Generalization of Information Spreading Forensics via Sequential Dependent Snapshots

Kechao Cai[†] Hong Xie[‡] John C.S. Lui[†]
[†]Department of Computer Science & Engineering
[†]The Chinese University of Hong Kong
[‡]School of Computing, National University of Singapore

ABSTRACT

Learning the characteristics of information spreading in networks is crucial in communication studies, social network sentiment analysis and epidemic investigation. Previous work on information spreading has been focused on the information source detection using either a single observation, or multiple but "*independent*" observations of the underlying network while assuming information spreads at a "*uniform spreading rate*". In this paper, we conduct the first theoretical and experimental study on information spreading forensics, and propose a framework for estimating information spreading rates, information source start time and location of information source by utilizing "*multiple sequential and dependent snapshots*" where information can spread at different rates. We prove that our estimation framework generalizes the rumor centrality [1], and we allow heterogeneous information spreading rates on different branches in d-regular tree networks. We further show that our framework can provide highly accurate estimates for the information spreading rates on different branches, the source start time, and more accurate estimate for the location of information source than rumor centrality and Jordan center in both synthetic networks and real-world networks (i.e., Twitter).

1. INTRODUCTION

Understanding information spreading in networks is an important task in various aspects of human life, i.e., communication advisers would like to know how fast information spreads (information spreading rates) in different networks or communities to be more successful in network marketing. Online media reporters would like to find out when the news or rumors start to spread (source start time) for better public sentiment analysis. Epidemiologist would try to locate the virus source to find out the reason for an epidemic. However, for all of these instances, such a task is very challenging in a large network because the complete temporal knowledge of information spreading, i.e., time index of when each individual (node) receives the information, is usually not available [2], and this makes forensics for the spreading rates, source start time or the information source difficult if not impossible. Moreover, a typical scenario of information spreading is that the source would spread information to different networks where the spreading rates are usually very different. For example, an epidemic usually has different spreading rates among different age groups and regions [3]. And the news or rumors have different spread rates among different communities or different networks [4]. Such a heterogeneity of spreading rates makes it more difficult to extract the information spreading characteristics. Thus, one has to consider how to estimate the information spreading rates, the source start time, and the location of the information source from one or more observations (or snapshots) of information spreading.

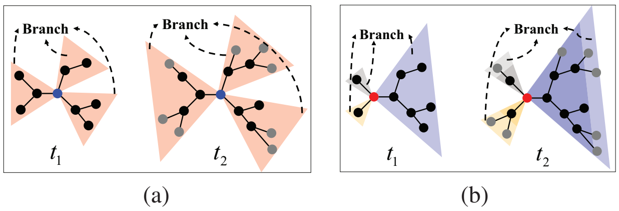

Fig. 1: Snapshot taken at t_1 (left with black dots) and snapshot taken at t_2 (right with gray and black dots) ($t_2 > t_1$) in (a) and (b) and different colored triangles represent different branches, respectively . (a): the blue dot represents the source estimate of rumor centrality. (b): the red dot represents the true information source.

In this paper, we present a framework to estimate the information spreading rates, the source start time and the information source with "*sequential and dependent snapshots*". We consider an unknown source starts spreading information with different spreading rates in a network. Specifically, the source first spreads information to different neighboring nodes with potentially different spreading rates, and then each of the neighbors spreads to other nodes along the edges with the spreading rate inherited from the source. We take sequential snapshots of the network at different times. The goal is to accurately estimate the spreading rates, source start time, and identify the information source based on these sequential snapshots.

Previous work mainly focused on the information source detection with a homogeneous information spreading rate. D. Shah and T. Zaman [1] first proposed the *rumor centrality* estimator using a single snapshot of information spreading. Later on, Wang *et al.* [5] presented a *union* rumor centrality that utilizes multiple *independent* snapshots of information spreading and yet proved that the sequential dependent snapshots will not improve the accuracy of source detection. Both are based on the assumption that information spreads at a uniform spreading rate which is not realistic for networks with different communities/groups [3]. Indeed, given the homogeneous information spreading rate, the rumor centrality and the union rumor centrality both give a source estimation that balances the sizes of the branches of the infected graphs [1, 5]. For example, Fig. 1(a) shows the infection 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 of rumor centrality. The underlying network (not drawn) is a 3-regular tree where every node in the tree is of degree 3. Fig. 1(b) illustrates the same two snapshots of the ground truth information spreading process at time t_1 and t_2 ($t_2 > t_1$) where the *red dot represents the true source* and different colored triangles represents different branches. Each snapshot has three branches drawn with different shades of colors rooted at the true source node (red dot). The rumor centrality or the union rumor centrality on these two snapshots would give wrong estimate for the source (the blue dot in the left and right of Fig. 1(a), respectively) that balances the sizes of branches (the pink triangles) as they fail to capture the different growth of branches in two sequential snapshots, while assuming that the information spreading has the same rate on these three branches. In contrast, we propose a general information spreading forensic model and propose a framework to estimate the different spreading rates, the source start time, and the information source using sequential snapshots.

Our estimation framework consists of four components as follows. 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 branches 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 for a node of 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 maximum likelihood as the source and give corresponding estimations for the spreading rates and source start time.

Our key idea is that a subsequent snapshot can reveal the information spreading rates for branches in earlier snapshots when one examines the "growth size" of the infected branches. The spreading rate estimates further help us infer the source start time and give likelihood estimate for the source. We illustrate this idea via Figure. 1(b). As shown in the two sequential snapshots in Fig. 1(b), the branch (blue triangles in Fig. 1(b)) with the largest growth size should have the largest spreading rate estimate. Such spreading rate estimates indicate that the branch in 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, and find a node (the red dot in Fig. 1(b)) in the first snapshot with the maximum likelihood that generates such branches, where the spreading rate estimates validate the growth sizes of the branches when the snapshots are taken. More importantly, we prove that our framework generalizes the rumor centrality [1] to the cases of heterogeneous spreading rates at different branches in d-regular tree networks, and show that our framework provides more accurate source estimates than rumor centrality and Jordan center [6], while giving highly accurate estimates for the information spreading rates and source start time. We have validated these claims in our experiments, both for synthetic and real-world networks (Twitter).

2. MODELS AND FRAMEWORK DESIGN

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 to describe the information spreading process. Specifically, at an *unknown* time t_0, an *unknown* information source node $v^* \in \mathcal{V}$ starts spreading information in $\mathcal{G}$. Each node in $\mathcal{G}$ can be either *susceptible* (not getting the info.) or *infected* (getting the info.). 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 with rate λ^{ij}). We focus on the case that $\theta_{ij}, \forall (i, j) \in \mathcal{E}$ 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 the subsequent infections, a node which was infected with rate λ^i would go on to infect its neighbors with 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 [7] as a highly accurate probabilistic model to capture the interaction behaviors between users in social networks. Our objective is to give maximum likelihood 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 $\mathcal{G}$.

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 contains all the infected nodes and the edges between every pair of 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 snapshot taking at time t_j (where $t_0 < t_1 < \ldots < t_j < \ldots < t_m$). Clearly, we have $G_j \subseteq G_m$ as these m snapshots are taken sequentially from the same spreading process on $\mathcal{G}$.

Framework Design. We now give the high level idea of our framework. We present a framework to estimate the spreading rates ($\boldsymbol{\lambda}$) on each branch, 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 snapshots $G_1, \ldots, G_m$ and the times $t_1, \ldots, t_m$ as the input.

• ***Branch Extraction.*** Upon taking the sequential snapshots, we further split them into d growing branches. 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 $T_v^i(t_j)$ for $1 \leq i \leq d$, 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 $1 \leq i \leq d$. The increment of the branch sizes of two consecutive times t_{j-1} and t_j of $T_v^i(t)$ is denoted by δ_j^i, i.e., $\delta_j^i = k_j^i - k_{j-1}^i$. As the spreading process has possibly different 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 at least one neighbor in $\mathcal{G} - T_v^i(t_j)$. 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 $b_j^i = |B_v^i(t_j)|$. From the boundary $B_v^i(t_{j-1})$ to the boundary $B_v^i(t_j)$, we can sample b_{j-1}^i ($b_{j-1}^i \geq 0$) paths that are disjoint with each other as the branches are trees. We denote $l_r (0 \leq r \leq b_{j-1}^i)$ as the length of the b_{j-1}^i disjoint path for $2 \leq j \leq m$ and $l_0 = 0$.

• ***Spreading Rates Estimation.*** We seek to derive the maximum likelihood estimators for the spreading rates on the different branches with the source being v. Consider the branch T_v^i. We have b_{j-1}^i *disjoint* 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 identical and independent Poisson process with rate λ^i. The maximum likelihood estimator for $\lambda^i (1 \leq i \leq d)$ can be expressed as:

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

where the right hand side is simply the joint probability 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 $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}^1, \ldots, \hat{\lambda}^d)$.

• ***Source Start Time Estimation.*** To estimate the source start time t_0, we consider the distribution of time $t = t_1 - t_0$ for the information spreading from the source v to the branch $T_v^i(t_1)$ given that the spreading rate in the branch is $\hat{\lambda}^i$, $1 \leq i \leq d$, during $(t_0, t_1]$. Note that there are k_1^i infections that occur during $(t_0, t_1]$ on the branch $T_v^i(t_1)$ and each new infection on branch $T_v^i(t_1)$ would introduce $d-2$ newly infectious edges. Moreover, the time follows an exponential distribution with rate $\hat{\lambda}^i$ on each edge. Hence, the total time $S_{k_1^i}(t)$ of infecting k_1^i nodes on the branch $T_v^i(t_1)$ is the sum of the k_1^i exponentially distributed random variables with rate $\hat{\lambda}^i, [1+(d-2)]\hat{\lambda}^i, \ldots, [1+(d-2)(k_1^i-1)]\hat{\lambda}^i$. Let $\hat{t}_0$ denote the estimator for t_0 and $\mathbf{P}_{k_1^i}(t_1|t_0) = S_{k_1^i}(t) - S_{k_1^i+1}(t)$ denote the probability that exactly k_1^i infections occur in the branch $T_v^i(t_1)$ during $(t_0, t_1]$ given that the spreading rate is $\hat{\lambda}^i$, $1 \leq i \leq d$. We give our derived source start time estimator $\hat{t}_0$ as follows,

$$\begin{aligned} \hat{t}_0 &= \arg\max_{t_0} \prod_{i=1}^{d} \mathbf{P}_{k_1^i}(t_1|t_0) \\ &= t_1 - \frac{1}{\sum \mathbb{1}_{\hat{\lambda}^i>0}} \sum_{i=1,\hat{\lambda}^i>0}^{d} \frac{\ln(1+(d-2)k_1^i)}{(d-2)\hat{\lambda}^i}, \end{aligned} \tag{2}$$

where $\mathbb{1}_{\hat{\lambda}^i>0}$ is an indicator function and only the branches with non-zero spreading rate $\hat{\lambda}^i > 0$ for $1 \leq i \leq d$ $(d > 2)$ in the summation term are considered. For $d = 2$, we have $\hat{t}_0 = t_1 - \frac{1}{\sum \mathbb{1}_{\hat{\lambda}^i>0}} k_1^i / \hat{\lambda}^i$. Eq. (2) suggests that the source start time estimator $\hat{t}_0$ is the difference between t_1 and the average spreading time from source v to the boundaries of different branches.

• ***Likelihood Estimation.*** We take advantage of the growing branches in the sequential snapshots and construct "*maximum likelihood estimators*" for the spreading rates, the source start time and the information source. In particular, assume that v is the source, we examine the infection process by considering the information spreading on each branch $T_v^i(t)$ with the spreading rate estimate $\hat{\lambda}^i$ during the time intervals $(\hat{t}_0, t_1], (t_1, t_2], \ldots$ and $(t_{m-1}, t_m]$ separately. Note that the infection process on each branch 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 maximum likelihood estimation for $\hat{v}$, $\hat{\boldsymbol{\lambda}}$, and $\hat{t}_0$ can be expressed as:

$$\begin{aligned} \{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} &= \arg\max_{v \in G_1} \prod_{i=1}^{d} \mathbf{P}[T_v^i(t_m), \ldots, T_v^i(t_1) | v, \hat{\boldsymbol{\lambda}}, \hat{t}_0] \\ &= \arg\max_{v \in G_1} \prod_{i=1}^{d} \prod_{j=1}^{m} \mathbf{P}[T_v^i(t_j) | T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0], \end{aligned} \tag{3}$$

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 $\hat{\boldsymbol{\lambda}}$, and $\mathbf{P}[T_v^i(t_j) | T_v^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ 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 $\mathbf{P}[T_{v'}^i(t_j) | T_{v'}^i(t_{j-1}), \hat{\boldsymbol{\lambda}}, \hat{t}_0]$ for $1 \leq j \leq m$. As such, the dependence of the branch $T_v^i(t_1)$ in G_1 on v carries over to the branches $T_v^i(t_j)$ in snapshots G_j for $2 \leq j \leq m$. Thus the branches $T_v^i(t_j)$ in snapshots G_j for $2 \leq j \leq m$ will impact on the estimates for the spreading rates $\hat{\boldsymbol{\lambda}}$, the source start time $\hat{t}_0$ and the source v in G_1. Therefore, we can give maximum likelihood estimates for the three information spreading characteristics that contribute to forming the branches in the first snapshot and branches in the following snapshots with the maximum probability. We then calculate the probability in Eq. (3), and give the explicit form of the maximum likelihood estimates of our framework as follows,

$$\{\hat{v}, \hat{\boldsymbol{\lambda}}, \hat{t}_0\} = \arg\max_{v \in G_1} R(v, G_1) \cdot C(\hat{\boldsymbol{\lambda}}, \hat{t}_0), \tag{4}$$

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

$$e^{-\sum\limits_{i=1}^{d} [(ak_m^i+1)t_m - \hat{t}_0]\hat{\lambda}^i} \prod_{i=1,\hat{\lambda}^i>0}^{d} \prod_{j=1}^{m} \left(e^{a\hat{\lambda}^i t_j} - e^{a\hat{\lambda}^i t_{j-1}}\right)^{\delta_j^i}, \tag{5}$$

for $1 \leq j \leq m$, $1 \leq i \leq d$, $a = d-2$, and only the branches with nonzero spreading rates $(\hat{\lambda}^i > 0)$ are considered in the product terms. $R(v, G_1)$ is the rumor centrality [1] of v in the first snapshot G_1. Note that our source estimator has an *additional scaling factor* $C(\hat{\boldsymbol{\lambda}}, \hat{t}_0)$ as compared with the previous work of rumor centrality [1]. This scaling factor characterizes the different spreading processes of the different branches. Moreover, the scaling factor is proved to be a constant regardless of where the source v is in G_1 when the spreading rates on the different branches are the same. Thus the rumor centrality is simply a special case and our framework generalizes it to the cases of heterogeneous spreading rates at different branches.

3. PRELIMINARY RESULTS

For synthetic networks, we run simulations on d-regular tree networks and power-law networks (where information spreads along the breadth-first-search tree). We take sequential snapshots of each spreading process at different times and apply our framework (the degree of each selected node is used as d $(d \geq 2)$ for power-law networks). Compared with the ground truth, the average estimation errors for the spreading rates are about 5%, and for the source birth times are about 9%. In addition, the average estimation errors of the source are within 1.1 hops of the true source, which is much lower (up to 50%) than the rumor centrality (1.8 hops) and Jordan center (2.0 hops). For real-world networks, we extract 274 information spreading graphs from the Twitter dataset [8] which includes the timestamps of the tweets as the ground truth of information spreading and we could achieve highly accurate estimates.

4. REFERENCES

[1] D. Shah and T. Zaman. Detecting sources of computer viruses in networks: theory and experiment. In *Proc. of ACM SIGMETRICS*, 2010.

[2] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? *ICWSM*, 10:34–41, 2010.

[3] 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. *Scientific Reports*, 1, 2011.

[4] B. Doerr, M. Fouz, and T. Friedrich. Why rumors spread so quickly in social networks. *Communications of the ACM*, 55(6):70–75, 2012.

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

[6] K. Zhu and L. Ying. Information source detection in the sir model: A sample-path-based approach. *Networking, IEEE/ACM Transactions on*, PP(99):1–1, 2015.

[7] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *Proc. of the ACM WSDM*, 2010.

[8] L. Weng. Prediction of viral memes on twitter. `http://carl.cs.indiana.edu/data/`.