degree of node v
$d_{v(inf)}(d_{v(un)})$	infected (uninfected) degree of node v
T_u^v	subtree rooted at node u and away from v
$\mathcal{P}$	a partition set of $\mathcal{V}$, that is, $\mathcal{P} = \{X_1, X_2, \ldots, X_L\}$
n_j	the number of nodes in module X_j
V_j	the volume of module X_j (the sum of degrees of nodes in X_j)
g_j	the number of inter-edges (edges with exactly one endpoint in X_j) of module X_j
$\mathcal{H}_{\mathcal{P}}(\mathcal{G})$	structural entropy of $\mathcal{G}$ by the partition $\mathcal{P}$
$\mathcal{H}(v, G_N)$	infected tree entropy of node v in G_N
$\mathbb{H}(v, G_N)$	equivalent infected tree entropy of node v in G_N
$T_{BFS}(v)$	the breadth-first-search tree rooted at node v
$\mathbf{P}_c$	correct detection probability of ITE estimator

Recently, Li and Pan [21] proposed the first metric for structural information. They claimed that the structure entropy provided the principle to detect the natural or true structure of a network. To fully exploit the network structure, our proposed source detection metric is based on this idea.

3 INFORMATION SOURCE ESTIMATOR

In this section, we first introduce the information spreading model and formulate the source detection problem. Then, we formally describe structural entropy [21] to lead to the ITE based source estimator in trees and general graphs, respectively. For convenience, we list the key parameters that will be used later in Table 1.

3.1 Spreading Model

We model the network as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes, and $\mathcal{E}$ is the set of edges of the form (i, j) for some i and j in $\mathcal{V}$. In this article, we limit our attention to the case where there is only one source node v^*.

We use the SI epidemic model for the information spreading, where the infected nodes are not allowed to recover. In the SI model, once a node i receives the information, it is called infected, and it independently attempts to infect each of its susceptible neighbors j. The spreading times associated with edges are independent random variables with identical exponential distribution with rate λ. Without loss of generality, we take $\lambda = 1$.

3.2 Source Detection Problem

Given the above spreading model, we observe the infected graph $G_N = (V_I, E_I)$ at some time t, where $|V_I| = N$. We have no prior knowledge of the value of t or the spreading time on each edge $e \in E_I$. All that we can utilize is the structure of the infected graph G_N, including the infected nodes $V_I \in \mathcal{V}$ and edges $V_I \times V_I \cap \mathcal{E}$ between them, as well as those edges on the boundary between infected and uninfected nodes, totally denoted by E_I. Assuming a uniform prior probability of the source node, the source detection problem can be formulated as the **maximum likelihood (ML)**

estimation problem given by

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(G_N | v). \tag{1}$$

3.3 Structural Information of a Network

Recall that we hope to make full use of structural features to infer the source. Also, as we have noted earlier in Section 1, structural entropy [21] is a measure that could fully capture the topological information of a network. Thus, we first briefly reproduce its main technical idea to facilitate our later usage of it for the derivation of our proposed source estimator.

In practice, there exist rich natural substructures in a complex network $\mathcal{G}$, such as various modules, components, or communities composed of different nodes and connections among them. They usually correspond to important subsets of the networks and form a partition $\mathcal{P}$ of the vertices. To characterize the structural information contained in $\mathcal{G}$ relative to $\mathcal{P}$, structural entropy aims to inquire the information content of the substructure $\mathcal{P}$ in $\mathcal{G}$. Specifically, since messages or information can be delivered between nodes through edges, we refer to a flow of messages from a sender m to a receiver n as a *call*, where $\{m, n\} \in \mathcal{E}$, and imagine an exogenous process continuously collects such calls uniformly at random. Hence, at any moment, the probability that a node v is the message receiver is $d_v/2|\mathcal{E}|$, where d_v is the degree of v. Considering both single node and substructures of the network as the receivers of calls, the authors [21] focused on the encoding of the network based on this probability distribution, committed to distinguishing the order from disorder in a noisy structure and identifying the true structure, which is defined as follows.

Definition 1 (Structural Information of a Network by a Partition). Given an undirected and connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, suppose that $\mathcal{P} = \{X_1, X_2, \ldots, X_L\}$ is a partition of $\mathcal{V}$; the structural information of $\mathcal{G}$ by $\mathcal{P}$ is as follows:

$$\begin{aligned}\mathcal{H}_{\mathcal{P}}(\mathcal{G}) &= \sum_{j=1}^{L} \frac{V_j}{2|\mathcal{E}|} H\left(\frac{d_1^{(j)}}{V_j}, \ldots, \frac{d_{n_j}^{(j)}}{V_j}\right) - \sum_{j=1}^{L} \frac{g_j}{2|\mathcal{E}|} \log_2 \frac{V_j}{2|\mathcal{E}|} \\ &= -\sum_{j=1}^{L} \frac{V_j}{2|\mathcal{E}|} \sum_{i=1}^{n_j} \frac{d_i^{(j)}}{V_j} \log_2 \frac{d_i^{(j)}}{V_j} - \sum_{j=1}^{L} \frac{g_j}{2|\mathcal{E}|} \log_2 \frac{V_j}{2|\mathcal{E}|},\end{aligned}$$

where V_j is the volume of module X_j, which is the sum of degrees of nodes in X_j; similarly, $2|\mathcal{E}|$ is the volume of $\mathcal{G}$, n_j is the number of nodes in X_j, $d_i^{(j)}$ is the degree of the i-th node in X_j, and g_j is the number of inter-edges, which are the edges with exactly one endpoint in module X_j.

The structural information of a module X_j consists of two levels: (a) from a module level, the information of the entire X_j as the receiver of messages, and (b) from a node level, the information of each single node $i \in X_j$ as the receiver. The key is that we can omit the module level code when the sender and receiver belong to the same module. Hence, for (a), the information of X_j as the receiver is $-\log_2 \frac{V_j}{2|\mathcal{E}|}$ with probability $\frac{g_j}{2|\mathcal{E}|}$ since we only need consider the deliveries whose senders are not in X_j. For (b), the information for all nodes in X_j as receivers is $H(\frac{d_1^j}{V_j}, \ldots, \frac{d_{n_j}^j}{V_j})$ with probability $\frac{V_j}{2|\mathcal{E}|}$, where $H(\cdot)$ is the entropy function, which is defined as $H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$ for a probability vector $\mathbf{p} = (p_1, \ldots, p_n)$, with $\sum_{i=1}^{n} p_i = 1$. Therefore, the structural entropy indeed captures the average number of bits needed to encode the receivers of the calls in a lossless way, which fully characterizes the structural information of a network with corresponding partitions.

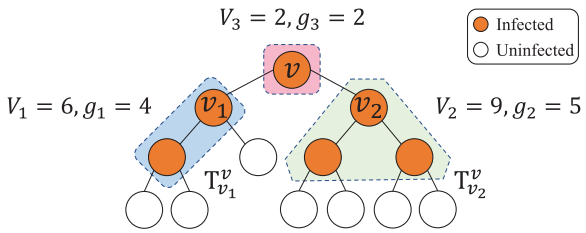


Fig. 1. An illustration of a partition of the infected tree. Node v is selected as the potential source for example, and the three derived modules are in the dotted boxes, respectively.

3.4 ITE Based Source Estimator: Tree Networks

Since the structural entropy captures the structural information of a graph $\mathcal{G}$ with any partition of the nodes, in this section, we introduce the structural entropy based source estimator for tree networks, which we name as the infected tree entropy based source estimator (ITE estimator in short).

Recall that our goal is to find the most likely source node given an observation of the infected tree G_N at some time t. To this end, we try to minimize the structural deviation between the cascades from all the potential sources $v \in G_N$ and the actual infection process. In this way, a natural and reasonable partition of G_N is needed to characterize the structure of spreading from any potential source. As mentioned earlier, there exists such an intrinsic structure of the infected tree, which is specified as follows. Suppose the g neighbors of the node $v \in V_I$ are $v_1, v_2, \ldots, v_g$, in which $d_{v(inf)}$ nodes are infected. For the fact that there exist no cycles in tree networks, the information starting from v will spread to $d_{v(inf)}$ disjoint subtrees, namely, $d_{v(inf)}$ disjoint modules, which form a spreading trajectory to construct G_N together with the node v itself. We call the trajectory determined by any potential source node $v \in G_N$ a *partition of the infected tree*, which is defined as follows.

Definition 2 (Partition of the Infected Tree by a Node). For any potential source $v \in G_N$, the partition of the infected tree G_N by the node v is that

$$\mathcal{P}_v = \left(v, T^v_{v_1}, \ldots, T^v_{v_{d_{v(inf)}}}\right),$$

which satisfies the following properties:

(1) given G_N, $\mathcal{P}_v$ is determined only by the location of v; and
(2) the modules in $\mathcal{P}_v$ are disjoint from each other;

where $T^v_{v_j}$ is the subtree rooted at the node v_j and away from the potential source node v.

To illustrate this definition, a simple example is shown in Figure 1, where we consider the potential source v. Since v has two infected neighbors, v_1 and v_2, the infected nodes are partitioned into three modules: the node v itself, the infected subtree rooted at v_1, and the infected subtree rooted at v_2, that is, $\mathcal{P}_v = \left(v, T^v_{v_1}, T^v_{v_2}\right)$.

Now that we have a partition of the infected nodes given a potential source node v, we can derive the structural information of the infected tree rooted at v, which we define for simplicity as the *infected tree entropy of* v as follows.

Definition 3 (Infected Tree Entropy). Considering that the information spreads in a tree network $\mathcal{G}(\mathcal{V}, \mathcal{E})$, we observe the infected tree G_N at some time. Then, the infected tree entropy of any infected node v, $\mathcal{H}(v, G_N)$, is defined by the structural information of G_N relative to $\mathcal{P}_v$,

ALGORITHM 1: Equivalent ITE Message Passing Algorithm

Input: Infected graph G_N
Output: Equivalent ITE for each node $u \in G_N$

Randomly choose a root node $v^* \in G_N$, and define $f(x, y) = x^{2(y-1)}$;
for *u in* G_N **do**
  if *u is a leaf node* **then**
    $l^{up}_{u \to par(u)} = 1$;
    $p^{up}_{u \to par(u)} = deg(u)$;
  else if *u is the root node* v^* **then**
    $L_{all} = \sum_{v' \in child(v^*)} l^{up}_{v' \to v^*} + 1$;
    $P_{all} = \sum_{v' \in child(v^*)} p^{up}_{v' \to v^*} + deg(v^*)$;
    $\mathbb{H}(v^*, G_N) = {P_{all}}^{2|child(v^*)|} \cdot \prod_{y \in child(v^*)} f(p^{up}_{y \to v^*}, l^{up}_{y \to v^*})$;
  else
    $l^{up}_{u \to par(u)} = \sum_{y \in child(u)} l^{up}_{y \to u} + 1$;
    $p^{up}_{u \to par(u)} = \sum_{y \in child(u)} p^{up}_{y \to u} + deg(u)$;
    $l_{par(u)} = L_{all} - l^{up}_{u \to par(u)}$;
    $p_{par(u)} = P_{all} - p^{up}_{u \to par(u)}$;
    $\mathbb{H}(u, G_N) = {P_{all}}^{2(|child(u)|+1)} \cdot f(p_{par(u)}, l_{par(u)}) \cdot \prod_{y \in child(u)} f(p^{up}_{y \to u}, l^{up}_{y \to u})$;
return $\mathbb{H}(u, G_N)$ for $u \in G_N$

that is,

$$\mathcal{H}(v, G_N) = \mathcal{H}_{\mathcal{P}_v}(G_N) = -\frac{d_v}{V}\log_2\frac{d_v}{V} - \sum_{j=1}^{d_{v(inf)}}\frac{g_j}{V}\log_2\frac{V_j}{V} - \sum_{j=1}^{d_{v(inf)}}\sum_{i=1}^{n_j}\frac{d_i^{(j)}}{V}\log_2\frac{d_i^{(j)}}{V_j}, \tag{2}$$

where d_v is the degree of v in $\mathcal{G}$; V is the volume of the infected tree G_N; g_j, V_j, and n_j are the number of inter-edges, the volume, and the size of j-th subtree of v, respectively; and $d_{v(inf)}$ is the number of infected neighbors of the node v.

Take the node v in Figure 1 as an example. Since the two neighbors of v, v_1, and v_2 are both infected, we have $d_{v(inf)} = 2$. As for the module $T^v_{v_1}$, there are two infected nodes with degree 3, then $V_1 = 6$. Meanwhile, as we can see, the number of inter-edges of $T^v_{v_1}$ is 4, hence, $g_1 = 4$. Similarly, we obtain $V_2 = 9, g_2 = 5; V_3 = 2, g_3 = 2$, and the volume of G_N is $V = V_1 + V_2 + V_3 = 17$. Therefore, the infected tree entropy of v will be

$$\mathcal{H}(v, G_6) = -\frac{2}{17}\log_2\frac{2}{17} - \frac{4}{17}\log_2\frac{6}{17} - \frac{5}{17}\log_2\frac{9}{17} - \frac{2\times 3}{17}\log_2\frac{3}{6} - \frac{3\times 3}{17}\log_2\frac{3}{9} \approx 2.179 \text{ (bits)}.$$

As the structural entropy described in Section 3.3, the infected tree entropy $\mathcal{H}(v, G_N)$ captures the average number of bits needed to encode the two-dimensional structure of G_N by the partition $\mathcal{P}_v$; however, the code itself is beyond our concern in this work. Since $\mathcal{P}_v$ is only determined by the location of node v, any potential source node $v \in G_N$ will determine a structural information of the infected tree. The smaller value of the ITE $\mathcal{H}(v, G_N)$, the lower extent to which the structure of G_N constructed by $\mathcal{P}_v$ deviates from the actual infection process starting from the original source, hence the probability that the node v is the original source of the information will be higher. As

such, we denote the source of our estimator by $\hat{v}$, and the ITE estimator can be formulated as

$$\hat{v} \in \arg\min_{v \in G_N} \mathcal{H}(v, G_N) \tag{3}$$

with ties broken uniformly at random.

3.5 ITE Based Source Estimator: General Graphs

For general graphs, owing to the lack of knowledge of the underlying spanning tree corresponding to the first time that each node gets infected, we use the BFS heuristic to deduce a tree network in the infected graph G_N. We assume that if the node $v \in G_N$ was the source, then the infection process was along the BFS tree rooted at v, $T_{BFS}(v)$. The intuition is that the BFS tree would correspond to all the closest neighbors of v being infected as soon as possible. We should notice that the removed edges by BFS will not be counted in the degree of both end nodes. With this heuristic, we obtain the following source estimator for a general graph.

$$\hat{v} \in \arg\min_{v \in G_N} \mathcal{H}(v, T_{BFS}(v)). \tag{4}$$

As we will empirically show in Section 6, this estimator indeed outperforms the baselines on different networks.

4 ALGORITHM FOR TREES

In order to efficiently find the potential source node with the minimum ITE, we propose a message-passing algorithm for tree networks. To do this, we first simplify the expression of $\mathcal{H}(v, G_N)$.

$$\begin{aligned}\mathcal{H}(v, G_N) &= -\frac{d_v}{V}\log_2\frac{d_v}{V} - \sum_{j=1}^{d_{v(inf)}}\frac{g_j}{V}\log_2\frac{V_j}{V} - \sum_{j=1}^{d_{v(inf)}}\sum_{i=1}^{n_j}\frac{d_i^{(j)}}{V}\log_2\frac{d_i^{(j)}}{V_j} \\ &= \frac{1}{V}\log_2\left[\frac{1}{\prod_{v' \in G_N} {d_{v'}}^{d_{v'}}} \cdot \mathcal{H}_2(v, G_N)\right].\end{aligned} \tag{5}$$

Note that the first term in the real number of the logarithmic function is a constant for each node v, so the value of $\mathcal{H}(v, G_N)$ is only determined by the second term $\mathcal{H}_2(v, G_N)$, where

$$\mathcal{H}_2(v, G_N) = V^{d_v + \sum_{j=1}^{d_{v(inf)}} g_j} \cdot \prod_{j=1}^{d_{v(inf)}} {V_j}^{V_j - g_j}. \tag{6}$$

Moreover, the following proposition states a structural property of the inter-edges in an infected tree.

Proposition 1. *In an infected tree, for any two infected nodes v_1 and v_2, we have*

$$\sum_{j=1}^{d_{v_1(inf)}} g_j(v_1) - \sum_{j=1}^{d_{v_2(inf)}} g_j(v_2) = [d_{v_1(inf)} - d_{v_1(un)}] - [d_{v_2(inf)} - d_{v_2(un)}], \tag{7}$$

where $d_{v(un)}$ denotes the number of uninfected neighbors of the node v.

Proof. The intuition is that the difference of the sum of inter-edges between two nodes is only determined by the respective number of infected and uninfected degrees. Denote the edges on the boundary between infected and uninfected nodes by $E_{boundary}$. Then, for any infected node v_1 in

G_N, based on the definition of inter-edges, we obtain that

$$\sum_{j=1}^{d_{v_1(inf)}} g_j(v_1) = g_1 + g_2 + \cdots + g_{d_{v_1(inf)}} = |E_{boundary}| - d_{v_1(un)} + d_{v_1(inf)}. \tag{8}$$

Similarly, for another infected node v_2, we have

$$\sum_{j=1}^{d_{v_2(inf)}} g_j(v_2) = g_1 + g_2 + \cdots + g_{d_{v_2(inf)}} = |E_{boundary}| - d_{v_2(un)} + d_{v_2(inf)}. \tag{9}$$

By Equations (8) and (9), we can derive

$$\sum_{j=1}^{d_{v_1(inf)}} g_j(v_1) - \sum_{j=1}^{d_{v_2(inf)}} g_j(v_2) = [d_{v_1(inf)} - d_{v_1(un)}] - [d_{v_2(inf)} - d_{v_2(un)}].$$

This completes the proof of Proposition 1. □

Based on Proposition 1, we can further simplify $\mathcal{H}_2(v, G_N)$ by omitting the constant term as follows:

$$\mathbb{H}(v, G_N) = V^{2d_{v(inf)}} \cdot \prod_{j=1}^{d_{v(inf)}} V_j^{2(n_j-1)}. \tag{10}$$

Therefore, the ITE estimator is transformed into finding the potential source node $\hat{v}$ with the minimum value of $\mathbb{H}(\hat{v}, G_N)$ in G_N, which we call the *equivalent ITE*. To calculate the equivalent ITE for each infected node u, we first traverse all infected nodes and record their degrees to obtain the volume V of G_N for the preparation step with a complexity of $O(N + |E_I|)$. In the next step, we select any node v^* as the root and calculate the size n_j and the volume V_j of all of its subtrees. This can be done by having each infected node u pass two messages to its parent node: the size of u's subtree $l^{up}_{u\to par(u)}$, and the corresponding volume $p^{up}_{u\to par(u)}$. The parent node adds those $l^{up}_{u\to par(u)}$ messages together and those $p^{up}_{u\to par(u)}$ messages together to obtain the size and the volume of its own subtree, respectively. These messages are then passed upward until the root node v^* receives all its children's messages, by which it will calculate its equivalent ITE.

Meanwhile, combining all these two messages of its children and the messages of itself, the root node can obtain two global values L_{all} and P_{all} that record the size N and the volume V of G_N, respectively. With L_{all}, each infected node u will then obtain the size of its parent's subtree by $l^{up}_{u\to par(u)}$ subtracted from L_{all}, which we call $l_{par(u)}$, and similarly, the volume of its parent's subtree can be obtained by $p_{par(u)} = P_{all} - p^{up}_{u\to par(u)}$. As a result, we can calculate the equivalent ITEs for any infected node u. The complexity of this step is $O(N)$. Thus, the message-passing algorithm is able to calculate the equivalent ITE for each node in G_N using only $O(N + |E_I|)$ computations, which is still the same order as the infection size even in the graphs whose scale grows exponentially with the diameter. The pseudocode for this message-passing algorithm is included in Algorithm 1 by omitting the preparation step, where we define $f(x, y) = x^{2(y-1)}$ to facilitate the delineation.

5 MAIN RESULTS

In this section, we present the main theoretical results of the ITE estimator under different graph structures.

5.1 Trivial Detection on Line Graphs

We start from a trivial structure, which is a line. Defining $\mathbf{P}_c$ as the correct detection probability of the ITE estimator under the infection size N, we will establish the following result.

Theorem 1. *Suppose the information spreads on a line graph where the degree of each node is* 2 *as per the SI model. Then we have*

$$\mathbf{P}_c = O\left(\frac{1}{\sqrt{N}}\right).$$

We can see that the correct detection probability scales as $N^{-1/2}$ on the line graph, which is trivial when N goes to infinity. The intuition for this result is that the structure of the line graph is so trivial that the ITE estimator could provide very little structural information of the source. We defer the proof of this theorem to Section 7.1.

5.2 Performance Guarantee on Regular Expanding Trees

We next consider the detection performance on regular expanding trees, where each node has degree $d \geq 3$. In this case, the tree expands quickly with the increase of the depth, and the structure is more complex than a line. We obtain the following result of our estimator.

Theorem 2. *Suppose the information spreads on a regular tree with degree* $d \geq 3$ *as per the SI model. Then*

$$0 < \lim_{N\to\infty} \mathbf{P}_c \leq \frac{1}{2}.$$

Intuitively, for such an infinite scale of infected trees, the correct detection probability for one randomly selecting an infected node tends to zero. However, due to the degree regularity and enough structural complexity in the network, our estimator could capture the structural features inside, and the left inequality in Theorem 2 indicates that ITE still performs the detection with a strictly positive probability even when the infection size N goes to infinity. The above result also says that the detection probability is bounded by 1/2. One can imagine that the source node has infected one neighbor. Because of the memoryless property of the spreading model, we can treat these two nodes as a new single enlarged source. Combined with the regularity of the tree, the later spreading is symmetric about these two nodes, and even the best estimator will not be able to distinguish between them, so the detection probability can never be greater than 1/2. Therefore, the performance of the ITE estimator is guaranteed on regular trees with $d \geq 3$. This theorem is proved in Section 7.2.

5.3 Asymptotically Complete Detection on Geometric Trees

Geometric trees are first introduced in [33], which grow polynomially in size with the diameter of the tree. They are parameterized by constants α, b, and c, with $\alpha \geq 0$, $0 < b \leq c$. Let $n^i(r)$ denote the number of nodes in the i-th subtree of the root node v^* at distance exactly r from the subtree's root node, and the degree of v^* is d_{v^*}; then we require that for all $1 \leq i \leq d_{v^*}$

$$br^\alpha \leq n^i(r) \leq cr^\alpha. \tag{11}$$

The condition of Equation (11) describes that each of the subtrees of the root node should satisfy polynomial growth with parameter $\alpha \geq 0$.

We consider the scenario where the information starts spreading from the root node of the geometric tree, and obtain the following result.

Theorem 3. *Consider a geometric tree with* $\alpha > 0$, $0 < b \leq c$, *and the root node* v^* *with degree* $d_{v^*} \geq 3$ *such that*

$$b(d_{v^*} - 1) > c \cdot \log_{(d_{v^*}-1)} c.$$

Suppose the information starts spreading from v^* as per the SI model. Then,

$$\lim_{N\to\infty} \mathbf{P}_c = 1.$$

We prove this result in Section 7.3. This theorem says that the ITE estimator has an asymptotic detection probability of 1, which is as good as the best possible estimator on this more general and heterogeneous tree network. In the sequel, we further demonstrate a remarkable advantage, where our estimator has more reliable detection performance compared to those centrality based algorithms.

5.4 Advantages with Heterogeneity of Geometric Trees

The parameter α characterizes the growth of the geometric trees, while the ratio c/b describes the heterogeneity of the subtrees. When $c/b \approx 1$, the subtrees are somewhat regular, whereas for c/b large enough, there is substantial heterogeneity in the subtrees. The following result demonstrates a further advantage of our estimator with increasing tree heterogeneity.

Theorem 4. *For a geometric tree with parameters $\alpha > 0$, $0 < b \le c$, and the root node v^* with degree $d_{v^*} \ge 3$. Let α, b, and d_{v^*} be fixed, then the performance of the ITE estimator is more robust to the increase of the ratio c/b compared to those centrality based algorithms under only moderate infection sizes.*

Remark: A more robust performance on detection under moderate infection sizes means that our estimator will be less affected by the increasing tree heterogeneity, and has a higher probability to correctly detect the source before the infection spreads to a wide range, which is of importance in reality. Intuitively, with the increasing heterogeneity in geometric trees, it is generally harder to correctly detect the source for any algorithm due to the more complex structures of the infected tree, where most centrality based algorithms [11, 32, 40] will probably be fooled to select the nodes with large degrees. In contrast, as we will prove in Section 7.4, the ITE estimator will not be completely dictated by the centrality of the potential source.

6 PERFORMANCE EVALUATION

In this section, we evaluate the performance of the ITE estimator on different networks.

6.1 Baseline Algorithms

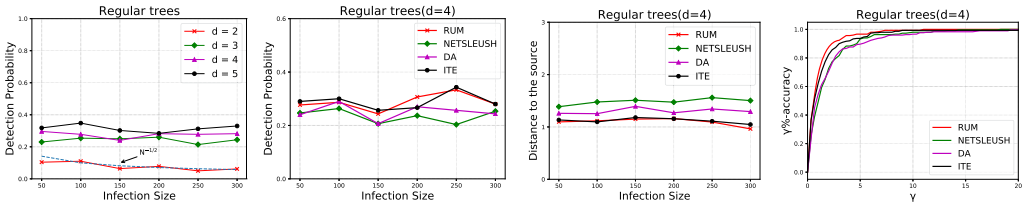
For fairness, we aim to select the algorithms proposed under the same SI model, requiring a single observation of the network and no side information, which are summarized as below.

- **RUM:** Find the node with maximum rumor centrality [32]. This classic algorithm is proved to be the maximum likelihood estimator on regular trees under the SI model, and can be applied to general graphs by BFS heuristic.
- **DA:** Find the node with maximum dynamic age [14]. Due to the high complexity in computing the dynamic age even for a single node, generally $O(N^3)$, let alone in finding the maximum one, this algorithm is not suitable for identifying sources in large-scale networks.
- **NETSLEUTH:** Find the node with maximum value in the smallest eigenvector of the submatrix constructed by infected nodes in the graph Laplacian matrix. This is a spectral graph theory based approach proposed in [28].

6.2 Evaluation Metrics

We evaluate the performance of the algorithms with the following metrics.

- Detection probability is the correct detection rate of the source estimator.
- Distance is the average number of hops from the estimated source to the original source, which is an often used metric for the source detection problem.



(a) $d = 2, 3, 4, 5$ (b) Detection probability (c) Distance to the source (d) γ%-accuracy

Fig. 2. Performance on regular trees. (a) ITE detection probability on regular trees. (b)–(d) Comparison between ITE and three baselines on regular trees when $d = 4$.

— γ% accuracy versus the rank percentage describes the probability that the original source is ranked among the top γ percent. Note that a source detection algorithm can not only provide a source estimator, but also can be used to rank the infected nodes by their likelihood to be the source. For example, RUM ranks the nodes in a descendant order according to their rumor centrality, whereas ITE ranks the nodes in an ascendant order of their infected tree entropy. We wish that the original source lies in the top ranked nodes with a high probability.

6.3 Regular Trees

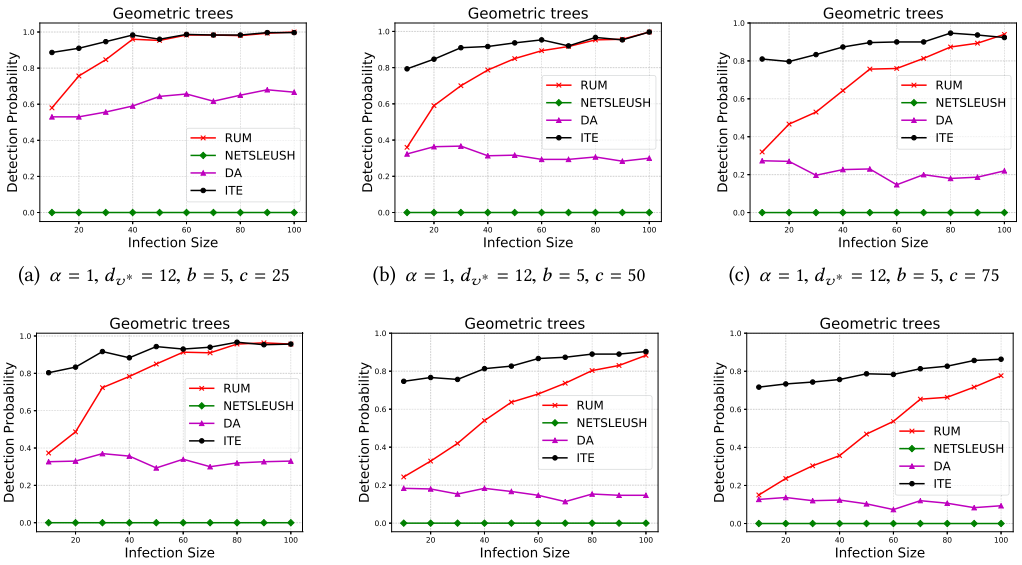
In this section, we evaluate the ITE estimators on regular trees. For each simulation, we select the source node uniformly at random and synthesize the spreading as per the SI model. We conduct 500 simulation runs for each configuration on each network.

The detection probability of the ITE estimator versus the infection size on different regular trees is shown in Figure 2(a). As can be seen, the detection rate scales as $N^{-1/2}$ as derived in Theorem 1 for line graphs, while for regular expanding trees with $d \geq 3$, the estimator has a non-trivial detection probability, which is less than $1/2$ and does not decay to 0 as predicted. For regular trees with $d = 4$, we compare the performance of the ITE estimator with the other three baselines. As shown in Figure 2(b), although rumor centrality [32] is proved to be the maximum likelihood estimator on regular trees, we can see that ITE performs very similarly to RUM. Moreover, even under some infection sizes, e.g., 50, 100, 150 and 250, ITE has a little higher detection probability than RUM. Meanwhile, compared to the other two algorithms, the detection probability of ITE is still higher. To make a further comparison, we also present the error distance and γ% accuracy of four algorithms in Figures 2(c) and (d), where the error distance curve of ITE almost coincides with that of RUM, which two have the lowest error distance. In addition, compared to the other two methods, ITE has a much closer ranking performance to RUM. Both results demonstrate that ITE performs similarly to the MLE estimator on regular trees.

6.4 Geometric Trees

In this section, we compare the performance of the algorithms on geometric trees. The results are averaged over 500 simulation runs. Figure 3 presents the results on geometric trees under different settings of c/b. In Figures 3(a)–(c) we fix $\alpha = 1$, $b = 5$, and $d_{v^*} = 12$, whereas $\alpha = 1$, $b = 10$, and $d_{v^*} = 12$ are fixed in Figures 3(d)–(f). We have the following three observations.

— Firstly, the detection probability of the ITE estimator is closer to 1 when the infection scale becomes larger as predicted by Theorem 3.
— Secondly, by comparing these subgraphs, we can explicitly see that ITE is less affected with the increasing ratio of c/b, and the gap of detection probabilities between ITE and the other three algorithms becomes wider under the same infection size.



(a) $\alpha = 1, d_{v^*} = 12, b = 5, c = 25$ (b) $\alpha = 1, d_{v^*} = 12, b = 5, c = 50$ (c) $\alpha = 1, d_{v^*} = 12, b = 5, c = 75$

(d) $\alpha = 1, d_{v^*} = 12, b = 10, c = 50$ (e) $\alpha = 1, d_{v^*} = 12, b = 10, c = 100$ (f) $\alpha = 1, d_{v^*} = 12, b = 10, c = 150$

Fig. 3. Performance on geometric trees.

— Thirdly, comparing the figures in the respective three columns, we take Figures 3(c) and(f) as an example. Under the same settings of $\alpha = 1$, $d_{v^*} = 12$, and $c/b = 15$, the only difference is the value of b. We observe that the gap mentioned in the second observation is even wider when $b = 10$ than when $b = 5$. The reason is that although c/b is the same, a larger b brings about more variance of the heterogeneity, which leads to more heterogeneity in each subtree of the root node.

Thus, our estimator has more advantages when there exists more heterogeneity in geometric trees, which is guaranteed by Theorem 4.

6.5 Graph Networks

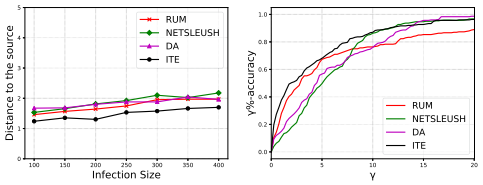
We next perform experiments on both synthetic and the following six real-world networks to demonstrate the efficacy of the ITE estimator on general graphs. Table 2 shows the statistics of the networks.

— **Power grid** network is the power grid of the Western states of the U.S.
— **LastFM** is a social network of LastFM users that was collected in March 2020.
— **IAS** is an **Autonomous Systems** (**AS**) peering graph inferred from Oregon route-views on March 31, 2001.
— **Facebook** is a page-page graph of verified Facebook sites collected in November 2017.
— **Enron** email network covers all the email communication within a dataset of around half million emails, where nodes represent email addresses.
— **Gowalla** is a location based social network of friendships over the period of February 2009 to October 2010.

6.5.1 Synthetic Networks. We select two very popular models for networks: small-world networks [37] and scale-free networks [3]. For both topologies, the underlying graphs contain 5,000 nodes. The small-world network is generated by rewiring edges and contains 25,000 edges, while

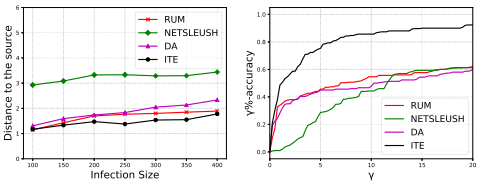
Table 2. Statistics of Networks

Networks	Description	$\lvert\mathcal{V}\rvert$	$\lvert\mathcal{E}\rvert$	$\lvert\mathcal{E}\rvert/\lvert\mathcal{V}\rvert$	Diameter
Small-world	Synthetic	5,000	25,000	5.0	9
Scale-free	Synthetic	5,000	9,996	2.0	9
Power grid	Infrastructure	4,941	6,594	1.3	46
LastFM	Online Social	7,624	27,806	3.6	15
IAS	Autonomous System	10,670	22,002	2.1	9
Facebook	Online Social	22,470	171,002	7.6	15
Enron	Communication	36,692	183,831	5.0	11
Gowalla	Friendship	196,591	950,327	4.8	14



(a) Distance to the source (b) γ%-accuracy

Fig. 4. Performance on the small-world network.



(a) Distance to the source (b) γ%-accuracy

Fig. 5. Performance on the scale-free network.

the scale-free network is generated by preferential attachment with 9,996 edges. We vary the infection size from 100 to 400 and run each simulation 300 times independently. In each simulation, the source node is chosen uniformly across node degrees to avoid the bias toward small degree nodes.

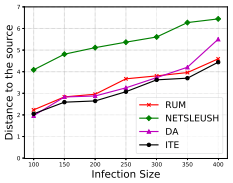
Figures 4 and 5 show the performance on the above two networks, respectively. For both the plots of γ% accuracy versus the rank percentage γ, we pick the infection size 400. As can be seen, the ITE estimator performs lower error distances to the original source compared to the other three algorithms in almost all cases. The improvement is more obvious in a small-world network than in scale-free network. For the small-world network used here, the average ratio of edges to nodes is 5, whereas for the scale-free network, the average ratio is 2. Thus, the small-world network is less tree-like. This may explain why ITE outperforms more apparently than the other three algorithms.

As for the γ% accuracy, ITE has similar or better performance on a small-world network, but clearly outperforms the baseline algorithms on the scale-free network. For example, the 10% accuracy of ITE is 85%, which is significantly higher than that of other algorithms, e.g., 55% for RUM, 48% for DA, and 44% for NETSLEUTH. The reason behind this may be the existence of many large degree *hubs* in the scale-free network, then the network has more heterogeneity compared to the small-world network.

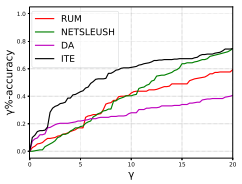
6.5.2 Real-World Networks. We conduct experiments on six real-world networks derived from SNAP [20] and KONECT [18], as shown in Table 2. Among them, the **Power grid** (**PG**) and LastFM networks are moderate size, and the remaining four networks are large scale. The source detection algorithms are evaluated systematically on these datasets, as they not only cover a wide range of categories, but also vary in size and the ratio of the number of edges to the number of nodes.

We adopt the similar settings as in Section 6.5.1 except for the infection size on the large-scale networks. In these four networks, we set the infection size range from 200 to 1,000, and remove the DA algorithm due to the rather high complexity as mentioned in Section 6.1.

Figures 6(a), 7(a), 8(a), 9(a), 10(a), and 11(a) show the distance versus the infection size. In general, owing to the more complex structures, all the algorithms yield a farther node from the source

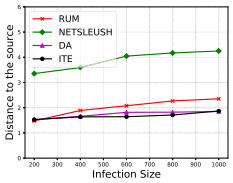


(a) Distance to the source

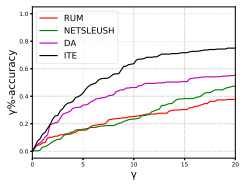


(b) γ%-accuracy

Fig. 6. Performance on the Power grid network.

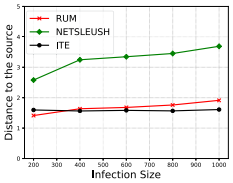


(a) Distance to the source

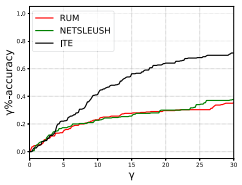


(b) γ%-accuracy

Fig. 7. Performance on the LastFM network.

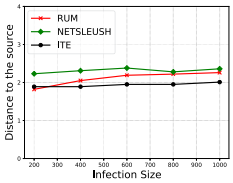


(a) Distance to the source

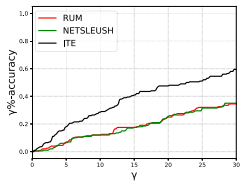


(b) γ%-accuracy

Fig. 8. Performance on the IAS network.

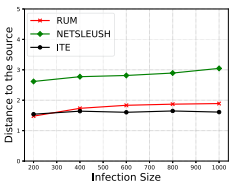


(a) Distance to the source

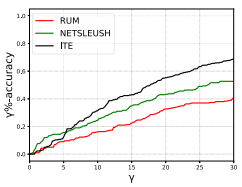


(b) γ%-accuracy

Fig. 9. Performance on the Facebook network.

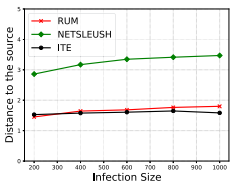


(a) Distance to the source

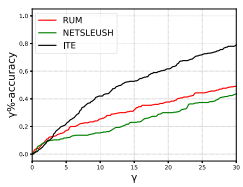


(b) γ%-accuracy

Fig. 10. Performance on the Enron network.



(a) Distance to the source



(b) γ%-accuracy

Fig. 11. Performance on the Gowalla network.

when the infected graph grows larger. Nonetheless, similar to the performance on synthetic networks, the ITE estimator establishes lower distances to the original source in almost all the cases. Moreover, as the infection size increases, we can see that ITE performs better and better than RUM. This phenomena is more obvious on the less tree-like networks, such as the Facebook network and the Enron network, where the average ratio of edges to nodes is 7.6 and 5, respectively.

In addition, the plots of γ% accuracy versus γ for 400 infection size are shown in Figures 6(b), 7(b), 8(b), 9(b), 10(b), and 11(b), respectively, from which we have the following observations. On one hand, we can see that ITE has a remarkable advantage of ranking nodes over the baselines on almost all the networks. In PG networks, the 10% accuracy of ITE is 60%, which is higher than that of other algorithms. However, as γ increases, NETSLEUTH becomes closer to ITE in terms of accuracy. The reason may be that the PG network has weak degree heterogeneity, which is similar to small-world networks in Figure 4(b). On the other hand, ITE improves accuracy for the LastFM network, and the 10% accuracy of ITE is close to 65%, far exceeding that of baseline algorithms. For example, it is even more than twice that of the classic RUM method. This is due to the fact that the LastFM network contains many hubs and has more heterogeneous degree distributions, and the ITE estimator is capturing the significant heterogeneity, similar to the scale-free networks in Figure 5(b). In addition, the same significant advantages can be seen on the IAS and the Gowalla networks. From this point of view, by leveraging more structural features, ITE can better rank the infected nodes by their likelihood to be the source in most cases.

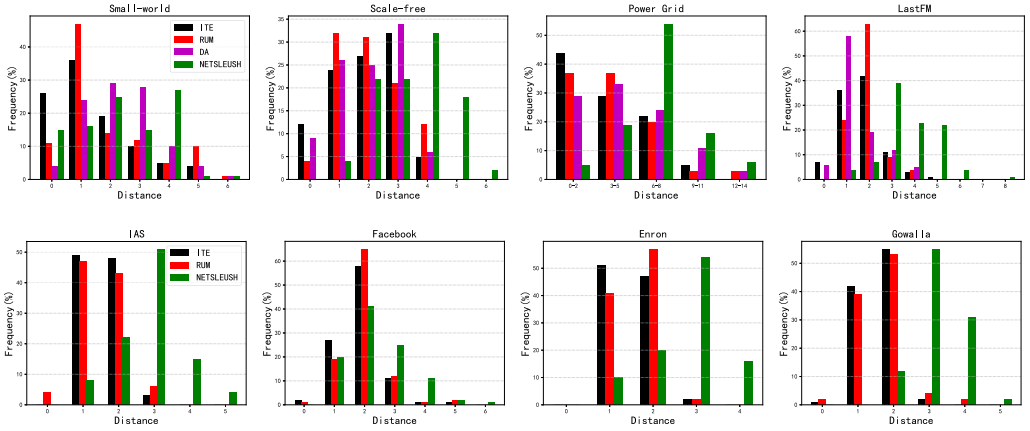
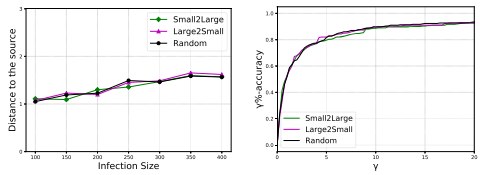
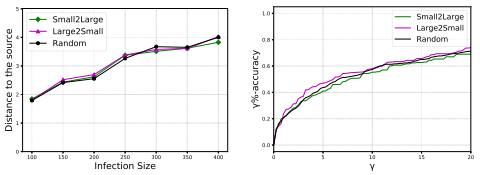


Fig. 12. Frequency of error distances across different topologies.



(a) Distance to the source (b) γ%-accuracy

Fig. 13. Comparison of three BFS policies on the scale-free network.



(a) Distance to the source (b) γ%-accuracy

Fig. 14. Comparison of three BFS policies on the Power grid network.

A further comparison on frequency of error distances across all the eight topologies is shown in Figure 12, where we pick $N = 400$. Our results show that ITE mostly and consistently find the actual source mostly within distance 2; besides, the probability of correct detection event (when error distance is 0) is generally higher than all the baseline methods. For the PG network, source detection becomes quite hard for all the methods. The reason for the same is the high sparsity of the PG network. Even then, ITE finds the source within two hops form the actual source with a probability around 45%.

6.5.3 Effects of Different BFS Trees. Since the ITE estimator on general graphs adopts BFS heuristic, and the BFS tree is not unique, we design the following three BFS policies at the same depth to demonstrate how different BFS trees affect the performance of ITE: (1) Random: traverse in random order, which we adopt in the above experiments; (2) Small2Large: traverse in ascending order of node IDs; and (3) Large2Small: traverse in descending order of node IDs. We select one synthetic network, scale-free network, and one real-world network, PG network, and other settings are the same as before. Figures 13 and 14 show the comparison results. The observation is that different BFS policies yield little variance on the performance of ITE, demonstrating that the selection of BFS tree has little effect on the ITE estimator.

7 PROOFS

This section establishes the proofs of Theorems 1–4. All of them utilize the ITE estimator to obtain the desired results.

7.1 Proof of Theorem 1

Before giving the proof of the result on line graphs, we first simplify the equivalent ITE in Equation (10) by the regularity of the trees. On a regular tree with degree $d \geq 2$, it is easy to see that $V = d \cdot N$ and $V_j = d \cdot n_j$, then

$$\mathbb{H}(v, G_N) = d^{2(N-1)} N^{2d_{v(inf)}} \cdot \prod_{j=1}^{d_{v(inf)}} n_j^{2(n_j-1)}. \tag{12}$$

By omitting the constant term $d^{2(N-1)}$, we denote the equivalent ITE on regular trees by $\mathbb{H}_r(v, G_N)$. That is,

$$\mathbb{H}_r(v, G_N) = N^{2d_{v(inf)}} \cdot \prod_{j=1}^{d_{v(inf)}} n_j^{2(n_j-1)}. \tag{13}$$

Recall that our goal is to find under what conditions that the original source v^* has the minimum value of $\mathbb{H}_r(v^*, G_N)$. On line graphs, for each infected node v, we have $d_{v(inf)} = 2$ except for two leaf nodes at the endpoints, so that we can omit the first term $N^{2d_{v(inf)}}$ for all non-leaf nodes. On the other hand, $N^{2d_{v(inf)}}$ only increases with the power of $d_{v(inf)}$, while the second term of $\mathbb{H}_r(v, G_N)$ grows exponentially, denoted by

$$\mathbb{H}_r^*(v, G_N) = \prod_{j=1}^{d_{v(inf)}} n_j^{2(n_j-1)}. \tag{14}$$

To begin with, we state the following lemma that characterizes the general form of the function in Equation (14) for further analysis. We present its proof later in this section.

Lemma 1. *For the function $g(\mathbf{x}) = \prod_{j=1}^{b} {x_j}^{2x_j-2}$, object to $\sum_{j=1}^{b} x_j = c$ (a constant), and $x_j > 0$, for any $j = 1, 2, \ldots, b$, then we have the following results:*

(1) $g(\mathbf{x})$ is strictly convex.
(2) $g(\mathbf{x})$ has the minimum value as $x_1 = x_2 = \cdots = x_b = \frac{c}{b}$.

As such, when the infection size N goes to infinity, the equivalent ITE of two leaf nodes will be much larger than that of non-leaf nodes. So we can only focus on the second term of all the non-leaf nodes, that is,

$$\mathbb{H}_r^*(v, G_N) = {n_1}^{2n_1-2} \cdot {n_2}^{2n_2-2},$$
$$\text{subject to } n_1 + n_2 = N - 1,\ n_1, n_2 \in \mathbb{Z}^+.$$

Based on Lemma 1, $\mathbb{H}_r^*(v, G_N)$ is strictly convex; then the intersection of the surface $\mathbb{H}_r^*(v, G_N)$ with the plane $n_1+n_2 = N-1$, denoted by l, is a strictly convex curve. As can be seen that $\mathbb{H}_r^*(v, G_N)$ is symmetric with the plane $n_1 = n_2$, so is the curve l. Therefore, considering the restriction that n_1 and n_2 are both positive integers, we conclude that $\mathbb{H}_r^*(v, G_N)$ has the minimum value under the following condition:

$$\begin{cases} n_1 = n_2, & N-1 \text{ is even.} \\ n_1 = n_2 + 1 \text{ or } n_2 = n_1 + 1, & N-1 \text{ is odd.} \end{cases}$$

Then, the correct detection probability $\mathbf{P}_c$ on the line graph is given by

$$\mathbf{P}_c = \mathbf{P}(n_1 = n_2) + \frac{1}{2}\left(\mathbf{P}(n_1 = n_2 + 1) + \mathbf{P}(n_2 = n_1 + 1)\right). \tag{15}$$

The same condition in Equation (15) has been proved by Shah and Zaman [32] that $\mathbf{P}_c$ scales as $O(1/\sqrt{t})$, where t is the observation time when the infection size is N. We note that the information

spreading on a line graph is divided into two independent Poisson processes starting from the original source and spreading in opposite directions with rate 1. Based on the property of the Poisson process, we will then conclude that

$$\mathbf{P}_c = O\left(\frac{1}{\sqrt{N}}\right), \tag{16}$$

which completes the proof of Theorem 1.

Proof of Lemma 1. Firstly, we transform the expression of $g(\mathbf{x})$ as follows.

$$g(\mathbf{x}) = \prod_{j=1}^{b} x_j^{2x_j-2} = e^{(2x_1-2)\ln x_1+(2x_2-2)\ln x_2+\cdots+(2x_b-2)\ln x_b}.$$

Denoting

$$h(\mathbf{x}) = (2x_1-2)\ln x_1 + (2x_2-2)\ln x_2 + \cdots + (2x_b-2)\ln x_b,$$

then we can obtain the Hessian matrix of $h(\mathbf{x})$:

$$\mathbf{A} = \begin{bmatrix} \frac{2x_1+2}{{x_1}^2} & 0 & \cdots & 0 \\ 0 & \frac{2x_2+2}{{x_2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2x_b+2}{{x_b}^2} \end{bmatrix}. \tag{17}$$

Since $x_j > 0$, for $j = 1, 2, \ldots, b$, the matrix $\mathbf{A}$ is positive-definite, and we derive that $h(\mathbf{x})$ is strictly convex. Considering the convexity of e^x, we obtain that $g(\mathbf{x}) = e^{h(\mathbf{x})}$ is strictly convex, which completes the first part of the proof.

Due to the monotonicity of e^x, $g(\mathbf{x})$ has the minimum value when $h(\mathbf{x})$ does. Based on the Lagrange Multipliers, we first define the Lagrange function as follows:

$$z(\mathbf{x}) \triangleq h(\mathbf{x}) + \lambda(c - x_1 - x_2 - \cdots - x_b).$$

Then we obtain

$$\begin{cases} \dfrac{\partial z}{\partial x_1} = 2\ln x_1 + 2 - \dfrac{2}{x_1} - \lambda = 0, \\ \dfrac{\partial z}{\partial x_2} = 2\ln x_2 + 2 - \dfrac{2}{x_2} - \lambda = 0, \\ \quad \vdots \\ \dfrac{\partial z}{\partial x_b} = 2\ln x_b + 2 - \dfrac{2}{x_b} - \lambda = 0, \\ x_1 + x_2 + \cdots + x_b = c. \end{cases} \tag{18}$$

Denote $p(x) = 2\ln x + 2 - \frac{2}{x}$, then $p'(x) = \frac{2x+2}{x^2} > 0$, so $p(x)$ is strictly increasing. As such, we observe the first b equations in Equation (18). Since the parameter λ remains the same, it can be concluded that $p(x_1) = p(x_2) = \cdots = p(x_b)$. Then the solution to the equations, that is, the condition when $g(\mathbf{x})$ has the minimum value, becomes

$$x_1 = x_2 = \cdots = x_b = \frac{c}{b},$$

which completes the second part of the proof. □

7.2 Proof of Theorem 2

To establish that on regular trees with $d \geq 3$, the probability of correct detection of the source using the ITE estimator is strictly positive and upper bounded by 1/2, irrespective of N, we need to find out under what conditions the source node v^* has the minimum $\mathbb{H}_r(v^*, G_N)$. Denote the d neighbors of v^* by $v_1, v_2, \ldots, v_d$, and let the random variable $T_i(t)$ be the number of infected nodes in the i-th subtree of v^* at time t. To find the lower bound, we first define a special case $S_n(t)$, under which the source node v^* is proved to be correctly detected. After that, we state that $S_n(t)$ is lower bounded by a strictly positive constant.

Define $S_n(t)$ as the event when all the d subtrees of the source have between n and $(d-1)n$ infected nodes. That is,

$$S_n(t) = \bigcap_{i=1}^{d} \{n \leq T_i(t) \leq (d-1)n\}, \text{ for } n > 0. \tag{19}$$

We shall make sure that $\mathbb{H}_r(v^*, G_N)$ is the minimum among all the infected nodes under this event. Considering that as t goes to infinity, n will be large enough, and the first term $N^{2d_{v(inf)}}$ only increases with the power of $d_{v(inf)}$, where $1 \leq d_{v(inf)} \leq d$. In this case, the value of the equivalent ITE of each infected node v is mainly determined by the exponential term $\mathbb{H}_r^*(v, G_N)$.

Next, we note that under the event $S_n(t)$, we have $T_i = n + c_i$ (t is omitted for simplicity) where $0 \leq c_i \leq (d-2)n$, for $1 \leq i \leq d$. Suppose w.l.o.g. that $c_d = \max(c_1, c_2, \ldots, c_d)$. Therefore,

$$T_{v^*}^{v_d} = (d-1)n + \sum_{i=1}^{d-1} c_i + 1 > (d-1)n \geq T_d. \tag{20}$$

Then the remaining $d-1$ subtrees of v_d have $n + c_d - 1$ infected nodes in all. Since $0 \leq c_d \leq (d-2)n$, we have

$$\frac{n + c_d - 1}{d-1} \leq n - \frac{1}{d-1} < n. \tag{21}$$

Based on the Lemma 2 stated below, to ensure the value of $\mathbb{H}_r^*(v_d, G_N)$ is as small as possible, the sizes of remaining $d-1$ branches should satisfy the nearest integer point from the minimum point $(\underbrace{\frac{n + c_d - 1}{d-1}, \frac{n + c_d - 1}{d-1}, \ldots, \frac{n + c_d - 1}{d-1}}_{d-1})$ as presented in Lemma 1. Considering Equation (21), the sizes of v_d's subtrees will be $(n - a_1, n - a_2, \ldots, n - a_{d-1}, T_{v^*}^{v_d})$, where $a_1, a_2, \ldots, a_{d-1}$ are all non-negative integers.

Lemma 2. *For the function $g(\mathbf{x}) = \prod_{j=1}^{b} x_j^{2x_j - 2}$, object to $\sum_{j=1}^{b} x_j = c$ (a constant), and $x_j \in \mathbb{N}^+$, for any $j = 1, 2, \ldots, b$. Then $g(\mathbf{x})$ has the minimum value when $(x_1, x_2, \ldots, x_b)$ reaches the nearest integer point from the minimum point $Q_0(\underbrace{c/b, c/b, \ldots, c/b}_{b})$.*

Different from Lemma 1, the variables x_j in Lemma 2 are all positive integers, and we present its proof later in this section. Further, the following Lemma 3 states a property of the positive, strictly convex, and monotonically increasing function that can be easily derived.

Lemma 3. *If $f(x)$ is positive, strictly convex, and monotonically increasing, then we have that*

$$f(x_1) \cdot f(x_2) \cdots f(x_k) < f(x_1 - b_1) \cdot f(x_2 - b_2) \cdots f(x_{k-1} - b_{k-1}) \cdot f(x_k + B),$$

where $x_1 \leq x_2 \leq \cdots \leq x_k$, $b_i \geq 0$, and $\sum_{i=1}^{k-1} b_i = B$.

Combining Equations (20) and (21) and Lemma 3, we obtain that $\mathbb{H}_r^*(v^*, G_N) < \mathbb{H}_r^*(v_d, G_N)$. Next, in the same way, for other neighboring nodes of v^*, it can be proved that, as $N \to \infty$,

$$\mathbb{H}_r^*(v_i, G_N) > \mathbb{H}_r^*(v_d, G_N) > \mathbb{H}_r^*(v^*, G_N), \text{ for } 1 \le i \le d-1.$$

From the later proof of Lemma 2, we can see that for an infected node v, with the infection size of each subtree as an integer coordinate $(x_1, x_2, \ldots, x_d)$, denoted by $C(v)$, its Euclidean distance to Q_0 is a critical factor of $\mathbb{H}_r^*(v, G_N)$, which is also the variance of $C(v)$. In other words, $\mathbb{H}_r^*(v, G_N)$ will be smaller with high probability when $C(v)$ has lower variance. Though we cannot derive a complete conclusion due to the asymmetry property of $g(\mathbf{x})$, this is an obvious trend because of the convexity of $g(\mathbf{x})$. Therefore, for other non-neighboring nodes v', as N goes to infinity, the variance of $C(v')$ will be much greater than that of the actual source v^*, hence we conclude that, as $N \to \infty$,

$$\mathbb{H}_r^*(v^*, G_N) < \mathbb{H}_r^*(v', G_N).$$

To sum up, we obtain that under $S_n(t)$, the ITE estimator correctly detects the source when N goes to infinity. Moreover, the probability of the event $S_n(t)$ was proved in Theorem 2 in [32] that is lower bounded by a strictly positive constant. As for the upper bound 1/2, we can easily derive from the regularity of the tree and the symmetry between the source and the first infected neighboring node. This completes the proof of Theorem 2.

Proof of Lemma 2. (1) If c/b is an integer, the conclusion is directly obtained from Lemma 1.

(2) If $c/b \in (a, a+1)$, where a is an integer, we assume that the nearest integer point from the minimum point $Q_0(\underbrace{c/b, c/b, \ldots, c/b}_{b})$ stated in Lemma 1 is $Q_1(\underbrace{a, a, \ldots, a}_{n}, \underbrace{a+1, \cdots, a+1}_{b-n})$, which satisfies

$$na + (b-n)(a+1) = c.$$

Bringing the coordinate of the point Q_1 into $g(\mathbf{x})$, we have

$$g_{Q_1}(\mathbf{x}) = a^{2n(a-1)} \cdot (a+1)^{2a(b-n)}.$$

Suppose $Q_2(\underbrace{a, a, \ldots, a}_{n+1}, \underbrace{a+1, \ldots, a+1}_{b-n-2}, a+2)$. Denote the distance between two points Q_i and Q_j by $d_{Q_iQ_j}$. Then it is easy to see that $d_{Q_0Q_2} > d_{Q_0Q_1}$, and

$$g_{Q_2}(\mathbf{x}) = a^{2(n+1)(a-1)} \cdot (a+1)^{2a(b-n-2)} \cdot (a+2)^{2(a+1)}.$$

Then we have

$$\frac{g_{Q_2}(\mathbf{x})}{g_{Q_1}(\mathbf{x})} = \frac{a^{2(a-1)}(a+2)^{2(a+1)}}{(a+1)^{4a}} > 1.$$

Hence, $g_{Q_2}(\mathbf{x}) > g_{Q_1}(\mathbf{x})$.

Similarly, for $Q_3(a-1, \underbrace{a, \ldots, a}_{n-2}, \underbrace{a+1, a+1, \ldots, a+1}_{b-n+1})$, we have $d_{Q_0Q_3} > d_{Q_0Q_1}$ and $g_{Q_3}(\mathbf{x}) > g_{Q_1}(\mathbf{x})$.

The same conclusion can be obtained for

$Q_4(\underbrace{a, \ldots, a}_{n}, a-1, \underbrace{a+1, \ldots, a+1}_{b-n-2}, a+3)$ and $Q_5(a-2, \underbrace{a, \ldots, a}_{n-2}, a+2, \underbrace{a+1, \ldots, a+1}_{b-n})$.

By induction, we conclude that $g_{Q_1}(\mathbf{x})$ is the minimum value of $g(\mathbf{x})$, which completes the proof. □

7.3 Proof of Theorem 3

The main idea is generally similar to that of Theorem 2, but is more challenging. As N goes to infinity, we wish to establish the effectiveness of the ITE estimator in the most extreme case, thus the performance is guaranteed in all cases.

For the reason that the volume of the j-th subtree, V_j, is uncertain for each potential source node, it is hard to directly analyze the equivalent ITE $\mathbb{H}(v, G_N)$ in Equation (10). To this end, we first focus on V_j in geometric trees. According to the definition of geometric trees, the average number of nodes in the r-th level is that $\mathbb{E}[n(r)] = (b+c)r^{\alpha}/2$. Denote the average degree of the nodes in the r-th level as $d(r)$. Then we have

$$d(r) = \frac{\mathbb{E}[n(r+1)]}{\mathbb{E}[n(r)]} + 1 = \left(\frac{r+1}{r}\right)^{\alpha} + 1 \xrightarrow{r\to\infty} 2. \tag{22}$$

As a result, with the increase of the level r, different branches of the geometric tree will grow somewhat like a line. In such a way, for the observation time t large enough, V_j can be approximated by $2n_j$; hence, let $\mathbb{H}^*_g(v, G_N)$ be the exponential term of the equivalent ITE on geometric trees; we obtain that

$$\mathbb{H}^*_g(v, G_N) \xrightarrow{t\to\infty} \mathbb{H}^*_r(v, G_N). \tag{23}$$

Next we will guarantee the performance in an extreme scenario. Let $T_i(t)$ denote the size of the infected subtree rooted at the i-th neighbor of v^* at time t. As derived in Theorem 4 in [33], for all $1 \le i \le d_{v^*}$, and a small enough $\epsilon > 0$,

$$\frac{b}{1+\alpha}(t(1-\epsilon)-2)^{\alpha+1} \le T_i(t) \le \frac{c}{1+\alpha}(t(1+\epsilon))^{\alpha+1}.$$

In the most extreme case, $d_{v^*}-1$ of these subtrees have minimal infection size $T_{min}(t)$ and the remaining one has maximal size $T_{max}(t)$, rooted at v_1, w.l.o.g., where

$$\begin{aligned} T_{min}(t) &= \frac{b}{1+\alpha}(t(1-\epsilon)-2)^{\alpha+1}, \\ T_{max}(t) &= \frac{c}{1+\alpha}(t(1+\epsilon))^{\alpha+1}. \end{aligned}$$

Based on Lemma 2, to ensure that $\mathbb{H}^*_g(v_1, G_N)$ is as small as possible, the infection coordinate $C(v_1)$ should satisfy

$$C(v_1) = \left\lceil (d_{v^*}-1)T_{min}(t) + 1, \underbrace{\frac{T_{max}(t)}{c}, \ldots, \frac{T_{max}(t)}{c}}_{c} \right\rceil.$$

Defining $\epsilon = t^{-\frac{1}{2}} + \delta$ for any small $0 < \delta < 1$, we have

$$\begin{aligned} \frac{\mathbb{H}^*_g(v^*, G_N)}{\mathbb{H}^*_g(v_1, G_N)} &= \frac{T_{min}(t)^{(2T_{min}(t)-2)(d_{v^*}-1)} \cdot T_{max}(t)^{2T_{max}(t)-2}}{[(d_{v^*}-1)T_{min}(t)+1]^{2(d_{v^*}-1)T_{min}(t)} \left(\frac{T_{max}(t)}{c}\right)^{2T_{max}(t)-2c}} \\ &= \frac{\left[\frac{b}{1+\alpha}(t-t^{\frac{1}{2}+\delta}-2)^{\alpha+1}\right]^{\left[\frac{2b}{1+\alpha}(t-t^{\frac{1}{2}+\delta}-2)^{\alpha+1}-2\right](d_{v^*}-1)}}{\left[\frac{(d_{v^*}-1)b}{1+\alpha}(t-t^{\frac{1}{2}+\delta}-2)^{\alpha+1}+1\right]^{\left[\frac{2(d_{v^*}-1)b}{1+\alpha}(t-t^{\frac{1}{2}+\delta}-2)^{\alpha+1}\right]}} \end{aligned}$$

$$
\begin{aligned}
&\cdot \frac{[\frac{c}{1+\alpha}(t+t^{\frac{1}{2}+\delta})^{\alpha+1}]^{[\frac{2c}{1+\alpha}(t+t^{\frac{1}{2}+\delta})^{\alpha+1}-2]}}{[\frac{1}{1+\alpha}(t+t^{\frac{1}{2}+\delta})^{\alpha+1}]^{[\frac{2c}{1+\alpha}(t+t^{\frac{1}{2}+\delta})^{\alpha+1}-2c]}} \\
&= [\frac{1}{d_{v^*}-1+\frac{1+\alpha}{b}(\frac{1}{t-t^{\frac{1}{2}+\delta}-2})^{\alpha+1}}]^{[\frac{2(d_{v^*}-1)b}{1+\alpha}(t-t^{\frac{1}{2}+\delta}-2)^{\alpha+1}]} \cdot c^{[\frac{2c}{1+\alpha}(t+t^{\frac{1}{2}+\delta})^{\alpha+1}]} \cdot \Theta(t^m) \\
&\overset{(j)}{=} [\frac{c^c}{(d_{v^*}-1)^{(d_{v^*}-1)b}}]^{\Theta(t^{\alpha+1})} \cdot \Theta(t^m) \\
&< 1, \qquad (24)
\end{aligned}
$$

for t large enough, where $m = 2(c - d_{v^*})(\alpha + 1)$ is a finite constant. Since by assumption $c^c < (d_{v*} - 1)^{(d_{v*}-1)b}$, as $t \to \infty$, then N tends to infinity, the first term in the right-hand side of the equation (j) will be far less than 1, and its exponential growth rate is much greater than that of the second term. As a result, we obtain that $\mathbb{H}^*_g(v^*, G_N) < \mathbb{H}^*_g(v_1, G_N)$ for t large enough. In a similar way to Theorem 2, we conclude that $\mathbb{H}^*_g(v^*, G_N)$ is the minimum among all the infected nodes. Therefore, if the performance of the ITE estimator is guaranteed in the most extreme scenario, then it will correctly find the source in all scenarios as N goes to infinity, such that

$$\lim_{N\to\infty} \mathbf{P}_c = 1.$$

This completes the proof of Theorem 3.

7.4 Proof of Theorem 4

We assume that the source node v^* first infects its neighbor v_i $(1 \le i \le d_{v^*})$ with degree d_{v_i}. For the memoryless property of exponential distribution, the spreading is then divided into two processes: (a) τ_1: starting from v^* and away from v_i with rate $(d_{v^*} - 1)\lambda$, and (b) τ_2: starting from v_i away from v^* with rate $(d_{v_i} - 1)\lambda$. Based on the definition of geometric trees, we obtain the expectation of the degree of v_i as follows.

$$\mathbb{E}(d_{v_i}) = \frac{(b + c)}{2} + 1. \qquad (25)$$

From Equation (25), we can see that if we fix the parameter b, then $\mathbb{E}(d_{v_i}) \propto c$. This indicates that the spreading rate of τ_2 will be higher with the increase of c, hence the information will be inclined to spread to the neighbors of v_i. As a result, v_i will have a larger infected degree.

Recall that the first term in Equation (10), $V^{2d_{v(inf)}}$, grows with the power of $d_{v(inf)}$, while the second term grows exponentially. Unlike the limiting case when the infection size goes to infinity in Theorem 3, when the infection size is only moderate, however, we cannot overlook the difference of $V^{2d_{v(inf)}}$ for any $v \in G_N$. Furthermore, for c/b large enough, then w.h.p. we have that

$$V^{2d_{v^*(inf)}} \ll V^{2d_{v_i(inf)}}. \qquad (26)$$

In addition, as mentioned in Section 7.2, the second term is highly related to the variance of each subtree's size, hence it characterizes the structural centrality of the potential source in a way. In this case, the source v^*, which, second to the node v_i, will have more balanced sizes of subtrees compared to those of other remaining infected nodes due to the spreading property.

Combining the above two factors, as the ratio of c/b increases, we will obtain that $\mathbb{H}(v^*, G_N) < \mathbb{H}(v_i, G_N)$ with higher probability and the source v^* will have the minimum ITE in G_N.

On the other hand, owing to the large infected degree of v_i, most centrality based estimators will probably be fooled to choose v_i as the source. By contrast, the ITE estimator will not be completely

dictated by the centrality of each potential source as mentioned, and will correctly find the source with higher probability. This derives a more reliable detection.

8 CONCLUSION

In this article, we propose a structural entropy based approach named ITE estimator for source detection under the SI model. Theoretically, we prove that on geometric trees, the ITE estimator not only has an asymptotically complete detection, but also remarkably yields more reliable detection under moderate infection sizes with the increasing tree heterogeneity, which has important practical significance. In addition, a non-trivial detection is guaranteed as the network grows to infinity on regular expanding trees. To improve the efficiency, we propose a message-passing algorithm with a complexity of $O(N + |E_I|)$, faster than most prior arts. By incorporating the BFS strategy on general graphs, extensive experiments with different metrics show that the ITE estimator outperforms other baselines on both synthetic and real-world networks.

REFERENCES

[1] Ameya Agaskar and Yue M. Lu. 2013. A fast Monte Carlo algorithm for source localization on graphs. In *Proceeding of the Wavelets and Sparsity XV*, Vol. 8858, 429–434.

[2] Kartik Anand and Ginestra Bianconi. 2009. Entropy measures for networks: Toward an information theory of complex topologies. *Physical Review E* 80, 4 (2009), 045102.

[3] Albert-László Barabási and Réka Albert. 1999. Emergence of scaling in random networks. *Science* 286, 5439 (1999), 509–512.

[4] Ginestra Bianconi. 2009. Entropy of network ensembles. *Physical Review E* 79, 3 (2009), 036114.

[5] D. Bonchev and N. Trinajstić. 1977. Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics* 67, 10 (1977), 4517–4533.

[6] Samuel L. Braunstein, Sibasish Ghosh, and Simone Severini. 2006. The Laplacian of a graph as a density matrix: A basic combinatorial approach to separability of mixed states. *Annals of Combinatorics* 10, 3 (2006), 291–317.

[7] Yun Chai, Youguo Wang, and Liang Zhu. 2021. Information sources estimation in time-varying networks. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2621–2636.

[8] Jaeyoung Choi, Sangwoo Moon, Jiin Woo, Kyunghwan Son, Jinwoo Shin, and Yung Yi. 2017. Rumor source detection under querying with untruthful answers. In *Proceedings of the IEEE INFOCOM*. 1–9.

[9] Jaeyoung Choi and Yung Yi. 2018. Necessary and sufficient budgets in information source finding with querying: Adaptivity gap. In *Proceedings of the IEEE ISIT*. 2261–2265.

[10] Yongwook Choi and Wojciech Szpankowski. 2009. Compression of graphical structures. In *Proceedings of the IEEE ISIT*. 364–368.

[11] Cesar Henrique Comin and Luciano da Fontoura Costa. 2011. Identifying the starting point of a spreading process in complex networks. *Physical Review E* 84, 5 (2011), 056105.

[12] Matthias Dehmer. 2008. Information processing in complex networks: Graph entropy and information functionals. *Applied Mathematics and Computation* 201, 1-2 (2008), 82–94.

[13] Wenxiang Dong, Wenyi Zhang, and Chee Wei Tan. 2013. Rooting out the rumor culprit from suspects. In *Proceedings of the IEEE ISIT*. 2671–2675.

[14] Vincenzo Fioriti and Marta Chinnici. 2012. Predicting the sources of an outbreak with a spectral technique. arXiv:1211.2333.

[15] Jacob Goldenberg, Barak Libai, and Eitan Muller. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters* 12, 3 (2001), 211–223.

[16] Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou. 2015. K-center: An approach on the multi-source identification of information diffusion. *IEEE Transactions on Information Forensics and Security* 10, 12 (2015), 2616–2626.

[17] Nikhil Karamchandani and Massimo Franceschetti. 2013. Rumor source detection under probabilistic sampling. In *Proceedings of the IEEE ISIT*. 2184–2188.

[18] Jérôme Kunegis. 2013. Konect: The koblenz network collection. In *Proceedings of the WWW*. 1343–1350.

[19] Theodoros Lappas, Evimaria Terzi, Dimitrios Gunopulos, and Heikki Mannila. 2010. Finding effectors in social networks. In *Proceedings of the ACM SIGKDD*. 1059–1068.

[20] Jure Leskovec and Andrej Krevl. 2014. *SNAP Datasets: Stanford Large Network Dataset Collection.* (June 2014). Retrieved May 10, 2021 from http://snap.stanford.edu/data.

[21] Angsheng Li and Yicheng Pan. 2016. Structural information and dynamical complexity of networks. *IEEE Transactions on Information Theory* 62, 6 (2016), 3290–3339.
[22] Yiwei Liu, Jiamou Liu, Zijian Zhang, Liehuang Zhu, and Angsheng Li. 2019. REM: From structural entropy to community structure deception. In *Advances in Neural Information Processing Systems*. 12938–12948.
[23] Andrey Y. Lokhov, Marc Mézard, Hiroki Ohta, and Lenka Zdeborová. 2014. Inferring the origin of an epidemic with a dynamic message-passing algorithm. *Physical Review E* 90, 1 (2014), 012801.
[24] Wuqiong Luo and Wee Peng Tay. 2013. Estimating infection sources in a network with incomplete observations. In *Proceedings of the IEEE GlobalSIP*. 301–304.
[25] Wuqiong Luo and Wee Peng Tay. 2013. Finding an infection source under the SIS model. In *Proceedings of the IEEE ICASSP*. 2930–2934.
[26] Wuqiong Luo, Wee Peng Tay, and Mei Leng. 2013. Identifying infection sources and regions in large networks. *IEEE Transactions on Signal Processing* 61, 11 (2013), 2850–2865.
[27] Pedro C. Pinto, Patrick Thiran, and Martin Vetterli. 2012. Locating the source of diffusion in large-scale networks. *Physical Review Letters* 109, 6 (2012), 068702.
[28] B. Aditya Prakash, Jilles Vreeken, and Christos Faloutsos. 2012. Spotting culprits in epidemics: How many and which ones?. In *Proceedings of the ICDM*. 11–20.
[29] Nicolas Rashevsky. 1955. Life, information theory, and topology. *Bulletin of Mathematical Biophysics* 17, 3 (1955), 229–235.
[30] C. Raychaudhury, S. K. Ray, J. J. Ghosh, A. B. Roy, and S. C. Basak. 1984. Discrimination of isomeric structures using information theoretic topological indices. *Journal of Computational Chemistry* 5, 6 (1984), 581–588.
[31] Martin Rosvall and Carl T. Bergstrom. 2008. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America* 105, 4 (2008), 1118–1123.
[32] Devavrat Shah and Tauhid Zaman. 2011. Rumors in a network: Who's the culprit? *IEEE Transactions on Information Theory* 57, 8 (2011), 5163–5181.
[33] Devavrat Shah and Tauhid Zaman. 2016. Finding rumor sources on random trees. *Operations Research* 64, 3 (2016), 736–755.
[34] Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27, 3 (1948), 379–423.
[35] Wenchang Tang, Feng Ji, and Wee Peng Tay. 2018. Estimating infection sources in networks using partial timestamps. *IEEE Transactions on Information Forensics and Security* 13, 12 (2018), 3035–3049.
[36] Zhaoxu Wang, Wenxiang Dong, Wenyi Zhang, and Chee Wei Tan. 2014. Rumor source detection with multiple observations: Fundamental limits and algorithms. In *Proceedings of the ACM SIGMETRICS*.
[37] Duncan J. Watts and Steven H. Strogatz. 1998. Collective dynamics of "small-world" networks. *Nature* 393, 6684 (1998), 440–442.
[38] Pei-Duo Yu, Chee Wei Tan, and Hung-Lin Fu. 2018. Rumor source detection in finite graphs with boundary effects by message-passing algorithms. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 175–192.
[39] Kai Zhu, Zhen Chen, and Lei Ying. 2016. Locating the contagion source in networks with partial timestamps. *Data Mining and Knowledge Discovery* 30, 5 (2016), 1217–1248.
[40] Kai Zhu and Lei Ying. 2014. Information source detection in the SIR model: A sample-path-based approach. *IEEE/ACM Transactions on Networking* 24, 1 (2014), 408–421.
[41] Kai Zhu and Lei Ying. 2015. Source localization in networks: Trees and beyond. arXiv:1510.01814.

Received 12 September 2021; revised 16 August 2022; accepted 9 October 2022