Viral Marketing for Product Cross-sell through Social Networks

Ramasuri Narayanam and Amit A. Nanavati

IBM Research, India.
Email ID: {nramasuri,namit}@in.ibm.com

Abstract. The well known *influence maximization problem* [1] (or viral market-ing through social networks) deals with selecting a few influential initial seeds to maximize the awareness of product(s) over the social network. In this paper, we introduce a novel and generalized version of the influence maximization prob-lem that considers simultaneously the following three practical aspects: (i) Often cross-sell among products is possible, (ii) Product specific costs (and benefits) for promoting the products have to be considered, and (iii) Since a company of-ten has budget constraints, the initial seeds have to be chosen within a given budget. We refer to this generalized problem setting as *Budgeted Influence Maxi-mization with Cross-sell of Products (B-IMCP)*. To the best of our knowledge, we are not aware of any work in the literature that addresses the B-IMCP problem which is the subject matter of this paper. Given a fixed budget, one of the key issues associated with the B-IMCP problem is to choose the initial seeds within this budget not only for the individual products, but also for promoting cross-sell phenomenon among these products. In particular, the following are the specific contributions of this paper: (i) We propose an influence propagation model to capture both the cross-sell phenomenon and product specific costs and benefits; (ii) As the B-IMCP problem is NP-hard computationally, we present a simple greedy approximation algorithm and then derive the approximation guarantee of this greedy algorithm by drawing upon the results from the theory of matroids; (iii) We then outline two efficient heuristics based on well known concepts in the literature. Finally, we experimentally evaluate the proposed approach for the B-IMCP problem using a few well known social network data sets such as WikiVote data set, Epinions, and Telecom call detail records data.

Keywords Social networks, influence maximization, cross-sell, budget constraint, matroid theory, costs, benefits, and submodularity.

1   Introduction

The phenomenon of viral marketing is to exploit the social connections among the individuals to promote awareness for new products [2–4]. One of the key issues in viral marketing through social networks is to select a set of influential seed members (also called as *initial seeds*) in the social network and give them free samples of the product (or provide promotional offers) to trigger cascade of influence over the network [5]. The problem is, given an integer k, how should we choose a set of k initial seeds so

that the cascade of influence over the network is maximized? This problem is known as *influence maximization problem* [1]. This problem is shown to be a NP-hard problem in the context of certain information diffusion models such as linear threshold model and independent cascade model [1]. Influence maximization problem is well studied in the literature [5, 6, 1, 7–15] in the context of a single product and multiple independent products; and we refer to the section on the relevant work for more details.

However, the existing work in the literature on viral marketing of products through social networks ignores the following important aspects that we often experience in several practical settings:

– *Cross-sell Phenomenon:* Certain products are complementary to each other in the sense that there is a possibility of cross-sell among these products,
– *Product Specific Costs and Benefits:* There is a cost associated with each product in order to provide promotional offers (or discounts) to each of the initial seeds. Similarly, there is a benefit associated with each product, when someone buys that product, and
– *Budget Constraint:* Companies have often have budget constraints and hence the initial seeds have to be chosen within a given budget.

To the best of our knowledge, we are not aware of any work in the literature that simultaneously deals with the cross-sell phenomenon, the product specific costs and benefits, and the budget constraint while addressing the influence maximization problem. We address this generalized version of the influence maximization problem in this paper.

In particular, we consider the following framework. Let P_1 be a set of t_1 independent products and similarly P_2 be another set of t_2 independent products. We assume that cross-sell is possible from the products in P_1 to the products in P_2 and, in particular, we consider the following specific form for the cross-sell phenomenon. The need for buying any product in P_2 arises only after buying some product in P_1. A real life instance of this type of cross-sell is as follows. Consider the context of computers and printers; in general, the need for buying any printer arises after buying a computer. Informally, cross-sell is the action or practice of selling an additional product (or service) to an existing customer. Typically, the additional product (i.e. in P_2) is of interest to the customer only because of a prior product (i.e. in P_1) purchased by the customer. From a social network diffusion model standpoint, the purchase of products in P_1 causes a lowering of threshold for buying certain products in P_2. It is this phenomenon that we explore in this paper.

We consider different costs and benefits for each product in P_1 and P_2. Since companies owning the products often have budget constraints, we can offer free samples (or promotional offers) of the products to the initial seeds within a given budget B. In this setting, one of the key issues is to choose the initial seeds not only for each individual product in P_1, but also for promoting the cross-sell phenomenon from the products in P_1 to the products in P_2. We refer to the above problem of selecting, within budget B, a set of initial seeds to maximize influence through the social network and hence to maximize the revenue to the company as the *budgeted influence maximization with cross-sell of products (B-IMCP)* problem. In what follows, we highlight the main challenges associated with the B-IMCP problem that make it non-trivial to address and also summarize the main contributions of this paper:

(A) Modeling Aspects: How to model the propagation of influence in social networks in the context of cross-sell of products? We note that the variants of the linear threshold model [16, 17, 1, 14] in their current form are not sufficient to model the cross-sell phenomenon.

In this paper, we address this issue by proposing a simple model of influence propagation over social networks in the context of cross-sell of products by naturally extending the well known linear threshold model [1]. We call this *linear threshold model for cross-sell of products (LT-CP)*. We then prove a few important properties of the LT-CP model such as monotonicity and submodularity.

(B) Algorithmic Aspects: We note that the B-IMCP problem, in the context of the LT-CP model, turns out to be a computationally hard problem, i.e. NP-hard (see Section 3 for more details). This calls for designing an approximation algorithm to address the B-IMCP problem. In this paper, we propose a greedy approximation algorithm to address the B-IMCP problem. Assume that B is the budget for the company. Let c_M^1 and c_M^2 be the maximum cost of providing a free sample of any product in P_1 and P_2 respectively. On similar lines, let c_m^1 and c_m^2 be the minimum cost of providing a free sample of any product in P_1 and P_2 respectively. We show that the approximation guarantee of the proposed greedy algorithm for the B-IMCP problem is $\frac{Bc_m}{B(c_M+c_m)+c_Mc_m}$, where $c_M = max(c_M^1, c_M^2)$ and $c_m = min(c_m^1, c_m^2)$. Interestingly, the approximation factor of the greedy algorithm is independent of the number of products and it only depends on (a) the budget B, and (b) the maximum and the minimum costs to provide free samples of products in P_1 and P_2 respectively.

We use the techniques from the theory of submodular function maximization over Matroids [18] to derive the approximation guarantee of the proposed greedy algorithm. We must note that the body of relevant work in the literature works with the framework of *approximations for maximizing submodular set functions [19]* to derive the approximation guarantee of the algorithms for the variants of the influence maximization problem with a single product and multiple independent products [1, 7, 14]. However, these techniques are not sufficient for the B-IMCP problem setting as (i) we have to work with cross-sell of products, (ii) product specific costs and benefits, and (ii) we have to deal with the budget constraint.

(C) Experimental Aspects: We experimentally observe that the proposed greedy approximation algorithm for the B-IMCP problem runs slow even on moderate size network data sets. We must note that similar observations are reported in the literature in the context of the greedy algorithm [1] for the well known influence maximization problem [7, 8]. The existing scalable and efficient heuristics [7, 8] for the influence maximization problem do not directly apply to the context of the B-IMCP problem. How to alleviate this scalability issue of determining a solution for the B-IMCP problem on large social network data sets? In this paper, we present an efficient heuristic for the B-IMCP problem. We experimentally evaluate the performance of the greedy approximation and heuristic algorithms using experimentation on several social network data sets such as WikiVote trust network, Epinions trust network, and Telecom call detail records data. We also compare and contrast the performance of the greedy approximation and the heuristic algorithms with that of two well known benchmark heuristics.

1.1 Novelty of This Paper

The primary contribution and the novelty of this paper is three fold: (i) Introducing the phenomenon of cross-sell of products and product specific costs and benefits while addressing the influence maximization problem, (ii) Naturally extending the well known linear threshold model to capture the cross-sell phenomenon, and (iii) Performing non-trivial analysis of the simple greedy algorithm for the B-IMCP problem to derive the approximation guarantee of the same.

2 Relevant Work

There are two well known operational models in the literature that capture the underlying dynamics of the information propagation in the viral marketing process. They are the linear threshold model [17, 16, 1] and the independent cascade model [20, 1]. In the interest of space constraints, we only present the most relevant work on the influence maximization problem in the literature and we categorize this into three groups as follows.

Influence Maximization with Single Product: Domingos and Richardson [5] and Richardson and Domingos [6] were the first to study influence maximization problem as an algorithmic problem. Computational aspects of the influence maximization problem are investigated by Kempe, Kleinberg, and Tardos [1] and they showed that the optimization problem of selecting the most influential nodes is NP-hard. The authors proposed a greedy approximation algorithm for the influence maximization problem.

Leskovec, et. al. [7] proposed an efficient algorithm for the influence maximization problem based on the submodularity of the underlying objective function that scales to large problems and is reportedly 700 times faster than the greedy algorithm of [1] and later Chen, Wang, and Yang [8] further improved this result. Even-Dar and Shapira [11], Kimura and Saito [9], Mathioudakis *et.al.* [10], Ramasuri and Narahari [21] considered various interesting extensions to the basic version of the influence maximization problem in social networks.

Influence Maximization with Multiple Products: Recently, Datta, Majumder, and Shrivastava [12] considered the influence maximization problem for *multiple independent products*.

Viral Marketing with Competing Companies: Another important branch of the research work in viral marketing is to study the algorithmic problem of how to introduce a new product into the market in the presence of a single or multiple competing products already in the market [13–15].

3 The Proposed Model for the B-IMCP Problem

Here we first present the LT-CP model for the B-IMCP problem. We must note that this model is a natural extension of the well known linear threshold model [17, 16, 1].

3.1 The LT-CP Model

Let $G = (V, E)$ be a directed graph representing an underlying social network where V represents a set of individuals and E is a set of directed edges representing friendship relations among these individuals. Let $|V| = n$ and $|E| = m$. For each edge $(i, j) \in E$, we define a weight w_{ij} and this indicates the probability with which the available information at individual i passes to individual j. Another interpretation of this weight w_{ij} is the probability with which individual j is influenced by the recommendation of the individual i. If there is no confusion, here onwards, we refer to individuals and nodes interchangeably. Similarly, we also refer to graphs and networks interchangeably.

In this context, we consider the following setting as introduced in Section 1. Note that $P_1 = \{1, 2, \ldots, t_1\}$ is a set of t_1 independent products and similarly $P_2 = \{1, 2, \ldots, t_2\}$ is another set of t_2 independent products. Cross-sell is possible from the products in P_1 to the products in P_2. For the simplicity of the technical analysis, we assume that (i) $t_1 = t_2$ and call this common value t; and (ii) there exists an onto function from P_1 to P_2, call it $H : P_1 \rightarrow P_2$, such that for each product $k \in P_1$ there exists exactly one product in P_2 (namely $H(k)$). That is, for each product $k \in P_1$, there exists a product $H(k) \in P_2$ such that cross-sell is possible from product $k \in P_1$ to product $H(k) \in P_2$. A company, call it M, owns the products in P_1 and P_2 and it has a fixed budget B for conducting viral marketing campaign for these products.

In particular, a free sample (or promotional offer) of product k in P_1 incurs a cost of c_k^1 and similarly a free sample (or promotional offer) of product z in P_2 incurs a cost of c_z^2. Also, when an item of product k in P_1 is sold, it leads to a benefit b_k^1 to the company. On the similar lines, when an item of product z in P_2 is sold, it leads to a benefit b_z^2 to the company. Now, we define a few important notations and terminology as follows:

- We call an individual (and is represented by a *node* in the graph) in the network *active* if he/she buys any product in P_1 or P_2, and *inactive* otherwise. For each node $i \in V$, let N_i be the set of its neighbors. Node i is influenced by any neighbor j according to a weight w_{ji}. Assume these weights are normalized in such a way that $\sum_{j \in N_i} w_{ji} \leq 1$.
- We define the following for each node $i \in V$. For each product $k \in P_1$ and for each $i \in V$, we define A_i^k to be the set of node i's active neighbors who bought product $k \in P_1$. On similar lines, for each product $z \in P_2$ and for each $i \in V$, we define A_i^z to be the set of node i's active neighbors who have initially bought product $H(z) \in P_1$ and then bought product $z \in P_2$.

We now define when the individual nodes buy the products in P_1 and P_2. Recall that the products in P_1 are independent. The decision of a node $i \in V$ to buy product $k \in P_1$ is based on a threshold function (f_i^k), which is dependent on N_i and a threshold (θ_i^k) chosen uniformly at random by node i from the interval $[0, 1]$. This threshold represents the minimum amount of influence required from the active neighbors of node i (who bought product $k \in P_1$) in order for node i to become active. Note that $f_i^k : 2^{N_i} \rightarrow [0, 1]$ is defined as $f_i^k(S) = \sum_{j \in S} w_{ji}, \forall S \subseteq N_i$. Now, we say that node i buys product $k \in P_1$ if $f_i^k(A_i^k) \geq \theta_i^k$.

Recall that cross-sell is possible from the products in P_1 to the products in P_2 and this cross-sell relationship is defined using the function H which is onto. For each

$z \in P_2$ and for each $i \in V$, we initially set the threshold θ_i^z of node i for buying the product z to be a very high quantity to model the fact that no node $i \in V$ buys product z in P_2 before buying product $H(z)$ in P_1. Now assume that node $i \in V$ has bought the product $H(z) \in P_1$ and since cross-sell is possible from product $H(z)$ to product $z \in P_2$, we decrease the threshold θ_i^z by defining that θ_i^z is chosen uniformly at random from the interval $[0, a]$ where $0 \le a < 1$. Now the decision of node i to buy product $z \in P_2$ is based on a threshold function (f_i^z), which is dependent on N_i and the threshold θ_i^z. This threshold represents the minimum amount of influence required from the active neighbors of node i (who have initially bought product $H(z) \in P_1$ and then bought product $z \in P_2$) in order for node i to become active. Note that $f_i^z : 2^{N_i} \to [0, 1]$ is defined as $f_i^z(S) = \sum_{j \in S} w_{ji}, \forall S \subseteq N_i$. Now, we say that node i buys product $z \in P_2$ if $f_i^z(A_i^z) \ge \theta_i^z$.

3.2 The B-IMCP Problem

In the presence of the above model, we now define the following. For each product $k \in P_1$, let S_1^k be the initial seed set. We define the influence spread of the seed set S_1^k, call it $\Gamma(S_1^k)$, to be the expected number nodes that buy the product $k \in P_1$ at the end of the diffusion process, given that S_1^k is the initial seed set. On similar lines, for each product $z \in P_2$, let S_2^z be the initial seed set. We define the influence spread of the seed set S_2^z, call it $\Delta(S_2^z)$, to be the expected number of nodes that initially buy the product $H(z) \in P_1$ and then buy the product $z \in P_2$ at the end of the diffusion process, given that S_2^z is the initial seed set. For any specific choice of the initial seed sets S_1^k for each $k \in P_1$ and S_2^z for each $z \in P_2$, we now define the objective function, call it $\sigma(S_1^1, \ldots, S_1^t, S_2^1, \ldots, S_2^t)$, to be the expected revenue to the company at the end of the diffusion process. That is,

$$\sigma(S_1^1, \ldots, S_1^t, S_2^1, \ldots, S_2^t) = \sum_{k \in P_1} \Gamma(S_1^k) b_k^1 + \sum_{z \in P_2} \Delta(S_2^z) b_z^2. \qquad (1)$$

Given the budget B, the B-IMCP problem seeks to find the initial seed sets $S_1^1, \ldots, S_1^t$, $S_2^1, \ldots, S_2^t$ such that the objective function is maximized.

We now show that the B-IMCP problem is a computationally hard problem.

Lemma 1. *The B-IMCP problem in the presence of the LT-CP model is NP-hard.*

Proof. By setting $|P_1| = t = 1$, $P_2 = \phi$, $c_k^1 = 1$ for each $k \in P_1$, and $b_k^1 = 1$ for each $k \in P_1$, we get that any arbitrary instance of the B-IMCP problem with the LT-CP model reduces exactly to an instance of the influence maximization problem with the linear threshold model [1]. It is already known that the influence maximization problem with the linear threshold model is NP-hard [1]. $\blacksquare$

3.3 Properties of the Proposed Model

We show that the objective function $\sigma(.)$ is monotone and submodular. The proof of monotonicity of $\sigma(.)$ is immediate, as the expected number of individuals that buy the product(s) does not decrease when we increase the number of initial seeds. Thus we

now focus on the proof of submodularity of $\sigma(.)$ and we note that this can be proved easily using the results from Kempe *et al.* [1]. However, for the sake of completeness, we give a sketch of the proof of this result.

Lemma 2. *For any arbitrary instance of the LT-CP model, the objective function $\sigma(.)$ is submodular.*

Proof Sketch: There are two main steps in this proof. First, for each $k \in P_1$, we have to show that $\Gamma(S_1^k)$ is submodular. Second, for each $z \in P_2$, we have to show that $\Delta(S_2^z)$ is submodular. Recall that (i) multiplying a submodular function with a positive constant results in a submodular function; and (ii) the sum of submodular functions is also a submodular function. This implies that the objective function $\sigma(.)$ is submodular.

Since the products in P_1 are independent, for each $k \in P_1$, it is straight forward to see that $\Gamma(S_1^k)$ is submodular due to Theorem 2.5 in Kempe *et al.* [1].

Claim 2: $\Delta(S_2^z)$ is submodular.

Since cross-sell is possible from the product $H(z) \in P_1$ to the product $z \in P_2$, we can compute $\Delta(S_2^z)$ after the diffusion of $H(z)$ finishes. We model the spread of $H(z)$ using the technique of live-edge paths as in Theorem 2.5 in [1]. Suppose that every node i picks at most one of its incoming edges at random, selecting the edge from neighbor j with probability w_{ji} and selecting no edge with probability $1 - \sum_j w_{ji}$. Each such selected edge in G is declared to be *live* and all other edges are declared to be *blocked*. Let G' be the graph obtained from the original graph G by retaining only the live-edges and let $\Pi(G)$ be the set of all such G'. Let $P(G'$ be the probability of obtaining G' from G using the process of live edges. Note that each G' models a possible trace of the spread of the product $H(z) \in P_1$.

Now consider $G' \in \Pi(G)$ and a node i in G'. Recall that each node i in the original graph G picks at most one of its incoming edges at random, selecting the edge from neighbor j with probability w_{ji} and selecting no edge with probability $1 - \sum_j w_{ji}$. For this reason, each node i in G' has at most one incoming edge. Now if node i in G' has an incoming edge, call it from node j, then i picks this only incoming edge with probability w_{ji} and picks no edge with probability $1 - w_{ji}$. Each such selected edge in G' is declared to be *live* and all the other edges are declared to be *blocked*. Using the arguments exactly similar to that in Kempe *et al.* [1], it turns out that proving this claim is equivalent to reachability via live-edge paths in G'. Let G'' be the graph obtained from G' by retaining only the live-edges and assume that $\Pi(G')$ is the set of all such G''. Let $P(G'')$ be the probability of obtaining G'' from G' using the process of live edges. For each $G'' \in \Pi(G')$, we define $a_{G''}(S_2^z, m)$ to be the number of nodes that are exactly m steps away on any path starting from some node in S_2^z in G''; i.e. $a_{G''}(S_2^z, m) = |\{v \in V \mid d_{G''}(S_2^z, v) = m\}|$ where $d_{G''}(S_2^z, v)$ is the length of the shortest distance from any node in S_2^z to v. We can now write $\Delta(S_2^z)$ as follows:

$$\Delta(S_2^z) = E_{G' \in \Pi(G)}\Big[E_{G'' \in \Pi(G')}\Big[\sum_{m=0}^{\infty} a_{G''}(S_2^z, m)\Big]\Big] \tag{2}$$

$$= \sum_{G' \in \Pi(G)} \sum_{G'' \in \Pi(G')} P(G')P(G'') \sum_{m=0}^{\infty} a_{G''}(S_2^z, m) \tag{3}$$

Now let us define $h(S_2^z, G'') = \sum_{m=0}^{\infty} a_{G''}(S_2^z, m)$ for all $S_2^z \subseteq V$. If we can show that $h(.)$ is submodular, then $\Delta(S_2^z)$ is also submodular as it is a non-negative linear combination of $h(.,.)$. It is not difficult to show that $h(.)$ is submodular and, in the interest of space, we do not present the proof of the same. ∎

Note: It is important to note that the submodularity of $\sigma(.)$ may break down if we change the way to model the cross-sell relationships among the products.

4 Greedy Approximation Algorithm for The B-IMCP Problem

Motivated by the greedy algorithm [1] for the well known influence maximization problem, we now present a simple greedy algorithm for the B-IMCP problem. Let

---

Algorithm 1 Greedy Algorithm to Select the Initial Seeds for the B-IMCP Problem

1: Initially set $S_1^k = \phi \quad \forall k \in P_1$ and $S_2^z = \phi \quad \forall z \in P_2$.
2: **while** $B > 0$ **do**
3: Pick a node $v_1 \in V \setminus \cup_{k \in P_1} S_1^k$ such that v_1 maximizes $valx = $
$\dfrac{\sigma\left(\cup_{k \in P_1} S_1^k \bigcup \{v_1\} \bigcup \cup_{z \in P_2} S_2^z\right) - \sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z\right)}{c_i^1}$, when we active it with
some product $i \in P_1$
4: Pick a node $v_2 \in V \setminus \cup_{z \in P_2} S_2^z$ such that v_2 maximizes $valy = $
$\dfrac{\sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z \bigcup \{v_2\}\right) - \sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z\right)}{c_j^2}$, when we active it with
some product $j \in P_2$.
5: **if** $valx \geq valy$ and $B - c_i^1 \geq 0$ **then**
6: Set $S_1^i \leftarrow S_1^i \cup \{v_1\}$, and $B \leftarrow B - c_i^1$
7: As cross-sell is possible from from product $i \in P_1$ to product $H(i) \in P_2$, update the
value of $\theta_{v_1}^{H(i)}$ for node v_1
8: **end if**
9: **if** $valy > valx$ and $B - c_j^2 \geq 0$ **then**
10: Set $S_2^j \leftarrow S_2^j \cup \{v_2\}$, and $B \leftarrow B - c_j^2$
11: **end if**
12: **end while**
13: Return $\left(S_1^1, S_1^2, \ldots, S_1^t, S_2^1, S_2^2, \ldots, S_2^t\right)$ as the initial seed set

---

$S_1^1, S_1^2, \ldots, S_1^t$ be the sets of initial seeds for the t products in P_1 respectively. Let $S_2^1, S_2^2, \ldots, S_2^t$ be the sets of initial seeds for the t products in P_2 respectively. Initially set $S_1^k = \phi$ for each $k \in P_1$ and $S_2^z = \phi$ for each $z \in P_2$. Algorithm 1 presents the proposed greedy algorithm and the following is the main idea of the same. The algorithm runs in iterations until the budget B is exhausted to select the initial seeds. The following is performed in each iteration of the algorithm:
(i) Let $v_1 \in V \setminus \cup_{k \in P_1} S_1^k$ be the next best seed for P_1 in the sense that when we activate it with product $i \in P_1$, v_1 maximizes the ratio of increase in the expected revenue gain
to the cost c_i^1; that is, v_1 maximizes $\dfrac{\sigma\left(\cup_{k \in P_1} S_1^k \bigcup \{v_1\} \bigcup \cup_{z \in P_2} S_2^z\right) - \sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z\right)}{c_i^1}$

and call this $valx$;

(ii) Let $v_2 \in V \setminus \cup_{z \in P_2} S_2^z$ be the next best seed in the sense that when we activate it with some product $j \in P_2$, v_2 maximizes the ratio of increase in the expected revenue gain to the cost c_j^1; that is, v_2 maximizes $\dfrac{\sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z \bigcup \{v_2\}\right) - \sigma\left(\cup_{k \in P_1} S_1^k \bigcup \cup_{z \in P_2} S_2^z\right)}{c_j^2}$ and call this $valy$;

(iii) If $valx \geq valy$ and $B - c_i^1 \geq 0$, then we add v_1 to the set of seeds for S_1^i and also decrease B by an amount c_i^1. To take care of the cross-sell phenomenon, we also update the threshold $\theta_{v_1}^{H(i)}$ for node v_1;

(iv) If $valy > valx$ and $B - c_j^2 \geq 0$, then we add v_2 to the set S_2^j and decrease B by an amount c_j^2.

Finally, the greedy algorithm returns $\left(S_1^1, S_1^2, \ldots, S_1^t, S_2^1, S_2^2, \ldots, S_2^t\right)$ as the initial seed set for the B-IMCP problem.

Running Time of Algorithm 1 Let $c = min\{c_1^1, c_2^1, \ldots, c_t^1, c_1^2, c_2^2, \ldots, c_t^2\}$ and $t = \frac{B}{c}$. Note that Algorithm 1 runs at most t rounds. In each iteration of this algorithm, we have to check at most $O(n)$ nodes to determine the best seed for P_1 and P_2. To determine $valx$ (or $valy$) in each iteration, we have to essentially count the number of nodes that are reachable from the initial seed elements using the live edges in the graph and it takes at most $O(m)$ time where m is the number of edges. Also, as underlying information diffusion process is a stochastic process, we have to repeat the experiment a number of times (call it R times) and take the average to determine values for $valx$ and $valy$ in each iteration. With all this in place, the running time of Algorithm 1 is $O(tnRm)$ where $t = \frac{B}{c}$.

4.1 Analysis of Algorithm 1

We now analyze Algorithm 1 and derive the approximation guarantee of the same. Our analysis uses results from matroid theory and Calinescu, *et al.* [18]. Towards this end, we first recall the definition of a matroid.

Matroid: A Matroid is a pair $M = (U, I)$, where $I \subseteq 2^U$ is a subset of the power set (all possible subsets) of U that satisfies the following constraints:

– I is an *independent set system*: $\phi \in I$ and $A \in I$, any set $B \subset A$ then $B \in I$ (All subsets of any independent set is also independent).
– $\forall A, B \in I$ and $|A| > |B|$: $\exists x \in A - B$ s.t. $B \cup \{x\} \in I$.

The first constraint defines an independent set system, and each set $S \in I$ is called an independent set. The problem of maximizing a sub-modular function on a Matroid is to find the independent set $S \in I$, s.t. $f(S)$ is maximum over all such sets S. If the input set is a Matroid, it is known that the sub-modular function maximization can be approximated within a constant factor ($\frac{1}{2}$ or $(1 - 1/e)$ in certain cases) using the greedy hill climbing approach (Nemhauser, Wolsey, and Fisher [19]).

The independent set system essentially defines the feasible sets over which the objective function is defined. In the context of B-IMCP problem, we call an initial seed

set $\left(S_1^1, S_1^2, \ldots, S_1^t, S_2^1, S_2^2, \ldots, S_2^t\right)$ *feasible*, when the sum of costs of providing free samples of the products in this initial seed set is within the budget B. It is easy to see that the feasible seed sets form an Independent set system, I. However, the feasible seed sets in I do not form a matroid since they can violate the second condition in the definition of matroid due to the budget constraint. The following example validates this fact.

Example 1. Consider a graph with 10 *individuals, i.e.* $V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. *There are* 4 *products in* P_1*; i.e.,* $P_1 = \{t_1^1, t_2^1, t_3^1, t_4^1\}$. *There are* 4 *products in* P_2*; i.e.,* $P_2 = \{t_1^2, t_2^2, t_3^2, t_4^2\}$. *Also let* $B = 20$, $c_1^1 = 4$, $c_2^1 = 3$, $c_3^1 = 5$, $c_4^1 = 6$, $c_1^2 = 2$, $c_2^2 = 4$, $c_3^2 = 2$, *and* $c_4^2 = 4$. *Now consider two feasible initial seed sets as follows:*

Consider a feasible seed set $\left(S_1^1, S_1^2, \ldots, S_1^4, S_2^1, S_2^2, \ldots, S_2^4\right)$ *as follows.* $S_1^1 = \phi$, $S_1^2 = \{2, 3, 4\}$, $S_1^3 = \{6\}$, $S_1^4 = \phi$, $S_2^1 = \phi$, $S_2^2 = \{2\}$, $S_2^3 = \{6\}$, $S_2^4 = \phi$. *Note that the cost of providing the free samples of products in the initial seed set* $\left(S_1^1, S_1^2, \ldots, S_1^4, S_2^1, S_2^2, \ldots, S_2^4\right)$ *is* $3 + 3 + 3 + 5 + 4 + 2 = 20$. *Given that* $B = 20$, *it is clear that* $\left(S_1^1, S_1^2, \ldots, S_1^4, S_2^1, S_2^2, \ldots, S_2^4\right)$ *is a feasible seed set.*

Consider another feasible seed set $\left(\bar{S}_1^1, \bar{S}_1^2, \ldots, \bar{S}_1^4, \bar{S}_2^1, \bar{S}_2^2, \ldots, \bar{S}_2^4\right)$ *as follows.* $\bar{S}_1^1 = \phi$, $\bar{S}_1^2 = \phi$, $\bar{S}_1^3 = \{3, 9\}$, $\bar{S}_1^4 = \{6\}$, $\bar{S}_2^1 = \phi$, $\bar{S}_2^2 = \phi$, $\bar{S}_2^3 = \phi$, $\bar{S}_2^4 = \{6\}$. *Note that the cost of providing the free samples of products in the initial seed set* $\left(\bar{S}_1^1, \bar{S}_1^2, \ldots, \bar{S}_1^4, \bar{S}_2^1, \bar{S}_2^2, \ldots, \bar{S}_2^4\right)$ *is* $5 + 5 + 6 + 4 = 20$. *Given that* $B = 20$, *it is clear that* $\left(\bar{S}_1^1, \bar{S}_1^2, \ldots, \bar{S}_1^4, \bar{S}_2^1, \bar{S}_2^2, \ldots, \bar{S}_2^4\right)$ *is a feasible seed set.*

Note that the cardinality of the first feasible set is $|\{2, 3, 4, 6\}| = 4$ *and that of the second feasible set is* $|\{3, 6, 9\}| = 3$. *Moreover, observe that node* 2 *is an initial seed for product* 2 *in* P_1 *in the first feasible seed set and it is not an initial seed for any product in the second feasible seed set. However, we cannot add node* 2 *to the seed set of any product in* P_1 *and* P_2 *in the second feasible seed set without violating the budget constraint.*

This example immediately leads to the following simple result.

Proposition 1. *For an arbitrary instance of the B-IMCP problem, the independent set system consisting of the feasible seed sets is not necessarily a Matroid.*

Hence we opt for a slightly weaker definition on an independent set system, called a p-system, defined as follows [12]. Informally, p-system says that for any set $A \subseteq V$, the sizes of any two maximal independent subsets of A do not differ by a factor more than p. Then according to Calinescu, *et al.* [18] the hill climbing gives a $\frac{1}{(p+1)}$ approximation of submodular function maximization for any p-system.

Towards this end, we first prove an useful result. Before proceeding further, we introduce the following notation. Assume that (i) c_M^1 is the maximum cost among all c_k^1, $k \in P_1$; i.e., $c_M^1 = max_{k \in P_1} \ c_k^1$, (ii) c_M^2 is the maximum cost among all c_z^2, $z \in P_2$; i.e., $c_M^2 = max_{z \in P_2} \ c_z^2$, (iii) c_m^1 is the minimum cost among all c_k^1, $k \in P_1$; i.e., $c_m^1 = min_{k \in P_1} \ c_k^1$, and (iv) c_m^2 is the minimum cost among all c_z^2, $z \in P_2$; i.e., $c_m^2 = min_{z \in P_2} \ c_z^2$.

Lemma 3. *The feasible seed sets for the B-IMCP problem form $c_M\left(\frac{1}{c_m} + \frac{1}{B}\right)$-system where $c_M = max(c_M^1, c_M^2)$ and $c_m = min(c_m^1, c_m^2)$.*

Proof. Consider an arbitrary instance of the B-IMCP problem. Note that V is the set of nodes in the graph G and let A be any subset of V. Let $\left(S_1^1, S_1^2, \ldots, S_1^4, S_2^1, S_2^2, \ldots, S_2^4\right)$ and $\left(\bar{S}_1^1, \bar{S}_1^2, \ldots, \bar{S}_1^4, \bar{S}_2^1, \bar{S}_2^2, \ldots, \bar{S}_2^4\right)$ be any two maximal feasible sets in A with maximum and minimum sizes respectively. Also let $S = S_1^1 \bigcup \ldots \bigcup S_1^t \bigcup S_2^1 \bigcup \ldots \bigcup S_2^t$ and $\bar{S} = \bar{S}_1^1 \bigcup \ldots \bigcup \bar{S}_1^t \bigcup \bar{S}_2^1 \bigcup \ldots \bigcup \bar{S}_2^t$. If $|S| \leq |\bar{S}|$, then S and $\bar{S}$ are of same size and hence the independent set system with all feasible seed sets forms a matroid. It is contradiction to Proposition 1 (see the Example 1). Hence we consider the case where $|\bar{S}| < |S|$. We will now bound how much the cardinality of S is greater than that of $\bar{S}$ in the worst case. We consider the following four cases.

Case 1 *($c_M^1 > c_M^2$ and $c_m^1 > c_m^2$):* The cardinality of S is much larger, in the worst case, than that of $\bar{S}$ when:

- All the seed elements in $\bar{S}$ are part of $\bar{S}_1^j$ for some product $j \in P_1$ such that $c_j^1 = c_M^1$.
- All initial seed elements of S are the initial seeds for some product $k \in P_2$ having cost $c_k^2 = c_m^2$.

Let $|S| = \alpha$ and $|\bar{S}| = \beta$. Then the above construction of S and $\bar{S}$ leads to the following inequality

$$\alpha c_m^2 \leq \beta c_M^1 + c_m^2. \tag{4}$$

Note that the term c_m^2 appears at right hand of the above inequality because there might be some budget left out after constructing the minimum cardinality feasible set $\bar{S}$ and it is at most c_m^2. Now equation (4) implies that

$$\Rightarrow \frac{\alpha}{\beta} \leq \frac{c_M^1}{c_m^2} + \frac{1}{\beta}. \tag{5}$$

Note that $\beta \geq \frac{B}{c_M^1}$. This fact and expression (5) imply that

$$\Rightarrow \frac{\alpha}{\beta} \leq \frac{c_M^1}{c_m^2} + \frac{c_M^1}{B} \quad \Rightarrow \frac{|S|}{|\bar{S}|} \leq c_M^1\left(\frac{1}{c_m^2} + \frac{1}{B}\right). \tag{6}$$

On similar lines as above, we can also deal with the remaining three cases: (a) $c_M^1 > c_M^2$ and $c_m^1 \leq c_m^2$; (b) $c_M^1 \leq c_M^2$ and $c_m^1 > c_m^2$; and (c) $c_M^1 \leq c_M^2$ and $c_m^1 \leq c_m^2$. This completes the proof of the lemma. ∎

Theorem 1. *The proposed greedy algorithm determines the initial seeds for the B-IMCP problem that is at least $\frac{Bc_m}{B(c_M+c_m)+c_Mc_m}$ times the optimum solution, where $c_M = max(c_M^1, c_M^2)$ and $c_m = min(c_m^1, c_m^2)$.*

Proof. It is known that the greedy hill climbing algorithm gives a $\frac{1}{(p+1)}$ approximation of submodular function maximization for any p-system (Calinescu, Chekuri, Pal, and Vondrak (2007)). The proof of this theorem follows immediately as a consequence of this fact and Lemma 3. ∎

In the view of the above results, we now present a few important observations as follows: (i) The approximation guarantee of the proposed greedy algorithm is independent of the number of products in P_1 and P_2. Also the approximation guarantee only depends on the budget and the minimum and maximum costs for providing free samples of the products in P_1 and P_2. (ii) If the cost of each product in P_1 and P_2 is 1, then $c_M = c_m = 1$. This implies that the greedy approximation algorithm returns a solution to the B-IMCP problem that is at least $\frac{B}{2B+1}$ times that of the optimum solution.

5   Heuristics for the B-IMCP Problem

We observe that the proposed greedy approximation algorithm runs slow even with data sets consisting of a few thousands of nodes (refer to Section 6 for more details). Though the design of a scalable heuristic for the B-IMCP problem is not the main focus of this paper, we here outline three heuristics for the proposed problem and we refer to the full version of this paper [22] for complete details about these heuristics.

(a) *Maximum Influence Heuristic (MIH):* The main idea behind this heuristic is motivated by Aggarwal *et al.* [23] and Chen *et al.* [8]. We now briefly present the main steps involved in the maximum influence heuristic as follows: (i) For each node $i \in V$, determine its influence spread; (ii) Sort the nodes in the network in non-increasing order of their influence spread values; and (iii) Pick nodes one by one as per the above sorted sequence and choose them as the initial seeds for appropriate products based on the ideas similar to that in Algorithm 1.

(b) *Maximum Degree Heuristic:* Following this heuristic, we first determine the nodes with high degree and then use steps similar to those presented in Algorithm 1 to construct the initial seed set.

(c) *Random Heuristic:* Following this heuristic, we select nodes uniformly at random to construct the initial seed set for the B-IMCP problem.

6   Experimental Evaluation

The goal of this section is to present experimental evaluation of the algorithms for the B-IMCP problem. We compare and contrast the performance of the proposed approximation algorithm, maximum influence heuristic, maximum degree heuristic and random heuristic. Throughout this section, we use the following acronyms to represent various algorithms: (i) *GA* to represent the proposed greedy approximation algorithm (i.e., Algorithm 1), (ii) *MIH* to represent the maximum influence heuristic, (iii) *MDH* to represent the maximum degree heuristic, and (iv) *Random* to represent the random heuristic. All the experiments are executed on a desktop computer with (i) Intel(R) Core (TM) i7 CPU (1.60 GHz speed) and 3.05 GB of RAM, and (ii) 32-bit Windows XP operating system. Each experimental result is taken as the average over $R = 1000$ repetitions of the same experiment. Further, we note that all the algorithms are implemented using JAVA.

6.1 Description of the Data Sets

Here we briefly describe the social network data sets that we use in our experiments.

WikiVote: This network data set contains all the users and discussion from the inception of Wikipedia till January 2008. Nodes in the network represent wikipedia users and a directed edge from node i to node j represents that user i voted on user j. This data set contains 7115 nodes and 103689 edges [24].

High Energy Physics (HEP): This is a weighted network of co-authorship between scientists posting preprints on the High-Energy Theory E-Print Archive between Jan 1, 1995 and December 31, 1999. This is compiled by Newman [25]. This network has 10748 nodes and 52992 edges.

Epinions: This is a who-trust-whom online social network of a general consumer review site Epinions.com. This data set consists of 75879 nodes and 508837 edges [26].

Telecom Call Data: This data set contains all the details pertaining to a call such as the time, duration, origin, and destination of the phone call. This data is collected from one of the largest mobile operators in a emerging market. We construct a graph from this data using the approach proposed in Nanavati *et al.* [27]. This data set consists of 354700 nodes and 368175 edges.

A summary of all the data sets described above is given in Table 6.1.

Data Set	Number of Nodes	Number of Edges
WikiVote	7115	103689
HEP	10748	52992
Epinions	75879	508837
Telecom Call Data	354700	368175

Table 1. Summary of the data sets used in the experiments

6.2 Experimental Setup

We follow the proposed LT-CP model of information diffusion. Given a weighted and directed social graph $G = (V, E)$ with a probability/weight w_{ij} for each edge (i, j) in E, we normalize these probabilities/weights as follows. Assume that node $i \in V$ has directed edges coming from nodes $j_1, j_2, \ldots, j_x$ with weights $q_{j_1 i}, q_{j_2 i}, \ldots, q_{j_x i}$ respectively. These weights represent the extent the neighbors of node i influence node i. Now let $q = q_{j_1 i} + q_{j_2 i} + \ldots + q_{j_x i}$; then the probability that node j_1 influence node i is given by $w_{j_1 i} = \frac{q_{j_1 i}}{q}$, the probability that node j_2 influence node i is given by $w_{j_2 i} = \frac{q_{j_2 i}}{q}$, and so on. Thus note that $w_{j_1 i} + w_{j_2 i} + \ldots + w_{j_x i} \le 1$ and this setting coincides with the proposed model in Section 3.1.

We have carried out the experiments with several configurations of the parameters and we obtained similar results for each of these configurations. In the interest of space, we in particular present the results for the following configuration of the parameters. We consider two products each in P_1 and P_2 respectively; i.e. $|P_1| = |P_2| = t = 2$. Also we consider that $B = 100$, $c_1^1 = 5$, $c_2^1 = 4.5$, $b_1^1 = 7$, $b_2^1 = 6.5$, $c_1^2 = 3$, $c_2^2 = 2.5$, $b_1^2 = 5$, $b_2^2 = 4.5$.
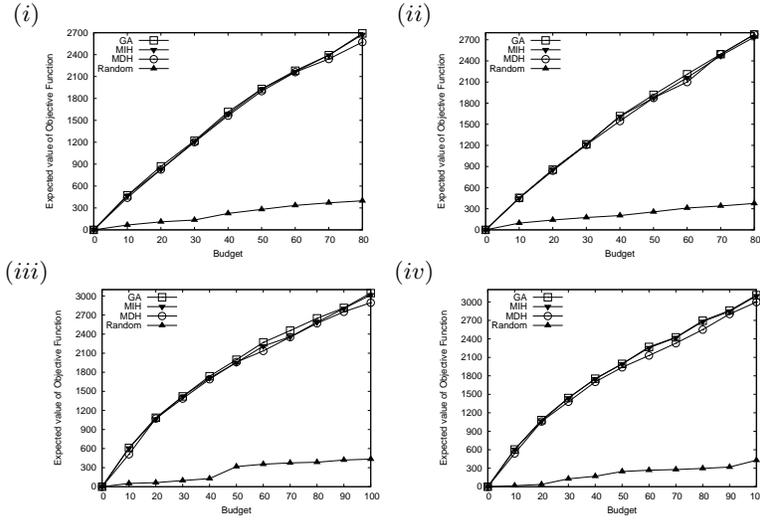
Fig. 1. Performance comparison of GA, MIH, MDH, and Random when the first variant of cross-sell is considered and (i) Dataset is HEP and cross-sell threshold is $[0, 0.5]$, (ii) Dataset is HEP and cross-sell threshold is $[0, 0.2]$, (iii) Dataset is WikiVote and cross-sell threshold is $[0, 0.5]$, and (vi) Dataset is WikiVote and cross-sell threshold is $[0, 0.2]$

We also work with two intervals for the cross-sell thresholds by setting $a = 0.2$ and $a = 0.5$ (refer to Section 3.1). This implies that the cross-sell thresholds come from two types of intervals, namely $[0, 0.2]$ and $[0, 0.5]$.

6.3 Experimental Results

We would like to compute the value of the objective function for the B-IMCP problem by varying the budget level using the four algorithms, namely GA, MIH, MDH, and Random. The experimental results in this setting are shown in Figure 1. These graph plots are obtained using HEP and WikiVote data sets and when the cross-sell thresholds come from the intervals [0,0.5] and [0,0.2] respectively. From all these graph plots, it is clear that the performance of MIH and MDH is almost same as that of GA. However, note that the performance of Random is very poor compared to that of GA.

Data Set	Running Time of GA (in Sec.)	Running Time of MIH (in Sec.)
HEP	66563	2590
WikiVote	148736	7200

Table 2. Running Times of GA and MIH on HEP and WikiVote Datasets
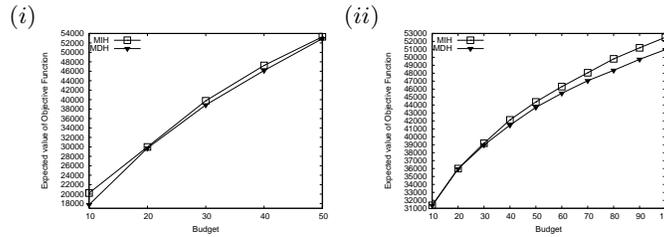
Fig. 2. Performance Comparison of MIH and MDH on Two Large Network Data Sets, namely (i) Epinions and (ii) Telecom Call Detail Records Data Sets

Experiments with Large Network Data Sets In this section, we focus on the running time of GA. Table 2 shows the running times of GA and MIH on HEP and WikiVote data sets. Clearly, the running time of GA is slower than MIH about 20 times.

We now present experimental results with large network data sets using MIH and MDH. Figure 2 shows the budget versus the expected value of the objective function curves for MIH and MDH using Epinions and Telecom data sets, when the first variant of cross-sell is used and cross-sell thresholds for nodes come from the interval $(0, 0.5)$. It is immediate to see that the performance is MIH is superior than that of MDH on these two data sets.

7 Conclusions and Future Work

In this paper, we introduced a generalized version of the influence maximization problem by simultaneously considering three aspects such as cross-sell phenomenon, product specific costs and benefits and budget constraints. We proposed a simple greedy algorithm to address this generalized version of the influence maximization problem. There are several ways to extend this work in this paper. First, it is interesting to examine other types of possibility of the cross-sell among the products. Second, we considered an onto function to represent the cross-sell relationships in this paper. However, it is important to work with other types of functions to represent the cross-sell relationships while retaining the properties of the diffusion model such as monotonicity and submodularity.

References

1. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of the 9th SIGKDD. (2003) 137–146
2. Rogers, E.: Diffusion of Innovations. Free Press, New York, USA (1995)
3. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets: Reasoning about a Highly Connected World. Cambridge University Press, Cambridge, UK (2010)
4. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
5. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proceedings of the 7th SIGKDD. (2001) 57–66

6. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proceedings of the 8th SIGKDD. (2002) 61–70
7. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: In SIGKDD. (2007) 420–429
8. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of the 15th ACM SIGKDD. (2009) 937–944
9. Kimura, M., Saito, K.: Tractable modles for information diffusion in social networks. In: Proceedings of 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD). (2006) 259–271
10. Mathioudakis, M., Bonchi, F., Castillo, C., Gionis, A., Ukkonen, A.: Sparsification of influence networks. In: Proceedings of the 17th SIGKDD. (2011)
11. Even-Dar, E., Shapira, A.: A note on maximizing the spread of influence in social networks. In: Workshop on Internet and Network Economics (WINE). (2007) 281–286
12. Datta, S., Majumder, A., Shrivastava, N.: Viral marketing for multiple products. In: Proceedings of IEEE ICDM. (2010) 118–127
13. Bharathi, S., Kempe, D., Salek, M.: Competitive in uence maximization in social networks. In: Workshop on Internet and Network Economics (WINE). (2007) 306–311
14. Borodin, A., Filmus, Y., Oren, J.: Threshold models for competitive influence in social networks. In: Workshop on Internet and Network Economics (WINE). (2010) 539–550
15. Carnes, T., Nagarajan, C., Wild, S., van Zuylen, A.: Maximizing in uence in a competitive social network: a follower's perspective. In: Proceedings of the 9th International Conference on Electronic Commerce (ICEC). (2007) 351–360
16. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology **83** (1978) 1420–1443
17. Schelling, T.: Micromotives and Macrobehavior. W.W. Norton and Company (1978)
18. Calinescu, G., Chekuri, C., Pal, M., Vondrak, J.: Maximizing a submodular set function subject to a matroid constraint. In: Proceedings of IPCO. (2007) 182–196
19. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions-1. Mathematical Programming **14**(1) (1978) 265–294
20. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters **12**(3) (2001) 211–223
21. Ramasuri, N., Narahari, Y.: A shapley value based approach to discover influential nodes in social networks. IEEE Transactions on Automation Science and Engineering **8**(1) (2011) 130–147
22. Ramasuri, N., Nanavati, A.: Viral marketing with competitive and complementary products through social networks. Technical report, IBM Research, India. URL: http://lcm.csa.iisc.ernet.in/nrsuri/Cross-Sell-Suri-Amit.pdf (2012)
23. Aggarwal, C., Khan, A., Yan, X.: On flow authority discovery in social networks. In: Proceedings of SIAM Conference on Data Mining (SDM). (2011) 522–533
24. Leskovec, J., Huttenlocher, D., Kleinberg, J.: Signed networks in social media. In: Proceedings of the 28th ACM SIGCHI Conference on Human Factors in Computing Systems (CHI). (2010) 1361–1370
25. Newman, M.E.J.: The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences (PNAS) **98** (2009) 404–409
26. Richardson, M., Agrawal, R., Domingos, P.: Trust management for the semantic web. In: Proceedings of Proceedings of ISWC. (2003)
27. Nanavati, A., Singh, R., Chakraborty, D., Dasgupta, K., Mukherjea, S., Das, G., Gurumurthy, S., Joshi, A.: Analyzing the structure and evolution of massive telecom graphs. IEEE Transactions on Knowledge Discovery and Data Engineering **20**(5) (2008) 703–718