Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems

School of Information Systems

1-2016

Information source detection via maximum a posteriori estimation

Biao CHANG
Singapore Management University, bchang@smu.edu.sg

Feida ZHU
Singapore Management University, fdzhu@smu.edu.sg

Enhong CHEN

Qi. LIU

DOI: https://doi.org/10.1109/ICDM.2015.116

Citation

CHANG, Biao; ZHU, Feida; CHEN, Enhong; and LIU, Qi.. Information source detection via maximum a posteriori estimation. (2016). *Proceedings 15th IEEE International Conference on Data Mining ICDM 2015*. 21-30. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3137

2015 IEEE International Conference on Data Mining

Information Source Detection via Maximum A Posteriori Estimation

Biao Chang[1], Feida Zhu[2], Enhong Chen[1], Qi Liu[1]
[1]University of Science and Technology of China, [2]Singapore Management University
[1]chbiao@mail.ustc.edu.cn, {cheneh,qiliuql}@ustc.edu.cn, [2]fdzhu@smu.edu.sg

***Abstract*—The problem of information source detection, whose goal is to identify the source of a piece of information from a diffusion process (e.g., computer virus, rumor, epidemic, and so on), has attracted ever-increasing attention from research community in recent years. Although various methods have been proposed, such as those based on centrality, spectral and belief propagation, the existing solutions still suffer from high time complexity and inadequate effectiveness. To this end, we revisit this problem in the paper and present a comprehensive study from the perspective of likelihood approximation. Different from many previous works, we consider both infected and uninfected nodes to estimate the likelihood for the detection. Specifically, we propose a Maximum A Posteriori (MAP) estimator to detect the information source for general graphs with rumor centrality as the prior. To further improve the efficiency, we design two approximate estimators, namely *Brute Force Search Approximation* (BFSA) and *Greedy Search Bound Approximation* (GSBA). BFSA tries to traverse the permitted permutations and directly computes the likelihood, while GSBA exploits a strategy of greedy search to find a surrogate upper bound of the probabilities of permitted permutations for a given node, and derives an approximate MAP estimator. Extensive experiments on several network data sets clearly demonstrate the effectiveness of our methods in detecting the single information source.**

***Keywords*—information source detection; maximum a posteriori; likelihood approximation; greedy search;**

I. Introduction

The boom of research on information network analysis, especially those on information diffusion such as influence maximization [16], [21], has brought ever-increasing attention to the topic of *information source detection* [33], which aims to identify the information source based on a single snapshot of the infected network (e.g., opinion, computer virus, rumor and epidemic). Its wide range of applications include epidemic outbreak prevention, Internet virus source identification and rumor source tracing in social networks [9], [12], [26], [29].

The research challenges of this problem come from a number of aspects: First, information diffusion is characteristic of high dynamicity and a great variety of patterns when initiated from different sources [14]. Second, the actual information diffusion model, which is often latent, has become the modelling target for many well-received pieces of works, such as Susceptible-Infected-Recovered (SIR) Model [17] and Independent Cascade (IC) Model [10]. Third, each diffusion process produces an infection sequence, giving out not only who are affected but also when each infection takes place. Unfortunately, the infection sequence is typically unavailable for the purpose of source detection. Nevertheless, various methods have been introduced along the years to overcome these challenges and detect the source of an diffusion for different situations, including methods based on centrality [15], [28], spectral [9], belief propagation [1], [2], [22], and so on.

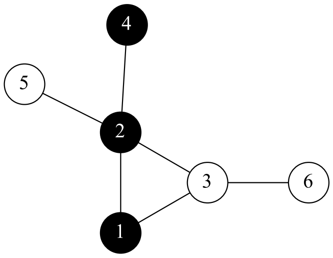

Fig. 1. A snapshot of the information diffusion on a toy graph, where black nodes are infected and others are not.

Despite all the research efforts, existing methods are still deemed inadequate due to their high computational complexity and yet-to-be-improved effectiveness. Therefore, in this paper we revisit the problem of *information source detection* and present a comprehensive study from a different perspective of likelihood approximation. The intuition behind our method is that uninfected (or susceptible) nodes also provide important insight for detecting the source. We illustrate this point with Figure 1 which shows a snapshot of an information diffusion example on an undirected graph. Since node 2 has two uninfected neighbors node 3 and node 5, if node 2 is indeed the source, node 3 and 5 would be more likely to be infected. Therefore, the presence of these two uninfected neighbors reduces the probability of node 2 being the source. Although this has been noticed by some work [22], [26], we exploit this intuition along a different direction.

Specifically, we first derive a Maximum A Posteriori (MAP) estimator to detect the information source for general graphs, which selects a node with the maximal posterior probability as the detected source. The Susceptible-Infected (SI) model [26], [28] is used to describe the process of information diffusion, which is a variant of SIR. It assumes that every node has two potential states, namely susceptible and infected. The MAP estimator applies *rumor centrality* [28] as the prior because Comin *et al.* [6] have shown that the source node tends to have higher centrality measurement values. Then we infer the exact formation of likelihood for general graphs, based on the hypothesis that the likelihood equals to the sum of probabilities of all permitted permutations starting with a

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.116

node. A permitted permutation [28] is corresponding to the node infection sequence which is generated by an information diffusion. For example, in Figure 1, $\{1, 2, 4\}$ is a permitted permutation if node 1 is the source. Combining the above prior and likelihood, we obtain the MAP estimator.

While a theoretically optimal MAP estimator is obtainable by enumerating all the permitted permutations, for greater efficiency , we design two approximate estimators, *Brute Force Search Approximation* (BFSA) and *Greedy Search Bound Approximation* (GSBA). Inspired by Shah and Zaman [28], BFSA uses a breadth-first search tree to estimate the spanning tree, and then enumerates the corresponding permitted permutations to get the approximate likelihood. Yet BFSA is still time-consuming as their research results have shown the factorial complexity of the number of permitted permutations for general trees. We therefore further propose GSBA which uses an upper bound to approximate the probabilities of permitted permutations starting with any given node v. This bound can be used to estimate the likelihood given that v is the source. To find this upper bound, GSBA exploits a strategy of greedy search to find a *surrogate* bound, which effectively avoids the enumeration of permitted permutations and drastically reduces the computational complexity. The experimental results on several network data sets validate the effectiveness of our methods.

To sum up, our contributions are listed as follows.

- We derive a Maximum A Posteriori estimator to detect the information source for general graphs. We show that we can obtain a theoretically optimal MAP estimator by enumerating all the permitted permutations.
- To improve the efficiency, we develop two approximation variants for the MAP estimator, namely *Brute Force Search Approximation* and *Greedy Search Bound Approximation*.
- We conduct comprehensive experiments on three networks to validate the effectiveness of our methods. The experimental results clearly demonstrate the effectiveness of our proposed approaches for single information source detection by outperforming the baselines .

Roadmap. The remainder of this paper is organized as follows. Section II provides a brief review of related works. Then we introduce some preliminaries of information source detection in Section III. Section IV and V give the details of our methods. In Section VI, we report the experimental results. Finally, we conclude the paper and discuss some future works in Section VII.

II. Related Work

In general, research work related to our problem can be discussed by two categories: information diffusion modeling and information source detection.

Information Diffusion Modeling. It is a fundamental problem to model information diffusion process, which has attracted research efforts from various communities including epidemiology, ethnography, and sociology [32]. Kermack and McKendrick introduced the Susceptible-Infected-Recovered (SIR) model to describe epidemic spreading [17]. The model assumes that every node has three possible states, i.e., susceptible, infected, and recovered. Once a susceptible node is infected, it can further infect its susceptible neighbors, but it may recover and never be infected again. Note that the Susceptible-Infected (SI) model used in this paper supposes that infected nodes would never recover, which is a special case of SIR. In social network analysis, Independent Cascade (IC) Model [10] and Linear Threshold (LT) Model [11] are widely used to characterise the information diffusion in social networks. Other models such as Susceptible-Infected-Susceptible (SIS) Model and models for diffusion of innovations can be found in [32].

Information Source Detection. Various methods (e.g., those based on centrality, spectral, belief propagation, and so on) have been proposed to identify the diffusion source for different situations. For example, Shah and Zaman [28], [29] are among the first to consider this problem. They proposed the notion of *rumor centrality* to implement the Maximum Likelihood Estimation for single rumor source detection under the SI model. For a node, its rumor centrality is the number of infection sequences which can span the given infected subgraph. We can see rumor centrality only considers the utility of infected nodes to detect the source. Under a specific continuous-time epidemic process, Pinto *et al.* [25] studied the problem when only a small fraction of nodes, instead of the whole graph, can be observed. Zhu and Ying [33] developed a sample-path-based approach to detect the source under the SIR model. The source is supposed to be the root of a sample path which is the node most likely resulting in the infected subgraph. They proved that for a tree graph, the output of their method is a Jordan center [15], which minimizes the maximum distance from a node to others. Dong *et al.* [7] explored the Maximum A Posteriori (MAP) estimation with different settings of the prior. For instance, the suspects may be all the infected nodes, or at most k infected nodes.

Recently, Lokhov *et al.* [22] made use of the infected and uninfected nodes to detect the source and introduced a time-consuming yet effective inference algorithm based on dynamic message passing (DMP) equations. It first uses DMP to estimate the marginal probability of a given node to be in a given state, and then exploits a mean-field-type approach to approximate the likelihood. Altarelli *et al.* [2] conducted Bayesian inference for this problem on a factor graph under the SIR model. They derived belief propagation (BP) equations for the probability distribution of system states conditioned on some observations, which is more accurate than DMP. They further considered this problem with noisy observations in [1]. Wang *et al.* [30] addressed the problem of rumor source detection with multiple independent observations, under the SI model. For trees, they found that multiple independent observations can dramatically increase the detection probability.

In addition, many researchers focused on detecting multiple information sources. Prakash *et al.* [26] started to explore the detection of multiple information sources under the SI model. They applied the Minimum Description Length (MDL) principle to automatically decide the number of source nodes, and then identified the best source nodes according to *exoneration* of infected nodes with many uninfected neighbors. Subsequently, Fioriti and Chinnici [9] proposed to use the *node dynamical importance* to estimate nodes' age, and designed a spectral technique to predict the sources of an outbreak. Dynamical importance of a node is the reduction of the largest eigenvalue of the adjacent matrix after it is removed from the network [27]. Luo *et al.* [23] extended rumor centrality for multiple sources detection, and they also tried to estimate the infection regions (i.e., nodes infected by each source).

Intuitively, information source detection can be viewed as the reverse process of information diffusion [16]. Lappas *et al.* [18] defined a similar problem, k-Effectors, which selects a set of k active nodes that can best explain the observed activation states in social networks. They proved that the k-Effectors(0) problem is NP-complete under the IC model, and gave two approximate solutions. Nguyen *et al.* [24] studied the k-Suspector problem which aims to find the top k most suspected sources of misinformation, and claimed NP-hardness of the problem under the IC model. Gundecha *et al.* [12] tried to seek the provenance of information for a few known recipients by recovering the information propagation paths in social media. Feng *et al.* [8] studied the problem of recovering other unknown recipients and seeking the provenance of information based on a few known recipients. They exploited frequent pattern propensity and node centrality measures to find important nodes.

In spite of all these existing work, we approach the problem of information source detection by maximizing a posteriori from a different perspective. Our solution makes full use of both the infected nodes and their uninfected neighbors, like [26], but from a different perspective. We assume that every infected node could be the source and use node centrality as the prior probability. Then we infer the exact formation of the likelihood for general graphs.

III. Preliminaries

In this section, we will introduce some preliminaries about information source detection, and give the problem formulation, and then revisit *Rumor Centrality*.

A. Problem Definition

Let $G(V, E)$ denote the undirected network, where V is the node set and E is the edge set. The information (such as opinion, rumor and epidemic) will spread on this network under a contagious information diffusion model. In this paper, we assume the information source consists of a single node, and apply the Susceptible-Infected (SI) model to describe the diffusion. Some important terms and notations are listed in Table I for easy reference.

SI is a variant of the popular Susceptible-Infected-Recovered model [17]. It assumes that every node has two possible states: susceptible and infected. Once a node i is infected (or receive the information), it will remain infected and never recover. Meanwhile, the node i will spread the information to its susceptible neighboring node j with probability λ_{ij}. All λ_{ij}'s are supposed to be independent. [3], [26]. In the following, we assume that $\lambda_{ij} = \lambda$ for convenience without loss of generality.

After the information has spread on the network for some time, there are N infected nodes, denoted by V_I, including the source node. These nodes and their inter-edges E_I can span an infected subgraph $G_I(V_I, E_I)$ of $G(V, E)$, which are referred to as G_I and G, respectively. G_I is connected because the information diffusion model is contagious. For example, recall the snapshot of information diffusion on the toy graph in Figure 1. Node 1, 2 and 4 are infected while others are susceptible. Therefore, the problem of *information source detection* is defined as follows [28].

TABLE I. Terms and Notations

Notations	Description
$G = G(V, E)$	an undirected graph
$G_I = G_I(V_I, E_I)$	an infected subgraph of G
v^*	the real information source
$\hat{v}$	the detected source
σ	a permitted permutation
$\sigma(1, \ldots, i-1)$	the first $i-1$ nodes of σ
$R(v, G_I)$	rumor centrality of node v
$G_s(U)$	a spanning graph of a node set U
$N(U)$	neighbors of a node set U
$d(U)$	degree of a node set U
$E(v, U)$	set of bridging edges between v and the node set U
$\Omega(v, G_I)$	set of permitted permutations starting with v and spanning G_I.

*Problem Definition 1 (**Information Source Detection**):* Given an undirected graph $G(V, E)$ and a snapshot of the infected subgraph $G_I(V_I, E_I)$ at some time stamp, the problem of *Information Source Detection* is to find the single information source v^* among all those infected nodes.

B. Rumor Centrality

After the information diffusion starts, it generates an infection node sequence $\sigma = \{v_1, \ldots, v_N\}$ $(1 \leqslant i \leqslant N)$, where $v_i \in V_I$ is the i-th infected node (i.e., $\sigma(i) = v_i$). This sequence corresponds to a permutation of N nodes, which is referred to as the *permitted permutation* in [28], and vice versa. This means that a permutation is permitted if it exactly matches the topological constrain specified by G_I. For example, if node 4 is the information source in Figure 1, {4,2,1} is a permitted permutation, but {4,1,2} is not because node 2 must be infected before node 1. Note that we also use σ to denote the nodes appearing in the sequence without ambiguity.

If G_I is a general tree, Shah and Zaman [28] have shown that *Rumor Centrality* of node v, i.e., the number of permitted permutations starting with v, denoted as $R(v, G_I)$, is defined as

$$R(v, G_I) = \prod_{u \in G_I} \frac{N!}{T_u^v}, \tag{1}$$

where u is a node of G_I and T_u^v is the number of nodes in the subtree rooted at u with v as the source. This centrality has been used to design a *Maximum Likelihood Estimator* (MLE) to estimate the likelihood probability $\mathbf{P}(G_I | v^* = v)$ given that node v is the information source.

Although rumor centrality is effective, it has two limitations. First, it only considers the infected subgraph and neglects other susceptible nodes which are also important for detecting the information source. For example, in Figure 1, $R(1, G_I) = R(4, G_I) = 1$, $R(2, G_I) = 2$. So node 2 is more likely to be the source according to rumor centrality no matter how many neighbors G_I has. But node 2 has two uninfected neighbors, node 3 and 5. If node 2 is the source, node 3 and 5 are more likely to be infected. Therefore, node 3 and 5 have reduced the chance of node 2 as the source, when we consider the global states of all nodes.

Second, rumor centrality assumes that the probabilities of all permitted permutation are equal for general graphs. It is easy to show that this assumption is not valid when the degrees of nodes are different, especially for graphs with loops.

To overcome these limitations and improve the accuracy for the detection, we propose a *Maximum a Posteriori* (MAP) estimator. It uses rumor centrality as the *prior* probability $\mathbf{P}(v^* = v)$, and then accurately infer the probability of a permitted permutation to obtain the *likelihood* $\mathbf{P}(G_I|v^* = v)$, which takes into account of the effects of both infected and uninfected nodes.

IV. Maximum a Posteriori Estimation

Based on Bayes Theorem, we can derive the posterior probability of node v being the real source given the infected subgraph G_I as follows.

$$\begin{aligned}\mathbf{P}(v^* = v|G_I) &= \frac{\mathbf{P}(G_I|v^* = v)\mathbf{P}(v^* = v)}{\mathbf{P}(G_I)} \\ &= \frac{\mathbf{P}(G_I|v^* = v)\mathbf{P}(v^* = v)}{\sum_{u\in G_I}\mathbf{P}(G_I|v^* = u)} \\ &\propto \mathbf{P}(G_I|v^* = v)\mathbf{P}(v^* = v),\end{aligned} \tag{2}$$

because the denominator $\mathbf{P}(G_I)$ is the sum of values appearing in the numerator and can be regarded as the normalization constant to be removed [5].

Given G_I, we can select the node maximizing the above posterior as the detected source. This is the following maximum a posteriori estimator.

$$\begin{aligned}\hat{v} &= arg\max_{v\in G_I}\mathbf{P}(v^* = v|G_I) \\ &= arg\max_{v\in G_I}\mathbf{P}(G_I|v^* = v)\mathbf{P}(v^* = v).\end{aligned} \tag{3}$$

Next we will derive the prior probability and likelihood in Equation (2), respectively.

A. Deriving the Prior

Although many works assume every node has the same prior to be the source [28], [22], [30], Comin *et al.* [6] have shown that the source node tends to have higher centrality measurement values. Therefore, we choose rumor centrality as the prior.

$$\begin{aligned}\mathbf{P}(v^* = v) &= \frac{R(v, G_I)}{\sum_{u\in V_I} R(u, G_I)} \\ &\propto R(v, G_I).\end{aligned} \tag{4}$$

For general graphs, we apply the following method used in [28] to compute $R(v, G_I)$,

$$R(v, G_I) \approx R(v, T_{bfs}(v)) \tag{5}$$

where $T_{bfs}(v)$ is a breadth-first search spanning tree starting with v. It uses the rumor centrality of $T_{bfs}(v)$ to approximate $R(v, G_I)$ of a graph.

B. Deriving the Likelihood

Formally, let $G_s(U)$, $N(U)$ and $d(U)$ be the spanning graph, the neighbor set and degree of a node set U, respectively. A spanning graph of a node set consists of these nodes and inter-edges among them. Note that $G_s(\sigma)$ is the spanning graph of nodes appearing in the permitted permutation σ. Let

$$\Omega(v, G_I) = \{\sigma|\sigma(1) = v, G_s(\sigma) = G_I\} \tag{6}$$

denote the set of permitted permutations each of which starts with v and could span the infected subgraph G_I. Let $E(v, U) = \{(v, u)|(v, u) \in E, u \in U\}$ be the set of bridging edges between v and the node set U. For example, in Figure 1, $G_s(\{1, 2, 4\}) = G_I$, $N(\{1, 2\}) = \{3, 4, 5\}$, $d(\{1, 2\}) = 4$, $\Omega(2, G_I) = \{\{2, 1, 4\}, \{2, 4, 1\}\}$, $E(3, \{1, 2\}) = \{(3, 1), (3, 2)\}$.

For an infection sequence or permitted permutation $\sigma \in \Omega(v, G_I)$, we have

$$\mathbf{P}(\sigma(i) = u|\sigma(1, \ldots, i-1)) = \frac{|E(u, \sigma(1, \ldots, i-1))|}{d(\sigma(1, \ldots, i-1))}, \tag{7}$$

where u is a neighbor of $\sigma(1, \ldots, i-1)$. It is the probability that u is selected to be the i-th infected node from the neighbors of the first $i-1$ nodes of σ. This means that every $u \in N(\sigma(1, \ldots, i-1))$ has the probability to be the next infected node, which is proportional to the number of back edges $E(u, \sigma(1, \ldots, i-1)$ between u and $\sigma(1, \ldots, i-1)$, as we assume that $\lambda_{ij} = \lambda$. For example, in Figure 1, $\mathbf{P}(\sigma(3) = 3|\sigma(1) = 2, \sigma(2) = 1) = 2/4 = 0.5$.

Accordingly, we can expand $\mathbf{P}(\sigma|v^* = v)$ and get the following based on Equation (7).

$$\begin{aligned}&\mathbf{P}(\sigma|v^* = v) \\ &= \mathbf{P}(\sigma(2)|\sigma(1))\cdots\mathbf{P}(\sigma(N)|\sigma(1, \ldots, N-1)) \\ &= \frac{|E(\sigma(2), \sigma(1))|}{d(\sigma(1))}\cdots\frac{|E(\sigma(N), \sigma(1, \ldots, N-1))|}{d(\sigma(1, \ldots, N-1))} \\ &= \prod_{i=2}^{N}\frac{|E(\sigma(i), \sigma(1, \ldots, i-1))|}{d(\sigma(1, \ldots, i-1))},\end{aligned} \tag{8}$$

where

$$\begin{aligned}&d(\sigma(1, \ldots, i-1)) \\ &\quad= \sum_{j=1}^{j=i-1}(d(\sigma(j)) - 2|E(\sigma(j), \sigma(1, \ldots, j-1))|),\end{aligned} \tag{9}$$

since every infected node contributes $d(\sigma(j)) - 2|E(\sigma(j), \sigma(1, \ldots, j-1))|$ new edges to the diffusion.

When the graph is a tree, every node has only one path to connect with others. This means $|E(\sigma(j), \sigma(1, \ldots, j-1))| = 1$. Thus Equation (8) becomes the following succinct form.

$$\mathbf{P}(\sigma|v^* = v) = \frac{1}{\prod_{i=2}^{N}\sum_{j=1}^{j=i-1}(d(\sigma(j)) - 2)}, \tag{10}$$

which is used for deriving the rumor centrality in [29].

Notice that the likelihood $\mathbf{P}(G_I|v^* = v)$ is the sum of probabilities of all permitted permutations which begin with

v [28]. Therefore, it can be decomposed as follows.

$$\begin{aligned}\mathbf{P}(G_I|v^*=v) &= \sum_{\sigma\in\Omega(v,G_I)} \mathbf{P}(\sigma|v^*=v) \\ &= \sum_{\sigma\in\Omega(v,G_I)} \prod_{i=2}^{N} \frac{|E(\sigma(i),\sigma(1,\ldots,i-1))|}{d(\sigma(1,\ldots,i-1))}.\end{aligned} \tag{11}$$

After substituting Equation (4) and (11) into (2), we obtain the following formation of the posterior probability.

$$\begin{aligned}&\mathbf{P}(v^*=v|G_I) \\ &\propto R(v,G_I) \sum_{\sigma\in\Omega(v,G_I)} \prod_{i=2}^{N} \frac{|E(\sigma(i),\sigma(1,\ldots,i-1))|}{d(\sigma(1,\ldots,i-1))}.\end{aligned} \tag{12}$$

It is clear that it indeed considers the states of both infected and susceptible nodes, and computes the probability for every permitted permutation from the global perspective.

In fact, when all nodes are infected (i.e., $G_I = G$), the MAP estimator defined by the above equation degenerates into the Maximum Likelihood Estimator in [28]. Because if $G_I = G$ and v is the source, the information must follow one permitted permutation of $\Omega(v, G_I)$ to spread such that the sum, $\mathbf{P}(G_I|v^* = v)$, in Equation (11) equals to 1. In other words, when $G_I = G$, every node could be the source and infect all the others as long as the information spreads for a sufficiently long period of time. At this moment, we can only use the prior knowledge depicted by rumor centrality in IV-A to distinguish these nodes.

If enumerating all the permitted permutations, Equation (12) says that we can get the theoretically optimal MAP estimator. Yet Equation (1) has shown the factorial complexity of the number of permitted permutations even for general trees, not to mention for general graphs. We will show how to get the approximate estimators to speed up the detection.

V. Approximate MAP Estimator

We propose two MAP approximation estimators in this section, namely *Brute Force Search Approximation* and *Greedy Search Bound Approximation*.

A. Brute Force Search Approximation

Brute Force Search Approximation (BFSA) tries to enumerate all the permitted permutations and get the MAP estimator. Algorithm 1 shows the pseudo codes. Specifically, it first initializes the likelihood, gets the breadth-first search spanning tree and the prior $R(v, G_I)$ for every infected node, from line 1 to 4. Then it calls Algorithm 2 to generate permitted permutations and obtains the likelihood. Finally, it gets the detected source by MAP according to Equation (3).

Algorithm 2 extends Heap's permutation generating algorithm [13]. During the generating process, it prunes the searching branches that do not follow the topological constrain by line 6. That means if $\sigma[p]$ is a descendant of $\sigma[i]$ in $T_{bfs}(\sigma[1])$, we can swap them to get a new permitted permutation.

BFSA can get the optimal MAP estimator for general trees, but may miss some permitted permutations for graphs with loops. Nevertheless, we will show the effectiveness of BFSA for source detection in the experiment part. However, its time complexity is exorbitantly high. To further improve the efficiency, we propose the following approximate estimator.

Algorithm 1: Brute Force Search Approximation(G,G_I)

```
   input : $G$ - the undirected graph
           $G_I$ - the infected subgrpah
   output: $\hat{v}$ - the detected source
 1 for $v \in V_I$ do
 2     $\mathbf{P}(G_I|v^* = v) = 0$;
 3     get the breadth first search spanning tree $T_{bfs}(v)$;
 4     get $R(v, G_I)$ by Equation (5);
 5 $\sigma$ = an array of $V_I$;
 6 getLikelihoodByBFSA($\sigma$,1,N);
 7 get $\hat{v}$ by Equation (3) and (11);
 8 return $\hat{v}$;
```

Algorithm 2: getLikelihoodByBFSA(σ, p, q)

```
   input : $\sigma$ - the infected node array
           $p$ - the starting index
           $q$ - the end index
 1 if $p == q$ then
 2     get $\mathbf{P}(\sigma|v^* = v)$ by Equation (8);
 3     $\mathbf{P}(G_I|v^* = v)$ += $\mathbf{P}(\sigma|v^* = v)$;
 4 else
 5     for $i = p; i \leq q; ++i$ do
 6         if $\sigma[p]$ is a node of the subtree rooted at $\sigma[i]$ of
           $T_{bfs}(\sigma[1])$ then
 7             swap($\sigma[p], \sigma[i]$);
 8             getLikelihoodByBFSA($\sigma, p + 1, q$);
 9             swap($\sigma[p], \sigma[i]$);
```

B. Greedy Search Bound Approximation

The basic idea of Greedy Search Bound Approximation (GSBA) is to find the upper bound of the appearinggenerative probability of a permitted permutation, and then to reduce the computational complexity of computing the likelihood.

Recall that $\Omega(v, G_I)$ is the set of permitted permutations beginning with v. Among $\Omega(v, G_I)$, there must be a permutation σ such that its appearing probability $\mathbf{P}(\sigma|v^* = v)$ in Equation (8) is maximal, which we denote as σ^v_{max}. Therefore, we have

$$\begin{aligned}&\mathbf{P}(\sigma|v^*=v) \leq \mathbf{P}(\sigma^v_{max}|v^*=v) \\ &\qquad = \prod_{i=2}^{N} \frac{|E(\sigma^v_{max}(i),\sigma^v_{max}(1,\ldots,i-1))|}{d(\sigma^v_{max}(1,\ldots,i-1))}.\end{aligned} \tag{13}$$

More importantly, if we adopt the permitted permutation generation method in Algorithm 2, there exists the following approximation,

$$|\Omega(v, G_I)| = R(v, G_I). \tag{14}$$

Note that the above is exact when the graph is a tree. Combining Equation (13) and (14) with (12), we get an upper bound approximation of the posterior like

$$\mathbf{P}(v^* = v|G_I) \leq R^2(v, G_I)\mathbf{P}(\sigma^v_{max}|v^* = v). \tag{15}$$

Accordingly, the MAP estimator of Equation (3) changes into

$$\hat{v} = arg \max_{v \in G_I} R^2(v, G_I)\mathbf{P}(\sigma^v_{max}|v^* = v). \tag{16}$$

We denote it as *maximum a posteriori upper-bound* (MAP-ub).

Now the only issue left is how to find σ^v_{max} and get the upper bound. When $\sigma(1, \ldots, i-1)$ is given, if exploiting the greedy search strategy to select $w \in N(\sigma(1, \ldots, i-1))$ to be the i-th infected node such that $\frac{|E(w,\sigma(1,\ldots,i-1))|}{d(\sigma(1,\ldots,i))}$ is maximal, we can get a permitted permutation σ^v_{gs}. If we set σ^v_{gs} as a *surrogate* of σ^v_{max}, and the estimator of Equation (16) becomes

$$\hat{v} = arg \max_{v \in G_I} R^2(v, G_I)\mathbf{P}(\sigma^v_{gs}|v^* = v). \tag{17}$$

So far, we get the final greedy search bound approximation. The pseudo codes of GSBA is mostly like BFSA's, except for line 6 and 7. GSBA will get the likelihood approximation by Algorithm 3 and detect the source according to the estimator in Equation (17). Its computational complexity is $O(N^2)$.

Algorithm 3: getLikelihoodByGSBA(G_I)

```
  input : $G_I$ - the infected subgrpah
1 for $v \in G_I$ do
2     $i = 2$;
3     initialize an empty queue $Q$;
4     add $v$ into $Q$;
5     while $i \leq N$ do
6         $u$ = remove the first node of $Q$;
7         compute Equation (7);
8         insert every unvisited neighbor $w$ of $u$ into $Q$
          according to the descending order of
          $\frac{|E(w,\sigma(1,\ldots,i-1))|}{d(\sigma(1,\ldots,i))}$;
9         ++$i$;
```

For every node in V_I, Algorithm 3 uses the greedy search to find σ^v_{gs}. When inserting a new node into the queue Q, we use the concept of insertion sort to ensure that the first node of Q has the maximal probability to be the next infected node. Its effectiveness will be shown in the experiment section.

GSBA is a trade-off between effectiveness and efficiency to approximate the likelihood. We left it to future work to explore other algorithms to find σ^v_{max} such as dynamic programing.

VI. Experiment

In this section, we present experimental results to compare our methods with some baselines with respect to single information source detection on different networks under three evaluation measures.

A. Datasets

Our datasets are simulations about information diffusion on three networks, namely *Scale-Free*, *Power-Grid* and *Wiki-Vote*. This kind of datasets is widely used in the literature [28], [22]. Specifically, a scale-free network is a connected graph, whose degree distribution nearly follows a power law. We generate it by Barabasi-Albert (BA) Model [4]. Power-Grid [31] is an undirected network containing information about the power grid of the Western States of the United States of America[1]. *Wiki-Vote* [19] is a who-voted-who graph on Wikipedia[2], downloaded from *Stanford Network Analysis Project* (SNAP) [20]. We assume that node u and v has an undirected edge if there is an edge between them in the original network. As we said in Section III-A, an infected graph should be connected. So we remove the disconnected nodes and keep the maximal connected component. After this filtering, their statistical information are listed in Table II.

TABLE II. Statistics of our data sets

Network	*Scale-Free*	*Power-Grid*	*Wiki-Vote*
Number of nodes	500	$4,941$	$7,066$
Number of edges	764	$6,594$	$100,736$

To simulate the information diffusion on a graph G, we adopt the following two strategies to run the aforementioned SI model with $\lambda_{ij} = \lambda$ for $\forall i, j$, respectively.

- Random test. We randomly select a node as the infection source from G.
- Full test. We let each node has a chance to be the source, because every node can be the source in real scenarios.

After selecting a source node, we run the SI model until the number of infected nodes equals to a given value. Repeat this process, and finally we have M infected subgraphs with a given size for each graph. For random test, we let $M = 100$, while for full test, $M = |V|$, where $|V|$ is the node number of G. Indeed, full test can validate the stability of a method more accurately. But we only use full test to compare RG, DC, JC, RC, RI and DI with GSBA in this paper, because others are time-consuming.

B. Baselines and Evaluation Measures

To validate our methods, namely *Brute Force Search Approximation* (BFSA) and Greedy Search Bound Approximation (GSBA), we compare them with the following methods.

1) *Random Guess* (RG). It randomly selects an infected node as the source.
2) *Distance Center* (DC). It selects an infected node which has the minimal distance centrality as the source. Distance centrality is a sum of the shortest distance from a node to any others. [28]
3) *Jordan Center* (JC). It selects an infected node which minimizes the maximum distance to others as the source. [15]
4) *Rumor Center* (RC). It selects an infected node which has the maximal rumor centrality as the source. Rumor centrality is defined as Equation (1). [28]
5) *NETSLEUTH* (NS). Suppose that $L(G)$ is the Laplacian matrix corresponding to G, and L_A is a submatrix of $L(G)$ corresponding to the infected subgraph G_I. It first gets an eigenvector of L_A corresponding to the smallest eigenvalue, and then selects an infected node which is corresponding to the maximal component of the eigenvector as the source. [26]

[1] http://konect.uni-koblenz.de/networks/opsahl-powergrid

[2] http://www.wikipedia.org/

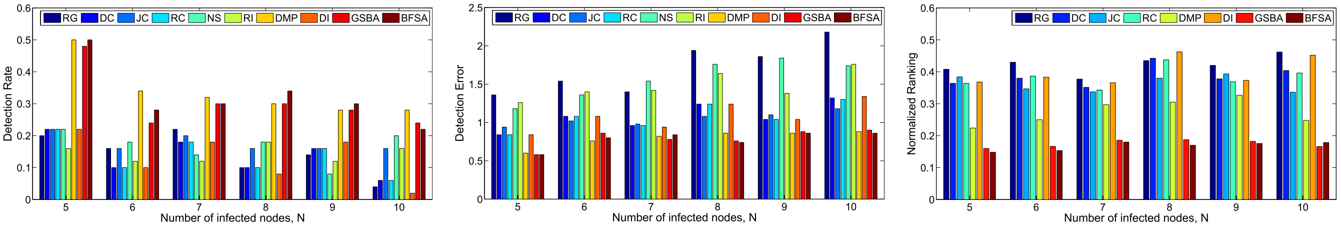

Fig. 2. Random test performance on Scale-Free Network with $\lambda = 0.2$.

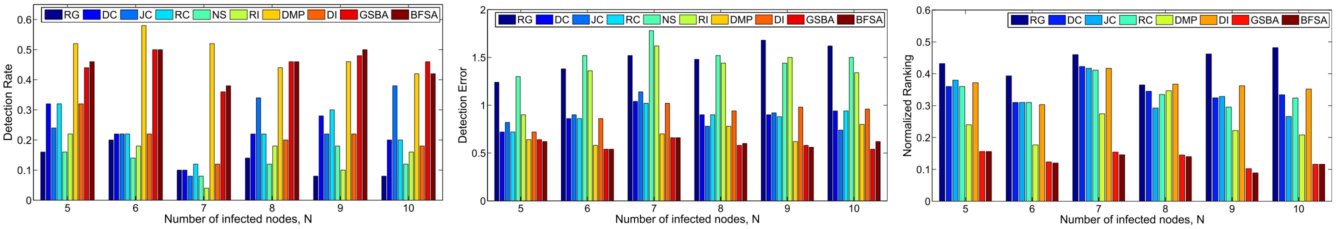

Fig. 3. Random test performance on Scale-Free Network with $\lambda = 0.5$.

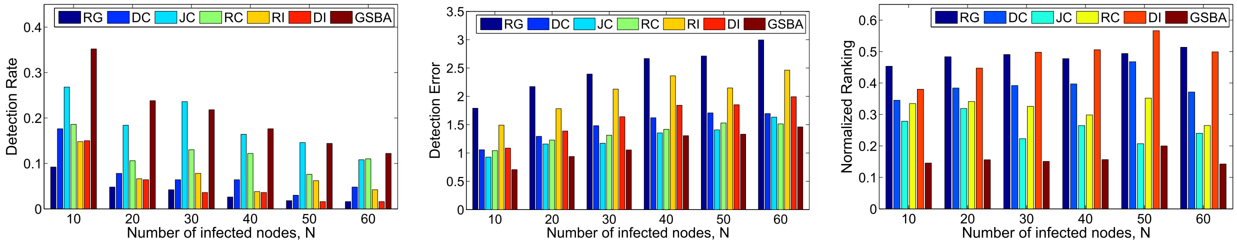

Fig. 4. Full test performance on Scale-Free Network with $\lambda = 0.5$.

6) *Reverse Infection* (RI). The algorithm lets every infected node broadcast its identity to the neighbors. Once a node receives a new identity, it will record the arriving time, and then broadcasts the identity to its neighbors. At last, the node which has received all the identities and the sum of their arriving times is minimal, is selected as the source. [33]
7) *Dynamic Message Passing* (DMP). The algorithm uses DMP equations to estimate the marginal probability of a given node to be in a given state, and then exploits a mean-field-type approach to approximate the likelihood. It selects an infected node corresponding to the maximal likelihood as the source. [22]
8) *Dynamic Importance* (DI). This is a method of the spectral family. It selects an infected node which has the maximal reduction of the largest eigenvalue of the adjacent matrix after it is removed from the network, as the source. [9]

Note that Lokhov *et al.* [22] have shown DMP is the state of the art among these methods.

We apply the following three widely used measures to evaluate the performance of different methods [28], [24], [22]. Let v^* be the real source and $\hat{v}$ be the detected source.

- *Detection Rate*. It is defined as

$$\text{Detection Rate} = \frac{M_T}{M}, \tag{18}$$

where M is the running number of tests and M_T is the number of tests which detect the source correctly.

- *Detection Error*. It is the average shortest distance between v^* and $\hat{v}$.
- *Normalized Ranking*. We rank the infected nodes in descending order by the probability to be the source. Normalized Ranking is defined as

$$\text{Normalized Ranking} = \frac{Ranking(v^*) - 1}{N_I}, \tag{19}$$

where N_I is the number of infected nodes and $Ranking(v^*)$ is the ranking of v^* in the sorted list.

To some degree, *Detection Rate* can reflect the detection precision of a method, and *Detection Error* shows how far the detected source is away from the real sources on the network, while *Normalized Ranking* can validate the accuracy that a method sorts the real source. Note that the larger *Detection Rate* is , the better performance the corresponding method achieves, but *Detection Error* and *Normalized Ranking* are opposite. *Normalized Ranking* is not applicable to NS and RI, because they do not assign a score to each node so that we cannot sort the infected nodes.

In the following, we will show the performance of these methods under different settings.

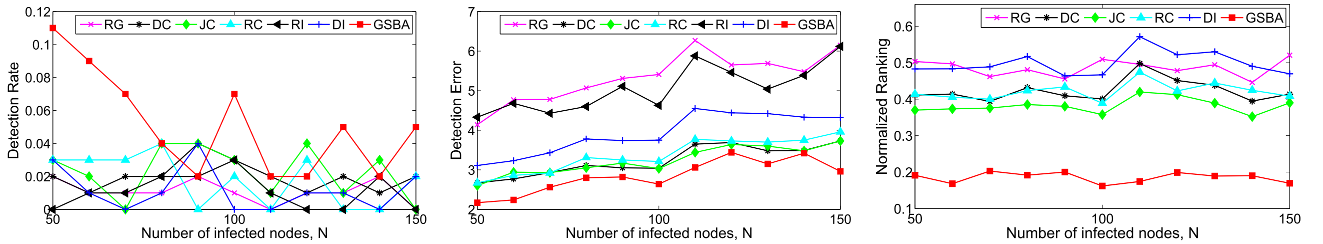

Fig. 5. Random test performance on Power-Grid Network with $\lambda = 0.2$.

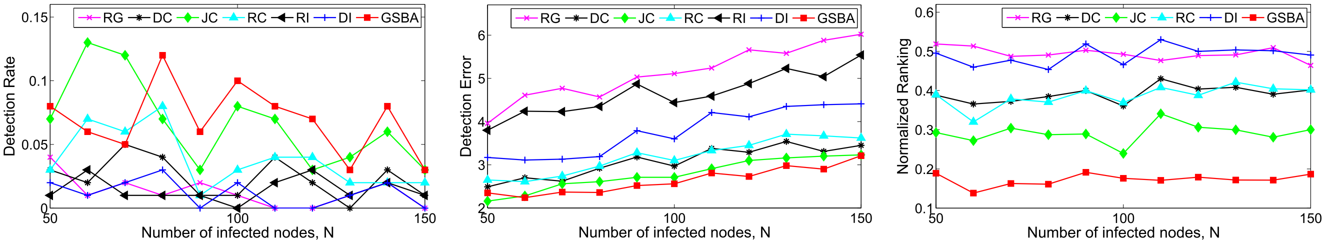

Fig. 6. Random test performance on Power-Grid Network with $\lambda = 0.5$.

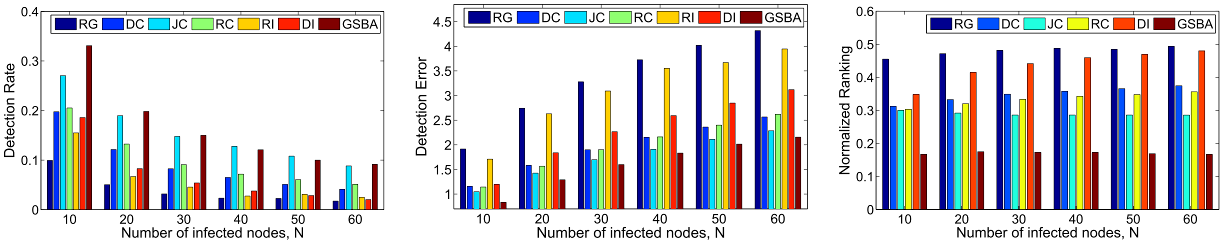

Fig. 7. Full test performance on Power-Grid Network with $\lambda = 0.5$.

C. Results on Scale-Free Network

In this subsection, we conduct experiments under three settings, namely random test with $\lambda = 0.2$ or $\lambda = 0.5$, and full test with $\lambda = 0.5$. For random test, $N \in [5, 10]$, while $N \in [10, 60]$ for full test. The results are shown in Fig 2, 3 and 4. We can have the following observations.

First, for all tests, our methods (GSBA and BFSA) achieve better performance than *Rumor Centrality* (RC) under all the three measures. The reason is that our methods consider the uninfected nodes to infer the probabilities of permitted permutations for general graphs. This proves once more that uninfected nodes are also helpful for detecting the source.

Second, Fig 2 and 3 clearly show the performance of DMP, GSBA and BFSA are nearly same with respect to *Detection Rate* and *Detection Error*, and all of them outperform other methods no matter λ equals to 0.2 or 0.5. But our methods achieve smaller Normalized Ranking than DMP. This means GSBA and BFSA can sort the real source more accurately. More importantly, GSBA is far more efficient than DMP and BFSA. The above indicates the feasibility of maximum a posteriori upper-bound and the greedy search strategy in Section V-B. In other words, after selecting rumor centrality as the prior, determining the likelihood like Section IV can improve the detection performance drastically.

Third, comparing the same measures in Fig 2 and 3, we found that different parameters of the information diffusion model apparently affect the performance of all methods. For GSBA and BFSA, the detection performance are better on scale-free networks with larger λ. The reason may be that scale-free networks with larger λ are more compact when we fix the number of infected nodes.

Fourth, under the strategy of full test, the performance of all methods change more regularly than ones under random test. For example, as the number of infected nodes increases, *Detection Rate* and *Detection Error* of all methods shown in Fig 4 have a tendency of becoming worse.

To sum up, our methods, GSBA and BFSA, can achieve nearly the same performance with DMP, but GSBA is far more efficient. These results prove the feasibility of our estimation approximations in Section V.

D. Results on Power-Grid Network

In the following, we want to display the experimental results for larger sizes of infected subgraphs. But as mentioned before, DMP, BFSA and NS are time-consuming, and BFSA behaves similarly to DMP on scale-free networks. Therefore, we only compare RG, DC, JC, RC, RI and DI with GSBA on Power-Grid and Wiki-Vote in random test, like in full test.

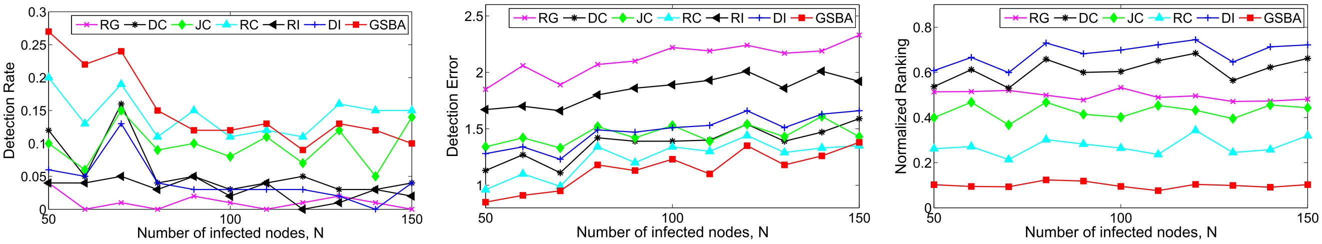

Fig. 8. Random test performance on Wiki-Vote Network with $\lambda = 0.2$.

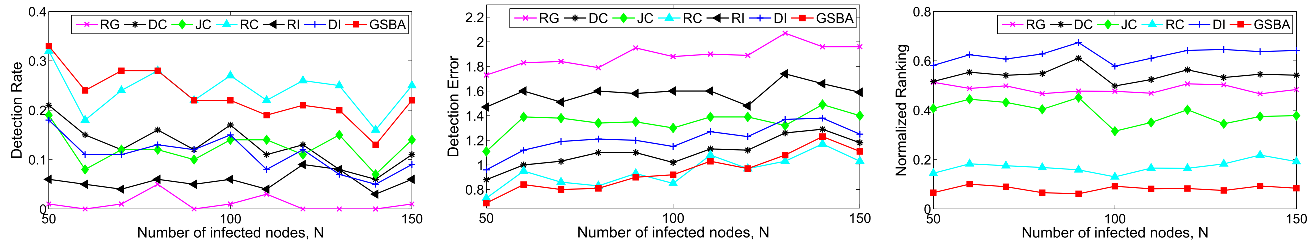

Fig. 9. Random test performance on Wiki-Vote Network with $\lambda = 0.5$.

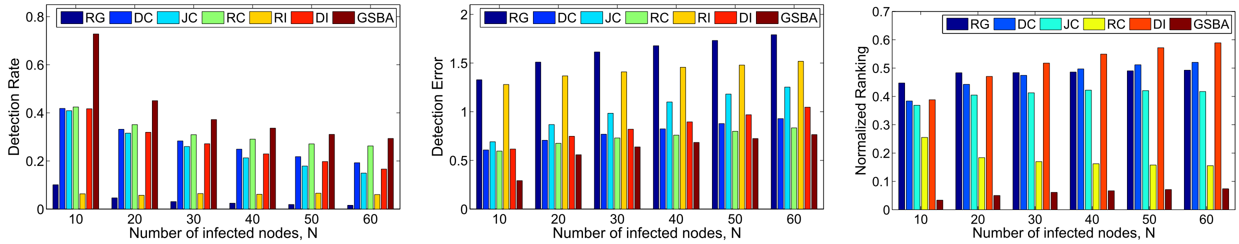

Fig. 10. Full test performance on Wiki-Vote Network with $\lambda = 0.5$.

For random test, $N \in [50, 150]$, while $N \in [10, 60]$ for full test. Fig. 5, 6 and 7 show the results on Power-Grid. We can get some interesting findings.

First, generally speaking, the results of random test and full test keep consistent. For example, our method, GSBA, always achieve the best performance of *Detection Error* and *Normalized Ranking*, for all situations shown in Fig. 5, 6 and 7. But *Detection Rate* of full test in Fig. 7 is much more smooth than those of random test in Fig. 5 and 6.

Second, both of *Detection Error* and *Normalized Ranking* of each method display an obvious regularity, under random test or full test. For example, *Detection Error* smoothly becomes larger as the number of infected nodes increases, such as in the middle figure of Fig. 7. While *Normalized Ranking* behaves very steadily and basically keeps invariant, which is an important property. It can be used to measure the ability of a method ranking the real source as higher as possible. GSBA has shown its superiority of *Normalized Ranking* under both of random test and full test.

Third, Fig. 7 shows that Jordan Centrality (JC) behaves similarly to GSBA on *Detection Rate* under full test with $\lambda = 0.5$, but GSBA achieves better performance of *Detection Error* and *Normalized Ranking* apparently.

To summarize, GSBA always achieves the best performance on the Power-Grid network, especially with respect to *Detection Error* and *Normalized Ranking*. For full test, Jordan Centrality and GSBA have similar *Detection Rate*.

E. Results on Wiki-Vote Network

To compare the methods on Wiki-Vote, we apply the same setting as experiments on Power-Grid. The results are shown in Fig. 8, 9 and 10. Obviously, we can draw some similar conclusions as the last subsection, such as the superiority of GSBA, the consistency of random test and full test, the increasing tendency of *Detection Error* and the invariance of *Normalized Ranking* with the number of infected nodes becoming larger. Except that, the performance of Rumor Centrality (RC) and Distance Centrality (DC) are better and more similar to GSBA's than Jordan Centrality's. Let us take $N = 50$ and $\lambda = 0.5$ as an example. For Power-Grid, the average diameter of infected subgraphs G_I is 9.5, and the average ratio of edges to nodes in G_I is 1.2. While for Wiki-Vote, the average diameter and ratio are 2.7 and 4.7, respectively. Indeed, the ratio of a tree is less than 1. Therefore, the infected subgraphs of Power-Grid are more tree-like. This may explain why RC and DC perform better than JC. In other words, GSBA $>$ RC $>$ DC $>$ JC on graphs less like trees, where $>$ means performing better. Besides, *Normalized Ranking* of GSBA is extremely less than other methods'.

In conclusion, the above experiments validate the effectiveness of our methods, especially when measured by *Normalized Ranking*. These also proves the feasibility of likelihood approximation like GSBA.

VII. Conclusion

In this paper, we revisited the problem of information source detection from the perspective of likelihood approximation. After deriving the Maximum A Posteriori (MAP) estimator, we design two approximation approaches, namely *Brute Force Search Approximation* (BFSA) and *Greedy Search Bound Approximation* (GSBA), to improve the efficiency. Experiments on several networks clearly show the superiority of our methods and that GSBA is nearly as effective as BFSA, but far more efficient.

Two directions are worth exploring as further study. First, we have so far derived our methods to detect the single information source under the Susceptible-Infected (SI) model. It is interesting to extend them to multiple information sources detection under other models, such as SIR. The second direction is to explore other more effective approaches to find the upper bound of likelihood, instead of greedy search.

Acknowledgment

This work was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010) and the Natural Science Foundation of China (Grant No. 61403358). It was also partially supported by the Pinnacle Lab at Singapore Management University and The DS LEE Fellowship.

References

[1] F. Altarelli, A. Braunstein, L. DallAsta, A. Ingrosso, and R. Zecchina, "The patient-zero problem with noisy observations," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2014, no. 10, p. P10016, 2014.

[2] F. Altarelli, A. Braunstein, L. DallAsta, A. Lage-Castellanos, and R. Zecchina, "Bayesian inference of epidemics on networks via belief propagation," *Physical Review Letters*, vol. 112, no. 11, p. 118701, 2014.

[3] N. T. Bailey, *The mathematical theory of infectious diseases and its applications*. Charles Griffin & Company Ltd, 5a Crendon Street, High Wycombe, Bucks HP13 6LE., 1975.

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

[5] C. M. Bishop *et al.*, *Pattern recognition and machine learning*. springer New York, 2006, vol. 4, no. 4.

[6] 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.

[7] W. Dong, W. Zhang, and C. W. Tan, "Rooting out the rumor culprit from suspects," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*. IEEE, 2013, pp. 2671–2675.

[8] Z. Feng, P. Gundecha, and H. Liu, "Recovering information recipients in social media via provenance," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ACM, 2013, pp. 706–711.

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

[10] 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.

[11] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, pp. 1420–1443, 1978.

[12] P. Gundecha, Z. Feng, and H. Liu, "Seeking provenance of information using social media," in *Proceedings of the 22nd ACM international Conference on Information & Knowledge Management*. ACM, 2013, pp. 1691–1696.

[13] B. Heap, "Permutations by interchanges," *The Computer Journal*, vol. 6, no. 3, pp. 293–298, 1963.

[14] J. L. Iribarren and E. Moro, "Branching dynamics of viral information spreading," *Physical Review E*, vol. 84, no. 4, p. 046116, 2011.

[15] C. Jordan, "Sur les assemblages de lignes," *J. Reine Angew. Math*, vol. 70, no. 185, p. 81, 1869.

[16] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2003, pp. 137–146.

[17] W. Kermack and A. McKendrick, "Contributions to the mathematical theory of epidemicsii. the problem of endemicity," *Bulletin of mathematical biology*, vol. 53, no. 1, pp. 57–87, 1991.

[18] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, "Finding effectors in social networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2010, pp. 1059–1068.

[19] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th international conference on World Wide Web*. ACM, 2010, pp. 641–650.

[20] J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.

[21] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and J. X. Yu, "Influence maximization over large-scale social networks: A bounded linear approach," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, ser. CIKM '14. New York, NY, USA: ACM, 2014, pp. 171–180.

[22] 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.

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

[24] D. T. Nguyen, N. P. Nguyen, and M. T. Thai, "Sources of misinformation in online social networks: who to suspect," in *Military Communications Conference, MILCOM*, 2012, pp. 1–6.

[25] 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.

[26] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones?" in *ICDM*, vol. 12, 2012, pp. 11–20.

[27] J. G. Restrepo, E. Ott, and B. R. Hunt, "Characterizing the dynamical importance of network nodes and links," *Physical Review Letters*, vol. 97, no. 9, p. 094102, 2006.

[28] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: theory and experiment," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 38, no. 1. ACM, 2010, pp. 203–214.

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

[30] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rooting our rumor sources in online social networks: The value of diversity from multiple observations," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 9, no. 4, pp. 663–677, June 2015.

[31] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[32] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[33] K. Zhu and L. Ying, "Information source detection in the sir model: A sample path based approach," in *Information Theory and Applications Workshop (ITA), 2013*. IEEE, 2013, pp. 1–9.