Which Targets to Contact First to Maximize Influence over Social Network

Kazumi Saito[1], Masahiro Kimura[2], Kouzou Ohara[3], and Hiroshi Motoda[4]

[1] School of Administration and Informatics, University of Shizuoka
52-1 Yada, Suruga-ku, Shizuoka 422-8526, Japan
k-saito@u-shizuoka-ken.ac.jp
[2] Department of Electronics and Informatics, Ryukoku University
Otsu 520-2194, Japan
kimura@rins.ryukoku.ac.jp
[3] Department of Integrated Information Technology, Aoyama Gakuin University
Kanagawa 229-8558, Japan
ohara@it.aoyama.ac.jp
[4] Institute of Scientific and Industrial Research, Osaka University
8-1 Mihogaoka, Ibaraki, Osaka 567-0047, Japan
motoda@ar.sanken.osaka-u.ac.jp

Abstract. We address a new type of influence maximization problem which we call "target selection problem". This is different from the traditionally thought influence maximization problem, which can be called "source selection problem", where the problem is to find a set of K nodes that together maximizes their influence over a social network. The very basic assumption there is that all these K nodes can be the source nodes, i.e. can be activated. In "target selection problem" we maximize the influence of a new user as a source node by selecting K nodes in the network and adding a link to each of them. We show that this is the generalization of "source selection problem" and also satisfies the submodularity. The selected nodes are substantially different from those of "source selection problem" and use of the solution of "source selection problem" results in a very poor performance.

Keywords: Information diffusion, influence degree, target node selection

1 Introduction

The emergence of Social Media such as Facebook, Digg and Twitter has provided us with the opportunity to create large social networks, which plays a fundamental role in the spread of information, ideas, and influence. Such effects have been observed in real life, when an idea or an action gains sudden widespread popularity through gword-of-mouthh or gviral marketingh effects. This phenomenon has attracted the interest of many researchers from diverse fields [11], such as sociology, psychology, economy, computer science, etc.

A substantial amount of work has been devoted to the task of analyzing and mining information diffusion processes in large social networks [15, 13, 1]. The main focus

over the past decade has been on optimization problems in which the goal is to maximize the spread of information through a given network, either by selecting a good subset of nodes to initiate the cascade [7] or by applying a broader set of intervention strategies such as node and link additions [18, 21]. Widely used information diffusion models in these studies are *independent cascade (IC)* [7], *linear threshold (LT)* [22] and their variants [8, 19, 6, 20]. These two models focus on different aspects of information diffusion. IC model is sender-centered (information push) and each active node *independently* influences its inactive neighbors with given diffusion probabilities. LT model is receiver-centered (information pull) and a node is influenced by its active neighbors if their total weight exceeds the threshold for the node. Basically the former models diffusion process of how a disease spreads and the latter models diffusion process of how an opinion or innovation spreads.

In this paper we deal with a new type of influence maximization problem. Traditionally this problem is defined to be finding a subset of nodes of size K that maximizes the influence degree with K as a parameter under a given information diffusion model and a given social network. It is unconditionally assumed that the information is guaranteed to start spreading from the selected K nodes. We call this problem as "Source selection problem" to distinguish it from our problem. We rather select K nodes and send information to these nodes. There is no guarantee that these nodes become the information source nodes. Suppose we want to spread our idea or opinion using a twitter, you must acquire reliable followers in the first place. To do this you have to carefully select the target users. Those users who have many followers already may not necessarily be good targets if they have many followees. Our problem is defined to be creating new links to a subset of nodes of size K from a new user such that the influence degree of this user is maximized. We call this problem as "Target selection problem" and analyze it for both LT and IC models.

"Target selection problem" also carries the same problem of 1) computational complexity of estimating influence degree of a given user which is defined to be the expected number of influenced nodes at the end of diffusion process that started from this user and 2) combinatorial explosion of search space in finding the optimal K target nodes. Fortunately, the influence degree is submodular, i.e. its marginal gain diminishes as the size K becomes larger in "Source selection problem", and the greedy solution has a lower bound which is 63% of the true optimal solution [7]. We prove that this submodularity also holds to "Target selection problem", and use a greedy algorithm at the expense of optimality. Various techniques have been devised to reduce the computational cost of solving "Source selection problem". These include bond percolation [9], pruning [8], lazy evaluation [14, 5], burnout [19], heuristics [2, 3], belief propagataion [16] and linear sytem approximation [23]. In this paper we use our own previous work, i.e. bond percolation [9], pruning [8] and burnout [19].

We compare the influence degree of "Target selection problem" with three other methods using four different real social networks. One is to use the solution of "Source selection problem" as target nodes. The other two are to use nodes selected from the largest out-degree and nodes randomly selected. In this paper we show only the results of LT model due to the page limitation. The results clearly show that the solution of "Target selection problem" is different from that of "Source selection problem" and the

influence degree using the solution of "Source selection problem" is only half of the influence degree of "Target selection problem".

2 Information Diffusion Models

We consider a network represented by a directed graph $G = (V, E)$, where V and E ($\subset V \times V$) are the sets of all the nodes and links, respectively. Below we revisit the definition of IC and LT models according to the literatures [7, 10]. In both models the diffusion process proceeds from an initial active node in discrete time-step $t \geq 0$, and it is assumed that nodes can switch their states only from inactive to active (*i.e.*, the SIR setting).

IC model has a *diffusion probability* $p_{u,v}$ with $0 < p_{u,v} < 1$ for each link (u, v) as a parameter. Suppose that a node u first becomes active at time-step t, it is given a single chance to activate each currently inactive child node v, and succeeds with probability $p_{u,v}$. If u succeeds, then v will become active at time-step $t + 1$. If multiple parent nodes of v first become active at time-step t, then their activation trials are sequenced in an arbitrary order, but all performed at time-step t. Whether u succeeds or not, it cannot make any further trials to activate v in subsequent rounds. The process terminates if no more activations are possible.

LT model has a *weight* $q_{u,v}$ (> 0) with $\sum_{u \in B(v)} q_{u,v} \leq 1$ for each link (u, v) as a parameter, where $B(v) = \{u \in V;\ (u, v) \in E\}$ is the set of parent nodes of node v. First, for any node $v \in V$, a *threshold* θ_v is chosen uniformly at random from the interval $[0, 1]$. An inactive node v is influenced by its active parent nodes. If the total weight from the active parent nodes of v at time-step t is at least the threshold θ_v, *i.e.*, $\sum_{u \in B_t(v)} q_{u,v} \geq \theta_v$, then v will become active at time-step $t+1$. Here, $B_t(v)$ stands for the set of all the parent nodes of v that are active at time-step t. The process terminates if no more activations are possible.

For a set of initial active nodes $W(\subset V)$, let $\varphi(W; G)$ denote the number of active nodes at the end of the random process. It is noted that $\varphi(W; G)$ is a random variable. We denote the expected value of $\varphi(W; G)$ by $\sigma(W; G)$, and call it the *influence degree of* W.

3 Target Selection Problem

We first give the formal definition of the source selection problem, or the traditional influence maximization problem [7, 14, 10, 3, 2]. Given a network $G = (V, E)$ and a constant K, the problem is to find a set of K nodes $W_K(\subset V)$ that maximizes the influence degree $\sigma(W_K; G)$, which is formally defined as follows:

$$\operatorname*{argmax}_{W_K \subset V} \sigma(W_K; G). \tag{1}$$

On the other hand, in the target selection problem tackled in this paper, we are given not only a network G and a constant K, but also an external information source node $x \notin V$ and values $\{r_{x,v} \mid v \in V\}$, each associated with link (x, v), where $r_{x,v} \in [0, 1]$ corresponds to a diffusion probability $p_{x,v}$ in case of IC model and a weight $q_{x,v}$ in case

of LT model. Then, we seek a set of K nodes $W_K \subset V$ that maximizes the influence degree of x in an extended network $G'(W_K)$ resulted from adding K links from u to each node $w \in W_K$ into G, which is formally defined as follows:

$$\operatorname*{argmax}_{W_K \subset V} f(W_K), \tag{2}$$

where $f(W_K) = \sigma(\{x\}; G'(W_K))$ and $G'(W_K) = (V \cup \{x\}, E \cup \{(x, w) | w \in W_K\})$. In case of LT model, we assume that each of the original weights to the target nodes, expressed as $q_{v,w}$ where $w \in W_K$ and $v \in B(w)$, is weakened to $(1 - r_{x,w})q_{v,w}$ due to the constraints on weights for LT model. Here we should emphasize that the target selection problem is a natural extension to the source selection problem because we obtain $\operatorname{argmax}_{W_K \subset V} \sigma(W_K; G) = \operatorname{argmax}_{W_K \subset V} f(W_K)$ by setting $r_{x,w} = 1$ for each $w \in W_K$. This is because all of the nodes selected in the target selection problem are definitely activated.

As mentioned in Section 1, since the function σ is submodular, i.e., $\sigma(W' \cup \{v\}; G) - \sigma(W'; G) \geq \sigma(W \cup \{v\}; G) - \sigma(W; G)$ if $W' \subseteq W$, we can approximately solve the source selection problem with a greedy method that recursively finds out W_k based on W_{k-1} by adding node v that maximizes $\sigma(W_{k-1} \cup \{v\}; G)$ to W_{k-1} starting from $W_0 = \emptyset$. Fortunately, in the target selection problem, the function f can be proven to be submodular from the following relation:

$$f(W_K) = \sum_{A \in 2^{W_K}} \sigma(A; G) \prod_{w \in A} r_{x,w} \prod_{w \in (W_K \setminus A)} (1 - r_{x,w}), \tag{3}$$

where 2^{W_K} denotes the power set of W_K. Recall that $r_{x,w}$ corresponds to a diffusion probability $p_{x,w}$ in case of IC model and a weight $q_{x,w}$ in case of LT model. Thus we can easily see that Equation (3) deals with each possible activation pattern A for the target set W_K with the probability that the pattern A happens. Here we should note that in case of LT model, each of the original weights to the target nodes $q_{v,w}$ is weakened to $(1 - r_{x,w})q_{v,w}$. Namely, under the condition that the external source node x fails to activate the target node w, the probability that the node v succeeds to activate the target node w is equivalent to $q_{v,w}$.

From Equation (3), since $f(W_K)$ is a non-negative linear combination of submodular functions $\sigma(\cdot)$, it is also submodular. Thanks to this property, we can solve the target selection problem in the same fashion as the source selection problem with a greedy method. As mentioned earlier, we can efficiently calculate such greedy solutions by using the techniques such as bond percolation [9], pruning [8] and burnout [19].

4 Experiments

Using large real-world networks, we experimentally evaluated the performance of the proposed method for solving the target selection problem on network $G = (V, E)$. We show only the results of LT model due to the page limitations. We chose to show LT model because this model is better suited to opinion spread where we came up with the notion of "target selection".

4.1 Datasets and Settings

In our experiments, we employed four datasets of real networks, where all the networks are represented as directed graphs. The first one is the Ameblo network, which is a reader network of Japanese blog service site "Ameaba" [1] (see [4] for more details). The Ameblo network has 56, 604 nodes and 734, 737 links. The second one is the Blog network, which is a trackback network of Japanese blogs used in [10]. The Blog network has 12, 047 nodes and 53, 315 links. The third one is the Cosme network, which is a fan-link network of "@cosme", [2] a Japanese word-of-mouth communication site for cosmetics (see [17] for more details). The Cosme network has 45, 024 nodes and 351, 299 links. The last one is the Enron network, which is derived from the Enron Email Dataset [12] (see [17] for more details). The Enron network has 19, 603 nodes and 210, 950 links.

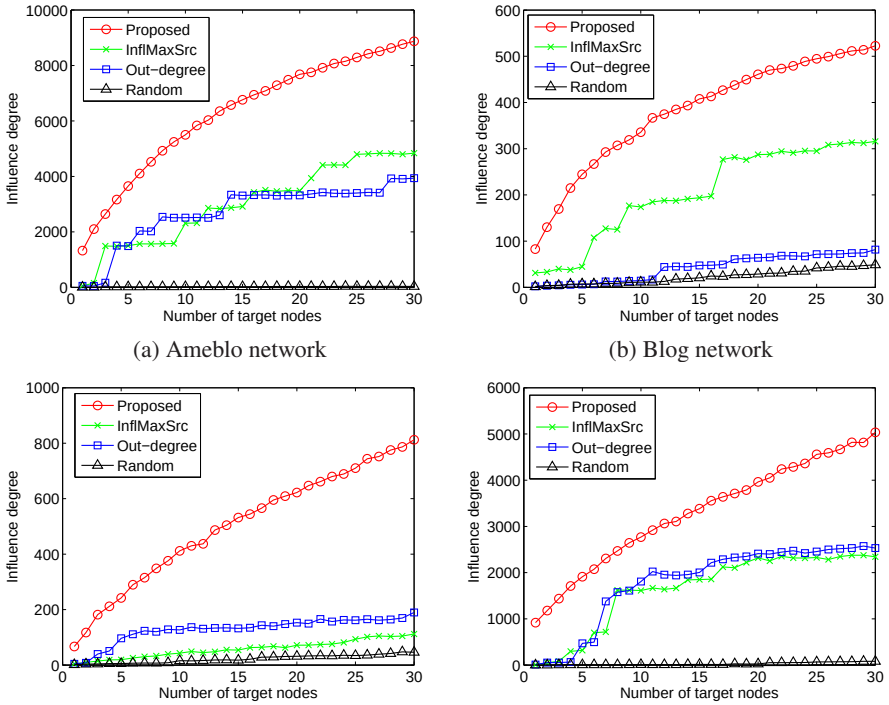
We compared the proposed method with three other heuristic methods as mentioned in Section 1. The first one is to use the solution of the source selection problem for the original network $G = (V, E)$ and add links to the selected K nodes from an external source node. Here, we employed the combined methods of our previous work (bond percolation [9], pruning [8] and burnout [19]). We refer to this method as the *InflMaxSrc* method. The second one is to select nodes in order of decreasing out-degrees, where the out-degree of a node means the number of outgoing links from the node. This is a method often used in the field of complex networks science. We refer to this method as the *Out-degree* method. The third one, which serves as the crude baseline, is to simply select nodes uniformly at random. We refer to this method as the *Random* method.

We evaluated the performance, $f(W_K) = \sigma(\{x\}; G'(W_K))$, where W_K is the selected K nodes by each method. The influence degree $\sigma(\{x\}; G'(W_K))$ was estimated by the empirical mean of the number of active nodes obtained from 10, 000 independent runs of information diffusion, with each run based on the bond percolation [9], pruning [8] and burnout [19]. For the parameters of LT model, we set $q_{u,v} = 1/B(v)$ $(\forall u, v \in V)$.

4.2 Experimental Results

Figures 1a, 1b, 1c and 1d show the results for the Ameblo, Blog, Cosme and Enron networks, respectively. Here, we plot the value of the objective function f (influence degree) as a function of the number k of target nodes, where the circles, crosses, squares and triangles indicate the results for the proposed, InflMaxSrc, Out-degree and Random methods, respectively. First, we see that the proposed method significantly outperformed the InflMaxSrc, Out-degree and Random methods for all four networks. The Random method is by far the worst. We can say that the proposed method can spread the information twice as much as the best of the other two methods can do We also note that the performance of the Out-degree method strongly depends on the characteristics of the network structure, and in some cases it is better than the InflMaxSrc method. We know that the InflMaxSrc method always outperforms the Out-degree and Random methods for the source selection problem (see [10]), but for the target selection problem it is not always the case that the InflMaxSrc method outperforms the Out-degree

[1] http://www.ameba.jp/
[2] http://www.cosme.net/



(a) Ameblo network (b) Blog network

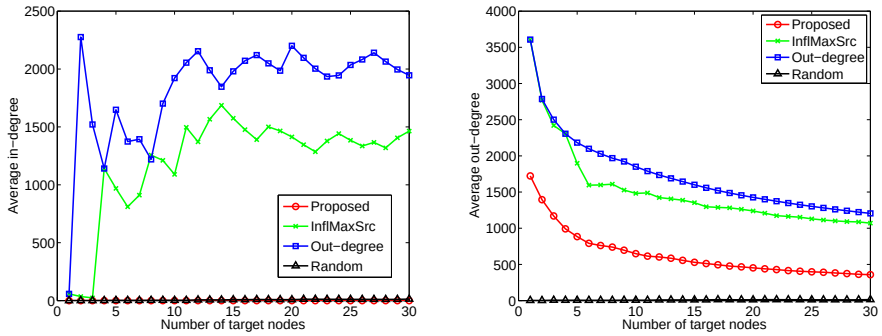
(c) Cosme network (d) Enron network

Fig. 1: Performance comparison for the target selection problem.

method. This is attributed to the fact that the information source node x is not necessarily able to activate all of the target nodes W_*. For instance, the influence degree f of the proposed method is 1.7 to 7.2 times as much as that of the InflMaxSrc method for $k = 30$.

This means that the selected nodes must be substantially different from each other for the four methods. To verify this we measured the solution similarity by F-measure $\mathcal{F}(k) = |W_k^* \cap W_k|/k$, where k stands for the number of target nodes for the target selection problem, and W_k^* and W_k are the solutions extracted by the proposed and one of the other three methods, respectively. The largest F-measure is 0.33 for Amebro network with W_3^* of InflMaxSrc. For the other networks, F-measure is much smaller, e.g., nearly 0 for Cosme network with all k and all other three methods. We confirmed that the proposed method found a solution dramatically different from that by the other three methods.

We next show the in- and out-degrees of the selected nodes in Fig. 2 to investigate why the influence degree achieved by the proposed method is much better than the influence degree by the other methods, *i.e.*, why the selected nodes are different. Here we only show the result of the Ameblo network due to a space limitation, but quite similar results have been obtained from the other networks. From this figure, it is found that, both the in- and out-degrees of the nodes selected by the InflMaxSrc and Out-



(a) Average in-degree (b) Average out-degree

Fig. 2: Average degrees of target nodes for the Ameblo network.

degree methods tend to be high, while the out-degree of the target nodes selected by the proposed method is not so high, but their in-degree is always low. The InflMaxSrc and Out-degree methods select the target nodes independently of their in-degree. This is self-evident for the Out-degree method by definition. In case of the InflMaxSrc method the target nodes are always active at the beginning of the information diffusion process by definition and the in-degree of the target nodes never affects their influence degree. Thus, it tends to select nodes that have many children as the target nodes. It is noted that in the LT model, nodes that have fewer parents have better chance to get activated than those that have many parents. This is because the weights from the parent nodes are larger in the former case and even a small number of active parents can activate the child nodes. Thus, the target node selected by the proposed method are more likely to get activated by the information source node than those selected by the InflMaxSrc and Out-degree methods. This is the main reason of the large difference in the selected target nodes and thus in the resulting influence degree.

5 Conclusion

In this paper we proposed a new type of influence maximization problem, which we call “target selection problem”. Traditionally influence maximization problem assumed unconditionally that the selected nodes can be the source nodes, e.g., can be activated, thus can be called “source selection problem”, and was the simplest model for viral marketing, e.g. which 1000 persons to send direct mails to promote a new product. We thought it more natural and realistic to view this problem from a slightly different angle. We maximized the influence of a new user (source node) who is outside of a community by selection a fixed number of target nodes in the existing community (social network) and adding a link to each of the target nodes. Acceptance of the information of the target nodes from the source node follows a probabilistic information diffusion model as well as the spread of information from the target nodes to the other nodes in the network does so. This “target selection problem” is a generalization of “source selection problem”

and carries similar properties, e.g. submodularity and high computational complexity of estimating influence degree which is the expected number of activated nodes at the end of information diffusion. We estimated the influence degree by the bond percolation and selected target nodes by a greedy algorithm. We solved "target selection problem" in four real world networks, each with slightly different characteristics. Our findings are 1) The solution of "target selection problem" is substantially different from the solution of "source selection problem, 2) Use of the selected nodes of "source selection problem" results in very poor performance (information spread is only half), 3) there is basically no or very few overlap of the nodes selected. This implies that care should be taken in selecting whom to contact first to maximize influence over a social network. We conjecture that such target nodes can be notable mediators, who play an important rolefor widely spreading information. Our immediate future work is to validate this claim using available real information propagation data.

Acknowledgments

This work was partly supported by Asian Office of Aerospace Research and Development, Air Force Office of Scientific Research under Grant No. AOARD-13-4042, and JSPS Grant-in-Aid for Scientific Research (C) (No. 23500194).

References

1. Bakshy, E., Hofman, J., Mason, W., Watts, D.: Everyone's an influencer: Quantifying influences on twitter. In: Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM2011). pp. 65–74 (2011)
2. Chen, W., Wang, C., Wang, Y.: 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 2010). pp. 1029–1038 (2010)
3. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010). pp. 88–97 (2010)
4. Fushimi, T., Saito, K., Kimura, M., Motoda, H., Ohara, K.: Finding relation between pagerank and voter model. In: Proceedings of the 11th International Workshop on Knowledge Management and Acquisition for Smart Systems and Services (PKAW 2012). pp. 208–222 (2010)
5. Goyal, A., Lu, W., Lakshmanan, L.: Influence spread in large-scale social networks - a belief propagation approach. In: Proceedings of the 20th International World Wide Web Conference (WWW2011). pp. 47–48 (2011)
6. Gruhl, D., Guha, R., Liben-Nowell, D., Tomkins, A.: Information diffusion through blogspace. SIGKDD Explorations 6, 43–52 (2004)
7. Kempe, D., Kleinberg, J., Tardos, E.: 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-2003). pp. 137–146 (2003)
8. Kimura, M., Saito, K., Motoda, H.: Efficient estimation of influence functions fot sis model on social networks. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI-09) (2009)

9. Kimura, M., Saito, K., Nakano, R.: Extracting influential nodes for information diffusion on a social network. In: Proceedings of the 22nd AAAI Conference on Artificial Intelligence (AAAI-07). pp. 1371–1376 (2007)
10. Kimura, M., Saito, K., Nakano, R., Motoda, H.: Extracting influential nodes on a social network for information diffusion. Data Mining and Knowledge Discovery 20, 70–97 (2010)
11. Kleinberg, J.: The convergence of social and technological networks. Communications of ACM 51(11), 66–72 (2008)
12. Klimt, B., Yang, Y.: The enron corpus: A new dataset for email classification research. In: Proceedings of the 2004 European Conference on Machine Learning (ECML'04). pp. 217–226 (2004)
13. Leskovec, J., Adamic, L.A., Huberman, B.A.: The dynamics of viral marketing. In: Proceedings of the 7th ACM Conference on Electronic Commerce (EC'06). pp. 228–237 (2006)
14. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007). pp. 420–429 (2007)
15. Newman, M.E.J., Forrest, S., Balthrop, J.: Email networks and the spread of computer viruses. Physical Review E 66, 035101 (2002)
16. Nguyen, H., Zheng, R.: Celf++: optimizing the greedy algorithm for influence maximization in social networks. In: Proceedings of the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012). pp. 515–530. LNAI 7524 (2012)
17. Ohara, K., Saito, K., Kimura, M., Motoda, H.: Effect of in/out-degree correlation on influence degree of two contrasting information diffusion models. In: Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP 2012). pp. 131–138 (2012)
18. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002). pp. 61–70 (2002)
19. Saito, K., Kimura, M., Motoda, H.: Discovering influential nodes for sis models in social networks. In: Proceedings of the Twelfth International Conference of Discovery Science (DS2009). pp. 302–316. Springer, LNAI 5808 (2009)
20. Saito, K., Kimura, M., Ohara, K., Motoda, H.: Learning continuous-time information diffusion model for social behavioral data analysis. In: Proceedings of the 1st Asian Conference on Machine Learning (ACML2009). pp. 322–337. LNAI 5828 (2009)
21. Sheldon, D., Dilkina, B., Elmachtoub, A., Finseth, R., Sabharwal, A., Conrad, J., Gomes, C., Shmoys, D., Allen, W., Amundsen, O., Vaughan, W.: Maximizing the spread of cascades using network design. In: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10). pp. 517–526. AUAI Press, Corvallis, Oregon (2010)
22. Watts, D.J., Dodds, P.S.: Influence, networks, and public opinion formation. Journal of Consumer Research 34, 441–458 (2007)
23. Yang, Y., Chen, E., Liu, Q., Xiang, B., Xu, T., Shad, S.: On approximation of real-world influence spread. In: Proceedings of the 2012 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2012). pp. 548–564. LNAI 7524a (2012)