Information Source Localization with Protector Diffusion in Networks

Jaeyoung Choi, Sangwoo Moon, Jinwoo Shin and Yung Yi

Abstract: **Recently, the problem of detecting the information source in a network has been much studied, where it has been shown that the detection probability cannot be beyond 31% even for regular trees if the number of infected nodes is sufficiently large. In this paper, we study the impact of an anti-information spreading on the original information source detection. We first show a negative result: the anti-information diffusion does not increase the detection probability under Maximum-Likelihood-Estimator (MLE) when the number of infected nodes are sufficiently large by *passive diffusion* that the anti-information starts to be spread by a special node, called the *protector,* after is reached by the original information. We next consider the case when the distance between the information source and the protector follows a certain type of distribution, but its parameter is hidden. Then, we propose the following learning algorithm: a) learn the distance distribution parameters under MLE, and b) detect the information source under Maximum-A-Posterior-Estimator (MAPE) based on the learnt parameters. We provide an analytic characterization of the source detection probability for regular trees under the proposed algorithm, where MAPE outperforms MLE by up to 50% for 3-regular trees and by up to 63% when the degree of the regular tree becomes large. We demonstrate our theoretical findings through numerical results, and further present the simulation results for general topologies (e.g., Facebook and US power grid networks) even without knowledge of the distance distribution, showing that under a simple protector placement algorithm, MAPE produces the detection probability much larger than that by MLE.**

Index Terms: **Information Source Localization, Epidemic diffusion model, Maximum Likelihood Estimator.**

I. INTRODUCTION

INFORMATION spread is universal in many types of online/offline and social/physical networks. Examples include the propagation of infectious diseases, the technology diffusion, the computer virus/spam infection in the Internet, and tweeting and retweeting of popular topics. Finding the source in those information spreads is one of the indispensable and useful tasks, arising in many different contexts, e.g., detecting a malicious agent, a patient zero, or an influential person, because pre-action can be taken by some authorities to limit the possible damages due to spreading of such diffused objects that are harmful, if spread in an uncontrolled manner. Since the seminal work by Shah and Zaman [1], extensive research efforts have been made [2–4], where the main focus has been on how to design an estimator and provide theoretical (positive and negative) limits on the detection performance. However, for example, it is shown [1] that in the regular tree topologies, the detection probability cannot be above 31% under Maximum Likelihood Estimator (MLE), and even worse, in other realistic topologies such as power grid graphs, scale-free graphs and Internet Autonomous System (AS) graphs, the detection probability is less than 5% under a MLE-based heuristic.

In this paper, our interest lies in how much detection performance can be improved by installing hidden agents, called *protectors* that spread "anti-information." The role of these protectors is to spread the information against the original one, vaccinate humans against infectious disease, or install security updates against computer virus. Intuitively, the existence of protectors and their infection with anti-information seem beneficial in detecting the original source, because they both block the original information spread and the snapshot of both protected and infected nodes, compared to that of only infected nodes, discloses more information to the detector. However, understanding which nodes should be estimated to be the information source and quantifying the detection performance in presence of protectors is far from trivial. In this paper, we assume that initially there exists a single information source and protector, where the the anti-information source responds passively in the sense that it is initially dormant and becomes active and starts to infect other nodes (with anti-one) only when the original one reaches itself.

Our main contributions are summarized in what follows:

1) First, we show that under MLE, the protector's anti-information spread does not improve the detection probability in regular trees under the passive diffusion. However, we show that this is not the case if some statistical feature on the distance between the two information sources is given. In particular, we assume that the distance distribution is of a specific type, where their parameters are *unknown.* In practice, the parameters can be learnt using certain prior records on the information source or the diffusion snapshot. If such prior records do not exist, one can use a learning algorithm such as MLE to estimate the parameters (see Section III for more details). We study three example distance distributions, Zipf, Geometric or Poisson, where the probability that the protector is located decays with distance in all distributions, but their decaying patterns are different.
2) Second, for a given quality in estimating the distribution parameters, we quantify how much the detection probability increases in regular trees under Maximum-A-Posterior-Estimator (MAPE) due to the usage of the protector's anti-information spreads. In particular, we show that the differ-

Part of this work was presented at the International Conference on Network Protocols (ICNP) Workshop on Machine Learning (NetworkML).

The authors are with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), email: email:{jychoi14, mununum, jinwoos, yiyung}@kaist.ac.kr.

Yung Yi is the corresponding author.

Digital Object Identifier xx.xxxx/JCN.2017.xxxxx

ence of the detection probabilities between MLE and MAPE is up to 50% for the 3-regular tree and up to 63% for the regular tree with infinite degree. This implies that if the protector is appropriately placed around the information source, the detection probability significantly increases.

3) Finally, we design a MAPE-based heuristic for general topologies such as *Erdös-Rényi* (ER) graph, small world graph, scale-free graph, as well as a Facebook ego network and a US power grid network, where we observe that the prior information based on the protector significantly helps to detect the information source.

Thus, we conclude that utilizing the anti-information is a simple way of detecting the original source better, where in literature several different approaches have been considered for a similar purpose, *e.g.*multiple observations [2], suspect set [3] and Jordan center-based [4] methods. We believe that ours shed new lights on this area, being of broad interest in the future.

Related work. The research on information source detection has recently received significant attentions. The first theoretical approach was done by Shah and Zaman [1, 5, 6] and they introduced the metric called *rumor centrality*, which is a simple topology-dependent metric. They proved that the rumor[1] centrality describes the likelihood function when the underlying network is a regular tree and the diffusion follows the SI (Susceptible-Infected) model, which is extended to a random graph network in [6]. Zhu and Ying [4] solved the rumor source detection problem under the SIR (Susceptible-Infected-Removed) model and took a sample path approach to solve the problem, where a notion of *Jordan center* was introduced, being extended to the case of sparse observations [12]. The authors [14], [16] studied the problem of estimating the source for random growing trees, where unlike aforementioned papers, they did not assume an underlying network structure. The authors in [20] inferred the historical diffusion traces and identifies the diffusion source from partially observed cascades, and similarly in [22], partial diffusion information is utilized. Recently, there has been some approaches for the general graphs in [21, 23] to find the information source of epidemic. All the detection mechanisms so far correspond to point estimators, whose detection performance tends to be low. There was several attempts to boost up the detection probability. Wang *et al.* [2] showed that observing multiple different epidemic instances can significantly increase the the detection probability. Dong *et al.* [3] assumed that there exist a restricted set of source candidates, where they showed the increased detection probability based on the MAPE (maximum a posterior estimator). The authors in [17, 28] introduced the notion of *set estimation* and provide the analytical results on the detection performance. Choi *et al.* [25, 28] studied the effects of querying to finding the source and they showed that how many queries are sufficient to achieve a target detection probability. Opposite to finding the source, [18, 19] consider the problem of hiding the source by introducing a new diffusion mechanism called *adaptive diffusion* from the source. The authors in [29] proposed a game model for both perspective of seeking and hiding the source.

[1]In this paper, the terms “rumor" and “information" are used interchangeably.

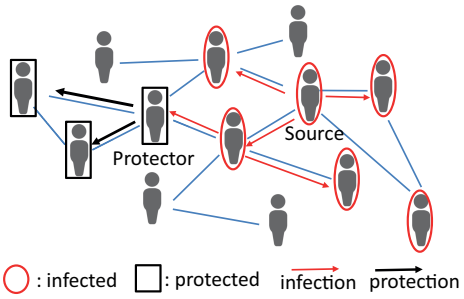


Fig. 1. Illustrative example of information spreading model for the existing of passive protector.

II. SYSTEM MODEL

In this section, we introduce a diffusion model of spreading two information and two estimators for finding the original information source as well. In addition to this, we review some prior works which are considered the source detection probability based on these estimators but different models.

A. Information Spreading Model

We consider an undirected graph $G = (V, E)$, where V is a set of nodes and E is the set of edges of the form (i, j) for $i, j \in V$. Each node represents an individual in human social networks or a computer host in the Internet, and each edge corresponds to a social relationship between two individuals or a physical connection between two Internet hosts. As in other works, e.g., [1], we assume a countably infinite set of nodes for avoiding the boundary effects.

There exist two spreading sources: an *information source* and a *protector*, which we denote by $v^*, p^* \in V$, respectively. The information source is the staring node which spreads an information such as rumor, and the protector corresponds to a node which spreads an “anti-information”, e.g., an anti-virus for virus spreading and a true fact for feigned information spreading. We consider the case when the protector is *passive* in the sense that the protector source is initially dormant, but becomes active and starts to infect its neighboring node only when the original information reaches it (see Fig. 1). As a model of spreading information and anti-information, we consider a variant of SI (Susceptible-Infected) model that each node is one of the following three states: *susceptible*, *infected*, or *protected*, where all nodes are initialized to be susceptible except the initially-infected information source v^* and the initially-protected protector p^*. Once a node i has an information (or an anti-information), it is able to spread it to another susceptible node j if and only if there is an edge between them, i.e., $(i, j) \in E$. We assume that once a node becomes either *infected* or *protected*, it does not change its state, as in the classical SI model. For each edge $(i, j) \in E$, let a random variable τ_{ij} be the time it takes for susceptible node j to receive the information (irrespective of being any two information) from non-susceptible node i. We assume τ_{ij} is exponentially distributed with rate $\lambda > 0$ independently with everything else. [2] Without

[2]This assumption omits the case that a susceptible node hears both information at the same time.

loss of generality, we assume that $\lambda = 1$.

B. Source Estimators: MLE and MAPE

MLE and MAPE. Let I_N and P_M be the sets of infected and protected nodes, respectively, when one intends to detect the information source. Here, the subscripts N and M are used to express the number of infected and protected nodes, respectively. To estimate the source v^*, we consider the following two popular estimators, MLE and MAPE:

$$\begin{aligned} v_{\texttt{ml}} &= \arg\max_{v\in I_N} P(I_N, P_M|v, p^*), \\ v_{\texttt{map}} &= \arg\max_{v\in I_N} P(v|I_N, P_M, p^*), \end{aligned} \tag{1}$$

where we assume that algorithms have the knowledge of the protector p^*. Note that the relation between MLE and MAPE can be explained by:

$$\begin{aligned} v_{\texttt{map}} &= \arg\max_{v\in I_N} P(v|I_N, P_M, p^*) \\ &\overset{(a)}{=} \arg\max_{v\in I_N} \frac{P(I_N, P_M|v, p^*)P(v, p^*)}{P(I_N, P_N, p^*)} \\ &= \arg\max_{v\in I_N} P(I_N, P_M|v, p^*)\cdot P(v, p^*), \end{aligned}$$

where (a) is from the Bayes' rule and $P(I_N, P_M|v, p^*)$ is the probability that the realizations I_N and P_M occur, given an information source v and the protector p^*. Therefore, MLE is equivalent to MAPE if $P(v, p^*)$ is assumed to be uniform over $v \in V$.

To further characterize two estimators, let $\sigma = (\omega_1 = v^*, \omega_2, ..., \omega_{M+N})$ be an infection sequence (also called a sample path) resulting in I_N, P_M, where the source v^* generates the information first, and all other nodes in the sequence $\omega_2, ..., \omega_{M+N}$ are arranged in ascending order of their propagation times. Then, we have

$$P(I_N, P_M|v, p^*) = \sum_{\sigma\in\Omega(v,p^*,I_N,P_M)} P(\sigma|v, p^*), \tag{2}$$

where $\Omega(v, p^*, I_N, P_M)$ be the set of all possible propagation sequences given I_N, P_M. Then, under regular tree G, one can follow the same approach as that in [1] and characterize MLE and MAPE based on the number of possible propagation sequences, i.e.,

$$v_{\texttt{ml}} = \arg\max_{v\in I_N} R(v, p^*, I_N, P_M), \tag{3}$$

$$v_{\texttt{map}} = \arg\max_{v\in I_N} R(v, p^*, I_N, P_M)\cdot P(v, p^*). \tag{4}$$

where

$$\begin{aligned} R(v, p^*, I_N, P_M) &= |\Omega(v, p^*, I_N, P_M)| \\ &= (M+N)! \prod_{u\in I_N\cup P_M} |T_u^v|^{-1}. \end{aligned} \tag{5}$$

We call $v_{\texttt{ml}}$ or $v_{\texttt{map}}$ a *protected center,* where unless confusion arises, we omit the name of a particular estimator. In the above, we let $|T_u^v|$ be the number of nodes in the subtree T_u^v rooted at node u when v is the information source. Then, the number of possible propagation sequences from v can be obtained by computing the product of the subtree size. One can compute $R(\cdot)$ for every infected node $v \in I_N$ in $O(M+N)$ time using a similar message passing algorithm to that in [1].

Detection Probability. We let C_{M+N} be the event of detecting the information source using a given estimator when there are $M+N$ infected and protected nodes in the graph, we are interested in the asymptotic case, *i.e.*,

$$\lim_{M+N\to\infty} P(C_{M+N}).^{3}$$

We denote by $\pi_d^{\texttt{ml}}$ and $\pi_d^{\texttt{map}}$ the detection probabilities of MLE and MAPE for d-regular tree, respectively. In addition, we use just π_d to refer to that of MLE without protected nodes (*i.e.*, $M = 0$) for a comparative purpose, where the following formula is known for π_d [6].

Lemma 1: ([6]) Under d-regular tree G,

$$\pi_d = \begin{cases} 0 & \text{if } d = 2, \\ 1 - d\left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right) & \text{if } d \geq 3, \end{cases}$$

where $I_x(\alpha, \beta)$ is the incomplete Beta function[4] with parameters α and β.

Using the above lemma, one can easily check that the detection probability for MLE without protector is at most 0.307 in the asymptotic case.

III. Main Result: Detection Probability

A. Maximum Likelihood Estimator

In this section, we provide the performance of MLE for detecting the information source in presence of opposite information, *i.e.*, protector, under regular trees.

Theorem 1: Under d-regular tree G,

$$\pi_d^{\texttt{ml}} = \pi_d \qquad \text{for all } d \geq 2.$$

The proof of Theorem 1 will be presented in Section IV-A. The implication is as follows. From the assumption for the diffusion rate $\lambda = 1$, we consider two cases. The first case is that the protector does not receive the information and second case is that it receives the information. To see this, let $N(t)$ be the number of infected nodes at time t on the path between two sources then it follows a Poisson process with rate λ and if the time goes to infinity, we have $\lim_{t\to\infty} \frac{N(t)}{t} = \lambda$. Hence, if $\lambda > 1$ then we $N(t)$ is larger than t which means the protector receives the information for sufficient large t and vise versa. However, for $\lambda = 1$, we need to consider both of cases. From the assumption that the network size is infinite and the protector is already in the network and the information is generated uniformly at random, the probability that the information source

[3]Note that since the distance between two sources are bounded by $L < \infty$, $M + N \to \infty$ implies that $M \to \infty$ and $N \to \infty$.

[4]The incomplete Beta function $I_x(\alpha, \beta)$ is the probability that a Beta random variable with parameters α and β is less than $x \in [0, 1]$, whose form is $I_x(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^x t^{\alpha-1}(1-t)^{\beta-1}dt$ where $\Gamma(\cdot)$ is the standard Gamma function [6].

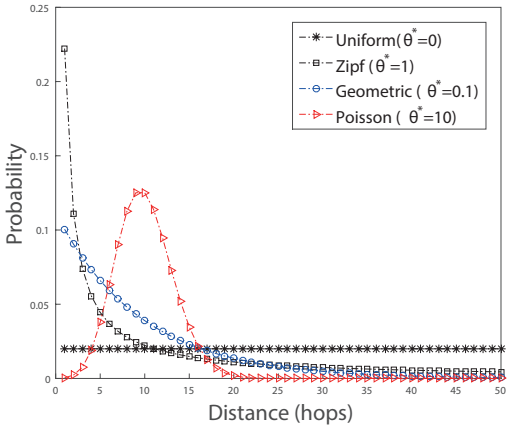


Fig. 2. Examples of distance distributions between two sources with L=50.

is located within the finite distance from the protector is zero. Hence, we see that the distance of two sources is infinite with probability one and this implies that the information does not reach to the protector. Therefore, if we use MLE to find the information source, there is no gain due to there can not see any diffusion of anti-information. As a second case, we consider the information reaches to the protector and there are some anti-information diffusion in the network. However, this theorem implies that the existence of anti-information does *not* improve the detection performance when there are sufficiently large infected and protected nodes in the graph. This seems somewhat counter-intuitive, because the diffusion of opposite information may provide a side information so as to enable better detection. This negative result can be explained for the following reasons. Depending on the distance between the protector and the original information source, two cases can be considered. First, when I_N is much larger than P_M, the MLE is highly likely to be equal to the original rumor center (without anti-information), resulting in the same detection probability. Second, however, when P_M is larger than I_N, the MLE is highly likely to be located in P_M, which leads MLE to estimate a border node between I_N and P_M (because the original information source should be in I_N), but the number of such border nodes is negligible (in fact, there exists a single border node in tree topologies), when N and M goes to infinity.

B. Maximum A Posteriori Estimator

From the result of MLE, we see that if there is no further information about the source, the detection can not be improved. Hence, in this section, we consider the case that a prior information[5] for the source node is given and then we use MAPE to find the source using this information. As such an information, we consider a distance between two sources as follows.

Distance Distribution. For computing MAPE, one has to know the probability $P(v, p^*)$ where v is the information source. To this end, we assume that the distance between v and p^* is a random variable following a specific distribution. In this paper, we consider three distributions: 'L-truncated' Zipf, Geometric or Poisson, where L is a non-negative integer constant, *i.e.*, for $1 \le l \le L$,

$$P\left(d(v,p^*)=l\right) \propto \begin{cases} 1/l^{\theta} & \text{for Zipf } (\theta \ge 0), \\ \theta(1-\theta)^{l-1} & \text{for Geometric } (0<\theta\le 1) \\ \theta^l e^{-\theta}/l! & \text{for Poisson } (\theta \ge 0), \end{cases} \tag{6}$$

and $P(d(v,p^*) = l) = 0$ for $l > L$.[6] (See Fig. 2) The main reason why we study these three distributions is because higher probabilities are assigned to nearer the original information source from the protector under them and these distributions will give how the *distance information* as the *priori* knowledge effect to detect the original information source. Hence, our goal is to *quantify* the improvement of the source detection probability for using these distance information. Nevertheless, our analytical results can be easily extended to other distributions. We also remark that it is reported the distance of two nodes follows Zipf distribution in some social networks [8–10].[7] Throughout this paper, we commonly use θ to mean the parameter of the distance distribution, where the true parameter $\theta = \theta^*$ might be unknown a priori and one has to run MAPE with an estimated parameter $\theta = \hat{\theta}$. As we mentioned before, since there is no enhancement for detecting the information source by MLE when the number of diffused nodes goes to infinity, we address *how much* it can increase the detection probability when there is a distance information of two sources as a priori. Hence, in this subsection, we provide the performance of MAPE for detecting the original information source in presence of anti-information under regular trees. It turns out that obtaining the exact formula of MAPE's detection probability, as in Lemma 1 in absence of protector, is technically challenging. However, we will provide a characterization of the lower bound of the detection probability with protector, as stated in Theorem 2, even when the *unknown* parameter of the distance distribution is not exactly equal to the *true* parameter θ^*.

Theorem 2: Let $\pi_d^{\mathtt{map}}(\hat{\theta})$ be the detection probability of MAPE for the learnt parameter $\hat{\theta}$ and the true parameter θ^*. Then for d-regular trees it follows that

$$\pi_d^{\mathtt{map}}(\hat{\theta}) - \pi_d \;\ge\; \left(p(\theta^*) - 1/2\right)^{\frac{2d-3}{3(d-2)}} \tag{7}$$

$$- 6|1 - p(\hat{\theta})/p(\theta^*)||\theta^* - \hat{\theta}|, \tag{8}$$

where

$$p(\theta) = \begin{cases} \frac{2^\theta}{2^\theta+1} & \text{for} \quad \text{Zipf}(\theta), \quad \theta \ge 0, \\ \frac{1}{2-\theta} & \text{for} \quad \text{Geometric}(\theta), \quad 0 \le \theta \le 1, \\ \frac{2+\theta}{2+2\theta} & \text{for} \quad \text{Poisson}(\theta), \quad \theta \ge 0. \end{cases} \tag{9}$$

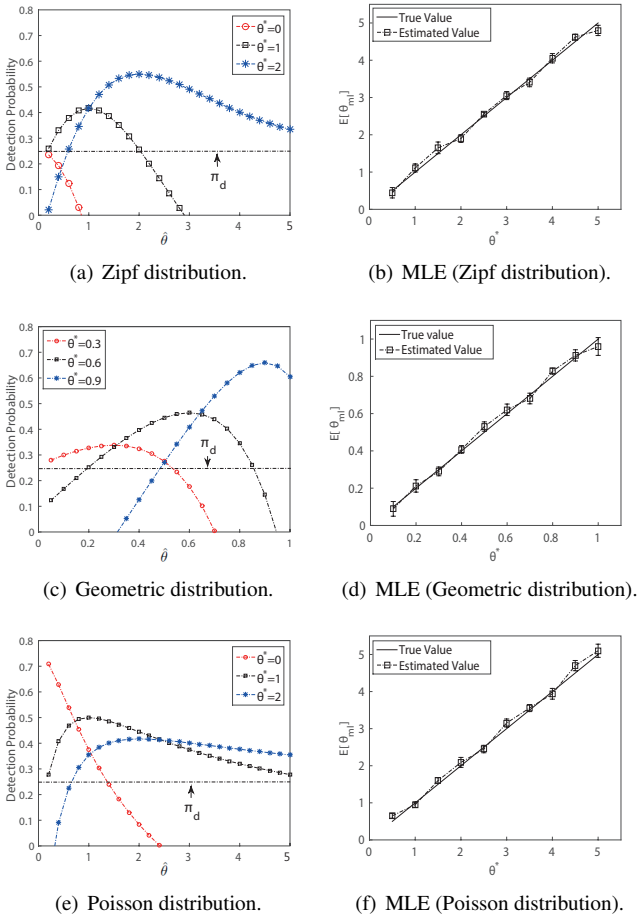
A few interpretations of Theorem 2 are in order.

(a) Theorem 2 states that depending on how well we learn the true parameter θ^*, the detection probability $\pi_d^{\mathtt{map}}$ is determined. In other words, $\hat{\theta}$ is far from θ^*, $\pi_d^{\mathtt{map}}$ may be lower than π_d.

[5] This can be obtained from some historical information of the sources or underlying graph topologies.

[6] In this paper, we consider sufficiently large L but finite.

[7] In practice, it may be not guaranteed that the distance of two sources is close. However, in this paper, we focus that how much the detection probability by MAP estimator will be increased if we have this information as a prior.



(a) Zipf distribution. (b) MLE (Zipf distribution).

(c) Geometric distribution. (d) MLE (Geometric distribution).

(e) Poisson distribution. (f) MLE (Poisson distribution).

Fig. 3. Theoretical results of Theorem 2 ((a), (c), (e)) and learning the parameter of each distributions by MLE ((b), (d), (f)) in Algorithm 1 for $d = 3$ and $L = 50$, respectively (We generate 100 random diffusion snapshots up to $M + N = 500$, and plot the average value of the estimated parameters with 95 percent confidence interval for the simulations) and $\pi_d = 0.25$ for $d = 3$.

(b) In fact, we see that the condition that the protector helps in detecting the information source, *i.e.*, $\pi_d^{\texttt{map}}(\hat{\theta}) - \pi_d \geq 0$, when the following condition holds:

$$|\theta^* - \hat{\theta}| \leq \frac{(p(\theta^*) - 1/2)^{\frac{2d-3}{3(d-2)}}}{6|1 - p(\hat{\theta})/p(\theta^*)|}. \tag{10}$$

For example, consider a Zipf distribution with $\theta^* = 1$ for 3-regular tree (*i.e.*, $d = 3$). Then, (10) holds if $|\theta^* - \hat{\theta}| \leq 0.8$, and the condition for other cases are similarly mild, where as we will present in Section III-C, $\hat{\theta}$ can be easily learnt with a high-accuracy and low-cost parameter learning algorithm.

(c) To roughly quantify our analytical result, if the distribution parameter estimation is almost perfect, *i.e.*, $\hat{\theta} \approx \theta^*$, then $\pi_d^{\texttt{map}}(\hat{\theta}) - \pi_d \gtrsim (p(\theta^*) - 1/2)^{\frac{2d-3}{3(d-2)}}$. The value of $p(\theta^*)$ ranges in all three distributions as: $1/2 \leq p(\theta^*) \leq 1$. Thus, the detection performance *gap* from MLE without protector is up to 50% for $d = 3$ and up to 63% for $d \to \infty$. This gap will reduce slightly, depending on the value of true parameter θ^*.

We obtain the numerical result of Theorem 2 for the three distributions as in Fig 3 ((a),(c),(e)). As an example, we consider the three true parameter *i.e.*, $\theta^* = 0, 1, 2$ for Zipf distribution and change the learning parameter from $\hat{\theta} = 0$ to $\hat{\theta} = 5$. We see that if $\theta^* = 0$ (Uniform distribution) then there is no gain of detection probability for any learnt parameter $\hat{\theta}$ due to lack of any distance information of the two sources. However, if $\theta^* > 0$ then there exists quite enhancement of the detection probability from learning the parameter appropriately. The results of other distributions are similar as in Fig 3.

C. Learning θ^*

In practice, there is no knowledge of the true parameter θ^* as a priori. In this case, one can estimate it using prior records of information sources or the following MLE simply using the current 'snapshot':

$$\begin{aligned}
\theta_{\texttt{ml}} &= \arg\max_{\theta} P(I_N, P_M, p^*|\theta) \\
&= \arg\max_{\theta} \sum_{l=1}^{L} P(I_N, P_M, p^*|d(v^*, p^*) = l) P(d(v^*, p^*) = l|\theta) \\
&\overset{(a)}{=} \arg\max_{\theta} \sum_{l=1}^{L} \left(\sum_{k=1}^{|V_l|} P(I_N, P_M, p^*|v_{l,k} = v^*) \right) \\
&\qquad \cdot P(d(v^*, p^*) = l|\theta) \\
&\overset{(b)}{=} \arg\max_{\theta} \sum_{l=1}^{L} R(V_l) P(d(v^*, p^*) = l|\theta),
\end{aligned} \tag{11}$$

where $v_{l,k}$ is the k-th infected nodes at the distance l and V_l is the set of these nodes for $0 \leq k \leq |V_l|$. Then, (a) is from the fact that $d(v, p^*) = l$ which implies that $v \in V_l$. The equality (b) is from the fact that for each $v_{l,k} \in V_l$, $P(I_N, P_M, p^*|v_{l,k}) \propto R(v_{l,k}, p^*, I_N, P_M)$ as in (3) where $R(V_l) = \sum_{k=1}^{|V_l|} R(v_{l,k}, p^*, I_N, P_M)$. Since $R(V_l)$ can be obtained from the snapshot and $P(d(v, p^*) = l|\theta)$ is determined when the distribution is given, MLE is obtained by solving the optimization problem (11). To do this, let $f(\theta) := \sum_{l=1}^{L} R(V_l) P(d(v^*, p^*) = l|\theta)$ then we see that this function does not guarantee the concavity in terms of θ (thus not a convex program), but from the monotonicity and differentiability of $P(d(v^*, p^*) = l|\theta)$, it is easy to check that the function $f(\theta)$ is a differentiable unimodular function[8]. Thus, we can apply a popular algorithm for maximizing a unimodular function [13] to solve (11) as in Algorithm 1 where $\varepsilon > 0$ is the termination constraint. One can easily check that the algorithm is terminated in polynomial time of $M + N$ and $1/\varepsilon$.

Fig. 3 ((b),(d),(f)) show numerical results on the performance of learning θ^* for various values of θ^* in three distributions. For the graphs, we consider the total number of diffused nodes $M + N = 500$ under the 3-regular tree ($d = 3$) and we generate 100 random diffusion snapshots, and plot the average value of the estimated parameters with 95 percent confidence interval. Our

[8] $f(\theta)$ is differentiable unimodular $\partial f(\theta)/\partial\theta > 0$ for one side of some θ and $\partial f(\theta)/\partial\theta < 0$ for the other side.

Algorithm 1 Maximum Likelihood Estimation (MLE) of θ^* for Regular Trees

Input: $(I_N, P_M, d, L, \theta_{min}, \theta_{max}, \varepsilon, P(v, p^*))$
for $v \in I_N$ **do**
  Compute $R(v, p^*, I_N, P_M)$ by a message passing algorithm [1] and obtain $d(v, p^*)$ by a shortest path algorithm;
  $R(V_l) \leftarrow 0$; $(1 \le l \le L)$
  if $d(v, p^*) = l$ **then**
    $R(V_l) \leftarrow R(V_l) + R(v, p^*, I_N, P_M)$;
  end if
end for
Set $f(\theta) = \sum_{l=1}^{L} R(V_l) P(d(v, p^*) = l|\theta)$;
$\theta^{new} \leftarrow \frac{\theta_{min}+\theta_{max}}{2}$; (initialize)
while $|\nabla f(\theta^{new})| \ge \varepsilon$ **do**
  Use Brent method [13] to find the root of $\nabla f(\theta^{new})$;
end while
return $\theta_{\mathtt{ml}} = \theta^{new}$

numerical results reveal that MLE-based parameter estimation is highly accurate in the expectation sense.

IV. Proof of Theorems

A. Proof of Theorem 1

If the protector does not receive the information, it is trivial that there is no gain using MLE without protector. Hence, to prove the theorem, we only focus on the case that the protector receives the information *i.e.*, $M > 0$.

Polya's Urn. The description of Pólya's urn can be directly applied into diffusion spreading scenario in regular tree networks, in which the connection first appeared in [6]. Since we assume homogeneous spreading, *i.e.*, $\lambda = 1$ for all edges, at each infection epoch the next infecting node will be uniformly selected among the neighbors of currently infected nodes from the memoryless property of exponential random variable. This can be interpreted as uniform ball drawing in Pólya's urn. Furthermore, since it is assumed that the underlying network is d-regular tree, $d-2$ additional infection candidates are added in each infection time. This can be interpreted as ε additional balls in the urn. The number of balls drawn with each color is mapped into the number of each subtrees, spread from the information source. Initially, all subtrees have only one infection candidates, which makes $b_j = 1$ for all j and the total number of diffused nodes is $M + N$ (*i.e.*after $M + N - 1$ draws). Let X_j be the number of nodes, which are either infected or protected, in j-th subtree rooted at the information source, where $j = 1, \ldots, d$. Then, the joint probability of $(X_1, \ldots, X_d)$ is given by

$$P\left[\bigcap_{j=1}^{d}(X_j = x_j)\right] = \frac{(M+N-1)!}{x_1!x_2!\cdots x_d!} \cdot \frac{\prod_{j=1}^{d} 1(1+\varepsilon)\ldots(1+(x_j-1)\varepsilon)}{\prod_{k=1}^{M+N-1} d+(k-1)\varepsilon}, \tag{12}$$

where $\varepsilon = d - 2$ and $\sum_{j=1}^{d} x_j = M + N - 1$. In the proof, we will directly use the above for $d = 2$ (*i.e.*, line graph), but for $d > 2$, we take one step further as following. Let $Y_j \equiv X_j/(M+N)$, $(1 \le j \le d)$ then the distribution of the marginal distribution of Y_j for $M + N$ are sufficiently large is given by

$$\lim_{M+N\to\infty} P\left[Y_j \le x\right] = I_x\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right), \tag{13}$$

where $I_x(\alpha, \beta)$ is the incomplete beta function. We use this result to handle the asymptotic regime in the proof.

Proof of Theorem. In the proof, we denote by $P(v_{\mathtt{ml}} = v)$ the probability that v is $v_{\mathtt{ml}}$ when there are $M + N$ diffused nodes in the network and denote by $P(v = v^*)$ the probability that v is the information source, respectively. Then, the detection probability is given by

$$\begin{aligned} \pi_d^{\mathtt{ml}} &= \lim_{M+N\to\infty} P(\boldsymbol{C}_{M+N}) \\ &= \lim_{M+N\to\infty} \sum_{v\in I_N} P(v_{\mathtt{ml}} = v) P(v = v^*) \\ &\overset{(a)}{=} \lim_{M+N\to\infty} \frac{1}{N} \sum_{v\in I_N} P(v_{\mathtt{ml}} = v), \end{aligned} \tag{14}$$

where (a) is from the fact that $P(v = v^*) = 1/N$ for all infected nodes $v \in I_N$ in MLE. Hence, we need to obtain $P(v_{\mathtt{ml}} = v)$. To this end, we consider the following two cases: *(i)* $d = 2$ and *(ii)* $d \ge 3$.

(i) $d = 2$: In this case, we first consider the following lemma for the line graph.

Lemma 2: For the line graph ($d = 2$), when the information source $v^* \in I_N$ has m $(1 \le m \le 2)$ infected neighbor nodes then

$$\lim_{M+N\to\infty} P(v_{\mathtt{ml}} = v^*) = \mathbf{I}_{\{m=1\}}/2, \tag{15}$$

where $\mathbf{I}_{\{\cdot\}}$ is the indicator function.

This lemma says that when the source v^* is a boundary node between protected set and infected set, the detection probability is 1/2 otherwise, it becomes zero for the number of diffused nodes goes to infinite. We will provide the proof of this lemma in the Appendix. By using this result, we obtain

$$\pi_2^{\mathtt{ml}} = \lim_{M+N\to\infty} \frac{1}{N} \sum_{v\in I_N} P(v_{\mathtt{ml}} = v) \overset{(a)}{=} \lim_{M+N\to\infty} \frac{1}{2N} = 0,$$

where (a) is from the fact that the infected node whose distance from protector is one has $m = 1$ and the others have $m = 2$. Therefore, we conclude that $\pi_2^{\mathtt{ml}} = \pi_2 = 0$.

(ii) $d \ge 3$: From the result of Polya's urn (13), we first state the following lemma.

Lemma 3: For d-regular tree $(d \ge 3)$, when the information source $v^* \in I_N$ has m $(1 \le m \le d)$ infected neighbor nodes then we have

$$\lim_{M+N\to\infty} P(v_{\mathtt{ml}} = v^*) = 1 - m\left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right), \tag{16}$$

where $I_x(\alpha, \beta)$ is the incomplete Beta function.

This lemma is a similar result in [3] which considered the detection probability of MAPE with a suspect set. They obtained the detection probability of the information source with m $(1 \leq m \leq d)$ infected neighbor in the suspect set. In our case, the suspect set is given by the distance from the protector since we assume the finite L. We will give the proof of this lemma in Section V. This lemma implies that if the information source is a border node with small infected neighbors then MLE detect it more correctly[9] since this node will has a higher lielihood than others more frequently. From this result, we obtain the detection probability as follows. First, note that the infected node with one-hop from the protector has $m = d-1$ candidate infected neighbors due to the protected one and all other infected nodes except this have $m = d$. From this facts, the detection probability is given by

$$\begin{aligned}\pi_d^{\mathtt{ml}} &= \lim_{M+N\to\infty} \frac{1}{N} \sum_{v\in I_N} P(v_{\mathtt{ml}} = v) \\ &= 1 - d\left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right) = \pi_d.\end{aligned}$$

Hence, we complete the proof of Theorem 1.

B. Proof of Theorem 2

To prove this, we first consider the detection probability $\pi_d^{\mathtt{map}}(\hat{\theta})$ of MAPE for d-regular tree which is conditioned on the distance l as

$$\begin{aligned}\pi_d^{\mathtt{map}}(\hat{\theta}) &= \lim_{M+N\to\infty} \sum_{v\in\mathcal{V}_{\mathcal{L}}} P(v_{\mathtt{map}} = v) P(v = v^*) \\ &= \lim_{M+N\to\infty} \sum_{l=1}^{L} \left(\sum_{v\in V_l} P(v_{\mathtt{map}} = v)\right) P(d(v,p^*) = l|\theta^*) \\ &\overset{(a)}{=} \sum_{l=1}^{L} \left(\lim_{M+N\to\infty} \sum_{v\in V_l} P(v_{\mathtt{map}} = v)\right) P(d(v,p^*) = l|\theta^*) \\ &= \sum_{l=1}^{L} \varphi_l^{\mathtt{map}}(\hat{\theta}, d) P(d(v,p^*) = l|\theta^*),\end{aligned}$$

where $\mathcal{V}_{\mathcal{L}} := \cup_{l=1}^{L} V_l$ and $\varphi_l^{\mathtt{map}}(\hat{\theta}, d) := \lim_{M+N\to\infty} \sum_{v\in V_l} P($ $v_{\mathtt{map}} = v)$ is the detection probability of the MAPE when the distance of two sources is l $(1 \leq l \leq L)$. The equality (a) is from the fact that L is finite. Due to the technical challenge in obtaining the exact formula of MAPE's detection probability, we will obtain the lower bound of the probability. To do this, consider the following lemma.

Lemma 4: For the information source $v^* \in I_N$ with $d(v^*, p^*) = l$ $(1 \leq l \leq L)$, let $P_l := P(d(v^*, p^*) = l|\hat{\theta})$ then we have

$$\begin{aligned}\varphi_l^{\mathtt{map}}(\hat{\theta}, d) \geq 1 - (d-1)\left(1 - I_{p_l}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right) \\ - \left(1 - I_{q_l}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right),\end{aligned} \tag{17}$$

[9]Note that if $m < d$, the value in (16) will be greater than that of the case when $m = d$ and this will be occurred if the node is border node in the observation.

where $p_l = \frac{P_l}{P_{l+1}+P_l}$ and $q_l = \frac{P_l}{P_{l-1}+P_l}$ for any distributions.

This result is slightly different in Lemma 1. We will give the proof of details in the Section V. By using this, we will prove the theorem as following steps. First, we obtain the lower bound of difference of detection probabilities between MAPE with true parameter and MLE without protector. Second, we will obtain the upper bound of difference of detection probabilities between MAPE with true parameter and MAPE with an estimated parameter. By subtracting these two results, we will conclude the proof of theorem 2. To obtain the result of first part, we consider the following result.

Lemma 5: For d-regular trees $(d \geq 2)$,

$$\pi_d^{\mathtt{map}}(\theta^*) - \pi_d \geq (p(\theta^*) - 1/2)^{\frac{2d-3}{3(d-2)}}, \tag{18}$$

where $p(\cdot)$ is defined in (9) for each distributions, respectively.

This result shows that the lower bound of difference between two detection probabilities depends on the true parameter and degree. We see that the detection performance *gap* from the MLE without anti-information is up to 50% for $d = 3$ and up to 63% for $d \to \infty$, which are non-negligible quantities. We will give the proof of details in Section V. Next, to obtain the second part, we consider the following lemma.

Lemma 6: For d-regular trees $(d \geq 2)$,

$$\pi_d^{\mathtt{map}}(\theta^*) - \pi_d^{\mathtt{map}}(\hat{\theta}) \leq 6K(\theta^*, \hat{\theta})|\theta^* - \hat{\theta}|, \tag{19}$$

where $K(\theta^*, \hat{\theta}) = |1 - p(\hat{\theta})/p(\theta^*)|$.

This results implies that for any estimated value $\hat{\theta}$, the difference of detection probabilities depends on how well it estimates the true parameter *i.e.*, $|\theta^* - \hat{\theta}|$ under the true parameter. We also give the proof of details in the next section. Then, by combining (18), (19), we obtain the result of theorem 2 and this completes the proof.

V. Proof of Lemmas

This section provides the proofs of lemmas used for establishing Theorem 1 and Theorem 2.

A. Proof of Lemma 3

Before the proof of lemma, we denote $|T_u^v|$ as the number of infected or protected nodes in the subtree rooted at node u where v is the information source as shown in [1]. In order to obtain the result of the lemma, we first consider the following proposition which is a property of MLE in the case of existing the protectors.

Proposition 1: (*Property of MLE*) For the uniform priori distribution, a node v is $v_{\mathtt{ml}}$ if and only if $|T_u^v| \leq (M+N)/2$ for all $u, v \in I_N$.

Proof. It can be directly obtained from the result of Proposition 2.

The above proposition implies that when there are protectors in the networks, MLE has the same property with the rumor centrality as in [1] *i.e.*, it is located at the middle of the infected and protected graph. The only difference is that when there are more protected nodes than infected nodes, the MLE

chooses the boundary infected node of these two set I_N and P_M as the estimator because it has highest likelihood among the infected nodes. Based on this, let $E_j = \{X_j \leq (M+N)/2\}$ be the event that the number of infected and protected nodes in a j-th subtree of the information source v^* is less than $(M+N)/2$ for $1 \leq j \leq m$, where m is the number of infected neighbors of v^*. Then, from the result in [3] and Proposition 1, the detection probability of the source $v^* \in I_N$ with m $(1 \leq m \leq d)$ infected neighbor nodes in d-regular tree $(d \geq 3)$ is given by

$$\begin{aligned}\lim_{M+N\to\infty} P(v_{\mathtt{ml}} = v^*) &= \lim_{M+N\to\infty} (1 - \cup_{j=1}^{m} P(E_j^c)) \\ &\overset{(a)}{=} \lim_{M+N\to\infty} (1 - mP(E_1^c)),\end{aligned} \tag{20}$$

where (a) is due to the fact that the events E_j are identical and disjoint for all j. From the result in (13), we have

$$\lim_{M+N\to\infty} P(E_1^c) = 1 - I_{1/2}\left(1/(d-2), (d-1)/(d-2)\right),$$

and by putting this into (20), we obtain the result which completes the proof of Lemma 3. ■

B. Proof of Lemma 4

In order to prove this lemma, we first construct a property of MAPE which is a kind of generalization of the result in Proposition 1. To do this, let $\rho(v, w)$ be the set of nodes on the path between v and w [10] in I_N. Then, we have the following Proposition.

Proposition 2: (*Property of MAPE*) For a given distribution, a node v is $v_{\mathtt{map}}$ if and only if

$$\prod_{i\in\rho(v,w)} \left(\frac{|T_i^v|}{(M+N) - |T_i^v|}\right) \leq \frac{P(d(v,p^*))}{P(d(w,p^*))}, \tag{21}$$

for all $u, w \in I_N$.

This seems to be quite complex compared to Proposition 1. However, one can easily check that when $P(d(v,p^*)) = P(d(w,p^*))$ for all $v, u \in I_N$, it is directly obtained the result in Proposition 1 as a special case of MAPE due to $|T_{i+1}^v| \leq |T_i^v| - 1$ for any $i \in \rho(v,w)$ where $i+1$ is a neighbor node of i which the distance from the node v is greater than that of v.

Proof. First, we prove that if $v = v_{\mathtt{map}}$ then v satisfies (21). To see this, we consider the computation of MLE in (5). Let $R^{'}(v) = R(v, p^*, I_N, P_M)P(d(v,p^*))$ then for a neighbor node u of v, one can easily check that $|T_u^v| = (M+N) - |T_v^u|$. Since we assumed $v = v_{\mathtt{map}}$, it follows that $\frac{R^{'}(u)}{R^{'}(v)} = \left(\frac{|T_u^v|}{(M+N)-|T_u^v|}\right)\frac{P(d(u,p^*))}{P(d(v,p^*))} \leq 1$. Let $\chi_u^v = |T_u^v|/((M+N) - |T_u^v|)$ then for any node $w \in I_N$,

$$\begin{aligned}\frac{R^{'}(w)}{R^{'}(v)} = \left(\prod_{i\in\rho(v,w)} \chi_i^v\right)\frac{P(d(w,p^*))}{P(d(v,p^*))} \leq 1 \\ \Leftrightarrow \prod_{i\in\rho(v,w)} \chi_i^v \leq \frac{P(d(v,p^*))}{P(d(w,p^*))},\end{aligned}$$

where $\rho(v,w)$ is the set of nodes in the path between v and w, not including v. Then, for every $s \in I_N$ with the distance $|\rho(v,s)| = h$ $(1 \leq h \leq 2L-1)$, v satisfies (21). Conversely, if a node v satisfies (21) then $v = v_{\mathtt{map}}$. Hence, we complete the proof of Proposition 2. ■

Based on this result, we will give the proof only for Zipf distribution since the proofs for the others are similar. To do this, let $v \in I_N$ be the information source with $d(v,p^*) = l$ $(1 \leq l \leq L)$ and let C_h be the event which satisfies (21) for all nodes $u \in I_N$ with $|\rho(v,u)| = h > 0$. Then, in order to obtain $\varphi_l^{\mathtt{map}}(\theta, d)$ in (17), we need to find the probability of $\cap_h C_h$ for all $1 \leq h \leq 2L-1$ in Proposition 2. To do this, we divide $\cap_h C_h$ into the two parts such as $\cap_h C_h = E_l \cap F_l$ where $E_l \equiv \cap_{h=1}^{L-l} C_h$ is a part of larger distance from p^* than that of v and $F_l \equiv \cap_{h=1}^{l-1} C_h$ is a part of direction to p^*. Let C_l^{M+N} be the event that MAP estimator is the information source when the distance of two sources is l and there are $M+N$ infected and protected nodes. Then, we have

$$\begin{aligned}P(C_l^{M+N}) &= P\left(\left(\cap_{j=1}^{d-1} E_{l,j}\right) \cap F_l\right) = 1 - P\left(\left(\cup_{j=1}^{d-1} E_{l,j}^c\right) \cup F_l^c\right) \\ &\overset{(a)}{\geq} 1 - \sum_{j=1}^{d-1} P\left(E_{l,j}^c\right) - P\left(F_l^c\right) \\ &\overset{(b)}{=} 1 - (d-1)P\left(E_l^c\right) - P\left(F_l^c\right),\end{aligned} \tag{22}$$

where $E_{l,j}$ is the event for j^{th} $(1 \leq j \leq d-1)$ subtree of larger distance from p^* than that of v. First, consider that the inequality (a) is from the union bound of the probability due to the fact that the events $E_{l,j}^c$ and F_l^c are not disjoint by the following reason. For a neighbor node $u \in I_N$ of v with $d(u,p^*) = l+1$, we have $|T_u^v| \leq (M+N)p_l$ in the Proposition 2 where

$$p_l = \frac{P(d(v,p^*))}{P(d(u,p^*)) + P(d(v,p^*))} = \frac{(l+1)^\theta}{(l+1)^\theta + l^\theta},$$

and similarly, $q_l = \frac{(l-1)^\theta}{l^\theta + (l-1)^\theta}$. Then $p_l + q_l < 1$ for any $\theta > 0$ and this makes the non-disjoint events. [11] The equality (b) is from the disjoint events of $E_{l,j}$ for all $1 \leq j \leq d-1$. Hence, it remains to obtain the probability $P\left(E_l^c\right)$ and $P\left(F_l^c\right)$ in (22), respectively. To this end, consider $P\left(E_l^c\right) = P\left(\left(\cap_{h=1}^{L-l} C_h\right)^c\right) = P\left(\cup_{h=1}^{L-l} C_h^c\right)$ but, it is not easy to obtain the exact probability due to the fact that $\cap_{h=1}^{L-l} C_h \neq C_1$. However, by dividing the event $\cup_{h=1}^{L-l} C_h^c$ as disjoint events and by taking limit, we have

$$\begin{aligned}\lim_{M+N\to\infty} P\left(\cup_{h=1}^{L-l} C_h^c\right) &= \lim_{M+N\to\infty} P\left(C_1^c\right) \\ &+ \underbrace{\lim_{M+N\to\infty} \sum_{h=1}^{L-l} P\left(C_{h+1}^c \mid \cap_{j=1}^{h} C_j\right)}_{(*)} \\ &\overset{(a)}{=} 1 - I_{p_l}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right),\end{aligned}$$

where (a) is from the fact that the term $(*)$ will be vanished as $M+N$ goes to infinity and by applying the Polya's urn in (13)

[10] For the tree, there is a unique path between any two distinct nodes.

[11] If $\theta = 0$ *i.e.*, for the Uniform distribution, the parameters $p_l = q_l = 1/2$ for all $1 \leq l \leq L$ and this makes the events to be disjoint as in [6].

into $P\left(C_1^c\right)$ because

$$\begin{aligned}(*) &= \lim_{M+N\to\infty}\sum_{h=1}^{L-l}\left(1-P\left(C_{h+1}\mid \cap_{j=1}^{h}C_j\right)\right)\\ &= \sum_{h=1}^{L-l}\left(1-\lim_{M+N\to\infty}P\left(C_{h+1}\mid \cap_{j=1}^{h}C_j\right)\right)\\ &\overset{(a)}{=} \sum_{h=1}^{L-l}\left(1-\lim_{M+N\to\infty}P\left(C_{h+1}\mid C_h\right)\right)\overset{(b)}{=}0,\end{aligned}$$

where (a) is from the memoryless property of exponential distributions of homogeneous diffusion process and (b) is $\lim_{M+N\to\infty}P\left(C_{h+1}\mid C_h\right)=1$ from the Hoeffding inequality. Similarly, one can obtain the part of $P\left(F_l^c\right)$. Then, by putting these results into (22), we obtain (17) and this completes the proof of Lemma 4.

C. Proof of Lemma 5

First, consider that the detection probability is obtained by using the incomplete beta function as in Lemma 4. Accordingly, to obtain the difference of two detection probabilities, it is required the difference of two incomplete beta functions with different parameters. Let $D_d(x,y) := \left|I_x\left(\frac{1}{d-2},\frac{d-1}{d-2}\right)-I_y\left(\frac{1}{d-2},\frac{d-1}{d-2}\right)\right|$ be the difference of two incomplete beta functions with parameters x and y then we have the following proposition.

Proposition 3: For $0\le x,y\le 1$ and $d\ge 3$, we have

$$\frac{|x-y|^{\frac{2d-3}{4(d-2)}}}{d-2}\le D_d(x,y)\le\frac{c(d)(1+x^{1/(d-2)})|x-y|}{d-2},\tag{23}$$

where $c(d)=\frac{2d-3}{2(d-2)\Gamma(\frac{d-1}{d-2})}$.

This results guarantee mathematical tractability for calculating the difference of detection probability because it is a simple approximation to the incomplete beta function which is a complex form to handle, directly. This is one of polynomial approximation and then we can obtain the proper bound of detection probability by using this result which will be given in later of the section.

Proof. First, we will show that the upper bound in (23). From the definition of incomplete beta function, we have

$$\begin{aligned}D_d(x,y) &= \left|I_x\left(\tfrac{1}{d-2},\tfrac{d-1}{d-2}\right)-I_y\left(\tfrac{1}{d-2},\tfrac{d-1}{d-2}\right)\right|\\ &= \frac{\Gamma(\frac{1}{d-2}+\frac{d-1}{d-2})}{\Gamma(\frac{1}{d-2})\Gamma(\frac{d-1}{d-2})}\left|\int_0^x t^{\frac{1}{d-2}-1}(1-t)^{\frac{d-1}{d-2}-1}dt\right.\\ &\qquad\left.-\int_0^y t^{\frac{1}{d-2}-1}(1-t)^{\frac{d-1}{d-2}-1}dt\right|\\ &= \frac{1}{B(\frac{1}{d-2},\frac{d-1}{d-2})}\cdot\left|\int_x^y t^{\frac{1}{d-2}-1}(1-t)^{\frac{d-1}{d-2}-1}dt\right|\\ &\overset{(a)}{\le}\frac{1}{B(\frac{1}{d-2},\frac{d-1}{d-2})}\cdot\left|\int_x^y t^{\frac{1}{d-2}-1}dt\right|\le c(d)|x^{\frac{1}{d-2}}-y^{\frac{1}{d-2}}|\\ &\overset{(b)}{\le}\frac{c(d)(1+x^{\frac{1}{d-2}})}{d-2}|x-y|,\end{aligned}$$

where (a) is from the fact that $t^{\frac{1}{d-2}-1}(1-t)^{\frac{1}{d-2}}\le t^{\frac{1}{d-2}-1}$ for $0<t\le 1$ and (b) is from the fact that $x^{\frac{1}{d-2}}$ is a contraction mapping with $(1+x^{\frac{1}{d-2}})/(d-2)$. The terms $B(\frac{1}{d-2},\frac{d-1}{d-2})=\Gamma(\frac{1}{d-2})\Gamma(\frac{d-1}{d-2})/\Gamma(\frac{1}{d-2}+\frac{d-1}{d-2})$ is a beta function and $c(d)=(2d-3)/2(d-2)\Gamma(\frac{d-1}{d-2})$. By similar approaches in this proof, the lower bound also can be obtained as follows.

$$\begin{aligned}D_d(x,y) &= \tfrac{1}{B(\frac{1}{d-2},\frac{d-1}{d-2})}\cdot\left|\int_x^y t^{\frac{1}{d-2}-1}(1-t)^{\frac{d-1}{d-2}-1}dt\right|\\ &\overset{(a)}{\ge}\tfrac{1}{B(\frac{1}{d-2},\frac{d-1}{d-2})}\cdot\left|\int_x^y(1-t)^{\frac{d-1}{d-2}}dt\right|\\ &\overset{(b)}{\ge}\tfrac{2d-1}{(d-2)^2}\cdot\left|\int_x^y(1-t)^{\frac{d-1}{d-2}}dt\right|\\ &= \tfrac{2d-1}{(d-2)(2d-3)}\cdot\left|(1-x)^{\frac{2d-3}{d-2}}-(1-y)^{\frac{2d-3}{d-2}}\right|\\ &\overset{(c)}{\ge}\tfrac{1}{(d-2)}\cdot\left|x^{\frac{2d-3}{4(d-2)}}-y^{\frac{2d-3}{4(d-2)}}\right|\overset{(d)}{\ge}\tfrac{1}{(d-2)}\cdot|x-y|^{\frac{2d-3}{4(d-2)}},\end{aligned}$$

where (a) follows from the fact that $t^{\frac{1}{d-2}-1}(1-t)^{\frac{d-1}{d-2}-1}\ge(1-t)^{\frac{d-1}{d-2}}$ and (b) is from the fact that $B(x,y)=\frac{x+y}{xy}\prod_{n=1}^{\infty}\left(1+\frac{xy}{n(x+y+n)}\right)^{-1}\le\frac{x+y}{xy}$. The inequality (c) follows from the fact that $(1-x)^a-(1-y)^a\ge x^{a/4}-y^{a/4}$ for $0\le a\le 4$ and $0\le x,y\le 1$, and (d) is due to the fact that x^a is a concave function with respect to x when $a\le 1$. Hence, we complete the proof of Proposition 3. ■

Based on this result, we will finish the proof of Lemma. To do this, let $P_l^*=P(d(v^*,p^*)=l|\theta^*)$ and then from the fact that $\sum_{l=1}^{L}P_l^*=1$, we have

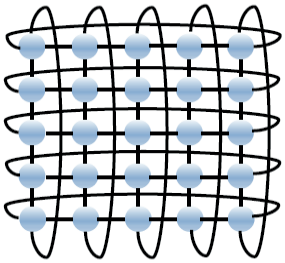
$$\begin{aligned}\pi_d^{\texttt{map}}(\theta^*)-\pi_d &= \textstyle\sum_{l=1}^{L}\left(\varphi_l^{\texttt{map}}(\theta^*,d)-\pi_d\right)P_l^*\\ &= \textstyle\sum_{l=1}^{L}\left((d-1)D_d(p_l^*,1/2)-D_d(q_l^*,1/2)\right)P_l^*\\ &\overset{(a)}{\ge}\textstyle\sum_{l=1}^{L}\left((d-1)\frac{|p_l^*-1/2|^{\frac{2d-3}{4(d-2)}}}{d-2}\right.\\ &\qquad\left.-\frac{c(d)(1+(q_l^*)^{\frac{1}{(d-2)}}|q_l^*-1/2|}{d-2}\right)P_l^*\\ &\overset{(b)}{\ge}\textstyle\sum_{l=1}^{L}\left((d-1)\frac{|p_l^*-1/2|^{\frac{2d-3}{4(d-2)}}}{d-2}-\frac{|p_l^*-1/2|^{\frac{2d-3}{4(d-2)}}}{d-2}\right)P_l^*\\ &= \textstyle\sum_{l=1}^{L}|p_l^*-1/2|^{\frac{2d-3}{4(d-2)}}P_l^*\overset{(c)}{\ge}(p(\theta^*)-1/2)^{\frac{2d-3}{3(d-2)}},\end{aligned}$$

where (a) is due to Proposition 3 and (b) follows from some algebra using the fact that $|q_l^*-1/2|\le|p_l^*-1/2|$ for all $1\le l\le L$. The inequality (c) is from the fact that $|p_l^*-1/2|^{\frac{2d-3}{4(d-2)}}\ge(p(\theta^*)-1/2)^{\frac{2d-3}{3(d-2)}}$ for all $1\le l\le L$ and $d\ge 3$ where $p(\theta^*)$ is defined in (9). Therefore, we complete the proof of Lemma 5.

D. Proof of Lemma 6

We also use the result of Proposition 3 to obtain the upper bound of the difference between detection probabilities with true

(a) US Power Grid Network.



(b) 5 x 5 Grid Torus Graph.

Fig. 4. Examples of the real world graph: (a) US power grid network and (b) Grid torus netowrk.

and estimated parameters. To see this, consider that

$$\begin{aligned}
\pi_d^{\mathtt{map}}(\theta^*) - \pi_d^{\mathtt{map}}(\hat{\theta}) &= \textstyle\sum_{l=1}^{L} \left(\varphi_l^{\mathtt{map}}(\theta^*, d) - \varphi_l^{\mathtt{map}}(\hat{\theta}, d)\right) P_l^* \\
&= \textstyle\sum_{l=1}^{L} \left((d-1)D_d(p_l^*, p_l) - D_d(q_l^*, q_l)\right) P_l^* \\
&\overset{(a)}{\leq} \textstyle\sum_{l=1}^{L} \Bigg(\frac{(d-1)c(d)(1+(p_l^*)^{\frac{1}{d-2}})|p_l^* - p_l|}{d-2} \\
&\qquad - \frac{c(d)(1+(q_l^*)^{\frac{1}{d-2}})|q_l^* - q_l|}{d-2} \Bigg) P_l^* \\
&\overset{(b)}{\leq} \textstyle\sum_{l=1}^{L} \left(\frac{(d-1)c(d)(1+(p_l^*)^{\frac{1}{d-2}})|p_l^* - p_l|}{d-2} \right) P_l^* \\
&= \textstyle\sum_{l=1}^{L} \left(\tilde{c}(d)(1+(p_l^*)^{\frac{1}{d-2}})|p_l^* - p_l| \right) P_l^* \\
&\overset{(c)}{\leq} |\theta^* - \hat{\theta}| \textstyle\sum_{l=1}^{L} K_l \left(\tilde{c}(d)(1+(p_l^*)^{\frac{1}{d-2}}) \right) P_l^* \\
&\overset{(d)}{\leq} 6K(\theta^*, \hat{\theta})|\theta^* - \hat{\theta}|,
\end{aligned}$$

where (a) is from Proposition 3 and (b) follows from the fact that $|q_l^* - q_l| \geq |p_l^* - p_l|$ for all l. The inequality (c) is due to the fact that p_l is a contraction mapping w.r.t. θ since p_l is a continuous differentiable and K_l is a corresponding a Lipschitz constant and $\widetilde{c}(d) \equiv c(d)(d-1)/(d-2)$. The inequality (d) follows from the facts that $p_l^* \leq p_1^* := p(\theta^*)$ for all $2 \leq l \leq L$ with $\widetilde{c}(d)(1+p(\theta^*)^{\frac{1}{d-2}}) \leq 3$ for all $d \geq 3$. For given θ^* and θ, one can easily check that $K_l \leq 2|1 - p(\hat{\theta})/p(\theta^*)|$ for all $1 \leq l \leq L$ by simple algebra and this completes the proof of Lemma 6.

VI. General Graphs and Simulation Results

We have so far assumed that the underlying graph is a regular tree, which is simply for analytical tractability as done in other related work. In this section, inspired by our analytical findings in earlier sections, we study the detection performance of a MAPE-based algorithm in more practical, general graphs. such as ER-random graph, small-world network, scale-free network, a Facebook network, and a US power grid network. In addition to those graphs under which we randomly place a protector following either of Zipf, Geometric, and Poisson. Different from this, we also

MAP-BFS estimator with θ^* learning. We first describe a heuristic estimator motivated by MAPE, which is necessary due to the computational intractability [12] of the problem MAPE in (1). Motivated by the heuristic in [1], we propose a heuristic algorithm based on Breadth-First Search (BFS), as described in what follows: Let σ_v be the infection sequence of the BFS ordering of the nodes in the given graph, then we estimate the source $v_{\mathtt{map}}^{\mathtt{b}}$ that solves the following:

$$v_{\mathtt{map}}^{\mathtt{b}} = \arg\max_{v \in G_N} P(\sigma_v | v, p^*) \Big[R(v, p^*, T_b(v)) \times P(d(v, p^*)) \Big],$$

where $T_b(v)$ is a BFS tree rooted at v and the information spreads along it and $d(v, p^*)$ is the shortest distance between v and p^*. Note that $P(d(v, p^*))$ uses an MLE-estimated parameter as in Section III-C based on $T_b(v)$, where computing the rumor centrality $R(\cdot)$ (in (11)) with $T_b(v)$ is the key component. This $T_b(v)$-based parameter learning is also a heuristic since obtaining the exact $\theta_{\mathtt{ml}}$ for a general graph is hard to solve. Except for the complexity in learning the distribution parameter, we can estimate the information source in $O(N(M+N))$ time.

Graphs. We consider (i) three *synthetic random* graphs: *Erdös-Rényi* (ER) random graphs, small-world (SW), scale-free (SF)

Algorithm 2 Distance Centrality-Based Algorithm (DSBA)

Input: (G, n, L)
Select a subgraph $G_L \subseteq G$ with diameter L randomly and generate an information source $v_i^* \in G_L$ uniformly at random up to $1 \leq i \leq n$;
for $v \in G_L$ **do**
Compute the distance $d(v, v_i^*)$ by a shortest path algorithm for all i and calculate the distance centrality of v by $C(v) = 1/\sum_{i=1}^{n} d(v, v_i^*)$;
end for
$P \leftarrow \phi$;
$v = \arg\max_{v \in G_L} C(v)$;
$P \leftarrow P \cup \{v\}$;
if $|P| > 1$ **then**
Choose $v \in P$ uniformly at random;
end if
$p^* \leftarrow v$;
return p^*

Algorithm 3 Degree Centrality-Based Algorithm (DGBA)

Input: (G, L)
Select a subgraph $G_L \subseteq G$ with diameter L randomly ;
Set $D(v)$ by the degree of node v in G_L;
$P \leftarrow \phi$;
$v = \arg\max_{v \in G_L} D(v)$;
$P \leftarrow P \cup \{v\}$;
if $|P| > 1$ **then**
Choose $v \in P$ uniformly at random;
end if
$p^* \leftarrow v$;
return p^*

[12] We can easily prove that this is ♯P-complete similarly to the proof of MLE without protectors in [1].

Table 1. Detection Probabilities with *known* distribution for General Graphs

Distribution	ER	SW	SF	Torus	FB	US
No protector	0.02	0.03	0.02	0.03	0.01	0.03
Uniform	0.03	0.05	0.03	0.07	0.02	0.04
Zipf	**0.13**	**0.10**	**0.11**	**0.21**	**0.07**	**0.11**
Geometric	**0.10**	**0.08**	**0.09**	**0.18**	**0.06**	**0.08**
Poisson	**0.11**	**0.11**	**0.10**	**0.14**	**0.09**	**0.09**

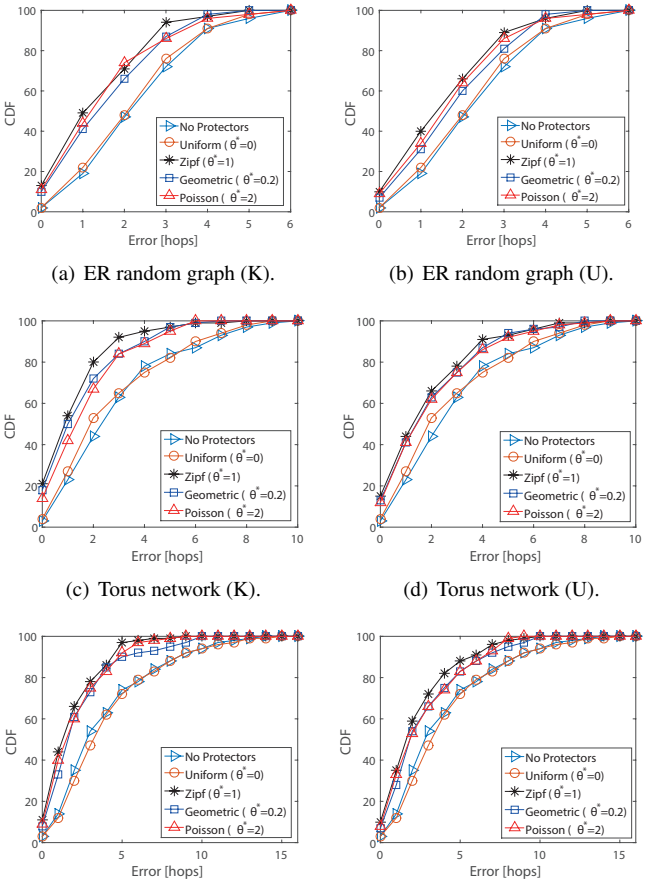
graphs, and Torus grid (TG) (see Fig. VI(b) as an example) and (ii) two *real-world* graphs; a Facebook (FB) ego network and a US power (US) grid network. First, in synthetic random graphs, we set the average degree as 4 when there are 2000 nodes in the networks. For the Torus grid network, we consider a 60×60 grid torus network (thus 3600 nodes). Second, the Facebook ego network [30] is a undirected graph consisting of 4039 nodes and 88234 edges where each edge corresponds to a social relationship (called FriendList) and the diameter is 8 hops. The US power grid network [31] consists of 4941 nodes and 6594 edges and the diameter is 46 hops.

Protector Selection Algorithms [13]**.** In practice, the distance distribution may not be known as a priori so that we need to estimate or to assume some proper distributions to obtain the detection behaviors by MAP-BFS estimator. In this simulation, we consider the following two scenarios: (i) Known distribution (K) and (ii) Unknown distribution (U), respectively. For the first case, since the distribution is given as a priori, we only need to estimate the hidden true parameter of the distribution by some heuristic learning algorithm as we mentioned earlier. However, in the second case, due to the lack of the knowledge of distribution, we use some statistical information about the history of location for previous information sources. Based on this, we provide two protector selection algorithms as follows. First, we consider an algorithm based on distance centrality (DSBA) of locations for them if the diameter of the network is huge. Second, we consider an algorithm based on the degree centrality (DGBA) of the networks, otherwise. In both algorithms, we use the notion G_L to denote a subgraph of G which the diameter is $L > 0$.

Setup. We use the true parameters: $\theta^* = 1$ for Zipf, $\theta^* = 0.2$ for Geometric and $\theta^* = 2$ for Poisson distributions and compare the results to the case of no protectors in the network and no priori information (*i.e.*, MLE). We just choose these parameters for a representative example and show their results due to space limitation, where we observe a similar trend in other parameter configurations. We use MATLAB for the simulations and generate 200 random graph samples for synthetic random graphs and a scenario-driven graph, where we diffuse an information and opposite information until we have $M + N = 600$. By considering the total network size, we set the value L as 50 % to the diameter of networks and we performed 100 iterations for all graphs to obtain the results.

Simulation Results. In the simulation, we obtain two different results as in the Fig. 5 for the known distribution (K) and unknown distribution (U), respectively. The x-axis of the figure

[13]Different to the case of regular trees, it become an important issue that where the protector is located in general graphs because the degrees of each node and the diameter of graphs are not same.



(a) ER random graph (K). (b) ER random graph (U).

(c) Torus network (K). (d) Torus network (U).

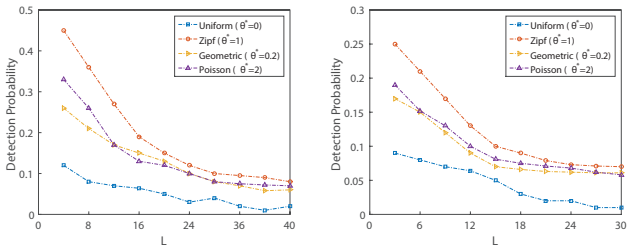
(e) US power grid network (K). (f) US power grid network (U).

Fig. 5. Simulation results of MAP-BFS detection performances where the Cumulative Distribution Function (CDF) of the distance between true source and estimator (Error) with 100 iterations under the general topologies when $M + N = 600$. (K: known, U:Unknown)

Table 2. Detection Probabilities with *unknown* distribution for General Graphs

Distribution	ER	SW	SF	Torus	FB	US
No protector	0.02	0.03	0.02	0.03	0.01	0.03
Uniform	0.02	0.03	0.04	0.04	0.02	0.03
Zipf	**0.10**	**0.08**	**0.10**	**0.15**	**0.06**	**0.10**
Geometric	**0.07**	**0.07**	**0.07**	**0.13**	**0.04**	**0.07**
Poisson	**0.09**	**0.09**	**0.08**	**0.12**	**0.05**	**0.08**

indicated that the number of errors between true source and estimator and the y-axis indicates that the Cumulative Distribution Function (CDF) of the errors. Clearly, the zero value of the error is the exact detection probability. For the second case, we use DGBA for ER, SW, SF and FB graphs, and use DSBA for Torus and US networks. The results show that if the distance distribution is known as a priori, the detection performances of MAP-BFS heuristic are better than that of the case of no protector and no priori information (*i.e.*Uniform distribution). It is hard to be beyond 5% for the case of no protector and no priori information but, if the distance information is given, we see that the detection probabilities can be beyond 10 % for the synthetic as well as real world topology even for our parameter setting with the estimated parameter (See Table 1). This means that the priori information of the distance between two sources is more



(a) Torus network. (b) US Power grid network.

Fig. 6. MAP-BFS detection probabilities with known distributions by varying L with 100 iterations under Torus and US power grid networks for $M + N = 600$.

helpful to find the information source in the general graphs as we expected. In the case of unknown distribution, the detection performances decrease compared to those of known distribution case. However, there are non-negligible enhancements from the result of no protector (See Table 2). This implies that if there is some information about the distance of two sources, then it is better to consider a protector which is located properly than that of no protector with assuming some proper distribution (like Zipf distribution). Finally, in Fig. 6, we obtain the simulation result as varying the value L and the results show that the MAP-BFS estimator under Zipf, Geometric and Poisson distribution outperform the case with no priori information, as expected, and the probabilities are larger than that of Uniform distribution even for the sufficient large setting of L.

VII. CONCLUSION

In this paper, we consider the information source detection problem in presence of passive protectors that spread the anti-information. We obtain MAPE with true parameter for a given priori distribution with hidden parameter and show that the detection probabilities become larger than that of no protectors and no priori information by learning the true parameter. Furthermore, we consider the case that the distance distribution is not given as a priori and for this case, we provide two protector locating algorithms which shows that there is also non-negligible enhancements for detecting the information source by MAPE-based heuristic estimator by assuming proper distributions. For the future work, we will consider the heterogeneous diffusion rates and active protector which start the diffusion simultaneously.

REFERENCES

[1] D. Shah and T. Zaman. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. In *Proceedings of ACM SIGMETRICS*, 2010.

[2] Z. Wang, W. Dong, W. Zhang and C. W. Tan. Rumor source detection with multiple observations: fundamental limits and algorithms. In *Proceedings ACM SIGMETRICS*, 2014.

[3] D. Wenxiang, Z. Wenyi and T. C. Wei. Rooting Out the Rumor Culprit from Suspects. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2013.

[4] Z. Kai and Y. Lei. Information Source Detection in the SIR Model: A Sample Path Based Approach. In *Proceedings of IEEE Information Theory and Applications Workshop (ITA)*, 2013.

[5] D. Shah and T. Zaman. Rumors in a Network: Who's the Culprit?. *IEEE Transactions on Information Theory*, 57(8):5163-5181, 2011.

[6] D. Shah and T. Zaman. Rumor Centrality: A Universal Source Estimator. In *Proc. ACM SIGMETRICS*, 2012.

[7] W. Luo, Wee P. Tay and M. Leng. Identifying Infection Sources and Regions in Large Networks. *IEEE Transactions on Information Theory*, 61(11):2850-2865, 2013.

[8] A. Odlyzko and B. Tilly. A refutation of Metcalfe's Law and a better estimate for the value of networks and network interconnections. In *Im-magic.com*, 2006.

[9] D. Braha, B. Stacey and Y. Bar-Yam. Corporate competition: A self-organized network. In *Elsevier Social Networks*, 2011.

[10] A. Clauset, C. R. Shalizi and M. E. J. Newman. Power-law distributions in empirical data. In *arXiv:0706.1062v2*, 2009.

[11] W. Luo, Wee P. Tay and M. Leng. How to Identify an Infection Source With Limited Observations. *IEEE Journal of Selected Topics in Signal Processing*, 8(4):586-597, 2014.

[12] Z. Kai and Y. Lei. A robust information source estimator with sparse observations. In *Proceedings of IEEE INFOCOM*, 2014.

[13] J. Solomon. Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics. In *CRC Press*, 2015.

[14] F. Michael and Yu and Pei-Duo. Rumor source detection for rumor spreading on random increasing trees. In *Electronic Communications in Probability*, 2015.

[15] W. Luo and W.P. Tay. Finding an infection source under the SIS model. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[16] S. Bubeck, L. Devroye and G. Lugosi. Finding Adam in random growing trees. *arXiv:1411.3317*, 2014.

[17] J. Khim and P.L. Loh. Confidence Sets for Source of a Diffusion in Regular Trees. *arXiv:1510.05461*, 2015.

[18] G. Fanti, P. Kairouz, S. Oh and P. Viswanath. Spy vs. Spy: Rumor Source Obfuscation. In *Proc. ACM SIGMETRICS*, 2015.

[19] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran and P. Viswanath. Rumor Source Obfuscation on Irregular Trees. In *Proc. ACM SIGMETRICS*, 2016.

[20] Mehrdad Farajtabar, Manuel Gomez-Rodriguez, Nan Du, Mohammad Zamani, Hongyuan Zha and Le Song. Back to the Past: Source Identification in Diffusion Networks from Partially Observed Cascades. In *Proc. AISTATS*, 2015.

[21] Kai Zhu and Lei Ying. Information source detection in networks: Possibility and impossibility result. In *Proc. IEEE INFOCOM*, 2017.

[22] Kai Zhu, Zhen Chen and Lei Ying. Catch'Em All: Locating Multiple Diffusion Sources in Networks with Partial Observations. In *Proc. AAAI*, 2017.

[23] Biao Chang, Feida Zhu, Enhong Chen and Qi Liu. Information source detection via Maximum A postreia Estimation. In *Proc. IEEE ICDM*, 2015.

[24] Zheng Wang, Chaokun Wang, Jisheng Pei and Xiaojun Ye. Multiple Source Detection without Knowing the Underlying Propagation Model. In *Proc. AAAI*, 2017.

[25] J. Choi, S. Moon, J. Woo, K. Son, J. Shin and Y. Yi. Rumor Source Detection under Querying with Untruthful Answers. In *Proc. IEEE INFOCOM*, 2017.

[26] J. Choi, S. Moon, J. Shin and Y. Yi. Estimating the Rumor Source with Anti-Rumor in Social Networks. In *Proc. IEEE ICNP Workshop on Machine Learning*, 2016.

[27] J. Choi, S. Moon, J. Shin and Y. Yi. Rumor Source Detection: Power of Protector. In *Proc. Proc. NetSci*, 2016.

[28] S. Moon, J. Choi, J. Shin and Y. Yi. Rumor Source Detection: Power of Querying. In *Proc. Proc. NetSci*, 2016.

[29] W. Luo, W. P. Tay and M. Leng. Infection Sprading and Source Identification: A Hide and Seek Game. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, VOL. 64, NO. 16, AUGUST 15, 2016.

[30] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. In *Proceedings of NIPS*, 2012.

[31] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440-442, 1988.