Information Source Detection in Networks: Possibility and Impossibility Results

Kai Zhu and Lei Ying

School of Electrical, Computer and Energy Engineering
Arizona State University
Tempe, AZ, United States, 85287
Email: kzhu17@asu.edu, lei.ying.2@asu.edu

***Abstract*—This paper studies information source detection in networks under the independent cascade (IC) model. Assume the spread of information starts from a single source in a network and a complete snapshot of the network is obtained at some time. The goal is to identify the source based on the observation. We derive the maximum a posterior (MAP) estimator of the source for tree networks and propose a Short-Fat Tree (SFT) algorithm for general networks based on the MAP estimator. The algorithm selects the Jordan infection center [1] and breaks ties according the degree of boundary infected nodes. Loosely speaking, the algorithm selects the node such that the breadth-first search (BFS) tree from it has the minimum depth but the maximum number of leaf nodes. On the Erdos-Renyi (ER) random graph, we establish the following possibility and impossibility results: (i) when the infection duration $< \frac{\log n}{(1+\alpha)\log \mu}$ for some $\alpha > 0.5$, SFT identifies the source with probability 1 (w.p.1) asymptotically (as network size increases to infinity), where n is the network size and μ is the average node degree; (ii) when the infection duration $> \lceil \frac{\log n}{\log \mu} \rceil + 2$, the probability of identifying the source approaches zero asymptotically under any algorithm; and (iii) when infection duration $< \frac{\log n}{(1+\alpha)\log \mu}$ for some $\alpha > 0$, asymptotically, at least $1-\delta$ fraction of the nodes on the BFS tree starting from the source are leaf-nodes, where $\delta = 3\sqrt{\frac{\log n}{\mu}}$, i.e., the BFS tree starting from the actual source is a fat tree. [1] Numerical experiments on tree networks, the ER random graphs and real world networks with different evaluation metrics show that the SFT algorithm outperforms existing algorithms.**

I. Introduction

The information source detection problem is to identify the source of information diffusion in networks based on available observations like the states of the nodes and the timestamps at which nodes adopted the information (or called infected). The solution of the problem can be used to answer a wide range of important questions. For example, in epidemiology, the knowledge of the epidemic source has been used to understand the transmission media of the disease [2]. For a computer virus spreading on the Internet, tracing the source helps locate the virus creator. For the news over the social media, locating the sources helps users verify the credibility of the news.

Because of its wide range of applications, the problem has gained a lot of attention in the last few years since the seminal work by Shah and Zaman [3]. A number of effective information source detection algorithms have been proposed under different diffusion models. Despite significant efforts and successes, theoretical guarantees have been established only for tree networks due to the complexity of the problem in non-tree networks. In this paper, we first develop a new information source detection algorithm, called the Short-Fat Tree algorithm, and then present a comprehensive performance analysis of the algorithm under the IC model for both tree networks and the ER random graph. To the best of our knowledge, SFT is the first algorithm that has provable performance guarantees on both tree networks and the ER random graph [4] (non-tree networks).

The fundamental possibility and impossibility results are summarized as follows.

1) For tree networks, we prove that the Jordan infection center with the maximum weighted boundary node degree (WBND) is the MAP estimator of the source under the heterogeneous IC model. Based on the derivation, we propose an algorithm called the Short-Fat Tree (SFT) algorithm which is applicable to both tree and general networks.
2) We analyze the performance of the SFT algorithm on the ER random graph. Under some mild conditions on the average node degree, we establish the following three results:
   (i) Assume the infection duration $< \frac{\log n}{(1+\alpha)\log \mu}$ for some $\alpha > 0.5$, SFT identifies the source with probability 1 (w.p.1) asymptotically (as network size increases to infinity).
   (ii) Assume the infection duration $\geq \left\lceil \frac{\log n}{\log \mu} \right\rceil + 2$, the probability of identifying the source approaches zero asymptotically under any information source detection algorithm, i.e., it is impossible to detect the source with non-zero probability.
   (iii) Assume the infection duration $< \frac{\log n}{(1+\alpha)\log \mu}$ for some $\alpha > 0$, asymptotically, at least $1-\delta$ fraction of the nodes on the BFS-tree starting from the source are leaf-nodes, where $\delta > 3\sqrt{\frac{\log n}{\mu}}$. This result does not provide any guarantee on the probability of correctly localizing the source, but states that the BFS-tree starting from the true source is a "fat" tree, which

[1] The results above hold under some other minor conditions, which will be presented in Section III.

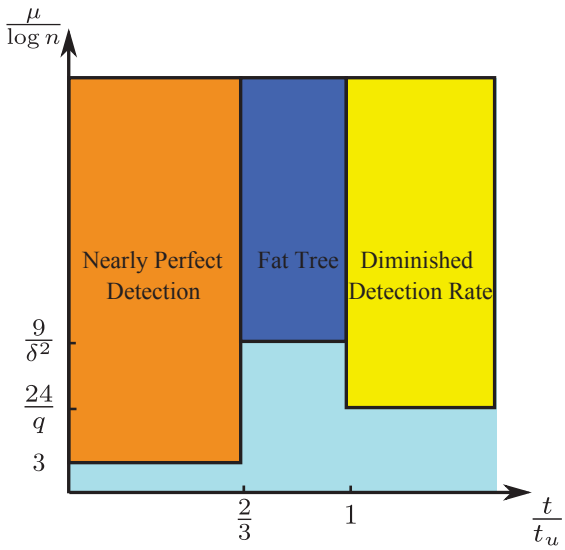


Fig. 1: Main Results Summary: q is the lower bound of the infection probability; there are $1-\delta$ fraction of nodes are boundary nodes on the BFS tree rooted at the source; $t_u = \left\lceil \frac{\log n}{\log \mu + \log q} \right\rceil + 2$ which is the lower bound of the observation time for the impossibility results in Theorem 5 (all algorithms fail when $t > t_u$.)

further justifies the SFT algorithm.

The results are summarized in Figure 1. We remark that results (i) and (iii) are highly nontrivial because a subgraph of the ER random graph is a tree with high probability only when the diameter is $\frac{\log n}{2 \log \mu}$, and (i) and (iii) deal with subgraphs that are not trees. To the best of our knowledge, these are the first theoretical results on information source detection on non-tree networks under probabilistic diffusion models.

3) One drawback of the WBND tie-breaking is that it requires the infection probabilities of all edges in the IC model. We simplify WBND to BND by using the boundary node degree in SFT. As shown in Section IV, the performance of BND tie-breaking is very close to WBND tie-breaking. We conducted extensive simulations on trees, ER random graphs and real world networks. SFT outperforms existing algorithms by having a higher detection rate and being closer to the actual source. We further evaluated the scalability of the algorithm by measuring the running time. Our results demonstrate that SFT achieves a better performance with a reasonably short execution time.

A. Related Work

[3] is one of the first papers that study the information source detection problem, in which a new graph centrality called rumor centrality was proposed and proved to be the maximum likelihood estimator (MLE) on regular trees under the susceptible-infected (SI) model. In addition, the detection probability (the probability that the estimator is the source) for regular trees was proved to be greater than zero and the detection probability for geometric trees approaches one asymptotically as the increase of the spreading time. Later, [5] quantified the detection probability of the rumor centrality on general random trees.

The rumor centrality has been further studied under different scenarios: 1) [6] extended the rumor centrality to multiple sources and showed that the detection probability goes to one as the number of infected nodes increases for geometric trees when there are at most two sources; 2) [7] proved a similar performance guarantee for the single source case when only a subset of infected nodes are observed; 3) [8] studied the detection probability when the prior knowledge of suspect nodes is available in the single source detection problem for trees; 4) [9] analyzed the detection probability of the rumor centrality for tree networks when there are multiple observations of independent diffusion processes from the same source.

[1] proposed the sample path based approach for the single source detection problem. Define the infection eccentricity of a node to be the maximum distance between the node and the infected nodes. [1] proved that on tree networks, under the homogeneous susceptible-infected-recovered (SIR) model, the root of the most likely sample path is a node with the minimum infection eccentricity (a Jordan infection center), which is within a constant distance to the actual source with a high probability. The approach has been extended to several directions: 1) [10] extended the approach to the case with partial observations and under the heterogeneous SIR model; 2) [11] extended the analysis to multiple sources under the SIR model and proved that the distance between the estimator and its closest actual source is bounded by a constant with a high probability in tree networks; 3) [12], [13] proved that the Jordan infection centers are the optimal sample path estimators under the SI model [12] and the susceptible-infected-susceptible (SIS) model [13] for tree networks, respectively.

Besides the rumor centrality and the Jordan infection center, several other heuristic algorithms based on a single snapshot of the network have been proposed in the literature: 1) [14] studied a similar problem under the independent cascade (IC) model [15] to minimize the l1 distance between the expected states and observed states of the nodes. A dynamic programming algorithm was proposed to solve the problem for tree networks and a Steiner tree heuristic was used for general networks; 2) [16] proposed an algorithm called NETSLEUTH which ranks the nodes according to an eigen vector based metric under the SI model. The algorithm was designed based on the Minimum Description Length principle; 3) [17] proposed a dynamic message passing algorithm based on the mean field approximation of the maximum likelihood estimation (MLE) of the source.

In addition, there exist several other algorithms which tackled the problem under the assumption that a subset of the infection timestamps are known: 1) [18] solved the MLE problem with partial timestamps for tree networks and extended the algorithm to general networks using a BFS tree heuristic; 2)

[19] proposed two rank based algorithms using a modified BFS tree heuristic for general graphs; 3) [20] proposed a simulation based Monte Carlo algorithm which utilizes the states of the sparsely placed observers within a fixed time window; 4) [21] obtained sufficient conditions on the number of timestamps needed to locate the source correctly under the deterministic slotted SI models.

Our paper establishes possibility and impossibility results of SFT beyond tree networks, which differs it from the existing work mentioned above. The rest of the paper is organized as follows. In Section II-A, we first introduce the IC model and formulate a MAP problem for information source detection and SFT will be presented in Section II-B. Section III summarizes the main theoretical results of the paper including the analysis on both tree networks and the ER random graph. The simulation based performance evaluation will be presented in Section IV. Due to space limitation, all the proofs are provided in our technical report [22].

II. Model and Algorithm

A. Model

Given an undirected graph g, denote by $\mathcal{E}(g)$ the set of edges in g and denote by $\mathcal{V}(g)$ the set of nodes in g. We consider the IC model [15] for information diffusion and assume a time-slotted system. Each node has two possible states: active (or called infected) and inactive (or called susceptible). At time slot 0, all nodes are inactive except the source. At the beginning of each time slot, an active node attempts to activate its inactive neighbors. If an attempt is successful, the corresponding node becomes active at next time slot; otherwise, the node remains inactive. The weight of each edge represents the success probability of the attempt, called the *infection probability* of the edge and each attempt is independent of others. Each active node only attempts to activate each of its inactive neighbors once. Denote by q_{uv} the infection probability of edge (u, v) and we assume $q_{uv} = q_{vu}$ throughout the paper since the graph is undirected. We assume that a complete snapshot $\mathcal{O} = \{\mathcal{I}, \mathcal{H}\}$ of the network at time t (called the *observation time*) is given, where $\mathcal{I}$ is the set of active nodes and $\mathcal{H}$ is the set of inactive nodes. Based on $\mathcal{O}$, we want to detect the source. We further assume the observation time t is unknown. The problem can be formulated as a MAP problem as follows,

$$\arg \max_{v \in \mathcal{V}(g)} \Pr(v|\mathcal{O}).$$

where $\Pr(v|\mathcal{O})$ is the probability that v is the source given the snapshot $\mathcal{O}$. The infected nodes form a connected component under the IC model, called the *infection subgraph* and denoted by g_i. Since the source must be an infected node, the MAP problem can be simplified to

$$\arg \max_{v \in \mathcal{I}} \Pr(v|\mathcal{O}),$$

and the search of the information source can be restricted to the infection subgraph. We assume the observation time t, which itself is a random variable, is independent of the source node.

B. The Short-Fat Tree Algorithm

In this section, we first present the SFT algorithm. We will show in Theorem 2 that the algorithm outputs the MAP estimator for tree networks, which motivates the algorithm. The performance on the ER random graph is studied in Theorems 3 and 4.

We first introduce several necessary definitions. Denote by d^g_{uv} the distance from node u to node v in graph g, where the distance is the minimum number of hops between two nodes. Define the *infection eccentricity* of an infected node to be the maximum distance from the node to all infected nodes on the infection subgraph g_i, denote by $e(v, \mathcal{I})$,

$$e(v, \mathcal{I}) = \max_{u \in \mathcal{I}} d^{g_i}_{uv}.$$

Recall that the *Jordan infection centers* of a graph are the nodes with the minimum infection eccentricity [1].

Consider a BFS tree T_v rooted at node v on the infection subgraph g_i. Denote by $\text{par}_v(u)$ the parent of node u in T_v. Define the set of *boundary nodes* of T_v to be

$$\mathcal{B}(v, \mathcal{I}) = \{w \in \mathcal{I} | d^{T_v}_{vw} = e(v, \mathcal{I})\},$$

which are the set of active nodes furthest away from node v in the infection subgraph.

The weighted boundary node degree (WBND) with respect to node v is defined to be

$$\sum_{(u,w) \in \mathcal{F}'_v} |\log(1 - q_{uw})|, \tag{1}$$

where

$$\mathcal{F}'_v = \{(u, w) | (u, w) \in \mathcal{E}(g), w \neq \text{par}_v(u), u \in \mathcal{B}(v, \mathcal{I})\}. \tag{2}$$

The SFT algorithm, presented in Algorithm 1, identifies the source based on the BFS trees on the infection subgraph. The algorithm is called the *Short-Fat Tree* algorithm because (1) it first identifies the *shortest* BFS tree; and (2) the shortest BFS tree that maximizes the WBND is then selected in tie-breaking, which is usually the tree with a large number of leaf-nodes, i.e., a *fat* tree. The pseudo codes of the algorithms are presented in Algorithm 1 and 2, which can be executed in a parallel fashion.

A simple example is presented in Figure 2 to illustrate algorithm. Each node has a unique node ID. The red nodes are infected and the white nodes are healthy. For simplicity, we assume the weights of all edges equal to $|\log(0.5)|$. The vector next to each infected node records the distance from it to all infected nodes. Initially at Iteration 0, each infected node only knows the distance to itself. For example, $[0 * * *]$ next to node 1 means that the distance from node 1 to itself is 0 and the distance from node 1 to node 2 is unknown. At Iteration 1, each infected node broadcasts its ID to its neighbors. Upon receiving the node ID from node 1, node 2 updates its vector to $[1\ 0 * *]$, and broadcasts node 1's ID to its neighbors in next iteration. The figure in the middle shows the updated vectors after all node ID exchanges occur at Iteration 1. At Iteration

2, node 1 and 2 do not receive any new node IDs. Therefore, node 1 and node 2 report themselves as the Jordan infection centers which are circled with blue in Figure 2. The boundary nodes of the BFS tree rooted at node 1 are 2,3,4. The WBND of node 1 is $13|\log(0.5)|$. Similarly, the boundary nodes of the BFS tree rooted at node 2 are 1,3,4 and the WBND is $9|\log(0.5)|$. Therefore, node 1 has a larger WBND and is chosen to be our estimator of the information source.

Algorithm 1: The Short-Fat Tree Algorithm

Input: $\mathcal{I}, g$;
Output: $v^\dagger$ (the estimator of information source)
Set subgraph g_i to be a subgraph of g induced by node set $\mathcal{I}$.
for $v \in \mathcal{I}$ **do**
  Initialize an empty dictionary D_v associating with node v.
  Set $D_v[v] = 0$.
end
Each node receives its own node ID at time slot 0.
Set time slot $t = 1$.
do
  for $v \in \mathcal{I}$ **do**
    if *v received new node IDs in $t-1$ time slot, where "new" IDs means node v did not receive them before time slot $t-1$* **then**
      v broadcasts the new node IDs to its neighbors in g_i.
    end
  end
  for $v \in \mathcal{I}$ **do**
    if *v receives a new node ID u which is not in D_v.* **then**
      Set $D_v[u] = t$.
    end
  end
  $t = t + 1$.
while *No node receives $|\mathcal{I}|$ distinct node IDs*;
Set $\mathcal{S}$ to be the set of nodes who receive $|\mathcal{I}|$ distinct node IDs.
for $v \in \mathcal{S}$ **do**
  Compute WBND of T_v using Algorithm 2.
end
return $v^\dagger \in \mathcal{S}$ with the maximum WBND.

Remark: Note Equation (1) requires the infection probabilities of all edges in the network which could be hard to obtain in practice. When the infection probabilities are not available, we can assume each edge has the same infection probability q and WBND becomes,

$$\left(\sum_{u\in\mathcal{B}(v,\mathcal{I})} \deg(u) - |\mathcal{B}(v,\mathcal{I})|\right)|\log(1-q)|,$$

where $\deg(u)$ is the degree of node u.

Algorithm 2: The WBND Algorithm

Input: v,D_v (Dictionary of distance from v to other nodes), g, $\mathcal{I}$, t;
Output: WBND(v)
Set $\mathcal{B}$ to be empty.
for *u in the keys of D_v* **do**
  if $D_v[u] = t$ **then**
    Add u to $\mathcal{B}$.
  end
end
Set $x = 0$;
for $w \in \mathcal{B}$ **do**
  Find the neighbor u of w such that $D_v[u] = t-1$.
  Set $x =$
  $x + \sum_{y\in\text{neighbors}(w)} |\log(1-q_{wy})| - |\log(1-q_{wu})|$.
end
return x.

Define the boundary node degree (BND) of node v to be

$$\sum_{u\in\mathcal{B}(v,\mathcal{I})} \deg(u) - |\mathcal{B}(v,\mathcal{I})| \tag{3}$$

which is only related to the degree of the boundary nodes and can be used to replace WBND as the tie-breaking among the Jordan infection center in SFT when the infection probabilities are unknown. As shown in Section IV, the performance using BND and WBND are similar. To differentiate the two algorithms, we call the algorithm which uses WBND as wSFT and the one which uses BND as SFT. Next, we analyze the complexity of the algorithm.

Theorem 1. *The worst case computational complexity of the SFT algorithm is $O(|\mathcal{I}|deg(\mathcal{I}))$ where $deg(\mathcal{I})$ is the total degree of nodes in $\mathcal{I}$ in graph g.*

The detailed proof can be found in Appendix A in our technical report [22].

III. Main Results

In this section, we summarize the main results of the paper and present the intuitions of the proofs.

A. Main Result 1 (The MAP Estimator on Tree Networks)

On tree networks, the Jordan infection center of the infection subgraph with the maximum WBND is a MAP estimator.

Theorem 2. *Consider a tree network. Assume the following conditions hold.*

- *The probability distribution of the observation time satisfies $\Pr(t) \geq \Pr(t+1)$ for all t.*
- *The source is uniformly and randomly selected, i.e., $\Pr(u) = \Pr(v)$.*

Denote by $\mathcal{J}$ the set of Jordan infection centers of the infection subgraph g_i. We have

$$\arg\max_{u\in\mathcal{J}} \sum_{(v,w)\in\mathcal{F}'_u} |\log(1-q_{vw})| \subset \arg\max_u \Pr(u|\mathcal{O}). \tag{4}$$

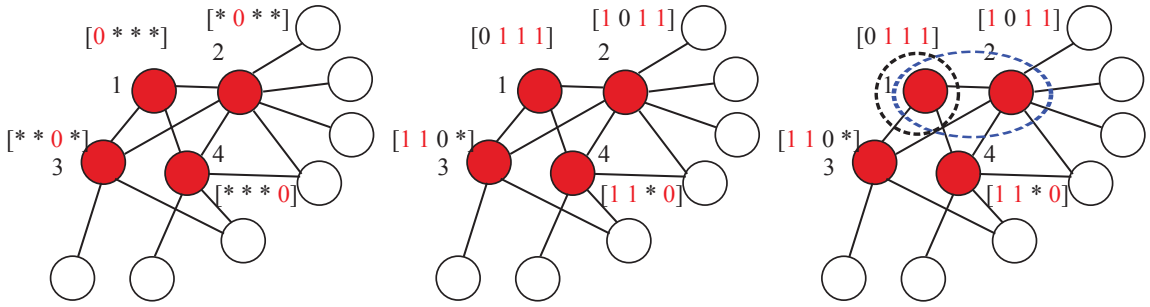


Fig. 2: An example of the Short-Fat Tree algorithm

where $\mathcal{F}'_u$ is defined in Equation (2).

The detailed proof can be found in Appendix B in our technical report [22]. The theorem has been proved in two steps: 1) We show that one of the Jordan infection centers maximizes the posterior probability on tree networks following similar arguments in [1]. In particular, for two neighboring nodes, we show the one with smaller infection eccentricity has a larger posterior probability of being the source. Since there exists a path from any node to a Jordan infection center on the infection subgraph, along which the infection eccentricity strictly decreases, we conclude that a MAP estimator of the source must be a Jordan infection center; 2) Consider the case where the tree network has more than one (at most two according to [23]) Jordan infection centers. When the observation time is larger than the infection eccentricity of the Jordan infection center, the probability of having the observed infected subgraph from any Jordan infection center is the same. When the observation time equals the infection eccentricity, we prove that the probability for a Jordan infection center to be the source is an increasing function of WBND of the BFS tree starting from it.

B. Main Result 2 (Detection with Probability One on the ER Random Graph)

Denote by n the number of nodes in the ER random graph and p the wiring probability of the ER random graph. Let $\mu = np$. Recall that t is the observation time. We show that the Jordan infection center is the actual source in the ER random graph with probability one asymptotically when $t < \frac{\log n}{(1+\alpha)\log \mu}$, which implies that SFT can locate the source w.p.1 asymptotically.

Theorem 3. *If the following conditions hold, source s is the only Jordan infection center on the infection subgraph with probability one asymptotically.*

- $\mu > 3\log n$.
- $t \leq \frac{\log n}{(1+\alpha)\log \mu}$, *for some* $\alpha \in (\frac{1}{2}, 1)$.

The proof is more than 10 pages long so omitted in this paper due to space constraint. We next present a brief overview of the proof and the details can be found in Appendix C in our technical report [22]. Note the infection eccentricity of the actual source is no larger than the observation time t. We show in the proof that the infection eccentricity of an infected node other than the source is larger than t. Consider the BFS tree $T^\dagger$ rooted at the actual source s. A node is said to be on level i if its distance to the source is i. Consider another infected node s'. Denote by $a(s')$ the ancestor of s' on level 1 of $T^\dagger$. As shown in Figure 3, the yellow area shows the level t infected nodes on subtree T_u^{-s}, which is the subtree of $T^\dagger$ rooted at node u, and the distance from s' to a node in the yellow area is larger than t if any path between the two nodes can only traverse the edges on tree $T^\dagger$. If s' has an infection eccentricity no larger than t, there must exist a path from s' to each node in the yellow area with length no larger than t. Such a path must contain edges that are not in $T^\dagger$ (we call these edges *collision edges*). We show in the proof that the number of nodes that are within t hops from s' via collision edges are strictly less than the number of nodes in the yellow area. Therefore, the infection eccentricity of s' must be larger than t, which implies that s is the only Jordan infection center.

Existing theoretical results in the literature on information source detection problems are only for tree networks. As shown in the proof of Theorem 3, the infection subgraph of the ER random graph is not a tree when $t > \frac{\log n}{2\log \mu}$. From the best of our knowledge, this result is the first one on non-tree networks.

C. Main Result 3 (The Fat Tree Result on the ER Random Graph)

Theorem 4. *If the following conditions hold,*

- $\mu > \frac{9}{\delta^2}\log n$.
- $t \leq \frac{\log n}{(1+\alpha)\log \mu}$, *for some* $\alpha \in (0, 1)$.

the leaf-nodes of the BFS tree starting from the actual source consists of at least $1-\delta$ fraction of the BFS tree asymptotically.

The detailed proof can be found in Appendix D in our technical report [22]. Consider the BFS tree from the source s in graph g. The boundary nodes are the nodes at level t and all boundary nodes must be infected at time t. If we ignore the presence of collision edges, the number of infected nodes roughly increases by a factor of $q\mu$ at each level where $q = \min_{(u,v)\in\mathcal{E}(g)} q_{uv}$. Due to this exponential growth nature,

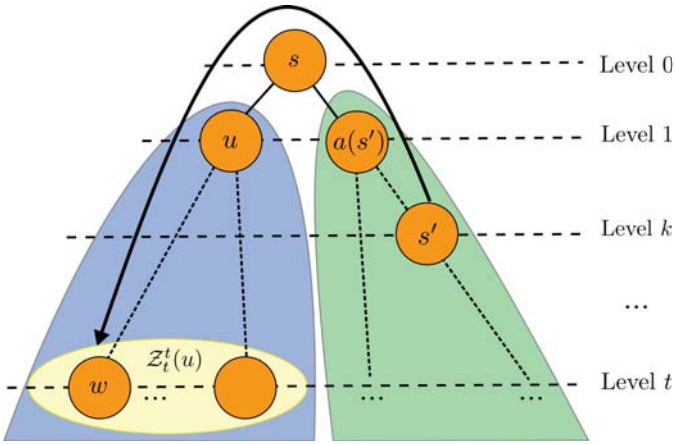


Fig. 3: A pictorial example of $\mathcal{Z}_t^t(u)$ in BFS tree $T^\dagger$

the total number of infected nodes is dominated by those infected at the last time slot. We show this property holds with the presence of collision edges. Theorem 4 suggests that the BFS tree rooted at the actual source is a "fat" tree and the BND of the actual source is large. Hence, in the tie breaking, the SFT algorithm has a good chance to select the actual source, which suggests that BND is a good tie breaking rule for the ER random graph.

D. Main Result 4 (The Impossibility Result on the ER Random Graph)

We next present the threshold of t after which it is impossible for any algorithm to find the actual source with a non-zero probability asymptotically. The result is based on the analysis of the diameter of an ER random graph in Theorem 4.2 in [24]. For clarity purpose, we rephrase that theorem with our notation in the following lemma.

Lemma 1. *If* $24\log n < np << \sqrt{n}$, *we have*

$$\lim_{n\to\infty} \Pr(\textit{Diameter}(g) \leq D+2) = 1,$$

where $D = \lceil \frac{\log n}{\log np} \rceil$.

We remark that in [24], the condition is $\log n << (n-1)p << \sqrt{n}$. We explicitly calculated the lower bound according to the proof in [24]. For the sake of completeness, we present the proof in Appendix E in our technical report [22].

Based on Lemma 1, we obtain the following impossibility result.

Theorem 5. *If* $24\log n < q\mu << \sqrt{n}$ *and* $q > 0$ *is a constant,*

$$\lim_{n\to\infty} \Pr(\mathcal{I} = \mathcal{V}(g)) = 1$$

when the observation time

$$t \geq \left\lceil \frac{\log n}{\log\mu + \log q} \right\rceil + 2 \triangleq t_u. \tag{5}$$

In other words the entire network is infected. In such a case, asymptotically, the probability of any node being the source is $1/n$.

The process to generate the ER random graph and the process of the information diffusion under the IC model can be viewed as a combined process. In this combined process, an edge exits only when the edge exists in the ER random graph and is live in the IC model. The detailed definition of the live edge could be found in Appendix B in our technical report [22]. Loosely speaking, an edge (u, v) is said to be live if node v is infected by node u under the IC model. When the observation time is larger than or equal to the diameter of the coupled ER random graph, all nodes in the network are infected. In such a case, the probability of a node being the source is $1/n$ as the source was uniformly chosen. Based on Lemma 1, the diameter of the combine network is smaller than $\lceil \frac{\log n}{\log q + \log\mu} \rceil + 2$ w.p.1 asymptotically.

Remark 1: We compare t_u in Equation (5) and the upper bound in Theorem 3. Since q is a constant, the ratio between t_u and the upper bound becomes $\frac{1}{1+\alpha}$ asymptotically. Since α can be arbitrarily close to $\frac{1}{2}$, the ratio becomes $\frac{2}{3}$. Therefore, the Jordan infection center is the actual source when the observation time is in the range of $(0, \frac{2}{3}t_u)$ and it is impossible to locate the source when the observation time is (t_u, ∞).

Remark 2: We compare t_u and the upper bound in Theorem 4 and asymptotically the ratio between t_u and the upper bound becomes $\frac{1}{1+\alpha}$ where $\alpha \in (0, 1)$. Since α can be arbitrarily close to 0 and the ratio are close to 1 which means the BFS tree from the source has large BND before it becomes impossible to locate the source. While the theorem does not provide any guarantee on the detection rate, it justifies the tie-breaking using BND and WBND.

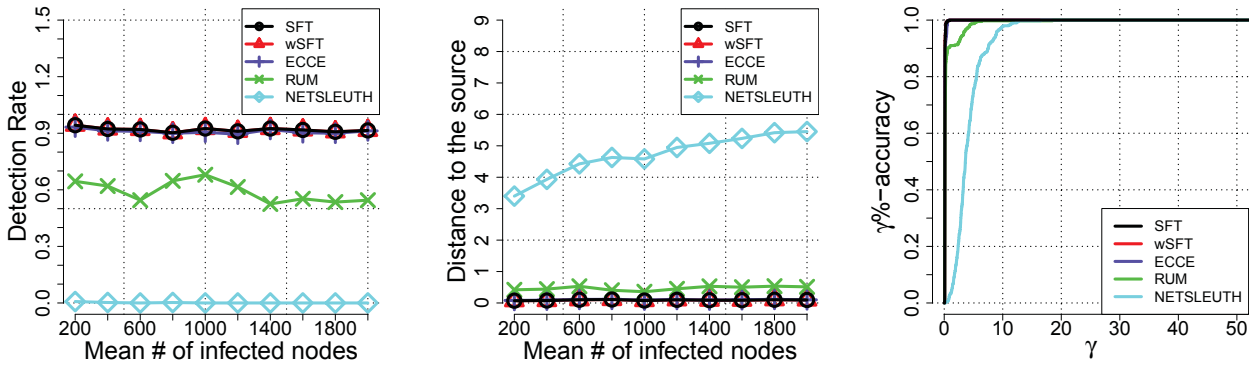
IV. Performance Evaluation

In this section, we compare the proposed algorithms with existing algorithms on different networks such as tree networks, the ER random graphs and real world networks.

A. Algorithms

Among all the existing algorithms discussed in Section I, we choose the algorithms which require only a single snapshot of the network but not the infection probabilities which could be difficult to obtain in practice. We compared SFT and wSFT with the algorithms summarized as follows.

- **ECCE:** Select the node with minimum infection eccentricity. Ties are breaking randomly. Recall the definition of the infection eccentricity is the maximum distance from the node to all infected nodes. [1] showed that the optimal sample path estimator on tree networks is the Jordan infection center of the graph under the SIR model.
- **RUM:** Select the node with maximum rumor centrality proposed in [3]. The rumor centrality was proved to be the maximum likelihood estimator on regular trees under the continuous time SI model in which the infection time follows exponential distribution.
- **NETSLEUTH:** Select the node with maximum value in the eigenvector corresponding to the largest eigenvalue of a submatrix which is constructed from the infected nodes



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

Fig. 4: Performance in the binomial trees

based on the graph Laplacian matrix. The algorithm was proposed in [16].

Among the selected algorithms, only wSFT requires the infection probabilities. We included wSFT to evaluate the importance of the knowledge of edge weights to our algorithm. We will see that the performance of SFT is almost identical to wSFT, so the infection probabilities are not important for our detection algorithm.

B. Evaluation Metrics

We evaluated the performance of the algorithms with three different metrics.

- Detection rate is the probability that the node identified by the algorithm is the actual source. A desired goal of the information source detection is to have a high detection rate.
- Distance is the number of hops from the source estimator to the actual source. The distance is an often used metric for information source detection.
- $\gamma\%$-accuracy is the probability with which the source is ranked among top γ percent. Note that besides providing a source estimator, an information source algorithm can also be used to rank the infected nodes according to their likelihood to be the source. For example, SFT can rank the nodes in an ascendant order according to their infection eccentricity and then breaks ties using BND. Other algorithms can be used to rank nodes as well. $\gamma\%$-accuracy is a less ambitious alternation to the detection rate. When the detection rates of all algorithms are low, it is useful to compare $\gamma\%$-accuracy as a high $\gamma\%$-accuracy guarantee that the actual source is among the top ranked nodes with a high probability.

C. Binomial Trees

In this section, we evaluate the algorithms on binomial trees. Denote by $\text{Bi}(m, \beta)$ the binomial distribution with m number of trials and each trial succeeds with probability β. A binomial tree is a tree where the number of children of each node follows a binomial distribution $\text{Bi}(m, \beta)$. In the experiments, we set $m = 20$ and $\beta = 0.5$. We adopted the IC model where the infection probability of each edge is assigned with a uniform distribution in $(0.2, 0.5)$. The lower bound on the infection probability is set to be 0.2 to prevent the diffusion process dies out quickly. We evaluated the performance for different infection size x. Under a discrete infection model, it is hard to obtain the diffusion snapshots with exact x infected nodes. Therefore, for each infection size x, we generate the diffusion samples where the number of infected nodes are in range $[0.75x, 1.25x]$. The source was chosen uniformly at random among all nodes in the network. We varied x from 200 to 2000 with a step size 200. For each infection size, we generate 400 diffusion samples.

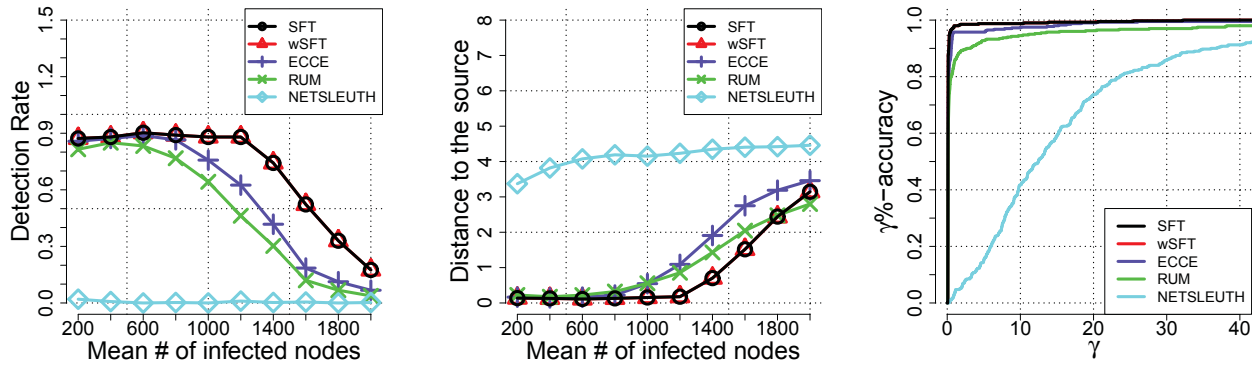
Figure 4a shows the detection rates for different infection sizes. The detection rates of ECCE, SFT and wSFT do not change for different infection sizes since the structure of the binomial tree is simple. SFT, wSFT and ECCE have the highest detection rate (more than 0.9) while the detection rate of RUM and NETSLEUTH are much lower.

The distance results are shown in Figure 4b. As expected, SFT, wSFT and ECCE outperform RUM, which are all much better than NETSLEUTH.

Figure 4c shows the $\gamma\%$-accuracy versus the rank percentage γ. We picked infection size 1,000. As shown in Figure 4c, all three algorithms based on infection eccentricity (ECCE, SFT, wSFT) have better performance than RUM and NETSLEUTH. Recall that the node identified by wSFT is a MAP estimator of the actual source.

D. The ER random graph

In this section, we compared the performance of the algorithms on the ER random graph. In the experiments, we generated the ER random graph with $n = 5,000$ and wiring probability $p = 0.002$. We again varied the infection network size from 200 to $2,000$. The infection probability of each edge is assigned with a uniform distribution in $(0.2, 0.5)$. We generated 400 diffusion samples.



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

Fig. 5: Performance in the ER random graph

Figure 5a shows the detection rate versus the infection size. The detection rate decreases as the infection size increases. SFT and wSFT have higher detection rates compared to other algorithms. Figure 5b shows the results on distance. As we expected, SFT and wSFT outperform other algorithms when the infection size is less than 1,600 nodes. As the size of the infected nodes increase, SFT and wSFT become close to RUM in term of distance to the source. However, the detection rate of both algorithms are still much higher than that of RUM. Another observation is that SFT and wSFT have identical performance which indicates that the performance of SFT is robust to edge weights.

Figure 5c shows the $\gamma\%$-accuracy versus the rank percentage γ with 1000 infected nodes. SFT and wSFT have similar or better performance compared to all other algorithms.

Although the performance of ECCE and SFT algorithms are similar in tree networks, SFT outperforms ECCE significantly on the ER random graphs. The observation indicates that BND is an effective tie breaking rule and increases the detection accuracy.

E. The Internet Autonomous System Network

The Internet autonomous systems (IAS) network[2] is the Internet autonomous system from Oregon route-views on March, 31st, 2001 with 10,670 nodes and 22,002 edges. The IAS network is a small world network. We adopted similar settings as in Section IV-D.

The detection rates are shown in Figure 6a. The detection rate of ECCE is low since the IAS graph is a small world network and there are multiple Jordan infection centers due to the small diameter of the network. With the tie breaking rule BND, the detection rate doubles in most cases which demonstrates the effectiveness of BND. While the detection rate of SFT is only 10% when the infection size is 1,000, the distance to the actual source is slightly more than one-hop away as shown in Figure 6b. In addition, the $\gamma\%$-accuracy versus γ for 1,000 infection size is shown in Figure 5c. The 10%-accuracies of SFT and wSFT are close to 70% which are significantly higher than that of other algorithms.

[2]Available at http://snap.stanford.edu/data/index.html

F. Running Time vs Performance

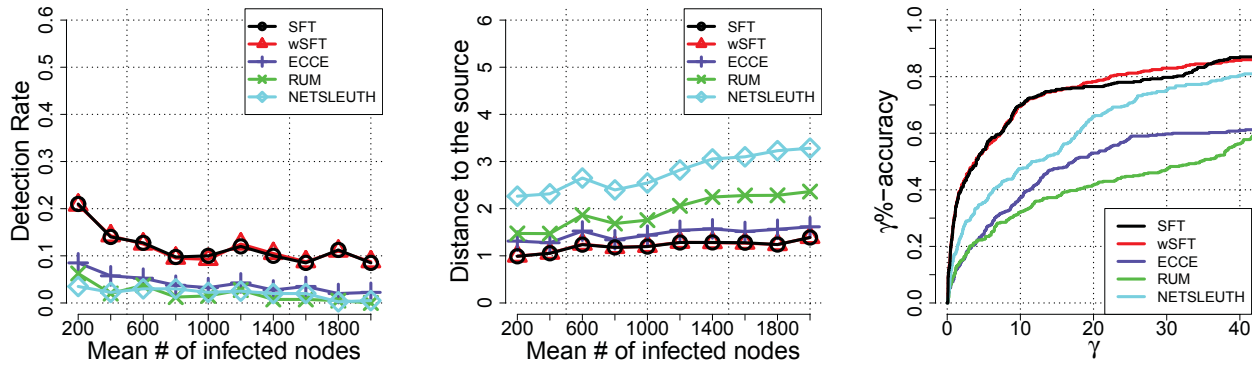
In this section, we evaluated the scalability of the algorithms by comparing the running time. The experiments were conducted on an Intel Core i5-3210M CPU with four cores and 8G RAM with a Windows 7 Professional 64 bit system. All algorithms were implemented with python 2.7. The ER random graphs with 5,000 nodes and $p = 0.002$ edge generation probability were used in the experiments. The infection probability of each edge is uniformly distributed over $(0.2, 0.5)$. We generated 100 diffusion samples for the experiments. Figure 7 show the average running time versus the detection rate. The infection size is chosen to be 1,000. SFT and wSFT took 1.11 seconds and achieves 0.87 detection rate while NETSLEUTH took 0.62 seconds with 0 detection rate and RUM took 14.86 seconds with 0.7 detection rate. The detection rate of SFT is much higher than NETSLEUTH and SFT is 14 times faster than RUM.

V. Conclusions

In this paper, we derived the MAP estimator of the information source on tree networks under the IC model. Based on that, the SFT algorithm has been proposed. We proved that the SFT algorithm identifies the information source with probability one asymptotically in the ER random graph when the observation time $t \leq \frac{2}{3}t_u$, which is the first theoretical guarantee on non-tree networks to our best knowledge. We evaluated the performance of SFT on tree networks, the ER random graph and the IAS network.

Acknowledgment

This work was supported in part by the U.S. Army Research Laboratory's Army Research Office (ARO Grant No. W911NF1310279).



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

Fig. 6: Performance in the IAS graph

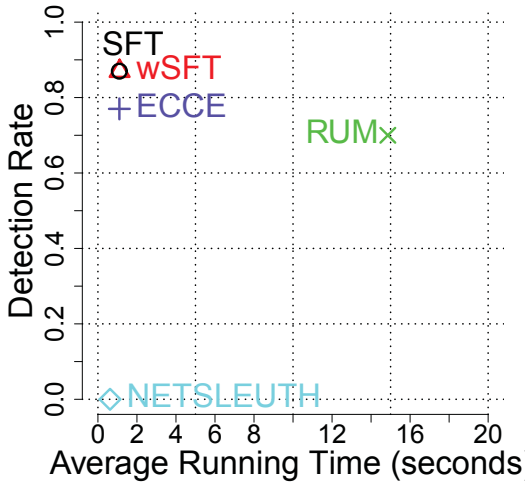


Fig. 7: Detection rate versus running time in the ER random graph

REFERENCES

[1] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample path based approach," *IEEE/ACM Trans. Netw.*, Nov. 2014. DOI: 10.1109/TNET.2014.2364972.

[2] J. Snow, "The cholera near Golden-square, and at Deptford," *Medical Times and Gazette*, 1854.

[3] D. Shah and T. Zaman, "Rumors in a network: Who's the culprit?," *IEEE Trans. Inf. Theory*, vol. 57, pp. 5163–5181, Aug. 2011.

[4] P. Erdos and A. Renyi, "On random graphs I," *Publ. Math. Debrecen*, vol. 6, pp. 290–297, 1959.

[5] D. Shah and T. Zaman, "Rumor centrality: a universal source detector," in *Proc. Ann. ACM SIGMETRICS Conf.*, (London, England, UK), pp. 199–210, 2012.

[6] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Trans. Signal Process.*, vol. 61, pp. 2850–2865, 2013.

[7] N. Karamchandani and M. Franceschetti, "Rumor source detection under probabilistic sampling," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, (Istanbul, Turkey), July 2013.

[8] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Proc. IEEE Int. Symp. Information Theory (ISIT)*, (Istanbul, Turkey), pp. 2671–2675, 2013.

[9] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: fundamental limits and algorithms," in *Proc. Ann. ACM SIGMETRICS Conf.*, (Austin, TX), 2014.

[10] K. Zhu and L. Ying, "A robust information source estimator with sparse observations," in *Proc. IEEE Int. Conf. Computer Communications (INFOCOM)*, (Toronto, Canada), April-May 2014.

[11] Z. Chen, K. Zhu, and L. Ying, "Detecting multiple information sources in networks under the SIR model," in *Proc. IEEE Conf. Information Sciences and Systems (CISS)*, (Princeton, NJ), 2014.

[12] W. Luo and W. P. Tay, "Estimating infection sources in a network with incomplete observations," in *Proc. IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, (Austin, TX), pp. 301–304, 2013.

[13] W. Luo and W. P. Tay, "Finding an infection source under the SIS model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, (Vancouver, BC), May 2013.

[14] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *Proc. Ann. ACM SIGKDD Conf. Knowledge Discovery and Data Mining (KDD)*, pp. 1059–1068, 2010.

[15] 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.

[16] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?," in *IEEE Int. Conf. Data Mining (ICDM)*, (Brussels, Belgium), pp. 11–20, 2012.

[17] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Phys. Rev. E*, vol. 90, p. 012801, Jul 2014.

[18] P. C. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Phys. Rev. Lett.*, vol. 109, no. 6, p. 068702, 2012.

[19] K. Zhu, Z. Chen, and L. Ying, "Locating contagion sources in networks with partial timestamps," *arXiv preprint arXiv:1412.4141*, 2014.

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

[21] S. Zejnilovic, J. Gomes, and B. Sinopoli, "Network observability and localization of the source of diffusion based on a subset of nodes," in *Proc. Annu. Allerton Conf. Communication, Control and Computing*, (Monticello, IL), 2013.

[22] K. Zhu and L. Ying, "Source localization in networks: Trees and beyond," *arXiv preprint arXiv:1510.01814*, 2015.

[23] F. Harary, *Graph theory*. Addison-Wesley, 1991.

[24] M. Draief and L. Massouli, *Epidemics and rumours in complex networks*. Cambridge University Press, 2010.