Information Spreading Forensics via Sequential Dependent Snapshots

Kechao Cai, *Member, IEEE*, Hong Xie, *Member, IEEE*, *ACM*,
and John C. S. Lui, *Fellow, IEEE*, *ACM*

***Abstract*— Mining the characteristics of information spreading in networks is crucial in communication studies, network security management, epidemic investigations, etc. Previous works are restrictive because they mainly focused on the information source detection using either a single observation, or multiple but *independent* observations of the underlying network while assuming a *homogeneous* information spreading rate. We conduct a theoretical and experimental study on information spreading, and propose a new and novel estimation framework to estimate 1) information spreading rates, 2) start time of the information source, and 3) the location of information source by utilizing *multiple sequential and dependent snapshots* where information can spread at heterogeneous rates. Our framework generalizes the current state-of-the-art rumor centrality [1] and the union rumor centrality [2]. Furthermore, we allow heterogeneous information spreading rates at different branches of a network. Our framework provides conditional maximum likelihood estimators for the above three metrics and is more accurate than rumor centrality and Jordan center in both synthetic networks and real-world networks. Applying our framework to the Twitter's retweet networks, we can accurately determine who made the initial tweet and at what time the tweet was sent. Furthermore, we also validate that the rates of information spreading are indeed heterogeneous among different parts of a retweet network.**

***Index Terms*— Information source estimation, information spreading forensics, sequential snapshots, conditional maximum likelihood estimators.**

I. Introduction

UNDERSTANDING information spreading in networks is a fundamental task in various aspects of human activities, e.g., advertisers would like to know how fast information spreads (*information spreading rates*) in different channels or communities in a network so as to design better network marketing strategies. Network security managers would like to find out when a computer virus starts to spread (*source start time*) so that they could rollback the system to a previous safe state to maintain a more reliable and trustworthy network. Epidemiologists would like to locate the patient zero (*information source*) in a social network so as to find out the reason for an epidemic. This form of study is called the *information spreading forensic*. However, such a forensic study is technically challenging in large-scale networks because the complete temporal knowledge of information spreading, i.e., the time index of when each individual (node) receives the information or gets infected is usually not available [3], and this makes information spreading forensics difficult. Moreover, a typical scenario of information spreading is that the source would spread information to different parts (or channels) of a network at different rates. For example, an epidemic usually has different spreading rates among different age groups [4], and news or rumors have different spreading rates among different communities [5]. Such a heterogeneity of spreading rates makes it more difficult to uncover the information spreading characteristics. In this work, we consider how to provide accurate estimates for the information spreading characteristics when we can have one or more sequential observations (or snapshots) of the information spreading process.

We propose a new and novel framework to estimate the information spreading rates, the source start time and the location of information source with "*sequential and dependent snapshots*". We consider an unknown source which starts spreading information at different spreading rates in a network. Specifically, the source first spreads information to different neighboring nodes at (potentially) different rates, and then each of these neighbors spreads to other nodes at the spreading rate inherited from the source. We assume that one can make sequential observations (or sequential snapshots) of the network at different times. The question is: *how can one accurately estimate the spreading rates, source start time, and identify the location of the information source based on these sequential snapshots?*

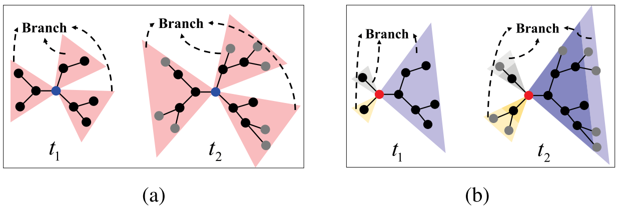
Previous work mainly focused on finding the location of the information source with a restrictive assumption of "*homogeneous*" information spreading rate in a network. Shah and Zaman [1] first proposed the *rumor centrality* estimator using a single snapshot of information spreading. Later on, Wang *et al.* [2] presented the *union* rumor centrality that utilizes multiple but *independent* snapshots of information spreading and claimed that the sequential dependent snapshots will not improve the accuracy of source detection. Both are based on the assumption that information spreads at a homogeneous spreading rate, and this is not realistic for networks with different communities/groups [4]. Indeed, given the homogeneous information spreading rate, the rumor centrality and the union rumor centrality both give a source estimate that balances the sizes of the branches of a spreading graph [1], [2].

Manuscript received November 6, 2016; revised June 27, 2017 and October 24, 2017; accepted December 10, 2017; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor X. Liu. Date of publication January 22, 2018; date of current version February 14, 2018. The work of J. C. S. Lui was supported in part by GRF under Grant 14630815 and in part by the Huawei's Research Grant. *(Corresponding author: John C. S. Lui.)*

The authors are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail: kccai@cse.cuhk.edu.hk; hxie@cse.cuhk.edu.hk; cslui@cse.cuhk.edu.hk).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors.

Digital Object Identifier 10.1109/TNET.2018.2791412



(a) (b)

Fig. 1. Infection graphs (snapshots taken) at time t_1 and t_2 (with $t_1 < t_2$). (a) Rumor centrality predicts that the "blue" node is the information source. (b) In reality, the "red" node is the true information source. The inaccuracy of (a) is due to the assumption that spreading rates on all branches are the same.

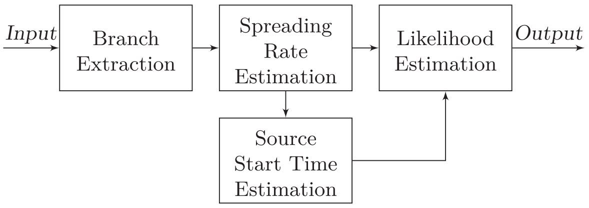


Fig. 2. Framework Diagram. *Input* are all the snapshots and the times when the snapshots were taken. *Output* are the conditional maximum estimates for the spreading rates, the source start time, and the source.

For example, Fig. 1(a) shows the information spreading graphs of two sequential snapshots taken at time t_1 and t_2 ($t_2 > t_1$), and the blue dot represents the source estimate by the rumor centrality. The ground truth, however, is shown in Fig. 1(b), which shows the same two snapshots of the information spreading process at time t_1 and t_2 ($t_2 > t_1$), where the *red dot is the true source* and different color-shaded triangles represent different branches. The rumor centrality or the union rumor centrality on these two snapshots would give a "*wrong estimate*" for the source's location (the blue dot in the left and right of Fig. 1(a), respectively) as they fail to capture the different growth sizes of branches in two sequential snapshots because they assumed the information spreading has the same rate on all branches. In contrast, we propose a general information spreading forensic model and propose a new and novel framework to estimate the different spreading rates, the source start time, and the location of the information source using sequential and dependent snapshots.

Our framework consists of four components as shown in Fig. 2. We take all the snapshots and the times as the *Input*. For each node in the first snapshot, at the *Branch Extraction* step, we extract its branches in the sequential snapshots. Then we examine the growth of each branch and give estimates for the spreading rates on different branches at the *Spreading Rate Estimation* step. Using the estimates for the rates, we estimate the source start time at the *Source Start Time Estimation* step. Finally, we calculate the likelihood of a node being the source at the *Likelihood Estimation* step. We obtain a likelihood estimation for each node in the first snapshot and then output the node with the conditional maximum likelihood as the source and give the corresponding estimates for the spreading rates and source start time.

Our key idea is that subsequent snapshots can reveal the information spreading rates of the branches detected in earlier snapshots when one examines the "growth size" of these branches. The spreading rate estimates on different branches further help us to infer the source start time and give a likelihood estimate for the source's location. We illustrate this idea via Fig. 1(b). As shown in the two sequential snapshots in Fig. 1(b), the branch (light blue shaded triangles in Fig. 1(b)) with the largest growth size should have the largest spreading rate estimate. Such spreading rates estimates indicate that the branch in light blue is highly likely the largest branch in the first snapshot. As such, we can estimate the relative sizes of branches rooted at the source, trace back to the start time and find the node (the red dot in Fig. 1(b)) in the first snapshot with the conditional maximum likelihood of a node generating such branches using the corresponding estimates of spreading rates and the estimate of start time.

Contributions: In this work, we prove that our framework generalizes both the rumor centrality [1] and the union rumor centrality [2] by allowing heterogeneous spreading rates at different branches, and we demonstrate that our framework improves the accuracy of source estimates compared with the state-of-the-art source estimators, namely, rumor centrality and Jordan center [6], [7]. In addition, our framework gives highly accurate estimates using conditional maximum likelihood estimators (CMLEs) for the information spreading rates and source start time. We validated these claims in our experiments, both for synthetic and real-world retweet networks in Twitter. Applying our framework to the Twitter's retweet networks, we can accurately determine who made the initial tweet and at what time the tweet was sent without any knowledge of the timestamps. Furthermore, we also discovered that the spreading rates are indeed heterogeneous among different parts of a retweet network and we provide accurate estimates of these spreading rates.

The organization of this paper is as follows. We present the information spreading model, observation model and our analysis framework in Sec. II. Derivation of various conditional maximum likelihood estimators are presented in Sec. III. In Sec. IV, we present experimental results of applying our framework to both synthetic networks and real-world networks. Related work is given in Sec. V. Sec. VI concludes the paper.

II. Models and Estimation Framework

We first present a model to characterize the information spreading process over a network, then we present an observation model to describe the snapshots that we can have on the information spreading process. Lastly, we present a general framework to estimate the spreading rates, the source start time, and the location of the information source.

A. Information Spreading Model

Consider an information spreading process over a network. The underlying network is modeled as an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote the vertex set and the edge set respectively. We use a continuous time Susceptible-Infected (SI) model [8] to describe the information spreading process. Compared with other spreading models such as the Susceptible-Infected-Recovered (SIR) model [8], the SI model is particularly suitable for modeling the information spreading process over networks, because a user either gets the information (i.e., gets *infected*) and then spreads the information to others in some time, or does not know about the information and thus won't spread the information (stays *susceptible*). Each node in $\mathcal{G}$ can be either *susceptible* (not getting the information) or *infected* (getting the information). Specifically, at an *unknown* time t_0, an *unknown* information source node $v^* \in \mathcal{V}$ starts spreading information in $\mathcal{G}$. Once a node is infected, it can infect (spread information to) a susceptible neighboring node and turn the neighboring node into an infected node. Consider an edge $(i,j) \in \mathcal{E}$, and suppose that node i is infected and node j is susceptible. Node i will infect node j after a random time θ_{ij}, which follows an exponential distribution with rate λ^{ij} (hereafter we say that node i infects node j at rate λ^{ij}). Furthermore, $\theta_{ij}, \forall(i,j) \in \mathcal{E}$ are independent distributed random variables. To construct the source estimator in a computationally tractable way, we consider the underlying graph $\mathcal{G}$ as an *infinite d-regular tree network* (each node has the same degree d and $d \geq 2$). The unknown source v^* infects each of its d neighbors with *unknown* rate $\lambda^1, \lambda^2, \ldots, \lambda^d$ at t_0 respectively. For subsequent infections, a node which was infected at rate λ^i would go on to infect its neighbors at rate λ^i for $1 \leq i \leq d$. Let $\boldsymbol{\lambda} = (\lambda^1, \lambda^2, \ldots, \lambda^d)$. Note that this continuous model was justified in [9] as a highly accurate probabilistic model to capture the interaction behaviors between users in social networks. Our objective is to give accurate estimates for the spreading rates $\boldsymbol{\lambda}$, the source start time t_0, and the information source v^* with a finite number of sequential snapshots on $\mathcal{G}$.

B. Observation Model

We now describe the observations that we can have on the information spreading process. We take *sequential snapshots* of the network $\mathcal{G}$ at different times during the spreading process. Each snapshot generates an infected graph which contains all infected nodes and infected edges (that connects the infected nodes) up to the time that snapshot was taken. More specifically, we consider m *sequential* snapshots ($m \geq 2$). Let $G_j \subseteq \mathcal{G}$ for $1 \leq j \leq m$ denote the j-th infected graph obtained at time t_j (where $t_0 < \cdots < t_j < \cdots < t_m$). Clearly, we have $G_j \subseteq G_m$ as these m snapshots are taken sequentially from the same spreading process on $\mathcal{G}$. Let $N_1 \leq \cdots \leq N_m$ be the number of infected nodes in $G_1, \ldots, G_m$ respectively.

C. Designing Estimation Framework

We now give the high level idea of our framework which provides estimates on the spreading rates ($\boldsymbol{\lambda}$) on each branches, the start time (t_0) and information source (v) using the branches split from the m snapshots for each node v in the first snapshot. Our framework consists of four components (described below) and takes the m sequential snapshots $G_1, \ldots, G_m$ and the times $t_1, \ldots, t_m$ as the input.

1) Branch Extraction Component: Upon taking the sequential snapshots, we further split them into d growing disjoint branches sharing no common nodes. Assume that the source node is v. Let $u_1, u_2, \ldots, u_d$ be the neighbors of v in $\mathcal{G}$. Let $T_v^i(t)$ denote the branch that is rooted at node v and does not contain u_{-i} (where $u_{-i} = \{\cup_1^d u_i\} \setminus u_i$) up to time t and $T_v^i(t_0) = v$. Thus, G_j is split into d different tree branches $T_v^i(t_j)$, $1 \leq i \leq d$, and each branch has a copy of source v. Let $k_j^i = |T_v^i(t_j)| - 1 \geq 0$ denote the number of infected nodes in $T_v^i(t_j)$ excluding v and $k_0^i = 0$ for $1 \leq i \leq d$ and $1 \leq j \leq m$. The increment size of $T_v^i(t)$ at two consecutive times t_{j-1} and t_j is denoted by δ_j^i, i.e., $\delta_j^i = k_j^i - k_{j-1}^i$. As the spreading process may have different spreading rates on the d branches, we denote that the spreading rate on the edges of the i-th branch as λ^i for $1 \leq i \leq d$. Moreover, for the branch $T_v^i(t_j)$, a node $u \in T_v^i(t_j)$ is a *boundary node* if u has no child in $T_v^i(t_j)$. We denote $B_v^i(t_j)$ as the boundary that consists of the boundary nodes of the branch $T_v^i(t_j)$ and let $b_j^i = |B_v^i(t_j)|$. Those are the nodes which are infected but have not infected other nodes. Thus, under this definition of boundary nodes, we can always sample b_{j-1}^i ($b_{j-1}^i \geq 0$) paths *that are disjoint with each other* as the branches are trees from the boundary $B_v^i(t_{j-1})$ to the boundary $B_v^i(t_j)$. Specifically, for each node u in $B_v^i(t_{j-1})$, we randomly select a path from u to $B_v^i(t_j)$. This way, we can sample b_{j-1}^i disjoint paths. We denote $l_r (0 \leq r \leq b_{j-1}^i)$ as the length of the r-th path in the b_{j-1}^i paths for $2 \leq j \leq m$ and $l_0 = 0$.

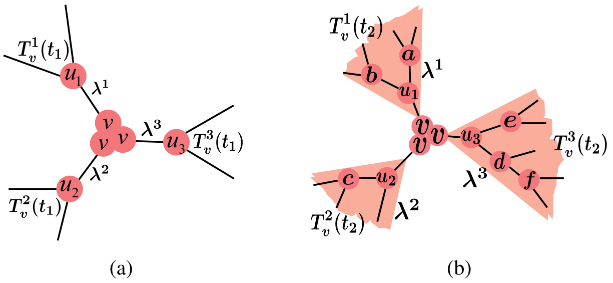


Fig. 3. Two sequential snapshots of a spreading process on a 3-regular tree where different branches have different spreading rates. (a) represents G_1, an infected tree at time t_1. (b) represents G_2, an infected tree with three larger branches at time t_2, with $t_2 > t_1$.

To illustrate, Fig. 3 shows two sequential snapshots of a spreading process on a 3-regular tree (or $d = 3$) where different branches have different spreading rates λ^1, λ^2 and λ^3. Assume that v is the source, Fig. 3(a) shows a snapshot G_1 taken at time t_1, and the branch $T_v^i(t_1)$ has nodes $\{v, u_i\}$ with $k_1^i = 1$ and spreading rate λ^i on each edge for $i \in \{1,2,3\}$. At time $t_2 > t_1$, we observe the network again and obtain the second snapshot G_2 as shown in Fig. 3(b). We see that the infected tree has grown and it has three larger and independent branches $T_v^i(t_2)$ for $1 \leq i \leq 3$, which are connected to the source node v with $k_2^1 = 3, k_2^2 = 2, k_2^3 = 4$ and the size increments $\delta_2^1 = 2, \delta_2^2 = 1, \delta_2^3 = 3$. u_3 is a boundary node of $T_v^3(t_1)$, i.e., $B_v^3(t_1) = \{u_3\}$. Node e and f are the boundary nodes of $T_v^3(t_2)$, i.e., $B_v^3(t_2) = \{e, f\}$. From $B_v^3(t_1)$ to $B_v^3(t_2)$, we can sample a single path, e.g., the path $u_3 - e$, and the length of this path is one.

2) Spreading Rate Estimation Component: This component provides point estimates for the spreading rates on different branches given that v is the source. Consider the branch T_v^i. We have b_{j-1}^i *disjoint* sample paths that connect the boundary $B_v^i(t_{j-1})$ with the boundary $B_v^i(t_j)$. Thus, the spreading process on each of the b_{j-1}^i paths is a Poisson process with rate λ^i and the spreading processes on different paths are independent. The estimator for $\lambda^i (1 \leq i \leq d)$ is:

$$\hat{\lambda}^i = \arg\max_{\lambda^i} \prod_{j=2}^{m} \frac{[\lambda^i(t_j - t_{j-1})]^{\sum_{r=0}^{b_{j-1}^i} l_r}}{e^{\lambda^i(t_j - t_{j-1}) b_{j-1}^i} \prod_{r=0}^{b_{j-1}^i} l_r!},$$

where the argument of $\arg\max$ is simply the joint density function of observing all the b_{j-1}^i disjoint paths in $T_v^i(t)$ during $(t_{j-1}, t_j]$ for $2 \leq j \leq m$. By letting the derivative of λ^i be 0, we obtain the spreading rate estimator $\hat{\lambda}^i$ as follows,

$$\hat{\lambda}^i = \frac{\sum_{j=2}^m \sum_{r=0}^{b_{j-1}^i} l_r}{\sum_{j=2}^m b_{j-1}^i (t_j - t_{j-1})}, \tag{1}$$

if $\sum_{j=2}^m b_{j-1}^i (t_j - t_{j-1}) \neq 0$. Otherwise, $\hat{\lambda}^i = 0$. Eq. (1) shows that the spreading rate estimator $\hat{\lambda}^i$ is in fact the average spreading rate of information spreading on all disjoint paths in the $m-1$ snapshots. Let $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^1, \ldots, \hat{\lambda}^d)$.

3) Source Start Time Estimation Component: Assume that the source is v, to estimate the source start time t_0, we consider the distribution of time $t_j := t_j - t_0$ for the information spreading from v to the branch $T_v^i(t_j)$ given that the spreading rate in the branch is $\hat{\lambda}^i$, $1 \leq i \leq d$, during $(t_0, t_j]$ for $1 \leq j \leq m$. Note that there are k_j^i infections that occur during $(t_0, t_j]$ on the branch $T_v^i(t_j)$. Moreover, the spreading time on each edge follows an exponential distribution with rate $\hat{\lambda}^i$. Each new infection would introduce new infectious edges thereby increasing the spreading rate of new infections (see details in Lemma 1). Hence, the total time $S_{k_j^i}(t)$ of infecting k_j^i nodes on the branch $T_v^i(t_j)$ is the sum of the k_j^i exponentially distributed random variables with increasing rates. Let $\hat{t}_0$ denote the estimator for t_0 and $\mathbf{P}_{k_j^i}(t_j|t_0)$ denote the probability that exactly k_j^i infections occur in the branch $T_v^i(t_j)$ during $(t_0, t_j]$ given that spreading rate is $\hat{\lambda}^i$, $1 \leq i \leq d$ (see details in Lemma 2). We notice that the dependence between the second snapshot and the subsequent snapshots does not contribute to the estimation of t_0 because spreading rates on the edges of branch $T_v^i(t)$ follow exponential distribution which has the memoryless property. Therefore, only the branch $T_v^i(t_1)$ is useful when using adjacent snapshots to estimate t_0. To give an estimate for t_0 while taking the branches of all the m snapshots into account, we consider the m spreading process instances in the m overlapping intervals $(t_0, t_j]$ $(1 \leq j \leq m)$ other than the m intervals $(t_{j-1}, t_j]$ $(1 \leq j \leq m)$.

Thus, we obtain the source start time estimator $\hat{t}_0$ as follows,

$$\hat{t}_0 = \arg\max_{t_0} \prod_{j=1}^m \prod_{i=1}^d \mathbf{P}_{k_j^i}(t_j|t_0). \tag{2}$$

We give the explicit form of $\hat{t}_0$ in Proposition 1.

Proposition 1: Given that the underlying graph is a d-regular tree, $v \in G_1$ is the source and the estimated spreading rates are $\hat{\lambda}^i$ for $1 \leq i \leq d$, the source start time estimator $\hat{t}_0$ is given by

$$\hat{t}_0 = \frac{1}{m} \sum_{j=1}^m \Big(t_j - \frac{1}{\sum \mathbb{1}_{\hat{\lambda}^i > 0}} \sum_{i=1, \hat{\lambda}^i > 0}^d \frac{\ln(1 + a k_j^i)}{a \hat{\lambda}^i} \Big), \tag{3}$$

where $a = d - 2$ $(d > 2)$, $\mathbb{1}_{\hat{\lambda}^i > 0}$ is an indicator function. For $d = 2$, we have $\hat{t}_0 = \frac{1}{m} \sum_{j=1}^m \big(t_j - \frac{1}{\sum \mathbb{1}_{\hat{\lambda}^i > 0}} k_j^i / \hat{\lambda}^i \big)$.

Proof: The result is derived from a special case in Lemma 2. See the proof for Lemma 2 in Sec. III-A. ■

Remark: Eq. (3) implies that the source start time estimator $\hat{t}_0$ is the average of the difference between t_j, $j = 1, \ldots, m$ and the average spreading time from source v to the boundaries of different branches.

4) Likelihood Estimation Component: Our straightforward objective is to give estimates for the spreading rates, source start time and the information source given the observed snapshots as shown in Eq. (4).

$$\begin{aligned} \{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} &= \arg\max_{v \in G_1, \boldsymbol{\lambda}, t_0} \mathbf{P}[v, \boldsymbol{\lambda}, t_0 | G_1, \ldots, G_m] \qquad (4) \\ &= \arg\max_{v \in G_1, \boldsymbol{\lambda}, t_0} \mathbf{P}[G_1, \ldots, G_m | v, \boldsymbol{\lambda}, t_0]. \qquad (5) \end{aligned}$$

Computing the probability in Eq. (4) is infeasible in general. Thus we assume each node in the first snapshot G_1 is equally likely to be the source, and we do not have any prior knowledge about the spreading rates $\boldsymbol{\lambda}$ and the source start time t_0, and treat $\boldsymbol{\lambda}$ and t_0 as independent variables. Thus $\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\}$ are exactly the maximum likelihood estimators of information source v, spreading rates $\boldsymbol{\lambda}$ and source start time t_0. This way, we convert the straightforward objective in Eq. (4) to the objective in Eq. (5) using Bayesian transformation.

However, v, $\boldsymbol{\lambda}$, and t_0 are *not independent*. As we have discussed in Sec. II-C.2 and Sec. II-C.3, $\boldsymbol{\lambda}$ depends on v ($\hat{\boldsymbol{\lambda}} := \boldsymbol{\lambda}(v)$), and t_0 depends on both $\boldsymbol{\lambda}$ and v ($\hat{t}_0 := t_0(\boldsymbol{\lambda}(v), v)$). We take advantage of such dependence among $v, \boldsymbol{\lambda}$, and t_0, and use the two estimators, $\hat{\boldsymbol{\lambda}}$ and $\hat{t}_0$, derived from the *spreading rate estimation component* and the *source start time estimation component* to give *conditional maximum likelihood estimators* (CMLEs) for v, $\boldsymbol{\lambda}$, and t_0 as follows,

$$\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} = \arg\max_{v \in G_1} \mathbf{P}[G_1, \ldots, G_m | v, \boldsymbol{\lambda}(v), t_0(\boldsymbol{\lambda}(v), v)]. \tag{5'}$$

Now we calculate the probability in Eq. (5'). Recall that each node v in G_1 is equally likely to be the source. We examine the spreading process by considering the information spreading on each branch $T_v^i(t)$ at the spreading rate estimate $\hat{\lambda}^i$ during the time intervals $(\hat{t}_0, t_1]$, $(t_1, t_2]$, ... and $(t_{m-1}, t_m]$ separately, $1 \leq i \leq d, 1 \leq j \leq m$. Moreover, the spreading process on each branch under our assumption is independent of each other and the branch $T_v^i(t_j)$ is only dependent on its earlier state, i.e., $T_v^i(t_{j-1})$ $(1 \leq j \leq m)$. Therefore, the conditional maximum likelihood estimators, $\hat{v}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{t}_0$ in Eq. (5') can be expressed as:

$$\begin{aligned} \{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} &= \arg\max_{v \in G_1} \prod_{i=1}^d \mathbf{P}[T_v^i(t_m), \ldots, T_v^i(t_1) | v, \hat{\boldsymbol{\lambda}}, \hat{t}_0] \qquad (6) \\ &= \arg\max_{v \in G_1} \prod_{i=1}^d \prod_{j=1}^m \mathbf{P}[T_v^i(t_j) | T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]. \qquad (7) \end{aligned}$$

where $\mathbf{P}[T_v^i(t_m), \ldots, T_v^i(t_1) | v, \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ is the joint probability (likelihood) that we observe the branches $T_v^i(t_j)$ at t_j for $1 \leq j \leq m$ given the source $v = T_v^i(\hat{t}_0)$ and the spreading rates $\hat{\boldsymbol{\lambda}}$, while $\mathbf{P}[T_v^i(t_j) | T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ is the conditional probability of observing the branch $T_v^i(t_j)$ at t_j given that the branch $T_v^i(t_{j-1})$ is observed at t_{j-1} with the source being v and the spreading rates being $\hat{\boldsymbol{\lambda}}$. Note that $\mathbf{P}[T_v^i(t_j) | T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ is dependent on the size of $T_v^i(t_{j-1})$, which is dependent on the location of the source v in G_1, the spreading rates $\hat{\boldsymbol{\lambda}}$, and the source start time $\hat{t}_0$. Any other node v' in G_1 would result in different branches $T_{v'}^i(t_{j-1})$ and give different conditional probabilities $\mathbf{P}[T^i_{v'}(t_j)|T^i_{v'}(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ for $1 \leq j \leq m$. As such, the dependence of the branch $T^i_v(t_1)$ in G_1 on v carries over to the branches $T^i_v(t_j)$ in snapshots G_j for $2 \leq j \leq m$. Thus the branches $T^i_v(t_j)$ in snapshots G_j for $2 \leq j \leq m$ will influence the estimates for the spreading rates $\hat{\boldsymbol{\lambda}}$, the source start time $\hat{t}_0$ and the source v in G_1. For every v in G_1, we give such a likelihood estimate. Therefore, we can give conditional maximum likelihood estimates for the three information spreading characteristics that contribute to the formation of the branches in the first snapshot and branches in the sequential snapshots.

III. Derivation of Estimators

Here we derive explicit expressions for the conditional maximum likelihood estimators stated in Eq. (7), namely, the spreading rate $\hat{\boldsymbol{\lambda}}$, the source start time $\hat{t}_0$ and the source $\hat{v}$ stated in Eq. (7). We also present a parallel algorithm used in the estimation components and give a detailed complexity analysis of the algorithm. To convey the core idea, we use an example to illustrate the benefit of two sequential snapshots in information source detection. Since the infected graph has d independent branches, we analyze the infection process as d independent infection processes. We first study the infection process in a single branch, and then combine these infection processes together to construct our estimators.

A. Information Spreading on a Single Branch

To begin, we study the infection process by dividing the infection process into m sub-processes based on time step t_{j-1} and t_j for $1 \leq j \leq m$. We first derive an explicit expression for $\mathbf{P}[T^i_v(t_j)|T^i_v(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ by inspecting the infection process on the i-th branch $T^i_v(t)$ for $t \in (t_{j-1}, t_j]$ with v being the source. $\mathbf{P}[T^i_v(t_j)|T^i_v(t_{j-1}), , \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ is actually the product of the number of "*permitted spreading orders*" and the probability of a permitted spreading order from $T^i_v(t_{j-1})$ to $T^i_v(t_j)$ since the permitted spreading orders from $T^i_v(t_{j-1})$ to $T^i_v(t_j)$ are equally likely and mutually exclusive because the underlying graph $\mathcal{G}$ is a d-regular tree and the infection rates on edges are modeled as identical distributed exponential random variables on each branch.

Definition 1: A **Permitted Spreading Order** is a valid spreading order of nodes constrained by the structure of the infected graph. E.g., let G be an infected graph with the root node being the source v. For a node u to be infected, its parent node u_p has to be infected first. So the permitted spreading order starts with v, and with u_p ahead of u.

To illustrate, consider the snapshot G_2 in Fig. 3(b). Suppose that v is the source, then one possible permitted spreading order in G_2 is v–u_1–u_2–a–u_3–b–c–e–d–f. More generally, a permitted spreading order from the branch $T^i_v(t_{j-1})$ to the branch $T^i_v(t_j)$ should always start from the nodes in $T^i_v(t_{j-1})$ which have infectious edges connecting to other susceptible nodes in $T^i_v(t_j)$. For example, in Fig. 3, one possible permitted spreading order from the branch $T^1_v(t_1)$ to the branch $T^1_v(t_2)$ is $u_1 - b - a$, where u_1 has two infectious edges in $T^1_v(t_1)$.

To compute the number of permitted spreading orders starting from $T^i_v(t_{j-1})$ to $T^i_v(t_j)$, we consider the branch $T^i_v(t_{j-1})$ as a *super source node* v^i_{j-1}, which consists of all the infected nodes and edges in $T^i_v(t_{j-1})$. Then the branch $T^i_v(t_j)$ becomes a *super branch* $\bar{T}^i_{v^i_{j-1}}(t_j)$ taking the branch $T^i_v(t_{j-1})$ as a single node v^i_{j-1}. Thus, the number of nodes in $\bar{T}^i_{v^i_{j-1}}(t_j)$ excluding the super node v^i_{j-1} is equal to $(k^i_j - k^i_{j-1})$. We can compute the number of permitted spreading orders in the super branch $\bar{T}^i_{v^i_{j-1}}(t_j)$ via the rumor centrality in [1].

*Definition 2 (**Rumor Centrality. [1]**):* Rumor centrality for a node v in a given tree G is the number of permitted spreading orders starting with v in G, which is given by

$$R(v, G) = \frac{N!}{\prod_{w \in G} T_{w|v}}, \tag{8}$$

where N is the number of nodes in G, $T_{w|v}$ is the number of nodes in the subtree rooted at node w and pointing away from v, with v being the source.

Example: consider the snapshot G_2 in Fig. 3(b). The sizes of the subtrees rooted at v are as follows: $T_{v|v} = 10, T_{u_1|v} = 3$, $T_{u_2|v} = 2, T_{u_3|v} = 4$, $T_{d|v} = 2, T_{a|v} = T_{b|v} = T_{c|v} = T_{e|v} = 1$. Therefore, the rumor centrality for node v is: $R(v, G_2) = \frac{10!}{10\times3\times2\times4\times2\times1\times1\times1\times1} = 7560$. We can also calculate the number of permitted spreading orders in the super branches using rumor centrality. For example, consider the branches $T^1_v(t_1)$ and $T^1_v(t_2)$ in Fig. 3. We have $k^1_2 - k^1_1 = 2$. Taking $T^1_v(t_1)$ as a super node v^1_1, the sizes of the subtrees in $\bar{T}^1_{v^1_1}(t_2) \setminus T^1_{v^1_1}(t_1)$ are $T_{a|v^1_1} = 1, T_{b|v^1_1} = 1$. Thus we can obtain $R(v^1_1, \bar{T}^1_{v^1_1}(t_2)) = \frac{2!}{1\times1} = 2$.

Note that $\bar{T}^i_{v^i_{j-1}}(t_j)$ is a realization of the spreading process on a branch rooted at v on the underlying graph $\mathcal{G}$ at t_j. To derive the probability of observing such a realization, we analyze how the $\delta^i_j = k^i_j - k^i_{j-1}$ infected nodes and the super node v^i_{j-1} form the super branch $\bar{T}^i_{v^i_{j-1}}(t_j)$. Specifically, starting from v^i_{j-1}, each infection on the super branch $\bar{T}^i_{v^i_{j-1}}(t_j)$ will introduce new infectious edges to the boundary nodes, creating new opportunities of infecting other susceptible nodes on the boundary and forming possibly different infection graphs other than $\bar{T}^i_{v^i_{j-1}}(t_j)$. Therefore, given that δ^i_j nodes are infected on a d-regular tree, the probability that we can observe these infected nodes form the super branch $\bar{T}^i_{v^i_{j-1}}(t_j)$ is dependent on how the infectious edges on the boundary infect new nodes. Let β^i_j denote the probability that δ^i_j nodes form the super branch $\bar{T}^i_{v^i_{j-1}}(t_j)$. In the following lemma, we provide an explicit expression of β^i_j.

Lemma 1: Given that δ^i_j ($\delta^i_j = k^i_j - k^i_{j-1} \geq 0$) infections occurred after observing the branch $T^i_v(t_{j-1})$ in a d-regular tree according to the information spreading model described in Sec. II-A, the probability that the δ^i_j infected nodes form the branch $\bar{T}^i_{v^i_{j-1}}(t_j)$ is

$$\beta^i_j = \begin{cases} 1, & \delta^i_j = 0, \\ \displaystyle\prod_{r=k^i_{j-1}}^{k^i_j - 1} \frac{1}{(1+ra)}, & \delta^i_j \geq 1, \end{cases} \tag{9}$$

for $1 \leq i \leq d$ and $1 \leq j \leq m$ where $a = d - 2$.

Proof: As we are proving common properties of the d branches, here we drop the superscript "i" on $\hat{\lambda}^i$, v_{j-1}^i, $\bar{T}_{v_{j-1}^i}^i(t_j)$, β_j^i, k_{j-1}^i and k_j^i unless a distinction needs to be made. Without loss of generality, we study the r-th ($k_{j-1} \leq r \leq k_j - 1$) infection in the super branch $\bar{T}_{v_{j-1}}(t_j)$. Note that every new infection would introduce $d-2$ new infectious edges (through which the information can infect new nodes). Thus there are $1+r(d-2)$ infectious edges at the beginning of r-th infection. That is, there are $1+r(d-2)$ ways to infect new node at the beginning of r-th infection. Therefore, the $k_j - k_{j-1}$ infections could happen in $\prod_{r=k_{j-1}}^{k_j-1}[1+r(d-2)]$ of possible ways where the branch $\bar{T}_{v_{j-1}}(t_j)$ is a single realization. Thus the probability that the $k_j - k_{j-1}$ infections forms the branch $\bar{T}_{v_{j-1}}(t_j)$ can be given by Eq. (9) by adding back the super script "i". This finishes the proof. ■

We now derive the probability of a specific permitted spreading order for the spreading process on the super branch $\bar{T}_{v_{j-1}^i}^i(t_j)$. For each permitted spreading order in $\bar{T}_{v_{j-1}^i}^i(t_j)$, there is a sequence of δ_j^i infections that occur during $(t_{j-1}, t_j]$. We denote the probability that a sequence of $k_j^i - k_{j-1}^i$ infections occur in the super branch $\bar{T}_{v_{j-1}^i}^i(t_j)$ during interval $(t_{j-1}, t_j]$ given that k_{j-1}^i infections have occurred before t_{j-1} as $\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})$.

This is exactly the probability of a single permitted spreading order on $\bar{T}_{v_{j-1}^i}^i(t_j)$. Thus by Lemma 1, the probability that δ_j^i infections occur during $(t_{j-1}, t_j]$ as well as forming the super branch $\bar{T}_{v_{j-1}^i}^i(t_j)$ for a single permitted order should be $\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i$. Note that there are $R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j))$ permitted spreading orders on $\bar{T}_{v_{j-1}^i}^i(t_j)$. As such, by multiplying such a probability by the number of permitted spreading orders on $\bar{T}_{v_{j-1}^i}^i(t_j)$, we have:

$$\mathbf{P}[T_v^i(t_j)|T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0] = R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j))\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i. \quad (10)$$

Finally, $\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})$ can be stated as follows.

Lemma 2: For the super branch $\bar{T}_{v_{j-1}^i}^i(t_j)$ in a d-regular tree, the probability that a sequence of δ_j^i ($\delta_j^i \geq 0$) infections occur at the rate $\hat{\lambda}^i > 0$ during $(t_{j-1}, t_j]$ is given by

$$\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1}) = \begin{cases} e^{-(1+k_{j-1}^i a)\hat{\lambda}^i t}, & \delta_j^i = 0, \\ \dfrac{(1-e^{-a\hat{\lambda}^i t})^{\delta_j^i}}{a^k \delta_j^i! e^{(1+k_{j-1}^i a)\hat{\lambda}^i t}} \prod\limits_{r=k_{j-1}^i}^{k_j^i-1}(1+ra), & \\ & \delta_j^i \geq 1, \end{cases} \quad (11)$$

for $1 \leq j \leq m$, where $a = d-2$ and $t = t_j - t_{j-1}$.

Proof: We refer the interested readers to our supplementary material available alongside the paper for the detailed proof.

B. Information Spreading on d Branches

Now we can combine the d spreading processes on the d branches together. Putting Eq. (10) into Eq. (7), we can see the conditional maximum likelihood estimators are equivalent to:

$$\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} = \arg\max_{v \in G_1} \prod_{i=1}^{d} \prod_{j=1}^{m} R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j)) \cdot \mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i. \quad (12)$$

Note that there is an interesting relationship between our source estimator in Eq. (12) and the previous work on rumor centrality [1]. To reveal the relationship, consider the snapshot G_{j-1} as a super node $v_{G_{j-1}}$ ($2 \leq j \leq m$). The difference graph $G_j - G_{j-1}$ together with those edges that connect to G_{j-1} (i.e., super node $v_{G_{j-1}}$) form a super graph $\bar{G}_j$. $\bar{G}_j$ contains d groups of branches. Branches in the i-th group share a common ancestor node v. We now show the relation between $R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j))$ and $R(v_{G_{j-1}}, \bar{G}_j)$ in the following lemma.

Lemma 3: For the d super branches in the super graph $\bar{G}_j$, the number of permitted spreading orders in each branch satisfies

$$\prod_{i=1}^{d} \frac{R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j))}{\delta_j^i!} = \frac{R(v_{G_{j-1}}, \bar{G}_j)}{(N_j - N_{j-1})!}, \quad (13)$$

for $1 \leq j \leq m$, where we define $R(v_{G_0}, \bar{G}_1) = R(v, G_1)$ and $R(v_{G_{j-1}}, \bar{G}_j)$ is the rumor centrality of the super node $v_{G_{j-1}}$ in the super graph $\bar{G}_j$. Moreover, $R(v_{G_{j-1}}, \bar{G}_j)$ is simply a ***constant*** regardless of which node is the source in G_1 given sequential snapshots G_{j-1} and G_j for $2 \leq j \leq m$.

Proof: We refer the interested readers to our supplementary material available alongside the paper for the detailed proof.

Likelihood Estimation: Applying the results in the Lemma 1 to Lemma 3, we give our conditional maximum likelihood estimators $\hat{v}$, $\hat{\boldsymbol{\lambda}}$ and $\hat{t}_0$ in Eq. (12) for m sequential snapshots in the following theorem.

Theorem 1: Given that the underlying d-regular tree graph, for sequential snapshots G_1, G_2,...,G_m, which are taken at t_1, t_2, ..., t_m respectively ($\hat{t}_0 < t_1 < t_2 < \ldots < t_m$), the conditional maximum likelihood estimation for the source $\hat{v}$, the source spreading rates $\hat{\boldsymbol{\lambda}}$ and the source start time $\hat{t}_0$ are

$$\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} = \arg\max_{v \in G_1} R(v, G_1) \cdot C(\hat{\boldsymbol{\lambda}}, \hat{t}_0), \quad (14)$$

where $C(\hat{\boldsymbol{\lambda}}, \hat{t}_0)$ is given by

$$e^{-\sum\limits_{i=1}^{d}[(ak_m^i+1)t_m - \hat{t}_0]\hat{\lambda}^i} \prod_{i=1, \hat{\lambda}^i > 0}^{d} \prod_{j=1}^{m} \left(e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}}\right)^{\delta_j^i},$$

and $a = d-2$, where $\hat{\lambda}^i$ and $\hat{t}_0$ are calculated by Eq. (1) and Eq. (3) respectively and only the branches with nonzero spreading rates ($\hat{\lambda}^i > 0$) are considered in the product terms.

Proof: Using the results in Lemma 1 to Lemma 3, the source estimator in Eq. (12) can be simplified as follows,

$$\begin{aligned} &\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} \\ &= \arg\max_{v \in G_1} \prod_{i=1}^{d} \prod_{j=1}^{m} R(v_{j-1}^i, \bar{T}_{v_{j-1}^i}^i(t_j)) \cdot \mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i \\ &\overset{(a)}{=} \arg\max_{v \in G_1} \prod_{i=1}^{d} \prod_{j=1}^{m} R(v_{G_{j-1}}, \bar{G}_j)\mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i \cdot \delta_j^i! \\ &\overset{(b)}{=} \arg\max_{v \in G_1} R(v, G_1) \prod_{i=1}^{d} \prod_{j=1}^{m} \mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\beta_j^i \cdot \delta_j^i!. \quad (15) \end{aligned}$$

In above, (a) is from the results in Eq. (13) and (b) comes from the fact that $R(v_{G_{j-1}}, \bar{G}_j)$ is a constant for $2 \leq j \leq m$ derived in Lemma 3. From Eq. (9) in Lemma 1 and Eq. (11) in Lemma 2, we have

$$
\begin{aligned}
&\prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} \mathbf{P}_{k_j^i|k_{j-1}^i}(t_j|t_{j-1})\delta_j^i! \cdot \beta_j^i \\
&= \prod_{i=1,\hat{\lambda}_i>0}^{d} \prod_{j=1}^{m} \frac{(1-e^{-a\hat{\lambda}_i(t_j-t_{j-1})})^{\delta_j^i}}{e^{(1+k_{j-1}^i a)\hat{\lambda}_i(t_j-t_{j-1})} \cdot a^{\delta_j^i}} \\
&= \frac{e^{-(t_m-\hat{t}_0)\sum\limits_{i=1}^{d}\hat{\lambda}^i}}{a^{N_m-1}} \prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} \frac{(1-e^{-a\hat{\lambda}^i(t_j-t_{j-1})})^{\delta_j^i}}{e^{a\hat{\lambda}^i k_{j-1}^i(t_j-t_{j-1})}} \\
&= \prod_{i=1,\hat{\lambda}_i>0}^{d} \prod_{j=1}^{m} \frac{(1-e^{-a\hat{\lambda}_i(t_j-t_{j-1})})^{\delta_j^i}}{e^{(1+k_{j-1}^i a)\hat{\lambda}_i(t_j-t_{j-1})} \cdot a^{\delta_j^i}} \\
&= \frac{e^{-(t_m-\hat{t}_0)\sum\limits_{i=1}^{d}\hat{\lambda}^i}}{a^{N_m-1}} \prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} \frac{(e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}})^{\delta_j^i}}{e^{a\hat{\lambda}^i(t_j k_j - t_{j-1}k_{j-1})}} \\
&= \frac{e^{-(t_m-\hat{t}_0)\sum\limits_{i=1}^{d}\hat{\lambda}_i}}{a^{N_m-1}} \prod_{i=1,\hat{\lambda}_i>0}^{d} e^{-a\hat{\lambda}_i t_m k_m^i} \prod_{j=1}^{m} \\
&\quad \times (e^{a\hat{\lambda}_i t_j} - e^{a\hat{\lambda}_i t_{j-1}})^{\delta_j^i} \\
&= \frac{e^{-\sum_{i=1}^{d}[(ak_m^i+1)t_m-\hat{t}_0]\hat{\lambda}^i}}{a^{N_m-1}} \prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} (e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}})^{\delta_j^i} \\
&= C(\hat{\boldsymbol{\lambda}}, \hat{t}_0)/a^{N_m-1}, \qquad (16)
\end{aligned}
$$

where

$$
C(\hat{\boldsymbol{\lambda}}, \hat{t}_0) = e^{-\sum\limits_{i=1}^{d}[(ak_m^i+1)t_m-\hat{t}_0]\hat{\lambda}^i} \prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} \times (e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}})^{\delta_j^i},
$$

and $a = d - 2$. Note that $C(\hat{\boldsymbol{\lambda}}, \hat{t}_0)$ is a scaling factor that depends on the source estimation $\hat{v}$, the nonzero spreading rates estimations $\hat{\boldsymbol{\lambda}}$ ($\hat{\lambda}^i > 0$), as well as the start time estimation $\hat{t}_0$ given the m snapshots. Thus by putting Eq. (16) back into Eq. (15) and taking $N_j (1 \leq j \leq m)$ as constants, we obtain the source estimator given in Theorem 1. ∎

Remark: Our conditional maximum likelihood estimators provide a source estimation that has an *additional scaling factor* as compared with the previous work of rumor centrality [1]. In particular, we have the following result.

Corollary 1: Suppose that the spreading rates $\lambda^1, \lambda^2, \ldots, \lambda^d$ are equal on all branches, i.e., $\lambda^1 = \lambda^2 = \ldots = \lambda^d$, then the conditional maximum likelihood estimator in Theorem 1 becomes,

$$
\hat{v} = \arg\max_{v \in G_1} R(v, G_1),
$$

which is equivalent to the rumor centrality in [10]. Moreover, the homogeneous spreading rate and source start time can be given by Eq. (1) and Eq. (3) respectively.

Proof: We refer the interested readers to our supplementary material available alongside the paper for the detailed proof. The above corollary states that our framework produces at least as good a source node estimator as the rumor centrality when we have a homogeneous spreading rate, but when spreading rates are different, our framework produces more accurate estimates than previous state-of-the-art schemes as shown in Sec. IV.

C. Extending to General Networks

Our framework provides estimates not only in d-regular trees but also for general networks. We like to point out that there is an *underlying information spreading tree* which corresponds to the first time each node gets infected for an information spreading process. Such a tree is also termed as the word-of-mouth propagation tree in the literature, e.g, [11]. Therefore, for general networks, one can use the *Breadth-First-Search (BFS) trees* of the snapshots to approximate the information spreading trees. Specifically, for our framework (as well as the rumor centrality) that can only perform estimation on trees, we pick each node in G_1 as a candidate source estimate, use the *maximum degree* of the nodes in G_1 as d in our framework in Eq. (14) (Note that we do not assume that the node of the maximum degree is the source node.) and use the corresponding BFS trees (rooted at the picked node) of the sequential snapshots to approximate the underlying information spreading tree.

D. Algorithm & Complexity Analysis

Algorithm 1 shows the implementation of our framework. Specifically, in line 1, for each v node in G_1, we build BFS trees of the m sequential snapshots $G_1 \ldots, G_m$ and extract branches from the built BFS trees (see line 2 to line 8). Then we sample disjoint paths and estimate the spreading rate on each branch (see line 9 to line 16). We average the spreading time on different branches and give an estimate for the source start time (see line 17 to line 21). After that, we calculate the likelihood of node v being the source (see line 22 to line 27). Each iteration in the for-loop in line 1 is independent of other iteration. Thus Algorithm 1 can be parallelized by dividing the nodes in G_1 into several groups and calculating the likelihoods for each node in each group in parallel. Finally, we take the node v and corresponding spreading rates and source start time with the conditional maximum likelihood as the output $\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\}$ of our framework (line 29).

Complexity Analysis: The total computational complexity of Algorithm 1 is $O(N_1 N_m)$. Specifically, for each round in Algorithm 1, the branch extraction step requires $O(N_m)$ operations as it involves breadth-first-search operations on the m sequential snapshots, and the spreading rates estimation step requires $O(N_m)$ computations. The computational complexity of source start time can be neglected as only d computations are needed. Moreover, the rumor centrality of each node in G_1 can be pre-computed with complexity $O(N_1)$, and the computational complexity of the scaling factors is also $O(N_1)$. We run above procedures in Algorithm 1 for N_1 times to evaluate the likelihoods for all the nodes in G_1. Thus the total computational complexity of our framework is $O(N_1 N_m)$, and it provides *conditional maximum likelihood estimates* for the

Algorithm 1 Est. Framework With Sequential Snapshots

```
Input: $G_1, G_2, \ldots, G_m$ and $t_1, t_2, \ldots, t_m$
for $v \in G_1$ do
    Branch Extraction. Construct $m$ BFS trees $T_{1,\ldots,m}$ from $v$ to other nodes in $G_1, \ldots, G_m$ sequentially.
    Initialize $u_1, \ldots, u_d$ as the $d$ neighbors of $v \in T_m$, $d$ is the maximum degree in $T_m$.
    for $i = 1, \ldots, d$ do
        for $j = 1, \ldots, m$ do
            $T_v^i(t_j) =$ subtree in $T_j$ rooted at $u_i$
            $k_i^j = |T_v^i(t_j)|$
            $B_v^i(t_j) =$ nodes that have no child in $T_v^i(t_j)$
    Spreading Rate Estimation. Sample $b_{j-1}^i = |B_v^i(t_j)|$ paths from $B_v^i(t_{j-1})$ to $B_v^i(t_j)$ for $1 \leq i \leq d$ and $2 \leq j \leq m$. The length of the $r$-th path is set as $l_r$ ($1 \leq r \leq b_{j-1}^i$).
    Set $\hat{\lambda}^i = 0$, $t_s = 0$ and $l_s = 0$.
    for $i = 1, \ldots, d$ do
        for $j = 2, \ldots, m$ do
            $t_s = t_s + b_{j-1}^i(t_j - t_{j-1})$
            for $r = 1, \ldots, b_{j-1}^i$ do
                $l_s = l_s + l_r$
        $\hat{\lambda}^i = l_s / t_s$
    Source Start Time Estimation. Set $t = 0$, $c = 0$
    for $i = 1, \ldots, d$ do
        if $\hat{\lambda}^i > 0$ then
            $t = t + \ln(1 + (d-2)k_1^i/(\hat{\lambda}^i(d-2)), c = c + 1$
    $\hat{t}_0 = t_1 - t/c$.
    Likelihood Estimation. Calculate the rumor cen-. trality $v.rc$ of node $v$ in $G_1$ (See details in [10]). Set $a = d - 2$ and $t_0 = \hat{t}_0$
    $C_\lambda = e^{-\sum_{i=1}^{d}[(ak_m^i+1)t_m - \hat{t}_0]\hat{\lambda}^i}$
    for $i = 1, \ldots, d$ do
        if $\hat{\lambda}^i > 0$ and $d > 2$ then
            for $j = 1, \ldots, m$ do
                $C_\lambda = C_\lambda \cdot (e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}})^{\delta_j^i}$
    $v.e = v.rc \cdot C_\lambda$
Output: $v$, $\hat{t}_0$ and $\hat{\lambda}^1, \ldots, \hat{\lambda}^d$ with the conditional maximum likelihood
```

location of information source, the spreading rates and source start time using m sequential snapshots.

E. Illustration via an Example

Fig. 4 illustrates an example of information spreading on a 4-regular tree where the true source is v', the true source start time $t_0 = 0.7$, and the true spreading rates $\lambda^1 = 0.1$, $\lambda^2 = 1.5$, $\lambda^3 = 0.1$ and $\lambda^4 = 1$. The infected nodes observed in the first snapshot taken at $t_1 = 1$ are in black color, and additional infected nodes observed in the second snapshot taken at $t_2 = 2$ are in gray color. In Fig. 4(a), we suppose v is the source and in Fig. 4(b), we suppose v' is the source. Notice that in Fig. 4(b), with v' being the source, the branches $T_{v'}^1$ and $T_{v'}^3$ do not have any infected nodes (excluding v') at time t_2. Thus we mark these branches with dotted lines (for edges) and dotted circles (for nodes). We now demonstrate how to apply our estimator to discover the most likely source node. Consider Fig. 4(a) and suppose v is the source. The rumor centrality of v and the sizes of the branches rooted at v are: $R(v, G_1) = \frac{4!}{4\times1\times1\times1} = 6$, $k_1^1 = 1, k_1^2 = 0, k_1^3 = 1, k_1^4 = 1, k_2^1 = 3, k_2^2 = 3, k_2^3 = 3, k_2^4 = 2$. Applying Eq. (1), the spreading rate estimates for the branches are: $\hat{\boldsymbol{\lambda}} = (1, 2, 1, 1)$. (Note that there is only one sample path with length 2 from the boundary $B_v^2(t_1)$ to the boundary $B_v^2(t_2)$.) Using Proposition 1, the estimate for source start time is: $\hat{t}_0 \approx 0.59$. Now consider Fig. 4(b) and suppose v' is the source. We have $R(v', G_1) = \frac{4!}{4\times3\times1\times1} = 2$, $k'^1_1 = 0, k'^2_1 = 3, k'^3_1 = 0, k'^4_1 = 0, k'^1_2 = 0, k'^2_2 = 10, k'^3_2 = 0, k'^4_2 = 1$. Similarly, the spreading rate estimates for the branches are: $\hat{\boldsymbol{\lambda}}' = (0, 4/3, 0, 1)$. (Note that the lengths of sample paths are $1, 2$ and 1 respectively from the boundary $B_{v'}^2(t_1)$ to the boundary $B_{v'}^2(t_2)$.) The estimate for the source start time is: $\hat{t}'_0 \approx 0.64$. Applying Theorem 1, we calculate the ratio between the (unnormalized) likelihood of v' and the (unnormalized) likelihood of v of being the source

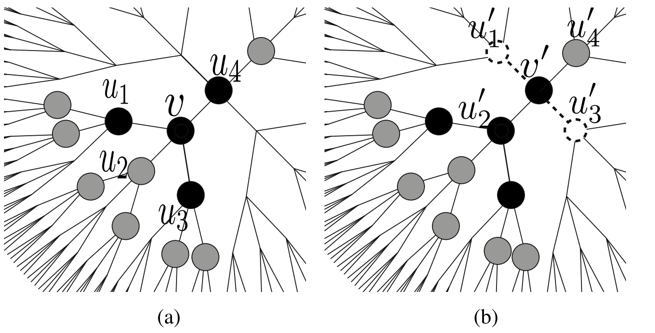


Fig. 4. Sequential snapshots at $t_1 = 1$ and $t_2 = 2$ on a 4-regular tree. (a) Suppose v is the source. (b) Suppose v' is the source.

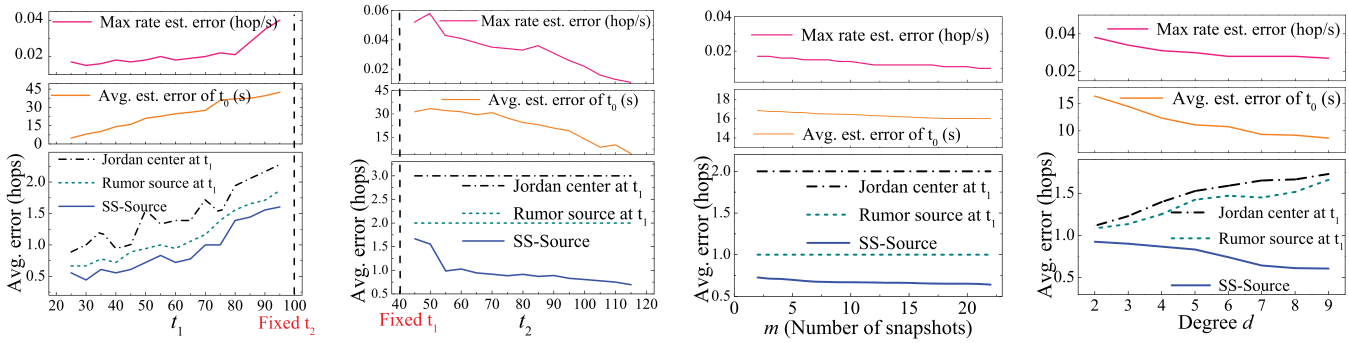
$$\frac{R(v', G_1) \cdot C(\hat{\boldsymbol{\lambda}}', \hat{t}'_0)}{R(v, G_1) \cdot C(\hat{\boldsymbol{\lambda}}, \hat{t}_0)} \approx \frac{2}{6} \cdot \frac{1.75}{0.172} \approx 3.4 \tag{17}$$

Thus the likelihood of v' being the source is over two times larger than the likelihood of v being the source. In fact, as shown in Fig. 4(b), v' is more likely to be the source since v' has a large branch $T_{v'}^2(t_1)$ at time t_1 and the branch $T_{v'}^2(t_1)$ is more likely to infect more nodes and form $T_{v'}^2(t_2)$.

Remark: In the above example, if we use the rumor centrality [1] or the Jordan center [6] as the source estimator, it would indicate that node v is the source if only one snapshot (either the first snapshot or the second one) is observed. However, as we have demonstrated that v' is more likely to be the source if we have two "*sequential snapshots*". This shows that our approach is more accurate and general than the previous work in [1] and [6] while considering the heterogeneity of the spreading rates. Furthermore, this also shows that the second snapshot provides us with more information about the information spreading, i.e., the spreading rates of the branches that connect to the source and the source start time. Thus this shows that taking sequential snapshots is important for detecting the information source.

IV. Experiments

In this section, we present experimental results of applying our framework to both synthetic networks and real-world retweet networks of different scales to estimate spreading rates, source start time and the location of information source.



(a) (b) (c) (d)

Fig. 5. Accuracy on d-regular tree networks (heterogeneous rates). ***Top***: estimation error of $\hat{\boldsymbol{\lambda}}$. ***Middle***: estimation error of $\hat{t}_0$. ***Bottom***: estimation error of SS-Source, Rumor source and Jordan center. (a) Accuracy vs. t_1. (b) Accuracy vs. t_2. (c) Accuracy vs. m. (d) Accuracy vs. d.

For synthetic networks, we generate the underlying graphs by various network generation models and select a node v^* as the true source. Starting from v^*, we apply the SI model described in Section II to infect v^*'s branches at different rates $\boldsymbol{\lambda}$. We keep track of the spreading process by taking a number of snapshots $G_1, \ldots, G_m$ (i.e., record the states of infected nodes) at different times $t_1, \ldots, t_m$. For different real-world networks, we extracted the retweet networks (networks of users who tweet and retweet the same piece of information) from a publicly available Twitter dataset [12]–[14] (small scale) and the Higgs Twitter dataset [15], [16] (large scale). As each tweet is timestamped in the datasets, we can identify the true source v^* with the minimum timestamp, calculate the true spreading rates and generate snapshots $G_1, \ldots, G_m$ at different times for each retweet network. Note that the synthetic networks we generated as well as the extracted retweet networks are *not constrained as regular trees*. For non-regular tree networks, we use the BFS-tree approximation method (see Sec. III-C) while applying our framework.

In the following, we investigate the impact of various parameter variations on our framework's estimation errors for the spreading rates, the source start time and the information source, which we refer to as *SS-Source* (Sequential-Snapshots Source). Specifically, the estimation error of the information source is measured by the shortest hop distance between the estimated source $\hat{v}$ and the true source v^*, i.e., $\text{dist}(v^*, \hat{v})$ (hops). We also compare the estimation errors of SS-Source with those of *Rumor source* (the source estimated using the rumor centrality [1]) and *Jordan center* [6], respectively. The estimation error of the spreading rates is measured by the maximum absolute difference between the true spreading rates $\boldsymbol{\lambda}$ and the estimated spreading rates $\hat{\boldsymbol{\lambda}}$, i.e., $\max_{1 \le i \le d} |\lambda^i - \hat{\lambda}^i|$ (hop per unit time). The estimation error of the source start time is measured by the absolute difference between the true source start time t_0 and the estimated source start time $\hat{t}_0$, i.e., $|t_0 - \hat{t}_0|$ (second or day).

A. Experiments on d-Regular Trees (Heterogeneous Rates)

We generate large d-regular trees as the underlying networks and simulate information spreading at different rates on different branches on these networks. As the information spreading is a random process, the sizes of the snapshots are random variables. Thus we will specify the information spreading rates in the simulations and study the impact of the controllable parameters, namely the snapshot times $t_1, \ldots, t_m$ and degree d on the accuracy of the source estimate $\hat{v}$, rates estimates $\hat{\boldsymbol{\lambda}}$ and start time estimate $\hat{t}_0$.

- **Impact of snapshot time t_1:** t_1 is an important parameter that determines the first snapshot G_1 which influences $R(v, G_1)$ in our estimation framework in Eq. (14). To find the impact of t_1 on the accuracy of our framework, we use two sequential snapshots $(m = 2)$, vary t_1 (from 25 seconds to 95 seconds in step size of 5), keep t_2 fixed at $t_2 = 100$ seconds (the vertical dash line in Fig. 5(a)) and start the information spreading process on an 4-regular tree at the specified spreading rates $\boldsymbol{\lambda} = (0.01, 0.02, 0.03, 0.04)$ (hop/second) at time $t_0 = 0$. Since the rumor centrality and Jordan center only consider a single snapshot in source estimation, we feed these two estimators with the snapshot taken at t_1 as the input. For each specific t_1, the estimation errors are averaged over 1000 times of information spreading. As shown in Fig. 5(a) (*bottom*), for SS-Source, Rumor source and Jordan center, the average estimation errors are increasing as t_1 increases. It is because that the increasing size of the first snapshot makes the probability of locating the source increasingly small. However, our SS-Source achieves the lowest source estimation error. This indicates that the growth sizes of the branches in the underlying graph become significantly different as t_1 increases and SS-Source can recognize such different growth sizes of the branches and give a better estimate of the source. On the contrary, neither Rumor source nor Jordan center captures the different growth sizes as both of the estimators do estimation based on a single static snapshot and neglect the dynamics in the spreading.

For the spreading rate estimation error and the source start time estimation error, as shown in Fig. 5(a) (*top* and *middle*), they both increase as t_1 continues to increase. This is because that the number of sample paths decreases when the time difference $t_2 - t_1$ becomes small and causes *the different growth sizes of branches* diminish. Nevertheless, our framework still makes good spreading rates estimate with the error within 0.04 hop/second and the source start time estimate with the error less than 35 seconds, while Rumor source and Jordan center cannot even provide such estimates.

• **Impact of snapshot time** t_2**:** In real life, the information spreading is usually unnoticed until the information spreads widely and catches enough attention. Thus the time t_1 when the first snapshot is taken is difficult to control. Hence we study the impact of t_2, i.e., the time of taking the second snapshot, on the accuracy of our framework. We simulate information spreading starting at $t_0 = 0$ on a 4-regular tree at rate $\boldsymbol{\lambda} = (0.01, 0.02, 0.03, 0.04)$ on the 4 branches. We keep t_1 fixed at $t_1 = 40$ and vary t_2 from 45 to 115 with step size 5. We feed the Rumor source and Jordan center with the snapshot taken at t_1 as the input (i.e., the first snapshot is fixed), and compare the estimation accuracy using the results averaged over 1000 times of information spreading starting from the fixed first snapshot. As shown in Fig. 5(b) (*bottom*), the estimation errors of both Rumor source and Jordan center are constant as the first snapshot is fixed. However, the estimation error of SS-Source *decreases* as t_2 increases. The reason is that the different growth sizes of the branches in the snapshot at t_2 can tell us the relative sizes the branches in the snapshot at t_1. Moreover, the larger t_2, the smaller estimation error of SS-Source as the larger difference $t_2 - t_1$ allows more sample paths from the first snapshot to the second snapshot and thus our framework provides a more accurate spreading rates estimate (Fig. 5(b) (*top*)) as well as a more accurate source start time estimate (Fig. 5(b) (*middle*)).

• **Impact of the number of snapshots** m**:** In general, we can have more than 2 snapshots. To study the impact of the number of snapshots, m, we keep t_1 fixed at $t_1 = 40$ (i.e., the first snapshot is fixed), take a sequential snapshot every 30 seconds and record the time t_m ($m \geq 1$ and m increases by 1 if a snapshot is taken). We feed SS-Source with the m snapshots that were taken before t_m (including t_m). For Rumor source and Jordan center, we use the snapshot at t_1 as the input and compare the estimation error averaged over 1000 times of spreading on a 4-regular tree at rates $\boldsymbol{\lambda} = (0.01, 0.02, 0.03, 0.04)$ on the 4 branches. From Fig. 5(c), we see the extra snapshots reduce the estimation errors of SS-Source, the source start time and the spreading rates. In conclusion, we see that using two snapshots that are separated by a large time interval ($30 \sim 40$) can provide estimates of low errors in our framework, and it is *the time interval between two snapshots*, rather than the number of snapshots, that captures the growth of branches since we get a minor improvement on the estimations by using more than 2 snapshots.

In summary, the results in Fig. 5(a), Fig. 5(b) and Fig. 5(c) indicate that our framework that uses an additional snapshot helps to decrease the estimation errors of the source, the source start time and the spreading rates significantly. In contrast, both Rumor source and Jordan center *fail to identify the growing branches in information spreading*, cannot provide an estimate for the spreading rate and source start time and give worse source estimate than that of SS-Source in our framework.

• **Impact of the degree** d**:** To study the impact of d on the accuracy of our framework, we run the information spreading process on the underlying d-regular tree networks with different d ranging from 2 to 9 at spreading rates $\boldsymbol{\lambda}$ on the d branches being $(0.01, \ldots, 0.01 * d)$. On each underlying d-regular tree, we take the first snapshot at $t_1 = 40$ and the second snapshot at $t_2 = 95$ respectively and average the estimation error over 1000 times of information spreading. For Rumor source and Jordan center, we feed them with the snapshot at t_1. Then we use Rumor source and Jordan center to estimate the source on the two snapshots separately.

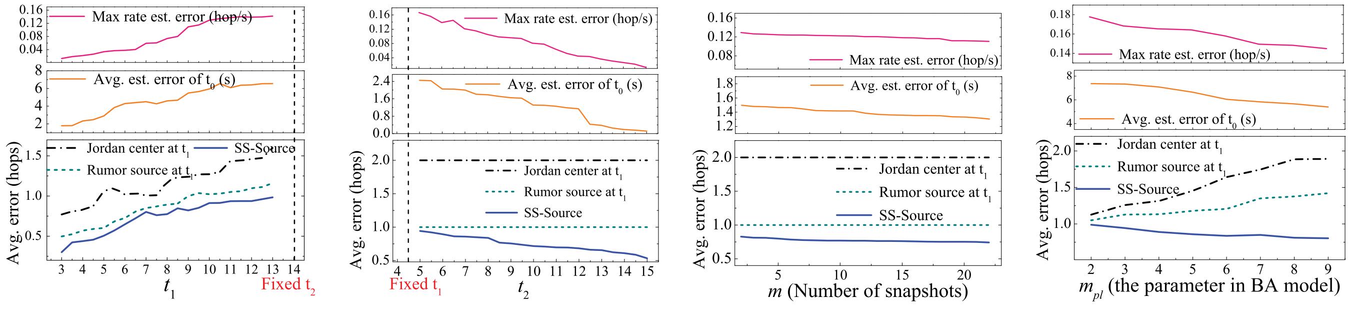
As shown in Fig. 5(d) (*bottom*), the average errors for Rumor source and Jordan center both increase as the degree d increases. It is because that the branches are more likely to get imbalanced when the number of branches is larger when the spreading rates on the branches are different. However, the estimation error of SS-Source decreases as d increases and achieves the lowest error among the three. This can be explained that the different growth sizes of branches appear with a higher probability when there are more branches with heterogeneous spreading rates. Therefore, there are higher chances that SS-Source finds the source using sequential snapshots. This is also the reason that the estimation errors of the spreading rates in Fig. 5(d) (*top*) and the source start time in Fig. 5(d) (*middle*) are getting smaller when d increases.

B. Experiments on Power-Law Graphs (Heterogeneous Rates)

We use the Barabási-Albert (BA) preferential attachment model [17] to generate the graphs with the power-law degree distribution, which is a typical feature of networks in the real world. There are two parameters n_{pl} and m_{pl} in the BA model, where n_{pl} is the number of nodes and m_{pl} is the number of edges to attach a new node to existing nodes. Here n_{pl} is set to be a large enough thus we can run the information spreading process and take sequential snapshots of the information spreading. We study the impact of the parameters $t_1, \ldots, t_m$ and m_{pl} on the accuracy of our framework, Rumor source and Jordan center. Note that the snapshots on power-law networks are not necessarily trees since we take all the edges among the nodes into the snapshots. As such, we use the BFS-tree approximation for our framework and the Rumor source.

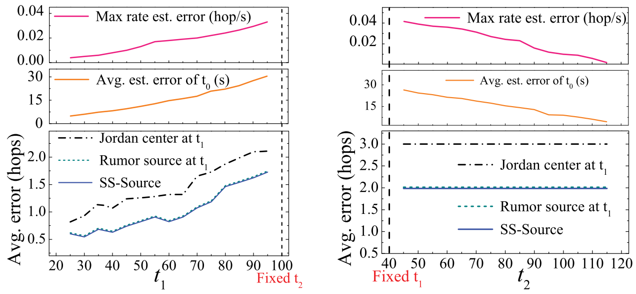
• **Impact of the snapshot times** $t_1, \ldots, t_m$ **and** m_{pl}**:** When studying the impact of $t_1, \ldots, t_m$, we keep the underlying network unchanged with a typical parameter setting for BA model: $n_{pl} = 80,000$, $m_{pl} = 2$. We evaluate the accuracy of our framework, Rumor source and Jordan center using the same methods described in Section IV-A. We *randomly* select a node of degree d ($d \geq 2$) as the source and simulate information spreading in the generated power-law networks with the spreading rates $\boldsymbol{\lambda} = (0.01, \ldots, 0.01 * d)$ on different branches. In the power-law networks, the degree of nodes can be very large. Thus the information can spread very fast in such networks. So we take snapshots in a short time period by fixing t_2 at 14 seconds when we study the impact of t_1 and fixing t_1 at 4.5 when we study the impact of t_2. Moreover, to study the impact of the number of snapshots m, we fix t_1 at $t_1 = 3$ and take sequential snapshots every 3 seconds. To study the impact of m_{pl} on the accuracy of the estimators, we keep t_1 and t_2 fixed ($t_1 = 3$ and $t_2 = 8$) and simulate information spreading on the power law graphs with m_{pl} varying from 2 to 9.

As shown in Fig. 6(a), Fig. 6(b), Fig. 6(c) and Fig. 6(d), we observe similar results for the power-law networks.



(a) (b) (c) (d)

Fig. 6. Accuracy on power-law networks (heterogeneous rates). ***Top***: estimation error of $\hat{\boldsymbol{\lambda}}$. ***Middle***: estimation error of $\hat{t}_0$. ***Bottom***: estimation error of SS-Source, Rumor source and Jordan center. (a) Accuracy vs. t_1. (b) Accuracy vs. t_2. (c) Accuracy vs. m. (d) Accuracy vs. m_{pl}.



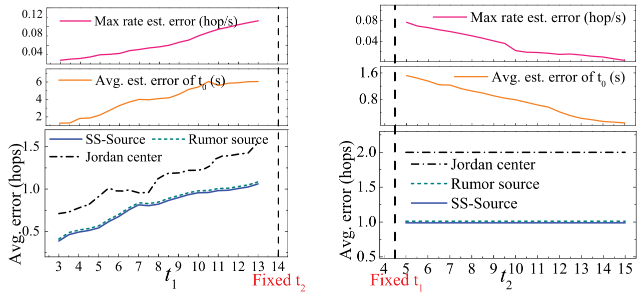
(a) (b)

Fig. 7. Accuracy on d-regular networks (homogeneous rate). ***Top***: estimation error of $\hat{\boldsymbol{\lambda}}$. ***Middle***: estimation error of $\hat{t}_0$. ***Bottom***: estimation error of SS-Source, Rumor source and Jordan center. (a) Accuracy vs. t_1. (b) Accuracy vs. t_2.

SS-Source achieves the lowest estimation error on the power-law graphs either we fixed t_1 or t_2. Also, for the accuracy of SS-Source, the extra number of snapshots can help (though very little) decrease the estimation error and the time interval $t_2 - t_1$ is more critical than the number of snapshots (two snapshots with interval ($3 \sim 5$ seconds) are enough). Therefore, we conclude that the different growth sizes of the branches also exist in power-law networks and our framework is able to give the best estimate of the source and provide good estimates for the spreading rates and using sequential snapshots in power-law networks.

C. Experiments on d-Regular Trees and Power-Law Graphs (Homogenous Rate)

To verify the theoretical result that our framework provides *the same source estimate with the rumor centrality estimator* when the spreading rates on different branches are the same, we apply our framework to both d-regular trees and power-law graphs by setting a constant spreading rate on different branches. For d-regular trees, we simulate information spreading starting at $t_0 = 0$ on a 4-regular tree at the same rate $\boldsymbol{\lambda} = (0.02, 0.02, 0.02, 0.02)$ on the 4 branches. For power-law networks, we randomly select a node of degree d ($d \geq 2$) as the source and simulate information spreading in the generated power-law networks at the spreading rates $\boldsymbol{\lambda} = (0.01, \ldots, 0.01)$ on different branches. In Fig. 7(a) (Fig. 8(a)), we keep t_2 fixed at $t_2 = 100$ ($t_2 = 14$) and study the impact of snapshot time t_1. In Fig. 7(b) (Fig. 8(b)), we keep t_1 fixed at $t_1 = 40$ ($t_1 = 4.5$) and study the impact of snapshot time t_2. As shown in the bottom of Fig. 7(a) and Fig. 8(a), our source estimator provides the same source estimate with the rumor source at different t_1. Fig. 7(b) (bottom) and and Fig. 8(b) (bottom) show that only the snapshot at t_1 is useful in estimating the source for both our framework and Rumor source. In addition, our framework utilizes the sequential snapshots and provides the spreading rate estimates (middle of Fig. 7(a), Fig. 7(b), Fig. 8(a), and Fig. 8(b)) and the source start time estimates (top of Fig. 7(a), Fig. 7(b), Fig. 8(a), and Fig. 8(b)).

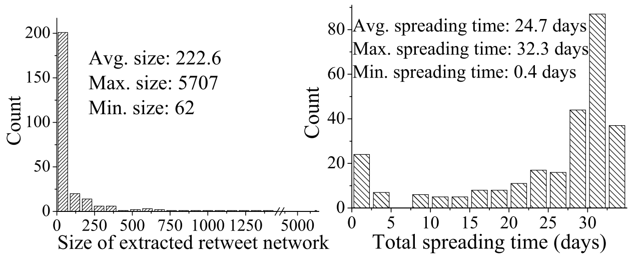


(a) (b)

Fig. 8. Accuracy on power-law networks (homogeneous rate). ***Top***: estimation error of $\hat{\boldsymbol{\lambda}}$. ***Middle***: estimation error of $\hat{t}_0$. ***Bottom***: estimation error of SS-Source, Rumor source and Jordan center. (a) Accuracy vs. t_1. (b) Accuracy vs. t_2.

D. Experiments on Real-World Networks

We apply our analytical framework to real-world retweet networks which are extracted from a publicly available dataset [12]–[14]. The dataset includes sequences of the tweeters, retweeters, and timestamps for observed hashtags in the public tweets from Twitter collected from March 24, 2012 to April 25, 2012. For tweets that contain a hashtag (e.g., *#wheniwaslittle*, *#niallfact*, *#thoughtsduringschool*) which was created by a tweeter (the source), the data maintain a *retweet timeline*, which keeps track of the information spreading from the tweeter to retweeters and then from retweeters to retweeters by recording who (anonymized IDs) at what time (timestamps in Unix time) retweets to whom (anonymized IDs) with the that hashtag. The data contain the entire information of the spreading process of all hashtags on Twitter, and it facilitates a simpler retweet network extraction method than epidemic history reconstruction in [18] and diffusion progression construction in [19].



(a)

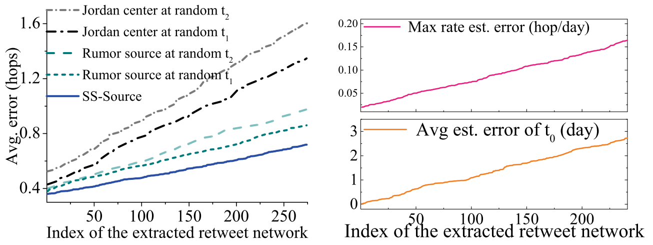
(b)

Fig. 9. Statistics of the extracted retweet networks. (a) Histogram of extracted retweet networks. (b) Histogram of total spreading time.

Retweet Networks Extraction: We treat each hashtag as a piece of information/rumor and extract the retweet network of each hashtag from the retweet timeline. Specifically, for each retweet timeline, we examine the timestamps in ascending order and connect the tweeter and retweeters to form a connected network. Such a network is formed by the tweeter who initially created the hashtag, and the retweeters who are connected (directly or indirectly) to the tweeter and retweeted with that hashtag later. Furthermore, we only consider the user and the timestamp that he/she tweets with the hashtag for the first time. Therefore, each extracted retweet network is actually the *information spreading tree* of a hashtag in Twitter, where the tweeter who tweets the earliest is the true source. Besides, in the extraction procedure, we omit the extracted information spreading trees that are of small size with the number of nodes less than 20 or with the diameter less than 4 as these networks are so small that information forensics on them are not worth looking into.

We extracted 274 retweet networks. For each extracted retweet network, we calculate the total spreading time by subtracting the minimum timestamp from the maximum timestamp in that extracted network. We give the histograms of the sizes of the extracted networks and the total spreading times in Fig. 9. From Fig. 9(a), we see that most of the extracted retweet networks have very small size (less than 500 users). Moreover, as shown in Fig. 9(b), the total spreading time on most of the extracted networks tends to be long (more than 25 days). This indicates that the source is usually among a small set of nodes and we have enough time to observe and take sequential snapshots of the information spreading process.

Note that on each extracted retweet network, every node (user) has a timestamp to indicate when he/she tweeted the information. We first shift all timestamps by subtracting the minimum timestamp (the timestamp of the source) from them thus making the source start time t_0 zero. Then we divide all the timestamps by $86,400$ thus transforming the timestamps from Unix time format to float numbers that represent the time (in days) when the users tweet. Then we count the largest hop-distance in each of the branches rooted at the source and divide the distance by the maximum timestamp. In this way, we obtain the ground truth of the spreading rates (hop/day) on the branches rooted at the source. Moreover, for each extracted retweet network, we select several timestamps as $t_1, \ldots, t_j, \ldots, t_m$ between the minimum and the maximum timestamp. These timestamps are equally spread and spanning over the entire total spreading time of a retweet network. We obtain the snapshot at a specific t_j by generating the subgraph containing nodes with smaller timestamps than t_j from the retweet network. We feed the snapshots to our framework (with BFS-tree approximation), Rumor centrality, and Jordan center.



(a)

(b)

Fig. 10. (a): source estimation errors on the extracted retweet networks (sorted by size). (b): source start time and information spreading rates on the extracted retweet networks (sorted by size).

For each extracted retweet network, results are similar to the results shown in Fig. 5 and Fig. 6. Namely, despite the impact of t_1 or t_2, the SS-Source in our framework has the lowest estimation error for the source. Also, additional snapshot helps to reduce the source estimation error. Moreover, the time interval between snapshots is more critical than the number of snapshots. Thus using only two snapshots that are separated by a large time interval ($3 \sim 4$ days in retweet networks) can provide good estimates. In addition, we found that the spreading rates on the branches rooted at the source in real-world networks are indeed *heterogeneous*. The typical spreading rates on different branches of the source vary from 0.01 hop/day to 2.53 hop/day in a single retweet network. For the estimation of the source start time, the estimation errors never exceed 3.0 days, which are considered very small compared with the 25-day average total spreading time.

As these results are very similar to the results shown in Fig. 5 and Fig. 6, we do not include the results for brevity. Instead, we select out 241 retweet networks each of which with spreading time larger than 3 days from the 274 retweet networks. Then we randomly select t_1 and t_2 with time interval 2.5 days for 1000 times on these 241 retweet networks, using the snapshots at t_1 and t_2 (Rumor source and Jordan center do estimation on the snapshots at t_1 and t_2 separately) and then calculate the average estimation errors of the source, the spreading rates and the source start time. The results are shown in Fig. 10.

From Fig. 10(a), we see that SS-Source has the lowest estimation error despite that the estimation error increases as the size of the network increases. Although the difference between SS-Source and Rumor source is small (about 0.1), this result indicates that on each extracted retweet network, we can take two snapshots at two randomly selected times with fixed interval, and provide a source estimate with the lowest estimation error. Moreover, using merely two sequential snapshots, we can give accurate estimates for the spreading rates and the source start time.

- **Justification for high accuracy:** We now use two snapshots (Fig. 11(a) and Fig. 11(b)) on an extracted real retweet network (with the same hashtag *#mustfollow*) to provide insight into the high estimation accuracy of our framework. We see in Fig. 11, the true source (black dot) has a very large

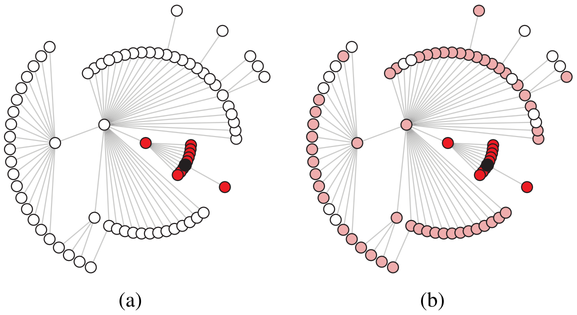


Fig. 11. Two snapshots of a retweet network. The black dot represents the true source. The white dots represent nodes that are not infected yet at the time the snapshot is taken. (a): the red dots represent the users tweeted before t_1. (b): The pink dots represent the users that tweeted during $(t_1, t_2]$.

branch in both snapshots. The message he/she tweeted spreads out at a very large spreading rate to other users via his/her friend (the red dot in the middle). However, the message does not spread to other branches during $(t_1, t_2]$. Our framework adds more weight to the real source by considering the growth of branches and estimating the spreading rates using sequential snapshots, while rumor source and Jordan center consider the snapshots statically. Hence, in real-world networks, our framework (SS-Source) is able to identify the information source with a higher accuracy and provides accurate spreading rates estimates on the branches.

E. Experiments on a Large-Scale Retweet Network

We further apply our framework to a large-scale retweet network extracted from the Higgs Twitter dataset [15], [16]. The Higgs Twitter dataset contains a single retweet network that describes the spreading process of the discovery of a new particle with the features of the elusive Higgs boson from 1st July, 2012 to 7th July, 2012. The dataset provides the retweet activities of who retweets whom at when in Twitter. Using the same techniques described in Sec. IV-B, we reconstruct the retweet network that has $15,915$ connected users with the total spreading time as long as 7 days.

Similar to Sec. IV-D, we take snapshots of the extracted retweet network, and our estimation of the spreading rates validate that the rates on the branches rooted at the source are very different and vary from 0.01 hop/day to 3.7 hop/day. Compared with the estimation errors of Rumor source and Jordan center, our SS-Source achieves the lowest estimation error (within 0.4 hops on average). For the source start time estimation, our framework also provides a source start time estimate with the error less than 1.1 days.

V. Related Work

Due to its theoretical importance and practical values, information spreading forensics has gained a lot of interest and attention in recent years. Most of the work focused on detecting the location of the information source. Lappas *et al.* [20] proposed k-EFFECTORS to find the most influential individuals rather than the information source. They assume prior knowledge of the interacting probabilities among users in networks under the independent cascade model. In the seminal work, T. Zaman *et al.* introduced an estimator known as the *rumor centrality* to detect the information source under the SI model using a single snapshot of the infected nodes [1], [10], [21]. The source detected by the rumor centrality aims to balance the size of branches [1] when the information spreads in a homogeneous rate over a network. However, it is very likely that the branches are not balanced in a snapshot (as we illustrated in Fig. 1(b)) if the spreading rates are heterogeneous in different branches.

Following [1], Dong *et al.* [22] studied the problem of rooting out the rumor source given prior knowledge on a set of suspected source nodes. Based on different centrality measures, Comin and da Fontoura Costa [23] showed that the source node tends to have the highest centrality measurement values. In [24], with observers in the network placed beforehand, Pinto *et al.* estimated the source location by assuming that the direction and the times of infections are known. Zhu and Ying [6] proposed an elegant sample-path-based method to detect the source in SIR model with homogeneous infection probability. They proved that the source node minimizes the maximum distance to other infected nodes and then used the Jordan center [25] as the source estimator. Then Chen *et al.* [7] incorporated the sample-path-based method with novel clustering and localization techniques to detect multiple sources in networks. Lokhov *et al.* [26] developed a computationally expensive algorithm, dynamic message passing (DMP), to estimate the probability that a given node produces the observed snapshot and output the node with the highest probability. Prakash *et al.* [27] and [28], provided an algorithm for multiple sources detection based on a coding-theoretic method but the computation complexity increases exponentially with the number of nodes. Wang *et al.* [2] proposed the union rumor centrality measure and considered source detection with multiple but independent snapshot observations. Farajtabar *et al.* [29] proposed an estimation method using the information cascade which requires fine-grained observations of the infection times for nodes and edges. In contrast, Fanti *et al.* [30] and [31] proposed messaging protocols by controlling the information spreading rates among different users so to have perfect obfuscation of the rumor source.

In our work, we perform information spreading forensics using sequential dependent observations while considering heterogeneous spreading rates in a network. We provide conditional maximum likelihood estimates of the information source as well as the spreading rates and the source start time. Thus our approach of source detection differs fundamentally from those using only a single observation [1], [6], [10], [21], [22] or multiple yet independent observations [2].

VI. Conclusion and Future Work

To the best of our knowledge, this paper is the first theoretical and experimental study on the information spreading forensics using *sequential and dependent* observations. We derived conditional maximum likelihood estimators for information spreading rates, source start time and information source in d-regular tree networks from the perspective of "*growing branches*" in sequential observations. For general networks, we designed a parallel algorithm for the estimators in our framework. By applying our framework to d-regular tree networks, power-law networks, and real-world retweet networks in Twitter, we demonstrated that our framework can take advantage of the growing branches in information spreading, and provides highly accurate estimates for the these three metrics.

Our framework also generalizes rumor centrality [10] and the union rumor centrality [2] by allowing information spreads at heterogeneous rates, and it opens the door for future research on information spreading forensics, e.g., consider limitations on the number of the sequential snapshots as taking snapshots comes at a certain cost, and consider the cases where full snapshots of the network are not available in a noisy environment. We are also interested in incorporating our framework with clustering or community detection techniques to extend our framework to give estimates in information spreading forensics with multiple sources.

REFERENCES

[1] D. Shah and T. Zaman, "Detecting sources of computer viruses in networks: Theory and experiment," in *Proc. ACM SIGMETRICS*, 2010, pp. 203–214.

[2] Z. Wang, W. Dong, W. Zhang, and C. W. Tan, "Rumor source detection with multiple observations: Fundamental limits and algorithms," in *Proc. ACM SIGMETRICS*, 2014, pp. 1–13.

[3] M. De Choudhury *et al.*, "How does the data sampling strategy impact the discovery of information diffusion in social media?" in *Proc. ICWSM*, 2010, pp. 34–41.

[4] M. Laskowski, L. C. Mostaço-Guidolin, A. L. Greer, J. Wu, and S. M. Moghadas, "The impact of demographic variables on disease spread: Influenza in remote communities," *Sci. Rep.*, vol. 1, Oct. 2011, Art. no. 105.

[5] B. Doerr, M. Fouz, and T. Friedrich, "Why rumors spread so quickly in social networks," *Commun. ACM*, vol. 55, no. 6, pp. 70–75, 2012.

[6] K. Zhu and L. Ying, "Information source detection in the SIR model: A sample-path-based approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 1, pp. 408–421, Feb. 2016.

[7] Z. Chen, K. Zhu, and L. Ying, "Detecting multiple information sources in networks under the SIR model," *IEEE Trans. Netw. Sci. Eng.*, vol. 3, no. 1, pp. 17–31, Jan. 2016.

[8] N. T. Bailey *et al.*, *The Mathematical Theory of Infectious Diseases and its Applications*. London, U.K.: Griffin, 1975.

[9] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. ACM WSDM*, 2010, pp. 241–250.

[10] D. Shah and T. Zaman, "Rumor centrality: A universal source detector," in *Proc. ACM SIGMETRICS*, 2012, pp. 199–210.

[11] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida, "On word-of-mouth based discovery of the Web," in *Proc. ACM IMC*, 2011, pp. 381–396.

[12] L. Weng. *Prediction of Viral MEMES on Twitter*. [Online]. Available: http://carl.cs.indiana.edu/data/

[13] L. Weng, F. Menczer, and Y.-Y. Ahn, "Virality prediction and community structure in social networks," *Sci. Rep.*, vol. 3, Aug. 2013, Art. no. 2522.

[14] L. Weng, F. Menczer, and Y.-Y. Ahn, "Predicting successful memes using network and community structure," in *Proc. 8th AAAI ICWSM*, 2014, pp. 535–544.

[15] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi, "The anatomy of a scientific rumor," *Sci. Rep.*, vol. 3, Oct. 2013, Art. no. 2980.

[16] M. De Domenico, A. Lima, P. Mougel, and M. Musolesi. (2015). *Higgs Twitter Dataset*. [Online]. Available: https://snap.stanford.edu/data/higgs-twitter.html

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

[18] P. Rozenshtein, A. Gionis, B. A. Prakash, and J. Vreeken, "Reconstructing an epidemic over time," in *Proc. ACM SIGKDD*, 2016, pp. 1835–1844.

[19] E. Sefer and C. Kingsford, "Diffusion archaeology for diffusion progression history reconstruction," in *Proc. IEEE ICDM*, Dec. 2014, pp. 530–539.

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

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

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

[23] C. H. Comin and L. da Fontoura Costa, "Identifying the starting point of a spreading process in complex networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 84, no. 5, p. 056105, 2011.

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

[25] C. Jordan, "Straight lines, symmetry, self similarity," *J. die Reine Angewandte Math.*, vol. 70, pp. 185–190, 1869.

[26] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "Inferring the origin of an epidemic with a dynamic message-passing algorithm," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 90, no. 1, p. 012801, 2014.

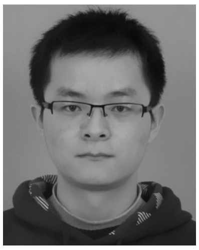
[27] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Spotting culprits in epidemics: How many and which ones," in *Proc. IEEE ICDM*, Dec. 2012, pp. 11–20.

[28] B. A. Prakash, J. Vreeken, and C. Faloutsos, "Efficiently spotting the starting points of an epidemic in a large graph," *Knowl. Inf. Syst.*, vol. 38, no. 1, pp. 35–59, 2014.

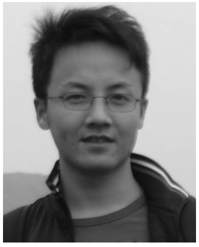
[29] M. Farajtabar, M. G. Rodriguez, M. Zamani, N. Du, H. Zha, and L. Song, "Back to the past: Source identification in diffusion networks from partially observed cascades," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, vol. 38. May 2015, pp. 232–240.

[30] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath, "Spy vs. spy: Rumor source obfuscation," in *Proc. ACM SIGMETRICS*, 2015, pp. 271–284.

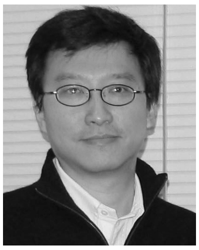
[31] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran, and P. Viswanath, "Rumor source obfuscation on irregular trees," in *Proc. ACM SIGMETRICS*, 2016, pp. 153–164.



Kechao Cai received the B.S. and M.S. degrees from the Huazhong University of Science and Technology, China, in 2010 and 2013, respectively. He is currently pursuing the Ph.D. degree with The Chinese University of Hong Kong under the supervision of Prof. J. C. S. Lui. His current research area includes distributed protocol design, data analytics of online social networks, and online learning algorithms.



Hong Xie (M'14) received the B.Eng. degree from the School of Computer Science and Technology, University of Science and Technology of China, in 2010, and the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2015, under the supervision of Prof. J. C. S. Lui. He is currently a Post-Doctoral Research Fellow with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His research interests include applied machine learning, data science, network economics, and stochastic modeling. He is a member of the ACM.



John C. S. Lui (F'10) received the Ph.D. degree in computer science from the University of California at Los Angeles. He was a Chairman with the CSE Department from 2005 to 2011. He is currently the Choh-Ming Li Chair Professor with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include communication networks, system security (e.g., cloud security, mobile security, etc.), network economics, network sciences, large-scale distributed systems, and performance evaluation theory. He is an Elected Member of the IFIP WG 7.3, a fellow of the ACM, and a Croucher Senior Research Fellow. He was a recipient of the various departmental teaching awards and the CUHK Vice-Chancellors Exemplary Teaching Award. He was also a co-recipient of the IFIP WG 7.3 Performance 2005 and IEEEIFIP NOMS 2006 Best Student Paper Awards. He serves in the Editorial Board of the IEEE/ACM TRANSACTIONS ON NETWORKING, the IEEE TRANSACTIONS ON COMPUTERS, the IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, the *Journal of Performance Evaluation*, and the *International Journal of Network Security*.