Influence Maximization by Link Activation in Social Networks

Wenjing Yang, Leonardo Brenner, Alessandro Giua

Influence Maximization by Link Activation in Social Networks

1st Wenjing Yang
Laboratoire d'Informatique et Systèmes
Aix-Marseille Université
Marseille, France
wenjing.yang@lis-lab.fr

2nd Leonardo Brenner
Laboratoire d'Informatique et Systèmes
Aix-Marseille Université
Marseille, France
leonardo.brenner@lis-lab.fr

3rd Alessandro Giua
Department of Electrical and Electronic Engineering
University of Cagliari
Cagliari, Italy
giua@diee.unica.it

***Abstract*—The propagation of innovations in social networks has been widely studied recently. The previous work mostly focuses on either maximizing the influence by identifying a set of initial adopters, or minimizing the influence by link blocking under a certain diffusion model. In our case, we address an influence maximization problem considering the link activation under the Independent Cascade model. For this problem, we propose an approximate solution based on a cost-degree coefficient for searching the relatively effective links to be activated. Evaluated on a real network, our algorithm is experimentally demonstrated to perform well in both aspects of effectiveness and efficiency.**

***Index Terms*—Social networks, Influence maximization, Independent Cascade model, Link activation**

I. Introduction

In recent years, a large number of social network sites have appeared to connect people and groups together. Networks have been proved to be a good tool to obtain information and communicate ideas. Besides, they are becoming an effective marketing platform, through which it is possible to spread information or products to a large scale with a high speed.

The studies on the diffusion of innovations in social networks [1]–[5] began in the middle of the 20th century. Motivated by the application of *viral marketing*, Domingos and Richardson [6] proposed a general framework for the application of data mining and modeled the social network as a Markov random field. Kempe et al. [7] proposed two diffusion models, namely the *Independent Cascade model* and the *Linear Threshold model*, to model the propagation of innovations. Besides, they also formulated the issue of choosing influential sets of individuals as a discrete optimization problem. It aims to identify a small subset of initial adopters in a social network to maximize the influence propagation under a given diffusion model. They also proved this *influence maximization* problem is NP-hard and gave a greedy approximation algorithm which guarantees, under certain conditions, that the influence spread is within $(1 - 1/e)$ of the optimal influence spread. However, this approach requires long time to run the simulation, thus later much effort was devoted to derive more efficient algorithms [8]–[10].

In this work we also study a problem of influence maximization considering in particular the Independent Cascade model. However, unlike previous approaches, we assume that the network is given and the set of initial adopters is randomly selected. The decision variable that we can choose to maximize the influence spread are the active links in the network. Activating a link has a certain cost and we have a limited budget for that.

We believe that the Independent Cascade model can describe quite well the decision of a company that has a given budget for a publicity campaign. The links of the network could represent different ways in which the influence may propagate and the objective of the campaign is supporting the most successful ones so as to maximize the publicity spread.

We control the links in order to maximize the influence spread. Previously, the link control has been used only for influence minimization.

Most of the work which aimed to solve the influence minimization problem by link control applied "link blocking" [11]–[17]. They proposed various approaches to limit the spread of negative innovation such as injection, rumor, virus, etc. in either preventive or reactive way. A preventive strategy focuses on the network topology modification in order to make the network more resistant to a negative innovation. The algorithm proposed by Tong et al. [15] optimized the leading eigenvalue of the network adjacency matrix to control the influence dissemination process. Unlike the preventive strategy, a reactive strategy takes the initially affected nodes into account to guide the link removal operation. Kimura et al. [13], [18] combined the bond percolation method with the greedy algorithm to approximately solve this problem and compared with the link-removal heuristics based on betweenness and out-degree. Nandi et al. [17] proposed mixed-integer programming formulations of four network interdiction models for removing a set of links from a network to minimize the negative influence spread.

To the best of our knowledge, this is the first study on the influence maximization by means of activating links under the Independent Cascade model. Different from other work

This work was supported in part by the China Scholarship Council (CSC), the project PHC CAI YUANPEI 2017 from Campus France under Project No. 38908TK, and the National Natural Science Foundation of China under Grant No. 61703321 and 61672400.

on the influence maximization which focused on selecting a certain number of relatively optimal initial adopters, we aim to activate the most effective links within a limited budget to achieve influence maximization. Unfortunately, the heuristic we consider cannot provide an optimal solution to this problem but, in general, only a suboptimal one. In addition, we evaluate our heuristic on a real network and compare its performance with a random approach.

The rest of the paper is organized as follows. Section 2 reviews the Independent Cascade model and formulates the problem of influence maximization with a limited budget by activating links. Section 3 proposes a cost-degree heuristic to find a suboptimal solution to this problem. Section 4 presents a series of experimental results. We conclude the paper in Section 5.

II. Problem Formulation

In this paper, we assume that the potential links connecting individuals in a network require a cost to be activated. On account of a limited budget, only a certain set of links can be activated. The goal is to maximize the final influence dissemination under a limited budget. We apply the Independent Cascade model to describe the propagation of an innovation through networks.

A. Independent Cascade Model

Consider a directed graph $G = (V, E)$, in which V is the set of nodes and an edge $(i, j) \in E$ denotes that node i influences node j directly [7][1]. In the Independent Cascade model, every edge $(i, j) \in E$ is associated with a *propagation probability* $p : (V \times V) \to (0, 1]$, where $p_{i,j}$ represents the probability that node j is influenced by node i through the edge (i, j) at step k when node i is activated at step $k-1$. Thus, we denote an Independent Cascade model by a triple $G_{IC} = (V, E, p)$.

The *in-neighbors* of node j denoted as $N_j^{in} = \{i \in V | (i, j) \in E\}$ represent all individuals with direct influence on node j. The *out-neighbors* of node j denoted as $N_j^{out} = \{i \in V | (j, i) \in E\}$ represent the individuals on which node j has direct influence.

Let us define ϕ_0 as the *seed set*, i.e., the set of nodes which have adopted the innovation at step $k = 0$. Then the innovation propagates from the seed set. We assume that nodes can switch from being inactive to being active, but can not switch in the other direction. Moreover, every active node has only one chance to influence each of its out-neighbors. If there are many in-neighbors of inactive node j that are activated at step $k-1$, the order in which they attempt to activate node j at step k does not affect the probability of node j being activated [19].

The activation process ends when no more nodes adopt the innovation. The probability that node $j \in V$ is activated during the dynamic evolution is defined as the *activation probability* of node j, denoted as π_j. Assuming the set of active links is $E_a \subseteq E$, the final influence propagation $\sigma(E_a)$ is defined as follow:

[1]In this paper we use the terms individual or node, link or edge interchangeably.

$$\sigma(E_a) = \sum_{j \in V} \pi_j \tag{1}$$

B. Influence Maximization by Link Activation

In this part, we give a mathematical definition of the influence maximization problem considering link activation based on the Independent Cascade model.

In a given Independent Cascade network $G_{IC} = (V, E, p)$, the edges $(i, j) \in E$ are normally inactive but may activated by an external control agent. Assume we are given a *cost vector* $\mathbf{c} \in V \times V$, where *activation cost* $c_{i,j}$ denotes the cost for activating the link between node i and node j. We activate a set of links $E_a \subseteq E$ to construct an *active graph* and the total cost should not exceed a budget $\mathcal{K}$. Then the seed set ϕ_0 is randomly selected from this active graph. The goal is to maximize the final influence spread $\sigma(E_a)$ by the activation of the set of links E_a. We formalize this problem as follow:

Problem 1: Given an Independent Cascade model $G_{IC} = (V, E, p)$, let $\mathbf{c} \in V \times V$ be a cost vector. The seed set ϕ_0, i.e., the initial state of the network, is not given and we assume will be randomly selected at runtime. Activate a set of edges $E_a \subseteq E$, such that the final influence propagation $\sigma(E_a)$ is maximized and the total cost for activating the edges $(i, j) \in E_a$ is no more than a budget $\mathcal{K}$, i.e.,

$$\begin{aligned} max \quad & \sigma(E_a) && (2)\\ s.t. \quad & \sum_{(i,j) \in E_a} c_{i,j} \leq \mathcal{K} && (3)\\ & E_a \subseteq E && (4)\\ & \mathcal{K} \in \mathbb{R}_+ && (5) \end{aligned}$$

Note that the seed set ϕ_0 is not specified in Problem 1, but randomly selected in the active graph generated by the set of links $E_a \subseteq E$. In fact, this is a stochastic optimization problem where both the input data (the seed set ϕ_0) and the system's performance index (the final influence spread) are random variables.

III. Methodology

In this section, we propose an approach considering both propagation probability and activation cost for solving this problem.

Our goal is to maximize the final influence propagation on the condition that the total cost for link activation is no more than a given budget. Thus, the edges with big propagation probability but small activation cost are considered to be activated firstly. Denote $\Theta_{i,j} = p_{i,j}/c_{i,j}$ as the *cost-degree* of link (i, j), then select the link (i, j) with the maximum value of $\Theta_{i,j}$ to be activated in each iteration. The iteration process stops when the total cost for link activation does not satisfy the budget constraint. The procedure in detail is shown in Algorithm 1.

The time complexity of Algorithm 1 is $O(m)$ with $m = |E|$ (the number of edges) since the algorithm computes $\Theta_{i,j}$ for each edge $(i, j) \in E$.

Algorithm 1 Cost-Degree Algorithm

Input: An Independent Cascade network $G_{IC} = (V, E, p)$; a cost vector **c**; a budget $\mathcal{K}$

Output: An active edge set E_a

Initialize $E_a = \emptyset$

Compute $\Theta_{i,j} = p_{i,j}/c_{i,j}$ for $\forall (i,j) \in E$

while $\sum_{(i,j)\in E_a} c_{i,j} \leq \mathcal{K}$ **do**

  $E_a = E_a \cup \{ \underset{(i,j)\in E\setminus E_a}{argmax} (\Theta_{i,j}) \}$

end while

return E_a

IV. Experiment

Based on a real-world network, we experimentally evaluate the performance of the proposed heuristic.

A. Data Sets

We use a real-world dataset—**airportsinUS**[2], which is a benchmark network widely used in social network analysis [20]. It is a weighted network of the 500 airports with the largest amount of traffic from publicly available data in the United States. Nodes represent US airports and edges represent air travel connections among them. There are 5960 edges in total. Based on the weights $w_{i,j}$ of edges, we obtain $p_{i,j}$ by $w_{i,j}/\sum_i w_{i,j}$. Besides, activation cost $c_{i,j}$ is uniformly selected from the interval $(0, 10)$ at random.

B. Experiment Setup

We compare the proposed heuristic based on the cost-degree coefficient with a random method. Different budget values $\mathcal{K} = \{1000, 5000, 10000, 15000, 20000, 25000\}$ are considered in our experiment. Furthermore, the seed sets in different sizes $|\phi_0| = \{1, 5, 10, 15, 20, 25, 30\}$ are randomly selected in the active graphs generated by either of the two considered algorithms under a certain budget value. The tested algorithms are briefly described as following:

Random: Randomly select a set of edges to be activated unless the total cost is more than the budget limit $\mathcal{K}$.

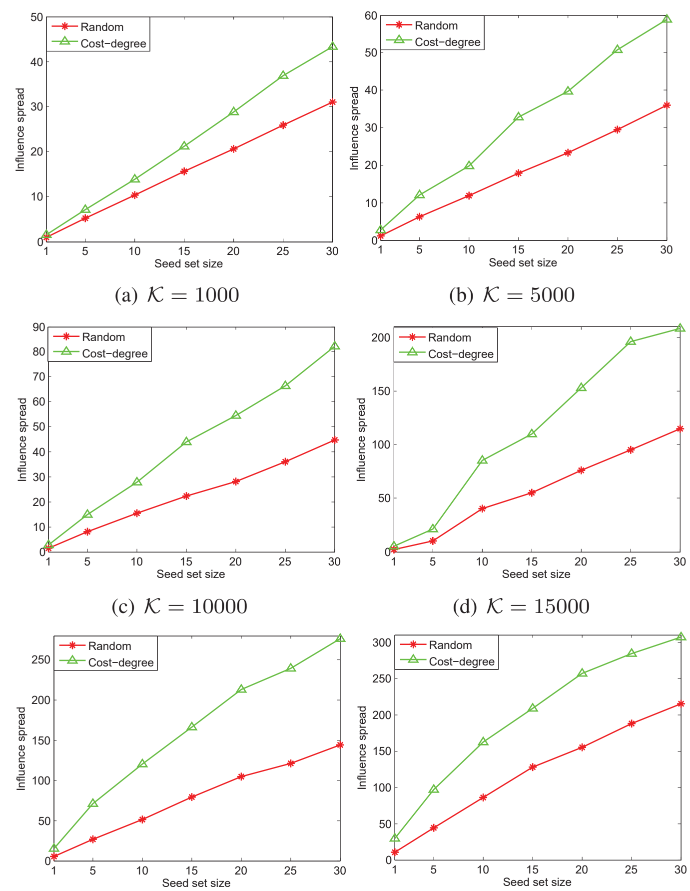
Cost-degree: Our cost-degree algorithm proposed in Section 3.

All approaches are implemented in MATLAB. All experiments are run on a PC with 2.40GHz Intel Core i5 Processor and 8GB memory.

C. Experiment Results

We evaluate the algorithms on the airportinUS network under the Independent Cascade model in terms of the influence spread and the running time. The influence spread is denoted with the total activation probabilities of all nodes in the network. After selecting the set of links to be activated by either of these two algorithms, the influence spread is computed approximately by the *SteadyStateSpread* algorithm [21], [22]. The algorithm terminates when the aggregate change in the absolute probabilities between two consecutive iterations is less than 0.001. We simulate the process 100 times by randomly re-selecting the seed set ϕ_0 in a certain size from the active graphs generated by either of the two algorithms. Then limited by a budget $\mathcal{K}$, the final influence spread $\sigma(E_a)$ with each size of seed set is computed by the average of the results in these 100 times simulations.

The final influence spread given different sizes of seed sets under the budget values $\mathcal{K} = \{1000, 5000, 10000, 15000, 20000, 25000\}$ is shown in Fig. 1. Obviously, the influence spread increases with the budget since more links can be activated. Moreover, We can observe that our proposed algorithm performs much better than the random algorithm under the same budget value. It is more clear in Fig. 2.



(a) $\mathcal{K} = 1000$ (b) $\mathcal{K} = 5000$

(c) $\mathcal{K} = 10000$ (d) $\mathcal{K} = 15000$

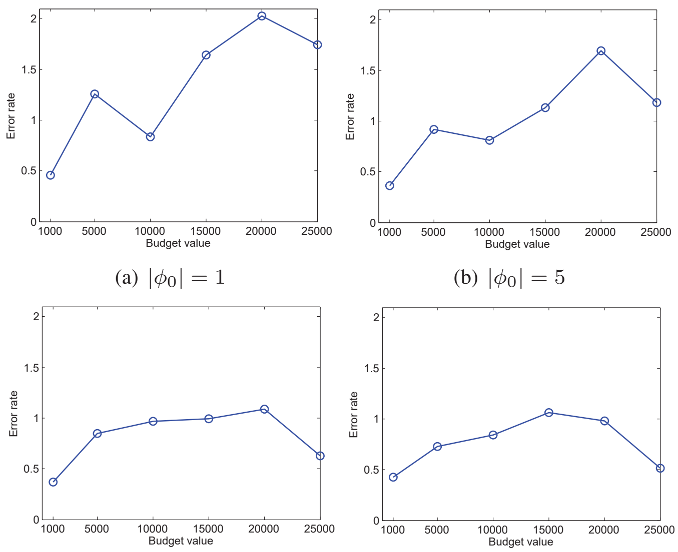
(e) $\mathcal{K} = 20000$ (f) $\mathcal{K} = 25000$

Fig. 1. Influence spread of the tested algorithms given different budget values

Fig. 2 shows the error rate of influence spread after the link activation by the two algorithms with $|\phi_0| = \{1, 5, 15, 25\}$, i.e., $(\sigma(E_a)^{cost-degree} - \sigma(E_a)^{random})/\sigma(E_a)^{random}$ with a same seed set size and budget value. We can observe that the cost-degree algorithm can choose a more successful set of active links than the random approach regardless of the budget value and the seed set size.

Secondly, we compare the running time for selecting the set of links to be activated by the two algorithms under different

[2] https://sites.google.com/site/cxnets/usairtransportationnetwork



(a) $|\phi_0| = 1$ (b) $|\phi_0| = 5$

(c) $|\phi_0| = 15$ (d) $|\phi_0| = 25$

Fig. 2. Error rate of influence spread after link activation by the tested algorithms given different seed set sizes

budget values $\mathcal{K}$, shown in Table I. Both tested algorithms take a bit longer time when the budget $\mathcal{K}$ increases. Besides, our algorithm almost takes as the same little time as the random algorithm under a same budget value, while it can achieve a much larger influence spread.

TABLE I
RUNNING TIME FOR SELECTING LINKS TO BE ACTIVATED UNDER DIFFERENT BUDGET VALUES

Running time (s)	$\mathcal{K} = 1000$	$\mathcal{K} = 5000$	$\mathcal{K} = 10000$	$\mathcal{K} = 15000$	$\mathcal{K} = 20000$	$\mathcal{K} = 25000$
Random	0.02	0.09	0.11	0.17	0.20	0.24
Cost-degree	0.04	0.09	0.16	0.22	0.24	0.29

V. Conclusion

Most previous work focuses on the approaches to either influence maximization considering the initial adopters or influence minimization based on link blocking. Differently in this paper, we formulate an influence maximization problem within a limited budget considering to activate the relatively most effective links. For approximately solving this problem, we propose a heuristic associated with a cost-degree coefficient. Moreover, experiments on a real network show that our approach works effectively and efficiently.

There are several future directions for this work. First, the propagation probability and cost vector for our current diffusion model are fixed and unrelated. We plan to generalize our model in terms of associating these two parameters together in order to better describe the real-world scenarios. Second, our cost-degree algorithm does not consider the seed set during the identification of links to be activated. We aim to propose new approaches which involve the whole diffusion process while selecting the links to be activated.

References

[1] B. Ryan and N. C. Gross, "The diffusion of hybrid seed corn in two iowa communities." *Rural sociology*, vol. 8, no. 1, p. 15, 1943.

[2] J. S. Coleman, E. Katz, and H. Menzel, *Medical innovation: A diffusion study*. Bobbs-Merrill Co, 1966.

[3] N. E. Friedkin and E. C. Johnsen, "Social influence networks and opinion change," *Advances in Group Processes*, vol. 16, pp. 1–29, 1999.

[4] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England journal of medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[5] Y. Bouanan, G. Zacharewicz, J. Ribault, and B. Vallespir, "Discrete event system specification-based framework for modeling and simulation of propagation phenomena in social networks: application to the information spreading in a multi-layer social network," *SIMULATION*, 2018.

[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*. ACM, 2001, pp. 57–66.

[7] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[8] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 420–429.

[9] 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*. ACM, 2009, pp. 199–208.

[10] C. Zhou, P. Zhang, J. Guo, X. Zhu, and L. Guo, "Ublf: An upper bound based approach to discover influential nodes in social networks," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 907–916.

[11] E. A. Enns and M. L. Brandeau, "Link removal for the control of stochastically evolving epidemics over networks: A comparison of approaches," *Journal of theoretical biology*, vol. 371, pp. 154–165, 2015.

[12] J. He, H. Liang, and H. Yuan, "Controlling infection by blocking nodes and links simultaneously," in *International Workshop on Internet and Network Economics*. Springer, 2011, pp. 206–217.

[13] M. Kimura, K. Saito, and H. Motoda, "Blocking links to minimize contamination spread in a social network," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 3, no. 2, p. 9, 2009.

[14] J. Marcelino and M. Kaiser, "Critical paths in a metapopulation model of h1n1: Efficiently delaying influenza spreading through flight cancellation," *PLoS currents*, vol. 4, 2012.

[15] H. Tong, B. A. Prakash, T. Eliassi-Rad, M. Faloutsos, and C. Faloutsos, "Gelling, and melting, large graphs by edge manipulation," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 245–254.

[16] C. J. Kuhlman, G. Tuli, S. Swarup, M. V. Marathe, and S. Ravi, "Blocking simple and complex contagion by edge removal," in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 399–408.

[17] A. K. Nandi and H. R. Medal, "Methods for removing links in a network to minimize the spread of infections," *Computers & Operations Research*, vol. 69, pp. 10–24, 2016.

[18] M. Kimura, K. Saito, and H. Motoda, "Minimizing the spread of contamination by blocking links in a network." in *AAAI*, vol. 8, 2008, pp. 1175–1180.

[19] D. Rosa and A. Giua, "On the spread of innovation in social networks," *IFAC Proceedings Volumes*, vol. 46, no. 27, pp. 322–327, 2013.

[20] V. Colizza, R. Pastor-Satorras, and A. Vespignani, "Reaction–diffusion processes and metapopulation models in heterogeneous networks," *Nature Physics*, vol. 3, no. 4, p. 276, 2007.

[21] C. C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks," in *Proceedings of the 2011 SIAM International Conference on Data Mining*. SIAM, 2011, pp. 522–533.

[22] W. Yang, L. Brenner, and A. Giua, "Computation of activation probabilities in the independent cascade model," in *Control, Decision and Information Technologies (CoDIT), 2017 International Conference on*. IEEE, 2017.