2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Estimating Influence of Social Media Users from Sampled Social Networks

Kazuma Kimura, Sho Tsugawa
Graduate School of Systems and Information Engineering, University of Tsukuba,
1–1–1 Tennodai, Tsukuba, Ibaraki 305–8573, Japan
Email: kimura@mibel.cs.tsukuba.ac.jp, s-tugawa@cs.tsukuba.ac.jp

***Abstract*—Several indices for estimating the influence of social media users have been proposed. Most such indices are obtained from the topological structure of a social network that represents relations among social media users. However, several errors are typically contained in such social network structures because of missing data, false data, or poor 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 indices for estimating the influence of social media users. We compare the estimated influence of users, as obtained from a sampled social network, with their actual influence. Our experimental results show that using biased sampling methods, such as sample edge count, is a more effective approach than random sampling for estimating user influence, and that the use of random sampling to obtain the structure of a social network significantly affects the effectiveness of indices for estimating user influence, which may make indices useless.**

***Index Terms*—social networks, social media, node ranking, influence estimation, network sampling.**

I. Introduction

Research on social network analysis (SNA) has been actively pursued[1], [2], [3]. In SNA, several indices for estimating influence of social media users have been proposed [4], [5], [6], [7], [8], [9]. Such indices are expected to be useful for identifying influential users who can spread information to many other users in viral marketing on social media. Most indices are obtained from the topological structure of a social network that represents relations among social media users. In social networks, social media users are represented by nodes, and friendship or follow relations are represented by links. User influence is estimated from the network structure. Popular indices for estimating user influence include centrality (degree centrality, closeness centrality, and betweenness centrality) [10], PageRank [11], and the k-core [12] index of nodes corresponding to social media users.

Existing studies evaluate the effectiveness of indices for estimating influence when a social network structure is completely available. However, it is difficult to obtain the entire structure of large social networks. Several errors are typically contained in social network due to missing data, false data, or poor node and link sampling from the social network [13]. The effects of such incompleteness in social networks on indices estimating user influence are not known.

This paper investigates the effects of node sampling from a social network on the effectiveness of indices obtained from the sampled network for estimating the influence of social media users. Because social networks in social media, such as Twitter and Facebook, are huge, their structures are generally estimated by node sampling. Assuming such a situation, we evaluate the effectiveness of influence indices when the social network is obtained by several sampling methods through comparison between estimated and actual user influences. Following Pei et al. [9], we evaluate the effectiveness of indices by using actual records of information cascades.

Our main contributions are as follows:

- We extensively investigate the effects of node sampling on the influence indices. While existing studies have investigated the effects of random errors in a social network on influence indices [13], we investigate the effects of both random and non-random node sampling on the indices.
- We show that the effect of using biased (i.e., non-random) sampling methods is generally small for identifying influential users in social media. Our results suggest that when the social network is available from only a limited number of node samplings, using biased sampling methods such as sample edge count [14] is an effective approach for identifying influential users.

The remainder of this paper is organized as follows: In Section II, we introduce existing studies related to using social network structures to estimate the influence of social media users. Section III explains the methodology of the experiments, and Section IV shows the results. Finally, Section V concludes this paper and discusses future work.

II. Related Work

A. Indices for estimating the influence of social media users

Although several indices for estimating the influence of social media users have been proposed, centrality (degree centrality, closeness centrality, and betweenness centrality) [10], PageRank [11], and k-core index [12] of nodes in a social network are most popular and widely used [5], [6], [7], [8], [9]. A social media user is represented by a node in the social network, and these node indices are used as indices of user influence. This section provides definitions of these indices.

Degree centrality estimates the influence of a node based on its degree. For a directed network, both in-degree centrality and out-degree centrality can be defined; we use the in-degree

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-4799-7492-4/15/$31.00

centrality in this paper. The in-degree centrality of node v_k is defined as:

$$C_d(v_k) = \frac{\sum_{i=1}^{n} a(v_i, v_k)}{n-1}, \quad (1)$$

where n is the number of nodes in the network, and $a(v_i, v_k) = 1$ if a link from node v_i to v_k exists, otherwise 0.

Closeness centrality estimates the influence of a node based on the distance between the nodes and other nodes in a network. The closeness centrality of node v_k is defined as:

$$C_c(v_k) = \frac{n-1}{\sum_{i=1}^{n} d(v_k, v_i)}, \quad (2)$$

where $d(v_k, v_i)$ is the shortest path length from node v_k to v_i.

Betweenness centrality estimates the influence of a node based on the proportion of shortest paths between all other node pairs passing through the node. The betweenness centrality of node v_k is defined as:

$$C_b(v_k) = \frac{\sum_{i<j}^{n}\sum^{n} \sigma_{ij}(v_k)}{\sum_{i<j}^{n}\sum^{n} \sigma_{ij}}, \quad (3)$$

where σ_{ij} is the number of shortest paths from node v_i to v_j, and $\sigma_{ij}(v_k)$ is the number of shortest paths from node v_i to v_j through node v_k.

While PageRank was originally proposed for estimating the importance of a web page, it is also widely used for estimating the influence of nodes in social networks [9]. The PageRank of node v_k is defined as:

$$PR(v_k) = \frac{1-d}{n} + d \sum_{v_i \in T(v_k)} \frac{PR(v_i)}{L(v_i)}, \quad (4)$$

where d is a damping factor, which can be set between 0 and 1, $T(v_k)$ is a set of nodes that have a link to node v_k, and $L(v_i)$ is the number of links originating from node v_i.

The k-core index estimates the influence of a node based on the size of a dense subnetwork, called the core, to which the node belongs. The k-core of node v in network G is defined as follows: Let H be a subnetwork of network G, and let $\delta(H)$ be the degree of a node whose degree is the minimum among the nodes belonging to subnetwork H. In other words, the degree of each node belonging to H is at least $\delta(H)$. Subnetwork H is a k-core of network G if $\delta(H) \geq k$. The k-core index of node v is defined as the maximum k of the k-core to which node v belongs.

B. Effectiveness of influence indices

The effectiveness of indices for estimating influence has been evaluated through both simulations using influence cascade models [8] and experiments using actual records of information cascades [9]. In the simulations, the actual influence of a node corresponding to a user is defined as the number of nodes affected by an influence cascade originating from that node. Most studies use susceptible infectious recovered (SIR) [15], independent cascade (IC) [16], or linear threshold (LT) [17] models. Chen et al. [8] evaluate the effectiveness of popular influence indices by comparing the actual influence as obtained from simulations of the SIR model with the estimated influence obtained from the indices.

While most existing studies use influence cascade models for evaluating the effectiveness of influence indices, as mentioned above, Pei et al. [9] use the actual records of information diffusion for the evaluation. Pei et al. define the actual influence of a user as the number of reposts of that user's posts. Following Pei et al., we use the actual records of an information cascade for evaluating influence indices. Previous studies have evaluated the effectiveness of influence indices when the complete structure of a social network is available. In contrast, we evaluate the effectiveness of influence indices when the structure of a social network is only available from node samplings.

C. On the robustness of influence indices

Borgatti et al. [13] and Frantz et al. [18] have investigated the robustness of centrality against the addition or deletion of nodes and links. These studies investigate the stability of node rankings based on centrality measures against the random addition and deletion of nodes and links. Our study is similar in that it investigates the characteristics of centrality measures in incomplete networks. However, we particularly focus on the effectiveness of centralities for identifying influential nodes rather than their ranking stability. Moreover, we investigate the effects of both random and non-random intentional node samplings on influence indices. To the best of our knowledge, the effects of non-random node sampling on influence indices have not been evaluated in the existing studies.

III. Methodology

A. Overview

We investigate the relation between user influence as estimated from a sampled incomplete social network and the actual user influence. First, by sampling a fraction of nodes from a complete social network, we obtain an incomplete social network. Second, we estimate the influence of each node by calculating its degree centrality, closeness centrality, betweenness centrality, PageRank, and k-core index in the obtained incomplete social network. Third, we compare the influence of each user as estimated from the incomplete social network and the volume of information diffusion originating from the user, which corresponds to the actual user influence. Following Pei et al. [9], the actual user influence is defined as the capability of the user for spreading information to other users. We obtain user rankings based on both their estimated influence and the volume of information diffusion originating from them, and calculate the consistency of these rankings. Following Borgatti et al. [13] and Pei et al. [9], we define $\text{Overlap}_{1\%}$ as a measure for user ranking consistency. $\text{Overlap}_{1\%}$ is defined as $|E_{1\%} \cup A_{1\%}|/|E_{1\%}|$, where $E_{1\%}$ is the

TABLE I: Statistics for each dataset

Dataset	Number of nodes	Number of links	Number 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 Journals	247,675	856,864	45,684,601

set of users in the top 1% of a ranking based on the estimated influence and $A_{1\%}$ is the set of users in the top 1% of a ranking based on actual influence.

B. Datasets

We use four types of datasets: one is our collected dataset called *Twitter-follow*, and three publicly available datasets called *Twitter-mention* [1] [2], *Facebook* [19] [2] and *APS Journals* [2]. Several statistics of these datasets are shown in Table I.

The details of each dataset are described below.

Twitter-follow This dataset contains a social network that represents follow relations of Twitter users and the records of retweets posted by them during a specific period, described below. In the social network, a 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 user i's tweets at least once. If user j retweets user i's tweet, we consider that information has spread from user i to j, and the number of users retweeting each user's tweet is used as proxy of user influence. This dataset was collected by the following process using the Twitter API:

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 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, a 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 j. As a measure for quantifying the actual influence of user i, we used the number of users who have retweeted user i's tweets at least once.

Facebook This dataset contains a social network representing user friendships on Facebook, and the records of posts by them during a period of September 25, 2006, to January 22, 2009. In this network, a Facebook user is represented as a node, and a friendship between user i and j is represented as an undirected link between node i and 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 has spread from user i to j, and the number of users who posts to the wall of each user is used as a proxy of user influence.

APS Journals This dataset contains a social network representing co-authorships in APS journals, and records of citations of papers published until 2005. In this network, an author is represented as a node, and a co-authorship of author i and j is represented as an undirected link between node i and 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 j, and the number of citations of papers of each author is used as a proxy of the influence of the author.

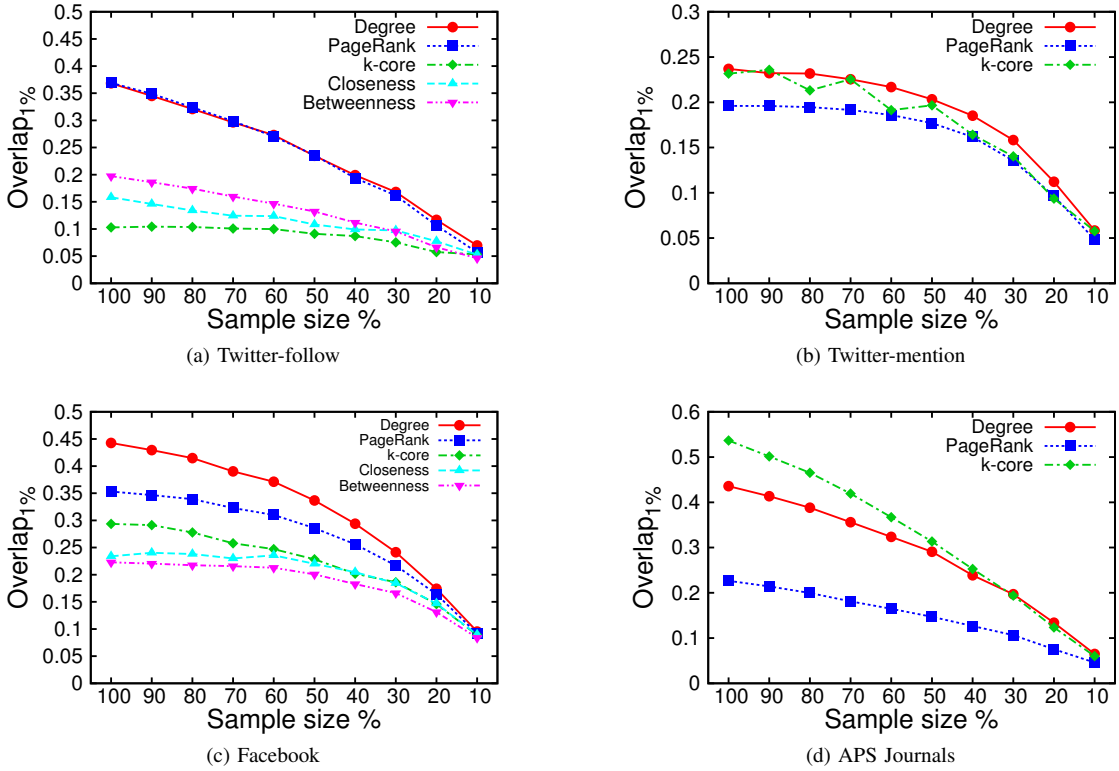
C. Sampling methods

This study uses four typical network sampling methods: breadth-first search (BFS), depth-first search (DFS), sample edge count (SEC) [14], and random sampling. This experiment investigates the effects on influence indices of applying these sampling methods to the social networks. Compared to random sampling, BFS, DFS, and SEC are known to be biased to sample high-degree nodes [14]. Since influential nodes tend to have high degree [6], the bad effects of those biased sampling methods on influence indices are expected to be smaller than the effect of random sampling.

Overviews of the sampling methods are described below. The sampling methods repeatedly obtain nodes until the number of obtained nodes reaches a 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 links between those nodes are used for estimating their influence.

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

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



(a) Twitter-follow

(b) Twitter-mention

(c) Facebook

(d) APS Journals

Fig. 1: Relations between sample size and $\text{Overlap}_{1\%}$ in each dataset when using random sampling.

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.

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 [14]. 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 [14].

SEC SEC aims to obtain high-degree nodes without global knowledge of the network, greedily obtaining a node with highest expected degree. Let S be a set of obtained nodes. Initially, S contains a randomly selected node. SEC greedily obtains a node with the most links from the nodes in S [14]. This method greedily obtains the node with the highest expected value of degree.

In the following results, for each sampling method, we repeated the node sampling process 30 times, and obtained average value of $\text{Overlap}_{1\%}$.

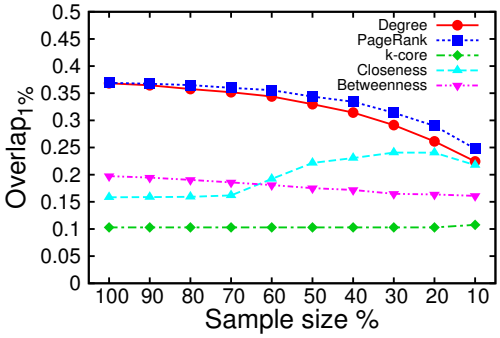
IV. Results

First, we investigate the effects of random sampling on influence indices. We rank nodes based on their influence as estimated from the sampled social network, and calculate the $\text{Overlap}_{1\%}$ between the rankings, based on the estimated influence and the ranking based on the actual influence. Figure 1 shows the relations between sample size and $\text{Overlap}_{1\%}$ in each dataset. Note that the results for closeness centrality and betweenness centrality in the Twitter-mention and APS Journals datasets are not shown, due to the high computational costs that would be required.

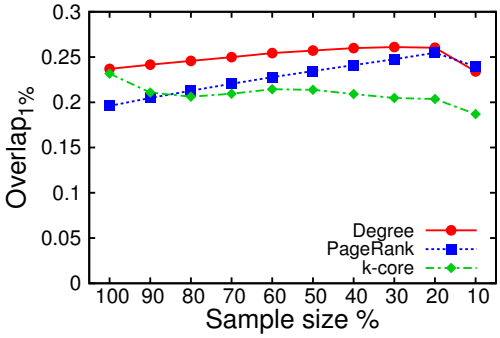
These results show that $\text{Overlap}_{1\%}$ gradually decreases as sample size decreases in all datasets and for all influence indices. If the nodes are sampled at random, then the probability that the sampled subnetwork contains influential nodes decreases linearly against the sample size, a natural result. Therefore, the decrease of $\text{Overlap}_{1\%}$ against sample size is natural. Focusing on the results when the sample size is 10%, we find that $\text{Overlap}_{1\%}$ is approximately 0.05 for all datasets and all indices. This indicates that using a sampled network obtained from 10% random sampling is not effective for finding influential nodes, since even randomly selecting 1% of nodes from a network without using its topological structure achieves $\text{Overlap}_{1\%}$ of 0.01.

We next investigate the effects of other sampling methods on influence indices. The relations between sample size and $\text{Overlap}_{1\%}$ when using SEC, BFS, and DFS sampling are shown in Figs. 2, 3, and 4, respectively.

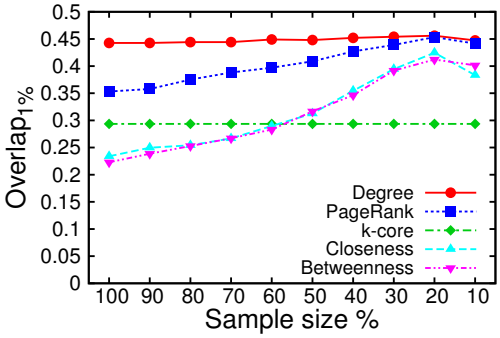
From Fig. 2, we find that $\text{Overlap}_{1\%}$ when using SEC sampling is higher than that when using random sampling. Even when the sample size is 10%, $\text{Overlap}_{1\%}$ is comparable



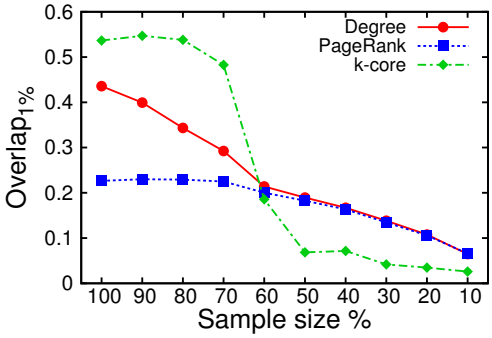
(a) Twitter-follow



(b) Twitter-mention

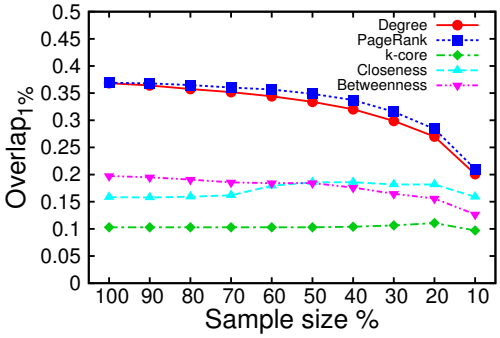


(c) Facebook

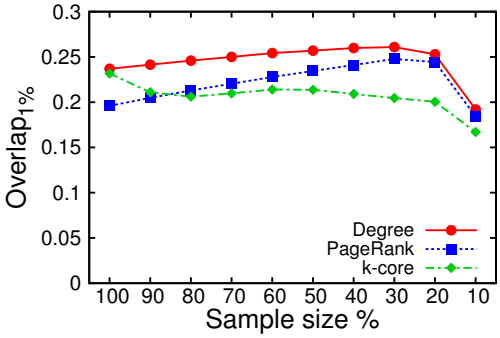


(d) APS Journals

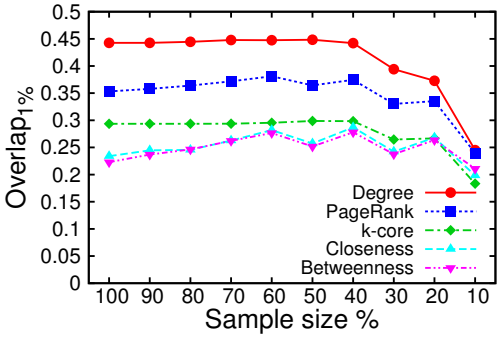
Fig. 2: Relations between sample size and $\text{Overlap}_{1\%}$ in each dataset when using SEC.



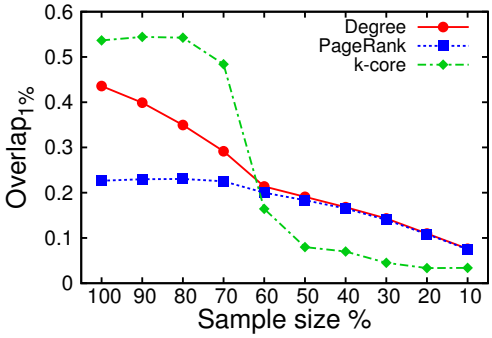
(a) Twitter-follow



(b) Twitter-mention

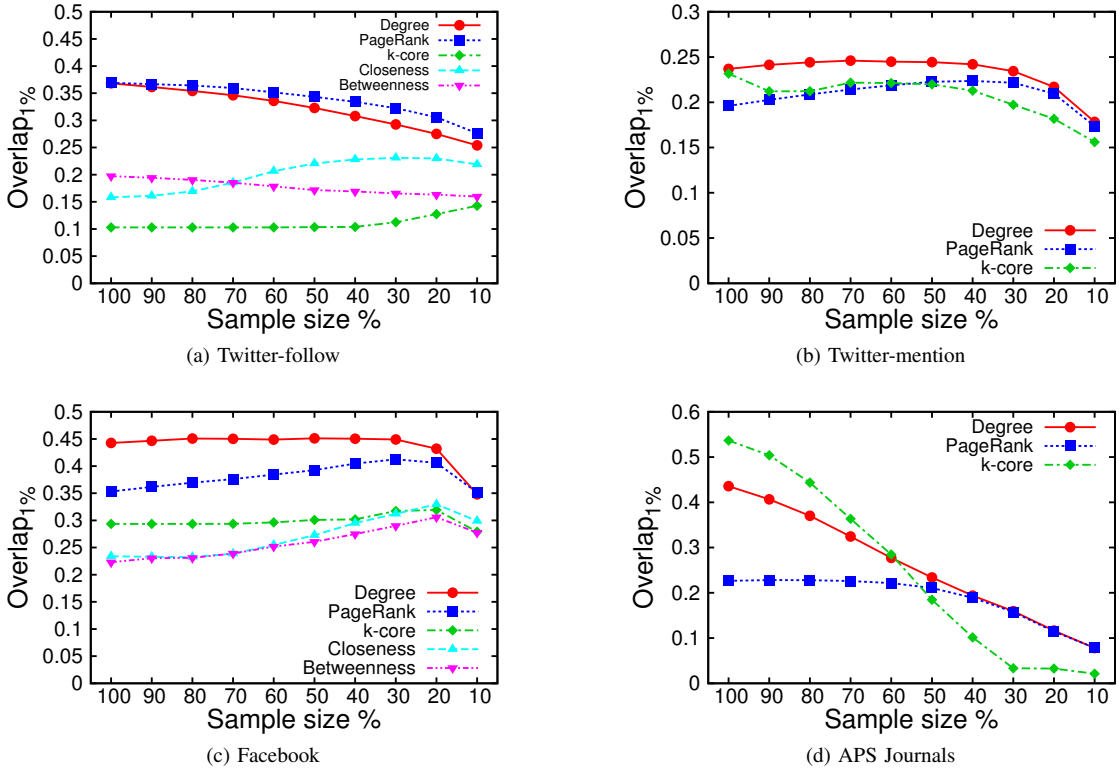


(c) Facebook



(d) APS Journals

Fig. 3: Relations between sample size and $\text{Overlap}_{1\%}$ in each dataset when using BFS.



(a) Twitter-follow

(b) Twitter-mention

(c) Facebook

(d) APS Journals

Fig. 4: Relations between sample size and $\text{Overlap}_{1\%}$ in each dataset when using DFS.

with that when the sample size is 100% (i.e., when using the entire network), except for the case of the APS Journals dataset. This is because SEC sampling successfully finds high-degree nodes, which are expected to have high influence. We can also find similar results when using BFS and DFS (Figs. 3 and 4).

These results indicate that when the structure of a social network is available from sampling alone, using SEC, BFS, or DFS rather than random sampling is effective for finding influential nodes in the network. However, in the results for the APS Journals dataset (Figs. 1(d) and 2(d)) with small sample sizes, $\text{Overlap}_{1\%}$ is relatively low even when using SEC, DFS, or BFS. More detailed investigation is needed to clarify the cause of this, but this might be because these methods typically traverse a specific area of the network even when influential nodes are widely distributed in the network.

Comparing the difference among influence indices, the effects of sampling on each index are similar. Focusing on the values of $\text{Overlap}_{1\%}$, degree centrality—which only uses local information of the network—achieves comparable or even higher $\text{Overlap}_{1\%}$ values than do PageRank or k-core, which use global information of the network. This is in agreement with [6], which reports that degree centrality is effective for estimating the influence of social media users.

In the results for the Twitter-mention and Facebook datasets, we can find that, for some indices, the $\text{Overlap}_{1\%}$ with a small sample size is slightly higher than that when the sample size is 100%. We expect that this is because node sampling reduces noise in estimating the influence of nodes.

V. Conclusion and Future Work

In this paper, we extensively investigated the effects of node sampling from a social network on the effectiveness of indices for estimating influence of social media users. While existing studies have investigated the effects of random errors in a social network on influence indices [13], we investigated the effects of both random and non-random node sampling on the indices. Our experimental results shows that the effect of using biased (i.e., non-random) sampling methods is generally small for identifying influential users in social media. Our results suggest that when the social network is available from only a limited number of node samplings, using biased sampling methods such as sample edge count [14] is an effective approach for identifying influential users.

In future work, we plan to propose a robust index for estimating the influence of social media users from only a sampled social network. It is also our important future work to investigate the effects of other types of incompleteness in social networks on the influence indices. For instance, it would be an interesting future direction to consider a situation where the complete neighbours of a node are not always available when the node is sampled.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Number 16K20931 and the Telecommunications Advancement

Foundation.

REFERENCES

[1] D. J. Watts, "A twenty-first century science," *Nature*, vol. 445, no. 7127, p. 489, Feb. 2007.

[2] S. P. Borgatti, A. Mehra, D. J. Brass, and G. Labianca, "Network analysis in the social sciences," *Science*, vol. 323, no. 5916, pp. 892–895, Feb. 2009.

[3] S. Tsugawa, Y. Matsumoto, and H. Ohsaki, "On the robustness of centrality measures against link weight quantization in social networks," *Computational and Mathematical Organization Theory*, vol. 21, no. 3, pp. 318–339, Sep. 2015.

[4] J. Shetty and J. Adibi, "Discovering important nodes through graph entropy the case of Enron email database," in *Proceedings of the 3rd International Workshop on Link Discovery*. ACM, 2005, pp. 74–81.

[5] S. P. Borgatti, "Identifying sets of key players in a social network," *Computational & Mathematical Organization Theory*, vol. 12, no. 1, pp. 21–34, 2006.

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

[7] A. Pal and S. Counts, "Identifying topical authorities in microblogs," in *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM'11)*. ACM, 2011, pp. 45–54.

[8] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica A: Statistical Mechanics and Its Applications*, vol. 391, no. 4, pp. 1777–1787, 2012.

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

[10] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1979.

[11] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems*, vol. 30, no. 1, pp. 107–117, 1998.

[12] S. B. Seidman, "Network structure and minimum degree," *Social Networks*, vol. 5, no. 3, pp. 269–287, 1983.

[13] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social Networks*, vol. 28, no. 2, pp. 124–136, 2006.

[14] A. S. Maiya and 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)*. ACM, 2011, pp. 105–113.

[15] W. Kermack and A. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A*, vol. 115, no. 772, pp. 700–721, 1927.

[16] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Letters*, vol. 12, no. 3, pp. 211–223, 2001.

[17] D. J. Watts, "A simple model of global cascades on random networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 9, pp. 5766–5771, 2002.

[18] T. L. Frantz, M. Cataldo, and K. M. Carley, "Robustness of centrality measures under uncertainty: Examining the role of network topology," *Computational and Mathematical Organization Theory*, vol. 15, no. 4, pp. 303–328, 2009.

[19] B. Viswanath, A. Mislove, M. Cha, and 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.