Price Trade-offs in Social Media Advertising

Milad Eftekhar[†], Saravanan Thirumuruganathan[‡], Gautam Das[‡], and Nick Koudas[†]
[†]Department of Computer Science, University of Toronto
Toronto, ON, Canada
[‡]Computer Science and Engineering Department, University of Texas at Arlington
Arlington, Texas, USA
milad@cs.toronto.edu, saravanan.thirumuruganathan@mavs.uta.edu,
gdas@cse.uta.edu, koudas@cs.toronto.edu

ABSTRACT

The prevalence of social media has sparked novel advertising models, vastly different from the traditional keyword based bidding model adopted by search engines. One such model is topic based advertising, popular with micro-blogging sites. Instead of bidding on keywords, the approach is based on bidding on topics, with the winning bid allowed to disseminate messages to users interested in the specific topic.

Naturally topics have varying costs depending on multiple factors (e.g., how popular or prevalent they are). Similarly users in a micro-blogging site have diverse interests. Assuming one wishes to disseminate a message to a set V of users interested in a specific topic, a question arises whether it is possible to disseminate the same message by bidding on a set of topics that collectively reach the same users in V albeit at a cheaper cost.

In this paper, we show how an alternative set of topics R with a lower cost can be identified to target (most) users in V. Two approximation algorithms are presented to address the problem with strong bounds. Theoretical analysis and extensive quantitative and qualitative experiments over real-world data sets at realistic scale containing millions of users and topics demonstrate the effectiveness of our approach.

Categories and Subject Descriptors

I.1.2 [**Computing Methodologies**]: Symbolic and algebraic manipulation, Algorithms, Analysis of algorithms; J.4 [**Computer Applications**]: Social and behavioral sciences, Economics; G.1.2 [**Mathematics of Computing**]: Numerical analysis, Approximation

Keywords

Social advertising; Topic-based advertising; Alternate topics

1. INTRODUCTION

Online advertising is a multi-billion dollar business and has attracted a lot of attention among many advertisers all over the world. Online display ads are ubiquitous (e.g. popular on prevalent sites such as CNN, BBC, Reuters, blogs, search engines' result pages, etc.). Several methods are utilized to deliver ads, the most popular approach is keyword bidding. Popular web portals and search engines have created platforms (e.g., Google AdWords) to display online ads based on a keyword bidding methodology. Typically multiple people may bid on a keyword and an auction is held for each keyword. The advertiser with the maximum bid wins the auction and its ad is shown to users who search for that keyword.

Social networks had expansive growth over the last decade. Facebook with over 1 billion users and Twitter with half a billion registered users are just two examples of successful social platforms hosting billions of messages posted every week. Users spend considerable time on social networks. Thus advertisers recently started to focus on advertising opportunities on such platforms.

Since time spent on social networks does not involve information search (keywords queries) but information production and consumption (generating posts, reading posts from social connections, and interacting with social connections), new models of advertising emerged. For example, recently Twitter introduced a new advertising platform [25] that provides advertisers several options for user targeting. One of them is to design advertising campaigns on specific topics (topic-based advertising). Utilizing this feature, an advertiser chooses a topic, places a bid value, and provides a tweet (called a "promoted tweet") to the system. If the bid is granted, the tweet provided is shown to a set of related users. In other words, the tweet is shown to a user (*appears in user's timeline*) if the chosen topic is relevant to that user. We say that these users are *targeted* by the chosen topic. Moreover, we refer to this set of users, as the *target set* of the topic. Similarly Facebook utilizes promoted stories with overall functionality related to that of promoted tweets.

Since social platforms have hundreds of millions of users, the type of topics in which these users produce or consume contents is expected to be highly diverse. In a micro-blogging platform for example, one would typically produce content on topics one knows well (maybe profess) and also consume content in topics one is interested in, by following other users who are producers of contents of such topics. Thus, if a user u is a producer (or consumer) of topics such as "soccer" and "computer science", we may target u by advertising on either "soccer" or "computer science". It is evident that there is not just a single way to target a user, but indeed, several ways exist utilizing different topics that are relevant to u.

Different topics have different costs however (exactly as different keywords have varying costs in the keyword based advertising model). Given that a user can be targeted possibly by multiple topics, an interesting question to ask is the following: Given a topic t with a target set S_t (the set of users targeted by t), is it possible to reach the same target set S_t by bidding on topics other than t in a more economical way? If that is possible and the new topics

COSN'14, October 1–2, 2014, Dublin, Ireland.

ACM 978-1-4503-3198-2/14/10 ...$15.00.
http://dx.doi.org/10.1145/2660460.2660462.

are less expensive compared to t, obviously this would be beneficial. We aim to identify a set of less expensive topics that target approximately (for a quantitatively measurable notion of approximation) the same set of users as the target set of t (i.e., they have approximately the same target sets). In doing so, we are interested to avoid targeting users outside S_t as that would not be beneficial. In particular we focus on a tight targeting model. Under this model, we aim to locate a set of topics with a target set as close as possible to t's target set. The key property is to prevent targeting users who are not in t's target set (e.g., users for whom t is not relevant). We penalize the method to avoid spamming these users. Therefore, a penalty cost (according to a *penalty cost function*) is associated with any instance of targeting a user outside t's target set. We aim to identify an alternative topic set R (obviously not including t) such that the number of users in t's target set that are targeted by at least one topic in R is maximized provided that the sum of the costs of topics in R and the sum of the penalty costs is not greater than a maximum budget.

The problem of identifying alternative topics is inspired by Twitter and Facebook advertising platforms. However, we would like to emphasize that as the details of these social advertising platforms are not known to public, the problem we discuss in this paper is a general problem and is *not* designed for or based on *any* specific social media platform including Twitter and Facebook.

Under this model, we show that if the penalty cost function is non-decreasing and convex, we identify solutions and propose algorithms with guaranteed approximation bounds.

Our techniques create a win-win situation for both advertisers and the advertising platforms. By providing more options (i.e., topics with approximately the same audience) for each advertiser to target, we prevent the situation where a single popular topic (that is very expensive) exists alongside several cheaper topics that no one bids on. Therefore by utilizing our techniques, more advertisers afford to target their desirable audience. Hence the revenue of the advertising platform may significantly increase (since more advertisers pay) while advertisers also obtain more savings per advertisement.

2. RELATED WORKS

Social based analytics: Many works have been done on micro-blogging platforms in recent years. Sankaranarayanan et al [24] use these platforms to identify breaking news as well as to consume news [17]. Micro-blogging platforms have also been used to monitor trends with novel applications such as predicting stock prices [23]. They have also been used to detect communities based on interests [14] or bursts [11] and to rank users based on their influence [27] within their community or based on their topical expertise [21]. Behavior of users on the social platforms and communities has also been studied [1, 19].

Advertising: Twitter has joined the likes of Google and Facebook to start an online advertising platform [25]. Recent research has shown that Twitter users respond favorably to advertising [6]. Broadly, existing work on social networks have studied three different types of advertising. The first is behavioral targeting [2, 29] where the aim is to show relevant advertisements based on user behavior over a given site or over a set of mutually co-ordinating sites. The second is influence based [4, 7, 18, 28] advertising. In this approach, the aim is to identify influential users whose tweets or posts serve as an endorsement influencing his/her followers to indulge in an activity. The final type of advertisement is topic based [8, 13, 16, 22, 27]. In this approach, advertisers bid on a topic and a promoted tweet is shown to users who are interested in the topic. In this paper, we focused on such an approach as it is closer to the Twitter advertising platform.

Set, Max and Budgeted Coverage Problems: From a theoretical perspective, our solutions are akin to the set cover and its variants - Max-Cover and Budgeted Set cover all of which have been proven to be NP-Complete [20]. Refer to [26] for a discussion on efficient approximation algorithms for set cover. Khuller et al., [15] proposed two approximation algorithms for the budgeted maximum coverage problem. We adopt these algorithms as a basis towards designing algorithms to address Problem 1. The online variant of set cover has been studied in [3] while [9] studied adoptions of the approximation algorithm for set cover to very large datasets. Bonchi et. al. [5] studied decompositions of a single query to a small set of queries whose result union approximates the original query result.

3. THE TARGETING PROBLEM

The online advertising platform offered by micro-blogging services enables advertisers to target users based on topics. The cost of advertising on different topics is, clearly, not the same. Some topics are costly since they are popular and attract the attention of advertisers, while some other topics are cheaper. On the other hand, a user may be targeted by many topics. If a user belongs to the target set of several topics, advertising on any of these topics will target this user.

Let U represent a set of users and T represent a set of topics. For a topic $t \in T$, let S_t represent the target set of t. The target set S_t is the set of users who are targeted by bidding on topic t. The target sets can be identified by different approaches such as user-defined lists [8, 10]. We note that the sets U, T, and the target sets S_t are inputs of the problem and can be computed by any means one prefers without changing any part of the problem and the algorithms proposed in this paper.

Let the cost of advertising on t be C_t. The cost of a topic depends on the payment method adopted. Two popular payment methods are *pay per impression* and *pay per click*. Utilizing *pay per impression* method, the advertiser pays an amount b_t (identified based on bid values) for each impression of its ad. In *pay per click* method, the advertiser pays an amount b_t when a user clicks on its ad. Assuming that the ad is shown to all users in the target set S_t, the cost of t is $C_t = |S_t| \times b_t$ in case of pay per impression technique and $C_t = |S_t| \times b_t \times f_t$ in case of pay per click technique. Here, f_t is the fraction of users in S_t who click on the ad.

Suppose one wishes to advertise on topic t with a budget of B at hand and $C_t > B$. Two cases exist. First, the cost C_t should be paid by the advertiser so that the advertisement carry on. Second, the ad of the winning bidder is shown to users in S_t till the budget is exhausted. In the former case, one cannot advertise on t as one does not have enough budget to do so. In the latter, one can target just a portion $\frac{B}{C_t}$ of the target set S_t, for both *pay per impression* and *pay per click* methods providing that users in S_t click on the ad uniformly at random.

Given that users can be targeted by multiple topics, a natural question arises. Is it possible to target more users in S_t by determining alternative topics without exceeding the budget? For example, one might conclude that when the goal is to advertise on topic "music", by choosing topic "wine" instead, we reach 70% of users in the target set of "music" while we pay half the cost of C_{music}.

By identifying these alternative topics, the advertising platform can provide the advertiser with multiple options to choose with different cost and coverage values. As explained in Section 1, providing these options is beneficial to both advertisers and the advertising platform. We note that the topics that are targeted by advertisers

are unknown to the users. In other words, a targeted user u just sees the ad (e.g., the promoted tweet) and is unaware of the topic that is utilized by the advertiser to target u. Therefore, utilizing alternative topics leads to no change in users' experience. As a further step, an advertiser may choose to target the alternative topic while creating advertisements on events related to both the main topic and the alternative topic (e.g., events on "music & wine"). Our intuition is that adopting this strategy even increases the engagement of users on the ad, hence increasing the click through rate (compared to the case where the ad just talks about music), as the ad would excite users in $S_{music} \cap S_{wine}$ (users who are interested in both music and wine).[1]

We aim to identify topics to (1) target as many users in S_t as possible and (2) avoid targeting users outside S_t. More formally, we associate a penalty cost when targeting users outside S_t (*unwanted targeting*). This penalty, that aids to avoid spamming these users, depends on the number of users targeted outside S_t, and the number of times each of these users is targeted. Let $u \notin S_t$; assume u is targeted x_u times. We denote the penalty cost as $f(x_u)$. Such cost depends on the number of times u is targeted. The goal of this cost is to capture the intuition that if a user does not belong to the target set of t, it is not supposed to be targeted for content related to t. Therefore, each time u is targeted incorrectly, we associate a penalty. This penalty increases as x_u increases. In particular we associate a penalty with a positive marginal increase (i.e., an increase following a convex trend) when the number of times a user is targeted increases. We utilize a function $f(x_u)$ that is (1) non-decreasing (the penalty cost does not decrease as a function of x_u), and (2) convex (the marginal cost does not decrease as a function of x_u). The penalty cost function captures the intuition that the penalty incurred when targeting a single user u, say, three times when $u \notin S_t$ for a topic t is higher than that of targeting three users not in S_t for a topic t only once. A non-decreasing convex function is appropriate to capture this behavior. We aim to maximize the number of users targeted in S_t with the lowest cost possible.

Problem 1. Let T be a set of topics, t be a specific topic, S_t be the target set of topic t, and B be the budget. Let $f(x_u)$ be the penalty cost for each user where x_u determines the number of times that the user u not in S_t is targeted. Identify a set $R \subseteq T - \{t\}$ to maximize

$$|S_R \cap S_t|$$

subject to $C_R + C'_R \leq B$ where $S_R = \bigcup_{r \in R} S_r$ is the union of the target set of all topics in R, $C_R = \sum_{r \in R} C_r$ is the cost of targeting all topics in R, $C'_R = \sum_{u \in S_R - S_t} f(x_u)$ is the total penalty cost, and for any user u outside S_t ($u \in S_R - S_t$), $x_u = |\{r | r \in R, u \in S_r\}|$ is the number of times u is targeted incorrectly (the number of topics in R that u belongs to their target set).

A reduction from the Set Cover problem shows that Problem 1 is NP-hard even in a very simple case where there is no penalty cost and the cost of targeting each topic is 1. The reduction is as follows. Consider a universe U, m sets $A_1, \cdots, A_m$, and an integer k. The set cover decision problem aims to determine whether there exists k sets among $A_1, \cdots, A_m$ with a union equal to U. We create an instance of Problem 1. Let t be a topic with $S_t = U$. For each A_i ($1 \leq i \leq m$), let t_i be a topic in T with $S_{t_i} = A_i$, and $C_{t_i} = 1$. Moreover, let $B = k$ and $f(x) = 0$ for any x. The answer to the set cover decision problem is *yes* if and only if $|S_R| = |S_t|$.

[1]Validation of this intuition needs psychological experiments and is out of the scope of this paper.

We present two algorithms to address Problem 1 and identify set R. Section 3.1 explains TG, a faster algorithm that provides a $1 - 1/\sqrt{e}$ approximation factor. Section 3.2 presents TG3 that provides a tighter bound of $1 - 1/e$.

3.1 The Tight Greedy algorithm (TG)

Let t be the given topic, and *coverage* of any set $A \subseteq T$ be the number of users that are targeted in set S_t when advertising on topics in A. Thus, the coverage of set A is $|S_A \cap S_t|$ where S_A is the union of the target set of all topics in A. The main idea in TG is (1) to identify a set of topics R_1 by iteratively adding the topic t' achieving the maximum ratio of marginal coverage over marginal cost ($\frac{|S_{R_1 \cup \{t'\}} \cap S_t| - |S_{R_1} \cap S_t|}{C_{t'} + C'_{R_1 \cup \{t'\}} - C'_{R_1}}$) as long as $C_{R_1 \cup \{t'\}} + C'_{R_1 \cup \{t'\}} \leq B$, (2) to identify a topic $q^* \in T$ with the maximum coverage (i.e., $|S_{q^*} \cap S_t|$) such that $C_{q^*} + C'_{q^*} \leq B$, and (3) to report the set with the maximum coverage, among R_1 and $\{q^*\}$, as the set R. The pseudo code of TG is presented as Algorithm 1.

Algorithm 1: The Tight Greedy algorithm (TG) for alternative topic set identification

Input: t: the original topic,
T: the set of topics (not including t),
U: the set of users,
$S_{t'}$: the target set of any arbitrary topic t',
$C_{t'}$: the cost of targeting any arbitrary topic t',
$C'_{t'}$: the penalty cost of any topic t',
B: budget
Output: R^*: a subset of topics

```
q* = arg max_{q∈T} |S_t ∩ S_q| s.t. C_q + C'_q ≤ B
R_1 = {}
while T is not empty do
    t* = arg max_{t'∈T} (|S_{R_1∪{t'}} ∩ S_t| − |S_{R_1} ∩ S_t|) / (C_{t'} + C'_{R_1∪{t'}} − C'_{R_1})
    if C_{R_1∪{t'}} + C'_{R_1∪{t'}} ≤ B then
        R_1 = R_1 ∪ {t*}
    T = T − {t*}
return R* = arg max_{R∈{{q*},R_1}} |S_R ∩ S_t|
```

As Algorithm 1 shows TG first identifies a set R_1 created by greedily adding the best available topic; second it identifies the topic q^* with maximum coverage; and finally it compares the coverage of these two options to identify the alternative topic set. A simpler algorithm that just identifies the set R_1 and reports it as the alternative topic set (we call it *simpleGreedy*) leads to arbitrarily bad approximation results as the following example clarifies.

Example 1. Assume the original topic is t with a target set of $S_t = \{u_1, u_2, \cdots, u_n\}$ and a very high cost. Suppose there exist two topics t_1 and t_2. Topic t_1 has a target set of $S_{t_1} = \{u_1\}$ and a cost of $C_{t_1} = 1$. Topic t_2 has a target set of $S_{t_2} = \{u_2, u_3, \cdots, u_n\}$ and a cost of $C_{t_2} = 2n$. Moreover, the budget is $B = 2n$. The *simpleGreedy* algorithm reports $\{t_1\}$ as the alternative set with a coverage of 1, while the optimal answer is $\{t_2\}$ with a coverage of $n - 1$. Thus, the approximation factor in this example is $\frac{1}{n-1}$. Clearly the approximation factor approaches 0 when n approaches infinity.

By comparing the set R_1 with the optimal topic q^*, we show that TG can lead to an approximation bound of $1 - 1/\sqrt{e}$.

THEOREM 1. *Utilizing any non-decreasing convex penalty function $f(x)$ in Problem 1 (i.e., $\frac{\partial f}{\partial x} \geq 0$ and $\frac{\partial^2 f}{\partial^2 x} \geq 0$), algorithm TG identifies an alternative topic set with an approximation factor of $1 - 1/\sqrt{e}$.*[2]

THEOREM 2. *The run time complexity of TG is $\mathcal{O}(|T|^2 \times |U|)$ where $|T|$ is the number of topics and $|U|$ is the number of users.*

PROOF. Line 1 takes $\mathcal{O}(|T| \times |U|)$ time since we measure the coverage of each topic; there are $|T|$ topics and calculating the coverage takes $\mathcal{O}(|U|)$ (note that the maximum size of a target set S can be $|U|$).

The while loop runs for $\mathcal{O}(|T|)$ iterations since in each iteration we remove exactly one topic from T and there are $|T|$ topics. In each iteration, we calculate the marginal increase in coverage and cost. This calculation takes $\mathcal{O}(|U|)$ for each topic. Hence, line 4 takes $\mathcal{O}(|T| \times |U|)$. The calculations in lines 5-7 takes $\mathcal{O}(|T|)$. Thus, The while loop in lines 3-7 takes $\mathcal{O}(|T|^2 \times |U|)$.

Overall, the run time complexity of TG is $\mathcal{O}(|T|^2 \times |U|)$. □

3.2 The Tight Greedy algorithm on a basis of 3 (TG3)

As Theorem 1 suggests the approximation bound of TG is $1 - 1/\sqrt{e}$. We can improve this bound utilizing algorithm TG3. The intuition in TG3 is to consider all sets of size 3, expand these sets greedily, and identify the set with the highest coverage. The algorithm TG3 (1) locates a subset R_1 of size not greater than 3 with maximum coverage such that $C_{R_1} + C'_{R_1} \leq B$, (2) locates sets R_2 that are created by iteratively adding topic t' achieving the maximum ratio of marginal coverage over marginal cost to any initial set of size 3 as long as the sum of the total cost and the total penalty cost does not exceed the budget B, and (3) reports the set with the highest coverage, among R_1 and all R_2 sets, as the set R. The pseudo code of TG3 is presented as Algorithm 2.

THEOREM 3. *Utilizing any non-decreasing convex penalty function $f(x)$ in Problem 1 (i.e., $\frac{\partial f}{\partial x} \geq 0$ and $\frac{\partial^2 f}{\partial^2 x} \geq 0$), algorithm TG3 results in an approximation factor of $1 - 1/e$.*

THEOREM 4. *The run time complexity of TG3 is $\mathcal{O}(|T|^5 \times |U|)$ where $|T|$ is the number of topics and $|U|$ is the number of users.*

PROOF. To identify R_1 we need to compute the coverage for any subset T with a size at most 3. There are $\mathcal{O}(|T|^3)$ subsets and for each subset it takes $\mathcal{O}(|U|)$ to compute the coverage. Hence identifying R_1 takes $\mathcal{O}(|T|^3 \times |U|)$.

To identify R_2, we need to expand all subsets of T of size 3 using the while loop. There are $\mathcal{O}(|T|^3)$ subsets. For each subset, the while loop runs for $\mathcal{O}(|T|)$ iterations. Each iteration evaluates all topics in T_{temp} that takes $\mathcal{O}(|T| \times |U|)$. Hence the second part of the algorithm (identifying R_2) takes $\mathcal{O}(|T|^3 \times |T| \times |T| \times |U|) = \mathcal{O}(|T|^5 \times |U|)$.

Therefore, the run time complexity of TG3 is $\mathcal{O}(|T|^5 \times |U|)$. □

4. EXPERIMENTS

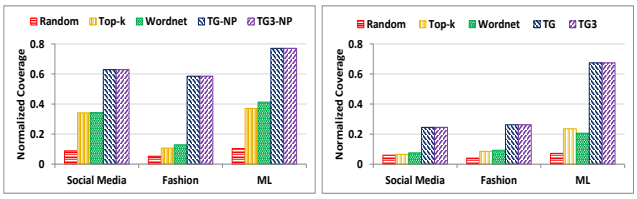
We conduct a comprehensive set of performance and quality experiments using realistic, large scale datasets derived from Twitter. We first describe our dataset in Section 4.1, followed by quantitative results on the run time, coverage, and cost of all proposed algorithms in Section 4.2; qualitative results of their output are discussed in Section 4.3.

[2]Long proofs are omitted due to space limitations. Please see our technical report for these proofs [12].

Algorithm 2: The Tight Greedy algorithm on a basis of 3 (TG3) to identify an alternative topic set

Input: t: the original topic,
T: the set of topics (not including t),
U: the set of users,
$S_{t'}$: the target set of any arbitrary topic t',
$C_{t'}$: the cost of targeting any arbitrary topic t',
$C'_{t'}$: the penalty cost of any topic t',
B: budget
Output: R: a subset of topics

1 $R_1 = \arg \max_{X \subseteq T \ \& \ |X| \leq 3 \ \& \ C_X + C'_X \leq B} |S_X \cap S_t|$
2 $R_2 = \emptyset$
3 **foreach** $X \subseteq T$ *s. t.* $|X| = 3$ *and* $C_X + C'_X \leq B$ **do**
4 $J = X$
5 $T_{temp} = T - X$
6 **while** $|T_{temp}| > 0$ **do**
7 Select $t' \in T_{temp}$ maximizing $\frac{|S_{J \cup \{t'\}} \cap S_t| - |S_J \cap S_t|}{C_{t'} + C'_{J \cup \{t'\}} - C'_J}$
8 **if** $C_{J \cup \{t'\}} + C'_{J \cup \{t'\}} \leq B$ **then**
9 $J = J \cup \{t'\}$
10 $T_{temp} = T_{temp} - \{t'\}$
11 **if** $|S_J \cap S_t| > |S_{R_2} \cap S_t|$ **then**
12 $R_2 = J$
13 **if** $|S_{R_1} \cap S_t| > |S_{R_2} \cap S_t|$ **then**
14 **return** R_1
15 **else**
16 **return** R_2



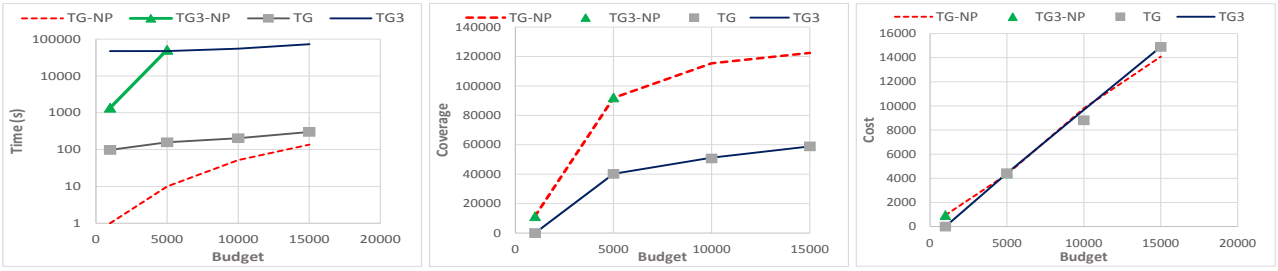
(a) Coverage with no penalty (b) Coverage with penalty

Figure 1: Comparison of our algorithms with baseline algorithms with and without penalty

4.1 Experimental Setup

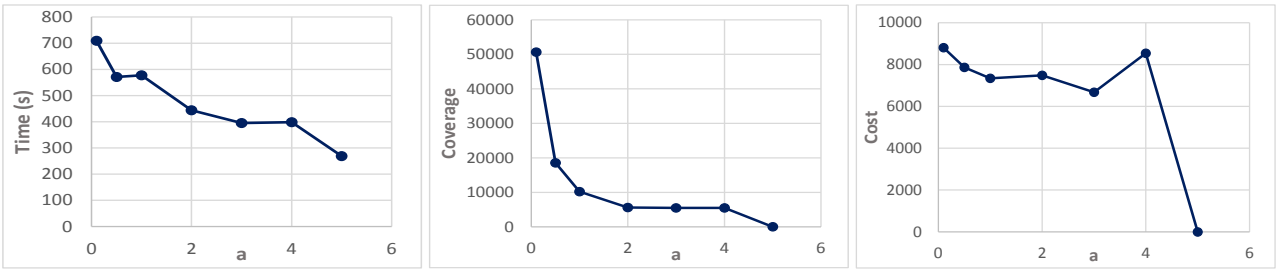
Hardware and Platform: The algorithms were coded in Java and evaluated on a quad core 2.4 GHz computer (AMD Opteron™ Processor 850) with 100 GB on memory running CentOS 5.5 with kernel version 2.6.18-194.11.1.el5. All algorithms are single-threaded.
Topics and Users : Recall that the major input to our problem is a set of topics and the target sets. Other relevant parameters include the expected bidding costs for each of the topics and a penalty function that determines the penalty cost of unwanted targeting.

For the case of Twitter, utilizing the standard APIs, we collected all Twitter lists and the users belonging to these lists. List names are adopted to identify topics [8, 10]. We collected a set of approximately 4.5 million topics and their target sets. For the case of our experiments, target sets include users that belong to these lists. Overall, the total number of users in these target sets is 150 million of which about 13.5 million accounts are distinct. On average, each user is in the target set of 11 topics.
Cost Model for Topics: While collecting users and topics was relatively straightforward, identifying the costs was not. Most com-



(a) (b) (c)

Figure 2: Impact of budget over time, cost, and coverage ($\gamma = 0.2$)



(a) (b) (c)

Figure 3: Impact of linear penalty cost over time, coverage, and bidding cost ($\gamma = 0.1$)

panies including Twitter do not reveal the bidding costs for their topics. Hence we adopt a diverse set of analytical but realistic cost models to estimate the cost of a topic. At a high level, our cost models can be partitioned into those that are independent of the target set size and those that are dependent on it.

For the former case, we generated costs for topics based on uniform and normal distributions. For the latter case, we generated costs based on a power law (size of target sets for different topics follows a power law) cost model. The rationale is that topics that are generic and have a large target set have a higher cost.

Penalty Function: We study two cases a) the penalty for any instance of unwanted targeting (covering a user outside S_t) is 0, and b) it is not. We studied three intuitive penalty functions. First a *linear* penalty function that assigns a penalty as ax where a is a constant (10 cents in our experiments except Figure 3) and x is the number of times the user was incorrectly targeted. Second is the *polynomial* cost function that assigns penalty as x^a where the parameters are as defined above. We also evaluated our algorithm on *exponential* cost functions according to a^x.

Performance Measures: There are multiple relevant metrics that could be used to evaluate our algorithms. The first is *runtime performance* which measures the time it takes to run our targeting algorithm. The second is the *coverage*, namely, how many users in the target set of t (the original query) are present in the target set of the alternative topic set R. Since our objective is to replace an expensive topic with multiple others relatively inexpensive ones, this is a crucial metric. Third is the *bidding cost* of the alternative topic set R (i.e. C_R). We would also like to reduce our *penalty cost* by minimizing the number of instances of unwanted targeting that are caused by our alternate topic set (C'_R). Experiments are performed on several original topics. The results are consistent with those presented below. In this section, due to space limitations, we report the results of the experiments done for the original topic of `social media` which has a target set of approximately 160,000 users.

Algorithms Evaluated: In this section, we evaluate two major algorithms that trade-off approximation bounds for speed: algorithms TG and TG3. In addition, both algorithms are affected by penalty function and we evaluated both scenarios with zero (TG-NP and TG3-NP where NP means no penalty) and non-zero penalty cost (TG and TG3). To speed up the algorithms, we also apply pruning techniques to remove irrelevant or costly topics. Given a topic t, the first pruning technique (we name it the *coverage-based* pruning technique) is to remove all topics with a coverage less than γ fraction of S_t. The alternative pruning technique (*ratio-based*) is to remove all topics with a coverage over cost ratio less than γ fraction of the maximum coverage over cost value among all topics. We have shown that these pruning techniques provide an approximation guarantee and a significant speedup in run time of the algorithms.[3] In all experiments, loading target sets in memory and conducting pruning require, respectively, 7 and 1 minute.

We compare these algorithms with baseline algorithms (Random, Top-k, WordNet) and demonstrate that the proposed algorithms outperform the baselines.

In all experiments except those reported in Figure 2, the budget is set to $10000.

4.2 Performance Analysis

Comparison with baseline algorithms: We start by comparing our algorithms TG and TG3 with 3 baseline algorithms:

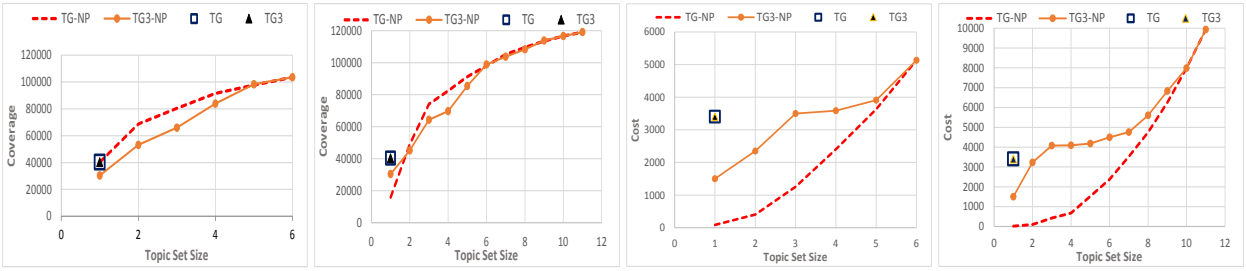
Random: Randomly pick topics until the budget is exhausted. Repeat this process for 10 times and pick the best.

Top-k: Order candidate topics based on their coverage. Pick topics in this order until the budget is exhausted.

WordNet: Given a query, do basic stemming, perform synonym expansion using Lucene-WordNet index and order results based on similarity. Pick topics in this order until the budget is exhausted.

Figure 1 reports the normalized coverage of the alternative topic sets identified by different algorithms when a coverage-based pruning technique is utilized with a pruning fraction of $\gamma = 0.5$. The normalized coverage of a topic set S with respect to a query topic t is the fraction of users in the target set of t that is targeted by S. The results for other pruning fraction values are consistent with those in Figure 1. Figure 1(a) displays the results when the penalty

[3]Please see [12] for an extensive set of experiments on the accuracy and performance of these pruning techniques.



(a) coverage - $\gamma = 0.5$ (b) coverage - $\gamma = 0.3$ (c) cost - $\gamma = 0.5$ (d) cost - $\gamma = 0.3$

Figure 4: Impact of topic set size on coverage and cost for different pruning fractions

for any unwanted targeting is zero; Figure 1(b) depicts the results adopting a linear penalty function. We observe that in both cases our algorithms TG and TG3 significantly outperform all baseline algorithms. While the baseline algorithms have normalized coverage values of 7%, 20%, and 21% in average, our algorithms result in normalized coverage values of up to 80%.

Impact of budget over time, cost, and coverage: We test how budget impacts the run time, cost, and coverage of the alternative topic set R. We decided to run experiments not taking more than a few hours. Figure 2 shows the results. As expected, as the budget increases, it is possible to afford a larger alternative topic set which in turn increases the run time, cost, and coverage. As the budget increases, the running time of the algorithms increases as they have to run additional iterations to choose more alternative topics (Figure 2(a)). The total cost also increases linearly, according to Figure 2(c), with budget increases. These changes are linear as the algorithms could utilize all the budget to cover more users. Note that since our algorithms choose topics with higher coverage to cost ratio in the first iterations, as we proceed we cover less and less new users by paying more and more, that explains the concave shape of coverage in Figure 2(b).

Impact of penalty cost over time, cost and coverage: We evaluate how the different penalty cost models affect the outcome of algorithms. We start with a *linear* penalty cost function $f(x_u) = a \times x_u$ for a non-negative constant a where x_u is the number of times user u is targeted by different topics. The results are provided in Figures 3(a)-3(c). When the cost of incorrect targeting (parameter a) increases, the algorithms become "risk-averse" and try to choose only topics that are very similar to the query topic and the size of the alternative topic set R would be smaller. This results in a drop in run time, coverage, and bidding cost C_R and an increase in penalty cost C'_R. We also evaluated our algorithms for other cost functions such as polynomial and exponential cost functions $f(x) = x^a$ and $f(x) = a^x$. We found the behavior to be similar to the linear function except the fact that the drop rate in run time, coverage, and bidding cost is much sharper.

Impact of alternative topic set size on coverage and cost: We also aim to understand how total coverage and cost changes when the algorithms add more topics in subsequent iterations to the alternative topic set R. We evaluate this experiment utilizing different pruning fractions. Figure 4 details this behavior. As we add more topics, coverage follows a concave shape while the total cost of this set increases following a *convex* behavior. This is expected since in later iterations the algorithms add topics with lower coverage to cost ratio. Further, we can observe that as the pruning fraction decreases, the size of target set increases (from a size of 6 for a pruning fraction 0.5 to a size of 11 for a pruning fraction 0.3) thereby increasing both the cost and coverage. Intuitively, a less aggressive pruning strategy results in more topics that are not necessarily cost or coverage optimal.

Table 1: Case Study of Alternate Topics (the words are stemmed)

Machine Learning	Fashion	Social Media
strata	beauti fashion	market pr
machinelearn(ing)	fashion peopl	socialmedia
ai	style fashion	communiti
info engin(e)	fashion blog	seo
ai ppl	shoe	blog
researchnew	fashion world	onlin(e) market
nosql	apparel	
nlp ml	stylist	
inform(ation) retriev(al)	fashion brand	
analytics research data dev		
data analyt(ics)		
aier		
fourtytwo		
data scientist		

4.3 Qualitative Results

In this section, we show that the output of our algorithms are quite realistic using three sample topics. For this purpose, we choose three diverse topics - `social media`, `fashion` and `machine learning`. Table 1 shows the alternate topics identified by our algorithms. We can see that the topics are intuitively similar to the original topic and expected to have users of related expertise. For example, our algorithms identify that users who produces content in topics such as `strata` (a data analysis language), `ai`, `ml`, etc. are also producing content in the topic `machine learning`. Also, topics such as `apparel`, `shoe`, `fashion blog` are good proxies if you wish to target producers in `Fashion`. Moreover, topics such as `online market`, `seo`, `blog` target similar users as `social media`.

5. CONCLUSION AND FUTURE WORKS

In this paper, we initiate a study into a targeting problem in social media advertising. We introduced a taxonomy of relevant parameters (such as cost and penalty function) and studied the feasibility of our problem for various scenarios. We show that the problem is NP-hard, present two algorithms to solve it, and prove that they provide good approximation guarantees. Finally, we conduct a comprehensive set of experiments that demonstrate the efficacy of our algorithms and the quality of the results.

As a future work, we are interested to analyze the impact of our techniques when all advertisers adopt our proposed strategy and to study costs' changes in equilibrium state. We also aim to evaluate the practicality of our approaches when real cost values are available.

6. REFERENCES

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Privacy enhancing technologies*, pages 36–58. Springer, 2006.

[2] A. Ahmed, Y. Low, M. Aly, V. Josifovski, and A. J. Smola. Scalable distributed inference of dynamic user interests for behavioral targeting. In *KDD*, pages 114–122, 2011.

[3] N. Alon, B. Awerbuch, and Y. Azar. The online set cover problem. In *STOC*, pages 100–105. ACM, 2003.

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on Twitter. WSDM, pages 65–74. ACM, 2011.

[5] F. Bonchi, C. Castillo, D. Donato, and A. Gionis. Topical query decomposition. In *SIGKDD*, pages 52–60. ACM, 2008.

[6] A. L. Brooks and C. Cheshire. Ad-itudes: Twitter users & advertising. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 63–66. ACM, 2012.

[7] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

[8] A. Cheng, N. Bansal, and N. Koudas. Peckalytics: Analyzing experts and interests on Twitter. SIGMOD Demo Track, pages 973–976. ACM, 2013.

[9] G. Cormode, H. Karloff, and A. Wirth. Set cover algorithms for very large datasets. In *CIKM*, pages 479–488. ACM, 2010.

[10] M. Eftekhar and N. Koudas. Some research opportunities on Twitter advertising. *IEEE Data Eng. Bull.*, 36(3):77–82, 2013.

[11] M. Eftekhar, N. Koudas, and Y. Ganjali. Bursty subgraphs in social networks. In *WSDM*, pages 213–222. ACM, 2013.

[12] M. Eftekhar, S. Thirumuruganathan, G. Das, and N. Koudas. Price trade-offs in social media advertising. Technical Report TR–DM–UT–14–05–00, University of Toronto, 2014. Available at: `http://www.cs.toronto.edu/~milad/paper/economical_bidding-TR.pdf`.

[13] D. Ferreira, M. Freitas, J. Rodrigues, and V. Ferreira. Twitviz-exploring Twitter network for your interests. *UMA*, pages 1–8, 2009.

[14] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: understanding microblogging usage and communities. In *WebKDD and SNA-KDD*, pages 56–65. ACM, 2007.

[15] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39 – 45, 1999.

[16] D. Kim, Y. Jo, I.-C. Moon, and A. Oh. Analysis of Twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*, 2010.

[17] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.

[18] C. Lee, H. Kwak, H. Park, and S. Moon. Finding influentials based on the temporal order of information adoption in Twitter. In *WWW*, pages 1137–1138. ACM, 2010.

[19] K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using facebook. com. *Social networks*, 30(4):330–342, 2008.

[20] R. G. Michael and D. S. Johnson. Computers and intractability: A guide to the theory of NP-completeness. *WH Freeman & Co., San Francisco*, 1979.

[21] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54. ACM, 2011.

[22] M. Pennacchiotti, F. Silvestri, H. Vahabi, and R. Venturini. Making your interests follow you on Twitter. CIKM '12, pages 165–174, New York, NY, USA, 2012. ACM.

[23] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes. Correlating financial time series with micro-blogging activity. In *WSDM*, pages 513–522. ACM, 2012.

[24] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51. ACM, 2009.

[25] Twitter. Start Advertising | Twitter for Business. `https://business.twitter.com/start-advertising`.

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

[27] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential Twitterers. In *WSDM*, pages 261–270. ACM, 2010.

[28] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on Twitter. In *WWW*, pages 705–714. ACM, 2011.

[29] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? WWW, pages 261–270, New York, NY, USA, 2009. ACM.