Influence Maximization Problem for Unknown Social Networks

Shodai Mihara*, Sho Tsugawa†, and Hiroyuki Ohsaki‡
* Graduate School of Science and Technology, Kwansei Gakuin University
† Faculty of Engineering Information and Systems, University of Tsukuba
‡ School of Science and Technology, Kwansei Gakuin University
{ E-mail: *dai@kwansei.ac.jp, †s-tugawa@cs.tsukuba.ac.jp, ‡ohsaki@kwansei.ac.jp }

***Abstract*—We propose a novel problem called *influence maximization for unknown graphs*, and propose a heuristic algorithm for the problem. Influence maximization is the problem of detecting a set of influential nodes in a social network, which represents social relationships among individuals. Influence maximization has been actively studied, and several algorithms have been proposed in the literature. The existing algorithms use the entire topological structure of a social network. In practice, however, complete knowledge of the topological structure of a social network is typically difficult to obtain. We therefore tackle an influence maximization problem for unknown graphs. As a solution for this problem, we propose a heuristic algorithm, which we call IMUG (Influence Maximization for Unknown Graphs). Through extensive simulations, we show that the proposed algorithm achieves 60–90% of the influence spread of the algorithms using the entire social network topology, even when only 1–10% of the social network topology is known. These results indicate that we can achieve a reasonable influence spread even when knowledge of the social network topology is severely limited.**

***Keywords*—social network; influence maximization; viral marketing; heuristic algorithm**

I. Introduction

Social media, such as Twitter and Facebook, are increasingly popular worldwide. As of early 2013, 200 million users were posting over 400 million messages on Twitter each day [1], and in March 2014 there were 1.28 billion monthly active users on Facebook [2].

Successful social media platforms are attractive not only for communication but also for information dissemination and so-called viral marketing. Social media users share information, some of which is disseminated to many other users by *word-of-mouth*. Such word-of-mouth information diffusion in social media is regarded as an important mechanism that influences public opinion and can affect product market share [3].

Detecting *influential* users is important for effective and efficient information dissemination and viral marketing in social media. For instance, suppose that a company develops a new product and would like to give free samples to social media users. The company hopes that sample recipients will post information about the new product, thereby spreading information about it among a large number of users and increasing its popularity. There is typically a limited budget for giving away samples, so giving samples to a small number of users who can influence many others is important for marketing success.

Motivated by applications such as viral marketing, influence maximization algorithms have been actively studied [4-11]. Influence maximization is a problem to detect a small set of influential nodes in a social network, which is a graph representing social relationships among individuals [4-7]. Given a social network, an influence cascade model, and a small number k, an influence maximization algorithm aims to find a set of k influential (seed) nodes in the network such that the expected number of nodes influenced by the seed nodes is maximized under the given cascade model [4, 5].

Existing influence maximization algorithms use the entire topological structure of the social network to detect seed nodes. For instance, Kempe et al. formulated the influence maximization problem as an optimization problem, and proposed an approximation algorithm to detect seed nodes from a social network [4]. Chen et al. proposed heuristic algorithms based on node degrees in social networks [5]. More recently, several improved approximation algorithms [8-11] and heuristic algorithms [12, 13] have been proposed. All of these algorithms use the entire topological structure of a social network.

However, complete knowledge of a social network's topological structure is typically difficult to obtain [14-16]. Social networks representing relationships among social media users are very large, and access to network data is typically limited, so we can typically only obtain a part of its structure [15, 16]. Even if we expend the time required to gather network data, it remains difficult to know the current state of highly dynamic social networks [14].

We therefore tackle the problem of influence maximization in unknown graphs, and propose a heuristic algorithm for the problem. Existing influence maximization algorithms assume that the entire social network topological structure is given. In contrast, our problem assumes only limited knowledge of the topological structure, which is obtained by *probing* (Fig. 1). We formulate this problem, and propose a heuristic algorithm called influence maximization for unknown graphs (IMUG) to solve it. We verify the effectiveness of IMUG through extensive simulations.

Our main contributions are summarized as follows.

- We address a novel and challenging problem, in which links between nodes are known from only a small number of limited probes. This is different from existing influence maximization problems, and designing efficient algorithms for probing and seed node selection from limited probing results are challenging

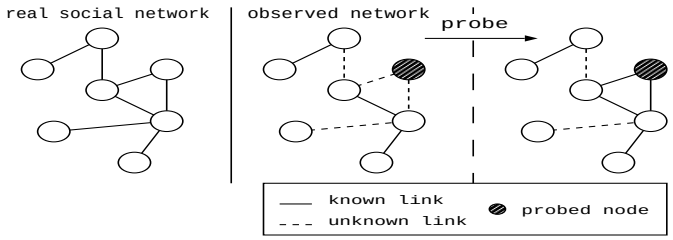


Fig. 1. An example of probing an unknown network

issues. Tackling this challenging problem is important for realizing effective influence spread in large social networks.

- We show that a reasonable solution can be obtained from limited knowledge of a social network's topological structure. Although the proposed algorithm adopts simple and straightforward approaches, we can achieve a 60–90% influence spread as compared with algorithms using the entire social network topology even when only 1–10% of the social network topology is known.

The remainder of this paper is organized as follows. In Section II, we introduce works related to influence maximization. In Section III, we formulate the influence maximization problem for unknown graphs. The proposed IMUG algorithm is presented in Section IV. Section V outlines experiments conducted on synthetic networks and real social networks to evaluate the effectiveness of IMUG. Finally, Section VI concludes this paper and discusses future work.

II. Related Work

Influence maximization has been actively studied in the literature [4-7, 17, 18]. The influence maximization problem of detecting a set of influential individuals was first proposed by Domingos et al. [6, 7] and formulated as an optimization problem by Kempe et al. [4]. Since the influence maximization problem is NP-hard, several approximation algorithms [4, 8-11] and heuristic algorithms [5, 12, 13] have been proposed. Recent research aims at designing faster approximation algorithms [8-11] and more effective heuristic algorithms [12, 13] than the existing ones.

Several related and variant problems have also been studied. Goyal et al. studied problems that minimize time or costs for a given size of influence spread [17]. Other related problems have also been studied, including influence maximization under multiple and competing influence cascades [18], influence maximization considering diversity of population [19], time-critical influence maximization [20], and topic-aware influence maximization [21].

These studies assume that the entire topological structure of a social network is available for obtaining solutions. However, as discussed in Section I, it is typically not easy to obtain complete knowledge of social network topological structures. We therefore tackle the influence maximization problem for unknown graphs.

The most closely related work is influence maximization in dynamic social networks [14]. Zhuang et al. studied influence maximization for dynamically changing graphs [14]. In their problem, complete knowledge of the social network's topological structure is initially available, but the network changes dynamically. They assume that the amount of noise in the network increases as time elapses due to network changes. Our study is similar to [14] in that both aim to maximize influence using incomplete information of the network topology. In contrast, however, we assume that the social network topological structure is unknown from the beginning, and only a limited amount of topology information can be obtained.

Several influence cascade models have been applied to the influence maximization problem. The independent cascade (IC) model, the weighted cascade (WC) model, and the linear threshold (LT) model are each popular and widely used [4]. In the IC model, each active (influenced) node spreads influence to its adjacent nodes with a predefined influence spread probability. In the WC model, each active node spreads influence to its adjacent nodes with an influence spread probability determined by its degree. In the LT model, each node becomes active if the ratio of active nodes among its adjacent nodes exceeds a predefined threshold.

III. Problem Formulation

We propose a problem called influence maximization for unknown graphs. The original influence maximization problem is for finding a set of seed nodes in a network such that the expected number of nodes influenced by the seed nodes is maximized under a given cascade model in a given social network. Our problem is similar, but does not assume that the social network's topological structure is given; the network topology is initially unknown, only the number of nodes is given, and a limited amount of probing is allowed to obtain a partial structure of the social network. Probing a node, for instance, corresponds to access via an application programming interface (API) for obtaining a list of friends of the target user. We assume multiple *rounds* in which probing, seed node selection, and to trigger influence spread from the selected seed nodes are allowed. Hence, the results of influence spread and probing in past rounds are available for seed node selection and further probing. We formulate this problem in detail below.

Following Chen et al. [5], we study influence spread on an unweighted, undirected graph $G = (V, E)$. In the initial state, only the set of nodes $V = \{v_1, \ldots, v_N\}$ is known, and the set of links E is unknown. Here, N is the number of nodes in graph G. Each node in graph G represents a user, and each link represents a relationship between two users in social media.

We assume R rounds, and that in each round, probing m nodes, an selecting κ seed nodes, and to trigger influence spread from the selected κ seed nodes are allowed. Probing node v obtains a list of nodes adjacent to v. In each round, we can repeat this probing process m times. We can also select κ seed nodes. The selected seed nodes become activated, spreads influence to its adjacent node, and each newly activated (influenced) node recursively repeats the influence spread process according to a given cascade model. This problem does not focus on designing an algorithm for a specific cascade model, and any influence cascade models such as the IC model, the LT model, and the WC model can be applied to this problem. We can obtain a list of activated nodes in each round. In what

follows, we call a node that has been activated at least once *active* node, and a node that has never been activated *inactive* node.

In summary, influence maximization for an unknown graph involves selecting m nodes to be probed and κ seed nodes for each round such that the expected number of active nodes in the Rth round is maximized. When selecting nodes to be probed, the results of probing and the results of influence spread in past rounds are available. When selecting seed nodes, information related to the selection of probed nodes and the results of probing in the current round are available.

IV. Influence Maximization Algorithm for Unknown Graphs

As a solution for the influence maximization problem for unknown graphs, we propose a heuristic algorithm called IMUG.

The basic ideas of IMUG are (1) greedily probing the node with the highest expected degree from unprobed nodes, and (2) greedily selecting the inactive node with the highest expected degree as a seed node by using the results of past probing and influence spread. Since we cannot obtain complete knowledge of the entire topological structure of the graph, we rely instead on available local information about node degrees. Since high-degree nodes tend to spread more influence than low-degree nodes do, and the effectiveness of degree-based heuristic algorithms have been shown [5], using node degree is a straightforward approach in situations of limited information. Note that the degree of each node is not available in the problem studied here. We therefore estimate each node degree from the results of past probing.

As a probing strategy, we adopt a biased sampling strategy called sample edge count (SEC), which is also known as a snowball sampling strategy [15]. SEC greedily probes the node with the highest expected degree. Given a set of already probed nodes D, SEC probes the node $v \in \overline{D}$ with the most links with nodes in D. Namely, the SEC strategy estimates the expected degree of node v in the original graph G as the degree of node v in the induced subgraph of $D \cup \{v\}$, and probes the node with the highest expected degree [15]. SEC is shown to be effective at finding hub nodes of large degree in several types of social network [15]. For each round, we repeat SEC probing m times.

In the initial state, IMUG estimates each node expected degree d_v as k_v $(v = 1, 2, \cdots, N)$, where k_v are IMUG parameters. If some knowledge of node degrees, such as the average degree, is available in advance, we can determine k_v by using that knowledge. When we have no information about node degree, a simple option is to set k_v to 0.

In each round, IMUG updates the expected degree of each node by using the results of SEC probing. When node v_p is probed, since the true degree of node v_p is known, d_{v_p}, which is the degree of node v_p, is fixed to the known true degree. Moreover, for each node v_i adjacent to node v_p, since node v_i is revealed to have at least one link (with node v_p), d_{v_i}, which is the expected degree of node v_i, is incremented by one unless the degree of node v_i is fixed.

TABLE I. Symbols used in the explanations of IMUG

G	undirected unweighted graph
V	set of nodes in graph G
N	number of nodes in graph G
R	number of rounds
S	set of seed nodes
D	set of probed nodes
A	set of active nodes
A_r	set of activated nodes in round r
d_v	expected degree of node v
k_v	initial value of d_v
v_p	node to be probed
v_s	seed node

Algorithm 1 IMUG

```
initialize A = φ and D = φ
for each node v ∈ V do
    initialize d_v = k_v
end for
for r = 1 to R do
    initialize S = φ
    for 1 .. m do
        select v_p = arg max_{v∈V} { d_v | v ∈ D̄ }
        D = D ∪ {v_p}
        d_{v_p} = actual degree of v_p
        for each neighbor u of v_p do
            d_u = d_u + 1 unless u ∈ D
        end for
    end for
    for 1 .. κ do
        select v_s = arg max_{v∈V} { d_v | v ∈ Ā }
        S = S ∪ {v_s}
        A = A ∪ {v_s}
    end for
    output S
    obtain A_r
    for each node v ∈ A_r do
        A = A ∪ {v}
    end for
end for
```

In each round, IMUG performs as follows. The pseudo-code of the proposed IMUG is shown in Algorithm 1, and symbol definitions are given in Table I.

1) Repeat the following procedures m times
   a) Select node v_p with the highest expected degree among nodes that have never been probed
   b) Probe node v_p to obtain a list of adjacent nodes
   c) Fix the degree of node v_p as the number of its adjacent nodes, and increment by one the expected degree of each adjacent node v_i unless node v_i has been already probed.
2) Rank inactive nodes by sorting them in descending order of their expected node degree, and select the top κ nodes in the ranking as seed nodes.
3) Activate the seed nodes, spread influence from them, and obtain a list of activated nodes in this round.

V. EXPERIMENT

In this section, we extensively evaluate the effectiveness of IMUG at spreading influence on unknown social networks. We perform simulation experiments on synthetic networks and real social networks to examine how much influence can be spread with limited knowledge of the network topology.

A. Methodology

We use two types of synthetic networks obtained from popular network generation models, the Erdös–Rényi (ER) model [22] and the Barabási–Albert (BA) model [23]. The ER model generates random graphs where the degree distribution follows a Poisson distribution, and the BA model generates scale-free graphs where the degree distribution follows a power-law distribution.

We also use five real social networks: NetHEPT[1] [5], DBLP[2] [24], Amazon[3] [25], Facebook-small[4] [26], and Facebook-large[5] [27]. NetHEPT and DBLP represent co-authorship among researchers, Amazon represents co-purchasing relationships among customers of an electronic commerce site, and Facebook-small and Facebook-large represent friendships among Facebook users. These are widely used as benchmark datasets for influence maximization problems [4, 5, 8, 9, 12, 28-31]. Table II summarizes characteristics of the networks used in the experiment. Directed networks are converted to undirected networks by simply ignoring the link direction. Multiple links are also simply ignored.

We perform simulations of influence spread on the introduced synthetic networks and real social networks, and investigate the number of active nodes in each round. Following [8, 10, 14], we use the IC model as an influence cascade model, and for influence spread probability we simply use p for all node pairs and for all rounds. In the IC model, each node is either activated or non-activated. Selected seed nodes become activated, and each activated node v_i spreads influence to its adjacent node v_j with a probability p if node v_j is non-activated in the current round. Each newly activated node recursively repeats the influence spread process. We assume that influence spread is independent between rounds. Namely, in each round each node may be activated and may spread influence to its neighbors regardless of whether the node was activated in a past round. We run simulations 100 times and take the average number of active nodes in each round. Note that *active node* is a node that has been activated at least once. For simulations of the ER and BA networks, we generate a network with the models for each simulation run.

Since influence maximization for unknown graphs is a new problem, we compare IMUG with naive approaches and an existing algorithm using the entire network topology. Specifically, the following algorithms are compared:

- **DegreeDiscountIC:** A heuristic algorithm shown to be effective when the network topology is completely known [5].
- **random**: A naive algorithm that randomly selects κ seed nodes from inactive nodes in each round without using knowledge of the network topology.
- **random-degree:** An improved naive algorithm that randomly probes m nodes from unprobed nodes and selects κ seed nodes with highest degree from already probed inactive nodes in each round.

Following [14], we chose DegreeDiscountIC as a comparison algorithm from among the effective algorithms [5, 8-13]. Note that DegreeDiscountIC is not designed to select seed nodes in each round, so in the simulation of DegreeDiscountIC we first select $R\kappa$ seed nodes and rank them using DegreeDiscountIC. Then in each round we activate κ nodes in the ranked order.

In the following simulation results, we used $k_v = 0$ $(v = 1, 2, \cdots, N)$ as IMUG parameters, assuming that knowledge of the network topology is not available in advance. Unless explicitly stated, we used $R = 100$ as the number of rounds, $\kappa = 1$ as the number of seed nodes for each round, and $m = \lceil 0.001N \rceil$ as the number of nodes to be probed for each round, where N is the number of nodes in graph G.

B. Results and Discussion

We first investigate the effectiveness of IMUG for spreading influence on the synthetic networks (ER and BA). Figure 2 shows the number of active nodes in each round for the ER and BA networks when the influence spread probability $p = 0.01$. Note that DegreeDiscountIC uses the entire network topology, IMUG and random-degree only use information obtained from probing m nodes in each round, and random does not use network topology.

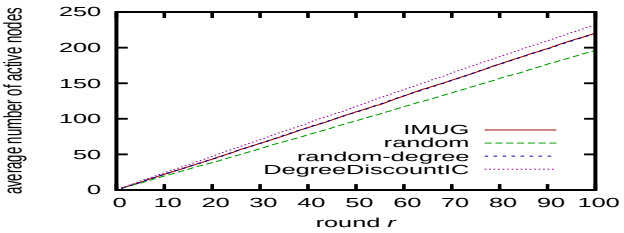
Figure 2 shows that IMUG achieves reasonable influence spread despite IMUG having only limited network topology knowledge. Particularly in earlier rounds (i.e., when the number of seed nodes is small), IMUG achieves comparable influence spread with DegreeDiscountIC, which uses the entire network topology. We find that random-degree also achieves influence spread comparable with DegreeDiscountIC and IMUG on the ER network. This is due to the degree distribution of the ER network, where node degrees are highly similar. In contrast, IMUG achieves a larger influence spread on the BA network than do the naive algorithms (random-degree and random) because the BA network has hub nodes with significantly large degrees. IMUG can successfully find such hub nodes with SEC probing, which improves the influence spread of IMUG over that of the naive algorithms. IMUG is thus expected to be an effective algorithm for influence maximization on unknown networks with hub nodes.

We next investigate the influence spread of each algorithm on synthetic networks when the influence spread probability $p = 0.1$ (Fig. 3). These results show that all algorithms including IMUG achieve large influence spread, and differences in influence spread among the algorithms when $p = 0.1$ are smaller than those when $p = 0.01$. Similar results, where the influence spread is not so sensitive to different algorithms when relatively large influence spread probability, are also reported in [4, 5]. The cause of this has been explained by the existence of a connected giant component after removing every edge with probability $1 - p$ [4, 5].

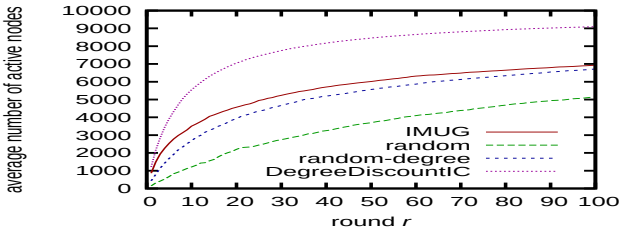
[1] http://research.microsoft.com/enus/people/weic/graphdata.zip
[2] http://snap.stanford.edu/data/com-DBLP.html
[3] http://snap.stanford.edu/data/amazon0302.html
[4] http://snap.stanford.edu/data/egonets-Facebook.html
[5] http://socialnetworks.mpi-sws.org/data-wosn2009.html

TABLE II. CHARACTERISTICS OF NETWORKS (VALUES FOR ER AND BA ARE AVERAGED FOR 100 DIFFERENT NETWORKS)

	ER	BA	NetHEPT	DBLP	Amazon	Facebook-small	Facebook-large
Number of nodes	10,000	10,000	15,233	317,083	262,114	4,039	63,731
Number of links	250,000	250,000	58,891	1049,870	1234,881	88,234	1545,686
Average degree	50	50	7.732	6.622	9.423	43.691	48.507
Standard deviation of degree	6.999	96.090	14.153	10.008	5.919	52.414	75.817
Coefficient of variation of degree	0.140	1.922	1.830	1.511	0.628	1.200	1.563
Clustering coefficient	0.005	0.065	0.429	0.732	0.202	0.617	0.059
Average path length	2.771	2.644	5.840	6.792	8.831	3.693	4.322
Degree assortativity coefficient [32]	0.0001	−0.058	0.389	0.267	−0.002	0.064	0.177

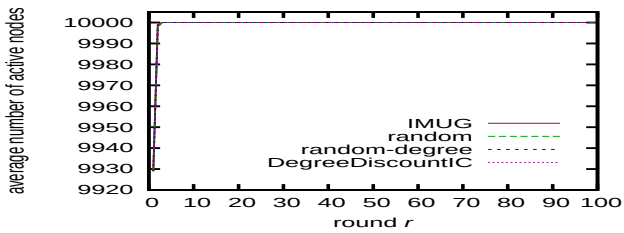


(a) ER

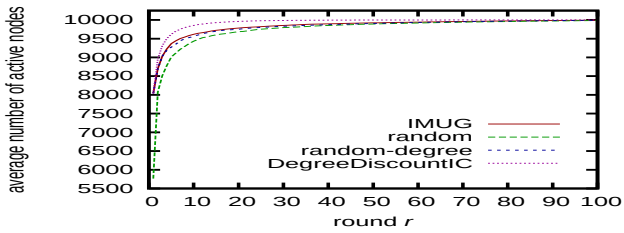


(b) BA

Fig. 2. Active nodes in each round on synthetic networks (influence spread probability: $p = 0.01$)



(a) ER



(b) BA

Fig. 3. Active nodes in each round on synthetic networks (influence spread probability: $p = 0.1$)

We next investigate the effectiveness of IMUG on real social networks. Many real social networks are known to have hub nodes [23, 33], which suggests that IMUG would be effective. However, real social networks have several characteristics that synthetic networks do not, such as small average path length, high clustering coefficients, and degree assortativity (see Table II). These characteristics may affect the performance of IMUG, so we perform simulations on several real social networks. Figure 4 shows the number of active nodes in each round for five real social networks when the influence spread probability $p = 0.01$.

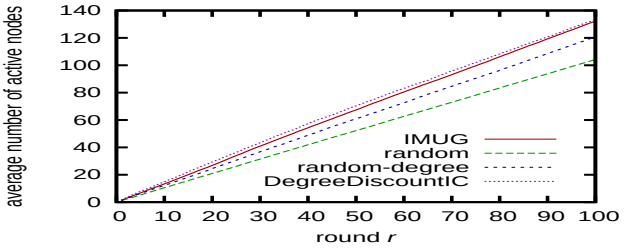
Figure 4 shows that IMUG achieves a larger influence spread than naive algorithms on most real social networks, and achieves comparable influence spread with DegreeDiscountIC, particularly on NetHEPT, DBLP, and Amazon. On Facebook-large and Facebook-small, the differences of influence spread between IMUG and DegreeDiscountIC are larger than those on other networks, which suggests room for improvement of IMUG. However, this does not indicate that IMUG is ineffective on these networks, since IMUG only uses very limited knowledge of network topology. For instance, in the 50th round IMUG only probes 5% of network nodes. IMUG achieves reasonable influence spread on these networks, considering its limited knowledge. Influence spread can be increased when more information about the network topology is available. We investigate the effects of the number of probed nodes on influence spread later in this section.

Figure 4(d) also shows that random-degree outperforms IMUG on Facebook-small. We need a more detailed investigation to reveal the cause of this, but it might be explained by the high average degree and relatively low coefficient of variation of degree in Facebook-small. In Facebook-small, there are many nodes with large degree (suggesting fewer nodes with small degree), so random probing may find high-degree nodes from a wider range of networks than does SEC probing, which may result in a larger influence spread under random-degree than IMUG. This suggests that IMUG can be improved by combining a random jump strategy with the SEC strategy when probing networks like Facebook-small. Such improvement of IMUG is an important area for future work.

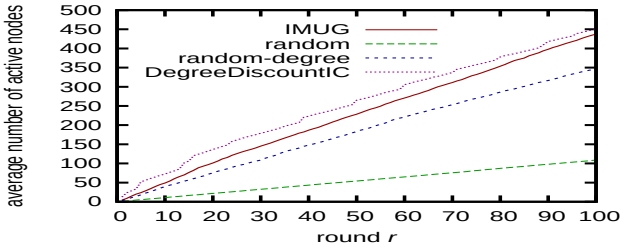
Comparing algorithms that do not use the entire network topology, we see that our strategy of selecting high-degree nodes as seed nodes and probes nodes with high expected degree is effective. IMUG and random-degree significantly outperform random, which indicates the effectiveness of selecting high-degree nodes as seed nodes. We can observe significant differences between random-degree and IMUG, particularly on DBLP and Amazon, which indicates the effectiveness of the SEC as a probing strategy. The SEC strategy helps us to find hub nodes efficiently, and this leads more active nodes.

We next investigate the influence spread of each algorithm when the influence spread probability $p = 0.1$. Figure 5 shows the number of active nodes in each round on the five real social networks.

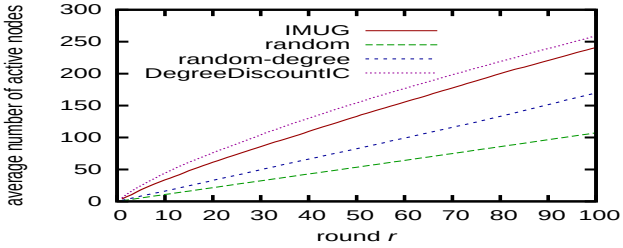
These results again show that reasonable influence spread can be achieved with very limited knowledge of the network



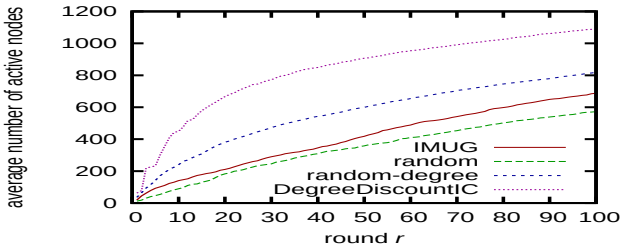
(a) NetHEPT



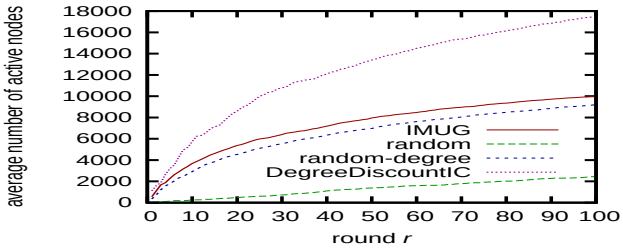
(b) DBLP



(c) Amazon

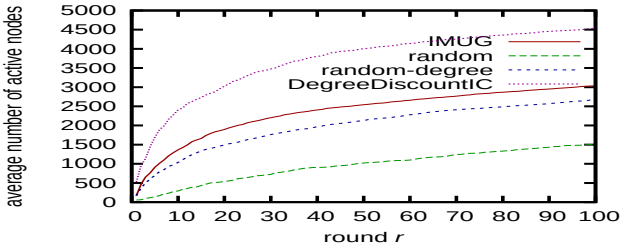


(d) Facebook-small

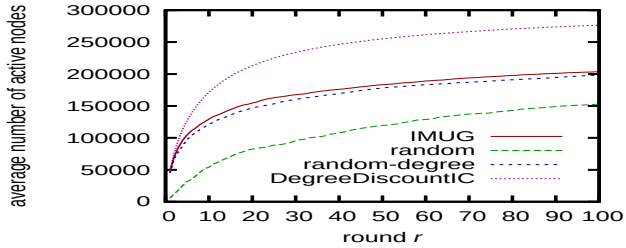


(e) Facebook-large

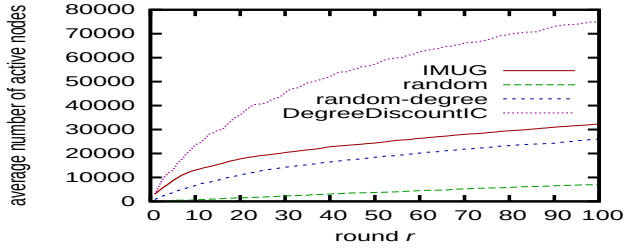
Fig. 4. Active nodes in each round on real social networks (influence spread probability: $p = 0.01$)



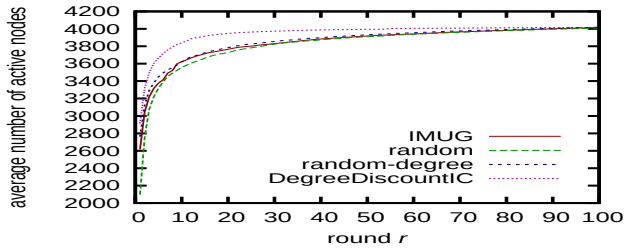
(a) NetHEPT



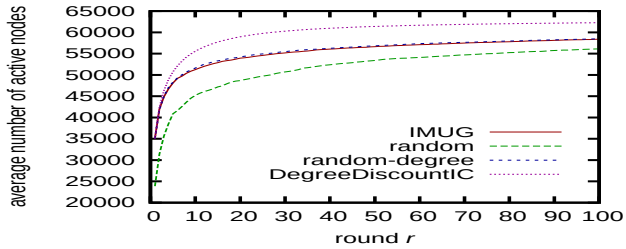
(b) DBLP



(c) Amazon



(d) Facebook-small



(e) Facebook-large

Fig. 5. Active nodes in each round on real social networks (influence spread probability: $p = 0.1$)

topology. IMUG outperforms naive algorithms on most networks. However, contrary to the results on synthetic networks, differences in influence spread between IMUG and DegreeDiscountIC when $p = 0.1$ are larger than those when $p = 0.01$ particularly on NetHEPT, DBLP, and Amazon. This suggests that these real social networks do not have giant components after removing every edge with probability $1 - p$ even when $p = 0.1$. Moreover, when the influence spread probability is large, small differences in seed node degree significantly affect the size of influence spread. Therefore, DegreeDiscountIC using global knowledge of network topology has an advantage over IMUG. In reality, the influence spread probability is considered to be not so large, and existing influence maximization studies often use a low influence spread probability such as $p = 0.01$ [5, 14, 28]. Therefore, IMUG in real situations is expected to achieve comparable performance with DegreeDiscountIC in most cases.

Finally, we investigate the relation between the amount of information used for seed node selection and influence spread. Namely, we investigate influence spread by changing the number of probed nodes m. Figure 6 shows the relation between the fraction of probed nodes until a specific round r to the number of nodes in the graph and the normalized number of active nodes in the round. The normalized number of active nodes is defined as the number of active nodes with IMUG divided by that with DegreeDiscountIC. Note that the number of normalized active nodes can be greater than 1 since DegreeDiscountIC is a heuristic and not an optimization algorithm.

These figures also show that in the previous results probing only 1–10% of nodes in the networks is sufficient for IMUG to achieve comparable influence spread with DegreeDiscountIC on NetHEPT, DBLP, and Amazon. For instance, approximately 90% of the influence spread of DegreeDiscountIC is achieved only from 10%, 5%, and 10% probing at the 10th, 50th, and 100th rounds, respectively.

On Facebook-small, the influence spread of IMUG increases with the number of probed nodes. Approximately 70–80% of the influence spread of DegreeDiscountIC is achieved by probing only 10–20% of nodes in the network.

On Facebook-large, the normalized number of active nodes is smaller than those on other networks, indicating that there is still room for improvement in IMUG. As already discussed in this section, one improvement may be combining random jumps with the SEC strategy and probing a wider range of the network. Although IMUG still has room for improvement, the influence spread with IMUG is considered to be reasonable. Looking at the (not normalized) number of active nodes, IMUG influences approximately 10,000 nodes (out of 63,731) with 100 seed nodes and probing 10% of the nodes in the network. This result may be enough when performing a viral-marketing campaign in practical situations.

From the above results, while the proposed IMUG has still room for improvement, we find that a reasonable influence spread can be achieved with surprisingly little probing. Specifically, IMUG achieves 60–90% of the influence spread of DegreeDiscountIC by probing only 1–10% of the nodes in a network. In real social networks, we can reach hub nodes with a small amount of probing due to the so-called *small-world phenomenon* [34], which allows a large influence spread with a small amount of probing.

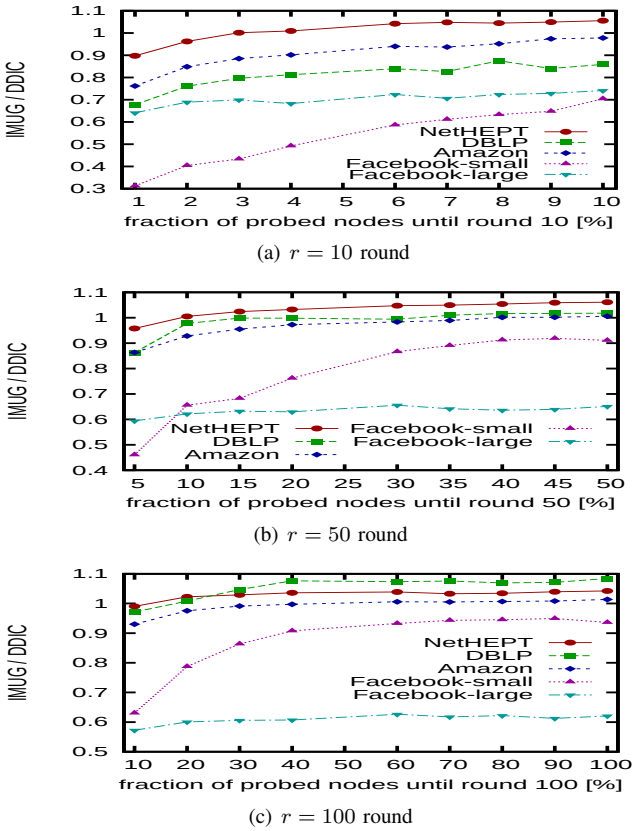


Fig. 6. Fraction of probed nodes until a specific round r vs. the number of active nodes in round r normalized by the number of active nodes with DegreeDiscountIC (influence spread probability: $p = 0.01$)

VI. Conclusion and Future Work

We proposed a novel problem called influence maximization for unknown graphs, and also proposed a heuristic algorithm, which we call IMUG, to address the problem. Unlike the original influence maximization problem, the entire topological structure of the social network is not given, and only limited knowledge of the topological structure is obtained through probing. We investigated the effectiveness of IMUG through extensive simulations, and results indicated that IMUG is effective for influence maximization on unknown social networks. Specifically, we have shown that IMUG achieves 60–90% of the influence spread of algorithms using the entire network topology by probing only 1–10% of the network nodes. Although we adopt simple and straightforward approaches in the algorithm, our results show that the straightforward algorithm achieves reasonable influence spread even when knowledge of the social network topology is severely limited. Our findings should be positive and important for practitioners conducting viral marketing on large social networks.

Since our current algorithm is simple and straightforward we plan to improve IMUG, for instance by incorporating past influence spread results in seed node selection and probing, and by using a random jump strategy for probing areas of interest. Investigating the effectiveness of the improved

IMUG for influence spread on larger-scale social networks is also important future work. Moreover, we are planning to investigate the effectiveness of IMUG under other influence cascade models than the IC model. We are also interested in considering the problem where the number of seed nodes and the number of nodes to be probed can be changed for each round.

ACKNOWLEDGMENTS

This work was partly supported by JSPS KAKENHI Grant Number 25280030 and the Telecommunications Advancement Foundation.

REFERENCES

[1] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh, "WTF : The who to follow service at Twitter," in *Proceedings of the International Conference on the World Wide Web (WWW'13)*, May 2013, pp. 505–514.

[2] "Facebook Reports First Quarter 2014 Results," `http://investor.fb.com/releasedetail.cfm?ReleaseID=842071`.

[3] E. Bakshy, J. M. Hofman, W. A. Mason, and 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)*, Feb. 2011, pp. 65–74.

[4] D. Kempe, J. M. Kleinberg, and E. 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)*, Aug. 2003, pp. 137–146.

[5] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*, Jun. 2009, pp. 199–208.

[6] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'01)*, Aug. 2001, pp. 57–66.

[7] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02)*, 2002, pp. 61–70.

[8] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Near-optimal time complexity meets practical efficiency," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD'14)*, Jun. 2014, pp. 75–86.

[9] A. Goyal, W. Lu, and L. V. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proceedings of the 20th International Conference Companion on World Wide Web (WWW'11)*, Mar. 2011, pp. 47–48.

[10] N. Ohsaka, T. Akiba, Y. Yoshida, and K. Kawarabayashi, "Fast and accurate influence maximization on large networks with pruned Monte-Carlo simulations," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI'14)*, Jul. 2014, pp. 138–144.

[11] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation: Scaling up with guarantees," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*, Jul. 2014, pp. 629–638.

[12] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large scale social networks," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'10)*, Jul. 2010.

[13] K. Jung, W. Heo, and W. Chen, "IRIE: Scalable and robust influence maximization in social networks," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM'12)*, Dec. 2012, pp. 918–923.

[14] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, "Influence maximization in dynamic social networks," in *Proceedings of the 13th IEEE International Conference on Data Mining (ICDM '13)*, Dec. 2013, pp. 1313–1318.

[15] 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)*, Sep. 2011, pp. 105–113.

[16] M. Gjoka, M. Kurant, C. T. Butts, and 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)*, Mar. 2010, pp. 1–9.

[17] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "On minimizing budget and time in influence propagation over social networks," *Social Network Analysis and Mining*, vol. 3, no. 2, pp. 179–192, Jun. 2013.

[18] S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in *Proceedings of the 3rd international Workshop on Internet and Network Economics (WINE'07)*, Dec. 2007, pp. 306–311.

[19] F. Tang, Q. Liu, H. Zhu, E. Chen, and F. Zhu, "Diversified social influence maximization," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'14)*, Aug. 2014, pp. 455–459.

[20] W. Chen, W. Lu, and N. Zhang, "Time-critical influence maximization in social networks with time-delayed diffusion process," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI'12)*, Jul. 2012, pp. 592–598.

[21] C. Zhang, J. Sun, and K. Wang, "Information propagation in microblog networks," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'13)*, Aug. 2013, pp. 190–196.

[22] P. Erdös and A. Rényi, "On the evolution of random graphs," in *Publications of the Mathematical Institute of Hungarian Academy of Sciences*, vol. 5, 1960, pp. 17–61.

[23] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[24] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," in *Proceedings of the 12th IEEE International Conference on Data Mining (ICDM'12)*, Dec. 2012, pp. 745–754.

[25] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Transactions on the Web (TWEB)*, vol. 1, no. 1, pp. 5:1–5:39, May 2007.

[26] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Proceedings of the Neural Information Processing Systems (NIPS'12)*, Dec. 2012, pp. 539–547.

[27] 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)*, Aug. 2009, pp. 37–42.

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

[29] H. Lamba and R. Narayanam, "A novel and model independent approach for efficient influence maximization in social networks," in *Proceedings of the 14th International Conference on Web Information Systems Engineering (WISE'13)*, Oct. 2013, pp. 73–87.

[30] Z. Huiyuan, T. N. Dinh, and M. T. Thai, "Maximizing the spread of positive influence in online social networks," in *Proceedings of the 33rd IEEE International Conference on Distributed Computing Systems (ICDCS'13)*, Jul. 2013, pp. 317–326.

[31] 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 *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM'14)*, Nov. 2014, pp. 171–180.

[32] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, pp. 208 701:1–208 701:4, Nov. 2002.

[33] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, "Structure and tie strengths in mobile communication networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, May 2007.

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