Accurate Path-based Methods for Influence Maximization in Social Networks

Yun-Yong Ko
Department of Computer and Software
Hanyang University, Korea
yyko@hanyang.ac.kr

Dong-Kyu Chae
Department of Computer and Software
Hanyang University, Korea
kyu899@hanyang.ac.kr

Sang-Wook Kim*
Department of Computer and Software
Hanyang University, Korea
wook@hanyang.ac.kr

ABSTRACT

This paper proposes a novel approach to target-oriented influence estimation, which remedies the drawback of state-of-the-art, thereby understanding information diffusion more accurately in a social network.

Keywords

Influence maximization, information diffusion, social networks

1. INTRODUCTION

As viral marketing in online social networks has attracted a lot of interest in these days, the problem of *influence maximization* (*IM*) has been extensively studied. Given a network G and a limited budget k, IM is a problem of finding a seed set S consisting of k seeds that maximize influence spread over G [3]. However, finding its optimal solution is *NP*-hard because it is required to compare all possible S. Kempe et al. [3] presented a greedy algorithm (referred to as *SimpleGreedy* henceforth) that guarantees to find 63% of the optimal solution [3]. However, *SimpleGreedy* still suffers from the low performance due to its running expensive Monte-Carlo (MC) simulations to estimate influence spread of a seed set. Instead of costly MC simulations, *SIMPATH* [2] and *IPA* [4] exploit the weights of every path starting from each seed (hereafter, we call these methods as *path-based methods*). They provide accuracy comparable to *SimpleGreedy*, but are an order of magnitude faster. In this paper, we point out the problem with the path-based methods in terms of accuracy and propose a novel approach to address the problem.

2. MOTIVATION

Eq. (1) describes how path-based methods compute influence of a given node v:

$$\sigma(\{v\}) = \sum\nolimits_{p \in \text{valid path of } v} W_p \quad (1)$$

Here, p represents an acyclic path starting from v and W_p represents the weight of p. W_p is computed by multiplying the weights w_e of all the edges e in p, i.e., $W_p = \prod_{e \in p} w_e$. Because it is known that the problem of finding all such paths is $\#P$-hard [2], they prune a path p once W_p becomes smaller than a pre-defined threshold.

Next, the influence spread of a seed set S is computed by the *linear sum* of all seeds' influence spread as in Eq. (2).

$$\sigma(S) = \sum\nolimits_{v \in S} \sigma(\{v\}) \quad (2)$$

The idea with existing methods is (1) to compute the amount of influence each individual seed (i.e., *source*) gives over all the non-seed nodes and then (2) to aggregate those of all the seed nodes thus computed. However, we claim this idea should be changed: (1) to compute the amount of influence each individual non-seed node (i.e., *target*) receives from a whole set of seed nodes and then (2) to aggregate those of all the non-seed nodes.

The reasoning behind this claim is as follows: The seed set obtained from *SimpleGreedy* is often considered as a *ground truth* for evaluating other approximate (or heuristic-based) *IM* algorithms; It defines the total amount of influence spread by a seed set as the number of non-seed nodes (i.e., *targets*) influenced (i.e., activated) by the seed set; This indicates that *SimpleGreedy* computes the influence spread by taking into account the influence received by an individual target node rather than that given by a source node.

In this paper, we propose a novel approach to target-oriented influence estimation that is able to remedy the drawback of existing path-based methods and thus provides more accurate results in *IM*.

3. PROPOSED APPROACH

When computing the influence spread of a seed set, our approach considers the amount of influence spread that all the non-seed nodes receive from the seed set. We first define $\sigma_d(S)$, the *aggregated* value of influence that a non-seed node $d \in V - S$ receives from all the seed nodes in S. The aggregation scheme in a non-seed node depends on which diffusion model is employed, namely *linear threshold* (LT) model or *independent cascade* (IC) model. In this paper, we focus on the IC model. Under the IC model, considering every seed node's influence toward a given target node independently, $\sigma_d(S)$ is computed as follows:

$$\sigma_d(S) = 1 - \prod\nolimits_{p \in P_{s \to d}} (1 - W_p) \quad (3)$$

Next, by taking the linear sum of influences received by every non-seed node, i.e., $\sigma_d(S)$, we get the total amount of

*Corresponding author

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889407.

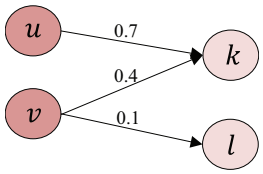

Figure 1: An example.

influence spread of the seed set as follows:

$$\sigma(S) = \sum_{d \in V-S} \sigma_d(S) \quad (4)$$

As a result, our approach successfully computes the influence spread received by non-seed (target) nodes, correctly following the spirit of *SimpleGreedy*.

Example: Figure 1 shows a seed set $S = \{u, v\}$ influencing non-seed nodes k and l. u influences k with a probability of 0.7 and v does k and l with probability of 0.4 and 0.1, respectively. The influence spread of S by existing source-oriented and our target-oriented approaches are as follows:

Source-oriented approach: $\sigma(S) = \sum_{s \in S} \sigma(s) = \sigma(u) + \sigma(v) = 0.7 + (0.4 + 0.1) = 1.2$

Target-oriented approach: $\sigma(S) = \sum_{d \in V-S} \sigma_d(S) = \sigma_k(S) + \sigma_l(S) = (1 - (1 - 0.7)(1 - 0.3)) + 0.1 = 0.89$

Note that both u and v try to influence k *at once*. In case of the source-based approach, however, they do not consider influence of u and v toward k independently; rather, they take the linear sum of the two influences. As a result, they fail to preserve the original characteristics of the IC model, thereby predicting influence spread incorrectly. In contrast, our approach computes influence spread correctly by considering it in the view point of target (non-seed) nodes.

4. EVALUATION

In this section, we evaluate the effectiveness of our approach with real-world datasets. Specifically, our experiments are to answer the following question: "*does our target-oriented approach provide a more accurate result than the source-oriented approach*?"

Dataset. We used two real world data sets consisting of *DBLP* and *Stanford* web graph.

Diffusion Model. We exploited the weighted cascade (WC) model [3], which is a widely-used variation of the IC model. It assigns a propagation probability to an edge (u, v) by $P_{uv} = \frac{1}{d_{in}(v)}$, where d_{in} is the *in-degree* of a node v.

Algorithms. We compared the following algorithms: *Random* selects nodes randomly for seeds (a baseline); *SDD* (*single degree discount*) selects nodes of a high degree (whenever a node is selected as a seed, the degree of its neighbors decreases by 1 [1]); *IPA* is the state-of-the-art under the IC model, which uses the source-oriented approach [4]; *TOA* (*target-oriented approach*) is the proposed one.

We found 100 sets of top-k seeds with k set as 1∼100 by employing each algorithm, and then ran 10,000 MC simulations with each seed set in order to understand its influence spread over a network.

Results. Figure 2 (a)∼(b) show the experimental results on influence spread (y-axis) according to the size of a seed set (x-axis). Among the four algorithms, *Random* provides the lowest influence spread. *SDD* provides influence spread higher than that of *Random*, but still lower than *IPA* and *TOA*. Our *TOA* shows the biggest influence spread. The difference of influence spread between *TOA* and *IPA* is more

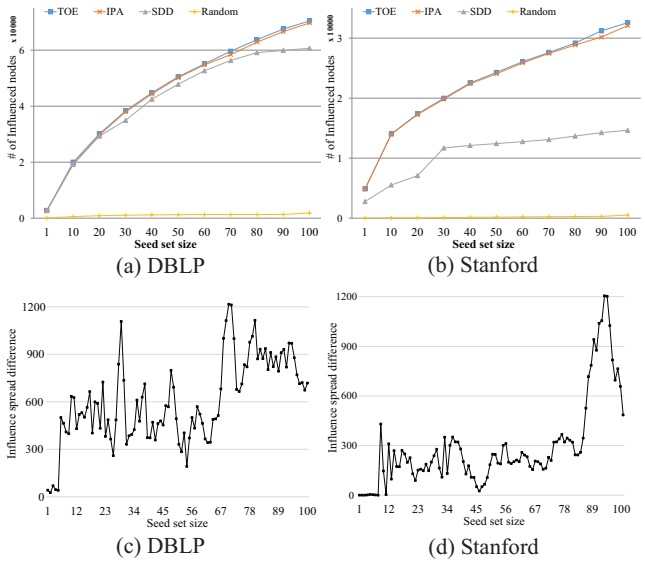

Figure 2: Experimental results.

clearly shown in Figure 2 (c)∼(d) where the x-axis indicates the seed set size and the y-axis indicates $(T - I)$ where T and I indicate the influence spread of *TOA* and *IPA*, respectively. The y-axis shows always positive values, which represents our *TOA* consistently outperforms *IPA* at any size of seed sets. The biggest differences are shown in *Stanford*, where the seed set of our *TOA* provides 4% larger spread than that of *IPA*.

In addition, up to the 60th seed node, *TOA* and *IPA* show similar amount of influence spread. However, after the 60th seed node, their differences tend to become bigger as the size of a seed set gets larger. This is because it is more likely that a non-seed node receives influences from multiple seed nodes as the size of a seed set gets larger. In this case, as seen in our example, *IPA* understands influence spread incorrectly. On the other hand, our *TOA* correctly interprets the influence spread and thus finds a more accurate result of top-k seeds that provides higher influence spread over a whole network.

5. ACKNOWLEDGEMENTS

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2014R1A2A1A10054151 and No.2015R1A5A7037751), and the Institute for Information and Communications Technology Promotion(IITP) grant funded by the Korean Government(MSIP) (No.R22121500070001002).

6. REFERENCES

[1] W. Chen et al. Efficient influence maximization in social networks. In *SIGKDD*, pages 199–208. ACM, 2009.

[2] A. Goyal et al. Simpath: An efficient algorithm for influence maximization under the linear threshold model. In *ICDM*, pages 211–220. IEEE, 2011.

[3] D. Kempe et al. Maximizing the spread of influence through a social network. In *SIGKDD*, pages 137–146. ACM, 2003.

[4] J. Kim et al. Scalable and parallelizable processing of influence maximization for large-scale social networks? In *ICDE*, pages 266–277. IEEE, 2013.