The Journal of Systems and Software 98 (2014) 1–8

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

“With a little help from new friends”: Boosting information cascades in social networks based on link injection

Dimitrios Rafailidis[a], Alexandros Nanopoulos[b,*], Eleni Constantinou[a]

[a] Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
[b] University of Eichstätt-Ingolstadt, Germany

ARTICLE INFO

Article history:
Received 11 November 2013
Received in revised form 16 July 2014
Accepted 14 August 2014
Available online 23 August 2014

Keywords:
Information cascades
Viral marketing
Social networks
Matrix factorization

ABSTRACT

We investigate information cascades in the context of viral marketing applications. Recent research has identified that communities in social networks may hinder cascades. To overcome this problem, we propose a novel method for injecting social links in a social network, aiming at boosting the spread of information cascades. Unlike the proposed approach, existing link prediction methods do not consider the optimization of information cascades as an explicit objective. In our proposed method, the injected links are being predicted in a collaborative-filtering fashion, based on factorizing the adjacency matrix that represents the structure of the social network. Our method controls the number of injected links to avoid an “aggressive” injection scheme that may compromise the experience of users. We evaluate the performance of the proposed method by examining real data sets from social networks and several additional factors. Our results indicate that the proposed scheme can boost information cascades in social networks and can operate as a “people recommendations” strategy complementary to currently applied methods that are based on the number of common neighbors (e.g., “friend of friend”) or on the similarity of user profiles.

1. Introduction

Information cascades take place in a social network by following a viral fashion, having users being influenced by the actions of their social contacts and adopting their choices (Liben-Nowell and Kleinberg, 2003). The spread of influence among users in social networks and the viral nature of information cascades are being extensively studied in the context of several applications; for instance: predicting customer churn through interpersonal influence (Zhang et al., 2012) or determining the most influential individuals within a network for purposes of viral marketing (Kempe et al., 2003; Hinz et al., 2011; Zhang et al., 2013). In this manuscript, we focus on information cascades in the context of viral marketing applications.

Viral marketing is a form of electronic word-of-mouth advertising with the help of social media. It creates a process where consumers replicate the viral spread of the marketing message in a way that bears similarity to pathological and computer viruses (Kiss and Bichler, 2008). Viral marketing is considered as a strategy that encourages people to share messages with other people (Kim and Lowrey, 2010), mostly between friends, colleagues and other acquaintance, capitalizing on the “personal recommendation” that is far more credible than anonymous information (Ulmanen, 2011). Thus, a chain reaction with exponential growth can be created, which results in rapid transmission by voluntary action, with low cost and high commercial impact, producing several benefits, such as brand awareness, long term profitability, increased sales, and customer loyalty (Clifford-Marsh, 2009).

The spread of viral marketing is controlled by the factor of *influence* that exists internally in social networks, reflecting the fact that when users perform an action, they can influence their social contacts to do the same (Friedkin, 2006). For this reason, research on viral marketing has focused on the *seed-selection* problem, which identifies and targets the most influential users in a social network, so that they can be used as a seed that will activate a chain-reaction of influence driven by word-of-mouth that will reach a very large portion of network (Bonchi et al., 2011).

1.1. Motivation

Additionally to the well-studied seed-selection problem, it is also important to understand the reasons that may hinder the

* Corresponding author. Tel.: +49 841 937 21872.
E-mail addresses: draf@csd.auth.gr (D. Rafailidis), alexandros.nanopoulos@ku.de, alexandros.nanopoulos@gmail.com (A. Nanopoulos), econst@csd.auth.gr (E. Constantinou).

http://dx.doi.org/10.1016/j.jss.2014.08.023
0164-1212/

cascade of viral marketing. Social networks often contain online communities with users pursuing mutual interests or goals. A tightly knit community can lead to "isolation" of its members, by preventing information to flow in and out of it (David and Jon, 2010). For instance, users of a community about a specific mobile operating system, like Blackberry OS, cannot easily get informed about other systems, like Android (by Google) or iOS (by Apple). This problem has been identified by recent research, which explains that densely connected communities can become "natural obstacles" causing information cascades to stop (David and Jon, 2010).

A direct way towards overcoming this problem, is to select more seed nodes that scatter over all communities in a social network. Nevertheless, this results in a large number of seed nodes, which contradicts the main advantage of viral marketing: i.e., a spread over a large fraction of the network initiated only from a very small number of seed nodes.

An alternative approach is to *inject* new links among users of a social network in an attempt to join communities with bridges. Social networks are currently using several "people recommendation" schemes based on simple criteria, such as the number of common neighbors (e.g., "friend of friend"[1]), similarity of user profiles, etc. Additionally, there exist several more advanced methods for the problem of link prediction (Liben-Nowell and Kleinberg, 2003). All these methods, however, are not designed to directly optimize the cascade of information in social networks (Chaoji et al., 2012). What is, therefore, required is the development of novel link injection schemes that can substantially improve the effectiveness of information cascades, which can boost applications like viral marketing.

1.2. Contribution and outline

In this manuscript, we propose a novel method for performing link injection in a social network. In our study, we focus on social networks that allow users to perform interaction with items. Such item interactions manifest themselves in purchases or product evaluations, e.g. consumer reviews, product ratings or both. In the social domain, users interact with other users. Social interactions manifest themselves in observed communication links; for instance, "followers" on Twitter or "friends" on Facebook. Connected users in the social domain may exchange information relevant to the item domain. As a result, the information exchanged in the social domain can influence behavior in the item domain. Examples of websites which offer social interactions among users alongside the item domain include: Ciao (for online product reviewing) and Last.fm (a social site for online music listening). In our experimental evaluation we use data from the aforementioned social networks.

Our approach aims at predicting injected links in order to boost the cascade of a piece of information, by making it spread virally over the network as much as possible. The injected links are being predicted in a collaborative-filtering fashion based on factorizing the adjacency matrix that represents the structure of the social network. Our method controls the number of injected links to avoid an "aggressive" link-injection scheme that may compromise the experience of users in a social network.

Compared to recent related work (Chaoji et al., 2012), our method does not depend on prior knowledge in the form of users' profiles, because such information may not be available, e.g., for reasons of privacy preservation. We evaluate the performance of the proposed method by examining real data sets from social networks and several additional factors. Our results indicate that the proposed link-injection scheme can boost information cascades in social networks and can operate as an alternative "people recommendations" approach in social networks, which can operate complementary existing methods, e.g., those based on the number of common neighbors (e.g., "friend of friend") or on the similarity of user profiles (Chen et al., 2009; Guy et al., 2009).

The rest of this manuscript is organized as follows: Section 2 summarizes the related work. Section 3 describes the problem formulation, followed by the description of the proposed methodology in Section 4. Our experimental evaluation is presented in Section 5. The manuscript is concluded in Section 6.

2. Related work

2.1. Seed selection and influence maximization

Viral marketing can be stated as an influence-maximization problem that is concerned with selecting the set of seed nodes in a social network who will initiate the spread of a piece of information and will cause the largest possible number of activations among remaining users (Kempe et al., 2003). The number of seed users is predefined and represents the cost to initiate the spread of information, i.e., the larger the seed set size, the higher the cost. According to this formulation, it has been shown that the influence maximization problem is NP-hard which means that finding the seed set that will achieve the maximal possible number of activated nodes requires execution time that grows exponentially with the size of the network (Kempe et al., 2003). Since real social networks are very large, finding the optimal seeds requires prohibitively large computational cost. To address this problem, a greedy hill-climbing algorithm has been proposed (Kempe et al., 2003). The popularity of the paradigm proposed in Kempe et al. (2003) resulted in the development of several extensions that scale-up seed selection in very large social networks (Chen et al., 2009).

The aforementioned approaches exploit knowledge about pairwise influence between users, called influence factors. Prior knowledge of influence factors is not possible in several real-world applications, because influence is intangible and hard to measure. A data-driven approach has been proposed recently (Goyal et al., 2011) which estimates influence factors based on actions that have been previously performed by users, such as buying a product or joining a community. This approach can be used effectively in the case when the recorded actions are relevant to the current viral-marketing campaign. Otherwise, the estimated influence factors will not be representative. Another problem is that several forms of recorded user actions comprise private data that are often not allowed to be used for marketing purposes. To avoid issues related to acquiring knowledge about influence factors, recent research in marketing proposed to select seeds according to their centrality in the social network structure, assuming that globally central users can initiate information diffusion with high chances to reach a larger part of the network (Hinz et al., 2011) than the non-central users do.

In our approach, we follow the direction of Hinz et al. (2011), by assuming no prior knowledge of influence factors. However, our investigation is complementary to the problem of seed selection, because the latter focuses on the selection of the most influential among the existing nodes in the network, which can maximize spread of a viral process, whereas we focus on the injection of new links between nodes.

2.2. Link injection and people recommendation

Most social networks use schemes for "people recommendation" based on similarity of user profiles or on algorithms that

[1] http://www.facebook.com/notes/facebook/people-you-may-know/15610312130.

resemble the "friend of friend" scheme (Chen et al., 2009; Guy et al., 2009). Standard link prediction algorithms, which predict new connections among nodes that are likely to occur in the near future (Liben-Nowell and Kleinberg, 2003), have been also examined for the same purpose (Schifanella et al., 2010). Nevertheless, such algorithms aim at predicting links that will occur, without taking into account the increase of information cascades. Twittomender is an alternative approach proposed to recommend users to follow on Twitter based on a combination of information cascade and collaborative filtering type features (Hannon et al., 2010).

More closely related to our approach is the recent work of Chaoji et al. (2012), which introduces algorithms for recommending connections among users, aiming at boosting information cascades in social networks. Although we share the same overall objective with Chaoji et al. (2012), our proposed method differs as follows. We exclusively consider users' friendships and we do not assume prior knowledge of influence factors such as users' profiles, interests, updates, etc., since it is extremely difficult to consider and continuously monitor all these influence factors for each node/user. The difficulty is that influence is intangible and hard to measure, since users' private data are not necessarily publicly available, depending on the privacy policy that each social networks follows (see Section 2.1). Additionally, in contrast to Chaoji et al. (2012) our method does not assume that a predefined number of new links should be created for *each* node/user in the network. This is of great importance, since in this case it would be required for each node/user in the network to accept all the recommended connections, a case which not necessarily reflects to real-world users. Moreover, in our study we examine different cascade models compared to Chaoji et al. (2012).

3. Problem formulation

Following standard notations, we use capital italic letter for matrices (e.g. A), lower-case bold letters for vectors (e.g. $\mathbf{a}$) and calligraphic fonts for sets (e.g. $\mathcal{A}$). Let $A \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$ be the adjacency matrix of the graph representing the structure of the social network, where $\mathcal{N}$ is the set of nodes and $\mathcal{L}$ is the set of links (pairwise social relationships). We consider the graph undirected, and thus the adjacency matrix A is symmetric, which means that if $U_i, U_j \in \mathcal{N}$ and $(U_i, U_j) \in \mathcal{L}$ then $(U_j, U_i) \in \mathcal{L}$. The values in A are initially 1 or 0, denoting the existence or the absence of a link between two nodes, respectively. Notice that A matrix is usually very sparse, with $|\mathcal{L}| \ll |\mathcal{N}|^2$.

We investigate the problem of extending the initial set of links $\mathcal{L}$ into a new $\mathcal{L} \cup \mathcal{L}^+$ set that contains additional (injected) links between the existing nodes of $\mathcal{N}$. Our goal is to choose the injected links in $\mathcal{L}^+$ in such way that they will boost the viral processes and help it spread more over the network. Additionally, we want to control the number of injected links $|\mathcal{L}^+|$, by expressing it as a factor f of the number of initial links $|\mathcal{L}|$; i.e., $|\mathcal{L}^+| = f \times |\mathcal{L}|$.

In real applications of a link-injection strategy, a set of link recommendations are going to be provided to the users in set $\mathcal{N}$ for adding new links among them, until $|\mathcal{L}^+|$ such new (injected) connections have been created. In order to increase the accuracy of such recommendations, we perform link prediction based on a collaborative-filtering approach (see Section 4.1), in order to predict the links that are more likely to be created. The predefined upper bound of the number of injected links guarantees that the number of injected links is kept controlled in order not to affect the experience of users that are not willing to receive a large number of recommendations to connect to other users. We have to note that our experimental evaluation (see Section 5) did not involve a user study where such recommendations could be provided to real users. We thus perform an indirectly experimental evaluation based on the premise that the generated links can be directly injected, which is justified by the high recommendation accuracy of the used link-prediction method. In order to evaluate the accuracy of the examined link-prediction methodology based on the NMF, the Area Under the ROC Curve (AUC) measure[2] has been applied to two datasets, namely Ciao and Last.fm which will be described in Section 5. We have randomly partitioned the datasets into testing and training subsets, where the half of the dataset consist the training subset and the rest of the dataset consists the testing subset. According to this partition, AUC for the two datasets was above 0.9. This result demonstrates that the examined link-prediction method is effective and can be used as a basis of the proposed approach.

4. Proposed methodology

4.1. Link injection based on collaborative filtering

In this section, we present the proposed link-injection method, which is based on the collaborative filtering paradigm. Following this paradigm, if two users have the same friends, then the users tend also to become friends. To capture this effect, we used the matrix factorization technique of non-Negative Matrix Factorization (Berry et al., 2007) (NMF), in order to reveal the latent associations between the users and to establish new users' connections, achieving thus link injection in the graph. By factorizing the $A \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$ adjacency matrix according to NMF, a new enhanced matrix $A' \in \mathbb{R}^{|\mathcal{N}|\times|\mathcal{N}|}$ is generated based on $A' = WH$, where $W \in \mathbb{R}^{|\mathcal{N}|\times k}$, $H \in \mathbb{R}^{k\times|\mathcal{N}|}$ and k is the number of latent factors, usually expressed as a percentage of $\%|\mathcal{N}|$.

NMF calculates an approximation of A, where the result of the factorization is the A' matrix, with the new stored values in A' denoting the recalculated weights of the initial $\mathcal{L}$ set of links, along with the weights of the injected set of links, denoted by $\mathcal{L}^+$. The factorization problem of NMF in the squared version is stated as follows, given adjacency matrix A, and the non-negative matrices W and H, the goal is to minimize the function $F(W, H) = ||A - WH||_F^2$, where $||\cdot||_F^2$ denotes the Frobenius norm. In our approach, we used the alternating non-negative least square algorithm of Liu et al. (2013), which follows an iterative approach, providing solution for the NMF problem and reaching to a global convergence. Therefore, the outcome of the NMF is the reconstruction of the initial A adjacency matrix of a $\mathcal{L}$ set of links, resulting into a new generated A' adjacency matrix of $\mathcal{L} \cup \mathcal{L}^+$ set of links.

Alternatively to NMF, standard factorization methods, such as the Singular Valued Decomposition (SVD) and Latent Semantic Analysis (LSA) methods could be used. SVD in this case would involve first a factorization of matrix A as $A = U\Sigma V^T$ and then a low-rank matrix approximation $A' = U\Sigma' V^T$, where Σ' is the same matrix as Σ except that it contains only the k largest singular values (the other singular values are replaced by zero). In case matrix A is weighted, LSA uses SVD to reduce the number of columns while preserving the similarity structure among rows. Users can be compared by taking the cosine of the angle between the two vectors formed by any two rows. The reason for selecting NMF instead of SVD is because, by definition, NMF generates a non-negative matrix A', which is important in our case, since the initial weights of $\mathcal{L}$ in A are positives and thus the recalculated weights of $\mathcal{L} \cup \mathcal{L}^+$ set of links in A' should also preserved positives.

The result of NMF is the A' reconstructed matrix, which by the definition of NMF is a full adjacency matrix, with the weights of $|\mathcal{L} \cup \mathcal{L}^+| = |\mathcal{N}|^2$ ranging from very low positives values (close to 0) to high values (close to 1). Therefore, depending on the link injection strategy that we would like to follow, given the injected set

[2] AUC is defined at: http://en.wikipedia.org/wiki/Receiver_operating_characteristic.

of links $\mathcal{L}^+$ in the A' reconstructed adjacency matrix, we filter out the weights of the links in $\mathcal{L}^+$ that exceed a predefined threshold. In our experiments, we varied the aforementioned threshold so as to compute $|\mathcal{L}^+| = f \times |\mathcal{L}|$, where f is a constant factor, i.e. ($\times 2$, $\times 4$, $\times 6$, $\times 8$), expressing thus the number of injected links $|\mathcal{L}^+|$ as a multiplication of the f constant factor with the number of initial links $|\mathcal{L}|$.

4.2. Diffusion model

In order to examine viral processes for experimental evaluation of the proposed link-injection method, we follow the Independent Cascade (IC) model (Kempe et al., 2003). IC is a widely used model for information diffusion in social networks. IC starts by activating the users in the seed set and then continues in discrete steps. When a user v becomes active in a time step t, it has a single chance to activate each of the neighbor user w connected to it that are currently inactive. User v succeeds in activating w with probability p_{vw} equal to the (normalized) influence factor that v has on w. If v succeeds then w becomes active in step $t+1$ and recursively tries to activate its neighbours. Otherwise, v makes no further attempts to activate w. The process runs until no more activations are possible. Therefore, IC models the individual influence each user has on all its neighbours.

In IC, probability p_{vw} represents the influence that users have on each other. For each tie between u and w we determine the value of p_{vw} according to the similarity between v and w, i.e., p_{vw} increases with increasing number of commonly preferred items by v and w. The basis for this is 'homophily', since users with more similar behavior tend to have more influence on each other (Peres et al., 2010). Therefore, for each tie from user v to w we first compute as weight for the tie, the number of commonly preferred items by v and w, and then compute p_{vw} as the normalized weight by dividing on the sum of all weights of incoming ties to w.

Social influence is the only factor that determines the activation of users in IC. However, the inherent preference that users have about the diffused information (product, brand, etc.) is also a factor that determines activations (Kim and Lowrey, 2010). We extend IC accordingly, by associating each user w with a random variable θ_w that follows Beta distribution $\theta_w \sim Beta(\alpha, \beta)$.

Thus, θ_w takes values in the range [0, 1], with values closer to 0 indicating higher inherent preference; i.e., w is more likely to get activated. When a user v tries to activate another user w, the probability of activation is, thus, calculated as following:

$$p_{vw} + \gamma \times \max(\theta_w; 1 - \theta_w) \times \theta_w \tag{1}$$

In this way, activation of each user w depends both on the social influence (weight p_{vw}) and on the inherent preference θ_w. Additionally, we also consider the fact that users with stronger inherent preferences tend to resist changing them (Mussweiler and Strack, 2000). In Eq. (1), factor $\gamma \times \max(\theta_w; 1 - \theta_w)$ quantifies this fact: the more extreme the inherent preference of w is, i.e., the closer θ_w is to 0 or 1, the less willing is w to change its inherent preference. Thus, γ is set to be +1 if θ_w is not less than 0.5 (more positive inherent preference), and to be -1 if θ_w is greater than 0.5 (more negative inherent preference).

IC assumes that all activated users will try to activate their neighbors. Nevertheless, not all activated users – no matter how positive their opinion about a product or a brand is – will pass on the message by trying to activate their neighbor users, because they may just keep it to themselves or forget about the whole experience all together. To take this into account, we assume that all users have a 'stopping probability' (equal for all nodes) and when they become activated, they try in turn to activate their neighbors according to this 'stopping probability'. Therefore, the higher the 'stopping probability', the higher is the 'difficulty' of the social network since more users are reluctant to participate in the viral process.

We have emphasize that, in contrast to existing research (Chaoji et al., 2012), we assume *no* knowledge of influence factors (i.e., the aforementioned probabilities of the form: p_{vw} for each pair of users v and w) during the prediction of the injected links. We examine influence factors in the IC model *only* during the evaluation of the effectiveness of the injected links.

5. Experimental evaluation

We performed an experimental evaluation having the following objectives:

- Initially, we want to investigate the impact of different seed selection strategies and examine how the 'difficulty' of the network reduces the spread of a viral process (see Section 4.2).
- Our main objective is to test the effectiveness of link injection with respect to the seed selection and network 'difficulty'.

5.1. Data sets

The first evaluation data set is the Ciao data set.[3] Ciao is a product review site, where users can rate items by writing reviews and establish trust networks with their like-minded users. Users give ratings from 1 to 5 for each product. The Ciao data set consists of 6262 users, with 167,320 ratings on 20,416 product items and 109,524 users trust-relations. The second evaluation data set is the Hetrec 2011 Last.fm data set,[4] which contains 92,800 listening records of 17,632 artists from 1892 users and 12,717 bi-directional user-friend relations. Therefore, the goal is to maximize the percentage of activated users. The characteristics of these data sets are described in Table 1.

For each data set, we create a graph that represents the social ties. The users' histograms of the evaluation data sets are presented in Fig. 1. In the case of the Ciao data set, we consider that the users who definitely express positive preference are those whose average rating is equal to largest point in the given rating scale. For the Last.fm data set, these are the users with average artist-listenings greater than 2000. These selected users (in both data sets) are considered to have higher inherent preference (see Section 4.2). To avoid noise existing in user interactions (rating or listening events), we focused on the more dense part of each evaluation data set, in order to examine the collaborative effects between users in the proposed link injection strategy. Thus, we applied the commonly used technique of p-core filtering (Batagelj and Zaversnik, 2011). The p-core of level p has the property, that each user and product/artist has/occurs in at least p observations. In our experiments we set p equal to 0.01 of the total number of users' ratings/artist listenings. Since the diffusion model is of probabilistic nature, we report average result out of 30 trials.

5.2. Results for seed selection strategies

In this set of experiments, we evaluate the following seed selection strategies without performing any link injection:

- *Random*: users are randomly selected as seed size, irrespective of their topological features in the network.

[3] http://www.public.asu.edu/~jtang20/datasetcode/truststudy.htm.
[4] http://www.grouplens.org/node/462.

Table 1
Characteristics of two real-world data sets.

Name	Type	Users	Social ties	Items	Feedback
Last.fm	Music listening	1892	12,717	17,632 (artists)	92,834 (listen events)
Ciao	Consumer reviews	2380	57,544	16,900 (products)	36,065 (reviews)

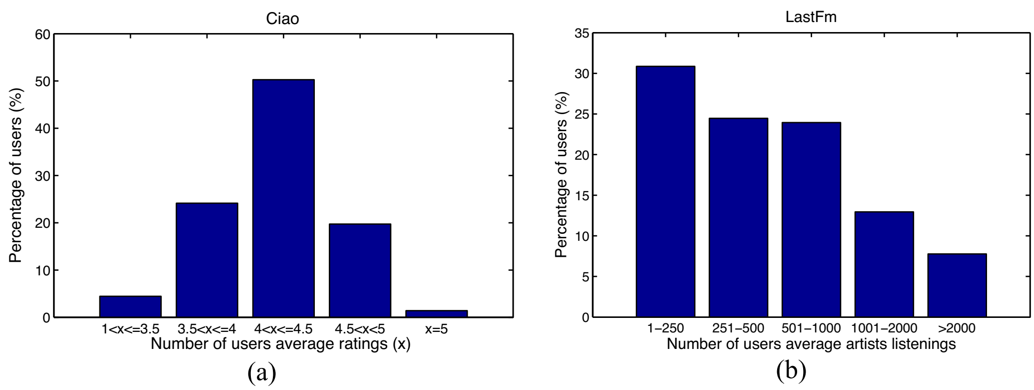

Fig. 1. The users' histograms of the (a) Ciao and (b) Last.fm evaluation datasets.

- *Degree centrality*: is defined as the number of links incident upon a user (Newman, 2010). Users/nodes with the highest degree centrality are selected as seed size.
- *Betweenness centrality*: is a measure of a user's centrality in a network (Newman, 2010). It is equal to the number of shortest paths from all vertices to all others that pass through that node, expressing how each user controls the flow within the network. Users/nodes with the highest betweenness centrality are selected as seed size.
- *Page rank*: is the widely known Google's Page Rank measure (Page et al., 1999), which estimates the importance of a user in the network. Users/nodes with the highest Page Rank values are selected as seed size.

In Section 4.2 we described α as an important parameter of the examined diffusion model, which controls the Beta distribution that assigns the θ_w weights that determine the inherent preference of nodes. As described, lower values of α parameter make the network more susceptible to the cascade, which means that users are more willing to accept the influence of the cascade, because they tend to have increased inherent preferences. In contrast, higher values of parameter α result in more 'difficult' network, since the users are more unwilling to the acceptance of the cascade. In Fig. 2, we present the experimental results for the aforementioned seed selection strategies by varying parameter α. As expected, higher values of α reduce the percentage of activated users for all seed selection strategies. Moreover, the seed selection strategy based on the highest PageRank values clearly outperforms the rest seed selection strategies, with the random selection strategy achieving the lowest numbers of users' activation, since it does not exploit any topological information of the users in the network. In the rest of experiments we set α equal to 10, reflecting to the challenging scenario, where users are generally not suspectable to the cascade.

In Fig. 3, we present the experimental results by varying the 'stopping probability'. As described in Section 4.2, larger values of the 'stopping probability' reduce the percentage of activated users in the market for all different seed selection strategies, because less users are willing to take part in the viral process. Similarly to the previous experiments, the seed selection strategy based on PageRank values performs higher than the rest selection strategies, with the random selection strategy achieving the lowest number of activated users. Following this finding, in the rest of our experiments, we set the 'stopping probability' equal to 0.75.

In Fig. 4, we show the experimental results by varying the percentage of users participating the seed. As expected, the final percentage of activated users is increased along with the increase of the seed size. The reason is that the cascade starts from the selected seed, thus a larger seed can propagate the cascade more effectively. In the rest of our experiments we set the percentage users as seed size equal to $3\%|\mathcal{N}|$. Moreover, we confirm the experimental results of Figs. 2 and 3, based on which PageRank is the selection seed strategy of highest performance. Therefore, in the rest of our experiments we set PageRank as the default seed selection strategy.

5.3. Results for link injection

In the following set of experiments we evaluate the impact of the proposed link injection strategy of Section 4.1. As described in Section 5.2, the chosen seed selection strategy is PageRank. Therefore, we denote as PageRank-NMF the method resulting from using PageRank for seed selection in combination with the enhanced adjacency matrix, derived by NMF ($k = 0.2|\mathcal{N}|$).

In Fig. 5 we evaluate the performance of link injection in terms of percentage of activated users, by varying the constant factor f that determines the number of injected links (see Section 4.1). The reason for considering in this experiment the constant factor f ranging from $\times 2$ to $\times 8$, is that lower values of f cannot achieve effective link injection and the propagation of the cascade is limited to the local neighborhoods of users that have been selected as seeds. On the contrary, higher values of f can result in aggressive link injection (in the form of recommendation) that can negatively impact the experience of users in a social network.

Based on the experimental results of Fig. 5, PageRank-NMF clearly outperforms the baseline PageRank method, by solving the extreme sparsity that occurs in the initial A adjacency matrix with the help of the enhanced A' matrix having f times more (injected) links. Additionally, as expected, performance is increased along with the increase of f. This happens because the more links are injected, the more the sparsity problem is solved.

Therefore, in the experiments of Figs. 6 and 7, we set the link injection strategy of (f=8). Analogously, the proposed PageRank-NMF method clearly achieves higher number of users' activation in

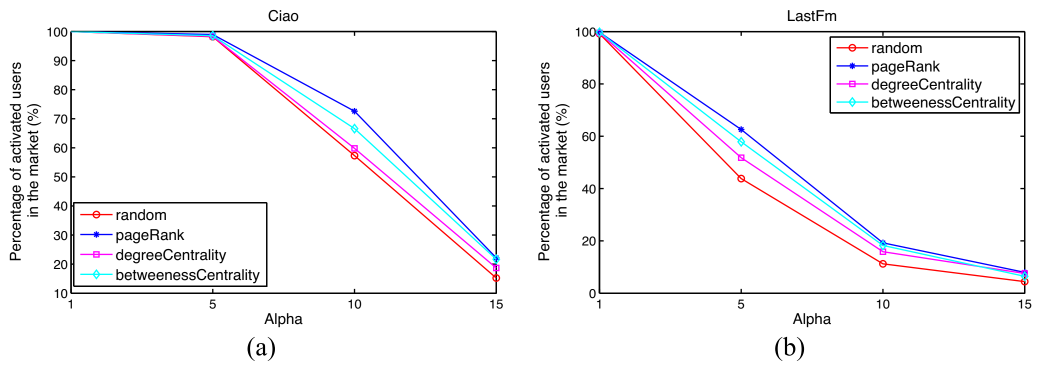

Fig. 2. The impact of parameter α on the (a) Ciao and (b) Last.fm evaluations data sets for the examined seed selection strategies.

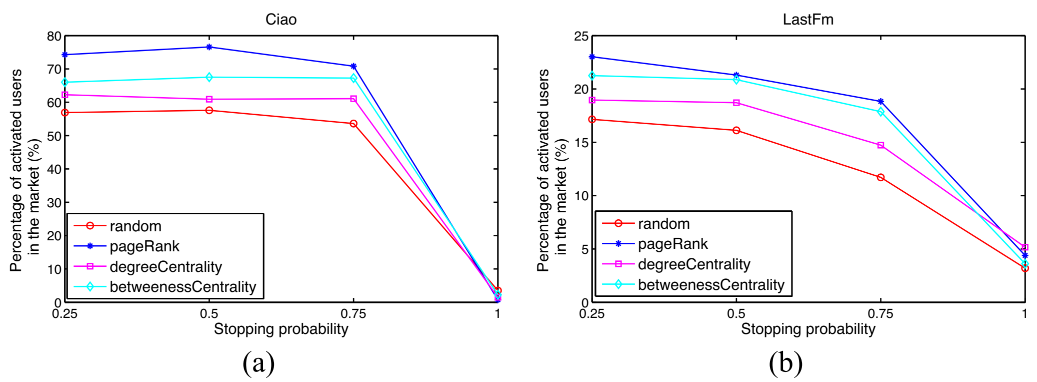

Fig. 3. The impact of 'stopping probability' on the examined seed selection strategies.

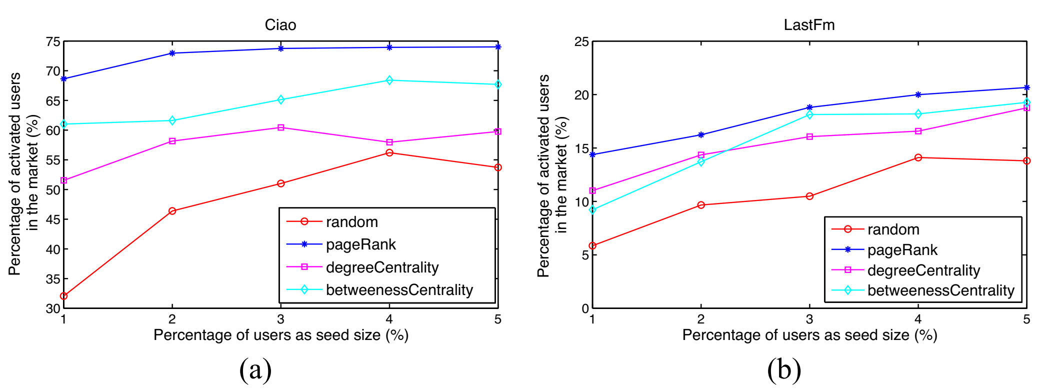

Fig. 4. The impact of the seed size.

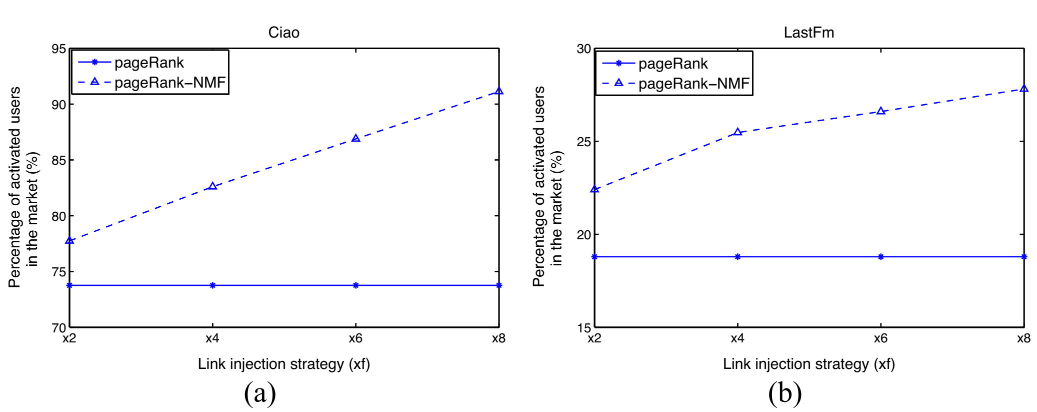

Fig. 5. The impact of link injection.

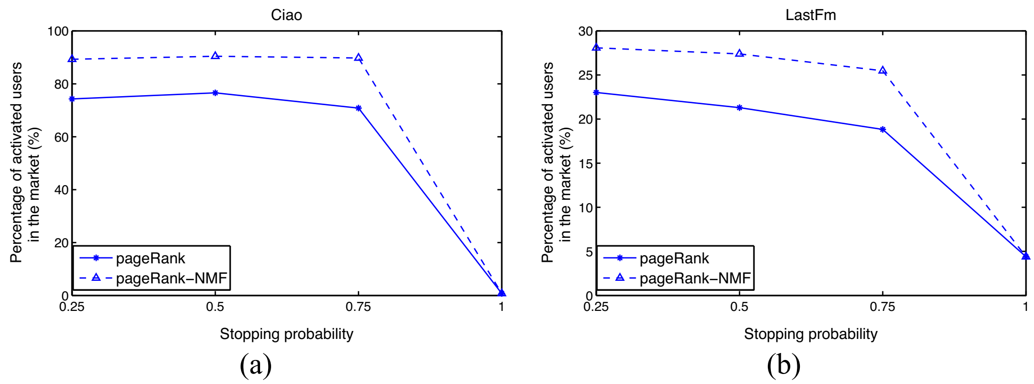

Fig. 6. Comparison of the proposed link injection strategy PageRank-NMF against the baseline PageRank method, by varying 'stopping probability'.

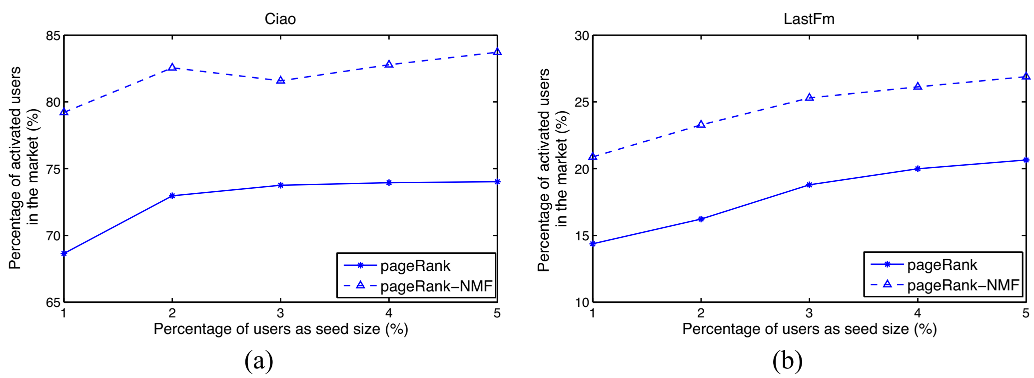

Fig. 7. Comparison of the proposed link injection strategy PageRank-NMF against the baseline PageRank method, by varying the seed size.

the market than the baseline PageRank method does, for both: (a) 'stopping probability' and (b) seed size.

6. Conclusions

In this paper, we proposed a novel method for injecting links in a social network. Our method follows a collaborative-filtering fashion, by being based on factorizing the adjacency matrix that represents the structure of the social network. We aim at predicting links that can boost the spread of information cascades.

The presented experimental results indicated that link injection can become an effective policy to overcome the factors that hinder cascades in social networks, which negatively impact the effectiveness of viral marketing. The proposed method can be implemented as an alternative "people recommendations" that can operate complementary existing methods, e.g., those based on the number of common neighbors (e.g., "friend of friend") or on the similarity of user profiles. In this way, a social network can enhance the structure of the social relations among its users in order to directly help the cascade of information that support applications such as that of viral marketing.

An important decision related to the application of the proposed method concerns the amount of injected links, which is controlled by the factor f. As we have seen, increased values of f can result in significant gains in terms of activated nodes. Nevertheless, a more "aggressive" link-injection scheme can compromise the experience of the users in a social network. Relatively small values of f (ranging between 2 and 8 as in our experiments), result to a small number of links added to the profile of users. Since the number of social connections for individual users can exceed one hundred,[5] the injected links can complete the existing ones without significantly altering the user profiles. Moreover, through the injected links, users can exploit new possibilities and get exposed to novel type of information, therefore they may be willing to accept such additional links. We have to note that in our experimental study, we examined injected links directly in terms of their number. In real applications of the proposed method, users have to first accept the injected links (a forced injection scheme may be considered unsuitable for most of social networks). This corresponds to a challenge that has to be addressed by future work, in order to develop injection schemes that both improve the effectiveness of cascades in social networks but also tend to predict links that will be accepted by users.

Finally, we have to consider that viral marketing can bring negative impacts due to the inherent lack of control: it is not possible to determine how the spread will be developed and to whom. Different users may perceive very differently the message of a viral-marketing campaign, which means that there exist always the possibly to understand the message in a wrong way, either by considering it as spam or – worse – initiating backlashes based on negative word of mouth that can negatively impact the reputation of the campaign initiator (Kaikati and Kaikati, 2004). In such cases, injected links can increase the spread of a backlash. This is another reason to have a controlled number of injected links.

In our future work, we plan to extend the proposed methodology, aiming at a more limited link injection strategy. This will

[5] http://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859

be achieved by extracting a vulnerability score for each node/user, expressing the affect that nodes/users will have in the spread of influence. With the help of the proposed link injection strategy and a link assignment technique, the goal will be to efficiently distribute a limited set of injected links to the identified nodes/users of high vulnerability score, in order to increase the spread of cascade, while highly reducing the number of injected links.

References

Batagelj, V., Zaversnik, M., 2011. Generalized Cores, arxiv.org/abs/cs/0202039.

Berry, M., Browne, M., Langville, A., Pauca, V., Plemmons, R., 2007. Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. 52, 155–173.

Bonchi, F., Castillo, C., Gionis, A., Jaimes, A., 2011. Social network analysis and mining for business applications. ACM Trans. Intell. Syst. Technol. 2 (3), http://dx.doi.org/10.1145/1961189.1961194, 22:1–22:37.

Chaoji, V., Ranu, S., Rastogi, R., Bhatt, R., 2012. Recommendations to boost content spread in social networks. In: Proceedings of the 21st International Conference on World Wide Web, WWW '12, ACM, New York, NY, USA, pp. 529–538, http://dx.doi.org/10.1145/2187836.2187908.

Chen, J., Geyer, W., Dugan, C., Muller, M., Guy, I., 2009. Make new friends, but keep the old: recommending people on social networking sites. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, ACM, New York, NY, USA, pp. 201–210, http://dx.doi.org/10.1145/1518701.1518735.

Chen, W., Wang, Y., Yang, S., 2009. Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on KNOWLEDGE Discovery and Data Mining, KDD '09, pp. 199–208.

Clifford-Marsh, E., 2009. Viral Marketing. Revolution. http://search.proquest.com/docview/231164510?accountid=41205

David, E., Jon, K., 2010. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York, NY, USA.

Friedkin, N., 2006. A Structural Theory of Social Influence, Structural Analysis in the Social Sciences. Cambridge University Press http://books.google.de/books?id=agQbBpFSGIEC

Goyal, A., Bonchi, F., Lakshmanan, L.V.S., 2011. A data-based approach to social influence maximization. In: Proceedings of the VLDB Endowment, vol. 5(1), pp. 73–84, arXiv:1109.6886.

Guy, I., Ronen, I., Wilcox, E., 2009. Do you know?: recommending people to invite into your social network. In: Proceedings of the 14th International Conference on Intelligent User Interfaces, IUI '09, ACM, New York, NY, USA, pp. 77–86, http://dx.doi.org/10.1145/1502650.1502664.

Hannon, J., Bennett, M., Smyth, B., 2010. Recommending twitter users to follow using content and collaborative filtering approaches. In: Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, ACM, New York, NY, USA, pp. 199–206, http://dx.doi.org/10.1145/1864708.1864746.

Hinz, O., Skiera, B., Barrot, C., Becker, J.U., 2011. Seeding strategies for viral marketing: an empirical comparison. J. Market. 75 (6), 55–71.

Kaikati, A.M., Kaikati, J.G., 2004. Stealth marketing: how to reach consumers surreptitiously. CA Manage. Rev. 46 (4), 6–22.

Kempe, D., Kleinberg, J., Tardos, E., 2003. Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 137–146.

Kim, Y., Lowrey, T.M., 2010. Marketing Communication on the Internet. John Wiley and Sons, Ltd., http://dx.doi.org/10.1002/9781444316568.wiem04062.

Kiss, C., Bichler, M., 2008. Identification of influencers – measuring influence in customer networks. Decis. Support Syst. 46 (1), 233–253, http://dx.doi.org/10.1016/j.dss.2008.06.007.

Liben-Nowell, D., Kleinberg, J., 2003. The link prediction problem for social networks. In: Proceedings of the Twelfth International Conference on Information and Knowledge Management, CIKM '03, ACM, New York, NY, USA, pp. 556–559, http://dx.doi.org/10.1145/956863.956972.

Liu, H., Li, X., Zheng, X., 2013. Solving non-negative matrix factorization by alternating least squares with a modified strategy. Data Min. Knowl. Discov. 26 (3), 435–451.

Mussweiler, T., Strack, F., 2000. Numeric judgments under uncertainty: the role of knowledge in anchoring. J. Exp. Soc. Psychol. 36, 495–518.

Newman, M., 2010. Networks: An Introduction. Oxford University Press, Inc., New York, NY, USA, ISBN: 0199206651, 9780199206650.

Page, L., Brin, S., Motwani, R., Winograd, T., 1999. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab, http://ilpubs.stanford.edu:8090/422/

Peres, R., Muller, E., Mahajan, V., 2010. Innovation diffusion and new product growth models: a critical review and research directions. Int. J. Res. Market. 27 (91), 91–106.

Schifanella, R., Barrat, A., Cattuto, C., Markines, B., Menczer, F., 2010. Folks in folksonomies: social link prediction from shared metadata. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining, WSDM '10, ACM, New York, NY, USA, pp. 271–280, http://dx.doi.org/10.1145/1718487.1718521.

Ulmanen, H., 2011. Antecedents of and Their Effect on Trust in Online Word-of-mouth: Case Finnish Discussion Forums. University of Jyv"askylä (M.Sc. thesis).

Zhang, X., Zhu, J., Xu, S., Wan, Y., 2012. Predicting customer churn through interpersonal influence. Knowl.-Based Syst. 28, 97–104, http://dx.doi.org/10.1016/j.knosys.2011.12.005.

Zhang, X., Zhu, J., Wang, Q., Zhao, H., 2013. Identifying influential nodes in complex networks with community structure. Knowl.-Based Syst. 42, 74–84, http://dx.doi.org/10.1016/j.knosys.2013.01.017.

Dimitrios Rafailidis was born in Larissa, Greece, in 1982. He received the Diploma in Informatics from the Computer Science Department, the M.Sc. degree in Information Systems and the Ph.D. degree in Information Retrieval from the Aristotle University of Thessaloniki, Greece in 2005, 2007 and 2011, respectively. His main research interests include Machine Learning and Pattern Recognition, Multimedia Information Retrieval, Databases, Social Media and Artificial Intelligence Systems..

Alexandros Nanopoulos is currently assistant professor at Catholic University of Eichstaett-Ingolstadt, Germany. His research interests include Machine Learning, Data mining and Web Information Retrieval. He received the B.Sc. and Ph.D. degrees from the Department of Informatics of Aristotle University of Thessaloniki, Greece, where he worked as a lecturer from 2004 to 2008. From 2005 to 2008, he was also an academic council of the Hellenic Open University, Greece. From 2008 to 2013 he was an assistant professor at University of Hildesheim, Germany. Dr Nanopoulos is the coauthor of more than 120 articles in international journals and conference proceedings. He has also coauthored the monographs Advanced Signature Indexing for Multimedia and Web Applications and R-Trees: Theory and Applications, both published by Springer Verlag. He has coedited the volume Wireless Information Highways, published by Idea Group, Inc. In 2008, he has served as a cochair of the European Conference of Artificial Intelligence (ECAI) Workshop on Mining Social Data and, in 2006 and 2007, as a cochair of the Advances in Databases and Information Systems (ADBIS) Workshops on Data Mining and Knowledge Discovery. Dr Nanopoulos has also served as a program committee member of several international conferences on data mining (e.g., ACM KDD, ACM RecSys).

Eleni Constantinou is a Ph.D. candidate in the Department of Informatics, Aristotle University of Thessaloniki, Greece. She holds a B.S. in Informatics from Aristotle University of Thessaloniki and an M.Sc. in Information Systems from Aristotle University of Thessaloniki. Her research interests include software reuse, program comprehension and architecture recovery.