ITE: A Structural Entropy Based Approach for Source Detection

Chong Zhang, Qiang Guo, Luoyi Fu, Xiaoying Gan and Xinbing Wang
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University, Shanghai, China
Email: {zhangchong18, johnnyguo, yiluofu, ganxiaoying, xwang8}@sjtu.edu.cn

***Abstract*—This paper studies the problem of source detection, which is to infer the source node out of an aftermath of a cascade, i.e., 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 paper, we propose a new metric which 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 [1], 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. 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. Experiments on various network topologies confirm the superiority of ITE to the baselines. For example, ITE returns an accuracy of 75% ranking the source among top 5%, far exceeding 45% of the classic algorithms on scale-free networks.**

I. Introduction

The ubiquity of many types of online/offline and social/physical networks has fundamentally changed the landscapes of information spreading, nowadays. Unfortunately, the same channels can be utilized to amplify isolated risks such as rumors, malware or an isolated failure in a power grid network that cause pernicious effects. Therefore, inferring the initiator 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 [2], which, along with numerous following efforts [3]–[10], investigates the problem in a common paradigm: given an observation $\mathcal{O}$ 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 $\mathbf{P}(\mathcal{O}|\hat{v})$. 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* [3], where it is simply believed that the source node is the one surrounded by the most infected neighbors, (ii) *distance* [4]–[6], that select the potential source based on the minimum infection eccentricity, or (iii) *infection size* [2], [7]–[10], 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 [11]–[17] 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. Our design of the new metric is mainly inspired by the structural entropy [1], 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) between nodes uniformly at random. In this manner, the structural entropy provably [1] 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 Section II-C). 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, 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 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) And 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 (Section IV and V), bring about improved source detection performance. Then, we extend this framework in general graphs by a BFS heuristic. To the best of our knowledge, we are the first to apply structural entropy in this problem. We summarize our main contributions 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. In contrast, for regular expanding trees, the ITE estimator can still guarantee a non-trivial detection even when the infection size goes to infinity.
- Numerical results on both synthetic and real-world networks confirm the superiority of ITE to other different source estimators. For example, the 5% accuracy of ITE is close to 75%, which is significantly higher than that of other algorithms on a scale-free network.

Related works. It is known that source detection problem is highly challenging. In the seminal work [2], 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 susceptible-infected (SI) model. This work was extended in [18] for random trees, where the detection probability was quantified. Later, the rumor centrality has been further studied under different models or assumptions [7]–[10].

Besides the rumor centrality, several other algorithms based on a single snapshot of the network have been proposed. Zhu and Ying [4] proposed a sample path based approach to detect the single source under susceptible-infected-recovered (SIR) model, while a message passing algorithm was proposed under the same scenario by Lokhov et al. [19]. In [20], Lappas et al. analyzed the detection problem under the independent cascade (IC) model [21] by minimizing the distance between the expected states and the observed states of the nodes. Prakash et al. [22] 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 [23] 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 which utilized side information for source detection problem, such as timestamps of the infected nodes [11]–[15], or directions from which a node gets infected [16], [17]. All these methods are unable to exploit the structural characteristics as much as possible.

The measures of graph entropy have been extensively studied in [24]–[27], where each of them is a specific form of the Shannon entropy [28] for different types of distributions. While Li and Pan [1] proposed the structure entropy minimization principle to detect the natural structure in a network. Our proposed metric is based on this idea.

Organization. The rest of this paper is organized as follows. We introduce the ITE estimator in Section II. For tree-type networks, we propose an efficient algorithms for its evaluation in Section III. Section IV summarizes the main theoretical results. The simulation based performance evaluations will be presented in Section V, and all the proofs are provided in Section VI. We conclude the paper in Section VII.

II. Information Source Estimator

A. 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 paper, we limit our attention to the case where there is only one source node v^*.

We use the *susceptible-infected* (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 $\lambda = 1$.

B. 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

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

C. 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 I, structural entropy [1] 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 substructures in a complex network $\mathcal{G}$, such as various modules, components or communities, which form a partition $\mathcal{P}$ of the vertices. To characterize the structural information contained in $\mathcal{G}$ relative to $\mathcal{P}$, it makes sense to inquire the information content of the substructure $\mathcal{P}$ in $\mathcal{G}$. Imagine that messages can be delivered between nodes through edges. A *call* is a flow of message from a sender m to a receiver n, where $\{m, n\} \in \mathcal{E}$, and an exogenous process is launched to continuously collect 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. The authors [1] focused on the encoding of the network based on this probability distribution, committed to distinguishing the order from disorder in a noisy structure and detecting 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. Critically, 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. 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.

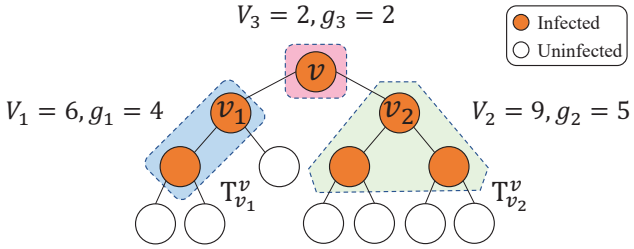


Fig. 1. Illustration of Partition of the Infected Tree.

D. 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 the 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, 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.*
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 Fig. 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 $v \in G_N$, $\mathcal{H}(v, G_N)$, is defined by the structural information of G_N relative to $\mathcal{P}_v$, that is*

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

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 Fig. 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 $T^v_{v_1}$ are 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 II-C, 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 constructed by $\mathcal{P}_v$ deviates from the actual infection process starting from the source, hence the probability that the node v is the actual source of the information will be higher. As such, we denote the source of our estimator by $\hat{v}$, then the ITE estimator can be formulated as:

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

with ties broken uniformly at random.

E. 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 breadth first search (BFS) heuristic to deduce a tree network in the infected graph G_N.

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

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.

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

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

III. 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 try to simplify the expression of $\mathcal{H}(v, G_N)$.

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

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

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

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*

$$\begin{aligned} &\sum_{j=1}^{d_{v_1(inf)}} g_j(v_1) - \sum_{j=1}^{d_{v_2(inf)}} g_j(v_2) \\ &= [d_{v_1(inf)} - d_{v_1(un)}] - [d_{v_2(inf)} - d_{v_2(un)}], \end{aligned} \quad (7)$$

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

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

$$\mathbb{H}(v, G_N) = V^{2d_{v(inf)}} \cdot \prod_{j=1}^{d_{v(inf)}} V_j^{2(n_j - 1)}. \quad (8)$$

Therefore, the ITE estimator is transformed into finding the potential source node $\hat{v}$ with the minimum value of $\mathbb{H}(\hat{v}, G_N)$, 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 a tuple to its parent node, denoted by $l^{up}_{u \to par(u)}$. The first item of the tuple is the size of u's subtree, and the second item is the corresponding volume. The parent node adds the $l^{up}_{u \to par(u)}$ tuples to obtain the size and volume of its own subtree. These tuples are then passed upward until the root node v^* receives all its children's tuples, by which it will calculate its equivalent ITE.

Meanwhile, adding all tuples of its children and the tuple of itself, $[1, deg(v^*)]$, the root node can obtain a global tuple L_{all} that records the size N and the volume V of G_N. With L_{all}, each infected node u will then obtain the tuple of its parent's subtree by $l^{up}_{u \to par(u)}$ subtracted from L_{all}, which we call $p_{par(u)}$ and helps to calculate their equivalent ITEs. 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 $f(x) = x[1]^{2(x[0]-1)}$.

IV. Main Results

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

A. 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*

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

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. This is a special case for regular trees, so we omit the proof of this theorem.

B. 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*

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

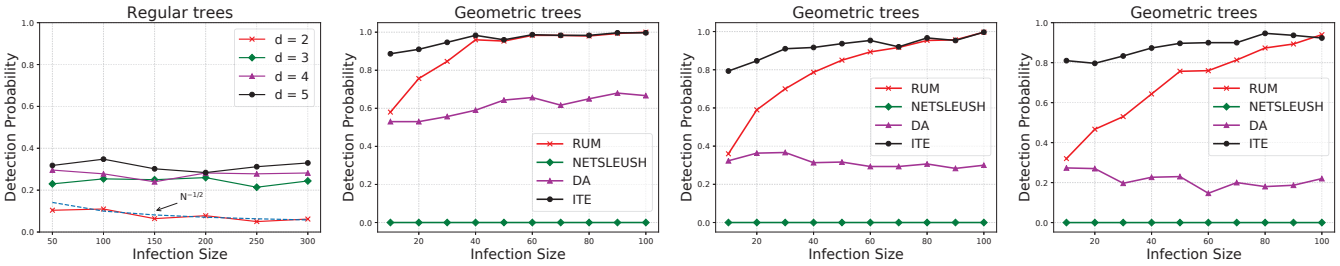
Due to the degree regularity and enough structural complexity in the network, our estimator could capture the structural features inside, and still perform the detection with a strictly positive probability even when the infection size N goes to infinity. Such is not the case for one randomly selecting an infected node. The above result also says that the detection probability is bounded by $1/2$. Therefore, the performance of ITE estimator is guaranteed on regular trees with $d \geq 3$. This theorem is proved in Section VI-A.

C. Advantages with Heterogeneity of Geometric Trees

Geometric trees are first introduced in [18], 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 for all $1 \leq i \leq d_{v^*}$

$$br^\alpha \leq n^i(r) \leq cr^\alpha. \quad (9)$$

The condition of (9) describes that each of the subtrees of the root node should satisfy polynomial growth with parameter $\alpha \geq 0$. The parameter α characterizes the growth of the geometric trees, while the ratio c/b describes the heterogeneity



(a) $d = 2, 3, 4, 5$ (b) $\alpha = 1, d_{v^*} = 12, b = 5, c = 25$ (c) $\alpha = 1, d_{v^*} = 12, b = 5, c = 50$ (d) $\alpha = 1, d_{v^*} = 12, b = 5, c = 75$

Fig. 2. Performance on tree networks.

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.

We consider the scenario where the information starts spreading from the root node of the geometric tree, and obtain the following result, which demonstrates a further advantage of our estimator with the increasing tree heterogeneity.

Theorem 3. *For a geometric tree with parameters $\alpha > 0$, $0 < b \leq c$, and the root node v^* with degree $d_{v^*} \geq 3$. Let α, b and d_{v^*} be fixed, then as the value of c (or to say c/b) increases, the ITE estimator will yield more reliable detection compared to those centrality based algorithms under only moderate infection sizes.*

Remark: A more reliable detection under moderate infection sizes means that our estimator has a better chance to 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 [2]–[4] will probably be fooled to select the nodes with large degrees. In contrast, as we will prove in Section VI-B, the ITE estimator will not be completely dictated by the centrality of the potential source.

V. Performance Evaluation

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

A. Baseline Algorithms

For comparison, we choose the algorithms proposed in 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 [2]. This classic algorithm is proved to be the maximum likelihood estimator on regular trees under the SI model.
- **DA:** Find the node with maximum dynamic age [23], which is defined as the absolute difference between the maximum eigenvalue of the adjacency matrix and the maximum eigenvalue of the adjacency matrix after the node is removed.
- **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 [22].

B. Tree Networks

We first provide simulation results on trees to corroborate the theoretical results in Section IV. 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.

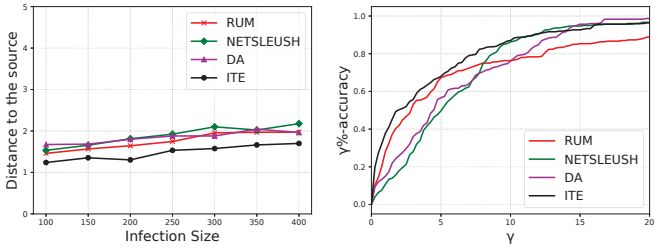
The detection probability of the ITE estimator versus the infection size on different trees is shown in Fig. 2. As can be seen, the detection rate scales as $N^{-1/2}$ as derived in Theorem 1 for line graphs. While for regular expanding trees with $d \geq 3$, the estimator has a non-trivial detection probability, which is less than $1/2$ and does not decay to 0 as predicted.

Figs. 2(b)-2(d) present the results on geometric trees under different settings of c/b, where we fix $\alpha = 1$, $b = 5$ and $d_{v^*} = 12$. We have the following two observations. Firstly, by comparing these three subgraphs, we can explicitly see that ITE is less affected with the increasing ratio of c/b. Secondly, the gap of detection probabilities between ITE and the other three algorithms becomes wider under the same infection size, thus, our estimator has more advantages when there exists more heterogeneity in geometric trees which is guaranteed by Theorem 3.

C. Graph Networks

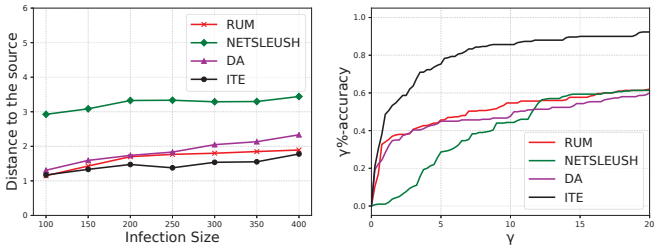
We next perform experiments on small-world networks [29], scale-free networks [30] and the US power grid (PG) networks [29]. The small-world network is generated by rewiring edges and contains 5000 nodes and 25000 edges, while the scale-free network is generated by preferential attachment with 5000 nodes and 9996 edges. The PG network is the electrical power grid of the western United States which contains 4941 nodes and 6594 edges. We vary the infection size from 100 to 400 and run each simulation 300 times. In each simulation, the source node is chosen uniformly across node degree to avoid the bias towards small degree nodes. We evaluate the performance of the algorithms with following metrics.

- Distance is the average number of hops from the estimated source to the actual source, which is an often used metric for source detection problem.
- $\gamma\%$-accuracy versus the rank percentage describes the probability that the actual source is ranked among top γ percent.



(a) Distance to the source (b) $\gamma\%$-accuracy

Fig. 3. Performance on the small-world network.



(a) Distance to the source (b) $\gamma\%$-accuracy

Fig. 4. Performance on the scale-free network.

For example, RUM ranks the nodes in an 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 actual source lies in the top ranked nodes with a high probability.

Figs. 3-5 show the performance on the three networks mentioned above respectively. For all the plots of $\gamma\%$-accuracy versus the rank percentage γ, we pick the infection size 400. From the perspective of distance, we observe that ITE performs better than the other three algorithms in almost all cases. The improvement is more obvious in small-world network than in scale-free network and PG network. For the small-world network used here, the average ratio of edges to nodes is 5, whereas for the scale-free network and PG network, the average ratio is 2 and 1.3 respectively. Thus, the small-world network is less tree-like. This may explain why ITE outperforms more apparently than the other three algorithms. From the perspective of $\gamma\%$-accuracy, ITE has similar or better performance compared to all other algorithms. Particularly on the scale-free network, the 5%-accuracy of ITE is 75%, which is significantly higher than that of other algorithms, e.g., 45% for RUM and DA, 29% for NETSLEUTH. The reason behind may be the existence of many large degree *hubs* in the scale-free network, then the network has more heterogeneity than the other two networks.

VI. Proofs

A. Proof of Theorem 2

Before giving the proof of the result on regular expanding trees, we first simplify the equivalent ITE in (8) by the regularity of the trees. On a regular tree with degree $d \geq 3$, it is easy to see that $V = d \cdot N$ and $V_j = d \cdot n_j$, then

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

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,

$$\mathbb{H}_r(v, G_N) = N^{2d_{v(inf)}} \cdot \prod_{j=1}^{d_{v(inf)}} n_j^{2(n_j-1)}. \quad (11)$$

To establish that on regular trees with $d \geq 3$, the probability of correct detection of the source using 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.

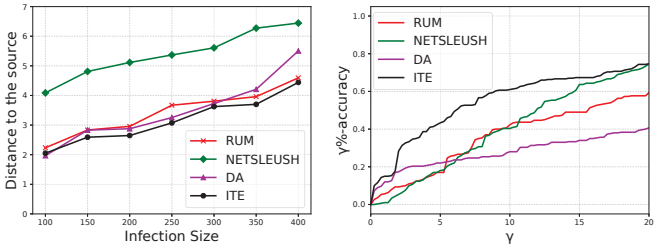
Define $S_n(t)$ as the event when all the d subtrees of the source have between n and $(d-1)n$ infected nodes. That is,

$$S_n(t) = \bigcap_{i=1}^{d} \{n \leq T_i(t) \leq (d-1)n\}, \quad \text{for } n > 0. \quad (12)$$

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 $d_{v(inf)}$, such that $1 \leq d_{v(inf)} \leq d$. In this case, the value of the equivalent ITE of each infected node v is mainly determined by the exponential term, denoted by

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

To begin with, we state the following Lemma which characterizes the general form of the function in (13) for the further analysis. We present its proof later in this section.



(a) Distance to the source (b) $\gamma\%$-accuracy

Fig. 5. Performance on the power grid network.

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

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

$$T_{v^*}^{v_d} = (d-1)n + \sum_{i=1}^{d-1} c_i + 1 > (d-1)n \ge T_d. \tag{14}$$

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

$$\frac{n + c_d - 1}{d-1} \le n - \frac{1}{d-1} < n. \tag{15}$$

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 (15), thus 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 $x_j's$ in Lemma 2 are all positive integers, and we present its proof later in this section. Further, the following Lemma states a property of the strictly convex function which 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 $x_1 \le x_2 \le \cdots \le x_k,\ b_i \ge 0$*, and* $\sum_{i=1}^{k-1} b_i = B$,

Combining (14) (15) and the 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 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 $\mathcal{C}(v)$, its Euclidean distance to Q_0 is a critical factor of $\mathbb{H}_r^*(v, G_N)$, which is also the variance of $\mathcal{C}(v)$. In other word, $\mathbb{H}_r^*(v, G_N)$ will be smaller with high probability when $\mathcal{C}(v)$ 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 $\mathcal{C}(v')$ 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 [2] 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 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 $h(\mathbf{x})$:

$$\mathbf{A} = \begin{bmatrix} \frac{2x_1+2}{{x_1}^2} & 0 & \cdots & 0 \\ 0 & \frac{2x_2+2}{{x_2}^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{2x_b+2}{{x_b}^2} \end{bmatrix}. \tag{16}$$

Since $x_j > 0$, for $j = 1, 2, \ldots, b$, the matrix $\mathbf{A}$ is positive-definite, 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,

$$\begin{cases} \dfrac{\partial z}{\partial x_1} = 2\ln x_1 + 2 - \dfrac{2}{x_1} - \lambda = 0, \\ \dfrac{\partial z}{\partial x_2} = 2\ln x_2 + 2 - \dfrac{2}{x_2} - \lambda = 0, \\ \quad \vdots \\ \dfrac{\partial z}{\partial x_b} = 2\ln x_b + 2 - \dfrac{2}{x_b} - \lambda = 0, \\ x_1 + x_2 + \cdots + x_b = c. \end{cases} \tag{17}$$

Denote by $p(x) = 2\ln x + 2 - \frac{2}{x}$, then $p'(x) = \frac{2x+2}{x^2} > 0$, so $p(x)$ is strictly increasing. As such, we observe the first b equations in (17). 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. □

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. □

B. Proof of Theorem 3

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.

$$\mathbb{E}(d_{v_i}) = \frac{(b+c)}{2} + 1. \tag{18}$$

From (18) 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 (8), $V^{2d_{v(inf)}}$, 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 2, 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

$$V^{2d_{v^*(inf)}} \ll V^{2d_{v_i(inf)}} \tag{19}$$

In addition, as mentioned in Section VI-A, 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 estimator 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.

VII. Conclusion

In this paper, 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 remarkably yields more reliable detection under moderate infection sizes with the increasing tree heterogeneity, which has important practical significance. Besides, 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, experiments with different metrics show that the ITE estimator outperforms other baseline algorithms on both synthetic and real-world networks.

Acknowledgment

This work was supported by National Key R&D Program of China 2018YFB2100302, and NSF China under Grant (No. 62020106005, 61822206, 61960206002, 61829201, 61832013, 62041205, 61532012).

REFERENCES

[1] A. Li and Y. Pan, "Structural information and dynamical complexity of networks," *IEEE Trans. on Information Theory*, vol. 62, no. 6, pp. 3290–3339, 2016.

[2] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?" *IEEE Trans. on information theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[3] C. H. Comin and L. da Fontoura Costa, "Identifying the starting point of a spreading process in complex networks," *Physical Review E*, vol. 84, no. 5, p. 056105, 2011.

[4] K. Zhu and L. Ying, "Information source detection in the sir model: A sample-path-based approach," *IEEE/ACM Trans. on Networking*, vol. 24, no. 1, pp. 408–421, 2014.

[5] W. Luo and W. P. Tay, "Estimating infection sources in a network with incomplete observations," in *Proc. IEEE GlobalSIP*, 2013, pp. 301–304.

[6] ——, "Finding an infection source under the sis model," in *Proc. IEEE ICASSP*, 2013, pp. 2930–2934.

[7] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Trans. on Signal Processing*, vol. 61, no. 11, pp. 2850–2865, 2013.

[8] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE ISIT*, 2013, pp. 2671–2675.

[9] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," in *Proc. ACM SIGMETRICS*, 2014.

[10] N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sampling," in *Proc. IEEE ISIT*, 2013, pp. 2184–2188.

[11] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Physical Review Letters*, vol. 109, no. 6, p. 068702, 2012.

[12] A. Agaskar and Y. M. Lu, "A fast monte carlo algorithm for source localization on graphs," in *SPIE Optical Engineering and Applications*, 2013.

[13] S. Zejnilović, J. Gomes, and B. Sinopoli, "Network observability and localization of the source of diffusion based on a subset of nodes," in *Proc. IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2013, pp. 847–852.

[14] K. Zhu, Z. Chen, and L. Ying, "Locating the contagion source in networks with partial timestamps," *Data Mining and Knowledge Discovery*, vol. 30, no. 5, pp. 1217–1248, 2016.

[15] W. Tang, F. Ji, and W. P. Tay, "Estimating infection sources in networks using partial timestamps," *IEEE Trans. on Information Forensics and Security*, vol. 13, no. 12, pp. 3035–3049, 2018.

[16] J. Choi, S. Moon, J. Woo, K. Son, J. Shin, and Y. Yi, "Rumor source detection under querying with untruthful answers," in *Proc. IEEE INFOCOM*, 2017, pp. 1–9.

[17] J. Choi and Y. Yi, "Necessary and sufficient budgets in information source finding with querying: Adaptivity gap," in *Proc. IEEE ISIT*, 2018, pp. 2261–2265.

[18] D. Shah and T. Zaman, "Finding rumor sources on random trees," *Operations Research*, vol. 64, no. 3, pp. 736–755, 2016.

[19] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Physical Review E*, vol. 90, no. 1, p. 012801, 2014.

[20] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *Proc. ACM SIGKDD*, 2010, pp. 1059–1068.

[21] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[22] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *Proc. ICDM*, 2012, pp. 11–20.

[23] V. Fioriti and M. Chinnici, "Predicting the sources of an outbreak with a spectral technique," *arXiv preprint arXiv:1211.2333*, 2012.

[24] S. L. Braunstein, S. Ghosh, and S. Severini, "The laplacian of a graph as a density matrix: a basic combinatorial approach to separability of mixed states," *Annals of Combinatorics*, vol. 10, no. 3, pp. 291–317, 2006.

[25] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *PNAS*, vol. 105, no. 4, pp. 1118–1123, 2008.

[26] M. Dehmer, "Information processing in complex networks: Graph entropy and information functionals," *Applied Mathematics and Computation*, vol. 201, no. 1-2, pp. 82–94, 2008.

[27] K. Anand and G. Bianconi, "Entropy measures for networks: Toward an information theory of complex topologies," *Physical Review E*, vol. 80, no. 4, p. 045102, 2009.

[28] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[29] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world'networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[30] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.