Accepted Manuscript

Identifying influencers from sampled social networks

Sho Tsugawa, Kazuma Kimura

PII: S0378-4371(18)30629-0
DOI: https://doi.org/10.1016/j.physa.2018.05.105
Reference: PHYSA 19645

To appear in: *Physica A*

Received date : 6 July 2017
Revised date : 23 March 2018

Please cite this article as: S. Tsugawa, K. Kimura, Identifying influencers from sampled social networks, *Physica A* (2018), https://doi.org/10.1016/j.physa.2018.05.105

*Highlights (for review)

- We investigate the effects of node sampling on the influencer identification.
- The negative effect of biased sampling on the identification of influencers is small.
- A small sample size is enough for identifying influencers in social media networks.
- For some networks, node sampling is beneficial in the identification of influencers.

Identifying Influencers from Sampled Social Networks[☆]

Sho Tsugawa[a,*], Kazuma Kimura[a]

[a]*Graduate School of Systems and Information Engineering, University of Tsukuba, 1–1–1 Tennodai, Tsukuba, Ibaraki 305–8573, Japan*

Abstract

Identifying influencers who can spread information to many other individuals from a social network is a fundamental research task in the network science research field. Several measures for identifying influencers have been proposed, and the effectiveness of these influence measures has been evaluated for the case where the complete social network structure is known. However, it is difficult in practice to obtain the complete structure of a social network because of missing data, false data, or node/link sampling from the social network. In this paper, we investigate the effects of node sampling from a social network on the effectiveness of influence measures at identifying influencers. Our experimental results show that the negative effect of biased sampling, such as sample edge count, on the identification of influencers is generally small. For social media networks, we can identify influencers whose influence is comparable with that of those identified from the complete social networks by sampling only 10–30% of the networks. Moreover, our results also suggest the possible benefit of network sampling in the identification of influencers. Our results show that, for some networks, nodes with higher influence can be discovered from sampled social networks than from complete social networks.

Keywords: social networks, social media, node ranking, influence, network sampling

[☆]An earlier version of this paper was presented at the workshop on social influence 2016 [1].

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

1. Introduction

Identifying influencers who can spread information to many other individuals from a social network is a fundamental research task in the network science research field [2, 3, 4, 5, 6, 7]. Widely spread information in a social network can affect public opinion, brand awareness, and product market share in society [8, 9]. In the information-spreading process, a few individuals (called influencers) are considered to play an important role in triggering a large-scale cascade of information diffusion [3, 10]. Therefore, the identification of influencers has attracted great interest from many researchers [2, 3, 4, 5, 8, 11, 12]. A typical application of the identification of influencers is viral marketing [5]. If influencers are known in viral marketing, advertisements can reach many people at low cost. Moreover, identifying influencers is expected to be useful for preventing the spread of rumors and misinformation [13].

Several measures have been proposed for estimating the influence of each node when identifying influencers in a social network [2, 3, 4, 6, 11, 12, 14, 15, 16, 7]. An excellent survey on such measures can be found in [7]. Such measures include centrality (degree centrality, closeness centrality, and betweenness centrality) [14], PageRank [15], and the k-core index [16]. We refer to these measures as *influence measures* in this paper.

Existing studies evaluate the effectiveness of influence measures when the social network structure is completely known [2, 3, 6, 11, 12, 7]. In these studies, the influence measures of each node are obtained from a complete social network, and nodes that are highly ranked with respect to the measures (e.g., the top 1% of nodes) are extracted as influencers. Then, the influence of the extracted nodes (i.e., influencers) is evaluated by using models of influence cascades, such as the susceptible–infected–removed (SIR) and susceptible–infected–susceptible (SIS) models [2, 6, 11, 7] or by using actual records of information cascades [3, 12].

However, it is difficult in practice to obtain the complete structure of a real social network [17, 18, 19]. Social networks typically contain several errors due to missing data, false data, or node/link sampling from the social network [17, 18, 19]. Such incompleteness in social networks should affect the effectiveness of influence measures for identifying influencers.

This paper focuses on social networks that are incomplete due to node sampling [19, 20, 21] and investigates the effects of node sampling on influence measures for identifying influencers. A promising area of application for identifying influencers is social media, such as Twitter and Facebook [8]. Because the social networks in social media are huge, their structures are typically estimated by node

sampling [21, 22, 23]. Therefore, investigating the effects of node sampling on influence measures is important when using these measures for identifying influencers in social media. We apply several node sampling strategies to social networks and identify influencers from the sampled social networks using influence measures. We then investigate the *actual* influence of the identified influencers. As in the work of Pei et al. [3], the actual influence of a node is empirically measured by the size of the information cascades triggered by the node. We obtain the sizes of these information cascades from records of information cascades, such as records of retweets on Twitter.

Our main contributions are as follows.

- We empirically investigate the effects of node sampling on the identification of influencers. While existing studies have investigated the effects of node sampling on the stability of influence measures [3, 19, 24], we investigate the effects of node sampling on the effectiveness of influence measures for identifying influencers.

- We show the necessary sample sizes for identifying influencers. We demonstrate that when using biased sampling strategies such as sample edge count [20], a small sample size (e.g., a sample size of 10-30%) is enough for identifying influencers in social media networks.

- We propose a possible benefit of node sampling in the identification of influencers. Our results suggest that, for some networks, nodes with higher influence can be discovered from sampled social networks than from complete social networks.

The remainder of this paper is organized as follows. In Section 2, we introduce existing studies related to using social network structures to identify influencers. Section 3 explains the research methodology. Section 4 shows the results. Section 5 discusses the implications of the results and limitations of the work. Finally, Section 6 contains our conclusions.

2. Related Work

The effectiveness of influence measures has been extensively evaluated through simulation using influence cascade models [2, 4, 6, 11, 25, 7]. In simulations, the actual influence of a node is defined as the number of nodes affected by an influence cascade originating from the node of interest. For example, Chen et al. [11]

evaluate the effectiveness of popular influence measures by comparing the actual influence (as obtained from simulation using the SIR model) with the influence estimated from the measures.

While most existing studies use influence cascade models for evaluating the effectiveness of influence measures, Pei et al. [3] adopt an empirical approach, using the actual records of information diffusion for the evaluation and defining the actual influence of a node as the number of nodes who repost the node's posts. Following Pei et al., we use the actual records of an information cascade for evaluating influence measures. As discussed in [3], the empirical validation of influence measures is important because model-based information diffusion and actual diffusion are suggested to be different. Previous studies have evaluated the effectiveness of influence measures when the complete structure of a social network is known. In contrast, we evaluate the effectiveness of influence measures when the structure of a social network is known only from node sampling.

Several studies have investigated the robustness of influence measures against incompleteness [17, 19, 26, 27, 28, 29], and these are closely related to our work. The robustness of influence measures against random errors has also been extensively studied [17, 26, 28]. Borgatti et al. [17] and Frantz et al. [26] have investigated the consistency of node rankings between a ground-truth network and a network with errors. The robustness of influence measures against node sampling has been also investigated [3, 19, 24]. Salamanos et al. [24] have addressed the problem of finding the top-k central nodes in a ground-truth social network from a sampled social network. These studies address the problem of identifying the top-ranked nodes on the basis of influence measures in a ground-truth social network from an incomplete social network. In contrast, the problem that we address is identifying actually influential nodes from an incomplete social network. Here, the actual influence of a node is empirically measured by the size of the information cascade triggered by the node.

3. Methodology

3.1. Overview

We investigate how accurately we can identify influencers from sampled social networks. First, by sampling a fraction of nodes from a complete social network G, we obtain an incomplete social network G'. Second, we calculate the influence measures of each node in G'. We use popular influence measures: degree centrality, closeness centrality, betweenness centrality [14], PageRank [15], and k-core index [16]. We also use recently proposed methodologies for identifying

Table 1: Statistics for examined datasets

Dataset	Num. of nodes	Num. of links	Num. of information cascades
Twitter-follow	50,000	331,270	214,532
Twitter-mention	3,907,682	5,399,949	1,000,221
Facebook	63,731	1,545,685	838,092
APS	247,675	856,864	45,684,601

multiple influencers: collective influence (CI) [30] and VoteRank [31]. The effectiveness of CI is shown through experiments using information diffusion logs [32] and diffusion models [33]. The effectiveness of VoteRank is also shown through experiments using diffusion models [31]. Third, we rank the nodes in descending order of their influence measures and extract the top $p\%$ nodes in the ranking as the influencers. Finally, we evaluate the actual influence of the extracted influencers. Following Pei et al. [3], the actual influence of a node is defined as the node's ability to spread information to other nodes. More precise definitions of the actual influence for each dataset will be given in the next subsection.

3.2. Datasets

We use four types of datasets: one is our collected dataset *Twitter-follow*, and the remaining three are publicly available datasets, *Twitter-mention* [1] [2], *Facebook* [34] [2], and *APS* [2]. Several statistics of these datasets are shown in Tab. 1.

The details of each dataset are described below.

Twitter-follow This dataset contains a social network that represents the follow relations of Twitter users and the records of retweets by them during a specific period, described below. In the social network, each Twitter user is represented as a node, and a follow relation from user i to user j is represented as a directed link from node i to node j. As a measure for quantifying the actual influence of user i, we used the number of users who have retweeted a tweet by user i at least once. If user j retweets user i's tweet, we consider the information in that tweet to have spread from user i to user j, and the number of users retweeting each user's tweet is used to characterize user influence. This dataset was collected through the following process using, the Twitter Application Programming Interface (API).

[1] http://trec.nist.gov/data/tweets/

[2] http://www-levich.engr.ccny.cuny.edu/webpage/hmakse/software-and-data/

1. We randomly selected 50,000 users who frequently retweeted posts from users meeting the following conditions:
   - Users who retweeted 10 or more tweets and whose number of retweets was between 10 and 100 during the period of December 11 to 17, 2013.
   - Users who posted tweets whose number of retweets was between 50 and 100 during the period of December 11 to 17, 2013.
2. We collected the follow relations among these 50,000 users during the period of December 16 to 19, 2013.
3. We also collected the retweets of tweets that were posted by these 50,000 users during the period of December 18, 2013, to January 31, 2014.

Twitter-mention This dataset contains a social network representing the mention relations of users on Twitter and the records of retweets posted by them during the period of January 23 to February 8, 2011. In this network, each Twitter user is represented as a node, and a mention from user i to user j is represented as a directed link from node i to node j. As a measure for quantifying the actual influence of user i, we used the number of users who have retweeted at least one of user i's tweets.

Facebook This dataset contains a social network representing user friendships on Facebook and the records of posts by them during the period of September 25, 2006, to January 22, 2009. In this network, each Facebook user is represented as a node, and a friendship between user i and user j is represented as an undirected link between node i and node j. As a measure for quantifying the actual influence of user i, we used the number of posts that were posted to user i's wall. If user j posts to user i's wall, we consider that information as having spread from user i to user j, and the number of users who post to the wall of each user is used to characterize user influence.

APS This dataset contains a social network representing co-authorships in APS journals and records of citations of papers published until 2005. In this network, each author is represented as a node, and co-authorship between author i and author j is represented as an undirected link between node i and node j. As a measure for quantifying the actual influence of author i, we used the number of citations of papers written by author i. If author j cites the paper of author i, we consider that information to have spread from

author i to author j, and the number of citations of papers of each author is used to characterize the influence of the author.

3.3. Sampling strategies

This study uses three biased sampling strategies: sample edge count (SEC) [20], breadth-first search (BFS), and depth-first search (DFS). For comparison purposes, we also use random sampling of nodes. SEC, BFS, and DFS are known to be biased towards sampling high-degree nodes [20]. Since influential nodes tend to have high degree [35], the negative effects of those biased sampling strategies on influence measures are expected to be smaller than the effects of random sampling.

Overviews of the sampling strategies are described below. The sampling strategies repeatedly obtain nodes until the number of obtained nodes reaches the desired sample size. We assume that when obtaining node i, the nodes linked to node i are known. In this experiment, only the sampled nodes and the links between those nodes are used for estimating node influence.

SEC SEC aims to obtain high-degree nodes without global knowledge of the network, greedily obtaining the node with the highest expected degree. Let S be a set of obtained nodes. Initially, S contains a randomly selected node. SEC greedily obtains the node with the most links from the nodes in S [20]. This method greedily obtains the node with the highest expected degree.

BFS BFS first obtains a randomly selected node. Then, BFS iteratively visits the neighbors of the visited node. At each iteration, BFS visits and obtains an unvisited neighbor of the earliest visited node [20]. This procedure is repeated until a specified number of nodes is obtained. Note that if there are no unvisited neighbors, a randomly selected unvisited node is newly obtained from the entire network.

DFS DFS uses a sampling method that is similar to that of BFS. DFS also iteratively visits unvisited neighbors of visited nodes. At each iteration, DFS obtains an unvisited neighbor of the *most recently* visited node, walking backward through the most-recent list as needed to find unvisited neighbors. [20].

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

In the following results, we repeated the node sampling process 30 times for each sampling strategy and obtained the average values of the evaluation metrics, which will be introduced in the next subsection.

3.4. Evaluation metrics

For evaluating the effectiveness of influence measures for identifying influencers, we use the following two metrics.

Overlap To evaluate the consistency of actual influencers and the identified influencers, we use the overlap between them defined in the following manner, which was also used in previous studies [3].

$$\mathrm{Overlap}_{p\%} = \frac{|T_{p\%} \cap A_{p\%}|}{|T_{p\%}|}, \tag{1}$$

where $T_{p\%}$ is the set of identified influencers (i.e., nodes in the top $p\%$ of a ranking based on the influence measures) and $A_{p\%}$ is the set of users in the top $p\%$ of a ranking based on actual influence in the complete network G.

Normalized influence To evaluate the strength of influence of the identified influencers, we use normalized influence, which is defined as

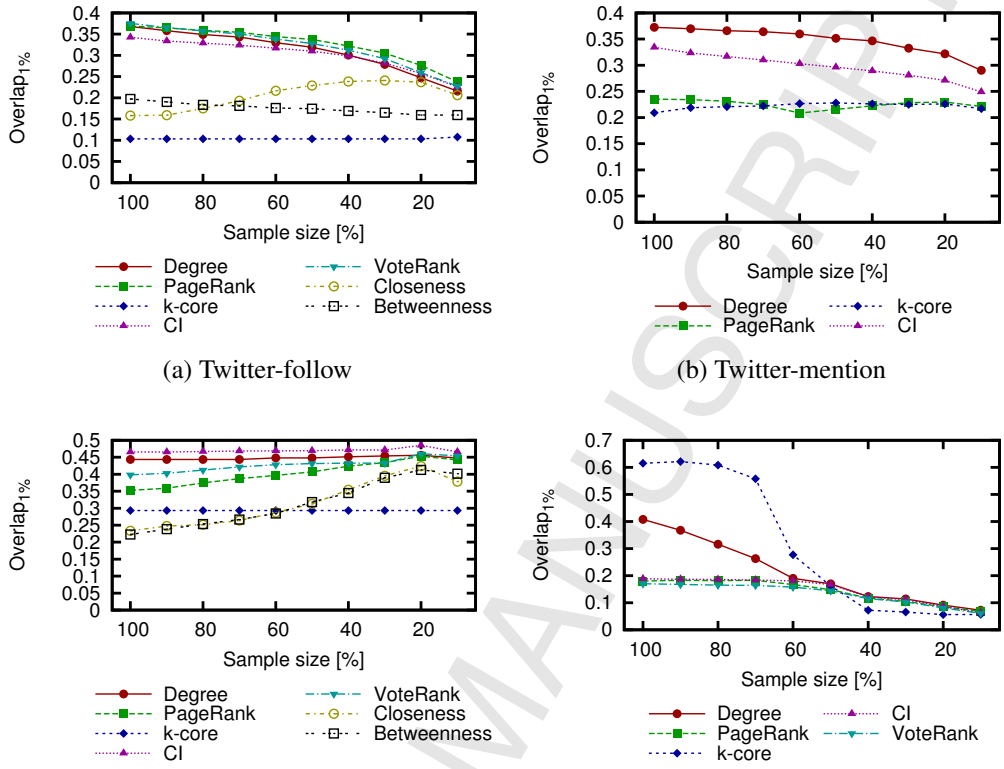
$$I_{p\%} = \frac{\sum_{v \in T_{p\%}} i(v)}{\sum_{u \in A_{p\%}} i(u)}, \tag{2}$$

where $i(v)$ is the actual influence of node v measured by the records of information cascades in the complete network G.

Unless explicitly stated otherwise, we use $p = 1$ for calculating the $\mathrm{Overlap}_{p\%}$ and normalized influence $I_{p\%}$. That is to say, we consider the problem of finding the top 1% of influencers from sampled social networks.

4. Results

We first present the results for $\mathrm{Overlap}_{1\%}$. Figures 1, 2, 3, and 4 show the relation between sample size and $\mathrm{Overlap}_{1\%}$ in each dataset when using SEC, BFS, DFS, and random sampling, respectively. Note that the results for closeness centrality and betweenness centrality in the Twitter-mention and APS datasets and



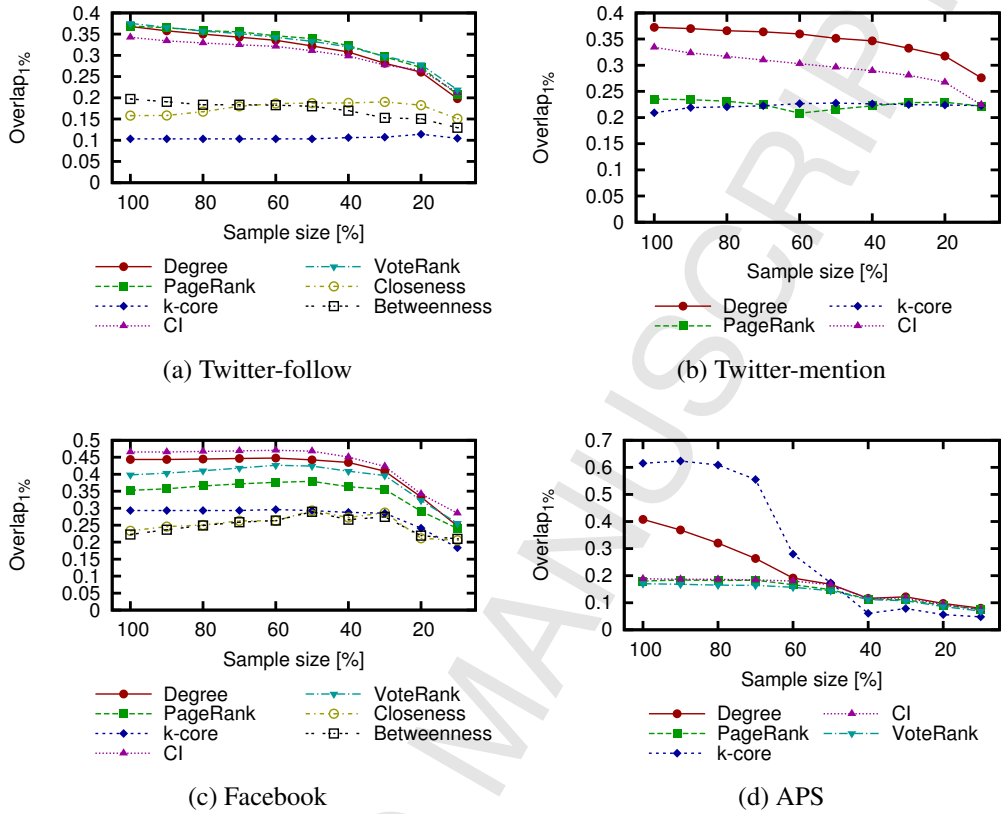
(a) Twitter-follow (b) Twitter-mention

(c) Facebook (d) APS

Figure 1: $\text{Overlap}_{1\%}$ vs. sample size in each dataset when using SEC.

the results for VoteRank in the Twitter-mention dataset are not shown, due to the high computational costs for calculating them. We used $l = 2$ as the parameter of CI, and $f = 1/\langle k \rangle$, where $\langle k \rangle$ is the average degree of the network, as the parameter of VoteRank, which are used in [31].

As shown in Fig. 1, we found that even when the sample size is 10%–30%, $\text{Overlap}_{1\%}$ is high, comparing with when the sample size is 100% (i.e., when using the entire network), except for the case of the APS dataset. This suggests that the negative effects of node sampling on the influence measures are small when using SEC sampling. This is because SEC sampling successfully finds high-degree nodes, which are expected to have high influence. For instance, in Facebook, approximately 93% of hub nodes (i.e., top 1% nodes based on degree) in the complete network are also contained in the sampled network even when the sample



(a) Twitter-follow

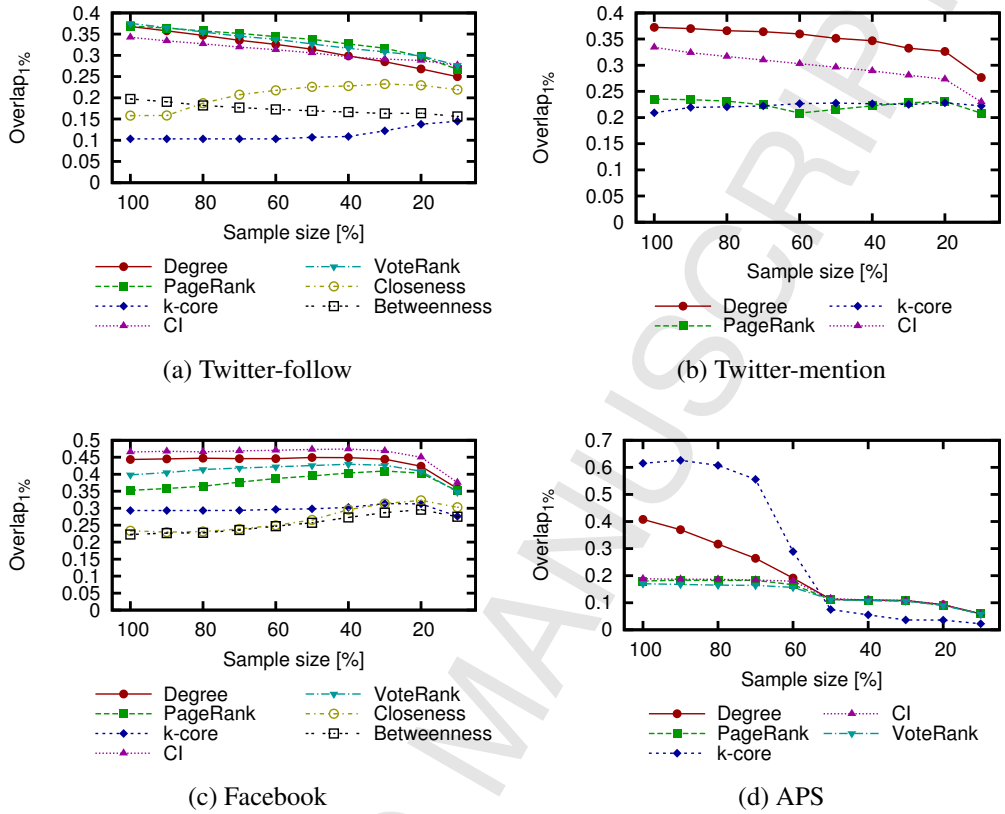
(b) Twitter-mention

(c) Facebook

(d) APS

Figure 2: $\text{Overlap}_{1\%}$ vs. sample size in each dataset when using BFS.

size is 10%. Therefore, $\text{Overlap}_{1\%}$ of degree when the sample size is 10% is comparable with that when the sample size is 100% in Facebook. In Twitter-follow, approximately 75% of hub nodes in the complete network are contained in the sampled network with 10% sample size. Therefore, in Twitter-follow, $\text{Overlap}_{1\%}$ of degree when the sample size is 10% is slightly lower than that when the sample size is 100%. We also found similar results when using BFS and DFS (Figs. 2 and 3). In contrast, when using random sampling and a sample size of 10%, we found that $\text{Overlap}_{1\%}$ was approximately 0.05 for all datasets and all influence measures (Fig. 4). These results indicate that for a social network known from sampling, using SEC, BFS, or DFS is more effective than random sampling at finding influential nodes in the network. However, in the results for the APS dataset with small sample sizes, $\text{Overlap}_{1\%}$ is relatively low even when using SEC,



(a) Twitter-follow

(b) Twitter-mention

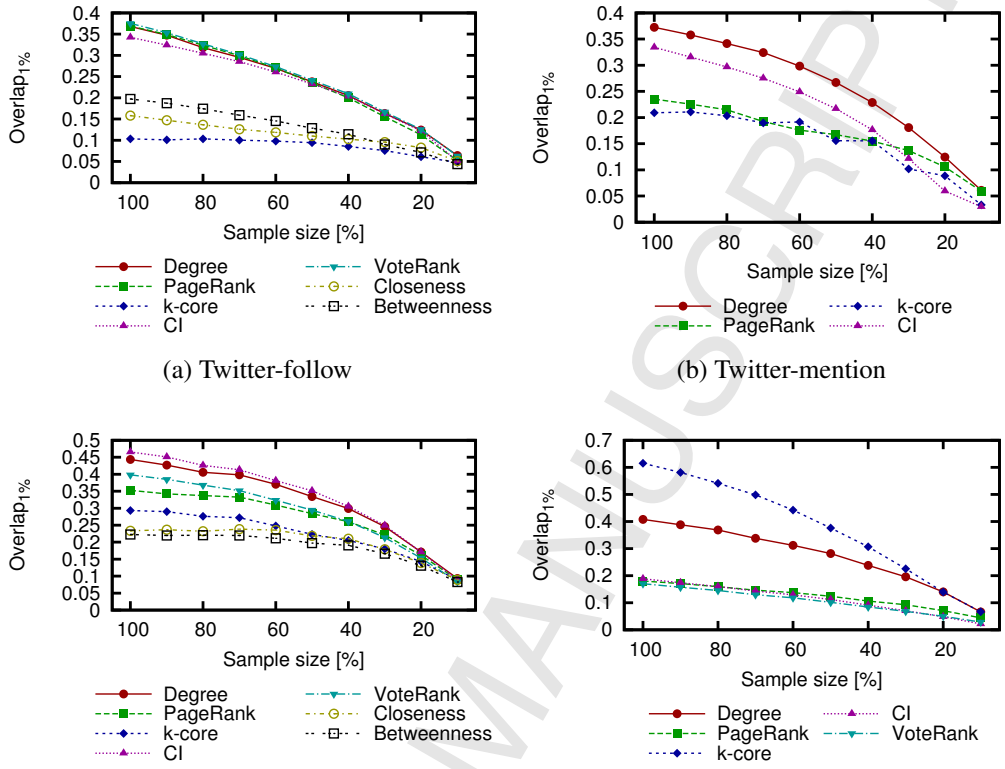
(c) Facebook

(d) APS

Figure 3: $\text{Overlap}_{1\%}$ vs. sample size in each dataset when using DFS.

DFS, or BFS. A sample size of 70% or more is necessary to achieve an $\text{Overlap}_{1\%}$ comparable with that when using the entire network. More detailed investigation is needed to clarify the cause of this, but this might be because these sampling strategies typically traverse a limited area of the network even when influential nodes are widely distributed in the network.

Comparing the differences among the influence measures, degree centrality, PageRank, CI, and VoteRank are effective for social media networks (i.e., Twitter-follow, Twitter-mention, and Facebook), and k-core is effective for a collaboration network (APS). In all networks, degree centrality, which uses only local information of the network, achieves comparable or even higher $\text{Overlap}_{1\%}$ values than do PageRank, betweenness, or closeness, which use global information of the network. This is in agreement with [35], which reports that degree centrality is



(a) Twitter-follow

(b) Twitter-mention

(c) Facebook

(d) APS

Figure 4: $\text{Overlap}_{1\%}$ vs. sample size in each dataset when using random sampling.

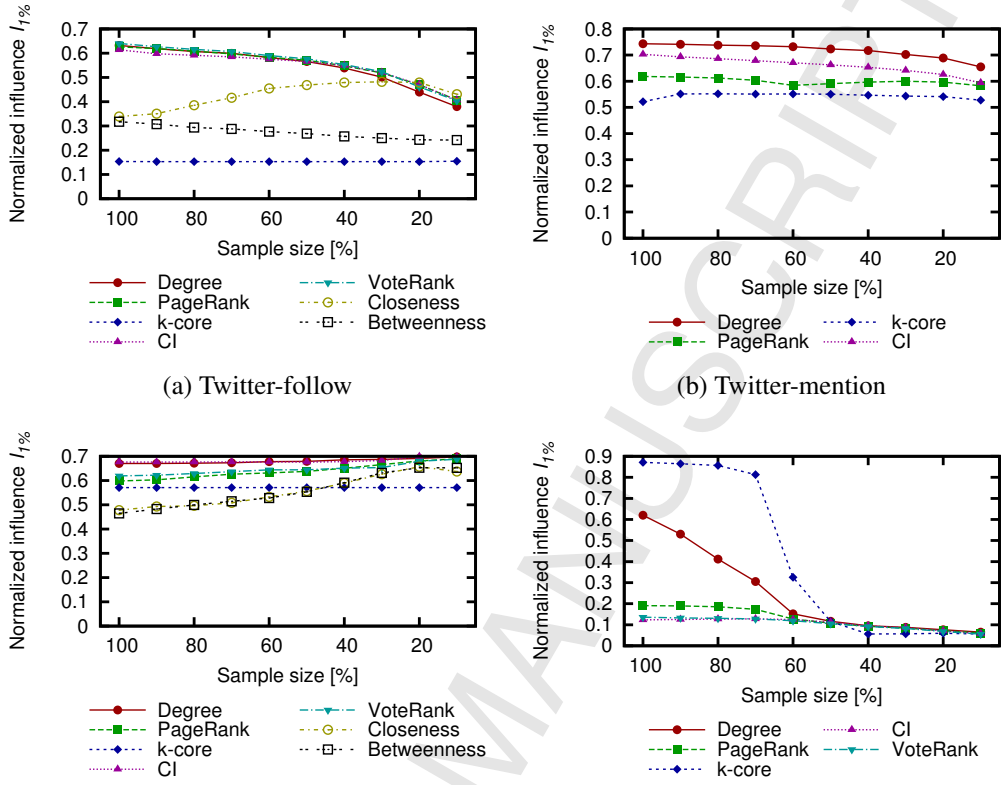
effective for estimating the influence of social media users. CI and VoteRank are also effective for social media networks. VoteRank achives the highest $\text{Overlap}_{1\%}$ for Twitter-follow, and CI achives the highest $\text{Overlap}_{1\%}$ for Facebook. The effectiveness of degree in social media networks might be explained by the definition of actual influence used in this paper. In social media networks, strong hubs who have extremely large number of links can get a lot of retweets and comments from their direct neighbors. Therefore, degree can be considered as a good predictor of influence. In contrast, in APS, hub nodes do not have the extreme number of links compared to the social media networks, and the number of citations do not heavily depends on the number of direct neighbors.

In the results for the Twitter-follow and Facebook datasets, to our surprise, we found that the $\text{Overlap}_{1\%}$ with a small sample size is slightly higher than that

when the sample size is 100% for some influence measures. In particular, the results of closeness centrality when using SEC (Fig. 1) show notable tendencies. We carefully investigated the node rankings based on closeness, and found that there exist low-degree and high-closeness nodes in Twitter-follow and Facebook. We particularly focused on Twitter-follow, and extracted the top 1% nodes based on closeness in the complete network (i.e., network with 100% sample size). We investigated the degree of the extracted nodes, and found that the first, second, and third quartiles of their degree are 2, 5, and 33, respectively. In contrast, for the sampled networks with 10% sample size, the first, second, and third quartiles of degree of the top 1% nodes based on closeness are approximately 38.1, 60.2, and 103.3, respectively. Typically, low-degree nodes tend to have low influence. Therefore, $\mathrm{Overlap}_{1\%}$ of closeness in the complete network is lower than that in the sampled networks with small sample sizes. The reason why low-degree nodes have high closeness in the complete network is as follows. In Twitter, high-degree hub nodes are followed by many low-degree nodes. Since the hub nodes naturally have high closeness, nodes adjacent to hub nodes also have high closeness. As a result, many low-degree nodes who follow high-degree nodes have high closeness in Twitter-follow. In contrast, when the sample size is small (e.g., sample size is 10%), such low-degree nodes are not included in the sampled networks since SEC preferentially obtains high-degree nodes. We found similar tendencies on Facebook. From above observations, we conclude that the existence of low-degree and high-closeness nodes, and the biases of SEC sampling are the causes of the counterintuitive tendencies of closeness. This suggests that network sampling sometimes reduces noise in estimating the influence of nodes.

We next investigated the normalized influence of the extracted influencers. Figure 5 shows the relation between sample size and normalized influence $I_{1\%}$ when using SEC sampling. This result again suggests that by using SEC sampling, we can identify influencers from a small sample of a social network, but not from the APS dataset. We also found that the values of normalized influence $I_{1\%}$ were high compared with the values of $\mathrm{Overlap}_{1\%}$. This means that highly influential nodes can be successfully identified.

Finally, we investigated the normalized influence $I_{p\%}$ while changing p. Figure 6 shows the relation between the fraction of extracted nodes p and the normalized influence $I_{p\%}$ when using SEC as a sampling strategy and degree centrality as an influence measure. Results with different sample sizes are compared in the figure. These results show that the normalized influence when sample sizes are 10%–30% is comparable with that when the sample size is 100% in social media networks regardless of the value of p. For APS, a sample size of 70% or more



(a) Twitter-follow

(b) Twitter-mention

(c) Facebook

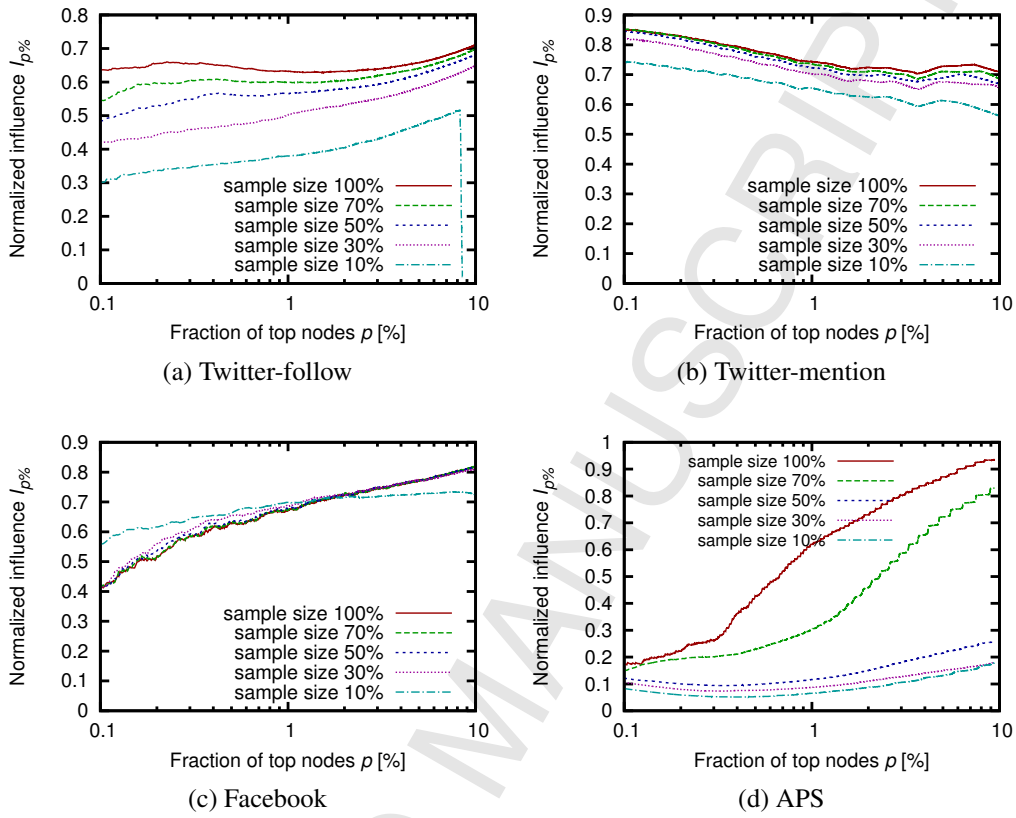
(d) APS

Figure 5: Normalized influence $I_{1\%}$ vs. sample size in each dataset when using SEC.

seems to be necessary for finding highly influential nodes. Again, to our surprise, these results show that with Facebook, 10% sampling finds more influential nodes than using the entire network. This suggests the potential benefit of network sampling in identifying influencers.

5. Discussion

Our results show that when using biased sampling strategies such as SEC, we can identify influencers in social media networks with a small sample size. Particularly for Twitter-mention and Facebook networks, a sample size of only 10% is enough for finding influencers comparable with those found from complete social networks. This is beneficial because social media networks are huge, and it is typically almost impossible to collect an entire network due to access restrictions



(a) Twitter-follow (b) Twitter-mention

(c) Facebook (d) APS

Figure 6: Normalized influence vs. fraction of top nodes p in each dataset when using SEC.

in the API [22]. Even when the complete network structure is available, using sampled networks is beneficial for calculating influence measures while keeping computational costs low [21].

Our results also show that in some cases, using sampled social networks more accurately identifies influencers than using the entire network does. This suggests that some nodes act as noise when identifying influencers, and that network sampling eliminates such nodes. Therefore, we expect that a better method for identifying influencers can be built by performing some pre-processing on the original network. Designing such pre-processing techniques is an interesting problem for future work.

Our study has several limitations. First, we should investigate the effects of other popular sampling strategies on the effectiveness of influence measures. For

instance, Metropolis–Hastings Random Walk [36], Forest Fire sampling [37], and Frontier sampling [38] are also widely used. These strategies aim to use unbiased sampling to accurately infer network characteristics such as degree distribution and clustering coefficients. A sampling strategy that samples a neighbor of a randomly selected node is used for the network immunization problem [39] and the detection of disease outbreaks [40]. To examine this strategy is also interesting. Rank Degree is a recently proposed sampling strategy that can be used for finding influential nodes [24]. However, our study and Rank Degree have different assumptions in the sampling scenario. Rank Degree assumes that the degree of the neighbors of a sampled node are known, whereas we assume that the degree of a node is known only when the node is sampled. This paper therefore does not include these sampling strategies, but investigating the effects of these strategies on finding influencers would be interesting.

Second, we should also investigate the effectiveness of other influence measures for identifying influencers. This paper uses conventional and widely used measures (i.e., degree, closeness, betweenness, PageRank, and k-core), but these are not state-of-the-art. The effectiveness of other recently proposed measures such as weighted LeaderRank [6], and interaction weighted k-core [12] should be investigated in future research.

Finally, the relation between the network structure and the necessary sample size is still unclear. The reason why APS is different from other network datasets should be investigated. Clarifying the differences between social media networks (e.g., Twitter-mention, Twitter-follow, and Facebook) and collaboration networks (e.g., APS) and why these differences affect the results of identifying influencers could be useful for designing better sampling strategies.

6. Conclusion

In this paper, we extensively investigated the effects of node sampling from a social network on the effectiveness of influence measures for identifying influencers in social networks. We applied the popular sampling strategies SEC, BFS, DFS, and random sampling to four social networks and identified influencers in the networks using degree, betweenness, closeness, PageRank, and k-core index. Our experimental results show that the effect of using biased (i.e., non-random) sampling strategies is generally small for identifying influencers in social media. For social media networks, we identified influencers whose influence is comparable with that of those identified from the complete social network by sampling only 10–30% of the networks with biased sampling strategies. In contrast, for

a collaboration network, a sample size of 70% or more was needed for identifying influencers. From above observations, we conclude that using biased sampling strategies is an effective approach for identifying influencers in large-scale social media networks. In contrast, for identifying influencers in collaboration networks, using network sampling is not a good approach, and we should obtain the network structure as completely as possible. Our results also suggest that network sampling is sometimes beneficial for identifying influencers. Particularly in Facebook-type networks, nodes with higher influence are discovered more readily from sampled social networks than from complete social networks. We believe that this is because network sampling eliminates noise encountered when identifying influencers.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Number 16K20931.

[1] K. Kimura, S. Tsugawa, Estimating influence of social media users from sampled social networks, in: Proceedings of Workshop on Social Influence (SI 2016), 2016, pp. 1302–1308.

[2] X. Zhang, J. Zhu, Q. Wang, H. Zhao, Identifying influential nodes in complex networks with community structure, Knowledge-Based Systems 42 (2013) 74–84.

[3] S. Pei, L. Muchnik, J. S. Andrade Jr, Z. Zheng, H. A. Makse, Searching for superspreaders of information in real-world social media, Scientific Reports 4 (2014) 5547.

[4] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, H. A. Makse, Identification of influential spreaders in complex networks, Nature Physics 6 (11) (2010) 888–893.

[5] D. Kempe, J. Kleinberg, É. Tardos, Maximizing the spread of influence through a social network, in: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03), 2003, pp. 137–146.

[6] Q. Li, T. Zhou, L. Lü, D. Chen, Identifying influential spreaders by weighted leaderrank, Physica A: Statistical Mechanics and its Applications 404 (2014) 47–55.

[7] L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, T. Zhou, Vital nodes identification in complex networks, Physics Reports 650 (2016) 1–63.

[8] E. Bakshy, J. M. Hofman, W. A. Mason, D. J. Watts, Everyone's an influencer: Quantifying influence on Twitter, in: Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11), 2011, pp. 65–74.

[9] S. Tsugawa, H. Ohsaki, On the relation between message sentiment and its virality on social media, Social Network Analysis and Mining 7 (1) (2017) 19.

[10] E. Katz, P. F. Lazarsfeld, Personal influence: the part played by people in the flow of mass communications., Free Press, 1955.

[11] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, T. Zhou, Identifying influential nodes in complex networks, Physica A: Statistical Mechanics and Its Applications 391 (4) (2012) 1777–1787.

[12] M. A. Al-garadi, K. D. Varathan, S. D. Ravana, Identification of influential spreaders in online social networks using interaction weighted K-core decomposition method, Physica A: Statistical Mechanics and its Applications 468 (2017) 278–288.

[13] C. Budak, D. Agrawal, A. El Abbadi, Limiting the spread of misinformation in social networks, in: Proceedings of the 20th International Conference on World Wide Web (WWW'11), 2011, pp. 665–674.

[14] L. C. Freeman, Centrality in social networks conceptual clarification, Social Networks 1 (3) (1979) 215–239.

[15] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Computer Networks and ISDN Systems 30 (1) (1998) 107–117.

[16] S. B. Seidman, Network structure and minimum degree, Social Networks 5 (3) (1983) 269–287.

[17] S. P. Borgatti, K. M. Carley, D. Krackhardt, On the robustness of centrality measures under conditions of imperfect data, Social Networks 28 (2) (2006) 124–136.

[18] D. J. Wang, X. Shi, D. A. McFarland, J. Leskovec, Measurement error in network data: A re-classification, Social Networks 34 (4) (2012) 396–409.

[19] P.-J. Kim, H. Jeong, Reliability of rank order in sampled networks, The European Physical Journal B-Condensed Matter and Complex Systems 55 (1) (2007) 109–114.

[20] A. S. Maiya, T. Y. Berger-Wolf, Benefits of bias: Towards better characterization of network sampling, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'11), 2011, pp. 105–113.

[21] N. K. Ahmed, J. Neville, R. Kompella, Network sampling: From static to streaming graphs, ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (2) (2014) 7.

[22] M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, Walking in Facebook: A case study of unbiased sampling of OSNs, in: Proceedings of the 29th IEEE International Conference on Computer Communication (INFOCOM'10), 2010, pp. 1–9.

[23] S. Mihara, S. Tsugawa, H. Ohsaki, Influence maximization problem for unknown social networks, in: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'15), 2015, pp. 1539–1546.

[24] N. Salamanos, E. Voudigari, E. J. Yannakoudakis, Deterministic graph exploration for efficient graph sampling, Social Network Analysis and Mining 7 (1) (2017) 24.

[25] J. Borge-Holthoefer, Y. Moreno, Absence of influential spreaders in rumor dynamics, Physical Review E 85 (2) (2012) 026116.

[26] T. L. Frantz, M. Cataldo, K. M. Carley, Robustness of centrality measures under uncertainty: Examining the role of network topology, Computational and Mathematical Organization Theory 15 (4) (2009) 303–328.

[27] S. Tsugawa, Y. Matsumoto, H. Ohsaki, On the robustness of centrality measures against link weight quantization in social networks, Computational and Mathematical Organization Theory 21 (3) (2015) 318–339.

[28] J. Platig, E. Ott, M. Girvan, Robustness of network measures to link errors, Physical Review E 88 (6) (2013) 062812.

[29] Q. Niu, A. Zeng, Y. Fan, Z. Di, Robustness of centrality measures against network manipulation, Physica A: Statistical Mechanics and its Applications 438 (2015) 124–131.

[30] F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (7563) (2015) 65–68.

[31] J.-X. Zhang, D.-B. Chen, Q. Dong, Z.-D. Zhao, Identifying a set of influential spreaders in complex networks, Scientific Reports 6 (2016) 27823.

[32] X. Teng, S. Pei, F. Morone, H. A. Makse, Collective influence of multiple spreaders evaluated by tracing real information flow in large-scale social networks, Scientific Reports 6 (2016) 36043.

[33] S. Pei, X. Teng, J. Shaman, F. Morone, H. A. Makse, Efficient collective influence maximization in cascading processes with first-order transitions, Scientific Reports 7 (2017) 45240.

[34] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi, On the evolution of user interaction in Facebook, in: Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09), 2009, pp. 37–42.

[35] J. Weng, E.-P. Lim, J. Jiang, Q. He, Twitterrank: finding topic-sensitive influential twitterers, in: Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM'10), 2010, pp. 261–270.

[36] M. Gjoka, M. Kurant, C. T. Butts, A. Markopoulou, Practical recommendations on crawling online social networks, IEEE Journal on Selected Areas in Communications 29 (9) (2011) 1872–1892.

[37] J. Leskovec, C. Faloutsos, Sampling from large graphs, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining (KDD'06), 2006, pp. 631–636.

[38] B. Ribeiro, D. Towsley, Estimating and sampling graphs with multidimensional random walks, in: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement (IMC'10), 2010, pp. 390–403.

[39] R. Cohen, S. Havlin, D. Ben-Avraham, Efficient immunization strategies for computer networks and populations, Physical Review Letters 91 (24) (2003) 247901.

[40] N. A. Christakis, J. H. Fowler, Social network sensors for early detection of contagious outbreaks, PloS One 5 (9) (2010) e12948.