Influence Maximization in the Independent Cascade Model

Gianlorenzo D'Angelo, Lorenzo Severini, and Yllka Velaj

Gran Sasso Science Institute (GSSI), Viale F. Crispi, 7, 67100, L'Aquila, Italy.
{gianlorenzo.dangelo,lorenzo.severini,yllka.velaj}@gssi.infn.it

Abstract. We present our ongoing work on the problem of increasing the information spread in a network by creating a limited amount of new edges incident to a given initial set of active nodes. As a preliminary result, we give a constant approximation algorithm for the case in which the set of initial active nodes is a singleton. Our aim is to extend this result to the general case. We outline some further research directions which we are investigating.

1 Introduction

Studying the processes by which ideas and influence propagate through a network has been one of the main goals in the field of social network analysis. The influence problem is motivated by many applications in different fields: from marketing, with the aim of maximizing the adoption of a new product [3], to epidemiology, in order to limit the diffusion of a virus or disease [11], going through the analysis of social networks to find influential users and to study how information flows through the network [1].

Different models of information diffusion have been introduced in the literature [5], two widely studied models are: the *Linear Threshold Model* (LTM) and the *Independent Cascade Model* (ICM). In both models, we can distinguish between active, or infected, nodes, called seeds, which spread the information, and inactive ones. Recursively, currently infected nodes can infect their neighbours with some probability. After a certain number of such cascading cycles, a large number of nodes becomes infected in the network. In LTM the idea is that a node becomes active if a large part of its neighbours is active. More formally, each node u has a threshold t chosen uniformly at random in the interval $[0, 1]$. The threshold represents the fraction of neighbours of u that must become active in order for u to become active. At the beginning of the process a small percentage of nodes of the graph is set to active in order to let the information diffusion

V. Biló, A. Caruso (Eds.): ICTCS 2016, Proceedings of the 17th Italian Conference on Theoretical Computer Science, 73100 Lecce, Italy, September 7–9 2016, pp. 269–274 published in CEUR Workshop Proceedins Vol-1720 at http://ceur-ws.org/Vol-1720

process start. In subsequent steps of the process a node becomes active if the fraction of its active neighbours is greater than its threshold. In ICM, instead, a seed u tries to influence one of its inactive neighbours but the success of node u in activating the node v only depends on the propagation probability of the edge from u to v (each edge has its own value). Regardless of its success, the same node will never get another chance to activate the same inactive neighbour. The process terminates when no further node gets activated.

An interesting question, in the analysis of the information spread through a network, is how to shape a given diffusion process so as to maximize or minimize the number of activated nodes at the end of the process by taking intervention actions. Many intervention actions have been studied in the literature, the most important one is: if we are allowed to add at most k seeds, which ones should be selected so as to maximize the number of active nodes resulting from the diffusion process [5]. Besides source selection, other intervention actions may be used to facilitate or limit the diffusion processes, such as inserting or deleting edges and adding or deleting nodes in the network.

To the best of our knowledge, under LTM, the problems that have been studied are the following: Khalil et al. [6] consider two types of actions, adding edges to or deleting edges from the existing network and they show that this network structure modification problem has a supermodular objective and therefore can be solved by algorithms with provable approximation guarantees. Zhang et al. [15] consider arbitrarily specified groups of nodes, and interventions that involve both edge and node removal from the groups. They develop algorithms with rigorous performance guarantees and good empirical performance. Kimura et al. [7] use a greedy approach to delete edges under the LTM without any analysis of the supermodularity of the objective, nor rigorous approximation guarantees. Kuhlman et al. [9] propose heuristic algorithms for edge removal under a simpler deterministic variant of LTM which is not only hard, but also has no approximation guarantee. Papagelis [12] and Crescenzi et al. [4] study the problem of augmenting the graph in order to increase the connectivity or the centrality of a node, respectively and experimentally show that this increases the expected number of eventual active nodes. Under ICM, the main results are the following: Wu et al. [14] consider intervention actions other then edge addition, edge deletion and source selection, such as increasing the probability that a node infects its neighbours. It can be shown that optimizing the selection of such actions with a limited budget tends to be NP-hard and is neither submodular nor supermodular. Sheldon et al. [13] study the problem of node addition to maximize the spread of information, and provide a counterexample showing that the objective function is not submodular. Bogunovic [2] addresses the node deletion problem providing a greedy algorithm. Kimura et al. [8] propose methods for efficiently finding good approximate solutions on the basis of a greedy strategy for the edge deletion problem under the ICM, but do not provide any approximation guarantees.

In this paper, we focus on the Independent Cascade Model and investigate the problem of adding a small number of edges incident to an arbitrary seed in order to increase the spreading of information in terms of number of nodes that become active. Thus, the problem we analyse differs from above mentioned ones and, as far as we know, similar problems have never been studied for the Independent Cascade Model.

The aim of this paper is reporting our ongoing research on which we wish to get feedback so as to possibly include these results in future publications.

2 Preliminary results

In this section we will give all the necessary definitions, introduce the problem that will be considered and show our preliminary results.

A social network is represented by a weighted directed graph $G(V,$ âĂĚ$E, p)$ where V represents the set of nodes, E represents set of relationships and $p : V \times V \to [0, 1]$ is the probability of an edge to propagate information. For each node u, N_u denotes the set of neighbours of u, i.e. $N_u = \{v|(u, v) \in E\}$.

The Independent Cascade Model [5] is an information diffusion model where the information flows over the network through cascade. Nodes can have two states, active: it means the node is already influenced by the information in diffusion, inactive: node is unaware of the information or not influenced. The process runs in discrete steps. At the beginning of ICM process, few nodes are given the information, they are known as seed nodes. Upon receiving the information these nodes become active. In each discrete step, an active node tries to influence one of its inactive neighbours. Regardless of its success, the same node will never get another chance to activate the same inactive neighbour. The success of node u in activating the node v depends on the propagation probability of the edge (u, v) defined as p_{uv}, each edge has its own value. The process terminates when no further node gets activated.

We define the influence of a set $A \subseteq V$ in the graph G, denoted by $\sigma(A, G)$, to be the expected number of active nodes at the end of the process, given that A is the initial set of seeds. Given a set S of edges not in E, we denote by $G(S)$ the graph augmented by adding the edges in S to G, i.e. $G(S) = (V, E \cup S)$.

Given a graph $G = (V, E)$, a vertex set $A \subseteq V$ and an integer k, the problem we are studying consists in finding a set S of edges incident to the nodes in A not in E (that is, $S \subseteq \{(a, v) : v \in V \setminus N_a, a \in A\}$) such that $|S| \leq k$ and $\sigma(A, G(S))$ is maximum.

In the paper we focus on the case $A = \{a\}$. We leave the case $|A| > 1$ as a future work. It has been shown [10] that for a monotone submodular function the following greedy algorithm provides a $\left(1 - \frac{1}{e}\right)$-approximation: start with the empty set and repeatedly add an element that gives the maximal marginal gain. The greedy algorithm can be extended to any monotone submodular objective function thanks to the following result.

Theorem 1 ([10]). *For a non-negative, monotone submodular function f, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the value of f. Then S provides a $\left(1 - \frac{1}{e}\right)$-approximation.*

In this paper, we exploit this result by showing that $\sigma(A, G(S))$ is monotone and submodular w.r.t. the possible set of edges incident to a.

Theorem 2. *$\sigma(A, G(S))$ is a monotonically increasing submodular function of the set S of edges to be added.*

Proof (sketch). We will use the definition of live-edge graph $X = (V, E_X)$ which is a directed graph where the set of nodes is equal to V and the set of edges is a subset of E. E_X is given by a edge selection process such that each edge is either live or blocked according to its propagation probability. We can assume that for each pair of neighbours in the graph, a coin of bias p_{uv} is flipped and the edges for which the coin indicated an activation are live, the remaining are blocked. It is easy to show that a diffusion model is equivalent to the reachability problem in live-edge graphs: given any seed set A, the distribution of active node sets after the diffusion process ends is the same as the distribution of node sets reachable from A in a live-edge graph.

We denote with $\chi(G)$ the probability space in which each sample point specifies one possible set of outcomes for all the coin flips on the edges, it is the set of all possible live-edge graphs. Let $R(A, X)$ denote the set of all nodes that can be reached from the nodes in A on a path consisting entirely on live edges: $R(A, X) = \bigcup_{a \in A} R(a, X)$.

The main idea to prove that the function in monotonically increasing is that, after an edge addition in G, the live-graph X has at least one more edge than the original live-edge graph, hence, the number of reachable nodes can not decrease. To prove submodularity, we note that the number of new reachable nodes from the seed after the edge addition in $G(T)$ is smaller or equal than the number of new reachable nodes in $G(S)$ since most of the nodes are already reachable by the edges in $T \setminus S$. We prove these conditions for all $X \in \chi(G)$. □

Note that, in the problem we are studying, the greedy algorithm can not evaluate the influence function exactly since $\sigma(A, G(S))$ is the expected number of activated nodes and it has been proven that evaluating this function is generally $\#P$-complete for ICM [3]. However, by simulating the diffusion process sufficiently many times and sampling the resulting active sets, it is possible to obtain arbitrarily good approximations to $\sigma(A, G(S))$ (see Prop 4.1 in [5] to bound the number of samples needed to obtain a $(1 + \delta)$-approximation). It is an extension of the result of Nemhauser et al. [10] that by using $(1 \pm \delta)$-approximate values for the function to be optimized where $\delta \geq 0$, we obtain $\left(1 - \frac{1}{e} - \epsilon\right)$-approximation, where ϵ depends on δ and goes to 0 as $\delta \to 0$.

Theorem 3. *For the problem of adding a set S of edges, not in E, incident to the node in $A = \{a\}$ such that $|S| \leq k$ and $\sigma(A, G(S))$ is maximum, there is a polynomial-time algorithm approximating the maximum influence to within a factor of $\left(1 - \frac{1}{e} - \epsilon\right)$ where ϵ is any positive real number.*

3 Future research

In this paper, we presented our ongoing work on the problem of increasing the information spread in a network considering the case in which the set of active nodes A is a singleton. We have analysed the properties of the influence function which is monotonically increasing and submodular and we propose a greedy approximation algorithm for efficiently computing a set of edges that a seed can decide to add to the graph in order to increase the expected number of influenced nodes. As future works, we plan to extend our approach to $|A| > 1$ and consider the insertion of edges incident to all the seeds in A. Moreover, we plan to analyse a generalization of the problem considered in this paper by allowing the deletion of edges incident to seeds. Finally, our intent is to study the same problem in a generalization of ICM, which is the Decreasing Cascade Model. In this model the probability of a node u to influence v is non-increasing as a function of the set of nodes that have previously tried to influence v. From the experimental point of view, our aim is to measure the efficiency of the greedy algorithm in term of expected number of influenced nodes.

References

1. E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proc. of the Fourth ACM Int. Conf. on Web Search and Data Mining*, WSDM '11, pages 65–74. ACM, 2011.
2. I. Bogunovic. Robust protection of networks against cascading phenomena, 2012.
3. W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2010.
4. P. Crescenzi, G. D'angelo, L. Severini, and Y. Velaj. Greedily improving our own closeness centrality in a network. *ACM Trans. Knowl. Discov. Data*, 11(1):9:1–9:32, 2016.
5. D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. *Theory of Computing*, 11:105–147, 2015.
6. E. B. Khalil, B. Dilkina, and L. Song. Scalable diffusion-aware optimization of network topology. In *Proc. of the 20th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, 2014.
7. M. Kimura, K. Saito, and H. Motoda. *Solving the Contamination Minimization Problem on Networks for the Linear Threshold Model.* 2008.
8. M. Kimura, K. Saito, and H. Motoda. Blocking links to minimize contamination spread in a social network. *ACM Trans. Knowl. Discov. Data*, 2009.
9. C. J. Kuhlman, G. Tuli, S. Swarup, M. V. Marathe, and S. Ravi. Blocking simple and complex contagion by edge removal. In *Proc. of the IEEE Int. Conf. on Data Mining (ICDM)*, 2013.
10. G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of approximations for maximizing submodular set functions–I. *Math. Program.*, 1978.
11. M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, Jul 2002.
12. M. Papagelis. Refining social graph connectivity via shortcut edge addition. *ACM Trans. Knowl. Discov. Data*, 10(2):12, 2015.

13. D. Sheldon, B. N. Dilkina, A. N. Elmachtoub, R. Finseth, A. Sabharwal, J. Conrad, C. P. Gomes, D. B. Shmoys, W. Allen, O. Amundsen, and W. Vaughan. Maximizing the spread of cascades using network design. *CoRR*, 2012.
14. X. Wu, D. Sheldon, and S. Zilberstein. Efficient algorithms to optimize diffusion processes under the independent cascade model. In *NIPS Work. on Networks in the Social and Information Sciences*, 2015.
15. Y. Zhang, A. Adiga, A. Vullikanti, and B. A. Prakash. Controlling propagation at group scale on networks. In *Proc. of the IEEE Int. Conf. on Data Mining (ICDM)*, 2015.