Optimizing Budget Allocation Among Channels and Influencers

Noga Alon
Tel Aviv University and Microsoft Research
nogaa@tau.ac.il

Iftah Gamzu
Tel Aviv University and Microsoft Research
iftah.gamzu@cs.tau.ac.il

Moshe Tennenholtz
Microsoft Research and Technion
moshet@microsoft.com

ABSTRACT

Brands and agencies use marketing as a tool to influence customers. One of the major decisions in a marketing plan deals with the allocation of a given budget among media channels in order to maximize the impact on a set of potential customers. A similar situation occurs in a social network, where a marketing budget needs to be distributed among a set of potential influencers in a way that provides high-impact.

We introduce several probabilistic models to capture the above scenarios. The common setting of these models consists of a bipartite graph of source and target nodes. The objective is to allocate a fixed budget among the source nodes to maximize the expected number of influenced target nodes. The concrete way in which source nodes influence target nodes depends on the underlying model. We primarily consider two models: a source-side influence model, in which a source node that is allocated a budget of k makes k independent trials to influence each of its neighboring target nodes, and a target-side influence model, in which a target node becomes influenced according to a specified rule that depends on the overall budget allocated to its neighbors. Our main results are an optimal $(1 - 1/e)$-approximation algorithm for the source-side model, and several inapproximability results for the target-side model, establishing that influence maximization in the latter model is provably harder.

Categories and Subject Descriptors

F.2 [**Theory of Computation**]: Analysis Of Algorithms And Problem Complexity

General Terms

Algorithms, Economics, Theory

Keywords

Approximation algorithms, budget allocation, influence models

1. INTRODUCTION

Brands and agencies use marketing as a tool to influence customers and make them into loyal buyers. One of the major decisions in a marketing plan deals with the allocation of a given budget among media channels in order to maximize the impact on a set of potential customers. Several examples of media channels are TV, newspapers, billboards, and websites. In many cases, a fine-grained decision is needed regarding the budget allocated within a specific channel, e.g., how to distribute the budget for radio commercials among the different major radio channels. In some cases, there are additional constraints on the amount of budget that can be spent on any specific outlet. Such constraints may be enforced by the policy makers or the regulations in an organization.

An underlying assumption is that the amount of budget allocated to a channel determines the chances of influencing particular customers. As a result, one may model this setting as a bipartite graph in which one side is the set of possible marketing channels, and the other is the population of customers. An edge between channel i and a customer j indicates that j may influence i with some probability that depends on the budget allocated to j. We emphasize that a similar situation occurs in a social network, where a marketing budget needs to be distributed among a set of potential influencers in a way that provides high-impact. Given that an influencer can effect the decisions of her neighbors in the network, and that her level of effort depends on the budget allocated to her, we get a setting that is similar to the one described above.

Both of the above-mentioned scenarios deal with the allocation of budget among different channels. A convenient way for modelling the fact that the probability of a channel to influence a customer depends on its allocated budget is by focusing on discrete budgets. The allocation of k units of budget to channel j corresponds to k attempts made by channel j to influence each of its corresponding customers. We note that each such attempt may have a different probability to influence a customer. These probabilities depend on the underlying model, as formally described in Subsection 1.1. We emphasize that all these models leave much freedom in the way one measures the expected number of customers that are influenced by a particular budget allocation. We also note that probabilistic models are natural since our knowledge of customers behavior is inherently limited, and thus, we can only have some probabilistic estimate for the typical impact of a budget allocation to some channel. We like to point out two major differences between our models and previous work (e.g., [9, 18, 19, 23, 3]):

1. Budgets – previous work on so-called budget allocation in similar contexts have only focused on the selection of a subset of influencing nodes. In particular, the allocation of budget among nodes was not treated. We

WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1229-5/12/04.

believe this to be a central issue, essential in the context of an algorithmic approach to marketing.

2. Propagation of influence – our work does not deal with the propagation of influence, which may happen in social networks. We emphasize that this is due to the fact that a major motivation for our work is the distribution of budget among marketing channels, although we believe that our models are applicable to social networks as well.

1.1 The influence models

We primarily focus on two models: a source-side influence model and a target-side influence model. The common setting in both models consists of a bipartite graph $G = (S, T, E)$, where S and T are collections of *source* and *target* nodes, respectively, and $E \subseteq S \times T$ is an edge set. Each source node s has a *capacity* $c_s \in \mathbb{N}_+$, and there is a (global) *budget* $B \in \mathbb{N}_+$. The objective is to distribute the budget among the source nodes in a discrete way that respects the capacities of nodes, and maximizes the expected number of target nodes that become *active* (or influenced). Specifically, each source node s should be allocated a budget $b_s \in \{0, 1, \ldots, c_s\}$ such that $\sum_{s \in S} b_s \leq B$. The process in which target nodes become active depends on the underlying model:

- **Source-side influence model** – Each source node $s \in S$ has a probability vector $p^s = \langle p^s_1, \ldots, p^s_{c_s} \rangle \in [0,1]^{c_s}$. A source node s that is allocated a budget of b_s makes b_s independent trials to activate each neighboring target node t. The probability that t is activated by s in the ith trial is p^s_i. Thus, if the set of source nodes $\Gamma(t)$ designates the neighbors of t in G then the probability that t becomes active is

$$1 - \prod_{s \in \Gamma(t)} \prod_{i=1}^{b_s} (1 - p^s_i) \ .$$

  We note that all the trials that s makes (to all its neighboring nodes) are independent.

- **Target-side influence model** – Each target node $t \in T$ has a probability vector $p^t = \langle p^t_1, \ldots, p^t_B \rangle \in [0,1]^B$. Suppose the set of source nodes $\Gamma(t)$ designates the neighbors of t in G, and let $B_t = \sum_{s \in \Gamma(t)} b_s$. Then, the probability that t becomes active is

$$1 - \prod_{i=1}^{B_t} (1 - p^t_i) \ .$$

  We emphasize that the probability that a target node t becomes active is a function of the overall budget B_t allocated to its neighbors.

Notice that our models are very general as they support completely arbitrary probability vectors, and not necessarily ones that are monotone or have any other restrictive structural properties. Also note that one can argue that in real-life situations the exact values of the probabilities are unknown. We believe however that one can almost always have some good estimate for these probabilities, and therefore, we assume that these probabilities are known.

1.2 Our results

We study both influence models and attain the following results:

Source-side influence model. We devise an efficient deterministic algorithm that has an approximation ratio of $1 - 1/e \approx 0.632$ for influence maximization in the source-side model. We observe that this result is best possible under the assumption that P $\neq$ NP. Specifically, this result matches the NP-hardness bound of $1 - 1/e$, which already holds for the special setting of maximum coverage [10]. We also discuss several generalizations of our model and techniques. In particular, we demonstrate that our approach can be employed in a general setting where the underlying graph is arbitrary and each source node has a different influence on each of its neighbors. This latter setting is well-motivated as it may abstract the question of influence maximization in social networks.

Target-side influence model. We demonstrate that influence maximization in this model extends both the maximum edge biclique and the dense k-subgraph problems. Consequently, we attain among some additional hardness results that our problem is hard to approximate within a factor of $\Omega(1/B^\epsilon)$ for some $\epsilon > 0$, assuming a certain hypothesis about the average-case hardness of random 3SAT [11]. This implies that influence maximization in the target-side model is provably harder than in the source-side model. In light of the above state of affairs, we then focus our attention on a modified model in which there are no capacity constraints on the source nodes. We develop an efficient deterministic algorithm that achieves a logarithmic approximation ratio. This establishes a computational separation between the two target-side models, and proves that the budget capacity constraints underlie the hardness of the problem in the former model.

1.3 Related work

Previous work dealing with influence in multi-agent systems considered issues such as finding a set of most influential individuals, and understanding the effects of the social structure on emergent behavior (see, e.g., [21] and the reference therein). More generally, the diffusion and spreading of opinions in societies has long been a topic of study in the social sciences [26, 15], and later got the attention of game-theorists [31, 22, 17] and AI researchers [28] among others. While a main interest in this research concerns with the need to influence individuals, the related literature does not consider the distribution of budgets among those individuals; rather, it primarily concentrates on issues such as information propagation. Our work seems to be the first to optimize the budget distribution beyond deciding the identity of the individuals that should be approached. Our notion of a budget is a discrete one. This is consistent with common practices in organizations (e.g., working in multiplications of some fixed value) and related simulations [27]. Other work relating to the subject of budget constraints in markets and advertising appears in [6, 14, 8, 2, 5, 29, 7].

2. A SOURCE-SIDE INFLUENCE MODEL

In this section, we devise an efficient deterministic algorithm that has an approximation ratio of $1 - 1/e \approx 0.632$ for the problem of maximizing the influence in the source-

side model. This result is best possible under the assumption that P $\neq$ NP. Specifically, this result matches the NP-hardness bound of $1-1/e$, which already holds for the special setting of maximum coverage [10]. The latter setting is equivalent to the case in which all the probability vectors are of the form $\langle 1, *, \ldots, * \rangle$, where each $*$ can be any valid value. In this case, the objective reduces to selecting B source nodes in a way that maximizes the cardinality of their neighbor set.

We note that our algorithm and analysis has similarities with the algorithm for maximizing a nondecreasing submodular set function subject to a knapsack constraint [30]. We emphasize that our problem is not an instance of the latter problem. In particular, in our problem, one may select a node multiple times (i.e., allocate it a budget other than 0 or 1), and these selections may not satisfy a decreasing marginal influence property as required for submodularity. Informally, our problem does not admit a decreasing marginal influence property since adding an extra unit of budget to some node may have a higher marginal influence with respect to prior budget increments.

Definition 1. Let $\{b_s\}_{s\in S}$ be a feasible budget allocation to the source nodes, and let s and $\Gamma(s)$ be a source node and its neighbors in G, respectively. The *marginal influence* of adding a feasible budget of k to s is the expected increase in the number of target nodes that become active. One can easily verify that this amounts to

$$\Delta(\{b_s\}_{s\in S}, s, k) = \sum_{t\in\Gamma(s)} \left(\prod_{v\in\Gamma(t)} \prod_{i=1}^{b_v} (1-p_i^v) \cdot \left(1 - \prod_{j=1}^{k} (1-p_{b_s+j}^s)\right)\right) .$$

In accordance, the *per unit marginal influence* of adding a budget of k to s is $\Delta(\{b_s\}_{s\in S}, s, k)/k$.

2.1 The algorithm

Let $\ell \geq 3$ be a fixed integer, and let $\mathcal{P}$ be the set of all solutions in which at most ℓ source nodes are allocated feasible budgets. Specifically, in each such solution, the allocated budgets respect the capacities of the corresponding nodes, and the overall allocated budget is at most B. One can easily verify that the cardinality of the set $\mathcal{P}$ is $O(|S|^\ell B^\ell)$, that is, polynomial in the parameters of the problem as long as ℓ is fixed.

Our algorithm enumerates over all the solutions in $\mathcal{P}$, where each solution is utilized as an initial budget allocation. Given some initial allocation, the algorithm completes it greedily. Specifically, let $\{b_s\}_{s\in S}$ be the initial allocation in which a set $S' \subseteq S$ was allocated an overall budget of $B' \leq B$. The greedy procedure completes the budget allocation of nodes in $S \setminus S'$ as follows:

Once the enumeration phase ends, the algorithm considers all the resulting budget allocations (each corresponds to some initial budget allocation from $\mathcal{P}$), and outputs the one whose expected influence is maximal.

2.2 Analysis

In what follows, we prove that our algorithm has an approximation ratio of $1-1/e \approx 0.632$. This result matches the NP-hardness bound of $1-1/e$, which holds for the special setting of maximum coverage [10]. We begin by introducing a notation and terminology that will be used later:

Algorithm 1 GreedyCompletion

Input: A feasible budget allocation $\{b_s\}_{s\in S}$, and a remaining budget $K = B - B'$
Output: An updated budget allocation

$\mathcal{Q} \leftarrow$ the set of all pairs $\langle s, k\rangle$ such that $s \in S \setminus S'$ and $k \in \{1, \ldots, \min\{K, c_s - b_s\}\}$
while $K \geq 0$ **do**
  for all $\langle s, k\rangle \in \mathcal{Q}$ **do**
    $\delta_{s,k} \leftarrow \Delta(\{b_s\}_{s\in S}, s, k)/k$
  end for
  let $\langle s, k\rangle \in \mathcal{Q}$ be a pair with a maximal $\delta_{s,k}$
  if $K \geq k$ **then**
    $b_s \leftarrow b_s + k$, $K = K - k$
    modify all pairs $\langle s, k'\rangle \in \mathcal{Q}$ to $\langle s, k' - k\rangle$
    remove all pairs $\langle s, k'\rangle \in \mathcal{Q}$ such that $k' \leq 0$
  else
    remove $\langle s, k\rangle$ from $\mathcal{Q}$
  end if
end while

- Let $\mathcal{A}^* = \{\langle s_1^*, b_1^*\rangle, \ldots, \langle s_{r^*}^*, b_{r^*}^*\rangle\}$ be an optimal solution for a given input instance of the problem. Here, each pair $\langle s_i^*, b_i^*\rangle \in S \times \mathbb{N}_+$ indicates that the source node s_i^* is allocated a (positive) budget of b_i^* in the optimal solution. We assume that the pairs in $\mathcal{A}^*$ are ordered according to a non-increasing marginal influence. Namely, the marginal influence of the pair $\langle s_1^*, b_1^*\rangle$ with respect to the empty solution is the highest among all other pairs, the marginal influence of the pair $\langle s_2^*, b_2^*\rangle$ with respect to the solution that consists of the pair $\langle s_1^*, b_1^*\rangle$ is the highest among all remaining pairs, and so on.

- Let $\mathrm{OPT} = \mathrm{OPT}_1 + \mathrm{OPT}_2$ be the expected number of target nodes that become active in the optimal solution. Here, OPT_1 indicates the overall marginal influence of the first ℓ pairs in $\mathcal{A}^*$ (with respect to the empty solution), and OPT_2 stands for the overall marginal influence of the remaining pairs in $\mathcal{A}^*$ (with respect to the solution that consists of the first ℓ pairs). Note that if $r^* \leq \ell$ then $\mathrm{OPT}_1 = \mathrm{OPT}$ and $\mathrm{OPT}_2 = 0$.

Recall that our algorithm enumerates over all the (initial) solutions in which at most ℓ source nodes are allocated feasible budgets. Hence, if $r^* \leq \ell$ then our algorithm finds an optimal budget allocation. Accordingly, in the remainder of this subsection, we deal with the case that $r^* > \ell$. We concentrate on the solution $\mathcal{A}$ that our algorithm generates with respect to the initial solution which consists of the first ℓ pairs in $\mathcal{A}^*$. Namely, we assume that $\mathcal{A}$ initially consists of $\{\langle s_1^*, b_1^*\rangle, \ldots, \langle s_\ell^*, s_\ell^*\rangle\}$. Clearly, if we prove that this solution is a $(1-1/e)$-approximation for the optimal outcome then our algorithm is guaranteed to have (at least) the same approximation ratio.

Let $\mathrm{ALG} = \mathrm{ALG}_1 + \mathrm{ALG}_2$ be the expected number of target nodes that become active in the above-mentioned solution $\mathcal{A}$. Here, ALG_1 indicates the overall marginal influence of the first ℓ pairs in $\mathcal{A}$ (with respect to the empty solution), and ALG_2 stands for the overall marginal influence of the remaining pairs in $\mathcal{A}$ (with respect to the solution that consists of the first ℓ pairs). Note that $\mathrm{ALG}_1 = \mathrm{OPT}_1$ due to our assumption regarding the enumeration step. Hence,

we are left to analyze the performance of the solution of the greedy completion procedure.

Recall that the greedy procedure is built around one main loop. In each iteration i of that loop, the algorithm extends the current solution with a pair $\langle s_i, k_i \rangle$ whose per unit marginal influence is maximal, namely, the budget of node s_i is increased by k_i. We mark the per unit marginal influence in that iteration by δ_i, and the corresponding marginal influence by $\Delta_i = \delta_i k_i$. Notice that if the required budget k_i is more than the remaining budget then the algorithm cannot extend the current solution. In such a case, the algorithm does not increase the budget of any node, the underlying pair is removed, and the loop continues. Also note that the algorithm may increase the budget of any node multiple times in different iterations.

Let $\langle s_L, k_L \rangle$ be the first pair for which the greedy procedure cannot extend the current solution. For the sake of the analysis, we may assume without loss of generality that $\langle s_L, k_L \rangle$ involves an increase towards the optimal solution. That is, suppose the current budget allocation is $\{b_s\}_{s \in S}$ then $b_s + k_L \leq b^*_s$. We note that our assumption is valid since if there is such a prior pair that does not involve an increase towards the optimal solution then excluding it from the potential pairs list $\mathcal{Q}$ does not change the greedy solution (with respect to the same initial solution), the optimal solution, and the analysis. Note that such a pair may be excluded by initially adjusting the budget constraint of the underlying node in an appropriate way. Lastly, we remark that in case that the greedy procedure does not experience a situation in which it cannot extend the current solution then we denote the last inspected pair by $\langle s_L, k_L \rangle$.

Lemma 1. $\Delta_i \geq k_i/K \cdot (\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j)$ in every iteration $1 \leq i \leq L$.

Proof. Let d_q be the difference between the budget allocated to node $s^*_{\ell+q}$ in the optimal solution and the budget allocated to it by the greedy procedure up to iteration i. In case the greedy procedure allocated more budget to node $s^*_{\ell+q}$ than the optimal solution then we set $d_q = 0$. Using a simple counting argument, one can attain that there is a node $s^*_{\ell+q^*}$ such that increasing its budget by d_{q^*} increases the overall influence by at least

$$\frac{d_{q^*}}{\sum_{q=1}^{r-\ell} d_q} \left(\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j \right) \geq \frac{d_{q^*}}{K} \left(\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j \right) ,$$

where the inequality follows since K is the overall budget allocated to the nodes $s^*_{\ell+1}, \ldots, s^*_{r^*}$ by the optimal solution. Since our greedy procedure selects a pair whose per unit marginal influence is maximal and the pair $\langle s^*_{\ell+q^*}, d_{q^*} \rangle \in \mathcal{Q}$, we get that $\delta_i \geq (\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j)/K$. □

Lemma 2. $\mathrm{OPT}_2 - \sum_{j=1}^{i} \Delta_j \leq \prod_{j=1}^{i} (1 - k_j/K) \cdot \mathrm{OPT}_2$ for every $0 \leq i \leq L$.

Proof. We prove this lemma by induction on i. The lemma clearly holds when $i = 0$. Assume that the lemma holds for $i - 1$ and notice that

$$\begin{aligned} \mathrm{OPT}_2 - \sum_{j=1}^{i} \Delta_j &\leq \mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j - \frac{k_i}{K} \cdot \left(\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j \right) \\ &\leq \left(1 - \frac{k_i}{K} \right) \cdot \left(\mathrm{OPT}_2 - \sum_{j=1}^{i-1} \Delta_j \right) \\ &\leq \prod_{j=1}^{i} \left(1 - \frac{k_j}{K} \right) \cdot \mathrm{OPT}_2 , \end{aligned}$$

where the first inequality follows from Lemma 1, and the last inequality is due to the induction hypothesis. □

Corollary 1. $\sum_{i=1}^{L} \Delta_i \geq (1 - 1/e) \cdot \mathrm{OPT}_2$.

Proof. Let us define $\varphi(x) = \ln(1 - x)$. Notice that φ is concave and monotonically decreasing in the range $(0, 1]$. This implies that given $x_1, \ldots, x_L \in (0, 1]$, we know that $\sum_{i=1}^{L} \varphi(x_i)/L \leq \varphi(\sum_{i=1}^{L} x_i/L)$ by Jensen's inequality. Substituting each x_i with k_i/K, we obtain that

$$\frac{1}{L} \sum_{i=1}^{L} \ln \left(1 - \frac{k_i}{K} \right) \leq \ln \left(1 - \frac{\sum_{i=1}^{L} k_i}{KL} \right) ,$$

or equivalently,

$$\prod_{i=1}^{L} \left(1 - \frac{k_i}{K} \right) \leq \left(1 - \frac{\sum_{i=1}^{L} k_i}{KL} \right)^L \leq \left(1 - \frac{1}{L} \right)^L ,$$

where the last inequality follows as $\sum_{i=1}^{L} k_i/K \geq 1$. Now, using the lemma, we conclude that

$$\begin{aligned} \mathrm{OPT}_2 - \sum_{i=1}^{L} \Delta_i &\leq \prod_{i=1}^{L} \left(1 - \frac{k_i}{K} \right) \cdot \mathrm{OPT}_2 \\ &\leq \left(1 - \frac{1}{L} \right)^L \cdot \mathrm{OPT}_2 \leq \frac{1}{e} \mathrm{OPT}_2 . \end{aligned}$$

□

We are now ready to prove the main theorem of this section.

Theorem 1. $\mathrm{ALG} \geq (1 - 1/e) \cdot \mathrm{OPT}$.

Proof. Recall that $\mathrm{ALG}_1 = \mathrm{OPT}_1$ due to our assumption regarding the enumeration step. In addition, notice that $\mathrm{ALG}_2 \geq \sum_{i=1}^{L-1} \Delta_i$ in case that the greedy procedure could not extend the solution in iteration L, and $\mathrm{ALG}_2 = \sum_{i=1}^{L} \Delta_i$, otherwise. In the former case, one can easily verify that $\Delta_L \leq \mathrm{OPT}_1/\ell$. This follows since the pair $\langle s_L, k_L \rangle$ involves an increase towards the optimal solution, and by our assumption regarding the enumeration step, the marginal influence of such an increase cannot be higher than the marginal influence of extending the solution that consists of the pairs $\bigcup_{j=1}^{i-1} \langle s^*_j, b^*_j \rangle$ with the pair $\langle s^*_i, b^*_i \rangle$, for all $1 \leq i \leq \ell$. As a result, $\mathrm{ALG}_2 \geq \sum_{i=1}^{L} \Delta_i - \mathrm{OPT}_1/\ell$. We can now conclude that

$$\begin{aligned} \mathrm{ALG} &= \mathrm{ALG}_1 + \mathrm{ALG}_2 \\ &\geq \mathrm{OPT}_1 + \sum_{i=1}^{L} \Delta_i - \frac{1}{\ell} \mathrm{OPT}_1 \\ &\geq \left(1 - \frac{1}{\ell} \right) \cdot \mathrm{OPT}_1 + \left(1 - \frac{1}{e} \right) \cdot \mathrm{OPT}_2 \\ &\geq \left(1 - \frac{1}{e} \right) \cdot \mathrm{OPT} , \end{aligned}$$

where the second inequality is due to Corollary 1, and the last inequality holds since $\ell \geq 3$. □

2.3 Discussion and extensions

Generalized graph models. A natural generalization of our model is when the edges, rather than the source nodes, are associated with probability vectors. In this setting, a source node s that is allocated a budget of b_s makes b_s independent trials to activate each neighboring target node t; the probability of activating t depends on the probability vector of the edge (s,t) in a similar way to the original model. Notice that an instance of the original model can be translated to an instance of the new model in which the probability vectors of all edges adjacent to any source node are identical. It is not hard to verify that our algorithm from Subsection 2.1 can be applied in this generalized model, attaining the same approximation ratio. In particular, one can easily validate that the counting argument from the proof of Lemma 1 is also applicable in this latter model.

An even more general model is when the underlying (directed) graph is arbitrary, rather than bipartite. In this setting, a node s that is allocated a budget of b_s makes b_s independent trials to activate each neighboring node t, where the probability of activating t depends on the probability vector of the edge $\langle s,t\rangle$. Furthermore, any node s may become influenced as a result of the budget allocated to it. This can be modelled by a self-loop in the underlying graph, that is, an edge $\langle s,s\rangle$, where this edge has a probability vector having the same interpretation as before. This general model is well-motivated as it may be used to abstract the question of influence maximization in social networks. Similarly to before, one can verify that our algorithm from Subsection 2.1 can be applied in this generalized model, attaining the same approximation ratio. Specifically, this case can be reduced to the bipartite case by (1) creating a source node copy s_v and a target node t_v copy in the bipartite graph for each graph node v, and (2) creating an edge (s_u,t_v) in the bipartite graph for each directed graph edge $\langle u,v\rangle$, having the same probability vector.

A tradeoff between running time and approximation. Our algorithm from Subsection 2.1, although efficient, may not be practical due to the enumeration step. Nonetheless, one can utilize it and its analysis to develop a simple algorithm that has a practical running time (essentially, the running time of the greedy completion procedure) with a somewhat worse approximation guarantee. This algorithm selects the budget allocation that achieves the maximal influence from $|S|+1$ solutions: the first solution is obtained by executing the greedy procedure on the given input instance, and the remaining $|S|$ solutions are obtained by allocating the maximal possible budget to each single source node. One can validate that this algorithm attains a $(e-1)/(2e) \approx 0.316$-approximation. Specifically, the outline of the analysis is the following: If the optimal solution has a single node whose marginal influence (with respect to the empty solution) is at least $(e-1)/(2e)\cdot\mathrm{OPT}$ then we are clearly through; otherwise, one can demonstrate that the greedy solution achieves an approximation ratio of $(1-1/e)-(e-1)/(2e)=(e-1)/(2e)$ by applying a similar reasoning to that presented in Theorem 1, while noting that $\Delta_L \leq (e-1)/(2e)\cdot\mathrm{OPT}$.

3. A TARGET-SIDE INFLUENCE MODEL

In this section, we prove that the problem of maximizing the influence in the target-side model is provably harder than in the source-side model. This is done by demonstrating that both the maximum edge biclique and the dense k-subgraph problems can be reduced to our problem. As a result, we attain that our problem is hard to approximate within a factor of $\Omega(1/B^{\epsilon})$ for some $\epsilon > 0$, assuming a certain hypothesis about the average-case hardness of random 3SAT [11]. We also establish some additional hardness results that depend on other computational complexity assumptions.

We then turn to consider a modified model in which there are no capacity constraints on the source nodes, or equivalently, one may assume that each $c_s = B$. We develop an efficient deterministic algorithm that achieves a logarithmic approximation ratio. This establishes a computational separation between both models, and proves that the budget capacity constraints underlie the hardness of the problem in the former model. We remark that this modified model is still NP-hard, and in fact, it is NP-hard to approximate to within a factor better than $1-1/e$ as it extends the maximum coverage problem [10]. The latter problem is equivalent to the case in which all the probability vectors are of the form $\langle 1,*,\ldots,*\rangle$, where each $*$ can be any valid value. In this case, there is no use for allocating a budget greater than 1 to any of the source node. Consequently, the objective reduces to selecting B source nodes in a way that maximizes the cardinality of their neighbor set.

3.1 Hardness results

A reduction from maximum edge biclique. We begin by demonstrating that our problem is as hard to approximate as the maximum edge biclique problem. As input for maximum edge biclique, we are given a bipartite graph $G=(S,T,E)$. Our goal is to find a biclique in G having a maximum number of edges. A vertex set $S'\cup T'$ such that $S'\subseteq S, T'\subseteq T$ is called a *biclique* if $(s,t)\in E$ for all $s\in S', t\in T'$. In what follows, we show a reduction to our problem from a variant of maximum edge biclique in which the cardinality k of the optimal subset of S-vertices is known in advance. Notice that the maximum edge biclique problem is polynomial-time reducible to this latter problem by enumerating over all $|S|$ possible values of k, and hence, this variant shares the same computational hardness as maximum edge biclique.

Given an input instance of the above-mentioned variant of maximum edge biclique, we construct an input instance of our problem that consists of the same bipartite graph G. Moreover, we set the capacity of each source node $s\in S$ to 1, the budget $B=k$, and the probability vector associated with each target node $t\in T$ to $\langle 0,\ldots,0,1\rangle$; here, the length of the prefix of 0's is $B-1$. Now, one can easily verify that a solution $S'\cup T'$ for maximum edge biclique with $|S'||T'| = k|T'|$ edges implies a budget allocation in the newly-created instance with an influence of $|T'|$. Specifically, the claimed influence is attained by allocating a unit of budget to each of the source nodes corresponding to S'. Conversely, it is not difficult to verify that given a valid budget allocation, one can perform a similar gap-preserving transformation in the opposite direction. In particular, notice that the subgraph induced by the source nodes that are allocated a unit budget and the target nodes that are influenced to an extent of 1 is indeed a biclique.

As a result of this gap-preserving reduction, and in conjunction with the hardness results presented by Feige [11] and Feige and Kogan [12], we attain two inapproximability

results. The first is based on a certain hypothesis about the average-case hardness of random 3SAT [11].

Theorem 2. The influence maximization problem in the target-side model is hard to approximate within a factor of $\Omega(1/B^{\epsilon})$ for some $\epsilon > 0$, assuming that there is no polynomial-time algorithm that refutes most 3CNF formulas with n variables and Δn clauses, and never wrongly refutes a satisfiable formula, where Δ is a sufficiently large constant independent of n.

The other is based on a plausible assumption that 3SAT has no subexponential algorithm [12].

Theorem 3. The influence maximization problem in the target-side model is hard to approximate within a factor of $\Omega(1/2^{(\log B)^{\delta}})$ for some $\delta > 0$, assuming that there is no algorithm for 3SAT that runs in time $2^{n^{3/4+\epsilon}}$ for some $\epsilon > 0$.

A reduction from dense *k*-subgraph. We proceed by presenting a simple reduction from the dense k-subgraph problem to our problem. An input instance for the dense k-subgraph problem [13] consists of a graph $G = (V, E)$ on n vertices, and a parameter $k \leq n$. The objective is to find a subset $V' \subseteq V$ of cardinality k that maximizes the number of edges having both endpoints in V'. Up until recently [1], the dense k-subgraph problem has "only" been shown not to admit a PTAS under various computational complexity assumptions [11, 20]. Nevertheless, it is a notorious problem that so far has resisted all attempts to provide good approximability results. In particular, the current best approximation ratio for this problem is roughly $\Omega(1/n^{1/4})$ [4]. Consequently, the following reduction proves that achieving a good approximation for our problem must result in a good approximation for the dense k-subgraph problem. For example, developing a $\Omega(1/B^{\epsilon})$-approximation algorithm for our problem would imply a $\Omega(1/k^{\epsilon})$-approximation for dense k-subgraph.

Given an input instance of dense k-subgraph, we construct an input instance of our problem that consists of a bipartite graph $G' = (S', T', E')$ such that $S' = V$, $T' = E$, and its edge set is defined as $E' = \{(v, e) : v \in S', e \in T',$ and v is one of the endpoints of e in $G\}$. Furthermore, we set the capacity of each source node to 1, the budget $B = k$, and the probability vector associated with each target node to $\langle 0, 1\rangle$. Now, one can easily verify that a solution V' for dense k-subgraph that covers ℓ edges implies a budget allocation in the newly-created instance with influence ℓ. Specifically, one should allocate a unit of budget to each of the source nodes corresponding to V'. Conversely, it is not difficult to verify that given a valid budget allocation, one can perform a similar value-preserving transformation in the opposite direction.

As a result of this value-preserving reduction, and in conjunction with the hardness results presented in [1], we attain the following inapproximability result. This result is based on a certain hypothesis regarding the hardness of the hidden clique problem.

Theorem 4. The influence maximization problem in the target-side model is hard to approximate within a factor of $\Omega(1/2^{(\log B)^{2/3}})$, assuming that there is no algorithm that runs in time $n^{o(logn)}$ and distinguished between a random graph $\mathcal{G}(n, 1/2)$ and a random graph $\mathcal{G}(n, 1/2)$ with a random clique of size $n^{1/3}$ placed in it.

3.2 An algorithm for a modified model

In light of the state of affairs presented in Subsection 3.1, we focus our attention on a modified model in which there are no capacity constraints on the source nodes, but the budget constraint B is still valid. We develop an efficient deterministic algorithm whose approximation ratio is logarithmic. We begin by introducing the maximum thresholds coverage problem. Subsequently, we show that given an instance of our problem, one can translate it to maximum thresholds coverage with multi-sets.

The maximum thresholds coverage problem. An input instance of the *maximum thresholds coverage* problem consists of a ground set of n elements $X = \{e_1, \ldots, e_n\}$, a collection of m subsets $X_1, \ldots, X_m \subseteq X$, and a budget B. Each element e_i is associated with a positive threshold value d_i, and a positive weight w_i. The objective is to select at most B of the subsets in a way that maximizes the overall weight of satisfied elements. An element e_i is *satisfied* if it appears in at least d_i of the selected subsets. When one is allowed to select multiple copies of any subset then this problem is refereed to as maximum thresholds coverage *with multi-sets*. We emphasize that the overall number of selected subset copies should be at most B in this case.

Given an input instance of our problem, we can translate it to maximum thresholds coverage with multi-sets as follows: Each target node t is translated to a collection of B elements $E_t = \{e_{t1}, \ldots, e_{tB}\}$, and each source node s is translated into a subset X_s that consists of all the elements corresponding to each of its neighboring target nodes. For example, if t is a neighbor of s in the bipartite graph then X_s consists of all the elements of E_t. The threshold of each element e_{tb} is set to b, and its weight is set to $f_t(b) - f_t(b-1)$, where $f_t(b) = 1 - \prod_{i=1}^{b}(1 - p_i^t)$ defines the probability that a target node t becomes active as a function of the overall budget b allocated to its neighbors.

One can easily validate that a solution for our problem defines a solution for maximum thresholds coverage with multi-sets whose overall weight is equal to the expected influence of the former solution. Specifically, each unit of budget assigned to a source node s translates to a selection of a copy of the subset X_s. In addition, notice that all the elements of each E_t are covered the same number of times as a result of our construction. Consequently, if the budget allocated to the neighbors of t was b then each of the elements of E_t are covered b times. Clearly, only the elements $e_{t1}, \ldots, e_{tb}$ are satisfied, and their overall weight contribution is $\sum_{i=1}^{b}(f_t(i) - f_t(i-1)) = f_t(b)$. Note that this is indeed equal to the probability that t becomes active when a budget of b is allocated to its neighbors, as required. Conversely, it is not difficult to verify that given a solution for maximum thresholds coverage with multi-sets, one can perform a similar value-preserving transformation in the opposite direction.

A logarithmic approximation algorithm. In the remainder of this subsection, we concentrate on the maximum thresholds coverage with multi-sets problem. We present a $\Omega(1/\log B)$-approximation algorithm for this problem, which implies the same performance guarantee for influence maximization in the target-side model by the above reduction. We emphasize that one may assume without loss of generality that the minimal element weight is at least 1. This assumption holds as one can normalize the weights (by di-

viding them with the minimal weight) without any consequences whatsoever. Our algorithm employs a classify-and-select approach. Specifically, it works as follows:

- We partition an input instance into $\lceil \log B \rceil$ classes based on the thresholds of the elements. More precisely, the ith class defines an input instance which is induced by the elements whose thresholds are in the range $I_i = [2^{i-1}, 2^i)$. This input instance is identical to the original instance with the modification that the collection of subsets is $X'_1, \ldots, X'_m$, where each X'_j consists only of the elements of X_j whose thresholds are in the range I_i.
- We find an approximate solution for each modified input instance. Specifically, given the input instance of class i, we consider a (weighted) coverage problem that consists of the same elements and subsets, but has a budget of $\max\{\lfloor B/2^i \rfloor, 1\}$. Note that there are no thresholds in this problem, and we are basically interested in maximizing the weight of covered elements. An element is covered if it appears in at least one selected subset. We execute the well-known greedy algorithm for coverage (see, e.g., [16]) on this problem to obtain a solution $\mathcal{S} \subseteq \{X'_1, \ldots, X'_m\}$. As a solution for class i, we return $\min\{2^i, B\}$ copies of each of the subsets in $\mathcal{S}$.
- As the solution of our algorithm, we return the solution that maximizes the overall weight of satisfied elements out of the $\lceil \log B \rceil$ solutions computed for the classes.

We next prove the main theorem of this section.

Theorem 5. The above algorithm has an approximation ratio of $\Omega(1/\log B)$.

PROOF. We prove that the solution that is computed for each class i attains a constant fraction of the weight that the optimal solution yields from the elements whose thresholds are in the range I_i. As a consequence, the above-mentioned approximation ratio follows since there is at least one class for which the optimal solution yields $\Omega(1/\log B)$ of the optimal weight. For ease of presentation, it would be convenient to assume that B is a power of 2, i.e., $B = 2^\ell$ for some $\ell \in \mathbb{N}$. We demonstrate how to neglect this assumption later on.

Consider some class i. We next argue that an optimal solution which is restricted to select either 0 or 2^i copies of any subset is only worse by some constant factor c from an optimal solution that can select any number of copies of any subset. Notice that proving this argument completes the proof of the theorem since our algorithm finds a constant factor approximation for the former setting. Specifically, The former setting is equivalent to the coverage problem considered by our algorithm, and the greedy algorithm is known to attain $(1 - 1/e)$-approximation for coverage [25, 24, 16]. Hence, our algorithm obtains a $c \cdot (1 - 1/e)$ fraction of the optimal weight for class i.

For the purpose of proving the above argument, we use a simple probabilistic argument. Let us concentrate on the optimal solution $\mathcal{S}^*$ for class i that can select any number of copies of any subset, and let E^* be the set of elements satisfied by this solution. We analyze the following approach: (1) randomly select $B/2^i$ subsets from the multiset $\mathcal{S}^*$, and (2) output a solution that consists of 2^i copies of any selected subset. Note that if several copies of a subset are selected, we still output only 2^i copies of it. Clearly, the resulting solution is restricted in the sense that either 0 or 2^i copies of any subset are selected. Furthermore, the number of copies in the solution is at most $B/2^i \cdot 2^i = B$. Finally, the expected weight of the solution is at least

$$\begin{aligned}
&\sum_{e \in E^*} \Pr\left[\begin{array}{c} X \text{ such that } e \in X \\ \text{is selected} \end{array}\right] \cdot w_e \\
&\geq \sum_{e \in E^*} \left(1 - \frac{\binom{B - 2^{i-1}}{B/2^i}}{\binom{B}{B/2^i}}\right) w_e \\
&\geq \sum_{e \in E^*} \left(1 - \left(1 - \frac{2^{i-1}}{B}\right)^{B/2^i}\right) w_e \\
&\geq \left(1 - \frac{1}{\sqrt{e}}\right) \sum_{e \in E^*} w_e \ ,
\end{aligned}$$

where the second inequality follows as $\binom{a-b}{c} / \binom{a}{c} \leq (1 - b/a)^c$. In other words, the expected weight of the solution is a constant fraction of the optimal weight of class i. As a result, the optimal solution which is restricted to select either 0 or 2^i copies of any subset is worse by at most a factor of $1 - 1/\sqrt{e} \approx 0.393$ from an optimal solution that can select any number of copies of any subset.

We now discuss how to deal with the general case that B is not a power of 2. Suppose that $\ell \in \mathbb{N}$ is the first integer such that $B < 2^\ell$. First notice that for class ℓ, our algorithm computes a greedy solution with a budget of 1 (i.e., selecting one subset whose elements have maximal weight), and then returns the solution that consists of B copies of that subset. Arguments similar to those presented in the above probabilistic argument, applied to the case in which one random subset of $\mathcal{S}^*$ is initially selected, can be used to prove that the weight of the optimal restricted solution for that class is at least 1/2 of the weight of the optimal unrestricted solution. In fact, applying the same probabilistic argument with respect to selecting just one random subset establishes that the weight of the optimal restricted solution for class $\ell - 1$ is at least 1/4 of the weight of the optimal unrestricted solution. Focusing on the remaining classes $i = 1, \ldots, \ell - 2$, one can demonstrate that applying the probabilistic argument in which we initially select $\lfloor B/2^i \rfloor$ subsets from $\mathcal{S}^*$ proves that the weight of the optimal restricted solution for class i is at least $1 - e^{-1/4} \approx 0.221$ of the weight of the optimal unrestricted solution. □

4. CONCLUDING REMARKS

We study the problem of allocating a given budget to various media channels to maximize the influence on a candidate audience. We present two models of influence, namely, a source-side and a target-side influence models. We develop an optimal approximation algorithm for the first model, while establishing that the latter model is provably harder. We also demonstrate that the budget capacity constraints underlie the computational hardness of the latter model.

We discuss several extensions of the source-side influence model in Subsection 2.3. We believe that it is worthwhile to further study those directions. We also think that it would be interesting to prove hardness results for the target-side influence model under more standard assumptions, like a P $\neq$ NP assumption. Finally, another worthy direction for

future research is to close the gap for the modified target-side influence model in which there are no capacity constraints.

5. REFERENCES

[1] N. Alon, S. Arora, R. Manokaran, D. Moshkovitz, and O. Weinstein. On the inapproximability of the densest k-subgraph problem. *Manuscript*, 2011.

[2] N. Archak, V. Mirrokni, and M. Muthukrishnan. Budget optimization for online advertising campaigns with carryover effects. In *6th Workshop on Ad Auctions*, 2010.

[3] S. Bharathi, D. Kempe, and M. Salek. Competitive influence maximization in social networks. In *Proceedings 3rd International Workshop on Internet and Network Economics*, pages 306–311, 2007.

[4] A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige, and A. Vijayaraghavan. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph. In *Proceedings 42nd ACM Symposium on Theory of Computing*, pages 201–210, 2010.

[5] S. Bhattacharya, G. Goel, S. Gollapudi, and K. Munagala. Budget constrained auctions with heterogeneous items. In *Proceedings 42nd ACM Symposium on Theory of Computing*, pages 379–388, 2010.

[6] C. Borgs, J. T. Chayes, N. Immorlica, M. Mahdian, and A. Saberi. Multi-unit auctions with budget-constrained bidders. In *Proceedings 6th ACM Conference on Electronic Commerce*, pages 44–51, 2005.

[7] N. Chen, N. Gravin, and P. Lu. On the approximability of budget feasible mechanisms. In *Proceedings 22nd Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 685–699, 2011.

[8] S. Dobzinski, R. Lavi, and N. Nisan. Multi-unit auctions with budget limits. In *Proceedings 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 260–269, 2008.

[9] P. Domingos and M. Richardson. Mining the network value of customers. In *Proceedings 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 57–66, 2001.

[10] U. Feige. A threshold of ln n for approximating set cover. *J. ACM*, 45(4):634–652, 1998.

[11] U. Feige. Relations between average case complexity and approximation complexity. In *Proceedings 34th ACM Symposium on Theory of Computing*, pages 534–543, 2002.

[12] U. Feige and S. Kogan. Hardness of approximation of the balanced complete bipartite subgraph problem. Technical report, Department of Computer Science and Applied Math., Weizmann Institute, 2004.

[13] U. Feige, D. Peleg, and G. Kortsarz. The dense k-subgraph problem. *Algorithmica*, 29(3):410–421, 2001.

[14] J. Feldman, S. Muthukrishnan, M. Pál, and C. Stein. Budget optimization in search-based advertising auctions. In *Proceedings 8th ACM Conference on Electronic Commerce*, pages 40–49, 2007.

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

[16] D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., 1997.

[17] M. Jackson and L. Yariv. Diffusion on social networks. *Economie Publique*, 16:69–82, 2005.

[18] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137–146, 2003.

[19] D. Kempe, J. M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In *Proceedings 32nd International Colloquium on Automata, Languages and Programming*, pages 1127–1138, 2005.

[20] S. Khot. Ruling out ptas for graph min-bisection, dense k-subgraph, and bipartite clique. *SIAM J. Comput.*, 36(4):1025–1071, 2006.

[21] J. Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In *Algorithmic Game Theory*. Cambridge University Press, 2007.

[22] S. Morris. Contagion. *Review of Economic Studies*, 67:57Ű78, 2000.

[23] E. Mossel and S. Roch. On the submodularity of influence in social networks. In *Proceedings 39th Annual ACM Symposium on Theory of Computing*, pages 128–134, 2007.

[24] G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Math. Operations Research*, 3(3):177–188, 1978.

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

[26] T. Schelling. *Micromotives and Macrobehavior*. Norton, 1978.

[27] H.-S. Shih and E. S. Lee. Discrete multi-level programming in a dynamic environment. In *Dynamical aspects in fuzzy decision making*. Springer, 2001.

[28] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: Modeling, analysis, and simulations. *Artif. Intell.*, 94(1-2):139–166, 1997.

[29] Y. Singer. Budget feasible mechanisms. In *Proceedings 51st Annual IEEE Symposium on Foundations of Computer Science*, pages 765–774, 2010.

[30] M. Sviridenko. A note on maximizing a submodular set function subject to a knapsack constraint. *Oper. Res. Lett.*, 32(1):41–43, 2004.

[31] H. P. Young. *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton University Press, 1998.