Crossing The Boundaries of Communities via Limited Link Injection for Information Diffusion In Social Networks

Dimitrios Rafailidis
Aristotle University of Thessaloniki
Thessaloniki, Greece
draf@csd.auth.gr

Alexandros Nanopoulos
Catholic University of Eichstät-Ingolstadt
Ingolstadt, Germany
alexandros.nanopoulos@ku.de

ABSTRACT

We propose a new link-injection method aiming at boosting the overall diffusion of information in social networks. Our approach is based on a diffusion-coverage score of the ability of each user to spread information over the network. Candidate links for injection are identified by a matrix factorization technique and link injection is performed by attaching links to users according to their score. We additionally perform clustering to identify communities in order to inject links that cross the boundaries of such communities. In our experiments with five real world networks, we demonstrate that our method can significantly spread the information diffusion by performing limited link injection, essential to real-world applications.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications

Keywords

Information diffusion; link injection; social networks

1. INTRODUCTION

The diffusion of information in social networks supports communication for the exchange of ideas or opinions. Although information can spread effectively inside tightly coupled communities, it may hardly propagate beyond their boundaries [2]. Despite the focus of research on the identification of influential spreaders, e.g. for viral-marketing applications [5], it may not be possible to frequently engage such influential users. Thus, the vast majority of information-diffusion processes are initiated by "regular" users (i.e., non-influential spreaders) and is bound to be restrained within the small circle of friends inside a community.

Our approach identifies a limited number of *new* social links that can be injected to help in spreading information outside community boundaries. In this respect, our approach is related to users-recommendation algorithms [3], such as Friend-of-Friend (FoF) schemes. However, existing user-recommendation algorithms do not focus on optimizing information diffusion. More related to our work the approach of [2] recommends connections to boost information diffusion based on knowledge about users' profiles and the content being shared among users, which may not be available due to, e.g., privacy issues. Our approach does not employ knowledge about users' preferences and is based only on the network structure. Furthermore, it controls more tightly the number of the injected links. In this paper, we extend our preliminary approach [1], by directly considering the structure of communities to inject links that can cross their boundaries.

2. PROPOSED APPROACH

Provided a graph G=(N, L) with N nodes and L edges and the respective adjacency matrix $A \in \mathbb{R}^{|N|\times|N|}$, our DNL model [1] performs link injection for boosting information diffusion, by constructing a new $A' \in \mathbb{R}^{|N|\times|N|}$ matrix, which corresponds to a new graph G'=(N, L'), on condition that the number of the new injected links $|L \cap L'|$ is small. DNL consists of the following steps:

Diffusion coverage score: the top-k set $S \subset N$ of nodes are identified, where $|S| \ll |N|$ is the subset of the most influential nodes with the highest *diffusion coverage*. The diffusion coverage for each node $j \in N$ is denoted as $\Delta\lambda(j)$ and represents the importance of node j for the flow of information that can spread over graph G. To measure the impact of each node we compute the robustness of graph G after the node removal. We follow the principle of *interlacing*, which is expressed by the Perron-Frobenius theorem and states that the first (largest) eigenvalue of the adjacency matrix reduces when removing a node or a link.

Non-negative matrix factorization: We factorize the adjacency matrix $A \in \mathbb{R}^{|N|\times|N|}$ according to a non-Negative Matrix Factorization (NMF) technique, generating a new matrix $A_{NMF} \in \mathbb{R}^{|N|\times|N|}$ with $A_{NMF} = WU$, where $W \in \mathbb{R}^{|N|\times D}$, $U \in \mathbb{R}^{D\times|N|}$, and D is the number of latent factors.

Link assignment: Finally, the new links of A_{NMF} with the highest likelihood are selected based on a link assignment algorithm, generating the final adjacency matrix A'. Depending on the link injection strategy, a predefined maximum number of links m is defined. Let L_S be the set of the links currently existing to the top-k nodes of the set S, with $L_S \subset L$, then the predefined threshold m is expressed as $m = |L_S| \times p$, where p is a constant factor. The inputs of the algorithm are (i) the $\Delta\lambda$ diffusion coverage scores of the top-k nodes of set S (step 1); (ii) matrix A_{NMF} (step 2); (iii)

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742733.

constant factor p for threshold m and (iv) an upper bound ub for each node, with the number of links that each node can be injected. After initializing A' with the existing links of A, the algorithm constructs a $B \in \mathbb{R}^{|S| \times |N|}$ matrix, where each i-th row corresponds to the top-k nodes of the set S and each j-th column to all N nodes. The algorithm scans all nodes in S and tries to inject links, i.e. insert new pairs in matrix A', based on the highest score $B(i,j) = \Delta\lambda(i) \cdot A_{\text{NMF}}(i,j)$ as well as with the highest $\Delta\lambda(j)$ score of the j-th column, on condition that the new $< i,j >$ link does not already exist in A' and does not violate any of the two constrains based on threshold m and upper bound ub for both i and j nodes.

L-DNL: Although the DNL model can increase the spread of cascades, it ignores the isolated communities of the network, requiring thus more links to cross the communities' boundaries. To handle this problem we propose a variant of DNL, namely L-DNL, where a clustering algorithm is included to identify users' communities. Thus, prior to the link assignment algorithm, the nodes of the graphs are clustered. Next, for each link in A_{NMF}, a matrix $C \in \mathbb{R}^{|N| \times |N|}$ is constructed with $C(i,j)$=1, if link $< i,j >$ connects two different clusters/communities and 0 otherwise. Finally, over the scan of the score matrix B, we set $B(i,j) = \Delta\lambda(i) \cdot A_{\text{NMF}}(i,j) + C(i,j)$, promoting thus new links $< i,j >$ that cross the communities boundaries.

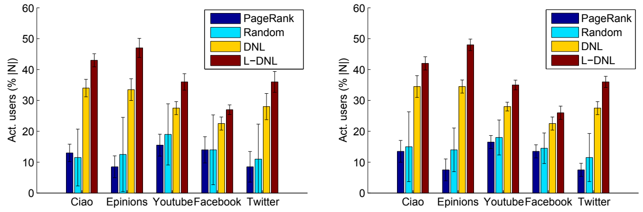
3. EXPERIMENTS

Datasets: We used five real datasets: Ciao and Epinions [7], Twitter [8], YouTube [9], and Facebook [10]. Their statistics are summarized in Table 1. Each connection between two users is weighted according to the information they share, e.g., amount of wall-posts in Facebook, retweets in Twitter, etc.

Table 1: The five networks.

Data Set	Users	Connections	Average Degree	Diameter	Clustering Coefficient
Ciao	7,317	177,727	23.106	10	0.218
Epinions	18,098	529,162	25.898	9	0.209
YouTube	13,723	167,253	10.176	12	0.159
Facebook	46,952	274,086	6.726	20	0.103
Twitter	456,631	14,855,875	28.642	11	0.1887

Results: In our experiments, we considered the PageRank algorithm as the default seed selection strategy in the information diffusion process. We evaluated the basic DNL model and its variant L-DNL against the case where no new connections are being inserted, denoted as *PageRank* and a *Random*-selection baseline, where the same number of links is assigned to randomly-selected graph nodes. We examine the Independent Cascade (IC) and Linear Threshold (LT) models [6]. The default value for the 'stopping probability' in the diffusion models is set to 0.25, i.e., a user propagates the information if 75% of his neighbors have been already activated. In our experiments we report average results out of 100 trials, since the examined diffusion models are probabilistic. In Figures 1(a)-(b) we evaluate the impact of the proposed DNL and L-DNL on the IC and LT models, in terms of the number of user activations, i.e., users that have received and accepted the information furnished by the underlying information diffusion model. In this experiment, we set $m = 1 \times |L_S|$ and $k = 0.1 \times |N|$, with $|L_S|$ being the number of the existing edges in of top-k nodes in S and $|N|$ the number of nodes in the network. For both IC and LT the proposed DNL and L-DNL boost information diffusion for all datasets, while L-DNL outperforms DNL by promoting injecting links that cross the communities' boundaries.



(a) IC model (b) LT model

Figure 1: Evaluation on (a) IC and (b) LT.

To evaluate the performance of L-DNL against DNL, in Figure 2 we examine the relative increase of activated users that L-DNL achieves, by varying the number of injected links. In particular, the constant factor is varied in p=[0.5 1 1.5 2] for threshold $m = p \times |L_S|$, corresponding to different percentages of link injection of the total graph's edges $\%|L|$.

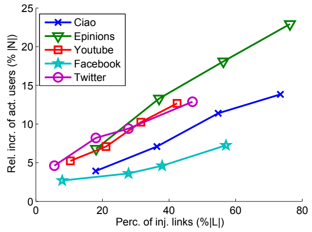


Figure 2: Rel. incr. of act. users (L-DNL vs DNL).

4. CONCLUSIONS

Our link injection method can significantly boost information diffusion in social networks, especially when promoting to inject links that cross the boundaries of users' communities. For future work we will evaluate the sensitivity of our method w.r.t. the probability that the (recommended) link injection by our collaborative-filtering approach would be accepted by users in a real case-study. Also, we will examine the performance of our method on large cascade sizes that can reach a large portion of the network [4].

5. REFERENCES

[1] S. Antaris, D. Rafailidis, and A. Nanopoulos. Link injection for boosting information spread in social networks. *Social Netw. Analys. Mining*, 4(1), 2014.

[2] V. Chaoji, S. Ranu, R. Rastogi, and R. Bhatt. Recommendations to boost content spread in social networks. In *WWW*, pages 529–538, 2012.

[3] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy. Make new friends, but keep the old: Recommending people on social networking sites. In *CHI*, pages 201–210, 2009.

[4] J. Cheng, L. A. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *WWW*, pages 925–936, 2014.

[5] A. Guille, H. Hacid, C. Favre, and D. A. Zighed. Information diffusion in online social networks: A survey. *SIGMOD Rec.*, 42(2):17–28, 2013.

[6] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146, 2003.

[7] `http://www.public.asu.edu/~jtang20/`.

[8] `http://www.public.asu.edu/~huanliu/, /GroupStructure/heterogeneous_network.html`.

[9] `http://socialnetworks.mpi-sws.org/datasets.html`.

[10] `https://snap.stanford.edu/data/higgs-twitter.html`.