A Greedy Monitoring Station Selection for Rumor Source Detection in Online Social Networks

Rong Jin, *Member, IEEE*, Priyanshi Garg, Weili Wu, *Senior Member, IEEE*, Qiufen Ni, and Rosanna E. Guadagno

***Abstract*—In monitoring station observation, for the best accuracy of rumor source detection, it is important to deploy monitors appropriately into the network. There are, however, a very limited number of studies on the monitoring station selection. This article will study the problem of detecting a single rumormonger based on an observation of selected infection monitoring stations in a complete snapshot taken at some time in an online social network (OSN) following the independent cascade (IC) model. To deploy monitoring stations into the observed network, we propose an influence-distance-based *k*-station selection method where the influence distance is a conceptual measurement that estimates the probability that a rumor-infected node can influence its uninfected neighbors. Accordingly, a greedy algorithm is developed to find the best *k* monitoring stations among all rumor-infected nodes with a 2-approximation. Based on the infection path, which is most likely toward the *k* infection monitoring stations, we derive that an estimator for the "most like" rumor source under the IC model is the Jordan infection center in a graph. Our theoretical analysis is presented in the article. The effectiveness of our method is verified through experiments over both synthetic and real-world datasets. As shown in the results, our *k*-station selection method outperforms off-the-shelf methods in most cases in the network under the IC model.**

***Index Terms*—Independent cascade (IC) model, influence distance, monitor deployment, rumor source detection.**

I. Introduction

DURING the last decade, the information source detection problem has drawn increased attention since the seminal study regarding rumor source detection by Shah and Zaman [1]. Most existing works on finding the source of information propagation in networks are under an information diffusion model, such as the independent cascade (IC) model and the epidemic model. Based on the assumed information propagation model, many approaches to detection have been developed. Some detection approaches rely on some applicable observations, such as the states of the nodes and the timestamps at which nodes received (or infected by) the information. The information source detection problem has been studied in many application domains after Shah and Zaman's: for example, detecting the source of an epidemic to control the spreading of disease infection [2], identifying a virus source in a computer network [3], locating the source of gas leakage with a wireless sensor network [4], finding multiple sources of information propagation in complex networks [5], and investigating misinformation sources in online social networks (OSNs) [6]. In this article, we are interested in one application of detecting the source of the rumor, which is actively sparked off deliberately for disseminating fake information in the current online social media environment. Some comprehensive studies regarding this application have been investigated by multiple recent surveys [7], [8], [9].

In our work, we use the IC model [10] for simulating the rumor infection process through the OSN, and the network is considered as an undirected graph. There are two possible states of each node in the graph: rumor-infected (or active) and uninfected (or inactive). Rumor-infected nodes are users who adopted the rumor, and they cannot be deactivated. Initially, only one rumor-infected node is the rumor source. All rest nodes in the graph are considered to be in the uninfected state. Since then, the rumor source started to infect its currently inactive neighborhood based on the propagation probability upon edges between them. The rumor spreads out in the network after some time, given a complete observation of the network taken at some time, which contains rumor-infected and uninfected nodes. Our goal is to select a set of the best k rumor-infected nodes to be our monitoring stations and to detect the rumor source based on these k infection monitoring stations, even though without knowing the first infection time of each selected monitoring station and the neighbor from which the infection of the rumor is accepted.

To overcome the problem, our main contributions are summarized in the following paragraph.

Consider an OSN $G = (V, E)$ with the IC diffusion model, and there is a single rumor source. We first proved the metric of influence distance for each edge associated

Manuscript received 7 November 2022; revised 12 March 2023; accepted 6 June 2023. Date of publication 7 July 2023; date of current version 3 April 2024. This work was supported in part by the National Science Foundation under Grant 1822985, in part by the National Natural Science Foundation of China under Grant 62202109, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110321 and 2022A1515010611, and in part by Guangzhou Basic and Applied Basic Research Foundation under Grant 202201010676. *(Corresponding author: Rong Jin.)*

Rong Jin is with the Department of Computer Science, College of Engineering and Computer Science, California State University, Fullerton, CA 92831 USA (e-mail: rong.jin@fullerton.edu).

Priyanshi Garg and Weili Wu are with the Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: priyanshi.garg@utdallas.edu; weiliwu@utdallas.edu).

Qiufen Ni is with the School of Computers, Guangdong University of Technology, Guangzhou 510006, China (e-mail: niqiufen@gdut.edu.cn).

Rosanna E. Guadagno is with the Department of Persuasive Information Systems, University of Oulu, 90570 Oulu, Finland (e-mail: Rosanna.Guadagno@oulu.fi).

Digital Object Identifier 10.1109/TCSS.2023.3284909

with a propagation probability that node u can influence node v after it is active under the IC model. This proved measurement allows the computation of influence propagation to be transformed into the computation of hop distance on regular tree graphs. Based on the influence distance, we propose a k-station selection approach to deal with the deployment of monitoring stations with a greedy-based 2-factor approximation algorithm. And, based on the infection path that most likely yields to the selected k monitoring stations, the rumor source estimator is derived from being the root node along this most-likely infection path, and such root node is called the Jordan infection center in the graph and it can be found by a polynomial time algorithm. Our performance evaluations have experimented on both synthetic and real-world datasets. It can be seen from the experimental results that our proposed greedy k monitoring station selection method performs better than all other three state-of-art baseline methods in most cases. To the best knowledge, our work is first studied on a new monitor selection method for rumor source detection problems in the OSN. Also, the estimator using eccentricity in graph theory to identify the rumor source is first discussed under the IC model in our work, as it has only been studied in the epidemic models [11], [12], [13] before.

The rest of the article is organized as follows. Section II introduces the most relevant works. The problem formulation and the metric of influence distance under the IC model are shown in Section III. The greedy k-station selection method and the rumor source estimator on tree networks under the IC diffusion model are discussed in Section IV. Section V shows the experimental evaluation on our method using both synthetic and real-world datasets. Section VI concludes this article.

II. Related Work

Extensive existing studies on the rumor source detection problem in OSNs are assumed with two classical diffusion models—epidemic models like SI, SIR, SIS, or SIRS and IC model.

Shah and Zaman [1] first studied the single rumor source detection problem on regular trees under the epidemic SI model, and they proposed a rumor centrality in the graph and proved it to be the maximum likelihood estimator (MLE). Later, [11] proposed an infected path-based approach on tree network under the epidemic SIR model for the single source detection problem with a complete snapshot, in which the source estimator is proved to be a Jordan infection center. The approach is also applied in other different problem settings, such as partial observation of a snapshot [14], multiple sources in the network [15], and under a different epidemic SIS model [12].

Apart from the epidemic models, [16] first studied the single rumor source detection problem under the IC model using the monitor observation approach on tree graphs. Specifically, they incorporated their proposed metric of rumor source candidates called the rumor quantifier with three classic monitor selection methods—random, incoming degree (ID), and betweenness centrality (BC) on the detection precision. They concluded that when the random monitor selection method is used, the accuracy of locating the rumor source averagely increases with an increasing number of monitors in the network from 50 to 2000. And, the larger number of monitors is more effective in detection in practice. However, their suggested range (with exact numbers) of monitors does not practically work for monitor selections in different sizes of networks. Also, all three methods take all nodes into monitor selection, which results in misselection on the uninfected nodes that are unuseful as they lack infection information, which, in turn, induces our work that monitors are only selected from rumor-infected nodes instead of all nodes in the network for all compared monitor selection methods, and multiple percentages in the monitor selection for different sizes of synthetic and real-world networks are also discussed.

The following are two additional inspiring works. Lim et al. [17] formulated a rumor source detection problem based on the k-minimum distance error in a network under the IC model. The objective of which is to find the optimal set of k candidate rumor source nodes to further minimize the average shortest hop distance error from the actual rumor source to any node in the set of k candidate sources. The impact of the value of k on the average shortest hop distance error was discussed. In [18], the L1 distance between the observed states and expected states of the nodes was aimed to minimize with several heuristic algorithms based on a snapshot of tree networks under the IC model.

III. Problem Formulation

A. IC Model for Rumor Spreading

An OSN is represented by an undirected graph $G = (V, E)$, where nodes $v \in V$ represent users, and each edge in E represents interactions between two users. Each node has two possible states: rumor-infected (active) and uninfected (inactive). Time is discretized into multiple slots. Denote the state of each node v in time-slot t as $X_v(t)$. And, $X[0, t] = \{X_v(\tau) : 0 \leq \tau \leq t, v \in V\}$ is defined to be an infection path of the rumor, on which the states of nodes are known from time 0 to t.

All nodes in the graph except the rumor source are in an uninfected state at the beginning of time. Under the IC model, at the beginning of each time slot t, a rumor-infected node u attempts to influence its uninfected neighbor node v. If u succeeds, then v changes its state into infected at the next time slot $t + 1$; otherwise, v remains uninfected. There is only one chance for each infected node to infect each of its uninfected neighborhoods. Each attempt is independent of the others. The process of infection propagation runs until no node can be further activated. There is the *infection probability* (or called *propagation probability*) associated with each edge, which is also representing the chance of success for each attempt. Let $p_{uv} \in [0, 1]$ denote the infection probability associated with the edge (u, v), the next inactive neighboring node of v will be infected with probability $1 - \prod_{u \in N_{\text{active}}(v)}(1 - p_{uv})$, where $N(v)$ denotes all neighbors of node v. Note that $p_{uv} = p_{vu}$, as we consider the undirected graph. In addition, we assume an infected node retains the rumor forever once it is infected.

A complete snapshot $O = \{X_v(t), v \in V\}$ is taken at a time t, such that

$$v \in \begin{cases} O_a, & \text{if } v \text{ is rumor-infected} \\ O_b, & \text{if } v \text{ is uninfected.} \end{cases}$$

Define O_a as a set of rumor-infected nodes and O_b to be a set of uninfected ones. Our method is to detect the rumor source s^* based on an observation of an optimal subset $M \subseteq O_a$ of k rumor-infected nodes to be our monitoring stations (or called *monitors*). The observation time t is assumed to be unknown.

B. Influence Distance Under the IC Model

In an OSN following the IC model, there exists a concept of influence distance that measures the distance with respect to the propagation probability of the edge between two nodes. For example, the influence distance d_{uv} measures the probability that the node u can influence the node v. Moreover, the smaller the value of d_{uv}, the higher probability that influence can be diffused from u to v.

We start off by showing the proof of how we extracted the influence distance for each edge with propagation probability from the IC model in graph G.

Suppose that at a time t, there is a weighted path $X_{v_1 v_k}[0,t] =< p_{12}, p_{23}, \ldots, p_{i(i+1)}, \ldots, p_{(k-1)k} >$ from node v_1 to v_k, where $p_{i(i+1)}$ is the propagation probability of each edge between two nodes along the path. The propagation probability of the path $X_{v_1 v_k}[0,t]$ is then defined to be

$$\mathtt{Pr}\big(X_{v_1 v_k}[0,t]\big) = \prod_{i=1}^{k-1} p_{i(i+1)} \tag{1}$$

where the term of the product is over a set of edges on the path. By intuition, along the path $X_{uv}[0,t]$, the chance of node u activating node v is $\mathtt{Pr}(X_{uv}[0,t])$ and all nodes along this path need to be activated. Let $\chi_{uv}^G(t)$ denote a set of all paths from u to v in graph G observed at time-slot t, for the case that there are multiple paths between u and v. Let $\bar{\mathtt{Pr}}(X_{uv}[0,t])$ denote the path in $\chi_{uv}^G(t)$ that has the largest propagation probability. And we say, $\bar{\mathtt{Pr}}(X_{uv}[0,t])$ is *maximum propagation path* from u to v, and $\bar{\mathtt{Pr}}(X_{uv}[0,t]) \in \chi_{uv}^G(t)$.

Since there are many alternative joint paths from u to v, $\bar{\mathtt{Pr}}(X_{uv}[0,t])$ obviously cannot completely capture the influence diffusion. To handle this hardness, we borrow the kth influence distance defined in [19].

Definition 1: Let $\chi_{uv}^k(t)$ be the set of k independent paths that includes the maximum propagation probability from node u to node v at an observation time t. Define the kth influence distance d_{uv}^k between u and v to be

$$d_{uv}^k = -\ln\left(1 - \prod_{X_{uv}[0,t] \in \chi_{uv}^k(t)} (1 - \mathtt{Pr}(X_{uv}[0,t]))\right). \tag{2}$$

Upon the structure of $\chi_{uv}^k(t)$, the paths in $\chi_{uv}^k(t)$ are edge-disjoint so that u activates v independently of these paths. Hence, $1 - \prod_{X_{uv}[0,t] \in \chi_{uv}^k(t)} (1 - \mathtt{Pr}(X_{uv}[0,t]))$ represents the probability that u can activate v through the paths in $\chi_{uv}^k(t)$. We assume the maximum propagation path is always unique. Thus, by (2), we have influence distance for $k = 1$

$$\begin{aligned} d_{uv}^1 &= -\ln\left(1 - \prod_{X_{uv}[0,t] \in \chi_{uv}^1(t)} (1 - \mathtt{Pr}(X_{uv}[0,t]))\right) \\ &= -\ln\big(\bar{\mathtt{Pr}}(X_{uv}[0,t])\big). \end{aligned} \tag{3}$$

Lemma 1: Given a graph $G = (V, E)$ under the IC model, from node u to node v, let the path $X_{uv}[0,t]$ be a set of disjoint edges with propagation probabilities at an observation time t. The influence distance $d_{uv} = -\ln(\mathtt{Pr}(X_{uv}[0,t]))$.

Lemma 2: Given an edge-disjoint path $X_{uv}[0,t]$ between two nodes u and v with weighted edges under the IC model in the graph G. The influence distance d_{uv} is the sum over the influence distance of edges of the path from u to v.

Proof: Take aforementioned weighted path $X_{v_1 v_k}[0,t]$ from node v_1 to v_k as an example, then by (1), we have

$$\mathtt{Pr}\big(X_{v_1 v_k}[0,t]\big) = p_{12} \cdot p_{23} \cdot \cdots \cdot p_{(k-1)k}.$$

Next, by Lemma 1, it turns out that

$$\begin{aligned} d_{v_1 v_k} &= -\ln\big(\mathtt{Pr}\big(X_{v_1 v_k}[0,t]\big)\big) \\ &= -\ln\big(p_{12} \cdot p_{23} \cdot \cdots \cdot p_{(k-1)k}\big) \\ &= -\ln(p_{12}) - \ln(p_{23}) \cdots - \ln\big(p_{(k-1)k}\big) \\ &= d_{12} + d_{23} + \cdots + d_{(k-1)k}. \end{aligned}$$

□

One can see that the calculation of influence probability becomes to calculate the influence distance in G under the IC model. A short influence distance between two users implies that the rumor is more easily spread between them. Note that $d_{uv} = d_{vu}$ as we consider the undirected graph. Our method uses *influence distance* as *distance.*

C. Infection Path-Based Estimation

For the following rumor source estimator, as the observation time t is unknown, on tree networks following the IC model, we propose an infection path-based method to identify the infection path $X^M[0,t]$ that most likely proceeds to the observed set M of k infection monitoring stations from 0 to t. That is

$$X^M[0,t] = \arg_t \max_{X[0,t] \in \chi(t)} \mathtt{Pr}(X[0,t]) \tag{4}$$

where $\chi(t) = \{X[0,t] | M \subseteq O_a\}$, and $\mathtt{Pr}(X[0,t])$ is the probability of the path that leads to all infection monitoring stations in M. The root node associated with $X^M[0,t]$ is considered the rumor source.

IV. Single Rumor Source Estimation for Tree Networks

In this section, the underlying network G is assumed on trees following the IC model with only one single source. The greedy k monitoring station selection method is proposed. We then derive that the root node of the infection path $X^M[0,t]$, which most likely yields to the set M of k selected infection monitoring stations up to time t, has minimum infection eccentricity in G. As such, the root node is the Jordan infection center in G which is an optimal estimator for the rumor source. First, let us introduce the following definitions.

Definition 2: Let d_{vu} denote the distance, that is, the shortest path between two nodes v and u in the graph. Given a set M ($|M| > 0$) of explicitly observed infection monitoring stations, define the longest distance between node v and any monitoring station u to be

$$\bar{d}(v, M) = \max_{u \in M} d_{vu} \tag{5}$$

where $\bar{d}(v, M)$ is called the infection eccentricity of node v, denoted by $\tilde{e}_M(v)$ [11]. Any node with minimum $\tilde{e}_M(v)$ is defined as the Jordan infection center in the graph.

Definition 3: Given that v is the source of the infection process and a set M ($|M| > 0$) of explicit observed infection monitoring stations. Let $X_v^M[0, t] \in \chi_v(t)$ to be the infection path that most likely toward M up to time t

$$X_v^M[0, t] = \arg_t \max_{X[0,t] \in \chi_v(t)} \mathtt{Pr}\big(X[0, t] | s^* = v\big) \tag{6}$$

where $\chi_v(t)$ is viewed as the set of all possible infection paths beginning at v and their endpoints in M at a time slot t. $\mathtt{Pr}(X[0, t] | s^* = v)$ is the likelihood of obtaining the path $X_v^M[0, t]$ given the infection source v.

A. Infection Path Propagation Time

Lemma 3: Given a rumor-spreading network $G = (V, E)$ following the IC model where the rumor source is v, and a set M ($|M| > 0$) of explicit observed infection monitoring stations. The rumor spreading time along the infection path that most likely yields to M is given by $t_v^M = \bar{d}(v, M)$.

Proof: We analyze the time duration of the most likely infection path such that

$$t_v^M = \arg_t \max_{X_v[0,t] \in \boldsymbol{\chi}_v(t)} \mathtt{Pr}\big(X_v[0, t] | s^* = v\big) \tag{7}$$

which means that we want the time t_v^M maximizing the likelihood of obtaining the path covering the set M of observed infection monitoring stations.

Time is assumed as discrete time slots. Within one time slot, the propagation of infection is at most one hop further from the source v. If observation time $t < \bar{d}(v, M)$, the infection of the rumor cannot infect the nodes in the M. Hence, it is not possible for every node in M to get infected. Therefore, we have the observation time $t \geq \bar{d}(v, M)$.

Intuitively, an infection path involves more infected nodes as later as an observation time t is. Each infected node can contribute a probabilistic factor of $(1-p)$ in each time-slot t to the overall probability of infection path to be infected. Thus, the infection probability associated with the path $X_v^M[0, t]$ is monotonically decreasing with respect to t.

From above-mentioned two conclusions, and according to Lemma 1, it can be proved that

$$t_v^M = \bar{d}(v, M).$$

□

We will have unique t_v^M for each $v \in V$.

B. Infection Source Estimator

In a complete snapshot graph G of a rumor-spreading network, the *infection subgraph* can be formed by connecting the rumor-infected nodes including the rumor source. As such, the detection of the rumor source can be limited to search on the infection subgraph. Let t be the observation time, which is a random variable independent of the source node.

Lemma 4: Suppose the rumor source is node v of G and the rumor infection spreading follows the IC model. For a set M ($|M| > 0$) of explicit observed infection monitoring stations, let g denote the minimum connected infection subgraph of G that includes M and v, and let $t_v^M = \bar{d}(v, M)$ for each $v \in g$. Then, for any pair of neighbors u and v in g, if $t_v^M < t_u^M$, we have

$$\mathtt{Pr}\big(X_v^M[0, t_v^M]\big) > \mathtt{Pr}\big(X_u^M[0, t_u^M]\big). \tag{8}$$

Proof: If $t_v^M < t_u^M$, by Lemma 3, we can easily have $\bar{d}(v, M) < \bar{d}(u, M)$. According to Definition 2, we have $\tilde{e}_M(v) < \tilde{e}_M(u)$ in the infection subgraph g. Further by Lemma 1, which indicates that the path beginning at a root with a smaller infection eccentricity is with a larger likelihood over edges of the path, which means that the path is more likely to occur, we then can prove Lemma 4. □

Lemma 5 ([11]): On a tree network with at least one rumor-infected node, there exists at most two Jordan infection centers. Furthermore, when the network exactly has two Jordan infection centers, these two centers must be adjacent.

Combining Lemmas 3–5, we come up with Theorem 6.

Theorem 6: Suppose the rumor infection spreading follows the IC model. Consider a tree network graph $G = (V, E)$, and a nonempty set M including observed infection monitoring stations. Then, the single rumor-infected source associated with $X^M[0, t]$ in (4) is estimated by the following equation:

$$s^* = \arg\min_{v \in V} \tilde{e}_M(v). \tag{9}$$

Intuition: We will assume one Jordan infection center s^*. According to Lemma 5, when there exist two Jordan infection centers, they can be considered as one single node. Next, we will show that, for any node $a \in g \setminus \{s^*\}$, where g is the minimum connected infection subgraph of G that includes M and the infection source, there exists an infection path from a to s^* on which the infection eccentricity monotonically decreases, indicating that s^* is supposed to be a Jordan infection center.

Proof: In Fig. 1, on the infection path from a to s^*, assume that u is the neighbor node of s^*. Denote $T_u^{-s^*}(g)$ as the subtree of g starting with a root node u but without a branch from s^*, and the nodes $s^*, u \in g$.

It is obvious that there exists an observed infection node w such that $d_{wu} = \tilde{e}_M(s^*) - d_{us^*}$, where $\tilde{e}_M(s^*) = d_{ws^*}$ is assumed, and the node $w \in T_u^{-s^*}(g) \cap M$. Considering a node $l \in T_u^{-s^*}(g) \cap M$, we will hold that $d_{lu} \leq d_{ws^*} - d_{us^*}$.

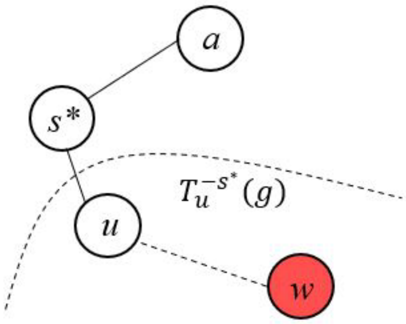


Fig. 1. Illustration for the proof in Theorem 6. Node w is viewed as an observed infection monitor in M.

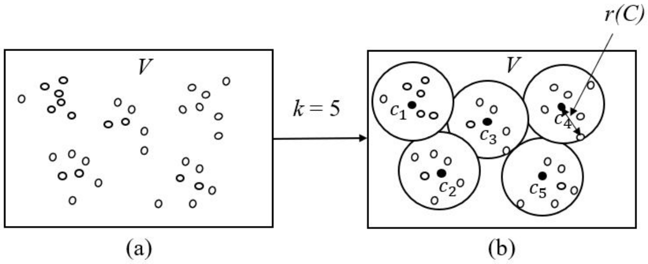


Fig. 2. Illustration of the k-center problem in a metric space with a finite set V of nodes, see Fig. 2(a). We can place 5 balls (value of k) of radius of $r(C)$ to cover V, where the set $C = \{c_1, c_2, c_3, c_4, c_5\}$, see Fig. 2(b).

Now consider $a \in g \setminus \{s^*\}$, and assume $a \in g \setminus T_u^{-s^*}(g)$, then, for any node $l \in T_u^{-s^*}(g) \cap M$, we have

$$\begin{aligned} d_{al} &= d_{as^*} + d_{s^*u} + d_{ul} \\ &\leq d_{as^*} + d_{s^*u} + d_{ws^*} - d_{us^*} \\ &= d_{as^*} + d_{ws^*} \end{aligned}$$

which denotes that $\tilde{e}(a) = d_{as^*} + d_{ws^*}$. Next, since $\tilde{e}(s^*) = d_{ws^*}$, it can be proved that the infection eccentricity decreases along the infection path from a to s^*. After that, by repeatedly using Lemma 4, we come up with a conclusion that the infection path with a root at node s^* is highly likely to happen than that rooted at node a from a to s^*. Therefore, this theorem holds. □

C. Monitoring Station Selection

In this section, let us discuss how to distribute k monitoring stations based on a complete snapshot $O = \{O_a, O_b\}$ of the undirected graph G taken at some time t. Given a positive integer k, a subset $M \subseteq O_a$ of k observed infection monitoring stations such that $|M| = k$. We expect the following formula to be smaller:

$$\min_{v \in V} \tilde{e}_{O_a}(v) - \min_{v \in V} \tilde{e}_M(v). \tag{10}$$

By Definition 2, $\tilde{e}_{O_a}(v) = \max_{u \in O_a} d_{vu}$, and $\tilde{e}_M(v) = \max_{u \in O_a \cap M} d_{vu}$, where d_{vu} denotes the influence distance between two nodes in our selection method. This expectation induces that the subset M is to be a solution to the k-center problem.

Definition 4 (The k-Center Problem): Given an undirected graph where a finite set V of nodes in space, and a positive integer $k \in |V|$, select a subset $C \subseteq V$ with $|C| = k$, called k cluster centers, such that the largest distance of any node in V to its cluster center is minimized. The problem is formally defined as follows:

$$\min_{C \subseteq V : |C| = k} \left(\max_{i \in V} d(i, C) \right) \tag{11}$$

where $d(i, C) = \min_{j \in C} d_{ij}$, and $\max_{i \in V} d(i, C)$ is called radius of C, that is the maximum value over all cluster center of C to their associated farthest neighbor node of V, denoted by the following equation:

$$r(C) = \max_{i \in V} d(i, C). \tag{12}$$

The k-center problem can be viewed by covering $|V|$ nodes using k balls. Given a node u in space, define $B(u, r)$ to be the ball of radius r whose center is at u. Suppose that C is any solution to the k-center problem, according to the definition of radius in (12), if we use balls of radius $r(C)$ to cover V, every node in V locates within the union of these balls. Also, $r(C)$ should be as small as possible. So for that one of the nodes of V will lie on the boundary of one of these balls. In other words, the neighbor nodes of each center $c_k \in C(k = 1, 2, \ldots, |C|)$ will lie within their associated ball. See Fig. 2.

Given the above-mentioned perspective of the k-center problem, we propose a k-station (monitor) selection problem as follows.

Definition 5 (k-Station Selection Problem): Here is a complete snapshot O of the OSN following the IC model for rumor propagation. Given a set $O_a \subseteq O$ of rumor-infected nodes and a positive integer $k \leq |O_a|$, find an optimal set M ($M \subseteq O_a$ with $|M| = k$) of balls of radius r as small as possible centered at k selected infection monitoring stations (nodes) such that O_a lies within the union of these k balls.

The selected k infection monitoring stations are considered representative centers covering all rumor-infected nodes in the network. Observation only in them can effectively help detect the source of the rumor.

Theorem 7: The k-station selection problem is NP-hard.

Proof: The k-center problem, like many clustering problems, is known as NP-hard. Identically, our proposed k-station selection problem is also NP-hard. □

As the proposed k-station selection problem is computationally hard, it is not possible to solve the problem exactly. Thus, we develop a greedy-based algorithm that achieves an approximation to the optimum value of radius r based on the influence distance in polynomial time in the worst case for general graphs.

Our greedy algorithm starts with selecting any node in O_a to be the initial monitoring station m_1. The following process is then repeated until we have the k monitoring stations. Denote $M_i = \{m_1, \ldots, m_i\}$ to be the current set of monitoring stations. Remember that, $r(M)$ is the maximum distance of any infected node in O_a from its nearest center in M. Let the node $u \in O_a \setminus \{m_1\}$ be the node that achieves this distance. Intuitively, u is the farthest user that accepts the rumor by its closest infection monitoring station. To satisfy u, the greediest approach is to directly select u to be the next monitoring station. In other

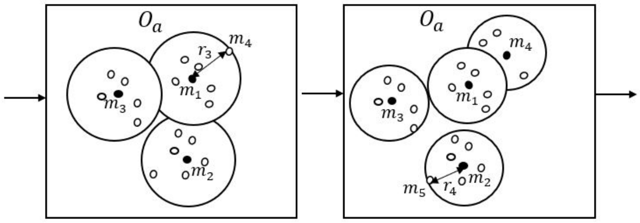


Fig. 3. Illustration of one step of the greedy approximation in Algorithm 1 to the proposed k-station selection problem.

words, we set that

$$m_{i+1} \leftarrow u$$

and

$$M_{i+1} \leftarrow M_i \cup \{m_{i+1}\}.$$

The pseudocode is presented in Algorithm 1. To be simplified, M is set to be empty. When the first monitoring station is selected, all the nodes of O_a then have infinite distances. Hence, the initial selection is arbitrary. And, the value of d_u indicates the influence distance from the node u to its closest monitoring station.

Algorithm 1 Greedy Influence-Distance-Based k-Station Selection

```
Input: $O_a \subseteq V$; a positive integer $k$
Output: $M \subseteq O_a$ with $|M| = k$
1: Initialize $M \leftarrow \emptyset$
2: for $u \in O_a$ do
3:     Initialize $d_u = \infty$
4: end for
5: for $i \leftarrow 1$ to $k$ do
6:     Let $u \in O_a \setminus \{m_1\}$ be the node s.t. $d_u$ is maximum
7:     $m_{i+1} \leftarrow u$
8:     $M_{i+1} \leftarrow M_i \cup \{m_{i+1}\}$
9:     for $v \in O_a$ do
10:        $d_v = \min(d_v, d_{vu})$ ▷ //update influence distance to nearest station
11:    end for
12:    $r = \max_{v \in O_a} d_v$ ▷ //update the radius $r$ of each station to its farthest neighbor node in $O_a$
13: end for
14: return $(M, r)$
```

Lines 5–8 of Algorithm 1 are illustrated in Fig. 3. Assuming that we have three monitoring stations $M = \{m_1, m_2, m_3\}$, let m_4 be the node that is farthest from its closest station, say m_1. Then, we create a station at m_4. And now, $M = \{m_1, m_2, m_3, m_4\}$. In anticipation of the next step, we find the node, say m_5, that maximizes influence distance to its nearest monitoring station. And if the algorithm continues, m_5 will be the location of the next monitoring station.

Theorem 8: Greedy influence-distance-based k-station selection algorithm runs in $O(k \cdot |O_a|)$.

Proof: According to Algorithm 1, we can simply figure out that its running time is $O(k \cdot |O_a|)$. Generally $k \leq |O_a|$, so for the worst case, its running time is $O(|O_a|^2)$. □

Theorem 9: Greedy influence-distance-based k-station selection algorithm is a 2-approximation solution for the k-station selection problem.

Proof: Let $M = \{m_1, \dots, m_k\}$ denote the set of monitoring stations computed by the greedy k-station selection algorithm, and $r(M)$ denotes the radius of M. Let $Y = \{y_1, \dots, y_k\}$ denote the optimum set of k monitoring stations such that $r(Y)$ is the smallest possible. We will show that the result produced by our greedy Algorithm 1 is guaranteed to be no more than twice the optimal value, that is

$$r(M) \leq 2r(Y).$$

Since Y and $r(Y)$ are unknown, our approach is to determine a lower bound $r_{\min}$ on the optimum value of the radius, that is, $r_{\min} \leq r(Y)$. Then we will show that our greedy Algorithm 1 generates a value of $r(M)$ that is at most twice this lower bound value of $r_{\min}$, that is, $r(M) \leq 2r_{\min}$. Thus, it will follow that $r(M) \leq 2r(Y)$. We will prove that based on three claims as follows.

Let us define M_i to be the set of monitoring stations selected by the greedy Algorithm 1 after its ith execution, and let $r_i = r(M_i)$ denote its overall radius—the farthest any node is from its nearest station in M_i. The algorithm stops with m_k, but for the purpose of the analysis, let us consider the next monitoring station to be appended if we execute it for an additional iteration. That is, let m_{k+1} denote the node of O_a that maximizes the value of $r(M_k)$ to its nearest station in M_k. Also, we define $M_{k+1} = \{m_1, \dots, m_{k+1}\}$.

Claim 9.1: For $1 \leq i \leq k+1$, $r_{i+1} \leq r_i$. That is, the sequence of the radius is monotonically nonincreasing.

Proof: Whenever a new monitoring station is added to the network, the distance to each node from its closest monitoring station will either be the same or will decrease. In Fig. 3, we also see the fact that the covering radii decrease with each step (e.g., $r_3 \geq r_4$). □

Claim 9.2: For $1 \leq i \leq k+1$, every pair of monitoring stations in M_i using greedy Algorithm 1 is separated by a distance of at least r_{i-1}.

Proof: Consider the ith step. By the induction hypothesis, the first $i-1$ stations are separated from each other by distances where $r_{i-2} \geq r_{i-1}$. By definition, the ith station is at the location with distance r_{i-1} from its closest station. Hence, it is at a distance of at least r_{i-1} from all the other monitoring stations. □

Claim 9.3: $r_{\min} \geq r(M)/2$.

Proof: We want to show that $r_{\min} = r(M)/2$, which means that $r_{\min} \geq r(M)/2$. To prove it, we will assume that $r_{\min} < r(M)/2$, then we use this assumption to derive a contradiction.

As $r_{\min}$ is assumed to be a lower bound on the optimum value of the radius, we then have $r_{\min} \leq r(Y)$. By Definition 4, we have $r(Y) \leq r(S)$ for any set S of k monitoring stations. Then, by our greedy Algorithm 1, we know that the maximum distance between any node and its closest monitoring station in M is at most $r(M)$. Consider any monitoring station y in the optimal solution Y. Since y is a cluster centering station, there must be at least one node x in the cluster such that $d_{yx} \leq r(Y)$. Now, let any monitoring station m in M be closest to x. By Definition 4, we know that $d_{xm} \leq r(M)$. Since d_{yx} and d_{xm} are nonnegative, and y and m are two distinct stations, by the triangle inequality, we have

$$d_{ym} \leq d_{yx} + d_{xm} \leq r(Y) + r(M).$$

Since y was an arbitrary monitoring station in Y, which implies that the maximum distance between any monitoring station in Y and its closest monitoring station in M is at most $r(Y) + r(M)$.

Now, we assume that $r_{\min} < r(M)/2$. Since we have $r_{\min} \leq r(Y)$, that implies $r(Y) > r(M)/2$; if we add $r(M)$ to both sides of this inequality, we then have $r(M)+r(Y) > r(M)/2+r(M)$, that is, $r(Y)+r(M) > \frac{3}{2}r(M)$, which contradicts the fact that the maximum distance between any monitoring station in Y and its closest monitoring station in M is at most $r(Y) + r(M)$. Therefore, we must have $r_{\min} \geq r(M)/2$. This completes the proof that $r_{\min} = r(M)/2$ if and only if M is an optimal solution to the k-station selection problem. □

Claim 9.4: Let $r_{\min} = r(M)/2$. Then for any set S of k monitoring stations, $r(S) \geq r_{\min}$.

Proof: By Definition 4, we know that every node of O_a lies within distance $r(S)$ of some point of S, and since $M \subseteq O_a$, this is true for M as well. Because $|M_{k+1}| = k+1$, by the pigeonhole principle, there exists at least two monitoring stations $m, m' \in M_{k+1}$ that is in the same neighborhood of some station $s \in S$, that is, $\max(d_{ms}, d_{m's}) \leq r(s)$. Since $m, m' \in M_{k+1}$, according to Claim 9.2, $d_{mm'} \geq r_k = r(M)$. Among the triple (m, s, m'), by applying the following two typical properties of any natural distance function, we then have

$$\begin{aligned} r(M) &\leq d_{mm'} \\ &\leq d_{ms} + d_{sm'} \text{(by triangle inequality)} \\ &= d_{ms} + d_{m's} \text{(by distance symmetry)} \\ &\leq r(s) + r(s) \\ &\leq r(S) + r(S) = 2r(S). \end{aligned}$$

Hence, $r(S) \geq r(M)/2$. By condition, $r_{\min} = r(M)/2$, then we have $r(S) \geq r_{\min}$, as desired. □

Now, by applying Claim 9.4 to the optimum set Y, we have that $r(Y) \geq r_{\min}$; by applying Claim 9.3, we have that $r_{\min} = r(M)/2$, and thus we can prove Theorem 9 that $r(M) \leq 2r(Y)$. □

D. Infection Path-Based Source Detection

An efficient algorithm to find the Jordan infection center has been discussed in [11] under the SIR model, which we will call infection path-based source detection (IPSD) algorithm under the IC model in this work. The details are described in Algorithm 2.

According to Theorem 6, the estimator for the rumor source based on the most-likely infection paths on trees is a Jordan infection center in the graph, which could be considered as a source candidate of the rumor.

The procedure of the IPSD algorithm is to find the Jordan infection centers based on the optimal infection paths toward the set of selected infection monitoring stations. First, let every selected infection monitoring station i in M disseminate the rumor message containing its station identification s_i to its neighbors. After every time slot, each neighbor node u will check whether the received message contains the station identification that was received before. If not received before, the node will record the new station identification as well as its received time of the message, say $t_u^{s_i}$. The node then spreads the message attached to the station identification to its neighborhood. When a node receives all k infection monitoring station identifications, it will announce itself as the rumor source, and the procedure terminates. If there are multiple nodes that receive all k infection monitoring station identifications at the same time, we break ties based on the infection closeness, which is defined as the inverse of the sum of distances from a node to all infected nodes in [11], which measures the efficiency of information propagating from a node to infected nodes. The IPSD algorithm breaks ties at random by selecting a Jordan infection center with the maximum infection closeness. In other words, we choose the node with the smallest $\sum t_u^{s_i}$ to break ties. And, the set R denotes the candidates of Jordan infection centers.

Algorithm 2 Infection Path-Based Source Detection

```
Input: set $M$ of $k$ infection monitoring stations; $g = (V, E)$
Output: the estimated rumor source $s^*$
Set $t = 1$
for $i \in M$ do
    $i$ disseminates the rumor message including its station identification $s_i$ to its neighbors
end for
do
    for $u \in V$ do
        if $u$ receives $s_i$ for the first time then
            Set $t_u^{s_i} = t$ and continue disseminating the rumor message including $s_i$ to its neighbors
        end if
    end for
    $t = t + 1$
while no node receives $k$ distinct station identifications
return ($s^* = \arg\min_{u \in R} \sum_{i \in M} t_u^{s_i}$), where $R$ is the set of nodes who receives $k$ distinct station identifications when the iteration terminates. Ties are randomly broken.
```

Theorem 10: The worst case complexity of Algorithm 2 is equal to $O(k \cdot |E|)$.

Proof: To implement Algorithm 2, we refer to some operations in [11]. For each node, we assign an array in the size k of integers and a counter. We also assign an index $i \in \mathbf{Z}$ to each infection monitoring station. Then, the data of the array equal to the influence distances from a node toward the infection monitoring stations. The counter records the number of distinct station indices received at the node. Let a message only include the index of an infection monitoring station. Every time when a node receives a new message, it will examine if the station index in the message has been received or not. If not, it will update the data, whose value equals to the influence distance from the current node to the infection monitoring station associated with the received station index, at the corresponding position of the array. Otherwise, the message will be discarded. At each iteration, each node propagates the new station indices to its neighborhood. After that, the counter of the node increases by one and checks whether its current value equals to k. And the complexity of the above-mentioned operation on one message is $O(1)$. Supposing that each edge handles to broadcast at most k messages in one direction, and thus there are at most $2k \cdot |E|$ messages in transmission. Therefore, our algorithm runs in $O(k \cdot |E|)$ in the worst case. □

V. Experiment

In this section, we conduct the evaluation on our proposed greedy influence-distance-based k-station selection method (referred to as **Greedy**) with our source estimator algorithm (IPSD) to identify the single rumor source based over synthetic d-regular trees and real-world networks. Also, we conduct the comparative analysis on our greedy monitoring station selection method with three monitor selection methods as discussed in [16] as follows.

1) *Random:* Select k monitors randomly in this method. Hence, the chance that $v \in O_a$ is chosen to be a monitor is $(k/|o_a|)$.
2) *Incoming Degree:* It is referred to as **Degree** in our experiment. In this method, select k nodes with the largest degrees as monitors. Ties are broken randomly.
3) *Betweenness Centrality:* In this method, select k nodes having the largest BC as monitors, and ties are broken randomly. We consider the hop distance to compute the shortest path between two nodes for calculating BC.

A. Experiment Design

Our experiment proceeds as follows: 1) the rumor source (the seed) is chosen at random on an undirected graph; 2) the IC model is applied to simulate the propagation of the rumor, and the infection probability of each edge is assigned; 3) either the rumor fails to propagate after two time-slots for d-regular trees, or it fails to reach 1% of all nodes for the real-world network, it will be considered as a negligible rumor and its cascade is discarded. A new rumor seed will then be chosen and the above-mentioned steps 1)–3) are repeated; 4) select k monitoring stations using greedy, random, degree, and BC-based selection methods, respectively. k is determined as follows: if there are x infected nodes at the end of propagation, $k = \lceil (y/100) * x \rceil$, where y is chosen to be $10, 40, 70,$ and 100 for d-regular trees and $10, 40,$ and 70 for real-world dataset; 5) the rumor source is identified using the IPSD algorithm for each monitor selection method; and 6) we simulate cascades 200 times for each value of y for both d-regular trees and real-world networks.

Detection rate and average hop distance are used to evaluate the effectiveness of each monitor selection method on the performance of our rumor estimator (IPSD). In our work, the detection rate is defined to be the percentage of experiments in which the predicated rumor seed matches the actual one. The hop distance between the predicated rumor seed and the actual rumor seed is determined for each cascade. The average is taken for all the cascades and is called the average hop distance.

TABLE I
Performance Analysis of Four Monitor Selection Methods Using the IPSD on Small-Infected Tree Networks. (a) Detection Rate (%). (b) Average Hop Distance

(a)

	10%	40%	70%	100%	Mean
Greedy	**10.00**	8.04	**10.50**	8.00	**9.135**
Random	9.50	**8.54**	8.50	8.00	8.635
Degree	9.50	7.54	9.00	8.00	8.510
BC	6.50	6.53	9.00	8.00	7.508

(b)

	10%	40%	70%	100%	Mean
Greedy	**1.62**	**1.40**	1.63	1.66	**1.578**
Random	1.74	1.50	1.63	1.66	1.633
Degree	1.94	1.55	**1.44**	1.66	1.648
BC	2.06	1.64	1.46	1.66	1.705

B. Performance Analysis on Synthetic Regular Trees

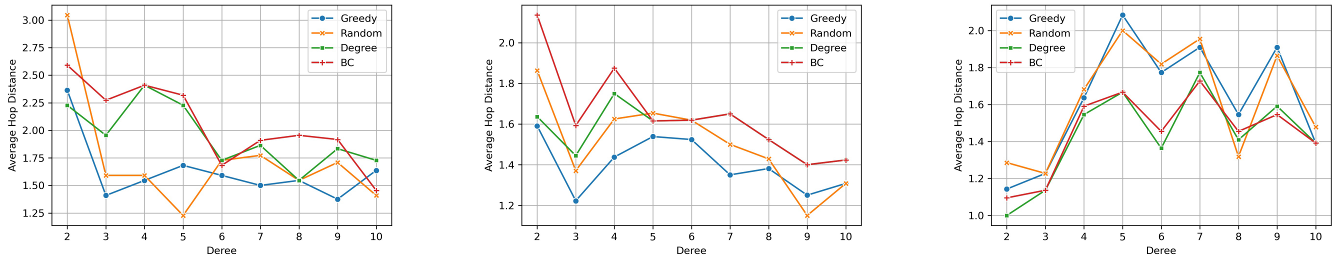
This section presents our assessment of the proposed greedy monitor selection method through contrasts with three other methods on synthetic $d-$regular trees, which are the trees where each node's degree is d other than the leaf nodes.

1) Small-Infected Trees: First, the performance was evaluated on the small-infected tree graphs. The infection probability p_{uv} for each edge was selected uniformly from $(0, 1)$. At the time slot t, a complete snapshot O was taken, and t was chosen uniformly from $(3, 20)$. At the end of propagation, the number of infected nodes was restricted to no more than 50. The average number of infected nodes among our experimental dataset being generated was 12. Degree d was varied from 2 to 10.

2) Large-Infected Trees: Next, the performance is evaluated on the large-infected trees. The infection probability p_{uv} for each edge was the same. For each rumor cascade, that probability was chosen randomly from $(0.5, 0.95)$. A complete snapshot O was taken at time-slot t and t was chosen uniformly from $(4, 100)$. Only those trees were selected that had more than 50 but no more than 500 infected nodes at the end of the rumor propagation. An average number of infected nodes was 190 among our experimental dataset being generated. Similar to the small-infected tree networks, degree d varied from 2 to 10.

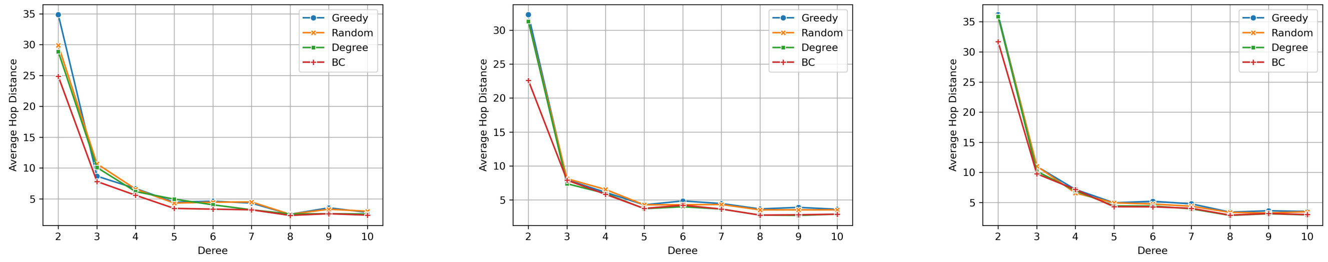
3) Experimental Results: Figs. 4 and 5 show the average hop distance between the predicted and actual rumor seeds as a function of degree d for all monitor selection methods with different numbers of monitors on the small-infected and the large-infected trees, respectively. In addition, Tables I and II show the statistical evaluation of detection rate (%) and average hop distance among all monitor selection methods with different numbers of monitors using our source estimation IPSD on the small-infected and the large-infected tree networks, respectively.

4) Major Observations: For small-infected tree networks, the average hop distance for the monitors selected through the



(a) (b) (c)

Fig. 4. Average hop distance as a function of degree (d), for small-infected tree networks, when 10%, 40%, and 70% of infected nodes are selected as monitoring stations (values of $|k|$), respectively. (a) 10%. (b) 40%. (c) 70%.



(a) (b) (c)

Fig. 5. Average hop distance as a function of degree (d), for large-infected tree networks, when 10%, 40%, and 70% of infected nodes are selected as monitoring stations (values of $|k|$), respectively. (a) 10%. (b) 40%. (c) 70%.

TABLE II

Performance Analysis of Four Monitor Selection Methods Using the IPSD on Large-Infected Tree Networks. (a) Detection Rate (%). (b) Average Hop Distance

(a)

	10%	40%	70%	100%	Mean
Greedy	**2.00**	0.47	**0.49**	1.00	**0.990**
Random	0.00	0.47	**0.49**	1.00	0.490
Degree	0.00	**1.90**	**0.49**	1.00	0.848
BC	0.00	0.47	0.00	1.00	0.368

(b)

	10%	40%	70%	100%	Mean
Greedy	8.02	7.73	8.76	7.89	8.100
Random	7.67	7.56	8.51	7.89	7.908
Degree	7.20	6.99	8.19	7.89	7.568
BC	**6.14**	**6.17**	**7.72**	7.89	**6.980**

proposed greedy k-station selection method is the least among all methods, as observed in Table I(b). Moreover, the least average hop distance noted is 1.40 using our proposed greedy selection method when merely 40% of infected nodes are selected as monitors. In addition, the average detection rate of the proposed greedy selection method is the highest compared to all other methods, as shown in Table I(a). We have also observed that, in general, the average hop distance decreases as the degree d increases for the cases of 10% and 40% of all infected nodes selected as monitors, while the average hop distance tends to increase as the degree increases when there are 70% infected nodes selected as monitoring stations, as shown in Fig. 4.

As seen in Table II for large-infected tree networks, the proposed greedy selection method has the highest average detection rate while it is smaller than that for the small-infected tree networks. When 10% of infected nodes is selected as monitoring stations, the average hop distance using the BC method is the least. We have also observed that the average hop distance decrease as the degree increases for all four monitor selection methods. And, the average hop distance is lower than 5 when degree $d > 6$, as displayed in Fig. 5.

TABLE III

Topological Properties of Real-World Networks. The Value of $|V|$ and $|E|$ Denote the Number of Nodes and Edges in a Network $G = (V, E)$, Respectively. ϕ Denotes the Diameter of the Network. $< d >$ Implies the Average Length of All Shortest Paths. $< k >$ Indicates the Average Network Degree

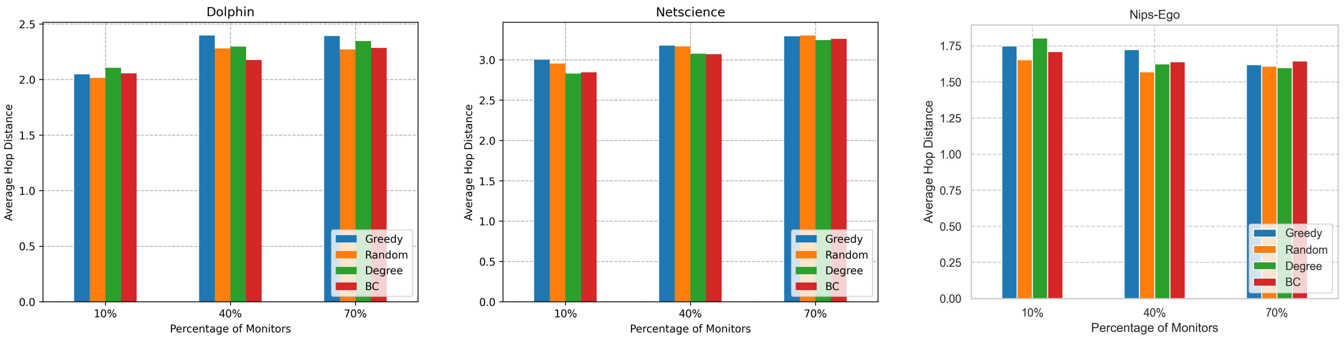
Dataset	$\lvert V\rvert$	$\lvert E\rvert$	ϕ	$< d >$	$< k >$
Dolphin	62	159	8	3.36	5.13
Netscience	379	914	17	6.04	4.82
Nips-Ego	2888	2981	9	3.87	2.06

C. Performance Analysis on Real-World Networks

This section compares our greedy monitor selection method with all three other methods on real-world networks.

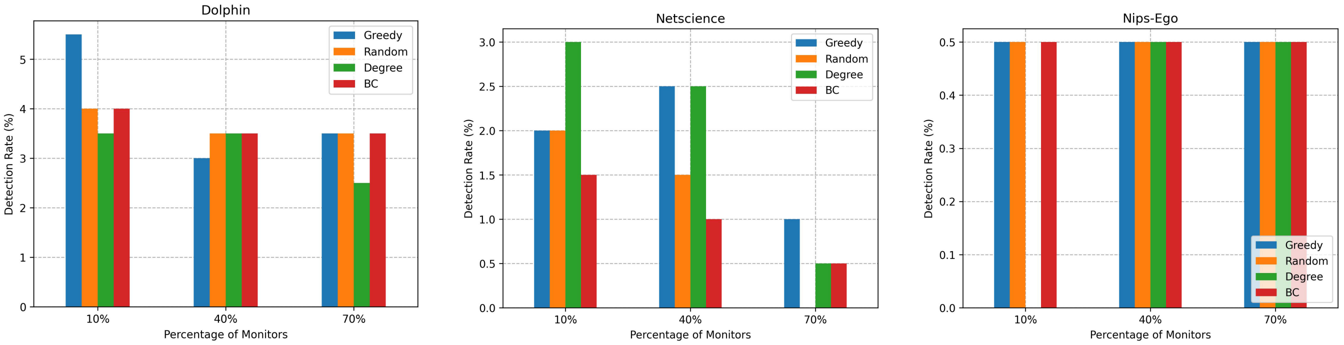
1) Datasets: We experiment on three datasets: Dolphin [20], Netscience [21], and NIPS-Ego (Facebook network) [22]. The basic topological information of them is displayed in Table III. All datasets are available online and can be downloaded from [23].

2) Parameter Settings: For all these networks, infection probability p_{uv} for each edge was set uniformly from $(0, 1)$. A complete snapshot was taken at time t chosen uniformly from $(3, 20)$. There was no restriction for the maximum size of infected nodes at the end of rumor diffusion. However, if the rumor failed to reach at least 1% of all nodes in the network, this propagation was discarded. Datasets Dolphin, Netscience,



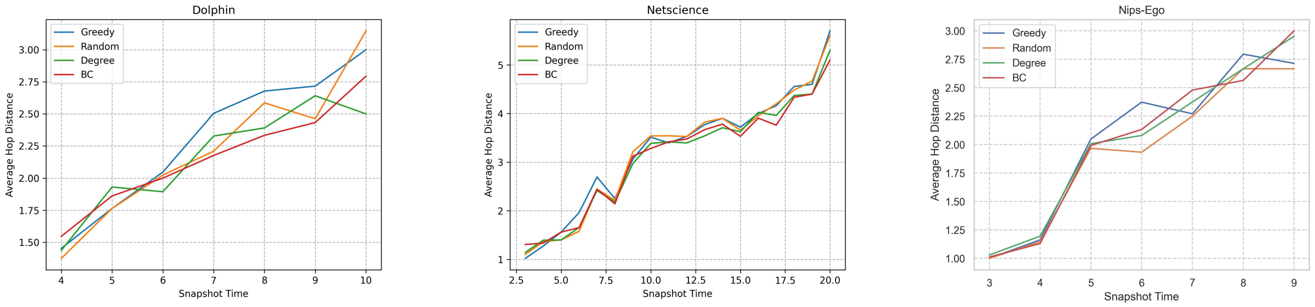
(a) (b) (c)

Fig. 6. Comparison of average hop distance as a function of the percentage of monitors between our greedy monitor selection method and the other three methods, when 10%, 40% and 70% of infected nodes are selected as monitors (values of $|k|$), respectively.



(a) (b) (c)

Fig. 7. Comparison of detection rate as a function of the percentage of monitors between our greedy monitor selection method and the other three methods, when 10%, 40%, and 70% of infected nodes are selected as monitors (values of $|k|$), respectively. The detection rate shows zero at 70% for the Random method (see Fig. 7(b)) and at 10% for the Degree method (see Fig. 7(c)).



(a) (b) (c)

Fig. 8. Average hop distance as a function of snapshot time between our greedy monitor selection method and the other three methods.

and Nips-Ego have an average of 51, 200, and 642 infected nodes, respectively.

3) Experimental Results: In Figs. 6 and 7, we compare average hop distance and detection rate as a function of the number of monitors between our greedy monitor selection method and all other three methods, respectively. In addition, we have also shown the average hop distance as a function of snapshot time in Fig. 8.

4) Major Observations: Fig. 6(a) and (b) shows that regardless of the monitor selection methods, the average hop distance increases as the selected number of monitoring stations increases. This is because in the relatively smaller dataset, even if more monitors are selected, they do not provide more evidence for the source estimator algorithm as they are likely to be closer to each other. Meanwhile, for the larger dataset, as observed in Fig. 6(c), the average hop distance decreases as the number of monitors increases, although the decrease is minor. Thus, we recommend using a reasonably smaller number of monitors as compared to using all infected nodes for estimating the rumor source no matter the size of the network. The overall observation shows our greedy method has a competitive performance with the other three monitor selection methods for all three datasets.

Fig. 7 shows that, in general, the detection rate tends to decrease as the network size increases. In Fig. 7(a) and (b), we observe that on average the detection rate of our proposed greedy selection method is higher in most cases, compared to other selection methods. In Fig. 7(c), we can see that as the dataset size grows in the number of nodes and edges, the detection rate decreases. This is also the case for an increase in the number of monitors. For example, the detection rate is 0 for 70% monitor selection in the Netscience dataset for the random method and for 10% monitor selection in the Nips-Ego dataset for the degree method. This is because there is an increase in the number of infection paths in the network for the source estimator algorithm as the number of edges and monitors increases. It can lead to the source estimation algorithm converging to another node (i.e., source estimator) and not the actual source.

In addition, we expect if the snapshot is taken at an earlier stage, we can predict the actual rumor source more accurately. This is evident in Fig. 8 for all three datasets, as the snapshot time increases average hop distance also increases.

Based on the observations for both synthetic and real-world datasets, we observe that average hop distance is a better metric to evaluate the performance of the rumor source detection algorithm, because inference from the result of the detection rate may be misleading. We can conclude that an appropriate monitor deployment strategy with a reasonably small size of monitoring stations selected from the infected nodes at some given snapshot is a more efficient way in terms of performance of detection and time complexity.

VI. Conclusion

In this article, for some OSNs under the IC model, we explain how to extract the influence distance that allows the computation of influence propagation to be transformed into the computation of hop distance. We first propose a novel greedy k-station selection method for the single rumor source detection problem in the undirected graph. The rumor source estimator is derived to be the Jordan infection center upon the optimal infection path that leads to the selected k monitoring stations. We validate the performance of our greedy monitor selection method with comparisons with other monitor selection methods using both synthetic and real-world networks. It can be observed from the results that our method outperforms the others in most cases for synthetic regular trees, and it is comparable to other methods for real-world networks. With a reasonably small number of monitoring stations itself, our rumor source estimator performs better in terms of performance and time complexity. In addition, we evaluate the performance of all methods as a function of snapshot time and we observe that the performance of source detection drops as later the snapshot is taken for all datasets. Last but not the least, using average hop distance as the performance metric for the rumor source detection compared to the use of detection rate is recommended.

For future work, we first expect to apply our greedy monitor selection method for other information diffusion models like a linear threshold model. Next, under the IC model, we expect to extend our selection and source estimation methods for multiple rumor source detection in OSNs. It should be noted that implementing our greedy selection method and the IPSD algorithm in considerably larger networks could pose a challenge. Therefore, we also expect that using machine learning techniques can provide opportunities to optimize monitor selection and improve the accuracy, efficiency, and scalability of rumor source detection problems in real-world significantly larger networks.

References

[1] 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.

[2] N. Antulov-Fantulin, A. Lančić, T. Šmuc, H. Štefančić, and M. Šikić, “Identification of patient zero in static and temporal networks: Robustness and limitations,” *Phys. Rev. Lett.*, vol. 114, no. 24, Jun. 2015, Art. no. 248701.

[3] D. Shah and T. Zaman, “Detecting sources of computer viruses in networks: Theory and experiment,” *SIGMETRICS Perform. Eval. Rev.*, vol. 38, no. 1, pp. 203–214, Jun. 2010.

[4] L. Shu, M. Mukherjee, X. Xu, K. Wang, and X. Wu, “A survey on gas leakage source detection and boundary tracking with wireless sensor networks,” *IEEE Access*, vol. 4, pp. 1700–1715, 2016.

[5] 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, Nov. 2011, Art. no. 056105.

[6] D. T. Nguyen, N. P. Nguyen, and M. T. Thai, “Sources of misinformation in online social networks: Who to suspect?” in *Proc. MILCOM IEEE Mil. Commun. Conf.*, Oct. 2012, pp. 1–6.

[7] J. Jiang, S. Wen, S. Yu, Y. Xiang, and W. Zhou, “Identifying propagation sources in networks: State-of-the-art and comparative studies,” *IEEE Commun. Surveys Tuts.*, vol. 19, no. 1, pp. 465–481, 1st Quart., 2017.

[8] S. Shelke and V. Attar, “Source detection of rumor in social network—A review,” *Online Social Netw. Media*, vol. 9, pp. 30–42, Jan. 2019.

[9] R. Jin and W. Wu, “Schemes of propagation models and source estimators for rumor source detection in online social networks: A short survey of a decade of research,” *Discrete Math., Algorithms Appl.*, vol. 13, no. 4, Aug. 2021, Art. no. 2130002.

[10] D. Kempe, J. Kleinberg, and É. Tardos, “Maximizing the spread of influence through a social network,” in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2003, pp. 137–146.

[11] 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.

[12] W. Luo and W. P. Tay, “Finding an infection source under the SIS model,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 2930–2934.

[13] W. Luo and W. P. Tay, “Estimating infection sources in a network with incomplete observations,” in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2013, pp. 301–304.

[14] K. Zhu and L. Ying, “A robust information source estimator with sparse observations,” in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2014, pp. 2211–2219.

[15] 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.

[16] W. Xu and H. Chen, “Scalable rumor source detection under independent Cascade model in online social networks,” in *Proc. 11th Int. Conf. Mobile Ad-Hoc Sensor Netw. (MSN)*, Dec. 2015, pp. 236–242.

[17] S. Lim, J. Hao, Z. Lu, X. Zhang, and Z. Zhang, “Approximating the k-minimum distance rumor source detection in online social networks,” in *Proc. 27th Int. Conf. Comput. Commun. Netw. (ICCCN)*, Jul. 2018, pp. 1–9.

[18] T. Lappas, E. Terzi, D. Gunopulos, and H. Mannila, “Finding effectors in social networks,” in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 1059–1068.

[19] G. A. Tong, S. Li, W. Wu, and D. Du, “Effector detection in social networks,” *IEEE Trans. Computat. Social Syst.*, vol. 3, no. 4, pp. 151–163, Dec. 2016.

[20] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, “The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations,” *Behav. Ecol. Sociobiol.*, vol. 54, no. 4, pp. 396–405, Sep. 2003.

[21] M. E. J. Newman, “Finding community structure in networks using the eigenvectors of matrices,” *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 74, no. 3, Sep. 2006, Art. no. 036104.

[22] A. L. Traud, P. J. Mucha, and M. A. Porter, “Social structure of Facebook networks,” *Phys. A*, vol. 391, no. 16, pp. 4165–4180, Aug. 2012.

[23] R. Rossi and N. Ahmed, “The network data repository with interactive graph analytics and visualization,” vol. 29, Mar. 2015. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/9277, doi: 10.1609/aaai.v29i1.9277.

Rong Jin (Member, IEEE) received the Ph.D. degree from the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA, in 2021.

She is currently an Assistant Professor at the Department of Computer Science, California State University, Fullerton, CA, USA. Her current research interests include computational influence problems, rumor source detection problems in online social networks, optimization, machine learning, deep learning, multimedia intelligent systems, digital realities, 3-D simulation, serious games, and computer vision.

Priyanshi Garg received the B.Tech. degree in information technology from the SRM Institute of Technology, Kattankulathur, India, in 2019. She is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. Her research interests are mainly in methodologies for rumor source identification in online social networks and the design and analysis of algorithms.

Weili Wu (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the Department of Computer Science, University of Minnesota, Minneapolis, MN, USA, in 1998 and 2002, respectively.

She is currently a Professor and the Director at the Data Communication and Data Management (DCDM) Laboratory, Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA. She has authored more than 228 journal papers and 102 conference papers in various prestigious journals and conferences such as IEEE TRANSACTIONS ON NETWORKING, *ACM Transactions on Knowledge Discovery in Data*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE International Conference on Distributed Computing Systems, INFOCOM, and *ACM Special Interest Group on Knowledge Discovery and Data Mining*. Her research interests are mainly in big data, social networks, blockchain technology, wireless sensor networks, IoT, and data mining.

Dr. Wu is an Associate Editor of the *International Journal of Bioinformatics Research and Applications* (IJBRA), *Computational Social Networks* (CSN), SOP *Transactions on Wireless Communications* (STOWC), *Journal of Combinatorial Optimization* (JOCO), and *Journal of Global Optimization* (JOGO).

Qiufen Ni received the Ph.D. degree from the School of Computers, Wuhan University, Wuhan, China, in 2020. She is currently pursuing the joint Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas, Richardson, TX, USA.

She is currently an Assistant Professor at the School of Computers, Guangdong University of Technology, Guangzhou, China. Her research interests are mainly in optimization problems in social networks and wireless networks, and the design and analysis of approximation algorithms.

Dr. Ni is a Technical Committee Member of the Theoretical Computer Science Branch Committee of the China Computer Federation.

Rosanna E. Guadagno received the Ph.D. degree in social psychology from Arizona State University, Tempe, AZ, USA, in 2003.

She completed her Post-Doctoral work at UC Santa Barbara, Santa Barbara, CA, USA, and previously taught at The University of Alabama, Tuscaloosa, AL, USA, UT Dallas, Richardson, TX, USA, UC Berkeley, Berkeley, CA, USA, UC Santa Cruz, Santa Cruz, CA, USA, and Stanford University, Stanford, CA, USA. She is also the Former Program Director at the National Science Foundation where she managed three programs: Social Psychology, the Science of Learning Centers, and Secure and Trustworthy Cyberspace (SaTC). She is currently an Associate Professor at the Department of Persuasive Information Systems, University of Oulu, Oulu, Finland. Her work has been published in journals, such as *Perspectives on Psychological Science*, *Psychological Inquiry*, *Personality and Social Psychology Bulletin*, *Computers in Human Behavior*, *Media Psychology*, *CyberPsychology, Behavior*, and *Social Networking*, and *Sex Roles*; covered in the press by: CBS News, The New York Times, The Atlantic Monthly, The New Yorker, The Associated Press, ESPN, New Scientist, MSNBC, and Alabama Public Radio. Her research interests focus on the confluence of three main areas: social influence and persuasion, mediated communication, and gender roles.