Information Cascade at Group Scale

Milad Eftekhar
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
milad@cs.toronto.edu

Yashar Ganjali
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
yganjali@cs.toronto.edu

Nick Koudas
Dept. of Computer Science
University of Toronto
Toronto, ON, Canada
koudas@cs.toronto.edu

ABSTRACT

Identifying the k most influential individuals in a social network is a well-studied problem. The objective is to detect k individuals in a (social) network who will influence the maximum number of people, if they are independently convinced of adopting a new strategy (product, idea, etc). There are cases in real life, however, where we aim to instigate groups instead of individuals to trigger network diffusion. Such cases abound, e.g., billboards, TV commercials and newspaper ads are utilized extensively to boost the popularity and raise awareness.

In this paper, we generalize the "influential nodes" problem. Namely we are interested to locate the most "influential ***groups***" in a network. As the first paper to address this problem: we *(1)* propose a fine-grained model of information diffusion for the group-based problem, *(2)* show that the process is submodular and present an algorithm to determine the influential groups under this model (with a precise approximation bound), *(3)* propose a coarse-grained model that inspects the network at group level (not individuals) significantly speeding up calculations for large networks, *(4)* show that the diffusion function we design here is submodular in general case, and propose an approximation algorithm for this coarse-grained model, and finally by conducting experiments on real datasets, *(5)* demonstrate that seeding members of selected groups to be the first adopters can broaden diffusion (when compared to the influential individuals case). Moreover, we can identify these influential groups much faster (up to 12 million times speedup), delivering a practical solution to this problem.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications; J.4 [**Social and Behavioral Sciences**]: Economics

Keywords

Influential Groups, Group Diffusion, Information Cascade, Social Networks

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

1. INTRODUCTION

Innovation diffusion (information cascade), the study of network entities' reactions against new objects and ideas, has been a hot topic in social sciences since 1890, and in game theory and computer science for the last two decades [5, 10, 13]. In this field of study, we are interested in how, why, and when the entities in a society adopt an innovation (an idea, information, a product, a behavior, a culture, an emotion, a virus, a disease, or other objects that are "perceived as new by an individual or other unit of adoption" [32]). We are also interested in how this adoption impacts friends and neighbors, and thus the overall cascade of innovation in a network. The spread of a particular health trend like obesity or happiness in a community [8], switching from a product or a service to another, support of specific political parties in an election, participating in political uprisings in unsteady societies are only a few examples emphasizing the importance of information diffusion studies.

Communication and interpersonal relationships play a principal role in spreading innovations among members of a community. To maximize this spread, wise selection of the first adopters is crucial. The problem of identifying the most influential set of k people in a social network (the "***seed set***" or the "***first adopters***") has received much attention in the literature. ***Final influence*** of a seed set S is defined as the number of people that will eventually adopt the innovation if S is the set of all members that initially adopt it. Thus, the goal is to find a seed set S with k members that has the highest final influence among all sets of size k. Clearly, the final influence of a seed set depends on the interpersonal relationship of the nodes in the network. The problem has been studied extensively [7, 23, 26, 31]. It has been shown that this problem is NP-hard for most models studied, even for very simple special cases [23]. Furthermore, there are models for which even approximating the optimal value within a factor of $n^{1-\epsilon}$ is NP-hard (n is the number of individuals) [30]. Thus, several heuristic and greedy algorithms have been proposed to approximate the best solution for models that are susceptible to achieve good approximation results [7, 10, 23, 35].

In this paper, we generalize this problem and study the problem of picking influential ***groups*** rather than individuals to target, as well as how this choice impacts the *final influence*. Here, we define each group as the set of people that can be targeted using a specific advertising medium: a given group can include a demographic in a city (targeted by a TV commercial); all highway drivers passing by specific billboards can form a group; another group can include

people who attend a seminar/conference, etc. As a result of targeting a typical group g, some members of g become convinced (these members would constitute the set of first or early adopters) and diffuse to others. In this paper, ***Individual Diffusion*** *($\mathcal{ID}$)* refers to a diffusion in the network where the seed set consists of individuals and ***Group Diffusion*** *($\mathcal{GD}$)* refers to a diffusion in the network where the seed set consists of groups. Given a set of groups $\mathbf{M}$, the goal is to identify the most influential set of l groups. We expect to achieve the maximum *final influence* when we target groups in the most influential set compared to any other subset of $\mathbf{M}$ with size l. Here, ***final influence*** of a set $\mathcal{M}$ is defined as the number of people that will eventually adopt the innovation if $\mathcal{M}$ is the set of all groups that are targeted for initial convincing attempts such as advertisements. To the best of our knowledge, this is the first study of influential groups, as opposed to individuals; a simple paradigm shift that offers several advantages:

(1) **Groups and associations are natural targets of initial convincing attempts in many real-life scenarios.** There are many cases in real-life where people and organizations target groups. One example is advertising campaigns: companies usually advertise their products by targeting large groups of people. TV commercials, billboards, newspaper ads, etc. are popular ways of advertising since they reach a wide range of audiences ranging from the people living in an area to those reading specific newspapers and magazines. Online social networks, such as Facebook and MySpace, are other homes for a large number of groups. These groups, that consist of people with a common interest or characteristics, offer well-defined targets for advertisements. In online social networks, people aggregate naturally in fan pages and events offering self formed groups.

(2) **The running time for identifying influential groups is reasonable.** The individual-case problem is NP–hard in most studied cases. Due to the large size of social networks, solving this problem requires extremely fast algorithms. In the group-based problem, assuming we have m groups, and considering these groups as the vertices of a new graph, the naïve algorithm of examining all possible subsets of groups to find the most influential one takes exponential time in m. Obviously, in case that $m = O(\log(n))$ (when we target very large groups), this simple algorithm runs considerably fast (linear in n). In Section 5, we show that our proposed algorithms for identifying influential groups are fast.

(3) **Targeting groups can be an economical choice.** Convincing specific individuals to ensure they adopt a new strategy and building personal loyalty can be extremely expensive. Even offering a sample product to someone [20, 23] (which can be very costly for some products) does not necessarily convince the target to use it and recommend it to others in the network. In contrast, the cost of targeting a group is not necessarily linear in the size of the group. For instance, if the cost of convincing a specific individual is x units (in dollars, hours, or any other cost metric), targeting a large group of people might lead to an expected number of r convinced individuals with a cost significantly less than $r \times x$. The key difference here is that when we target an individual in the individual diffusion case, we need to ensure that the targeted person is convinced, which can be extremely costly. In contrast, in group diffusion we might gain the benefits as long as a reasonable fraction of group members are convinced regardless of who they are, and without the need to convince every single member of the group. The following example shows that depending on the advertising medium, the advertising cost can be very low for a large group of people.

Example 1. The monthly rent for a billboard is around \$700-\$2500 [14]. Assume that a firm wants to put a billboard on a highway. Moreover, assume that an average of 2 cars per second use the highway. Thus, in one month around 5 million cars access the highway. If we assume that an average of 2 people are in each car, for the monthly rent of \$2500, the cost of advertising is \$1 per 4000 people. Comparing individual and group advertising, we realize that directly targeting a few individuals has a cost equal to advertising to millions of people using mass media. Likewise, the cost of a national TV or a local TV commercial is about \$10-\$50 and \$5, respectively, for 30 seconds per 1000 viewers.

Given the advantages pointed in the analysis of group diffusion, the question is when the budget for the initial persuasion attempts (*e.g.*, advertisements) is fixed, what approach results in higher *final influence*: targeting influential individuals or targeting influential groups? This is one question we answer in this paper.

Contributions and overview. This paper provides the first analysis of the innovation diffusion problem from a group perspective. We start by presenting a simple fine-grained group diffusion model ($\mathcal{FGD}$) to identify influential groups in a social network. We show that locating these groups is NP-hard (in the number of groups) and provide an algorithm (`topfgd`) to identify the set of influential groups (with a guaranteed approximation bound) (Section 3).

The proposed algorithm is not practical for large networks when there are hundreds of thousands of nodes in the network. We present a coarse-grained group diffusion model $\mathcal{CGD}$ (the main contribution besides the paradigm shift to studying group influences and cascades) that looks into network at the higher level of groups not at the level of individuals (Section 4). We show that identifying the most influential set of groups in this new model is still an NP-hard problem (in the number of groups). After proving the submodularity of the final influence function in this coarse-grained group diffusion model, we present an algorithm `topcgd` that significantly speeds up the calculations providing similar results.

We use real datasets to evaluate our algorithms (Section 5). We show that targeting groups can broaden diffusion and increase the final influence up to 11 times (than targeting individuals) in our experiments. Moreover, our evaluations demonstrate that the two algorithms `topfgd` and `topcgd` that run on group diffusion models respectively run up to 700 and 12 million times faster than the individual diffusion `topid` algorithm. We also show that our `topcgd` algorithm, although removes lots of details about individuals (hence results in higher speed), can identify groups as influential as the groups found by the detailed `topfgd` algorithm.

2. BACKGROUND

Several models have been proposed to analyze the individual diffusion in social networks. The widely-studied models can be generalized into the categories of *threshold models* [10, 16, 23] and *cascade models* [6, 7, 23]. In the former, an **acceptance threshold** θ and an aggregation function f are associated with each node. We call a node **active**,

if it adopts the innovation. If X_v denotes the active neighbors of v and if $f_v(X_v) \geq \theta_v$, v becomes active. In the traditional version of "linear threshold model", the aggregation function f is the sum of the weights of edges from X_v to v. The cascading model, on the other hand, uses **activation success probabilities** for edges instead of acceptance thresholds for nodes. A node tries once to activate its neighbors after adopting the innovation and succeeds with specified probabilities. These probabilities can either be independently determined ("independent cascade model") or depend on both sides of diffusion (*i.e.*, the active diffuser node and the targeted neighbor) and the history of previous activation attempts. A survey of these diffusion models can be found in [11, 30]. Other model variations have been also proposed, ranging from changing budget constraints [19, 31] to increasing decision options [21]. Other applications of the problem has also been studied such as analyzing networks resilience when failures cascade in the network. Blume et al. have proposed techniques to evaluate the maximum failure probability of d-regular graphs [4]. Some works are also done to infer the underlying influence network based on causal relationships [34] or when the links are unobserved [15, 25, 27].

Finding the most influential set of k nodes is a well-known problem in social networks analysis [3, 7, 9, 10, 17, 23, 24, 26]. Since this problem is NP-hard in most studied cases [23], several heuristic and greedy algorithms have been proposed; these include the naïve algorithm of randomly choosing k nodes, the heuristic approaches of k central nodes, k high out-degree nodes [35], degree-discount [7], and the greedy algorithm of hill climbing [10, 23]. The mentioned greedy hill climbing algorithm is the most well-known algorithm for finding the target set, however it is not fast enough for large networks. Several extensions have been proposed to alleviate this problem [6, 7, 26].

3. FINE-GRAINED GROUP DIFFUSION ($\mathcal{FGD}$)

We start by a simple model ($\mathcal{FGD}$) and an algorithm (`topfgd`) to identify the set of most influential groups in the network. $\mathcal{FGD}$ is a fine-grained design that determines the final influence of targeting each group set by simulating the individual diffusion process inside. The basic idea is to determine how advertising to a group translates into individual adopters. Having these first adopters, we then run individual diffusion to identify the final influence. Later, we show why this approach is not practical for large graphs and how we can enhance this model, to aggregate individual-level information and run the algorithm in group scale rather than individuals with results nearly identical to $\mathcal{FGD}$ but at a fraction of its run time (that is the main contribution of this paper after the paradigm shift from the individual influence problem to the group influence problem).

3.1 Modeling

We model a (social) network using a graph, $G_{ind} = (V_{ind}, E_{ind})$. In this graph, V_{ind} is the set of individuals (nodes) in the network, and E_{ind} contains existing edges between pairs of individuals (nodes). This graph can be either undirected or directed and either unweighted or weighted. In weighted graphs, edges contain the amount of influence that individuals can exert on each other (normalized such that the sum of the weight of all incoming edges to any node does not exceed 1). A set of m groups $\mathbf{M} = \{g_1, ..., g_m\}$ is also available where each $g_i \subseteq V_{ind}$ $(1 \leq i \leq m)$ represents an input group in this (social) network.

To trigger a diffusion, we primarily target some groups by initial persuasion attempts, such as advertisements. As a result of these persuasion attempts, some members of these groups become active (adopt the innovation). These active nodes start diffusion by attempting to activate their neighbors and this process continues for newly activated nodes. We assume that the process is *progressive*, i.e., when a node adopts the innovation, it will not switch back [30].

Let's take a deeper look at the details of the proposed diffusion model. As stated in Section 1, targeting groups for activation is more economical than targeting individuals at times. Reconsider Example 1. Assume the cost for directly convincing an individual is \$100 (for example, by offering sample products as suggested by [20, 23]) and also assume there is a total advertisement budget of \$2500. Hence, we can directly activate 25 individuals. In Example 1, we concluded that around 10 million passengers pass the billboard in a month. If 0.05% of these people adopt the innovation, there will be 5000 active people before diffusion affects the network. Comparing the 25 active people in the first case with 5000 in the second case, this shows that group advertising might naturally extend budget power.

We define a parameter, called ***escalation factor*** β, to show the extent to which group targeting can increase the power of budget and the number of first adopters (compared to the individual case). The value of β changes based on the problem structure, the size and shape of the network, the products to be sold, the brands involved, the set of competitors as well as the initial convincing methods (*e.g.*, advertising media). While it is as low as 1 in a simple case where we target members of a group individually, it can be 200 in the billboard advertising example (where $\beta = 5000/25 = 200$). In online advertising, as another example, the average CPM (Cost Per thousands iMpressions) is \$5 [22] and CTR (Click-Through Rate) varies from $0.2-3\%$ in average [33] to 2% for Google AdWords to $2-12\%$ in email marketing campaigns and $5-15\%$ in email newsletters [2]. Assume the cost of directly convincing an individual is \$100. With the same budget (\$100), we achieve 20000 impressions (when CPM is \$5). If CTR rate is 2%, we get a β of 400. In general, by choosing the right media to target groups, we can readily achieve high values of β.

The model follows: we assume each node a has an inherent acceptance threshold θ_a (Section 2) that is identified according to the threshold model [10] (a random variable in [0,1]) and determines the influence a should receive to acquire the innovation. Let's say we spend a budget of C_g units to target group g. We assume, in each targeted group, the budget is distributed uniformly between all members. An arbitrary node a will achieve a budget of $\mathcal{C}_a = \sum_{g \in \mathbf{M}} IsMember(a, g) \times \mathfrak{C}_g$, where the binary function $IsMember(a, g)$ is 1 if node a is a member of group g and 0 otherwise, and $\mathfrak{C}_g$ is the *enhanced initial budget per capita* in group g that is $\mathfrak{C}_g = \frac{C_g \times \beta_g}{N_g}$ where N_g is the size of group g, C_g is the initial budget assigned to it, and β_g is its escalation factor (depending on the media used to target group g). In this paper, adhering to the most common approach of budget assignment, we assign equal budget to target each selected group. Thus, if we spend a total budget of C to target l groups, $C_g = \frac{C}{l}$ for all groups g targeted for initial persuasion attempts and is zero for other groups.

As previously mentioned, initial persuasion attempts activate some members of the targeted groups. Let's assume we need a budget of x units to directly convince one individual. Therefore, all nodes a for which $\frac{c_a}{x} > \theta_a$ are activated and are the first adopters. Consider A as the set of these first adopters when we target group g. Having A, we utilize the individual diffusion threshold model to simulate how the diffusion proceeds in G_{ind}. Here the final influence of targeting group g in the group diffusion model ($FI_{\mathcal{FGD}}(\{g\})$) is equivalent to the final influence of targeting A in the individual diffusion model ($FI_{\mathcal{ID}}(A)$); *i.e.*, $FI_{\mathcal{FGD}}(\{g\}) = FI_{\mathcal{ID}}(A)$. We stress, however, that both the threshold and the cascade models [23] can be utilized here by our algorithms to simulate the individual diffusion process. It has been shown that these diffusion process models are equivalent in the general case [23]. Proposing the fine-grained group diffusion model, our goal is to identify the set of l groups leading to the highest final influence.

Note that an equivalent representation for this model is to create a new graph G' built on $G_{ind} = (V_{ind}, E_{ind})$, add a node for each group $g \in \mathbf{M}$ to V_{ind}, and insert edges between the node representing the group g and all nodes representing users belonging in the group g with a weight of $\frac{c_g}{x}$. The individual diffusion model can be executed on this new graph by restricting the initial seed selection to the nodes representing the groups.

3.2 The most influential groups on the Fine-grained Group Diffusion Model

We consider the following problem.

PROBLEM 1. *Let G_{ind} be a given (social) network and $\mathbf{M}$ be the set of input groups. Moreover, let l be an input integer. Identify a subset $\mathcal{M} \subseteq \mathbf{M}$ with a size l leading to the highest final influence if it is targeted for initial persuasion attempts.*

THEOREM 1. *Problem 1 is NP-hard.*[1]

THEOREM 2. *In the fine-grained group diffusion model, the final influence function $FI_{\mathcal{FGD}}(X)$ is monotone and submodular.*

topfgd: An algorithm to identify TOP influential groups on FGD. We generalize the greedy algorithm proposed for the individual case problem [23] to address Problem 1. The generalized algorithm, topfgd, runs in l iterations. In iteration i, it inspects the marginal influence increase ($FI_{\mathcal{FGD}}(\mathcal{M}_{i-1} \cup \{g\}) - FI_{\mathcal{FGD}}(\mathcal{M}_{i-1})$) of each group $g \in \mathbf{M} - \mathcal{M}_{i-1}$, selects the group g^* corresponding to the maximum value and sets $\mathcal{M}_i = \mathcal{M}_{i-1} \cup \{g^*\}$. $FI_{\mathcal{FGD}}(X)$ is estimated according to Section 3.1. Note that $\mathcal{M}_l$ is the solution.

COROLLARY 1. *Based on Nemhauser et al. [28], since $FI_{\mathcal{FGD}}$ is a non-negative monotone submodular function,* topfgd *approximates the optimal value to within a factor of $1 - 1/e$.*

THEOREM 3. *The run time of* topfgd *algorithm is $T = O(l \times m \times (n + |E_{ind}|) \times R)$ where R determines the number of iterations we should simulate the diffusion using different thresholds.*

[1]All proofs are removed to save space. See proofs at [12]

4. COARSE-GRAINED GROUP DIFFUSION ($\mathcal{CGD}$)

The run time of the $\mathcal{FGD}$ diffusion process and the topfgd algorithm is not practical for large networks with lots of nodes and edges. It is crucial, hence, to design a diffusion process that scales up to larger networks and is practical to be utilized in real world scenarios and provides approximation guarantees. In this section, we propose a coarse-grained group diffusion model ($\mathcal{CGD}$) that incorporates information about individuals in its calculations but does not run explicitly on the level of individuals. Our goal is to design a model that simulates this diffusion by merely looking at the groups not individuals. By aggregating the individual-level information in groups, we hope to preserve the accuracy of our algorithms, and at the same time reduce the run time. Section 4.1 details modeling of the network, the groups and their relations. We discuss the diffusion process in Section 4.2 and present an approximation algorithm to identify the most influential groups in this model in Section 4.3.

4.1 Graph Models

We model a (social) network using two directed and weighted graphs, the graph of individuals G_{ind} and the graph of groups G_{group}. The graph G_{ind} is created similar to Section 3. We define $G_{group} = (V_{group}, E_{group})$ to model inter-group influences in a social network. In this graph, vertices are the predefined (input) groups, and edges represent the aggregate influence of members of each group on other groups. Intuitively, the higher the weights of edges $\{(a,b); (a,b) \in E_{ind}, a \in A, b \in B\}$, the higher the weight of the edge (A, B) in E_{group}. We define the weight of the edge from group A to group B as:

$$w_{AB} = \frac{\sum_{i \in A-B, j \in B-A} w_{ij} + N_{A \cap B}}{N_B} \quad (1)$$

where $0 \leq w_{ij} \leq 1$ is the weight of the edge from node i to j in G_{ind} and N_B is the size of group B. We consider $w_{ij} = 0$ if $(i,j) \notin E_{ind}$. Moreover, $N_{A \cap B}$ is the number of individuals that are members of both groups. Adding $N_{A \cap B}$ to the enumerator, we indeed set the influence weight of each individual on itself to 1. Clearly, $w_{AB} \leq 1$.

Figure 1 depicts a typical graph of individuals G_{ind} in (a) and its corresponding graph of groups G_{group} in (b). Individuals who are targeted by the billboard constitute a group (group 1). Similarly, individuals who are targeted by TV and the newspaper, respectively, form group 2 and 3. The thickness of an edge in Figure 1(b) represents its weight (influence value).

G_{group} is a graph where nodes are the input groups and edges are the influence weights of groups on each other. The question is whether we can identify the top-l influential groups in a network by running the individual diffusion model on its corresponding G_{group}. Generalizing the diffusion process for this coarse-grained group model is not a trivial extension of the individual diffusion. In fact, it is impossible to utilize previously proposed algorithms for the individual case on the group graph without considering the following factors. *(i)* Unlike the individual diffusion model, the decision to adopt an innovation and become active is not binary for nodes in the graph of groups. A model, that assumes the only possible cases in a group are that no member adopts the innovation or the entire group adopts it, is not adequate. *(ii)* We need to use a more developed diffusion

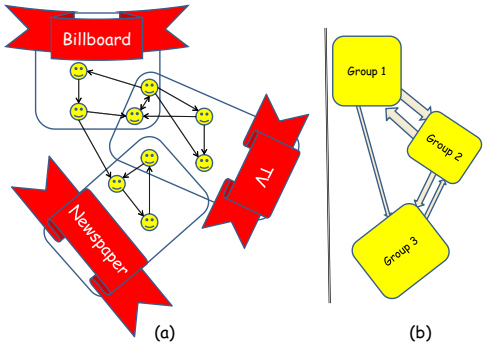


Figure 1: (a) A typical individual graph G_{ind}. (b) The corresponding group graph G_{group}. Thickness of edges in (b) represents the value of inter-group influences.

Table 1: Notations

Notation	Description
n – m	Total #individuals – Total #groups
k – l	Desired #individuals – Desired #groups
R	#iterations the diffusion process should be simulated by different threshold values.
t	Max #iterations for convergence (§ 4.2)
β – ρ	Escalation factor - Cohesion factor (§ 4.2)
λ	Progress fraction (§ 4.2)
d	Max #groups a node can belong to (§ 4.2)

model to analyze both intra-group and inter-group diffusion processes. *(iii)* The current techniques for the individual case greatly depend on the submodularity of the individual diffusion process; however, it is not easy to establish submodularity when we shift from the individual diffusion process to the coarse-grained group diffusion process. The following section describes the additional required steps. Table 1 summarizes the main notations used in this paper.

4.2 Diffusion Model

To trigger a diffusion in a network in $\mathcal{CGD}$, we primarily target some groups by initial persuasion attempts. Some members of these groups become active and start the diffusion process. We define a continuous variable, called a ***progress fraction*** λ, for each group to represent the fraction of group members that are currently active. We note that one cannot model group activation in the group diffusion model using a simple binary variable; usually, there is a mixture of active and inactive nodes in each group. Initially, the progress fraction value is zero for all groups except the targeted ones. There are several key factors that affect the initial progress fraction of targeted groups: the budget of initial persuasion attempts per capita, the group's structure, and the method of persuasion.

***(1)* Budget per capita spent to initially target groups.** We spend some budget to target each group. The budget, we allot on each group, is essential in determining how many nodes are active in that group. Up to a certain bound, we assume that higher budgets lead to higher progress fraction since, in this case, more time/money is spent to activate each group member.

***(2)* Group structure.** The cohesion of a group helps to increase its initial progress fraction. Within a cohesive (highly connected) group, members have such a high influence on each other that they can activate their group-mates more easily than in sparse groups. Thus, it is more likely that people, who are not ordinarily activated by initial convincing attempts, become active due to the influence of their group members. We define *group cohesion* ρ_i of group g_i as:

$$\rho_i = \frac{\sum_{l,j \in g_i} w_{lj}}{\sum_{k \in V_{ind}, j \in g_i} w_{kj}} \quad (2)$$

The denominator equals the group size N_{g_i}, when the sum of the weight of each node's incoming edges is 1.

***(3)* Power of the initial persuasion method.** As previously stated in Sections 1 and 3, targeting groups for activation can be more economical. Hence, the escalation factor β also plays an important role in initializing progress fraction values.

We define the initial progress fraction of group g_i as:

$$\lambda_{g_i}^{Init} = \min(1, \frac{\mathfrak{C}_{g_i}}{x \times (1 - p\rho_i)})$$

where $\mathfrak{C}_{g_i} = \frac{\beta_{g_i} \times C}{N_{g_i} \times l}$ is the *enhanced initial budget per capita* in g_i, ρ_i is its cohesion factor, x is the cost of directly convincing an individual, and p is the probability that each activation attempt succeeds (the *activation success probability* in the independent cascade model [23]). The intuition behind this definition follows. Merely based on system's initial persuasion attempts, $\alpha_i = \frac{\mathfrak{C}_{g_i}}{x}$ percent of g_i members become convinced (in fact α_i is the fraction of g_i members that can be directly convinced utilizing the enhanced budget). These convinced members diffuse inside the group. Since the success probability is p, the activations inside this group lead to an expected number of $\alpha_i \rho_i p$ additional convinced members. The newly activated members start the intra-group diffusion, and will succeed to increase the progress fraction by $(\alpha_i \rho_i p)\rho_i p$. Adhering to this intra-group diffusion process yields this progress fraction: $\lambda_{g_i}^{Init} = \alpha_i + \alpha_i(p\rho_i) + \alpha_i(p\rho_i)^2 + \alpha_i(p\rho_i)^3 + \cdots = \alpha_i \times \frac{1}{1-p\rho_i}$.

After calculating the initial values, to estimate the final progress fraction values (when diffusion has taken place between groups), we simulate the diffusion process by executing it in several iterations once the newly activated portion of each group tries to activate the neighboring groups. These attempts partly succeed according to the *activation success probability* p. Hence, it is expected that p percent of these activation attempts are successful. Based on the previous arguments, we propose the following coarse-grained group diffusion model once the process iteratively continues until it converges. In iteration $i+1$, the newly activated fraction of any group A (*i.e.*, $I = \lambda_A^i - \lambda_A^{i-1}$ where λ_A^i is A's progress fraction in iteration i) attempts to activate inactive members of neighboring groups B (inactive fraction is $J = 1 - \lambda_B^i$). As a result of this diffusion, some members of B become active (the fraction of the newly activated members is $I \times J \times w_{AB} \times p$) and try to activate their group-mates. This diffusion leads to a change in B's progress fraction:

$$\lambda_B^{i+1} = \min(1, \lambda_B^i + \frac{(\lambda_A^i - \lambda_A^{i-1}) \times (1 - \lambda_B^i) \times w_{AB} \times p}{1 - p\rho_B}) \quad (3)$$

We note that Equation 3 models both intra-group diffusion (the denominator) and inter-group diffusion (the numerator).

As previously mentioned, the goal is to find a seed set of l groups that maximizes the final influence. We approximate the final influence ($FI_{\mathcal{CGD}}$) utilizing Equation 4 (by considering inclusion-exclusion to take care of intersections):

$FI_{\mathcal{CGD}} =$

$$\sum_{j=1}^{d} (-1)^{j+1} \sum_{\substack{i1,\ldots,i_j: \\ 1\le i_1<i_2<\cdots<i_j\le|V_{group}|}} N_{\bigcap_{e=1}^{j} g_{i_e}} \prod_{e=1}^{j} \lambda^F_{g_{i_e}} \quad (4)$$

where $\lambda^F_{g_{i_e}}$ represents the final progress fraction of group g_{i_e}. Moreover, $N_{\bigcap_{e=1}^{j} g_{i_e}}$ is cardinality of the intersection of groups $g_{i_1}, \cdots, g_{i_j}$. We assume that each individual can be a member of a constant number (d) of groups.[2]. To clarify Equation 4, as an example for two non-disjoint graphs g_1 and g_2, $FI_{\mathcal{CGD}} = N_{g_1}\lambda^F_{g_1} + N_{g_2}\lambda^F_{g_2} - N_{g_1\cap g_2}\lambda^F_{g_1}\lambda^F_{g_2}$.

4.3 The most influential groups on $\mathcal{CGD}$

We now turn to study the problem of identifying the most influential groups. The goal is to find a set of l groups that has the maximum final influence under $\mathcal{CGD}$ model.

PROBLEM 2. *Let G_{ind} and G_{group} be, respectively, the graph of individuals and groups of a (social) network. Let l be an input integer. Identify a set of groups $\mathcal{M} \subseteq \mathbf{M}$ with size l that has the maximum value of $FI_{\mathcal{CGD}}$ among all subsets of $\mathbf{M}$ with a size of l. Here $FI_{\mathcal{CGD}}$ is the final influence (Equation 4).*

THEOREM 4. *Problem 2 is NP-hard.*

THEOREM 5. *The final influence function in $\mathcal{CGD}$ model ($FI_{\mathcal{CGD}}$) is monotone and submodular.*

topcgd: An algorithm to identify TOP influential groups on CGD. The proposed coarse-grained group diffusion model (Section 4.2) resolves the three concerns raised in Section 4.1. In this section, we propose a greedy algorithm that runs $\mathcal{CGD}$ process on G_{group} to approximate the most influential l groups. The algorithm runs similarly to topfgd by a difference in the diffusion function that it should simulate. We note that topcgd utilizes the diffusion process discussed in Section 4.2. Algorithm 1 presents the pseudo code.

COROLLARY 2. *Since $\mathcal{CGD}$ is non-negative, monotone and submodular (Theorem 5), topcgd approximates the optimal value to within a factor of $1 - 1/e$.*

THEOREM 6. *The run time complexity of topcgd is $T_{\texttt{topcgd}} = \theta(|E_{ind}| + lm(mt + n))$ where t is the maximum number of iterations we continue the diffusion process to converge.*

Section 5 shows that small values of t (about 10) are sufficient. Note that the run time of topcgd does not depend on R. Recall that parameter R determines the number of times the diffusion should be executed in topid (topfgd) with different threshold values to identify influential nodes (groups). However since in topcgd we have a top-level look into the diffusion process and because of the continuous nature of progress fraction values (compared to the binary nature of activation status of nodes in the individual diffusion model), one simulation is sufficient. In fact it is sufficient in topcgd to know that, when there are n activation attempts with probability p, after infinite repetitions the expected number of successes would be np, while in topid and topfgd we need to know the exact active nodes. This is one reason (besides the size of the network) that topcgd performs much faster than topid and topfgd.

[2]This assumption holds in real online social networks. For example, Facebook lets users join up to 300 groups [1]

Algorithm 1: topcgd

input : $G_{ind}, \mathbf{M}, \beta, C, x, l, t, p$
output: $\mathcal{M}$: the set of l influential groups
// Creating G_{group} and the cohesion factors
foreach g, g' *in* $\mathbf{M}$ **do**
  Calculate $W_{gg'}$ according to Equation 1.
end
foreach g *in* $\mathbf{M}$ **do**
  Calculate ρ_g according to Equation 2.
end
// Identifying influential groups
$\mathcal{M}^0 \leftarrow \emptyset$
for $i \in 1 \ldots l$ **do**
  foreach g *in* $\mathbf{M} - \mathcal{M}^{i-1}$ **do**
    $\mathcal{M}_g \leftarrow \mathcal{M}^{i-1} \cup \{g\}$
    // Diffusion
    foreach g' *in* $\mathcal{M}_g$ **do**
      $\lambda^1_{g'} = \min(1, \frac{C\beta_{g'}}{lN_{g'}x(1-p\rho_{g'})})$
    end
    for $j = 1$ *to* t **do**
      $\tau = \arg\max_{g' \in \mathbf{M}}(\lambda^j_{g'} - \lambda^{j-1}_{g'})$
      for $g' \neq \tau$ *in* $\mathbf{M}$ **do**
        $\lambda^{j+1}_{g'} = min(1, \lambda^j_{g'} + \frac{(\lambda^j_\tau - \lambda^{j-1}_\tau)w_{\tau g'}p(1-\lambda^j_{g'})}{1-p\rho_{g'}})$
      end
    end
    // Computing $influence_g = FI_{\mathcal{CGD}}(\mathcal{M}_g)$
    $influence_g \leftarrow 0$
    for g' *in* $\mathbf{M}$ **do**
      $influence_g += N_{g'} \times \lambda^{t+1}_{g'}$
    end
    for v *in* V_{ind} **do**
      foreach $s \subseteq groups(v)$ **do**
        $adjust = (-1)^{N_s+1}$ // N_s: size of s
        foreach g' *in* s **do**
          $adjust \times = \lambda^{t+1}_{g'}$
        end
        $influence_g += adjust$
      end
    end
  end
  $\mathcal{M}^i = \mathcal{M}^{i-1} \cup \{\arg\max_{g \in \mathbf{M} - \mathcal{M}^{i-1}} influence_g\}$
end
return $\mathcal{M}^l$

5. EXPERIMENTS

This section compares the final influence and the run time of the individual algorithm (topid) with group algorithms (topfgd, and topcgd). In order to have an apple-to-apple comparison, we use the individual diffusion to compare the final influence of the seed sets identified by each algorithm. In other words, if S_1 is the set of individuals identified by topid, and S_2 and S_3 are the group sets respectively returned

by topfgd and topcgd, we compare $FI_{\mathcal{ID}}(S_1)$, $FI_{\mathcal{FGD}}(S_2)$ and $FI_{\mathcal{FGD}}(S_3)$. Note that to calculate $FI_{\mathcal{FGD}}$, the $\mathcal{FGD}$ process utilizes $\mathcal{ID}$ too, hence a fair comparison is provided.

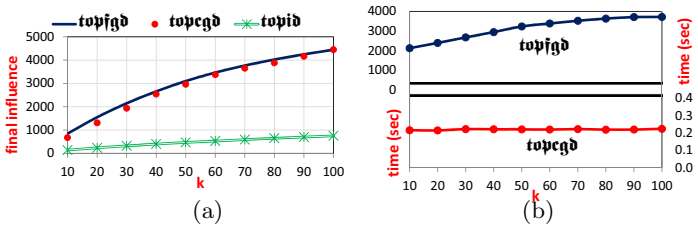
5.1 Experimental Datasets and Setup

In accordance with previous works [6, 7, 23], we employ a co-authorship dataset to compare our algorithms. The dataset is a network of co-authorship between scientists publishing papers or articles in computer science conferences or journals indexed by DBLP by Oct 21, 2012. As argued by Newman [29], co-authorship networks capture many key features of social networks. In the corresponding graph model, vertices are scientists, and edges represent co-authorship in at least one paper. To be more specific, assume P is the set of all papers with more than one author. Similar to [29], we define $\omega_{uv} = \sum_{p \in P} \frac{IS(p,u,v)}{coauthors(p)-1}$ where $coauthors(p)$ determines the number of coauthors who wrote paper p. Moreover, $IS(p, u, v)$ is 1 if u and v are both among coauthors of p and is 0 otherwise. To normalize weights (to make the sum of the weights of the incoming edges equal 1 for every node), we define the weight of edge uv as $w_{uv} = \frac{\omega_{uv}}{\sum_{u^* \in (V_{ind} - \{v\})} \omega(u^*, v)}$. This graph contains about 800 thousand nodes and 6.3 million directed weighted edges. Each conference or journal (*e.g.*, KDD, PVLDB) is a group. Each author is assigned to the top-3 groups that s/he has the highest number of publications in. The DBLP dataset contains about 3200 groups.

In addition, we have evaluated our algorithms on two other datasets. The first one is another co-authorship dataset in Physics [23, 29] with over 40 thousand nodes (authors) and 350 thousand directed edges (co-authorships). The second dataset is a dataset that consists of phone call records from one of the largest phone service providers in North America. This dataset contains the call records of over 1 million phone subscribers over the course of 2 months, which amounts to about 85 million phone calls and about 23 million edges. Here, the nodes are telephone numbers, and the edges represent the call history. Since information on group structures in the last two datasets is unknown, we created 1000 synthetic groups utilizing ROCK clustering algorithm [18] on the nodes. We observed similar trends in results over all 3 datasets hence we choose to report the results for the DBLP dataset that contains real groups.

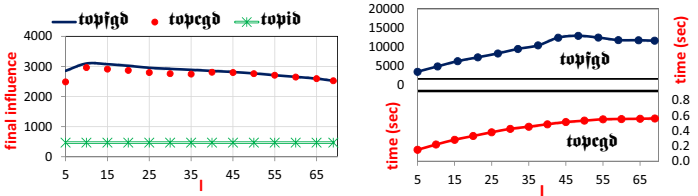
We conducted our experiments on a computer with 16 cores 2.6GHz (AMD OpteronTM Processor 850), 100GB of memory that is running CentOS 5.5. All algorithms are coded in Java and are single-threaded.

In the individual case of the problem, algorithms generate a set of k influential nodes as the output; the output in the group case is a set of l influential groups. The question we face is how to evaluate which approach leads to wider diffusion. Our strategy is to spend the same budget in all scenarios and evaluate the final influence. Assume $C = kx$ is the amount of budget we should spend to directly activate k influential nodes in the individual case. We spend the same budget to target l influential groups in the group cases. Assume $\mathcal{C}_a$ is the portion of the budget spent on node a (according to the $\mathcal{FGD}$ model, Section 3.1). To make meaningful comparisons, we start with a set of active nodes in all cases and measure the final influence using the individual diffusion techniques. For topid, the first adopters are the k influential individuals. In the grouping case (topfgd, topcgd), the first adopters are all nodes a for which $\frac{\mathcal{C}_a}{x} > \theta_a$



(a) (b)

Figure 2: Analyzing the impact of k



(a) (b)

Figure 3: Analyzing the impact of l

where θ_a is the acceptance threshold for a (the threshold model). For simplicity, we use the same β for all groups.

5.2 Experimental Results

Several experiments are conducted to compare the final influence and the run time of the algorithms. We study how these algorithms behave when various parameters change. As suggested by previous works [23], since the activation threshold values for individuals are not known, we determine the final influence for any seed set by running the diffusion process for $R = 10000$ times by re-choosing the activation thresholds δ, uniformly at random, and taking the average.

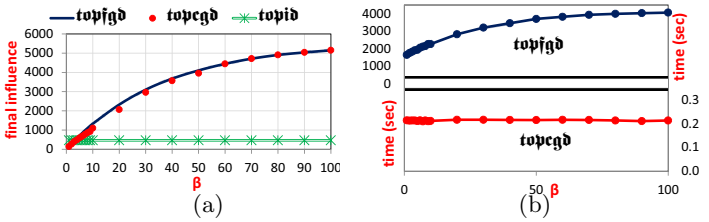
In our evaluations, we estimate final influence using the standard methods. We expect to see a similar decrease in the run time of all algorithms (topid, topfgd, topcgd) when the improvements of [6, 7, 26] (that are orthogonal to our work) are implemented.

5.2.1 *Individuals vs Groups: DBLP-1980 dataset*

Since topid and topfgd are not practical on large networks, in the first set of experiments we choose a subset of the DBLP dataset (the social graph created utilizing all publications published before 1980 referred as DBLP-1980) to compare topid, topfgd, and topcgd. This dataset consists of about 8 thousand nodes in 69 groups. The full DBLP dataset will be utilized later to evaluate topcgd. In the following experiments we employ the default setting of $k = 50$, $l = 10$, and $\beta = 30$. In each experiment, two parameters are fixed and the third one changes.

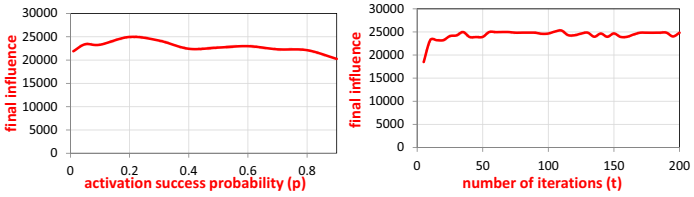
Impact of the initial convincing budget. Figure 2 depicts the final influence and the run time as a function of $k = \frac{\text{initial convincing budget}}{x}$ where x is the cost of directly convincing an individual. Increasing k (equivalently increasing the convincing budget) raises the final influence for all algorithms. Figure 2(a) offers a new insight: improvements in the group-based case are more significant when the budget is smaller due to the submodularity of the influence function.

Figure 2(b) reports the run time of the algorithms. Note the different scales in y-axis (kilosecond for topfgd vs. decisecond for topcgd). In fact, by increasing k from 10 to 100, the run time of topid raises from 50 hours to 700 hours and the run time of topfgd raises from 35 minutes to 61 minutes while the run time of topcgd remains about 200 milliseconds. The reason is that a higher value of k leads to more been activated nodes initially in topid and topfgd; hence, more edges



(a) (b)

Figure 4: Impact of the escalation factor



(a) Final influence vs. p (b) convergence

Figure 6: Impact of the topcgd parameters

are traversed while the diffusion is simulated. However, in the $\mathcal{CGD}$ model, increasing k only increases the progress fraction values of groups and the number of inter-group edges to be traversed does not change. We observe that topcgd is considerably faster than topfgd. Note that on this dataset with the default parameter setting, topid takes about 12 days while topfgd takes about 53 minutes and topcgd takes about 0.2 seconds to execute (including the modeling time to create the graph of groups G_{group} and ρ values). This suggests that utilizing topcgd can produces very similar results to the detailed topfgd algorithm much faster. Our experiments show that the run time of topcgd vs k is a constant curve while it is a concave downward curve for topfgd and a concave upward curve for topid. Hence, by increasing k the ratios of $\frac{\text{The run time of topid}}{\text{The run time of topfgd}}$ and $\frac{\text{The run time of topid}}{\text{The run time of topcgd}}$ significantly increase.

Impact of the desired number of groups to be targeted. Figure 3(a) shows that the final influence increases when l goes up to 10 and decreases afterwards. Thus, it is not the best strategy to spend all of our budget to target few very influential groups or split the budget on many groups. In the former (targeting a few number of groups), we restrict our influence coverage in the network. In the latter (targeting many groups), we spend a portion of our budget on less influential groups, hence obtaining less final influence.

Figure 3(b) studies the run time of group-based algorithms when l raises. We observe that as expected the run time of topfgd and topcgd increase when more influential groups should be identified. Since there is no parameter l in topid, its run time is 12 days for all values of l.

Impact of the escalation factor. Figure 4 shows the impact of the escalation factor β on the final influence and the run time. Note that β is not defined for topid hence we see a constant line. According to Figure 4(a), the group targeting algorithms (topfgd and topcgd) outperform the individual case (topid) in terms of final influence when β exceeds 3. Recall that based on our discussion in Section 3.1, the value of β for the billboard advertising in Example 1 is 200. Figure 4(a) shows that by targeting groups we achieve a final influence of twice as much as directly targeting individuals when $\beta = 10$ and this goes up to 11 times improvement when $\beta = 100$. Notice the concavity of results due to the submodularity of the final influence function.

By increasing β, the enhanced budget grows; hence, the run time of the topfgd increases, while the run time of the group-based topcgd algorithm remains unchanged (Figure 4(b)). The run time for topfgd starts from 1650 seconds when $\beta = 1$ and reaches 4100 seconds when $\beta = 100$, while the run time of topcgd remains about 200 milliseconds (same reason as in k's impact).

5.2.2 *Evaluation on the full DBLP dataset*

In this section, we evaluate the topcgd algorithm on the full DBLP dataset. Note that topid and topfgd are not practical on this large dataset. We compare topcgd with four baseline algorithms:

rnd: l groups are randomly selected to be targeted.
small: the l groups with the minimum sizes are selected.
big: the l groups with the maximum sizes are selected.
degree: the l groups that have the highest out-degree (defined as the weight on the outgoing edges) are selected. This influence for each group $g \in V_{group}$ is measured by $\sum_{g' \in V_{group}} w_{gg'}$ (Equation 1).

Figure 5 studies how topcgd behaves compared to the baseline algorithms when k, l, and β values are varied. Among the baseline algorithms, degree is the most accurate and small is the least accurate. We observe that in all experiments topcgd significantly outperforms all baseline algorithms.

We observed similar trends for run time of topcgd in the full DBLP dataset as DBLP-1980 dataset. We note that running topcgd with the default parameter setting $(\beta, k, l) = (30, 50, 10)$ takes less than 100 minutes in this dataset.

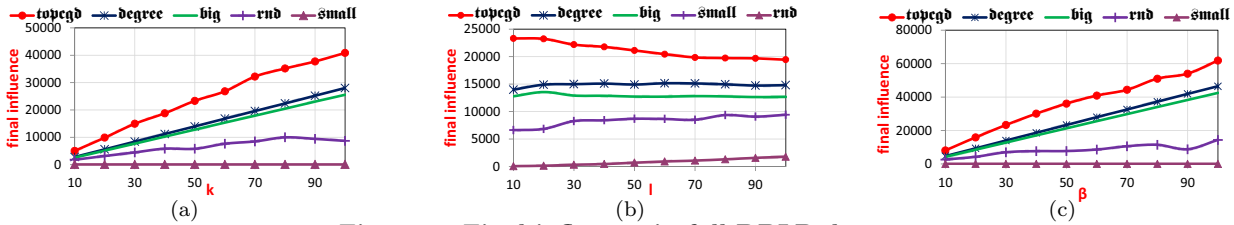
5.2.3 *How to adjust topcgd parameters?*

There are two other factors that should be evaluated for topcgd, namely (1) the activation success probability p and (2) the maximum number of iterations t to converge. We evaluated topcgd on the DBLP dataset for various values of p (varying from 0.01 to 0.9). The run time of topcgd increases from 94 minutes (when $p = 0.01$) to 122 minutes (when $p = 0.9$) and the final influence varies according to Figure 6(a) with the maximum occurring at $p = 0.2$. We also studied how the final influence of topcgd changes when we simulate the process for more number of iterations (Figure 6(b)). After t iterations pass, we identify the most influential l groups and measure the final influence when these groups are targeted. We observe that topcgd converges very fast within about 30 iterations.

6. CONCLUSION AND FUTURE WORK

This paper takes a closer look at innovation diffusion in networks when the focus is on groups rather than individuals. Under reasonable assumptions, we show that we can achieve wider diffusion and faster speeds by focusing on groups. We propose two models ($\mathcal{FGD}$ and $\mathcal{CGD}$) to simulate the diffusion in group scale. For each model, we present an approximation algorithm to identify the most influential groups. Our experiments on real datasets show that group algorithms (topfgd and topcgd) run much faster than the individual algorithm (topid): while it takes 30 days for topid in our most time-consuming experiment to determine influential individuals in the DBLP-1980 dataset, topfgd takes about one hour, and topcgd finishes in 0.2 seconds.

Although the $\mathcal{CGD}$ model aggregates the information about individuals and ignores many details (hence providing incredibly high speed), it results in a final influence comparable to the $\mathcal{FGD}$ model. In fact, in the individual diffusion



(a) (b) (c)

Figure 5: Final influence in full DBLP dataset

model as well as the fine-grained group diffusion model, we identify k entities (individuals or groups) that are highly influential for different values of nodes' thresholds θ. When we run $\mathcal{ID}$ and $\mathcal{FGD}$ for R times and obtain the average, we indeed consider the aggregate behavior of diffusion in the network. The aggregations in $\mathcal{CGD}$ produces very similar results in just one run. As a future work, we are interested to know how these algorithms behave when threshold values for nodes are known or they can be estimated. Another direction is to generalize the model to the cases where groups receive different budgets and/or the cost of advertising to each group is predetermined.

7. REFERENCES

[1] `http://www.facebook.com/help/199554316755501/`.

[2] Average email click-through rate. `http://bluesite.lyris.com/blog/85-Average-Email-Click-Through-Rate`.

[3] O. Ben-Zwi, D. Hermelin, D. Lokshtanov, and I. Newman. An exact almost optimal algorithm for target set selection in social networks. In *ACM Conf. on Electronic Commerce*, pages 355–362, 2009.

[4] L. Blume, D. Easley, J. Kleinberg, R. Kleinberg, and E. Tardos. Which networks are least susceptible to cascading failures. In *IEEE symposium on Foundations of Computer Science*, 2011.

[5] L. E. Blume. The statistical mechanics of strategic interaction. *Games and Economic Behavior*, 5:387–424, 1993.

[6] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1029–1038, 2010.

[7] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 199–208, 2009.

[8] N. A. Christakis and J. H. Fowler. *Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives*. Back Bay Books, USA, 2009.

[9] P. Dodds and D. Watts. Universal behavior in a generalized model of contagion. *Physical Review Letters*, 92(21):218701, 2004.

[10] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[11] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*, chapter 19. Cambridge University Press, USA, 2010.

[12] M. Eftekhar, Y. Ganjali, and N. Koudas. Information cascade at group scale. Technical Report `http://www.cs.toronto.edu/~milad/paper/GDA.pdf`.

[13] G. Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[14] Gaebler Ventures. Business Advertising. `http://www.gaebler.com/Business-Advertising.htm`.

[15] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 1019–1028, 2010.

[16] M. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

[17] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. Intl. World Wide Web Conf.*, 2004.

[18] S. Guha, R. Rastogi, and K. Shim. Rock: A robust clustering algorithm for categorical attributes. In *Proc. Intl. Conf. on Data Engineering*, pages 512–521, 1999.

[19] J. Hartline, V. S. Mirrokni, and M. Sundararajan. Optimal marketing strategies over social networks. In *Proc. Intl. World Wide Web Conf.*, pages 189–198, 2008.

[20] D. Iacobucci. Networks in marketing. pages 50–59. Sage, USA, 1996.

[21] N. Immorlica, J. Kleinberg, M. Mahdian, and T. Wexler. The role of compatibility in the diffusion of technologies through social networks. In *Proc. ACM Conf. on Electronic Commerce*, pages 75–83, 2007.

[22] A. Johnson. What's the average cpm for an online publisher? *Kikabink News, Internet Marketing News and Comment*, November 2010.

[23] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence in a social network. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[24] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM- The Blogosphere*, 47(12), December 2004.

[25] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proc. Intl. World Wide Web Conf.*, pages 641–650, 2010.

[26] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 420–429, 2007.

[27] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proc. Intl. Conf. on Information and Knowledge Management*, 2010.

[28] G. Nemhauser, L. Wolsey, and M. Fisher. An analysis of the approximations for maximizing submodular set functions. *Mathematical Programming, Springer Berlin*, 14:265–294, 1978.

[29] M. Newman. The structure of scientific collaboration networks. *National Academy of Sciences of the United States of America*, 98(2):404, 2001.

[30] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani. *Algorithmic Game Theory*, chapter 24. Cambridge University Press, USA, 2007.

[31] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proc. Intl. Conf. on Knowledge Discovery and Data Mining*, pages 61–70, 2002.

[32] E. M. Rogers. *Diffusion of Innovations*. New York: Free Press, USA, 1983.

[33] A. Stern. 8 ways to improve your click-through rate. *iMedia Connection*, February 2010.

[34] G. Ver Steeg and A. Galstyan. Information transfer in social media. In *Proc. Intl. Conf. on World Wide Web*, pages 509–518, 2012.

[35] S. Wasserman and K. Gaust. *Social Network Analysis*. Cambridge University Press, 1994.