IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received February 6, 2020, accepted April 11, 2020, date of publication April 20, 2020, date of current version May 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2988910

Benefits of Bias in Crawl-Based Network Sampling for Identifying Key Node Set

SHO TSUGAWA[1], (Member, IEEE), AND HIROYUKI OHSAKI[2], (Member, IEEE)
[1]Faculty of Engineering, Information and Systems, University of Tsukuba, Tsukuba 305-8573, Japan
[2]School of Science and Technology, Kwansei Gakuin University, Sanda 669-1337, Japan

Corresponding author: Sho Tsugawa (s-tugawa@cs.tsukuba.ac.jp)

This work was supported in part by the JSPS KAKENHI under Grant 19K11917.

ABSTRACT We study the problem of identifying a set of key nodes from a network when limited knowledge about its structure is available. Most studies assume complete knowledge of the given network when identifying a set of key nodes, but in current practice, networks of interest are often too huge to obtain their entire topological structures. When the complete structure of a network is not available, network sampling strategies are often used to obtain a partial structure of the network. We investigate how network sampling strategies affect the problem of identifying a key node set. Specifically, we investigate the effect of conventional network sampling strategies on the solutions found for two types of key node set identification problems: the minimum *p*-median problem and the influence maximization problem. Our results show that when the network is obtained using crawl-based network sampling strategies, both the minimum *p*-median and the influence maximization problems are effectively solved by simple heuristic algorithms with sampling ratios in the 10–20% range. We also find that among three conventional sampling strategies (random sampling, random walk sampling, and sample edge counts) checked in this paper, random walk sampling is the most robust strategy in terms of effectively identifying the key node sets of diverse types of networks.

INDEX TERMS Influence maximization problem, key node set identification, minimum *p*-median problem, network sampling, social networks.

I. INTRODUCTION

Identifying a set of *key* nodes in a given network is a fundamental research problem in network science research, and it has broad application [1]–[6]. Examples of the key node set identification problem include classical problems in graph theory such as minimum *p*-median (MM) and minimum *p*-center problems [3], [7]. The MM problem is motivated by applications to city planning, and is useful for determining facility locations [3]. The influence maximization (IM) problem is another popular key node set identification problem, which is expected to be useful for so-called "viral" marketing in social networks [4], [8]–[15]. IM aims to identify a small set of influential nodes (called seed nodes) for which the expected size of the influence cascade triggered by the seed nodes is maximized [8]. Note that the problem of identifying *k* most important nodes using centrality or other metrics [16]–[19] in a network is also considered to be a key node set identification problem.

One of the main research challenges in identifying key node sets has been to develop computationally efficient algorithms [4], [7], [20]. Because key node set identification problems are typically NP-hard [7], [8], naive algorithms for the problems are infeasible for use with large-scale networks. Many researchers have proposed efficient algorithms for key node set identification. Some algorithms offer theoretical guarantees on the quality of the solutions, and others are heuristic algorithms without theoretical guarantees [4], [7]. Thanks to the efforts of these researchers, scalable key node set identification algorithms are widely available [4], [7], [9], [13], [14].

However, some issues remain open in key node set identification. Notably, most existing studies assume complete knowledge is available for a given network whose key nodes are to be identified, but modern networks of interest are often too large for their entire topological structures to be efficiently known [21]–[29]. For instance, social networks,

The associate editor coordinating the review of this manuscript and approving it for publication was Ghufran Ahmed.

which represent relationships among social media users, are often very large, and access to network data is typically limited, so only a part of the network structure can be known [21], [24], [30]. It is also fundamentally difficult to obtain the current complete structure of the Internet due to both its scale and its distributed, heterogeneous nature [31], [32].

When the complete structure of a network is not available, a network sampling is often used to obtain a partial structure of the network [22], [23], [26]–[30], [33]. If arbitrary access to any nodes is allowed, random sampling seems to be a natural choice for simplicity because of its simplicity and neutrality. However, in many real applications, random access to nodes is not allowed [30], [33], so crawl-based sampling techniques have been widely used for analyzing the structure of several types of large-scale networks, such as online social networks [23], the world wide web [34], and peer-to-peer (P2P) networks [35]. When using crawl-based network sampling, it is assumed that only one node can be visited initially, but the neighbors of already visited nodes can be visited at each step. Popular crawl-based network sampling strategies include random walk (RW) sampling [36], breadth-first search (BFS), depth-first search (DFS), and sample edge counts (SEC) [22]. It is known that crawl-based sampling strategies have a *bias* toward high-degree nodes; that is, when using crawl-based network sampling strategies, the probability of visiting high-degree nodes is much higher than that of visiting low-degree nodes [22]. This bias in crawl-based network sampling is generally not desirable when estimating network characteristics, and therefore, considerable effort has been devoted to eliminating the bias in crawl-based sampling strategies [23], [29], [37]. In contrast with the problem of general characterization of network topology, the bias in crawl-based sampling might be beneficial when finding key nodes in a network. A pioneering work by Maiya and Berger-Wolf [22] suggested the benefit of the bias of crawl-based sampling strategies in identifying high-degree nodes. Their finding suggest the hypothesis that *the bias in crawl-based sampling is beneficial also for identifying the set of key nodes*.

This paper revisits the benefit of the biases in crawl-based network sampling strategies and examines how the crawl-based network sampling applied to a given network affects identification of the key node set in the network. It is naturally expected that when the sample size is small, identifying a key node set from such a partial network will be quite difficult or even impossible. However, because of the bias in crawl-based sampling, we expect that the key node set can be identified even from the limited knowledge obtained with crawl-based sampling. We address the following research questions in particular. (1) How do network sampling strategies affect the effectiveness of key node set identification algorithms? (2) How large a sample do we need in order to obtain a reasonable solution for key node set identification problems? To answer these questions, we formulate two types of key node set identification problems, assuming limited knowledge about the network structures. These two problems are variants of popular key node set identification problems: the MM and the IM problems. We apply simple heuristic algorithms to the problems, and examine the effects of network sampling on both of the MM and IM problems.

Our main contributions are summarized as follows.

- We show that the biases in crawl-based sampling strategies are beneficial in both the MM and IM problems. Although the definitions of the *key* nodes in the MM and IM problems are notably different, similar simple heuristic algorithms can achieve reasonable solutions to both problems when the partial structure of the network is obtained via crawl-based sampling.
- We demonstrate that crawl-based sampling strategies require only 10–20% sample sizes to obtain reasonable solutions of both the IM and MM problems. When a 10–20% sample size is available, heuristic algorithms can find a key node set that is comparable to the key node set obtained from examining the complete network.
- We reveal that a moderate level of bias in crawl-based sampling strategies is central to identifying the key node set effectively and robustly in an unknown network; that is, weak bias in the sampling strategies degrades the quality of solutions, and strong bias deteriorates the stability of the solutions.

This paper is organized as follows. In Section II, we provide definitions and give the formulations of the problems studied in this paper. In Section III, we explain the research methodology. In Sections IV and V, we present the results for the MM problem and the IM problem, respectively. In Section VI, we discuss the implications and future directions of this work. Finally, in Section VII, we conclude this paper.

II. PRELIMINARIES

A. DEFINITIONS

This paper considers the class of problems that require finding a set of key nodes in a *ground truth* network $G = (V, E)$ using only an *incomplete* subnetwork $G' = (V', E')$. The incomplete network G' is obtained by applying network sampling strategies to the ground truth network G. The graph G can be either directed or undirected, but for simplicity, in what follows, G is assumed to be undirected. Note that we consider network sampling and key node set identification *independently*. Although a problem of jointly optimizing both of the network sampling and key node set identification can be formulated, studying such problem is beyond the scope of this paper.

A network sampling strategy probes to obtain the nodes in $S \subseteq V(|S| = M)$, where M is called the sample size. Probing node v reveals the nodes adjacent to v. Let T be a set of nodes adjacent to nodes in S. Then, the set of nodes in the incomplete network G' is $V' = S \cup T$. The set of links in the incomplete network G' is $E' = \{(u, v) | u \in S \cap v \in T\}$.

B. MINIMUM p-MEDIAN PROBLEM FOR INCOMPLETE NETWORKS

Before introducing the MM problem for incomplete networks, we explain the original MM problem. Given an undirected unweighted network $G = (V, E)$, each node $v \in V$ has a demand $w(v)$. Let $d(u, v, G)$ be the shortest path length between nodes u and v in the network G, and let $D(v, X, G) = \min[d(u, v, G) : u \in X]$. Then, the MM problem is defined as follows.

Problem 1 Minimum p-Median Problem [3]: Given a network G and $w(v)$ for each node $v \in V$, find a set of p nodes X ($|X| = p, X \subseteq V$) such that the objective function (i.e., total cost) $f(X) = \sum_{v \in V} w(v)D(v, X, G)$ is minimized.

We now define the minimum p-median (MM) problem for incomplete networks. To the best of our knowledge, the MM problem under limited knowledge about the network is a novel problem that has not been studied before. In the original MM problem, the network G and the demand for all nodes are available. In contrast, in the MM problem for incomplete networks, only the network G' is available for finding the median node set X that minimizes the total cost $f(X)$. The MM problem for incomplete networks is defined as follows.

Problem 2 Minimum p-Median Problem for Incomplete Networks: A subnetwork $G' = (V', E')$ of the ground truth network G is obtained through sampling the nodes in a node set $S \subseteq V$. For each node $v \in V'$, its demand $w(v)$ is given. From these, find a set of p nodes X ($|X| = p$, $X \subseteq V$) that minimizes the objective function $f(X) = \sum_{v \in V} w(v)D(v, X, G)$.

C. INFLUENCE MAXIMIZATION PROBLEM FOR INCOMPLETE NETWORKS

The IM problem for incomplete networks is formulated analogously to the MM problem for incomplete networks. We first explain the original IM problem. Influence maximization problem is a combinatorial optimization problem on a graph that aims to identify a small set of influential nodes (known as seed nodes) such that the expected size of the influence cascade triggered by the seed nodes is maximized. While the IM problem has been studied under several types of influence cascade, the independent cascade (IC) model [8] is the most popular. This paper focuses on the IM problem using the IC model, although our problem formulation can be extended to the IM problem with other types of influence cascade. In the IC model, each node is either active or inactive. When node u becomes active at time step t, node u will influence each inactive neighbor node v ($(u, v) \in E$) with probability $p_{u,v}$ at the next time step $t + 1$. Namely, a node v becomes active with probability $p_{u,v}$. The parameter $p_{u,v}$ of the IC model is the probability of spreading influence between nodes u and v. Note that each node has a single chance to influence each of its neighbor. At time step 0, the nodes selected as seed nodes ($U \in V$) become active, and the other nodes are set as inactive. Then, the stochastic process explained above is repeated until it ends (i.e., no nodes are newly activated in the time step). Let W be a set of link weights representing the probability of influence spread, $U \subseteq V$ be a subset of nodes in graph G, and $\sigma(U, G, W)$ be the expected number of active nodes at the end of the process of the IC model on network G with probabilities W when U is the set of seed nodes. The IM problem is then defined as follows [8].

Problem 3 Influence Maximization Problem: Given a social network G, influence spread probabilities W, and an integer k, the aim is to find a set of seed nodes U ($U \subseteq V$, $|U| = k$) such that $\sigma(U, G, W)$, which we call the influence spread, is maximized under the IC model.

In contrast with the original IM problem, in the IM problem for incomplete networks, the ground truth network G is not available for use in finding a set of seed nodes. Only a subnetwork $G' = (V, E')$ of G is available. The IM problem for incomplete networks is then defined as follows.

Problem 4 Influence Maximization Problem for Incomplete Networks: Given an incomplete network G', influence spread probabilities $W' = \{p_{u,v} | (u, v) \in E'\}$, and an integer k, find a set of seed nodes U ($U \subseteq V$, $|U| = k$) such that $\sigma(U, G, W)$ is maximized under the IC model.

The IM problem under limited knowledge about the network was first proposed in our previous conference papers [24], [38] and has also been studied by other research groups [39], [40]. In this paper, through extensive experiments, we comprehensively investigate the effects of network sampling on both the IM problem and the MM problem.

III. METHODOLOGY

A. GENERATING INCOMPLETE NETWORKS

We generate an incomplete network G' from a given ground truth network G using the following three network sampling strategies.

1) SAMPLE EDGE COUNT (SEC) [22]

SEC aims to obtain high-degree nodes without global knowledge of the network by greedily taking the node with the highest expected degree among known but unselected nodes. Let S be a set of chosen nodes. Initially, S contains a randomly selected node. SEC greedily obtains the node with the most links from the nodes in S. This method greedily obtains the node with the highest expected degree. SEC is intended to have a strong bias toward high-degree nodes.

2) RANDOM WALK (RW) [36]

Initially, RW obtains and visits a randomly selected node. Then, RW repeatedly obtains and visits a randomly selected unvisited neighbor node of the most recently visited node until a specified number of nodes is obtained. If a visited node has no unvisited neighbor, then a randomly selected unvisited neighbor of some other visited node is obtained. RW does not intentionally visit high-degree nodes, but visited nodes still have a higher degree on average than the nodes that would be obtained from an unbiased random sampling.

Algorithm 1 Simple Greedy Algorithm for Minimum p-Median Problem on Incomplete Networks

```
initialize $X \leftarrow \emptyset$
while $|X| < p$ do
    select $u \leftarrow \arg\min_{v \in V' \setminus X} g(X \cup \{v\})$
    $X \leftarrow X \cup \{u\}$
end while
return $X$
```

3) RANDOM SAMPLING (RANDOM)

Random sampling repeatedly obtains a node uniformly at random from all nodes in a network until a specified number of nodes is obtained.

B. ALGORITHMS FOR FINDING THE KEY NODE SET

We use simple heuristic algorithms for both of the MM and IM problems. We apply existing MM and IM algorithms for complete networks to an incomplete network G'. Then, we investigate how the network sampling strategies affect the effectiveness of the existing key node set identification algorithms. This paper focuses on the effectiveness of the algorithms and does not experimentally investigate their computational cost. But note that the key node sets can be obtained from incomplete networks with lower computational cost than from the complete networks because the computational cost of the key node set identification algorithms depends on the size of the networks.

For the MM problem, we use a greedy algorithm for the original MM problem [41]. Since the MM problem is NP-hard, the greedy algorithm is often used for solving the MM problem [41]. Although the objective function $f(X)$ can be calculated using the ground truth network G and demand for all nodes in the original greedy algorithm, in the MM problem for incomplete networks, this information is not available. Therefore, in the MM algorithm for an incomplete subnetwork, the following objective function $g(X)$ is used instead of $f(X)$.

$$g(X) = \sum_{v \in V'} w(v) D(v, X, G') \tag{1}$$

In this, $g(X)$ can be calculated from an incomplete network G' and $w(v)(v \in V')$. Similar to the original greedy algorithm, the heuristic algorithm iteratively adds a node u to the median node set such that $g(\{X \cup u\})$ is minimized. Pseudocode for the algorithm is shown in Algorithm 1.

For the IM problem, we use TIM+ as an efficient approximation algorithm [10]. TIM+ is a state-of-the-art algorithm that achieves efficient computational cost and high effectiveness. We apply TIM+ to the incomplete network G'.

TABLE 1. Characteristics of networks used in the experiments for the MM problem.

	AS	P2P	PowerGrid
Number of nodes	10,670	6,299	4,941
Average degree	4.12	6.60	5.34
Density	0.00039	0.0010	0.0011
Clustering coefficient	0.456	0.015	0.11
Average shortest path length	3.64	4.64	19.0
Graph diameter	10	9	46

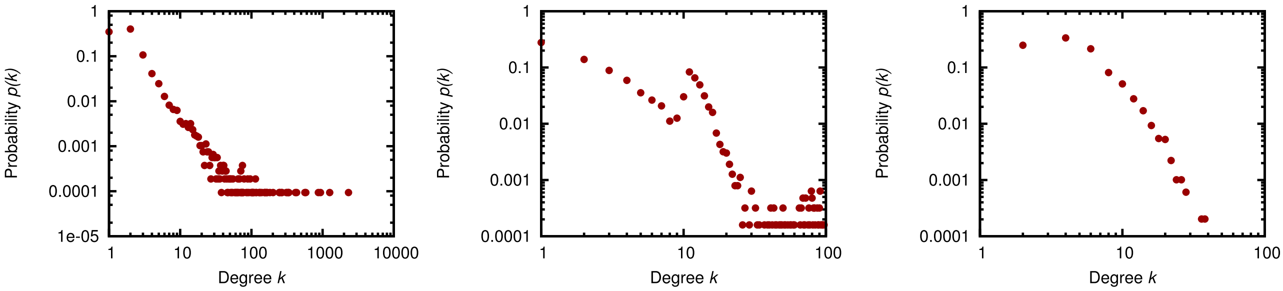
IV. RESULTS OF MINIMUM p-MEDIAN PROBLEM

A. DATASET AND PRELIMINARIES

As the ground truth networks G, we use (1) a network of Autonomous System (AS) [42],[1] (2) a P2P network of Gnutella (P2P) [43] ,[2] and (3) a network of the US powergrid (PowerGrid) [44].[3] Characteristics of each network are shown in Table 1, and the degree distributions for each network are shown in Fig. 1.

We randomly generated the demand of the nodes using a Zipf distribution, a normal distribution, and an exponential distribution. The results in [41] show that the degree of a node and the demand of the node are correlated, and therefore we use the following procedures to generate the demand of each node. We first generate $|V|$ random variables according to the given distribution. We then assign the i-th largest variable as the demand of the node with the i-th highest degree. We next swap the demand of each node with the demand of

[1] http://snap.stanford.edu/data/oregon1.html
[2] http://snap.stanford.edu/data/p2p-Gnutella08.html
[3] https://toreopsahl.com/datasets/#uspowergrid

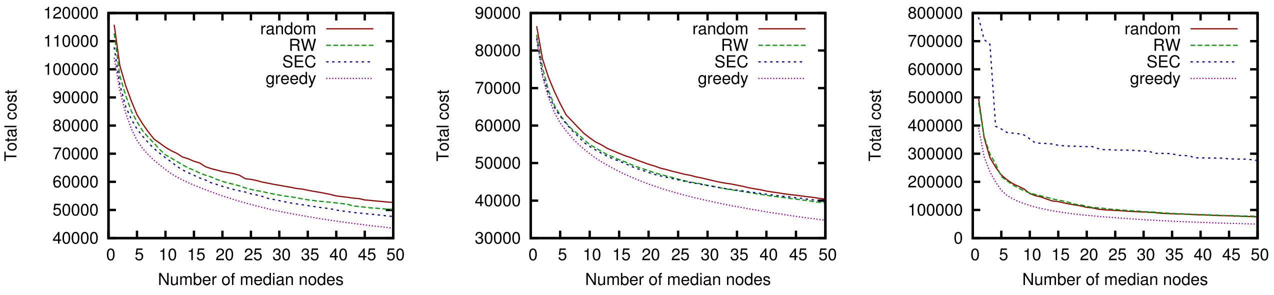


(a) AS

(b) P2P

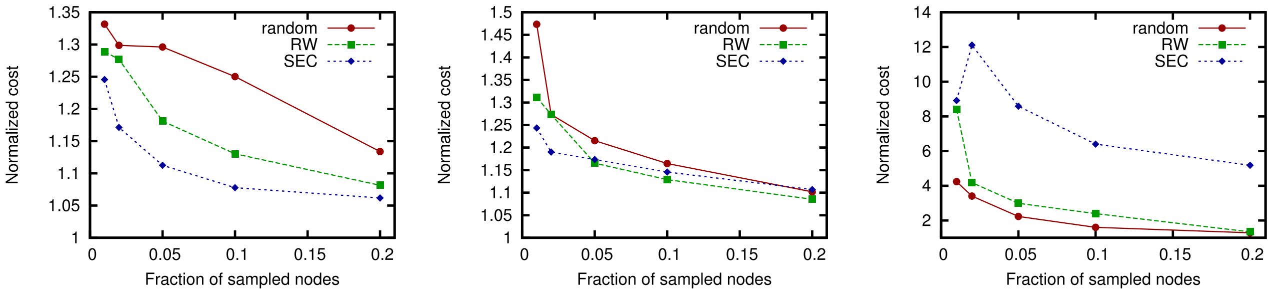
(c) PowerGrid

FIGURE 1. Degree distribution of each network used in the experiments for the MM problem.



(a) AS (b) P2P (c) PowerGrid

FIGURE 2. Total cost vs. number of median nodes (sample size: 10%; distribution of the demand: Zipf; parameter determining the correlation between degree of a node and its demand: $q = 0.9$): Total cost when using incomplete networks obtained with RW is comparable with the cost when using the complete network.



(a) AS (b) P2P (c) Powergrid

FIGURE 3. Normalized cost vs. sample size (distribution of the demand: Zipf; parameter determining the correlation between degree of a node and its demand: $q = 0.9$; number of median nodes: $p = 50$): A 20% sample size achieves a normalized cost of 1.1–1.3 for all three networks when using RW.

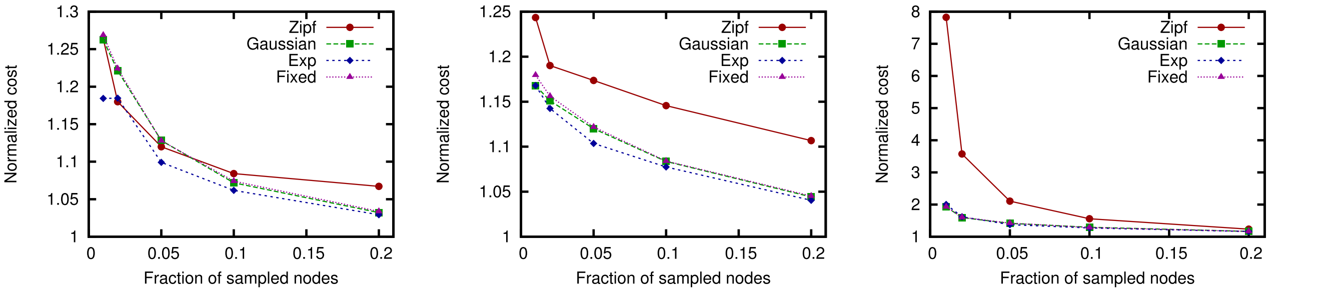
some other randomly selected node with probability $1 - q$ $(0 \leq q \leq 1)$. The parameter q controls the strength of the correlation between node demand and node degree. We used $\gamma = 2$ as the parameter of the Zipf distribution, mean $\mu = 1$ and standard deviation $\sigma = 0.1$ for the normal distribution, and mean $\lambda = 1$ for the exponential distribution.

We apply the algorithm introduced in Section III to the incomplete network G' obtained by SEC, RW, and random sampling, and obtain the set of median nodes X for each. To obtain the median node set X, we assume that the demand $w(v)$ is available for each node $v \in V'$. We then calculate the total cost $f(X)$ for the median node set X while changing the sample size, sampling strategy, and distribution of node demand. We generated the demands and obtained sample subnetworks for each parameter setting 20 times. The results shown from here are averaged over the 20 simulation runs for each configuration.

B. RESULTS

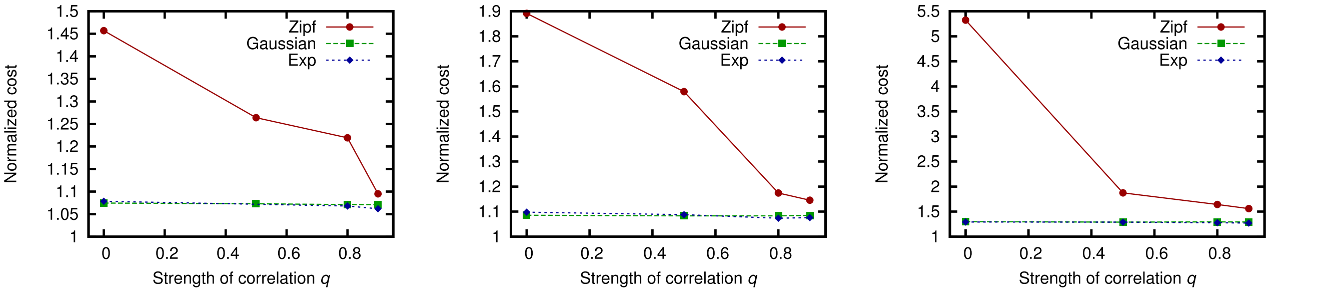
We first fixed the sample size as 10% of nodes ($M = 0.1|V|$), and investigated the total cost while changing the number of median nodes (Fig. 2). The distribution of node demand is the Zipf distribution. As the parameter of correlation between degree and demand, we used $q = 0.9$. The results when using the complete network and the demand of all nodes are included in the figures (denoted as greedy). These results show that the total cost when selecting median nodes from incomplete networks is comparable with the cost when using the complete network. When using the incomplete network obtained via SEC, the increase in cost relative to the cost for the complete network is only 10% for AS and 15% for P2P. In contrast, for PowerGrid, the cost when using the incomplete network obtained via SEC is significantly higher than the cost when using other sampling strategies. As shown in Fig. 1, there are no *strong hubs* that have significantly high degree in the PowerGrid, and therefore the benefit of finding high-degree nodes is smaller in PowerGrid than in AS and P2P. In the MM problem, selecting median nodes that are far from each other is generally desirable to achieve lower cost, but SEC typically traverses the network only near the starting node. This drawback of SEC also detrimentally affects the cost when applying SEC to PowerGrid for subnetwork selection.

We next investigate the relation between the sample size and the total cost. Fig. 3 shows the normalized cost when selecting 50 median nodes from the incomplete networks, compared against the sample size (characterized as the fraction of sampled nodes). The normalized cost is defined as the cost when selecting 50 median nodes from the incomplete networks divided by the cost when selecting 50 median nodes



(a) AS (SEC) (b) P2P (SEC) (c) PowerGrid (RW)

FIGURE 4. Comparison among different distributions of demand (the parameter determining the correlation between degree of a node and its demand: $q = 0.9$; number of median nodes: $p = 50$): The total cost when the node demand follows a Zipf distribution is higher than the cost when the node demand follows other distributions.



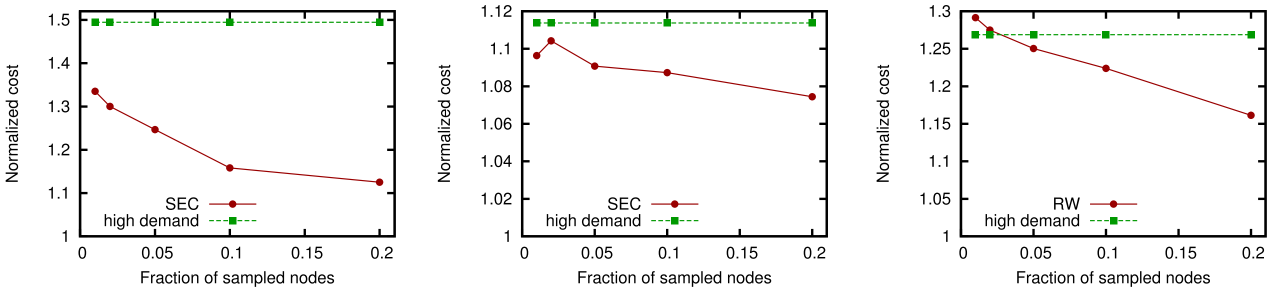
(a) AS (SEC) (b) P2P (SEC) (c) PowerGrid (RW)

FIGURE 5. Normalized cost vs. the parameter determining the correlation between degree of a node and its demand q (sample size: 10%; number of median nodes: $p = 50$): When the node demand follows the Zipf distribution, the total cost increases as the correlation between degree and demand of node decreases.

from the complete networks. These results show that when the incomplete network is obtained via RW, a 20% sample size achieves a normalized cost of 1.1–1.3 for all three networks. For AS and P2P, a 10% sample size is large enough to achieve a normalized cost of 1.2. This result suggests that when the subnetwork is obtained via RW, reasonable solutions for the MM problem can be obtained from only limited observations of the networks. The cost when using SEC is lower than that when using RW for the AS and P2P networks, but it is much higher than the cost when using RW for PowerGrid. These results suggest that when we crawl completely unknown networks to determine the median nodes, using RW and collecting 10–20% of samples is a good approach.

We next investigate the effects of node demand on the total cost. Fig. 4 shows the normalized cost for each demand distribution. For comparison purposes, the results when the demands of all nodes are fixed to 1 (denoted as Fixed) are also shown. Here, we use SEC for the AS and P2P networks, and RW for the PowerGrid network. The number of median nodes is fixed to 50. The results show that the total cost when the node demand follows a Zipf distribution, which has a heavy-tailed distribution, is higher than the cost when the node demand follows other distributions. Moreover, we also change the parameter q that controls the correlation between node degree and the node demand (Fig. 5). Here, the fraction of sampled nodes is 0.1, and the number of median nodes is 50. From these results, when the node demand follows the Zipf distribution, the total cost increases as the correlation between degree and demand of node decreases. When the correlation between node degree and node demand is low, low-degree nodes tend to have higher demand than when the correlation is strong. Low-degree nodes are more difficult to discover by sampling strategies than high-degree nodes are. As a consequence, high-demand and low-degree nodes are likely to be unknown when selecting the median nodes in the low-correlation scenario. This is why the correlation affects the total cost. We can also find that there is little difference between exponential and normal distributions. This could be explained by the fact that both distributions have an exponential tail, which implies that there are no extremely high-demand nodes. From this observation, the MM problem for incomplete networks requires a larger sample when the correlation between degree and demand is very low and the demand distribution is heavy-tailed.

We now tackle the problem of finding median nodes when the demand of *all* nodes is available but the topology is only partially known. To do this, we examine the benefit of

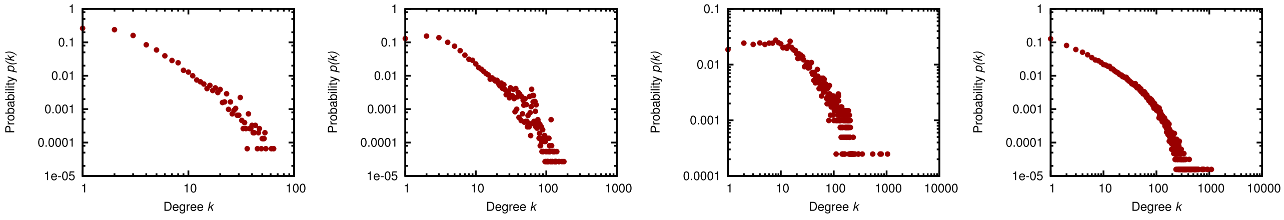


(a) AS (b) P2P (c) PowerGrid

FIGURE 6. **Total cost when the demand of all nodes is available (distribution of the demand: Zipf; the parameter determining the correlation between degree of a node and its demand: $q = 0$; the number of median nodes: $p = 50$): Using incomplete networks achieves a lower total cost than the high demand heuristic.**

TABLE 2. **Characteristics of each network used in the experiments for the IM problem.**

	NetHEPT	NetPHY	Facebook-small	Facebook-large
Number of nodes	15,233	37,154	4,039	63731
Average degree	4.23	9.74	43.7	48.5
Density	0.00028	0.00026	0.011	0.00076
Clustering coefficient	0.677	0.87	0.62	0.25
Average shortest path length	5.84	6.26	3.69	4.3
Graph diameter	22	19	8	15



(a) NetHEPT (b) NetPHY (c) Facebook-small (d) Facebook-large

FIGURE 7. **Degree distributions of networks used in experiments for IM problem.**

knowledge about the network topology for the MM problem. We use $h(X)$, defined as follows, as the objective function of the greedy algorithm instead of $g(X)$:

$$h(X) = \sum_{v \in V} w(v) D(v, X, G'). \qquad (2)$$

To calculate $h(X)$ for node $v \notin V'$, we let $D(v, X, G') = d_{max} + 1$, where d_{max} is the diameter of G'.

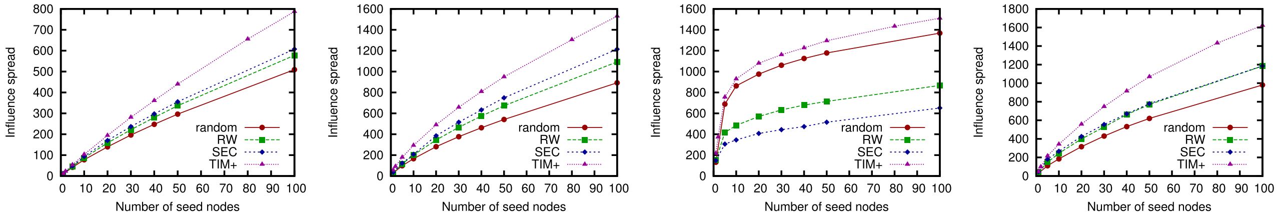
Fig. 6 shows the normalized total cost when the node demand follows the Zipf distribution, with $q = 0$ for the situation where the demand of all nodes is available. For comparison purposes, the results when selecting the 50 nodes having the highest demand as the median nodes are included in the figures (denoted as high demand). These results show that if node demand is available, a good solution can be obtained when there is no correlation between degree and demand. Comparing the results of high demand with the results when using incomplete networks shows that network topology is useful for achieving a lower total cost.

V. RESULTS OF INFLUENCE MAXIMIZATION PROBLEM

A. DATASET AND PRELIMINARIES

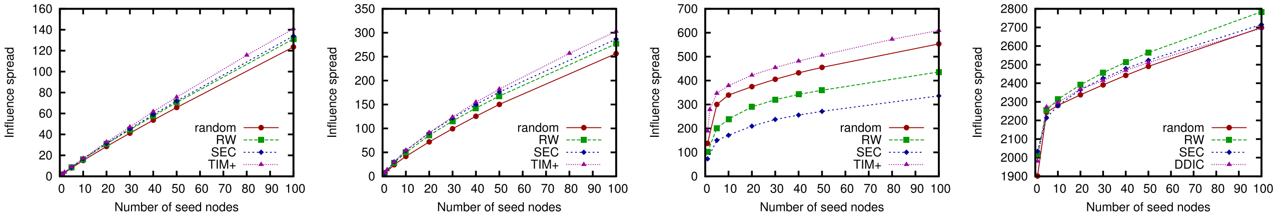
As the ground truth network G, we use four real social networks: NetHEPT [12], NetPHY [12], Facebook-small [45], and Facebook-large [46]. NetHEPT and NetPHY represent co-authorship among researchers, and Facebook-small and Facebook-large represent friendships among Facebook users. These are widely used as benchmark datasets for IM problems [8], [10]–[12], [47]–[51]. Multiple links are simply converted to a single link [24], [52]. Characteristics of each network are shown in Table 2, and the degree distribution of each network is shown in Fig. 7.

We synthetically generated the influence-spread probabilities of each link, using the weighted cascade (WC)



(a) NetHEPT (b) NetPHY (c) Facebook-small (d) Facebook-large

FIGURE 8. Influence spread vs. number of seed nodes (WC model): Influence spread when selecting seed nodes from incomplete networks is comparable with influence spread when using the complete network.



(a) NetHEPT (b) NetPHY (c) Facebook-small (d) Facebook-large

FIGURE 9. Influence spread vs. number of seed nodes (influence-spread probability $p = 0.01$): Influence spread when selecting seed nodes from incomplete networks is comparable with influence spread when using the complete network.

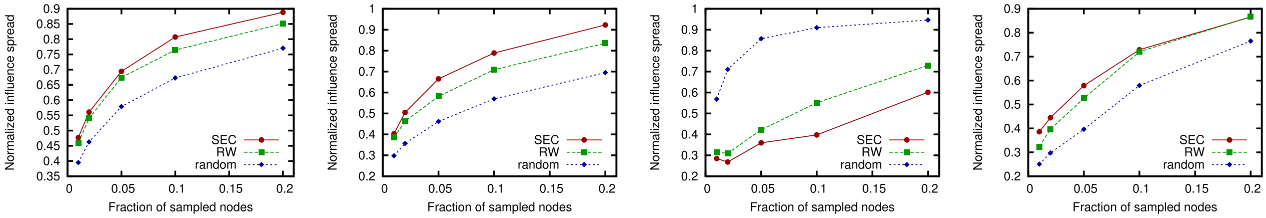
model [12]. Specifically, for each link (u, v), we let $p_{u,v} = 1/d_v$, where d_v is the in-degree of node v. The WC model is widely used for generating influence-spread probabilities for the evaluation of IM algorithms [10]–[12], [47], [53]. We also used $p_{u,v} = 0.01$ for all node pairs (u, v) for comparison.

We apply the algorithm introduced in Section III to incomplete networks G' obtained via SEC, RW, and random sampling and obtain the set of seed nodes U. As a parameter for TIM+ we used $\varepsilon = 0.1$. For the obtained seed node set U, we calculate the influence spread $\sigma(U, G, W)$ through simulation of the IC model on the ground-truth network G. We run each simulation 1,000 times and take the average of the influence spread.

B. RESULTS

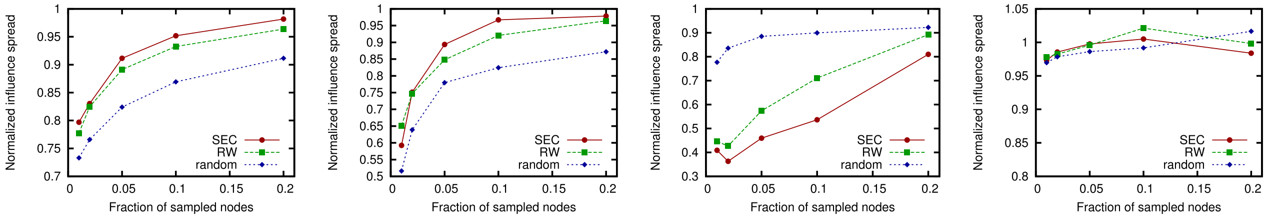
Similar to the approach in the previous section, we fixed the sample size as 10% of nodes ($M = 0.1|V|$) and investigated the influence spread while changing the number of seed nodes. Figs. 8 and 9 show the results when using, respectively, the WC model and the uniform ($p = 0.01$) model for influence-spread probabilities. The results when applying TIM+ to the complete network are included in the figures (denoted as TIM+). For Facebook-large with $p = 0.01$, we failed to obtain the results from TIM+ due to the scalability limits already reported in [54]. We therefore added the results when using degree discount IC [12] (denoted as DDIC), which is a lightweight heuristic algorithm, instead of the results of TIM+ for Facebook-large with $p = 0.01$. The obtained results show that influence spread when selecting seed nodes from incomplete networks is comparable with influence spread when using the complete network. Except for Facebook-small, using incomplete networks obtained via RW and SEC achieved higher influence spread than using incomplete networks obtained via random sampling. This result confirms the benefit of crawl-based sampling strategies for finding a key node set, as also seen in the results for the MM problem. For Facebook-small, using an incomplete networks obtained via random sampling achieved higher influence spread than using networks obtained via RW and SEC. In particular, when using SEC, the influence spread is much smaller than when using other strategies. As shown in Table 2, Facebook-small has a higher density than the other networks. The benefit of finding hub nodes is low in dense networks since influence can be easily spread, even from low-degree nodes. In contrast, the drawback of high locality with SEC can affect the selection of good seed nodes.

We next investigate the effects of sample size. Figs. 10 and 11 show the normalized influence spread when selecting 50 seed nodes from the incomplete networks. The normalized influence spread is defined as the influence spread when selecting 50 seed nodes from the incomplete networks divided by the influence spread when selecting 50 seed nodes from the complete network. For Facebook-large with $p = 0.01$, seed nodes in the complete network were selected using DDIC; for other settings, seed nodes in the complete network were selected using TIM+. Fig. 10 shows the results for the WC model, and Fig. 11 shows the results for $p = 0.01$. From these results, a normalized influence spread of 0.8–0.9 can be achieved by using only a small sample size of 10–20%.



(a) NetHEPT (b) NetPHY (c) Facebook-small (d) Facebook-large

FIGURE 10. Normalized influence spread vs. sample size (WC model; number of seed nodes: 50): A normalized influence spread of 0.8–0.9 can be achieved by using a sample size of 10–20%.



(a) NetHEPT (b) NetPHY (c) Facebook-small (d) Facebook-large

FIGURE 11. Normalized influence spread vs. sample size (influence spread probability: $p = 0.01$; number of seed nodes: 50): A normalized influence spread of 0.8–0.9 can be achieved by using a sample size of 10–20%.

These results suggest that we can obtain a set of influential seed nodes from a small sample. They also suggest that the crawl-based sampling strategies RW and SEC are effective for obtaining the partial structure of a network when identifying an influential node set. Only for Facebook-small was random sampling more effective than SEC and RW. However, when the sample size was 20%, RW achieved a normalized influence spread of 0.7 for the WC model and approximately 0.9 for $p = 0.01$. This suggests that when RW is used and 20% of nodes are sampled, sufficient influence spread can be achieved, and in many cases a smaller sample is sufficient.

VI. DISCUSSION

An important implication of our results is that the partial structure of a network obtained via crawl-based network sampling is sufficient for identifying key node sets. Our results suggest that a 10% sample size is enough in many cases. This is a good result for real applications since access to real networks is typically limited (e.g., by restrictions on the number of API calls for social media graphs).

Another implication is that using RW sampling with a moderate level of bias is a robust strategy when the ground truth network is completely unknown. SEC sampling is designed to find high-degree nodes, and the bias of visiting high-degree nodes in SEC is stronger than in RW sampling. Our results show that SEC outperformed RW in several networks. However, SEC was not robust, and sometimes performed poorly for several types of networks. For instance, for the PowerGrid network in the MM problem, and for the Facebook-small network in the IM problem, using the SEC strategy resulted in considerably worse results than using other strategies. When using the SEC strategy, the crawling process tends to be confined to a strongly clustered subnetwork, which sometimes results in very poor outcomes. These results suggest that a strong bias in SEC degrades its robustness. Therefore, when we do not have any knowledge about the ground truth network, RW sampling should be used rather than SEC sampling.

We recognize some limitations of this study and suggest them as future research directions. First, the generalizability of our findings to other types of key node set identification problems is still not clear. There are several types of key node set identification problems, such as several variants of the IM problem and the minimum p-center problem. Our results show that similar heuristics (i.e., applying algorithms for the complete network to the incomplete network) can be effective for both IM and MM problems on incomplete networks. We therefore expect that other types of key node set identification problems for incomplete networks can be solved with a similar approach. We are interested in validating our results against other types of key node set identification problems in future research. We are also interested in key node set identification problems for other types of networks such as dynamic networks and multi-layer networks. Second, our study is purely experimental, and theoretical verification of our results is also an important direction for future research. In particular, it would be worthwhile to derive the sample sizes necessary to achieve specific targets for practitioners using key node set identification algorithms. Third, we only consider some combinations of existing sampling strategies and existing key node set identification algorithms, and there is a room for development of more effective algorithms. We are interested in exploring how to choose pairs of sampling strategies and key node set identification algorithms for more robust key node set identification. Using more simple key node set identification algorithms such as using centrality measures of nodes would be a option for more robust key node set identification. Moreover, the sampling strategy could be adaptively changed during sampling. Such new adaptive sampling strategy could be effective for the key node set identification.

VII. CONCLUSION

We studied two variants of the problem of identifying a set of key nodes from a network under limited knowledge about its structure. Specifically, we investigated how conventional network sampling strategies affect the solutions obtained for the MM and IM problems, which are popular key node set identification problems, when only partial networks are known. The conventional crawl-based sampling strategies are known to have a bias toward high-degree nodes. Although these biases are not desirable for estimating the general characteristics of unknown networks, they are expected to be beneficial for identifying key node sets. Our results have shown the benefits of biases in crawl-based sampling strategies for the IM and MM problems. We showed that both the IM and MM problems are effectively solved by similar simple heuristic algorithms when the subnetworks were obtained by crawl-based sampling strategies. For many cases, we showed that a 10-20% sample size is enough to find key node sets that are comparable with the key node sets obtained from the complete networks. We also examined which sampling strategy should be used for identifying the key node sets. Our results suggest that using RW sampling is a good option. SEC sampling is sometimes better than RW but is sometimes considerably worse.

REFERENCES

[1] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," *Phys. Rep.*, vol. 650, pp. 1–63, Sep. 2016.

[2] S. P. Borgatti, "Identifying sets of key players in a social network," *Comput. Math. Org. Theory*, vol. 12, no. 1, pp. 21–34, Apr. 2006.

[3] B. C. Tansel, R. L. Francis, and T. J. Lowe, "State of the art—Location on networks: A survey. Part I: The p-center and p-median problems," *Manage. Sci.*, vol. 29, no. 4, pp. 482–497, Apr. 1983.

[4] Y. Li, J. Fan, Y. Wang, and K.-L. Tan, "Influence maximization on social graphs: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 10, pp. 1852–1872, Oct. 2018.

[5] H. Yu, X. Cao, Z. Liu, and Y. Li, "Identifying key nodes based on improved structural holes in complex networks," *Phys. A, Stat. Mech. Appl.*, vol. 486, pp. 318–327, Nov. 2017.

[6] F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," *Nature*, vol. 524, no. 7563, pp. 65–68, Aug. 2015.

[7] N. Mladenović, J. Brimberg, P. Hansen, and J. A. Moreno-Pérez, "The p-median problem: A survey of metaheuristic approaches," *Eur. J. Oper. Res.*, vol. 179, no. 3, pp. 927–939, Jun. 2007.

[8] 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 (KDD)*, 2003, pp. 137–146.

[9] L. Qiu, W. Jia, J. Yu, X. Fan, and W. Gao, "PHG: A three-phase algorithm for influence maximization based on community structure," *IEEE Access*, vol. 7, pp. 62511–62522, 2019.

[10] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, Jun. 2014, pp. 75–86.

[11] A. Goyal, W. Lu, and L. V. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web*, Mar. 2011, pp. 47–48

[12] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, New York, NY, USA, Jun. 2009, pp. 199–208.

[13] Y. Tang, Y. Shi, and X. Xiao, "Influence maximization in near-linear time: A martingale approach," in *Proc. ACM SIGMOD Int. Conf. Manage. Data (SIGMOD)*, May 2015, pp. 1539–1554.

[14] H. Wu, J. Shang, S. Zhou, Y. Feng, B. Qiang, and W. Xie, "LAIM: A linear time iterative approach for efficient influence maximization in large-scale networks," *IEEE Access*, vol. 6, pp. 44221–44234, 2018.

[15] J. Zhu, Y. Liu, and X. Yin, "A new Structure-Hole-Based algorithm for influence maximization in large online social networks," *IEEE Access*, vol. 5, pp. 23405–23412, 2017.

[16] M. Riondato and E. M. Kornaropoulos, "Fast approximation of betweenness centrality through sampling," *Data Mining Knowl. Discovery*, vol. 30, no. 2, pp. 438–475, Mar. 2016.

[17] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," *Nature Phys.*, vol. 6, no. 11, pp. 888–893, Nov. 2010.

[18] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Phys. A, Statist. Mech. Appl.*, vol. 391, pp. 1777–1787, Feb. 2012.

[19] Q. Li, T. Zhou, L. Lü, and D. Chen, "Identifying influential spreaders by weighted LeaderRank," *Phys. A, Stat. Mech. Appl.*, vol. 404, pp. 47–55, Jun. 2014.

[20] S. Tsugawa, "A survey of social network analysis techniques and their applications to socially aware networking," *IEICE Trans. Commun.*, vol. 102, no. 1, pp. 17–39, Jan. 2019.

[21] S. Tsugawa and K. Kimura, "Identifying influencers from sampled social networks," *Phys. A, Stat. Mech. Appl.*, vol. 507, pp. 294–303, Oct. 2018.

[22] A. S. Maiya and T. Y. Berger-Wolf, "Benefits of bias: Towards better characterization of network sampling," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Sep. 2011, pp. 105–113.

[23] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of OSNs," in *Proc. IEEE INFOCOM*, Mar. 2010, pp. 1–9.

[24] S. Mihara, S. Tsugawa, and H. Ohsaki, "Influence maximization problem for unknown social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2015, pp. 1539–1546.

[25] Y. Murase, H.-H. Jo, J. Török, J. Kertész, and K. Kaski, "Sampling networks by nodal attributes," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 99, no. 5, May 2019, Art. no. 052304.

[26] Y. Xie, S. Chang, Z. Zhang, M. Zhang, and L. Yang, "Efficient sampling of complex network with modified random walk strategies," *Phys. A, Stat. Mech. Appl.*, vol. 492, pp. 57–64, Feb. 2018.

[27] N. Blagus, L. Šubelj, and M. Bajec, "Empirical comparison of network sampling: How to choose the most appropriate method?" *Phys. A, Stat. Mech. Appl.*, vol. 477, pp. 136–148, Jul. 2017.

[28] A. Rezvanian and M. R. Meybodi, "Sampling social networks using shortest paths," *Phys. A, Stat. Mech. Appl.*, vol. 424, pp. 254–268, Apr. 2015.

[29] X.-K. Xu and J. J. H. Zhu, "Flexible sampling large-scale social networks by self-adjustable random walk," *Phys. A, Stat. Mech. Appl.*, vol. 463, pp. 356–365, Dec. 2016.

[30] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Practical recommendations on crawling online social networks," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 9, pp. 1872–1892, Oct. 2011.

[31] R. Oliveira, D. Pei, W. Willinger, B. Zhang, and L. Zhang, "The (In)completeness of the observed Internet AS-level structure," *IEEE/ACM Trans. Netw.*, vol. 18, no. 1, pp. 109–122, Feb. 2010.

[32] B. Donnet and T. Friedman, "Internet topology discovery: A survey," *IEEE Commun. Surveys Tuts.*, vol. 9, no. 4, pp. 56–69, 4th Quart., 2007.

[33] F. Chiericetti, A. Dasgupta, R. Kumar, S. Lattanzi, and T. Sarlós, "On sampling nodes in a network," in *Proc. 25th Int. Conf. World Wide Web (WWW)*, 2016, pp. 471–481.

[34] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," *Comput. Netw.*, vol. 33, nos. 1–6, pp. 309–320, 2000.

[35] A. H. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, and D. Stutzbach, "Respondent-driven sampling for characterizing unstructured overlays," in *Proc. IEEE 28th Conf. Comput. Commun. (INFOCOM)*, Apr. 2009, pp. 2701–2705.

[36] L. D. F. Costa and G. Travieso, "Exploring complex networks through random walks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 75, no. 1, Jan. 2007, Art. no. 016102.

[37] M. Kurant, A. Markopoulou, and P. Thiran, "Towards unbiased BFS sampling," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 9, pp. 1799–1809, Oct. 2011.

[38] S. Mihara, S. Tsugawa, and H. Ohsaki, "On the effectiveness of random jumps in an influence maximization algorithm for unknown graphs," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, 2017, pp. 395–400.

[39] B. Wilder, N. Immorlica, E. Rice, and M. Tambe, "Maximizing influence in an unknown social network," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 4743–4750.

[40] S. Stein, S. Eshghi, S. Maghsudi, L. Tassiulas, R. K. Bellamy, and N. R. Jennings, "Heuristic algorithms for influence maximization in partially observable social networks," in *Proc. 3rd Int. Workshop Social Influence Anal.*, 2017, pp. 20–32.

[41] Y. Li and M. T. Liu, "Optimization of performance gain in content distribution networks with server replicas," in *Proc. Symp. Appl. Internet*, 2003, pp. 182–189.

[42] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proc. 11th SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Apr. 2005, pp. 177–187.

[43] R. Matei, A. Iamnitchi, and P. Foster, "Mapping the Gnutella network," *IEEE Internet Comput.*, vol. 6, no. 1, pp. 50–57, Jan. 2002.

[44] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998.

[45] J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, May 2012, pp. 539–547.

[46] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi, "On the evolution of user interaction in facebook," in *Proc. 2nd ACM Workshop Online Social Netw. (WOSN)*, 2009, pp. 37–42.

[47] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Jul. 2010, pp. 1029–1038.

[48] X. Liu, M. Li, S. Li, S. Peng, X. Liao, and X. Lu, "IMGPU: GPU-accelerated influence maximization in large-scale social networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 136–145, Jan. 2014.

[49] H. Lamba and R. Narayanam, "A novel and model independent approach for efficient influence maximization in social networks," in *Proc. 14th Int. Conf. Web Inf. Syst. Eng. (WISE)*, Oct. 2013, pp. 73–87.

[50] H. Zhang, T. N. Dinh, and M. T. Thai, "Maximizing the spread of positive influence in online social networks," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst.*, Jul. 2013, pp. 317–326.

[51] Q. Liu, B. Xiang, E. Chen, H. Xiong, F. Tang, and Y. X. Jeffrey, "Influence maximization over large-scale social networks: A bounded linear approach," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, Nov. 2014, pp. 171–180.

[52] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, "Influence maximization in dynamic social networks," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 1313–1318.

[53] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 918–923.

[54] A. Arora, S. Galhotra, and S. Ranu, "Debunking the myths of influence maximization: An in-depth benchmarking study," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2017, pp. 651–666.

SHO TSUGAWA (Member, IEEE) received the M.E. and Ph.D. degrees from Osaka University, Japan, in 2009 and 2012, respectively. He is currently an Assistant Professor with the Faculty of Engineering, Information and Systems, University of Tsukuba, Japan. His research interests include network science, social network analysis, and computational social science. He is a member of ACM, IEICE, and IPSJ.

HIROYUKI OHSAKI (Member, IEEE) received the M.E. degree in information and computer sciences and the Ph.D. degree from Osaka University, Japan, in 1995 and 1997, respectively. He is currently a Professor with the Department of Informatics, School of Science and Technology, Kwansei Gakuin University, Japan. His research interests include design, modeling, and control of large-scale communication networks. He is a member of IEICE and IPSJ.