Social Marketing Meets Targeted Customers: A Typical User Selection and Coverage Perspective

Qi Liu[1], Zheng Dong[1], Chuanren Liu[2], Xing Xie[3], Enhong Chen[1,*], Hui Xiong[2]
[1]School of Computer Science and Technology, University of Science and Technology of China
qiliuql@ustc.edu.cn, zhengcro@mail.ustc.edu.cn, cheneh@ustc.edu.cn
[2]MSIS Department, Rutgers Business School, Rutgers University, USA
chuanren.liu@rutgers.edu, hxiong@rutgers.edu
[3]Microsoft Research
xing.xie@microsoft.com

***Abstract*—The emergence of social networks has provided opportunities for both targeted marketing and viral marketing. By concentrating the efforts on a few key customers, targeted marketing could make the promotion of the items (products) much easier and more cost-effective. On the other hand, viral marketing aims at finding a set of individuals (seeds) to maximize the word-of-mouth propagation of an item. However, these two marketing strategies can only exploit some specific characteristics of the social networks, and the problem of how to combine them together to build a better, stronger business is still open. To that end, in this paper, we propose a general approach for integrated marketing. Specifically, to market a given item, we first generate the item-specific candidate users by a recommendation algorithm, and then select the typical users who have the best balanced utility scores and consumption/social entropy. Next, treating typical users as targeted customers, we study the problem of maximizing information awareness in viral marketing with these constrained targets. Along this line, we define it as a constrained coverage maximization problem, and propose three solutions: GMIC, LMIC and QMIC. Finally, extensive experimental results on real-world datasets demonstrate that our integrated marketing approach could outperform the methods that consider only targeted marketing or viral marketing.**

I. Introduction

Due to the highly fragmented mass media in today's society, reaching consumers by traditional marketing strategies (e.g., via print advertising and TV commercials) is increasingly difficult [1]. Meanwhile, the rapid growth of online user population on the Internet and mobile social world (e.g., Facebook) has attracted a great deal of attention and interest from marketers [2]–[6].

Indeed, there are generally two reasons that online social networks have become new resources and platforms for marketing. First, as the worlds of daily life, online life and the Internet technology become more inextricably linked, a rich set of profiles (e.g., the consumption histories and the social status) from a large number of social users are available for creative use in automatic marketing. For instance, these users and their profiles could help generate customer segments accurately, and one step further, pave the way for successful targeted marketing. Specifically, targeted marketing identifies the typical customers and concentrates the marketing efforts on these customers. Thus, it could make the promotion of the items easier and more cost-effective [7]; Second, the user communications, and more generally, the diffusion of information in social networks can be utilized to design profitable viral marketing strategies, since the information diffusion of some customers may influence other's purchasing decisions. Particularly, viral marketing identifies the seed users with the strongest influence in the network to maximize the word-of-mouth propagation of a product [4].

Though both targeted marketing and viral marketing are more effective than marketing directly to a specific person [4] (e.g., via personalized recommendation [8]), they could only exploit some specific characteristics of the social networks and have limited marketing performance. For instance, targeted marketing may miss some potential customers with high utilities, and in contrast, viral marketing could waste the time and energy on a large scale of unprofitable social users. Thus, it is necessary to integrate these two marketing strategies together to build a better, stronger business. Actually, one straightforward way for an integrated marketing is to first find a set/segment of the targeted users for targeted marketing, and then conduct viral marketing on these targeted users and other users in the social network. As a matter of fact, there are several challenges inherent in designing and implementing such a combined marketing strategy. First, since the number of potential users to an item may be huge, a common challenge is to select only a set of the most interested users for targeted marketing. The implicit constraint is that these targeted users should be both relevant and diverse, as this is the only way to precisely cover as many different users as possible. Second, different from the traditional viral marketing strategies where a set of the global influential seed users are selected, in our scenario, the seed users not only should be influential in the entire network but also should "cover" as many targeted users as possible. Thus, how to measure the effectiveness of the seed users becomes another challenge.

To address the challenges mentioned above, in this paper, we propose an integrated marketing approach. Specifically, to market a given item, we first generate many candidate users by a recommendation algorithm (i.e., item-based collaborative filtering). Then, we select a small set of typical users for targeted marketing from these candidate users by balancing

*Contact Author.

1550-4786/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDM.2014.93

the utility scores and the entropy. Here, the utility of one candidate user is also measured by collaborative filtering, and the entropy of the entire user set could be computed based on multiple features (e.g., preference diversity and social diversity). Then, treating typical users as constraints, we study the problem of maximizing information awareness in viral marketing with these constrained targets. Specifically, we formulate this constrained viral marketing as a constrained coverage maximization problem. In addition to proposing a naive greedy solution, we also establish mathematically sound approximations and bounds, which lead to convex optimization and globally optimal solutions. Finally, extensive experimental results on real-world datasets demonstrate that our integrated marketing approach outperforms the methods considering only targeted marketing or viral marketing.

To the best of our knowledge, this is the first attempt on a comprehensive study of marketing strategies that integrate targeted marketing and viral marketing in online social networks. Specifically, our solution identifies the most profitable potential users and accordingly selects the most influential seeds to optimize the marketing performance. Meanwhile, the proposed integrated marketing approach is a general framework and each step is open to some other algorithms.

II. Related Work

The related literature can be grouped into two categories. The first category includes the related works of exploiting online social networks for improving marketing performance. The second category includes a brief discussion of the studies on maximum set coverage.

Marketing on Online Social Networks. Designing marketing strategies using social network analysis has been studied by employing various techniques and approaches, e.g., marketing through the media like emails [9], social events [10], social search engine [11] and social web pages [12]. However, the vast amount of information tends to overwhelm marketers [13], and it is essential to figure out one or a few customer segments to target. Therefore, comparing to the entire marketing process (a set of activities for choosing target markets, understanding user behavior and providing superior user value), we mainly focus on reviewing the marketing strategies by identifying the targeted customers.

Typically, the decision of whether or not to market to a particular user is based solely on his/her profiles or the population segment to which he/she belongs [4]. For each marketing effort, the first critical challenge is to accurately infer user profiles based on the data available in social networks (e.g., social connections) [14] or the implicit information revealed by the user generated content on the web (e.g., the style preferences) [15]. User profile not only includes the information about user demographics, but also consists of user's social status, personal interests and preferences, etc. For instance, for marketing to the enterprise customers, Zeng et al. proposed to infer users' employment affiliation information from social activities by a classification method [16]. Cho et al. developed a way of modeling social user mobility, and this model could reliably predict the locations and dynamics of customer movement [17]. Jamali et al. designed a random walk model combining the trust-based and the collaborative filtering approach for inferring customer preference [18]. Actually, most of the techniques like classification and sequence discovery in data mining could be applied to customer segmentation and selection [19]. After segmenting customers, marketers can now offer differentiated marketing strategies to the targeted customer groups, such as personalized recommendation [8], [20], price advertising [21] and viral marketing [22].

Viral marketing takes advantage of the fact that customers in social networks are strongly influenced by the opinions of their peers. Thus, we could inexpensively promote a new product by marketing primarily to a set of the seed users with the strongest influence [4]. Along this line, there are generally three types of research directions: measuring the information transition probability between two neighbors [23], modeling the information propagation process [6], [24]–[26], and applying social influence to viral marketing [22]. Though it is important to learn the influence propagation probability between two neighbors, this problem is beyond the scope of this paper, and we mainly focus on the second and the third directions. For describing the dynamics of information propagation, the idea of Independent Cascade(IC) model [25] and Linear Threshold (LT) model [26] are widely used. Unfortunately, the influence spread (i.e., the expected number of nodes that will be influenced) computation under these models is #P-hard [6]. Thus, Monte-Carlo simulation, which is very time-consuming, is employed to approximate the influence. To avoid Monte-Carlo simulation, Aggarwal et al. [27] proposed a stochastic information flow model, and Xiang et al. [28] proposed a linear social influence (Linear) model. Due to the inefficiency of traditional information propagation models, most of existing work on viral marketing (also called as social influence maximization) has to make a tradeoff between effectiveness and efficiency. The typical approaches include CELF [29], PMIC [6], SIMPATH [30], IRIE [31] and UBLF [32]. To the best of our knowledge, only Ref. [33] studies the problem of viral marketing with some constrained target users. Unlike our scenario, Ref. [33] tries to find the minimum size seed set for activating at least a given number of nodes in the targeted set.

Maximum Coverage Problem. Actually, the above social influence maximization problem (viral marketing) is a variant of the set cover problem which has been well studied [34]. The reason is that this type of problems has broad applications, such as recommendation [35], ensemble pruning [36], tag selection [37], and document summary [38]. Though these problems are generally NP-hard, the optimization functions are usually monotone and submodular. Thus, using a simple greedy algorithm we could achieve a solution that is guaranteed to have an approximation ratio of $(e-1)/e \approx 0.6321$ (here ,e is the base of the natural logarithm). However, if there are some constraints, e.g., budgets [39], prior domain knowledge [40] or the must-cover constraint (as shown in this paper), more complex solutions should be designed.

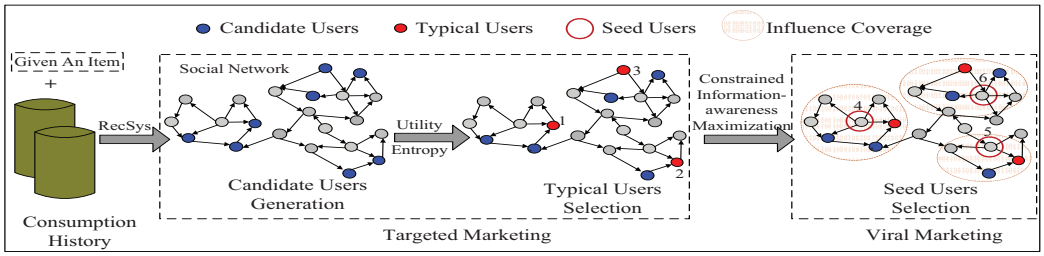


Fig. 1. Flowchart of the marketing strategy.

TABLE I
SEVERAL IMPORTANT MATHEMATICAL NOTATIONS.

Notations	Description
$f_{i\to j}$	influence from node i to j, j-th entry of $\mathbf{f}_i$
R	information-awareness matrix, $R_{ji}=1$ if node j is covered by i
I	the set of the items (e.g., products)
T	the set of the training users
U	the set of the typical users
S	the set of the seed users
K_U	size of the typical user set
K_S	size of the seed user set
$r(u)$	the utility score of a user u
$H_0(U)$	the normalized joint entropy of the entire set of users U
λ	parameter, balance the effect between utility and entropy

III. INTEGRATED MARKETING

In this section, we first give the preliminaries and problem formulation. Then, we describe the integrated marketing approach in detail. Actually, Fig. 1 shows the flowchart of the proposed marketing approach. Given the item for marketing, we first generate many candidate users that may like this item (e.g., the users with high utilities) by a recommendation algorithm. Since the number of these candidate users is usually too large for effective marketing, we then select a comparably small set of targeted users (typical users, e.g., users 1, 2, 3 in the figure) for targeted marketing by jointly modeling their utility scores and entropy. Next, we treat these targeted users as constraints when conducting viral marketing. Thus, we select a set of seed users (e.g., users 4, 5, 6 in the figure), the influence of whom could not only cover the targeted users but also cover as many other users as possible. In this way, marketing on these seed users will lead to both the maximum adoption and maximum information awareness of this item. Each step is illustrated in the following subsections. For better illustration, Table I lists some mathematical notations.

A. Problem Statement and Formulation.

1) Preliminaries: For better introducing our integrated marketing approach, we first show the general notations and issues related to the traditional social influence modeling and social marketing problem. Here, we start from viral marketing, i.e., the social influence maximization problem. Assume $G=(V,E,\mathbf{T})$ is a social network, where $V=\{1,2,...,n\}$ is the node/user set and edge set E represents all the connections between nodes. $\mathbf{T}=[t_{ij}]_{n*n}$ is the transition matrix for information propagation, i.e., t_{ij} represents the information propagation probability from node i to node j. If there is a connection from i to j in E, then $t_{ij}>0$, otherwise $t_{ij}=0$. Note that we assume $\mathbf{T}$ is given in this paper. Also, G is usually assumed to be directed, as influence propagation is specific to direction in the most general case [27].

With graph G, viral marketing uses the influence propagation models (including IC [25], LT [26] and Linear [28]) to compute the influence spread $\mathbf{f}_i$ for each node i following some defined rules. Specifically, $\mathbf{f}_i=[f_{i\to 1},f_{i\to 2},...,f_{i\to n}]'$, a $n\times 1$ vector, denotes the influence distribution of node i on each node in the network. Thus, the total influence spread of node i in network equals to the sum of the influence of node i to other nodes, namely $f_{i\to V}=\sum_{j\in V} f_{i\to j}$. We can see that $f_{i\to V}$ is actually the expected number of the nodes that will be influenced by i. Now, let's take Linear model as an example for illustrating the computation of $f_{i\to j}$ and $f_{i\to V}$ [28]: if i equals to j, then $f_{i\to j}=\alpha_i$ (e.g., 1) and α_i is the prior self-confidence of node i for spreading the information; Otherwise, $f_{i\to j}=d_j\sum_{k\in N_j} t_{kj}f_{i\to k}$, where $N_j=\{u\in V\mid (u,j)\in E\}$ and parameter $d_j\in(0,1]$ is the damping coefficient. Thus, the computation of $f_{i\to j}$ follows an iterative process, and the iterative computation for the entire $\mathbf{f}_i$ (also $f_{i\to V}$) quickly converges in $O(|E|)$ time.

Given the $f_{i\to V}$ computed by a specific influence propagation model, traditional work on viral marketing usually aims at finding a set of seed nodes S ($S\subset V$ and $|S|=K_S$) with the biggest $f_{S\to V}$. In this way, if using the seeds in S to spread the information, we could get the maximum influence spread, i.e., the expected number of successfully influenced nodes.

However, there are two limitations of the traditional methods. First, the seeds with the maximum influence spread may not result in the maximized information awareness on the network. Let's consider an example as shown in Fig. 2, where we have a toy social network and two seed candidates highlighted in black for viral marketing, i.e., node 1 and node 2. Fig. 2(a) and Fig. 2(b) illustrate the simulated influence spread results for these two candidate seeds, respectively. Since node 1 successfully made node 4, 5 and 6 active (Fig. 2(a)), and node 2 only influenced node 4 and 5 (Fig. 2(b)), thus node 1 will be chosen as the seed for influence spread. However, if we observe the network more carefully, we could find that it may be more appropriate to choose node 2 as the seed rather than node 1. The reason is that node 2's influence spread distribution is much more balanced than node 1, and thus it has the ability to make more people aware of the specific information. Here, we could measure the information awareness by the probability that targeted node knows about the information (e.g., about an item), and this is a relaxed definition of "being influenced"(which could be interpreted as really buying that item). Actually, in some marketing and advertising applications [3], [33], the service providers care

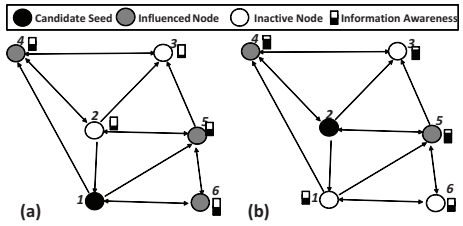


Fig. 2. A Motivating Example.

more about the number of people aware of their products rather than just the expected number of influenced ones. That is, we should also pay attention to the real distribution of the information spread $\mathbf{f}_i$ of candidate seed i.

Second, the selected seeds may not cover some important users with high utilities (i.e., the users that have high probability of buying this specific item). For instance, suppose we now want to promote a game product on the social network shown in Fig. 2 and we know that node 3 (a teenage boy) is possibly interested in this game, then we should also choose node 2 as the seed for marketing rather than node 1, since the influence of node 2 could cover the targeted node 3 very well.

In summary, in some real-world scenarios, it is important to make more people aware of the specific information, and also some of the important individuals (the targeted users selected in targeted marketing) must be covered (i.e., they should be influenced by the selected seeds so as to be aware of this information). To that end, in the following, we explain the formulations of our integrated marketing approach.

2) Problem Formulation: First, we introduce an asymmetry "information-awareness matrix" R, where $R_{ji} = 1$ if node j is covered by node i, and $R_{ji} = 0$ otherwise. Here, the "coverage" is measured by the pairwise influence strength, i.e., if $f_{i\to j} > t$ (t is a threshold (e.g., 0.1)) then we assume node j is covered by node i and $R_{ji} = 1$. In other words, under the influence of node i, node j becomes aware of this information. Actually, $f_{i\to j}$ can be computed by any existing influence model, e.g., Linear [28]. According to the previous discussion (e.g., the example shown in Fig. 2), the benefits of using R rather than $f_{i\to j}$ are: First, the balance of the influence distribution could be considered. Thus, we may evaluate the performance of the candidate seeds more precisely; Second, the final goal of the integrated social marketing could be more easily achieved, as we can conveniently judge if one targeted node is covered or not. However, there is also one underlying shortage of R: When evaluating a set of seeds, the effect of influence enrichment between these seeds are not included (i.e., $f_{i\to j} < t$ and $f_{k\to j} < t$, while $f_{\{i,k\}\to j}$ maybe larger than t). Since the seeds are usually far away from each other in the social network and there are limited influence enrichment, we just omit this phenomenon in this paper.

Then, the problem of the non-constrained information-awareness maximization problem can be formulated as:

$$\max_{p} \|Rp\|_0,$$

where p is a $n \times 1$ vector, representing whether the node i is selected as a seed ($p_i = 1$) or not ($p_i = 0$). In practice this maximization problem is often constrained by the number of selected seed nodes, e.g., K_S:

$$\mathbf{1}'p = \sum_i p_i = K_S,$$

where $\mathbf{1} \in \mathbb{R}^{n\times 1}$ is a vector of ones.

In targeted marketing, denote U as the set of targeted customers that must be covered by the selected seeds, then we have another constraint:

$$\{Rp\}_j \geq 1,\ \forall\, j \in U.$$

In summary, the problem that we would like to optimize for the integrated marketing is as follows

$$\begin{aligned} \max_{p}\ & \|Rp\|_0, \\ s.t.\ & \mathbf{1}'p = K_S, \\ & \{Rp\}_j \geq 1,\ \forall j \in U, \\ & p_i \in \{0,1\},\ \forall i \in V. \end{aligned} \tag{1}$$

In the following, we first illustrate the strategy of selecting typical users U in targeted marketing. Then, we propose several solutions for this constrained information-awareness maximization problem to get better marketing performance.

B. Targeted Marketing

In this subsection, we describe the way of selecting typical users U for targeted marketing by jointly modeling the customers' utility scores and their entropy.

1) Problem Statement and Solution: In this paper, we formulate the typical user selection problem as follows. Given an item for marketing and a small set of users T who have shown their preferences to this item (e.g., have bought this item in the training set), we seek another set of users U (who have not expressed their preferences to this item, i.e., $U \cap L = \varnothing$) that are the most "typical" to the marketer in the context of marketing this item on the social network.

It is possible to represent "typical" by many different *indicators*. Without loss of generality, this paper focuses on two of them, one user-level indicator: utility, and one set-level indicator: diversity. The intuition is that a useful selection of typical users contains *relevant* users referring to the given item. Furthermore, the users should be able to cover most of the segmentations of the candidate users (i.e., with *diverse* characteristics). Then, the goal is to find a subset U^* that maximizes the objective function $g(U)$:

$$g(U) = \lambda \sum_{u\in U} r(u) + (1-\lambda) H_0(U), \tag{2}$$

where $r(u)$ represents the utility score of a user u and the diversity function $H_0(U)$ is the normalized joint entropy of the entire set of users U. We balance them by specifying a parameter λ. The maximization problem is that:

$$U^* = \underset{U\cap L=\varnothing, |U|=K_U}{\arg\max}\ g(U). \tag{3}$$

Before introducing the way of computing r and H_0, we should note that the above formulation is similar to that in

Ref. [41], where a set of social media responses to online news articles are selected. Moreover, following the similar proof strategy given in Ref. [41], the entropy of a user set is a monotonic non-decreasing submodular function [42]. Combining with the way of computing r (shown later), we could easily conclude that $g(U)$ is also a monotonic non-decreasing submodular function. For this kind of function, we do not have to exhaustively search the space of all possible user subsets, but use a simple greedy algorithm which guarantees that the approximation is within $(1-1/e)$ of the optimal result. Initially, $U = \varnothing$ and we select a set of candidate users by computing $r(u)$. The candidate users are those who may prefer the given item in the near future. Having said that the number of candidate users are usually too large to market, and then we select the typical ones into U from these candidate users (following Eq. (2)): First, the user with the highest $r(u)$ is put into U. After that we iteratively add a new user u (selected from the candidate user set), as long as u could provide the biggest boost into objective function $g(U)$ if it was added into the set U. This iterative process will keep running until the targeted user set size is K_U.

2) Utility and Entropy Computation: We show the way of computing the utility score for a single user ($r(u)$) and the entropy score for a set of users ($H_0(U)$), respectively. First, given an item and a set of users T who have consumed this item, we use an item-based collaborative filtering [8] to compute $r(u)$. As a kind of recommendation algorithm, item-based collaborative filtering could address the data sparsity problem very well and thus generate high quality candidate/relevant user recommendations to each item. Since the focus of this paper is not to devise more sophisticated means to calculate user-item similarity, we choose Jaccard measure, which has performed well in binary preference data [43].

$$r(u) = sim(u, a) = |I_u \cap I_a|/|I_u \cup I_a|, \tag{4}$$

where I_u are the items that user u likes and I_a are the items that are most similar (i.e., often consumed together) to item a. After that, candidate user set is generated by selecting top users with the largest Jaccard similarities.

Second, we show the way of computing $H_0(U)$, which is motivated by the method in Ref. [41]. Given a user set U and the features (will be introduced later) used to collectively select typical users, we treat these features as binary random variables. Let $H_0(U)$ denote the normalized entropy of the user set U, and it is measured by $H_0(U) = H(U)/log(d)$, where d is the number of binary feature variables and $H(U) = -\sum_{i=0}^{d} p(f_i = 1)log(p(f_i = 1))$ and $p(f_i = 1)$ is the probability that feature f_i has the value of 1 given all users in U. The intuition is that we favor adding the users with different non-zero features from those already in U to increase diversity. In this paper, we focus on two types of features for computing the entropy, the consumption feature and the social feature. Specifically, for the consumption feature, each item in the system stands for a feature variable, and if the given user consumed this item then $f_i = 1$. For the social feature, each social user in the system is a feature variable, and if the given user has social connections with this user then $f_i = 1$. We can see that these two features capture the diversity of typical user set from different aspects, i.e., preference diversity and social diversity, respectively.

At last, λ could balance the effect between utility and entropy: if $\lambda = 0$, the entropy is highlighted; if $\lambda = 1$, only the utility is considered. We should note that, this targeted marketing framework is a general and open model which could handle more indicators and features.

C. Constrained Viral Marketing

We describe the solution for viral marketing to maximize the information awareness with the constraint that the targeted typical users must be covered. Unfortunately, optimizing the L_0 norm in Eq. (1) is NP-hard, thus in the following we explore both natural heuristics and mathematically sound relaxations to derive the optimal solutions. Specifically, we will introduce a naive greedy algorithm (GMIC), and algorithms using linear (LMIC) and quadratic (QMIC) programming, respectively.

1) Greedy Algorithm (GMIC): Our algorithm GMIC (**G**reedy for **M**aximum **I**nformation **C**overage) for Eq. (1) is a variant of the set cover solution. Initially, $S = \emptyset$. At each iteration, it adds a new node i_{max} into S (i.e., $p_{i_{max}} \leftarrow 1$), where i_{max} is selected from the nodes that maximize the increment on the coverage of the targeted users U. When there are multiple candidates maximizing the targeted coverage, we choose the one that leads to the maximum information coverage on the nodes of the entire network.

Algorithm 1: GMIC(R, K_S, U).

```
p ← 0 ∈ R^{n×1}.
for k = 1 to K_S do
    C ← {j | Σ_{i:p_i=1} R_ji > 0}.
    i_k ← arg max_{i:p_i=0} Σ_{j∈U\C} R_ji.
    i_max ← arg max_{i∈i_k} Σ_{j∈V\C} R_ji.
    Set i_max as any i ∈ i_k if multiple candidates returned.
    p_{i_max} ← 1.
end for
```

2) Linear Programming (LMIC): Having said that the optimization of L_0 norm is generally NP-hard, we can replace the L_0 norm with L_1 norm, as done in BP (Basis Pursuit) [44] and Lasso (least absolute shrinkage and selection operator) [45]. Therefore, we would like to maximize $\|Rp\|_1 = \mathbf{1}'Rp$ instead of $\|Rp\|_0$. Also, there are several ways to relax the discrete constraints $p_i \in \{0, 1\}$. Here we adopt a simple approach with $p_i \in [0, 1]$. Once the relaxed p is returned, we can select the K_S nodes with higher values in p. In summary, we would like to propose another solution LMIC (**L**inear programming for **M**aximum **I**nformation **C**overage) by

$$\begin{aligned} \max_{p} \quad & \mathbf{1}'Rp, \\ s.t. \quad & \mathbf{1}'p = K_S, \\ & \{Rp\}_j \geq 1, \ \forall j \in U, \\ & 0 \leq p_i \leq 1, \ \forall i \in V. \end{aligned} \tag{5}$$

3) Quadratic Programming (QMIC): Indeed, the L_1 norm is an upper bound of L_0 norm in our problem. For the maximization problem, it is intuitively better to work with a lower bound. To this end, we provide the following lower bound of $\|Rp\|_0$:

Theorem 1 *$\|Rp\|_0 \geq \frac{3}{2}\mathbf{1}'Rp - \frac{1}{2}p'Qp$, where $Q = R'R$.*

Proof: It is straightforward to see

$$\begin{aligned}\|Rp\|_0 =& \| \cup_{i:p_i=1} \{j|R_{ji}=1\}\| \\ \geq& \sum_{i:p_i=1} \|\{j|R_{ji}=1\}\| \\ & - \sum_{\substack{i_1,i_2:i_1<i_2 \\ p_{i_1}=p_{i_2}=1}} \|\{j|R_{ji_1}=1\} \cap \{j|R_{ji_2}=1\}\| \\ =& \frac{3}{2}\sum_{i:p_i=1} \|\{j|R_{ji}=1\}\| \\ & - \frac{1}{2}\sum_{i_1,i_2:p_{i_1}=p_{i_2}=1} \|\{j|R_{ji_1}=1\} \cap \{j|R_{ji_2}=1\}\|.\end{aligned}$$

Now we can conclude the proof with

$$\sum_{i:p_i=1} \|\{j|R_{ji}=1\}\| = \sum_{i:p_i=1} \mathbf{1}'R_{*i} = \mathbf{1}'Rp,$$

and

$$\begin{aligned}&\sum_{i_1,i_2:p_{i_1}=p_{i_2}=1} \|\{j|R_{ji_1}=1\} \cap \{j|R_{ji_2}=1\}\| \\ =& \sum_{i_1,i_2:p_{i_1}=p_{i_2}=1} Q_{i_1i_2} = p'Qp.\end{aligned}$$

■

With Eq. (1), similar to the formalization in Eq. (5), we can optimize

$$\begin{aligned}\max_{p} \quad & \frac{3}{2}\mathbf{1}'Rp - \frac{1}{2}p'Qp, \\ s.t. \quad & \mathbf{1}'p = K_S, \\ & \{Rp\}_j \geq 1, \ \forall j \in U, \\ & 0 \leq p_i \leq 1, \ \forall i \in V.\end{aligned} \tag{6}$$

In the following, we call this solution as QMIC (**Q**uadratic programming for **M**aximum **I**nformation **C**overage).

Actually, the biggest difference between the greedy algorithm (GMIC) and the approximating algorithms (LMIC and QMIC) is on their convexities. Both LMIC and QMIC are convex and will result in a global optimum respectively, only if the constraints could be satisfied. Here, we will leave the situation when those constraints may not be satisfied as a future work. In contrast, GMIC is non-convex and it will stuck at a local optimal solution, which might be far away from better alternatives. In summary, GMIC is able to return a result for any situations without performance guarantee, while LMIC and QMIC could return the global optimized result under some constraints. Notice that, since both LMIC and QMIC are the approximating methods, their globally optimized results may not be better than the local output of GMIC (this could be observed in the experiments).

IV. Experiments

We conduct experiments on real-world datasets to demonstrate: (1) The effects of parameter λ for typical user selection in targeted marketing; (2) The performance of our integrated marketing algorithms; (3) The overlap of the seed users.

A. Experimental Setup

Datasets. We choose two datasets from different domains: Ihou and Epinions. Ihou is an online Karaoke dataset that is collected from ihou.com[1], which contains all the singing (consumption) records of the users, and meanwhile, the users' social connections (follower-followee relationships) between July 2011 and April 2012; Epinions[2] is a dataset about the user ratings on the articles and it also contains a directed user-trust network [46]. Thus, the items for marketing are the songs or the articles, respectively. Detailed information of these datasets can be seen in Table II. We select the users who have consumed at least 5 items. Then, for each user's consumption record, we split it into a training set and a test set, by selecting the first 20 percentage of the consumption to be part of the training set and the remaining ones to be part of the test set. In this way, we could treat these items as the cold-start products that need social marketing.

TABLE II
Statistics of the data set.

Data	#Users	#Items	#Social Edges	#Consumptions
Ihou	86,192	9,588	455,424	766,861
Epinions	114,467	112,194	717,667	13,261,571

Benchmark Methods. Following their names in the step of viral marketing, we call the specific algorithms under our proposed integrated marketing solution as GMIC, LMIC and QMIC, respectively. Specifically, for each algorithm (e.g., GMIC), we first use Eq. (2) to find typical users, and then select the corresponding constrained viral marketing strategy (e.g., GMIC). We compare with several benchmark methods:

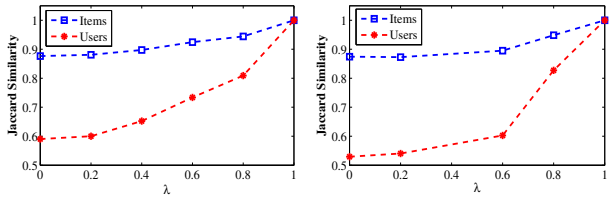
- *TU* is short for Targeted Users. In this method, we directly choose the typical users selected from targeted marketing as seeds for the following viral marketing.
- *CELF* is short for "Cost Effective Lazy Forward" [29] implemented under Independent Cascade (IC) model [25]. To the best of our knowledge, CELF (a greedy algorithm) is the most effective (though not the efficient one) solution for the traditional viral marketing.
- *RGMIC*, *RLMIC* and *RQMIC*. These three methods are the GMIC, LMIC and QMIC with randomly selected typical users.

We use TU and CELF, which considers only targeted marketing or viral marketing, respectively, to demonstrate the benefits provided by our integrated marketing approach. Meanwhile, the comparisons with RGMIC, RLMIC and RQMIC help us test the effectiveness of our selected typical users.

All the experiments were performed on a server of Windows 64-bit operating system with 2 GHz 24-Core Intel(R) XeonR E5-2620 CPU and 128GB of main memory.

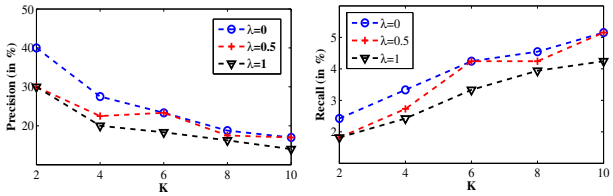
[1] http://www.ihou.com/

[2] http://www.trustlet.org/wiki/Extended_Epinions_dataset



(a) Ihou. (b) Epinions.

Fig. 3. The Jaccard similarity of the user sets and their consumed items.



(a) Precision. (b) Recall.

Fig. 4. The Precision and Recall results of the typical users (Ihou).

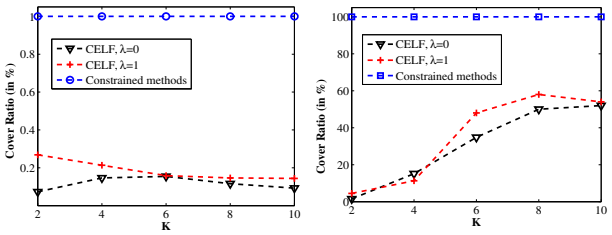
B. Typical User Selection

We compare the results of the typical user selection under different λ (Eq. (2)). For better measuring users' utilities by Eq. (4), we focus on the items with $|T|$ (number of the users who have consumed this item in the training set) larger than 10. First, Fig. 3 illustrates the average Jaccard similarities between the results under $\lambda = 1$ (i.e., no diversity is included) and other λ (in [0,1)). Specifically, for each item and λ, we run Eq. (2) to get a set of 10 typical users from the Top-50 candidate users with the highest $r(u)$. The Jaccard similarity of the user sets and the Jaccard similarity of the consumed items of these user sets are reported. Similar results could be observed from both Fig. 3(a) and Fig. 3(b): The smaller the λ, the bigger the difference between the user sets and the items. This implies that Eq. (2) and the features that we use are able to help select different typical users under different λ, and these users also have different item preferences.

Second, we compare the different typical user selection results in terms of user recommendation. Here, we directly test if the K typical users selected by Eq. (2) will finally consume the given item or not, and we choose the "Precision" and "Recall" [8] as the evaluation metrics. These recommendation results under different λ is shown in Fig. 4. We take Ihou as an example and similar results could be observed from Epinions. We can see that precision decreases (recall increases) when the number of typical users becomes larger. This implies that our item-based collaborative filtering could estimate the user preference very well as the higher ranked users have more probability to consume the item. Another interesting observation is that the best recommendation performance is achieved when $\lambda = 0$, while $\lambda = 1$ performs the worst. This means our method of introducing diversity could benefit both the recommendation accuracy and the information coverage of viral marketing (this will be shown in the following).

C. Marketing Performance Comparison

In this subsection, we compare the effectiveness and efficiency of each marketing solution. Specifically, we evaluate effectiveness from three different aspects: the coverage on the targeted typical users (Targets Coverage), the coverage on the ground truth users, i.e., the users in the test set (Test Set Coverage); and the coverage on other users in the entire network (Global Coverage). Without loss of generality, we compute the information cover matrix R by the Linear influence model [28] and the threshold t for measuring information awareness is set to be 0.1 for Ihou and 0.01 for Epinions (according to the sparsity), respectively. Notice that, even the Linear algorithm is fast enough, it is still impossible to compute the influence spread for all the nodes, since R will be too large. Thus, we first run PageRank [47] and only compute the influence spread $\mathbf{f}_i$ for 1,000 nodes with the highest PageRank values. In this way, we have 1,000 candidate seeds (columns) in R. For the sake of convenience, we let the size of the seed set be equal to the size of the typical user set ($K = K_U = K_S$) making sure that the constraints in LMIC and QMIC could be satisfied, and we choose comparatively small K ranging in [2,10].



(a) Ihou. (b) Epinions.

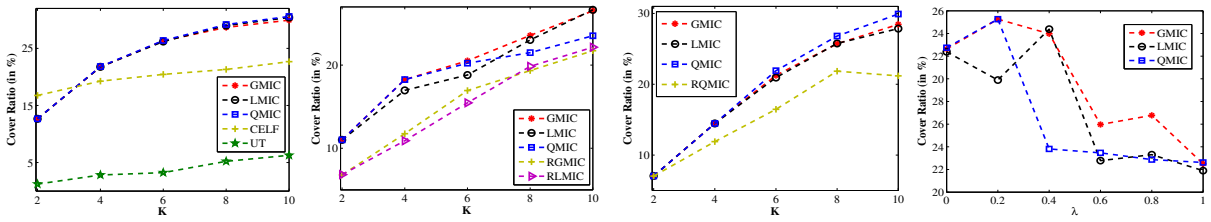
Fig. 5. Coverage on the targeted typical users.

Targets Coverage. Fig. 5 shows the coverage results on targeted typical users. Since most of the solutions (i.e., G-MIC, LMIC, QMIC, RGMIC, RLMIC, RQMIC and UT) treat typical users as constraints when selecting seeds, they could certainly get the 100% cover ratio on typical users as long as $K_S \geq K_U$. For consistency, we denote all these methods as "Constrained methods" in Fig. 5. In contrast, the global optimization method CELF does not suffer from such a constraint. Thus, CELF could not cover most of the typical users , no matter what λ (e.g., 0 and 1) and datasets are used.

Test Set Coverage. Fig. 6 illustrates the final information coverage on the ground truth users. Here, the ground truth users are those that consumed this specific item in the test set. Compared to the targets coverage, cover ratio on the test set is an even more straightforward metric for evaluating each marketing strategy. Fig. 6(a)-(d) in the first line are the test set coverage results for Ihou and the four sub-figures in the second line are for Epinions. Since only methods GMIC, LMIC and QMIC are sensitive to parameter λ, we draw them under different λ settings (i.e., 0, 0.5, 1). For better illustration[3], we split the benchmark methods into different subfigures, e.g., CELF and UT are only compared in the figure when $\lambda = 0$. From Fig. 6(a)-(c) we can see that our methods could cover much more users in the test set of Ihou for each λ. Similar results on Epinions could be also observed in Fig. 6(e)-(g).

Meanwhile, we investigate more details on the effect of tuning λ in terms of the cover ratio of the selected seeds.

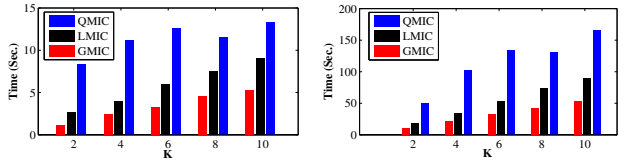
[3]Actually, for both Test Set Coverage and Global Coverage, we magnify the results of benchmark UT 100 times to make it be comparable with others.



(a) Ihou ($\lambda = 0$). (b) Ihou ($\lambda = 0.5$). (c) Ihou ($\lambda = 1$). (d) Ihou (K=6).

(e) Epinions ($\lambda = 0$). (f) Epinions ($\lambda = 0.5$). (g) Epinions ($\lambda = 1$). (h) Epinions (K=6).

Fig. 6. Coverage on the ground truth users in test set.



(a) Ihou. (b) Epinions.

Fig. 8. The average (per item) computational costs.

To this end, we set λ ranging from 0 to 1, with step 0.2, and compute the corresponding cover ratio on test set with K=6. The results on two datasets are shown in Fig. 6(d) and Fig. 6(h), respectively. From these figures we can see that adding diversity in our objective function Eq. (2) could generally help select more typical users, since much more ground truth users in the test set are covered. According to these two figures, it is better to set λ no bigger than 0.5.

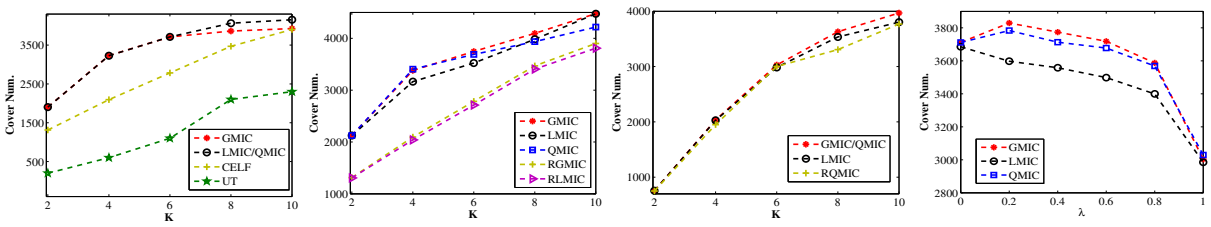
Global Coverage. Indeed, test set coverage only measures the observed/direct profit. When evaluating the marketing performance, the potential profit is also an important metric and this can be measured by the expected information coverage on the entire network. The performance of each method under this metric (i.e., $\|Rp\|_0$) is reported in Fig. 7. Actually, the arrangement in Fig. 7 is similar to that in Fig. 6. From this figure (i.e., Fig. 7(a)-(c) and Fig. 7(e)-(g)), we can see that our methods generally outperform the baselines. However, there is one exception, i.e., CELF also performs very well on Epinions in Fig. 7(e). The reason is that CELF outputs the seeds with the biggest influence spread $f_{S\to V}$ (i.e., $\sum_{S\in V} f_{S\to j}$) while our methods optimize the constrained information coverage with R. Thus, under this global coverage metric, CELF may perform better (e.g., as shown in Fig. 7(e) when the constraints dominate our selection) or worse (as shown in Fig. 7(a) when threshold t is large enough to make a big difference between influence spread and information coverage) than our methods. Also, the differences observed from some of the figures seems to be small, that's because the range of the y-coordinate (Cover Num.) is very large. Actually, these improvements are significant, e.g., GMIC could cover 930 users more than CELF when $K = 10$ in Fig. 7(e). Meanwhile, the global coverage also changes in terms of different λ (Fig. 7(d) and Fig. 7(h)), and the trend is similar to that in the test set coverage.

In summary, combining the results in Fig. 5, Fig. 6 and Fig. 7, we conclude that: (1) Compared with the methods using randomly selected typical users (i.e., RMIC, RLMIC and RQMIC), the viral marketing methods with the typical users selected based on Eq. (2) (i.e., GMIC, LMIC and QMIC) could better cover the users in the test set and other users in the network. This again demonstrates that our objective function (Eq. (2)) helps find both relevant and diversified typical users; (2) The traditional viral marketing method (i.e., CELF) can not deal with the scenarios when social marketing meets constrained customers, and the constrained customers are usually not influential enough to spread the specific information (i.e., UT). In contrast, our proposed solutions could maximize the information coverage with the constrained customers. (3) GMIC, LMIC and QMIC perform similarly, and their differences have been discussed in Section III-C.

Running Time. We compare the computational efficiency, and the results are shown in Fig. 8. We only present the computing time of GMIC, LMIC and QMIC because the time cost of RGMIC, RLMIC and RQMIC are the same with these algorithms, and UT almost has no time consumption while CELF is too time-consuming for a large number of Monte-Carlo simulations. Similar results could be observed from Fig. 8(a) and Fig. 8(b), where we can see that among our proposed algorithms, the naive greedy algorithm GMIC is the most efficient for locating influential seeds to market an item.

D. Overlap of Seed Users

We provide a further understanding of the relations between each marketing solution. First, we record the selected seed users of each algorithm, and the seeds for all the items are summarized together to stand for this specific method (e.g., GMIC). Then, the Jaccard similarity of these 7 seed sets are demonstrated in Fig. 9, and we do not show the results of UT algorithm, since its seeds (i.e., typical users) have little overlap with others. Here, we fix the size of each seed set as



(a) Ihou ($\lambda = 0$). (b) Ihou ($\lambda = 0.5$). (c) Ihou ($\lambda = 1$). (d) Ihou (K=6).

(e) Epinions ($\lambda = 0$). (f) Epinions ($\lambda = 0.5$). (g) Epinions ($\lambda = 1$). (h) Epinions (K=6).

Fig. 7. Coverage on the entire social network.

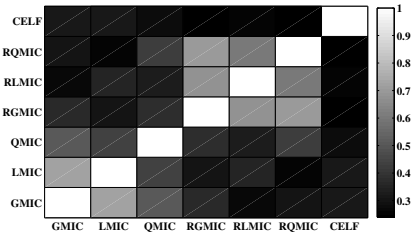


Fig. 9. The Jaccard similarity of the seed users (Epinions).

10, and fix $\lambda = 0$ for GMIC, LMIC and QMIC. Meanwhile, we just show the results on Epinions dataset and the similar results on Ihou are omitted due to the limited space.

Several interesting observations could be found in Fig. 9. For instance, the seeds output by GMIC, LMIC and QMIC have high overlap with each other, and the random version of these methods (i.e., RGMIC, RLMIC and RQMIC) also perform similarly. However, there is little overlap between these two types of algorithms. This implies that the different typical users will lead to different seed users for GMIC, LMIC and QMIC. Another observation is that CELF seems to be the most dissimilar one. Actually, as a non-personalized marketing strategy, CELF only selects the global influential nodes, and this is different from other methods which could generate seed users for each item. Meanwhile, it does not consider the influence distribution of each candidate seed.

V. Discussion

We discuss the advantages and limitations of this study. From the experimental results, we can see that the proposed integrated marketing approach works well for social marketing with constrained customers. Specifically, the selected typical users are both relevant and diverse, and thus could represent the potential customers of one specific item. Given these typical users as the constraints, our greedy method GMIC and the approximating algorithms (LMIC and QMIC) could lead to the maximum information coverage.

As a general framework, each step of our integrated marketing approach may be further improved in the future. First, in this paper, we only use limited information and metrics to select the typical users for targeted marketing. We believe this process should be much more complicated in the real-world marketing. Thus, we plan to incorporate more features and domain knowledge for the better definition and selection of typical users. Second, the algorithms designed for viral marketing may not perform well when the size of the seed set is smaller than the size of the typical users. Since the programming methods (LMIC and QMIC) may find no feasible solutions, while the GMIC method can only return local solution without performance guarantee. Then, how to find reliable marketing solutions for such a situation will be a very challenging research problem. Third, the performance (e.g., robustness) of our discoveries will be tested through more experiments. For instance, we plan to try more experimental settings (using other influence models to get R) on even larger datasets. Last but not least, for better marketing, we would like to figure out other factors(e.g., contexts [48] or significant events) beyond social influence that have impact on the consumption behaviors of social customers.

VI. Conclusion

In this paper, we provided a focused study on the integrated social marketing problem. Our target is to maximize the information coverage of some of the carefully selected typical users and maximize the information coverage on the entire social network simultaneously. Along this line, we first generated many candidate users by item-based collaborative filtering. Then, we selected the typical users from these candidate users for targeted marketing, and this was finished by balancing the users' utility scores and their entropy. Next, we treated these item-specific typical users as constraints and proposed three viral marketing solutions, GMIC, LMIC and QMIC, for finding a set of seed users to solve this constrained information awareness/coverage problem. Finally, extensive experimental results on real-world social network datasets demonstrated the effectiveness of our proposed marketing approach. We hope this study could lead to more future work.

ACKNOWLEDGEMENTS

This research was partially supported by grants from the National Science Foundation for Distinguished Young Scholars of China (Grant No. 61325010), the Natural Science Foundation of China (Grant No. 71329201), the Fundamental Research Funds for the Central Universities of China (Grant No. WK0110000042), the Anhui Provincial Natural Science Foundation (Grant No. 1408085QF110), and the National Science Foundation (NSF) via grant number CCF-1018151. Qi Liu gratefully acknowledges the support of the Youth Innovation Promotion Association, CAS.

REFERENCES

[1] K.-Y. Wang, I. Ting, H.-J. Wu *et al.*, "Discovering interest groups for marketing in virtual communities: An integrated approach," *Journal of Business Research*, vol. 66, no. 9, pp. 1360–1366, 2013.

[2] J. Hartline, V. Mirrokni, and M. Sundararajan, "Optimal marketing strategies over social networks," in *WWW*. ACM, 2008, pp. 189–198.

[3] J. M. Chan and R. Yazdanifard, "How social media marketing can influence the profitability of an online company from a consumer point of view," *Journal of Research in Marketing*, vol. 2, no. 2, pp. 157–160, 2014.

[4] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *SIGKDD*, 2002, pp. 61–70.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *SIGKDD*. ACM, 2003, pp. 137–146.

[6] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *SIGKDD*. ACM, 2010, pp. 1029–1038.

[7] L. Liu, Z. Yang, and Y. Benslimane, "Conducting efficient and cost-effective targeted marketing using data mining techniques," in *GCIS*. IEEE, 2013, pp. 102–106.

[8] P. B. Kantor, L. Rokach, F. Ricci, and B. Shapira, *Recommender systems handbook*. Springer, 2011.

[9] O. V. Pavlov, N. Melville, and R. K. Plice, "Toward a sustainable email marketing infrastructure," *Journal of Business Research*, vol. 61, no. 11, pp. 1191–1199, 2008.

[10] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han, "Event-based social networks: linking the online and offline social worlds," in *SIGKDD*. ACM, 2012, pp. 1032–1040.

[11] D. Horowitz and S. D. Kamvar, "The anatomy of a large-scale social search engine," in *WWW*. ACM, 2010, pp. 431–440.

[12] R. R. Yager, "Targeted e-commerce marketing using fuzzy intelligent agents," *Intelligent Systems and their Applications*, vol. 15, no. 6, pp. 42–45, 2000.

[13] R. D. Wilson, "Using online databases for developing prioritized sales leads," *Journal of Business & Industrial Marketing*, vol. 18, no. 4/5, pp. 388–402, 2003.

[14] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *WWW*. ACM, 2010, pp. 61–70.

[15] J. S. Alowibdi, U. A. Buy, and P. Yu, "Empirical evaluation of profile characteristics for gender classification on twitter," in *ICMLA*, vol. 1. IEEE, 2013, pp. 365–369.

[16] G. Zeng, P. Luo, E. Chen, and M. Wang, "From social user activities to people affiliation," in *ICDM*. IEEE, 2013, pp. 1277–1282.

[17] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: user movement in location-based social networks," in *SIGKDD*. ACM, 2011, pp. 1082–1090.

[18] M. Jamali and M. Ester, "Trustwalker: a random walk model for combining trust-based and item-based recommendation," in *SIGKDD*. ACM, 2009, pp. 397–406.

[19] E. W. Ngai, L. Xiu, and D. C. Chau, "Application of data mining techniques in customer relationship management: A literature review and classification," *Expert systems with applications*, vol. 36, no. 2, pp. 2592–2602, 2009.

[20] H. Zhu, H. Xiong, Y. Ge, and E. Chen, "Mobile app recommendations with security and privacy awareness," in *SIGKDD*. ACM, 2014, pp. 951–960.

[21] H. Estelami, "The computational effect of price endings in multi-dimensional price advertising," *Journal of Product & Brand Management*, vol. 8, no. 3, pp. 244–256, 1999.

[22] W. Chen, L. V. Lakshmanan, and C. Castillo, *Information and Influence Propagation in Social Networks*. Morgan and Claypool, 2013.

[23] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *SIGKDD*. ACM, 2008, pp. 7–15.

[24] P. Domingos and M. Richardson, "Mining the network value of customers," in *SIGKDD*. ACM, 2001, pp. 57–66.

[25] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing letters*, vol. 12, no. 3, pp. 211–223, 2001.

[26] M. Granovetter, "Threshold models of collective behavior," *American journal of sociology*, pp. 1420–1443, 1978.

[27] C. Aggarwal, A. Khan, and X. Yan, "On flow authority discovery in social networks," in *SDM*, 2011, pp. 522–533.

[28] B. Xiang, Q. Liu, E. Chen, H. Xiong, Y. Zheng, and Y. Yang, "Pagerank with priors: An influence propagation perspective," in *IJCAI*, 2013.

[29] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *SIGKDD*. ACM, 2007, pp. 420–429.

[30] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *ICDM*, 2011, pp. 211–220.

[31] K. Jung, W. Heo, and W. Chen, "Irie: Scalable and robust influence maximization in social networks," in *ICDM*. IEEE, 2012, pp. 918–923.

[32] 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 *ICDM*, 2013.

[33] P. Zhang, W. Chen, X. Sun, Y. Wang, and J. Zhang, "Minimizing seed set selection with probabilistic coverage guarantee in a social network," in *SIGKDD*. ACM, 2014.

[34] V. V. Vazirani, *Approximation algorithms*. springer, 2001.

[35] M. Hammar, R. Karlsson, and B. J. Nilsson, "Using maximum coverage to optimize recommendation systems in e-commerce," in *RecSys*. ACM, 2013, pp. 265–272.

[36] L. Xu, B. Li, and E. Chen, "Ensemble pruning via constrained eigen-optimization," in *ICDM*. IEEE, 2012, pp. 715–724.

[37] M. Wang, X. Zhou, Q. Tao, W. Wu, and C. Zhao, "Diversifying tag selection result for tag clouds by enhancing both coverage and dissimilarity," in *WISE*. Springer, 2013, pp. 29–42.

[38] H. Lin and J. Bilmes, "A class of submodular functions for document summarization." in *ACL*, 2011, pp. 510–520.

[39] A. Kulik, H. Shachnai, and T. Tamir, "Maximizing submodular set functions subject to multiple linear constraints," in *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2009, pp. 545–554.

[40] L. Xu, W. Li, and D. Schuurmans, "Fast normalized cut with linear constraints," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2866–2873.

[41] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes, "Automatic selection of social media responses to news," in *SIGKDD*. ACM, 2013, pp. 50–58.

[42] C.-W. Ko, J. Lee, and M. Queyranne, "An exact algorithm for maximum entropy sampling," *Operations Research*, vol. 43, no. 4, pp. 684–691, 1995.

[43] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *WWW*. ACM, 2007, pp. 271–280.

[44] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[45] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[46] P. Massa and P. Avesani, "Trust-aware bootstrapping of recommender systems," in *ECAI Workshop on RecSys*. Citeseer, 2006, pp. 29–33.

[47] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." in *Proceedings of WWW Conference*. Stanford InfoLab, 1999.

[48] H. Zhu, E. Chen, H. Xiong, H. Cao, and J. Tian, "Mobile app classification with enriched contextual information," *IEEE Transactions on Mobile Computing*, vol. 13, no. 7, pp. 1550–1563, 2014.