Activity-Driven Influence Maximization in Social Networks

Rohit Kumar[1,4], Muhammad Aamir Saleem[1,2], Toon Calders[1,3], Xike Xie[5], and Torben Bach Pedersen[2]

[1]Université Libre de Bruxelles, Belgium; [2]Aalborg University, Denmark; [3]Universiteit Antwerpen, Belgium; [4]Universitat Politécnica de Catalunya (BarcelonaTech), Spain; [5]University of Science and Technology of China, China

Abstract. Interaction networks consist of a static graph with a time-stamped list of edges over which interaction took place. Examples of interaction networks are social networks whose users interact with each other through messages or location-based social networks where people interact by checking in to locations. Previous work on finding influential nodes in such networks mainly concentrate on the static structure imposed by the interactions or are based on fixed models for which parameters are learned using the interactions. In two recent works, however, we proposed an alternative activity data driven approach based on the identification of influence propagation patterns. In the first work, we identify so-called information-channels to model potential pathways for information spread, while the second work exploits how users in a location-based social network check in to locations in order to identify influential locations. To make our algorithms scalable, approximate versions based on sketching techniques from the data streams domain have been developed. Experiments show that in this way it is possible to efficiently find good seed sets for influence propagation in social networks.

1 Introduction

Understanding how information propagates in a network has a broad range of applications like viral marketing [6], epidemiology and outdoor marketing [7]. For example, imagine a computer games company that has budget to hand out samples of their new product to 50 gamers, and want to do so in a way that achieves maximal exposure. In that situation the company would like to target those customers that have maximal influence on social media. For this purpose they monitor interactions between gamers, and learn from these interactions which ones are the most influential. Notice that for the company it is also important that the selected people are not only influential, but that their combined influence should be maximal; selecting 50 highly influential gamers in the same sub-community is likely less effective than targeting less influential users but from different communities. This example is a typical instance of the *Influence maximization problem* [6]. The common ingredients of an influence maximization problem are: a graph in which the nodes represent users of a social network, an

information propagation model, and a target number of seed nodes that need to be identified such that they jointly maximize the influence spread in the network under the given propagation model.

Earlier works in this area studied different propagation models, such as linear threshold (LT) or independent cascade (IC) models [3], the complexity of the influence maximization problem under these models, and efficient heuristic algorithms. For instance, Kempe et al. [3] proved that the influence maximization problem under the LT and IC models is NP-hard and they provided a greedy algorithm to select seed sets using maximum marginal gain. As the model was based on Monte Carlo simulations it was not very scalable for large graphs.

A critical issue in the application of influence maximization algorithms is that of selecting the right propagation model. Most of these propagation models rely on parameters such as the influence a user exerts on his/her neighbors. Therefore, a second important line of work deals with learning these parameters based on observations. For instance, in a social network we could observe that user a liking a post is often followed by user b, who is friend of a, liking the same post. In such a case it is plausible that user a has a high influence on user b, and hence that the parameter expressing the influence of a on b should get a high value. The parameter learning problem is hence, based on a record of activities in the network, estimate the most likely parameter setting for explaining the observed propagation. The resulting optimized model can then be used to address the problem of selecting the best seed nodes. Goyal et al. [2] proposed the first such data based approach to find influential users in a social network. They estimate the influence probability of one user on his/her friend by observing all the different activities the friend follows in a given time window divided by total number of activities done by the user.

All these works share one property: they are based on models and if activity data is used, it is only indirectly to estimate model parameters. Recently, however, new, model-independent and purely data-driven methods have emerged. Our two papers, [7] published at WSDM and [4] published at EDBT should be placed in this category of data-based approaches.

2 Data-Driven Information Maximization

In [4] we proposed a new time constrained model to consider real interaction data to identify influence of every node in an interaction network [5]. The central idea in our approach is to mine frequent *information channels* between different nodes and use the presence of an information channel as an indication of possible influence among the nodes. An *information channel*($ic(u, v)$) is a sequence of interactions between nodes u and v forming a path in the network which respects the time order. As such, an information channel represents a potential way information could have flown in the interaction network. An interaction could be bidirectional, for instance a chat or call between two users where information flows in both directions, or uni-directional where information flows from one user to another, for example in an email interaction or a re-tweet.

Figure 1 illustrates the notion of an information channel. There are interactions from user $a \to b$ and $c \to e$ at 9 AM, from $b \to d$ and $b \to c$ at 9:05 AM and $d \to f$ at 9:10 AM. These interactions form an interaction network. There is an information channel $a \to c$ via the temporal path $a \to b \to c$ but there is no information channel from $a \not\to e$ as there is no time respecting path from a to e. We define the *duration*$(dur(ic(u, v)))$ of an information channel as the time difference of the first and last interaction on the information channel. For example, the duration of the information channel $a \to b \to c$ is 10 minutes. There could be multiple information channels of different durations between two nodes in a network.The intuition of the information channel notion is that node u could only have sent information to node v if there exists a time respecting series of interactions connecting these two nodes. Therefore, nodes that can reach many other nodes through information channels are more likely to influence other nodes than nodes that have information channels to only few nodes. This notion is captured by the *influence reachability set*. The *Influence reachability set (IRS)* $\sigma(u)$ of a node u in a network $G(V, \mathcal{E})$ is defined as the set of all the nodes to which u has an information channel

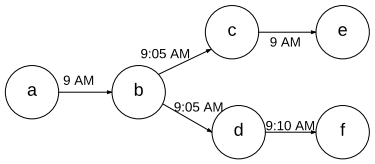

Fig. 1. Information channels between different nodes in the network. Every node is a user in a social network and the edges represents an interaction between users.

In [4] we presented a one-pass algorithm to find the IRS of all nodes in an interaction network. We developed a time-window based HyperLogLog sketch [1] to compactly store the IRS of all the nodes and provided a greedy algorithm to do influence maximization.

3 Finding Influential Locations

Outdoor marketing can also benefit from the same data based approach to maximize influence spread [7]. Recently, with the pervasiveness of location-aware devices, social network data is often complemented with geographical information, known as location-based social networks (LBSNs). In [7] we study navigation patterns of users based on LBSN data to determine influence of a location on another location. Using the LBSN data we construct an interaction graph with nodes as locations and the edges representing the users traveling between locations. For example, in Fig-

Check-ins

loc	Users t=1	t=2	t=3
T_1	b, c, e, f	a, h	f
T_2	a, h	f, g	a
M_1	g	i	d
H_1	–	b, c, d, e	i
H_2	d, i	–	–

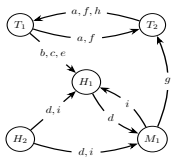

Fig. 2. Running example of a LBSN [7].Nodes in the graph are the locations visited by users a-h. Edges are the movement of user between locations in a time window.

ure 2 there is an edge from location T_1 to T_2 due to users a and f visiting both locations within one trip.

We define the influence of a location by its capacity to spread its visitors to other locations. The intuition behind this definition is that good locations to seed with messages such as outdoor marketing promotions, are locations from which its visitors go to many other locations thus spreading the message. Thus, location influence indirectly captures the capability of a location to spread a message to other geographical regions. For example, if a company wants to distribute free t-shirts to promote some media campaign in a city, it would get maximum exposure by selecting neighborhoods such that the visitors of these neighborhood spread to maximum other neighborhoods in the city. In [7] we provide an exact on-line algorithm and a more memory-efficient but approximate variant based on the HyperLogLog sketch to maintain a data structure called *Influence Oracle* that allows to greedily find a set of influential locations.

4 Conclusion

In both of our works, through simulation experiments, we have shown that the data driven approach is quite accurate in modeling influence spread in the network. We also used time window based variations of the HyperLogLog sketch as an alternative to capture the influence set of every node in the network enabling us to scale our algorithms to very high data volumes.

Acknowledgement

This work was supported by the Fonds de la Recherche Scientifique-FNRS under Grant(s) n T.0183.14 PDR. Xike Xie is supported by the CAS Pioneer Hundred Talents Program and the Fundamental Research Funds for the Central Universities

References

1. Flajolet, P., Fusy, É., Gandouet, O., Meunier, F.: Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm. DMTCS Proceedings (2008)
2. Goyal, A., Bonchi, F., Lakshmanan, L.V.: A data-based approach to social influence maximization. Proceedings of the VLDB Endowment 5(1), 73–84 (2012)
3. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD (2003)
4. Kumar, R., Calders, T.: Information propagation in interaction networks. In: EDBT (2017)
5. Kumar, R., Calders, T., Gionis, A., Tatti, N.: Maintaining sliding-window neighborhood profiles in interaction networks. In: ECML/PKDD (2015)
6. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: KDD (2002)
7. Saleem, M.A., Kumar, R., Calders, T., Xie, X., Pedersen, T.B.: Location influence in location-based social networks. In: WSDM (2017)