Poster: Information Source Detection with Limited Time Knowledge

Xuecheng Liu[1], Luoyi Fu[1], Bo Jiang[1], Xiaojun Lin[2], and Xinbing Wang[1]
[1]Shanghai Jiao Tong University [2]Purdue University
{liuxuecheng, yiluofu, bjiang, xwang8}@sjtu.edu.cn[1] linx@ecn.purdue.edu[2]

ABSTRACT

We study the source detection problem using limited timestamps on a given network. Due to the NP-completeness of the maximum likelihood estimator (MLE), we propose an approximation solution called infection-path-based estimator (INF), the essence of which is to identify the most likely infection path that is consistent with observed timestamps. The source node associated with that infection path is viewed as the estimated source $\hat{v}$. For the tree network, we transform the INF into integer linear programming and find a reduced search region using BFS, within which the estimated source is provably always on a path termed as *candidate path*. This notion enables us to analyze the accuracy of the INF in terms of error distance on arbitrary tree. Specifically, on the infinite g-regular tree with uniform sampled timestamps, we get a refined performance guarantee in the sense of a constant bounded $d(v^*, \hat{v})$. By virtue of time labeled BFS tree, the estimator still performs fairly well when extended to more general graphs. Simulations on both trees and general networks further demonstrate the superior performance of the INF.

1 PROBLEM SETUP

Many phenomenon can be modeled as information propagation in networks over time. Prevalent examples include spread of a disease through a population, transmission of information through a distributed network, and the diffusion of scientific discovery in academic network. In all these scenarios, it is disastrous once an isolated risk is amplified through diffusion in networks. Source detection therefore is critical for preventing the spreading of malicious information, and reducing the potential damages incurred.

In this paper, we study the source inference problem: given that a message has been diffused in network G, can we tell which node is the source of diffusion given some observations O_t at time t? Though previous studies have proposed some solutions assuming O_t is a snapshot under various diffusion models [5, 6], this problem remains underexplored in more practical settings.

Consider an undirected graph $G = (V, E)$ where V is the set of nodes and E is the set of edges of the form (i, j) for some node i and j in V. We use the *susceptible-infected* SI model in epidemiology to characterize the infection diffusion process. Suppose that time is slotted. Initially only one node $v^* \in V$ gets infected at the beginning of some unknown time-slot $t_0 \in \mathbb{Z}$. At the beginning of each time-slot $t > t_0$, each infected node attempts independently to infect each of its susceptible neighbors with success probability $p \in (0, 1]$. We define the *first infection timestamp* of node u as the time-slot t_u in which the state of node u changes from susceptible to infected.

At time t, we want to locate the source node with limited knowledge $\{t_s\}_{s\in\mathcal{S}}$ of the first infection timestamps, where $\mathcal{S}$ is the subset of reported nodes. This setting has many practical advantages over those using snapshot and direction information [5, 6]. First, it is time consuming, and sometimes impossible, to collect the full snapshot of the infected nodes at some time. For example, Twitter's streaming API only allows a small percentage (1%) of the full stream of tweets to be crawled. Second, sometimes the direction from which a susceptible node gets infected is hard to obtain. For example, in a flu outbreak a person often cannot tell with certainty who infected him/her. The same also goes for anonymous social networks [2], where the direction information is hidden. Finally, the partial timestamps are easy to access in most scenarios (such as online social network, etc.).

Assume that each node $v \in V$ has the same prior probability of being the source, the optimal MLE is given by

$$\hat{v} = \hat{v}(\mathcal{S}) \in \arg\max_{v\in V} \mathbf{P}(\{t_s\}_{s\in\mathcal{S}}|\mathcal{S}, v^* = v), \tag{1}$$

which can be reduced to the #P-C problem of counting linear extensions for a given poset [1, 5]. Therefore we aim to design the approximation algorithm.

2 ALGORITHM AND PERFORMANCE

Based on the setup described in Section 1, the diffusion process can be characterized by the first infection time of each node and the parent node of each infected node. If we add an directed edge from the parent node to each infected node, we can construct an directed rooted tree spanning all the infected nodes $\mathcal{S}$ which partially explains the diffusion process since the specific infection time for each node is unknown. We call such tree $T_{\mathcal{S}}(v)$ the *cascading tree* where v is the root node. If we assign the time labels $\mathbf{t}$ to each node that is consistent with partial timestamps $\{t_s\}_{s\in\mathcal{S}}$, the resulting tree $T_{\mathcal{S}}(v, \mathbf{t})$ is called the *labeled cascading tree*, which uniquely recovers the infection process.

Since both infection starting time t_0 and the labeled cascading tree are unknown, it is natural to treat them as variables to be estimated jointly. Therefore, the approximation idea is to find the most likely labeled cascading tree, and then view the root node as the estimated source node $\hat{v}$. We call such estimator the *infection-path-based* estimator (INF), given by

$$\hat{T}_{\mathcal{S}}(\hat{v}, \hat{\mathbf{t}}) \in \arg\max_{T_{\mathcal{S}}(v,\mathbf{t})} \mathbf{P}(T_{\mathcal{S}}(v, \mathbf{t})|\mathcal{S}, v^* = v) \tag{2}$$

Mobihoc '19, July 2–5, 2019, Catania, Italy

ACM ISBN 978-1-4503-6764-6/19/07…$15.00
https://doi.org/10.1145/3323679.3326626

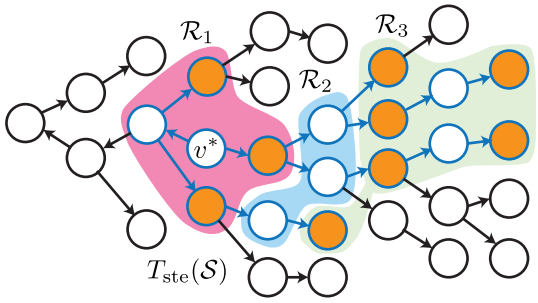


Figure 1: Illustration of the INF estimator in a tree.

If G is a tree, we can solve INF estimator in two steps:
Step 1: compute the likelihood of the most likely labeled cascading tree rooted at node $v \in V$. The cascading tree $T_{\mathcal{S}}(v)$ is unique, therefore it suffices to find the most likely time labels, which is the solution of the following integer linear programming (**ILP**)

$$\begin{array}{ll} \text{minimize (over } \mathbf{t}) & \sum_{(i,j)\in E(T_{\mathcal{S}}(v))} \mathbf{t}(j) - \mathbf{t}(i) \\ \text{subject to} & \mathbf{t}(u) = t_u \quad \forall u \in \mathcal{S} \\ & \mathbf{t}(j) - \mathbf{t}(i) \geq 1 \quad \forall (i,j) \in E(T_{\mathcal{S}}(v)) \\ & \mathbf{t}(u) \in \mathbb{Z} \quad \forall u \in V(T_{\mathcal{S}}(v)) \end{array} \tag{3}$$

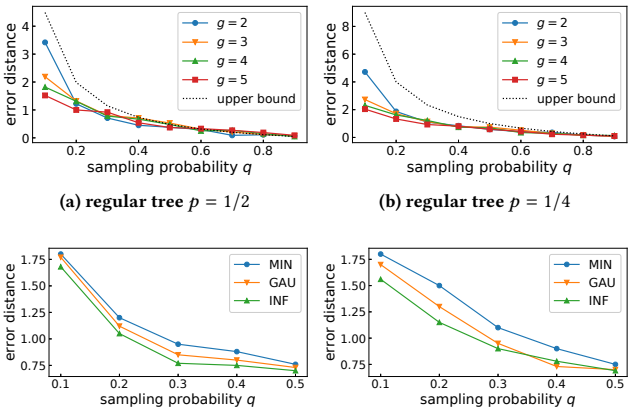
If **ILP**(3) is infeasible, there is no labeled cascading tree $T_{\mathcal{S}}(v, \mathbf{t})$ with positive likelihood. The **ILP**(3) and associated likelihood can be jointly solved in $O(N)$ time using message passing.
Step 2: find the reduced search region $\mathcal{V} \subset V$, the node in $\mathcal{V}$ with maximum likelihood is the estimated source $\hat{v}$. We define $d_{\mathcal{S}}(u, v)$ as the number of nodes with timestamps on the path between u and v. Inspired by the observe that the likelihood of $T_{\mathcal{S}}(v, \mathbf{t})$ is zero for any $\mathbf{t}$ if $d_{\mathcal{S}}(v, v^*) \geq 2$, we can prove that the estimated source $\hat{v} \in C \triangleq \{u | d_{\mathcal{S}}(v^*, u) = 1\} \cup \{u \in V(T_{\text{ste}}(\mathcal{S})) | d_{\mathcal{S}}(v^*, u) = 0\}$ where $T_{\text{ste}}(\mathcal{S})$ is the minimum Steiner tree spanning $\mathcal{S}$. However, C cannot be leveraged directly since the source v^* is unknown. We can find a slightly larger set $\mathcal{V}$ containing C using BFS starting from the node with minimum timestamp. The procedure is illustrated in Figure 1.

If G is a general network, it is computationally expensive to find the exact solution to INF since there are exponential number of possible cascading trees for each node. Therefore we utilize the BFS heuristic to find a partially labeled BFS tree $T_{\text{bfs}}(\mathcal{S}, v)$ which is consistent with the partial timestamps $\{t_s\}_{s\in\mathcal{S}}$ for each node $v \in V$. Then we compute the likelihood of the most likely labeled cascading tree using message passing. The estimated source is the node with maximum likelihood.

We characterize the accuracy of the INF estimator on the infinite g-regular trees in terms of error distance $d(\hat{v}, v^*)$. To do this, we prove in [3] the following two theorems:

Theorem 2.1. *The estimated source $\hat{v} \in \mathcal{P}^* \triangleq \bigcap_{s\in\mathcal{U}^*} \mathcal{P}(u^*, s)$ called* candidate path, *where $u^* = \arg\min_{u\in T_{ste}(\mathcal{S})} d(v^*, u)$ and $\mathcal{U}^*$ is a special set of nodes with timestamps.*

Theorem 2.2. *Assume that the nodes $\mathcal{S}$ are sampled uniformly at random with probability q. If $g = 2$, we have $\mathbf{P}(\hat{v} = v^*) = q + \frac{(1-q)pq(pq+3-3p)}{(pq+2-2p)(pq+1-p)}$ and $\mathbf{E}[d(\hat{v}, v^*)] \leq (1-q)\min\left\{\frac{1}{q}, \frac{2(1-p+pq)(1-p)^2}{pq(2-2p+pq)^2}\right\}$. If $g \geq 3$, we have $\mathbf{P}(d(v^*, \hat{v}) \leq D) \geq 1 - (1-q)(1-p+p(1-q)x_1)^g$ where x_1 is given by function iteration $x_D = 1$ and $x_i = h(x_{i+1}) = (1-p+p(1-q)x_{i+1})^{g-1}$ for $i = 1, 2, \ldots, D-1$. Denote the fixed point of $h(x)$ on $[0, 1]$ as x^*, then $\lim_{D\to\infty} x_1(D) = x^*$ and $x_1(D)$ is strictly decreasing with respect to D. Moreover, the INF estimator stochastically dominates MIN estimator in terms of error distance.*



(a) regular tree $p = 1/2$ **(b) regular tree $p = 1/4$**

(c) Erdös-Rényi (density 10.24) **(d) Scale-free (density 3.98)**

Figure 2: Simulation results on regular trees and general networks containing 1024 nodes.

3 SIMULATION RESULTS

We evaluate the accuracy of INF on different networks in terms of average error distance. As shown in Figure 2, the error distance decreases dramatically when q varies from 0.1 to 0.2 which implies that the INF performs fairly well even if the partial timestamps is small. We observe that the error distance is negatively correlated with degree g in Figure 2 (a)(b) indicating that the source with larger degree is more likely to be detected under INF. We compare the INF with the MIN and GAU [4] in Figure 2 (c)(d), and find that the INF performs no worse than GAU and MIN in almost all cases. This improvement is more obvious in scale-free network than in Erdös-Rényi network because the former is more tree-like.

ACKNOWLEDGMENTS

The work is supported by NSFC(61532012, 61822206, 61602303, 61829201), National Key R&D Program of China(2018YFB1004705), CCF Tencent RAGR(20180116), and Agri(X2017010).

REFERENCES

[1] G. Brightwell and P. Winkler. 1991. Counting Linear Extensions is #P-complete. In *STOC*. ACM, New York, NY, USA, 175–181.
[2] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran, and P. Viswanath. 2017. Hiding the Rumor Source. *IEEE Transactions on Information Theory* 63, 10 (Oct 2017), 6679–6713.
[3] X. Liu, L. Fu, B. Jiang, X. Lin, and X. Wang. 2018. Technical Report. https://www.dropbox.com/s/60l38w284j8rd2c/source-detection.pdf?dl=0. (Dec. 2018).
[4] P. C. Pinto, P. Thiran, and M. Vetterli. 2012. Locating the Source of Diffusion in Large-Scale Networks. *Phys. Rev. Lett.* 109 (Aug 2012), 068702. Issue 6.
[5] D. Shah and T. Zaman. 2011. Rumors in a Network: Who's the Culprit? *IEEE Transactions on Information Theory* 57, 8 (Aug 2011), 5163–5181.
[6] K. Zhu and L. Ying. 2016. Information Source Detection in the SIR Model: A Sample-Path-Based Approach. *IEEE/ACM Transactions on Networking* 24, 1 (Feb 2016), 408–421.