Price Trade-offs in Social Media Advertising

Milad Eftekhar
Department of Computer Science
University of Toronto

Saravanan Thirumuruganathan
Computer Science and Engineering Department
University of Texas at Arlington

Gautam Das
Computer Science and Engineering Department
University of Texas at Arlington

Nick Koudas
Department of Computer Science
University of Toronto

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. We propose techniques based on pruning and approximate calculations to speed up the execution of these algorithms while maintaining guaranteed approximation 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.

Keywords

Micro-blog Advertising, Topic-based targeting, 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. As users spend considerable time on social networks, naturally advertisers started in the last few years focusing 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 [30] 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 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 (Section 2). To scale the proposed algorithms to problems of arbitrary size, in Section 3.1 we propose various pruning techniques that enable them to work over much larger datasets. This is done, by pruning the datasets to a manageable size in a preprocessing step in such a way that the algorithms can be executed on the reduced data while offering approximation guarantees on the larger data sets. Section 3.2 shows that we can modify the proposed algorithms by relaxing the exact calculations to improve the performance of the proposed algorithms while still maintaining good approximation bounds.

As we have access to a large real dataset from Twitter containing about 4.5 million topics that target approximately 14 million distinct users, we evaluate all algorithms and proposed techniques on this dataset. Section 4 explains the system design. Section 5 reports our quantitative and qualitative findings demonstrating the overall efficiency and practical utility of our proposed solutions. An overview of related works is presented in Section 6, followed by Section 7 which concludes our discussion.

Our techniques create a win-win situation for both advertisers and the advertising platforms (e.g., Twitter and Facebook). 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. 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. Section 4 describes some approaches to identify the target sets in different social platforms. 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.

Assume one wishes to advertise on topic t with a budget of B at hand. Let the cost of advertising on t, denoted by C_t, be higher than the budget (i.e., $C_t > B$). In such a situation, one cannot advertise on t as one does not have enough budget to do so. Given that users can be targeted by multiple topics, a natural question arises. Is it possible to determine alternative topics to target a set of users that is as close to S_t as possible, without exceeding the budget? In particular, we aim 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. Examples of such functions follow:

- $f(x_u) = a \times x_u$ for a non-negative constant a. This linear function is utilized for the scenarios where we face a fixed penalty for any instance of targeting a user incorrectly.
- $f(x_u) = x_u^a$ for a non-negative constant a. This polynomial function is utilized when the marginal penalty cost increases when a user is targeted incorrectly multiple times.
- $f(x_u) = a^{x_u}$ for a non-negative constant a: an exponential penalty function to model scenarios where we want to apply a harsh increase in marginal penalty cost when we target a user several times incorrectly.

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. We present two algorithms to address Problem 1 and identify set R. Section 2.1 explains TG, a faster algorithm that provides a $1 - 1/\sqrt{e}$ approximation factor. Section 2.2 presents TG3 that provides a tighter bound of $1 - 1/e$.

2.1 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

$q^* = \arg\max_{q \in T} |S_t \cap S_q|$ s.t. $C_q + C'_q \leq B$
$R_1 = \{\}$
while *T is not empty* **do**
  $t^* = \arg\max_{t' \in T} \frac{|S_{R_1 \cup \{t'\}} \cap S_t| - |S_{R_1} \cap S_t|}{C_{t'} + C'_{R_1 \cup \{t'\}} - C'_{R_1}}$
  if $C_{R_1 \cup \{t'\}} + C'_{R_1 \cup \{t'\}} \leq B$ **then**
    $R_1 = R_1 \cup \{t^*\}$
  $T = T - \{t^*\}$
return $R^* = \arg\max_{R \in \{\{q^*\}, R_1\}} |S_R \cap S_t|$

As Algorithm 1 shows our approach 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 $1 - 1/\sqrt{e}$.*

PROOF. Refer to Appendix A for a complete proof. □

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 $\mathcal{O}(|T|^2 \times |U|)$. □

2.2 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 $1 - 1/e$.*

PROOF. Refer to Appendix B for a complete proof. □

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 $\mathcal{O}(|T|^5 \times |U|)$. □

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

3. SPEEDING UP THE ALGORITHMS

Theorems 2 and 4 imply that algorithms TG and TG3 may not be efficient on large datasets with millions of users and topics. In this section, we propose techniques to speedup the algorithms with guaranteed approximation bounds. Section 3.1 discusses two pruning techniques to reduce the number of topics under consideration and Section 3.2 explains how we can reduce the run time of the algorithms by performing approximate computations during the iterations of algorithms TG and TG3 instead of the exact calculations.

3.1 Pruning techniques

We first decrease the dimensionality of the problem while maintaining approximation guarantees on the quality of the alternative topic set R identified.

Real datasets contain millions of topics. Considering all of these topics is the main source of time complexity in algorithms TG and TG3. A majority of these topics however, are *unrelated* to the original topic t; i.e., there exists little overlap between the target set of these topics and the target set of the original topic t. Such unrelated topics are not going to contribute to the alternative topic set R. Thus removing them early on, can significantly decrease the dimensionality of the problem and the run time of the algorithms. Such removal however has to take place without impacting the overall quality of the final results.

We propose two techniques that significantly speed up algorithms TG and TG3 (as shown in Section 5) while maintaining approximation guarantees.

3.1.1 Coverage-based Pruning technique (CP)

In the first technique, (named CP) we prune topics based on their *coverage*. The coverage of a topic q (denoted by W_q) is the number of users, belonging to the target set of the original topic t, that q targets; i.e., $W_q = |S_q \cap S_t|$. Topics with low coverage would be withdrawn from consideration.

Algorithm 3: Coverage-based Pruning technique (CP)

```
1 W_max = max_{q∈T} W_q
2 for Topic q in T do
3     if W_q < θ × W_max then
4         Remove q from T
```

In Algorithm 3, we refer to θ as the *pruning fraction*. The value of θ defines a trade-off between the run time of the algorithm and the accuracy of the results, with a higher θ leading to more accurate outcomes.

THEOREM 5. *The CP technique introduces a* $(1-\theta B/\tilde{C}_{min})$ *approximation factor where* θ *is the pruning fraction,* B *is the budget, and* $\tilde{C}_{min}$ *is the min total cost for any topic; i.e.,* $\tilde{C}_{min} = \min_{q \in T} C_q + C'_q$. *The value of* C'_q *is the penalty cost where* q *is the only topic we choose.*

PROOF. Refer to Appendix C. □

COROLLARY 1. *Utilizing the CP technique with algorithms TG and TG3, respectively, yields a factor* $(1 - 1/\sqrt{e})(1 - \frac{\theta B}{C_{min}})$ *and a factor* $(1 - 1/e)(1 - \frac{\theta B}{C_{min}})$ *approximation algorithms.*

3.1.2 Ratio-based Pruning technique (RP)

Besides coverage, the cost of the topics also play an important role in building the final alternative topic set. In the second pruning technique (called RP), we prune the topics based on the ratio of their coverage over cost.

Algorithm 4: Ratio-based Pruning technique (RP)

```
1 r = max_{q∈T} W_q/C_q
2 for Topic q in T do
3     if W_q/C_q < rθ then
4         Remove q from T
```

The pruning fraction θ introduces a trade-off between run time and accuracy. The approximation bound of the RP technique is presented in Theorem 6.

THEOREM 6. *Let* q^* *represent the topic with the maximum ratio* $\frac{W_{q^*}}{C_{q^*}} = r$. *The RP technique is a factor* $(1 - \theta \frac{B}{C_{q^*}} \frac{W_{q^*}}{W_{max}})$ *approximation algorithm.*

PROOF. Refer to Appendix D. □

COROLLARY 2. *Utilizing the RP technique with algorithm TG and TG3, respectively, provides a factor* $(1-1/\sqrt{e})(1 - \theta r \frac{B}{W_{max}})$ *and a factor* $(1 - 1/e)(1 - \theta r \frac{B}{W_{max}})$ *approximation algorithm.*

Note that depending on the values of B, W_{max}, $\tilde{C}_{min}$, and r, we can identify a value for θ that provides the desired approximation guarantee.

3.2 Approximate calculations

Our second approach to reduce the run time of algorithms TG and TG3 is to speed up the calculations these algorithms perform. In each iteration, TG and TG3 evaluate all remaining topics to identify an optimal topic t^* with the maximum value of *marginal coverage over cost* (denoted by $\frac{MW_q}{MC_q}$ for a topic q). This optimal topic is utilized to expand the alternative topic set R. The idea is to locate a sub-optimal topic in each iteration. In particular, for a value of α ($0 < \alpha \leq 1$), we aim to locate a topic q such that $\frac{MW_q}{MC_q} \geq \alpha \frac{MW_{t^*}}{MC_{t^*}}$ where MW_q and MC_q are, respectively, the marginal coverage and the marginal cost of topic q at the current iteration.

In this section, we propose a technique to locate the sub-optimal topics. We emphasize that any other algorithm that identifies this sub-optimal topics can also be utilized.

3.2.1 APXCAL: The approximate calculation algorithm

The following lemma forms the core of APXCAL.

LEMMA 1. *Suppose* MW_q^i (MC_q^i) *is the marginal coverage (the marginal cost) of topic* $q \in T$ *at iteration* i *of algorithm TG or TG3. For any* $j > i$, $\frac{MW_q^i}{MC_q^i} \geq \frac{MW_q^j}{MC_q^j}$.

PROOF. Refer to Appendix A proof of Lemma 2. □

Let MCC represent $\frac{MW_q}{MC_q}$. Lemma 1 suggests that MCC of q does not increase after the execution of any iteration.

Thus, the MCC at iteration i ($\frac{MW_q^i}{MC_q^i}$) is an upper bound for the values at any iteration $j > i$.

APXCAL starts by creating a max-heap utilizing the topics in T (or T after the application of the pruning techniques introduced) based on their initial MCC values (i.e., $\frac{MW_q^0}{MC_q^0} = \frac{W_q}{C_q + C_q'^0}$ where $C_q'^0$ is the penalty cost when the only topic chosen is q). To identify a sub-optimal topic at each iteration, we retrieve the topic q^* at the max location of the heap (with a saved MCC value, m) and calculate its new MCC value m'. If $m' \geq \alpha m$, we select q^* as the sub-optimal topic of this iteration; otherwise, we update the MCC value of q^* and continue this process until a sub-optimal topic is identified. Note that if $m' \geq \alpha m$, q^* is a sub-optimal topic. To see this, assume at the current iteration, t^* is the optimal topic with an MCC value opt. Clearly $opt \leq m$, hence $m' \geq \alpha opt$. The pseudo code of APXCAL is presented as Algorithm 5.

Algorithm 5: APXCAL

```
1 while true do
2     q* ← max(heap)
3     m = q*.MCC
4     m' ← Calculate the new MCC of q*
5     if m' ≥ αm then
6         Remove q* from heap
7         return q*
8     q*.MCC = m' and update heap
```

THEOREM 7. *Utilizing APXCAL with TG results in an approximation factor of* $(1 - 1/\sqrt{e^{\alpha}})$.

PROOF. Refer to Appendix A. □

THEOREM 8. *Utilizing APXCAL with TG3 results in an approximation factor of* $(1 - 1/e^{\alpha})$.

PROOF. Refer to Appendix B. □

We can further improve the run time of TG3 for the case where $\alpha \leq ln(2)$ by executing it on subsets of size 2 (In Line 3 of Algorithm 2, change $|X| = 3$ to $|X| = 2$).

THEOREM 9. *Utilizing APXCAL with* $\alpha \leq ln(2)$*, a modified version of TG3 that runs on subsets of size 2 (instead of size 3) results in an approximation factor of* $(1 - 1/e^{\alpha})$.

PROOF. Refer to Appendix B. □

Note that the modified algorithm in Theorem 9 results in the same approximation factor as the original TG3 but with an improved run time complexity of $\mathcal{O}(|T|^4 \times |U|)$.

4. SYSTEM DESIGN

In the previous sections, we described efficient approximation algorithms to identify alternative topic sets for a given topic. In this section, we discuss the practical issues that would be encountered incorporating our algorithms for popular social media sites. Specifically, we discuss how to identify the set of topics T and how to identify the target set of each topic in different social platforms.

Identification of topic sets and users in a topic set for a given social media site is a fundamental operation before our algorithms could be applied. Given a topic t, there are multiple ways to define the set of users who comprise the target set t. One possible categorization with economic justification is based on the potential applications: (1) target sets consisting of *producers*, or (2) target sets consisting of *consumers*. Generally speaking, producers of a topic t are users who generate messages related to topic t, and consumers of topic t are users who follow messages related to t.

Targeting producers in advertising campaigns is popularized by influencer based marketing techniques to increase information spread and conduct viral marketing [19]. On the other hand, there are scenarios in which one aims to target consumers of a topic t directly (targeting end users instead of targeting producers). In this case, the target set of a topic would include the consumers of the topic.

There are numerous approaches to realize the target sets for each topic $t \in T$. One simple approach is to consider a popular keyword or tag in a message (e.g., hashtags in tweets) as a topic and place all users generating messages containing the keyword or the tag in the target set of the topic (if we are targeting producers), or place all users following those messages in the target set of the topic (if we are targeting consumers).

Another approach is to utilize machine learning techniques to uncover what topics each user produces or consumes. That would involve processing messages generated and followed by each user in a classification framework. One other approach is to extract information from users' profiles, groups they join, or lists they are part of. As an example in Facebook platform, many users post their interests in their profiles. Clearly each interest is a topic and all users with that interest constitute its consumer target set. In Twitter platform one may utilize Twitter lists, as will be elaborated shortly, to realize target sets. Since access to the Twitter platform was available to us, in this paper we utilize Twitter *lists* to realize target sets. We emphasize however that our entire proposal is completely orthogonal to the specific technique utilized to realize target sets. Any technique can be utilized and our framework applies equally without any modification.

While keyword based approaches are generic to any social media platform, it is possible to design more sophisticated and customized mechanisms for specific platforms such as Twitter. Twitter lists introduce a mechanism to enable user (tweet) filtering based on user-defined topics. A list is, in fact, a collection of users who share (according to the creator of the list) a common characteristic. This characteristic is typically disclosed by the list's title. For example, a user can create a list grouping twitter users "Lionel Messi", "Cris-

tiano Ronaldo", "Andres Iniesta", "Philipp Lahm", "Neymar", etc., and name the list "soccer". By restricting the tweets displayed only to accounts belonging in the list "soccer", a user can easily see information produced only by the members of the list and focus attention to tweets from the specific accounts. Essentially such a list encompasses producers (according to the list's creator) of "soccer" and is an easy way to focus attention to information generated by soccer producers. Looking at the entire collection of twitter lists, it is evident that it presents a crowd-sourced system for tagging twitter users based on the topics they produce contents in.

We utilize list names to identify topics. Starting from the name of a list, a series of textual preprocessing steps such as stemming, tokenization, stop word filtering, entity extraction, and related word grouping utilizing WordNet and Wikipedia is performed to extract a set of topics T.

For any topic $t \in T$, all users who belong to a list corresponding to topic t are identified as producers of t. Such producers constitute the target set S_t. In the case we aim to target consumers, identifying them for each topic can be determined utilizing the producers and the social connections between users. The basic premise is that if a user u follows a user v, who is known to be a producer of topic t, then u is a consumer of topic t. Thus for a topic t, we identify all producers and report their followers as the consumers of topic t. The target set S_t is the collection of these users.

Our algorithms can be adopted by the social platform itself or any third-party that has access to or can infer producers or consumers of different topics (e.g., by utilizing posts' contents, lists, groups, social connections, etc.). We emphasize that the algorithms presented in this paper accept the target sets as input. The method to identify target sets is orthogonal to our work and our algorithms work well without any modification utilizing *any* approach preferred to compute target sets.

5. 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 5.1, followed by quantitative results on the run time, coverage, and cost of all proposed algorithms in Section 5.2; qualitative results of their output are discussed in Section 5.3.

5.1 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. As described in Section 4, list names are adopted to identify topics. We collected a set of approximately 4.5 million topics and their target sets (for the case of our experiments, users that are producers of the topics as explained in Section 4). 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 companies 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.

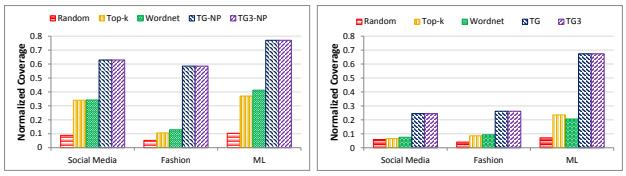
For the former case, we generated costs for topics based on uniform and normal distributions. In both cases, the average (which is representative) is 1000. We evaluated two normal distributions with low and high variance (with standard deviation being 10 and 100 respectively). Naturally, the latter results in a higher variance of costs.

An alternate approach is to model the cost in a way that is dependent on the target set size. In other words, topics that are generic and have a large target set have a higher cost. Another rationale is that in reality we expect that topics that are related will have lots of common users and hence will have similar target set size (and cost). Our last cost model captures this behavior. For this purpose we generated costs based on a power law (size of target sets for different topics follows a power law) cost model, defined as $\sqrt{target-set-size} * uniform(0,1)$.

Hereafter we refer to these cost models as *uniform*, *normal low*, *normal high*, and *power law*. In all experiments except those reported in Figure 4, the budget is set to \$10000.

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 5) 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



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

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

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). We evaluate pruning algorithms using measures introduced above. A good pruning heuristic significantly reduces the running time without incurring loss of alternate topics. 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. In all experiments except Figure 4 that evaluates the behavior of algorithms versus the budget, the budget is set to 10000. Loading target sets in memory and conducting pruning require, respectively, 7 and 1 minute.

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). We compare these algorithms with baseline algorithms (Random, Top-k, WordNet) and demonstrate that the proposed algorithms outperform the baselines. We also evaluate two pruning strategies - CP which is based on coverage while RP which is based on the ratio of coverage to cost.

5.2 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 CP pruning technique is utilized with a pruning fraction of 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 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 pruning fraction on run time, cost, and coverage: To evaluate the impact of different parameters on our proposed algorithms, we start with an experiment that demonstrates how beneficial pruning techniques are. We study how the behavior of our algorithms change, in the presence of different cost models as the pruning fraction varies. The first column of diagrams in Figure 2 (i.e., 2(a), 2(e), and 2(i)) depicts the algorithms' behavior when a *uniform* cost model is utilized. The second and third columns, respectively, relate to *normal low* and *normal high* cost models. Finally, the last column corresponds to the *power law* cost model. We decided to run experiments not taking more than a few hours (that occurs for pruning fractions above 0.3). Figure 2 represents the performance of CP with the pruning fraction varied from 0.3 to 0.5 (all algorithms follow a similar trend above 0.5); we measure the run time, coverage, and cost for all algorithms.

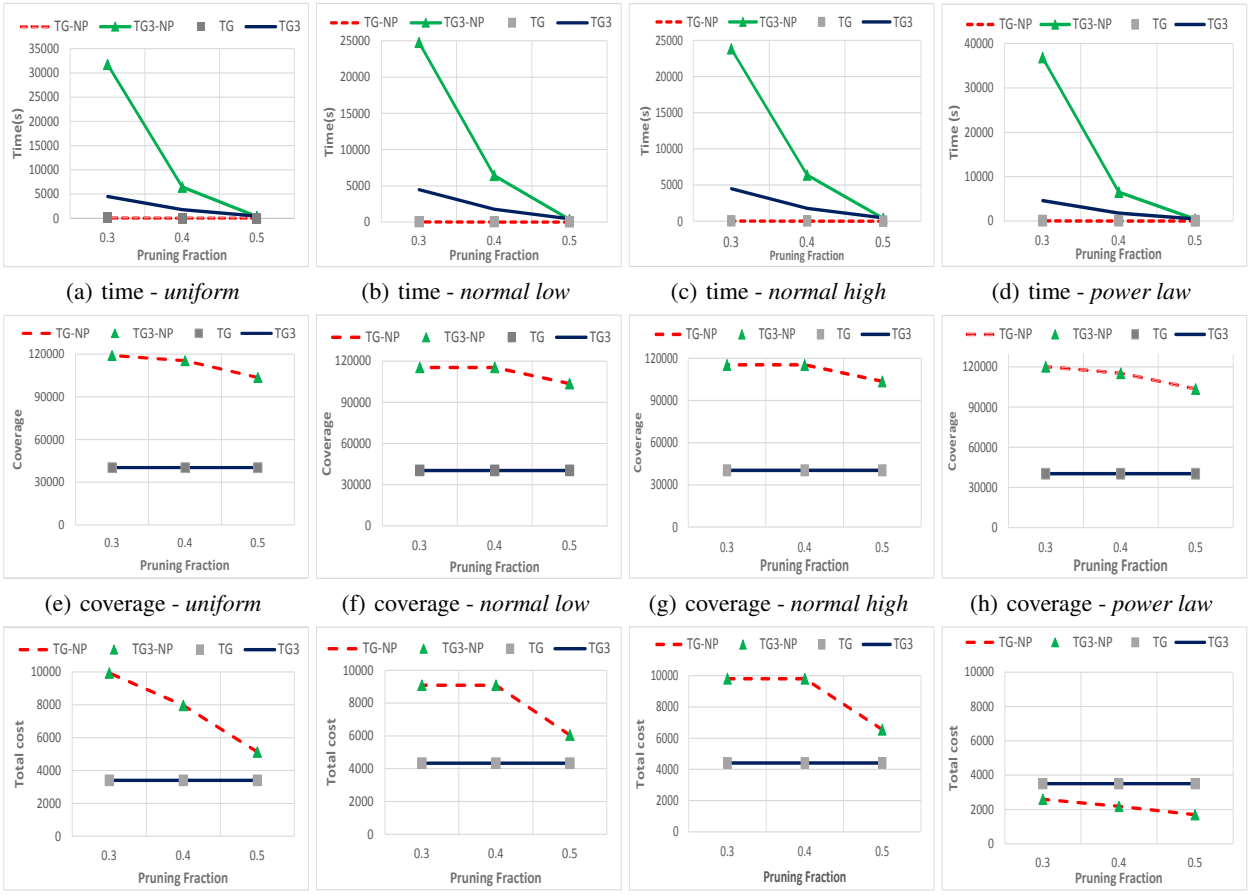
Figures 2(a) to 2(d) depict how the run time of the algorithms significantly decreases as pruning increases (from 10 hours with a pruning fraction of 0.3 to 7 minutes with a pruning fraction of 0.5). Our experiments show that algorithm TG for both scenarios - penalty of 0 and non zero - runs much faster than TG3. We also observe that this behavior is consistent across different cost models.

Figures 2(e) to 2(h) show the effect of pruning fraction to coverage. A higher pruning fraction has a dampening effect on coverage as potential alternate topics could be missed. However, the figures show the efficacy of CP in culling the irrelevant target sets as the overall coverage has only a minor drop (less than 6% on average) between pruning fractions of 0.3 and 0.5. Experiments show that when the penalty function is zero (TG-NP and TG3-NP), the coverage is higher (up to 75% of the target set of the original topic) as the algorithms could focus on identifying targets with only the budget constraint.

Figures 2(i) to 2(l) show how the total cost ($C_R + C'_R$) varies with the pruning fraction. In general, an aggressive pruning strategy actually reduces cost by removing all topics that are either irrelevant or not cost effective.

Our evaluations suggest that pruning significantly decreases the run time of the algorithms and the final cost while the coverage remains almost constant.

Comparative analysis of pruning techniques: While the previous experiments establish that pruning techniques are effective in general, CP and RP offer different trade-offs. We perform experiments utilizing all 4 different cost models that



(a) time - *uniform* (b) time - *normal low* (c) time - *normal high* (d) time - *power law*

(e) coverage - *uniform* (f) coverage - *normal low* (g) coverage - *normal high* (h) coverage - *power law*

(i) cost - *uniform* (j) cost - *normal low* (k) cost - *normal high* (l) cost - *power law*

Figure 2: Impact of pruning fraction over time, coverage, and cost on different cost models

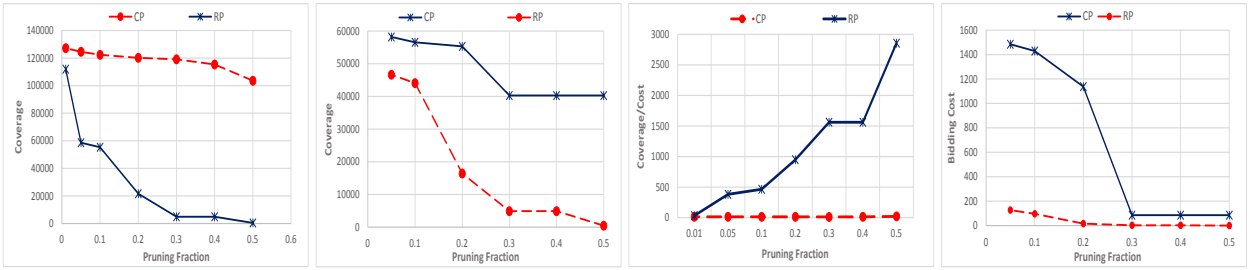
highlight and contrast the differing approaches.

CP and RP behave very similarly on *normal low* and *normal high* cost models. However, they depict differences on the *uniform* and *power law* cost models. The reason is that in the *normal* distribution, the variance of costs is smaller than *uniform* and *power law* distributions and the costs of different topics are very similar. Hence when the coverage of a topic is greater than another topic, most of the times its coverage to cost ratio, in the normal cost models, is also greater and vice versa. Therefore, both CP and RP prune similar topics, and thus the behavior of these techniques in terms of run time, coverage, and cost are similar. In the presence of *uniform* and *power law* distributions, however, as the costs associated to topics are more diverse, CP and RP show differences detailed below.

Figure 3 displays the results for the *uniform* cost model. Figure 3(a) shows how the coverage changes for different pruning fractions utilizing CP and RP for algorithm TG-NP. Figure 3(b) corresponds to the same experiment for TG. The results follow the same trend for TG3-NP and TG3. These figures display a possibly counter intuitive behavior: a wide difference between the coverage of CP and RP for high pruning fractions. For example, in Figure 3(a) when the pruning fraction is 0.5, CP has a coverage of close to 100K while RP is abysmally low. However, the relative performance improves as the pruning fraction reduces. Both techniques result in comparable high coverage values when a low pruning fraction (0.01) is utilized.

In order to explain this recall that CP is a coverage based technique that prunes topics with low coverage. In contrast, RP is a ratio based technique that drops topics that have a low coverage to cost value. When the cost of a topic is independent of size, CP removes low coverage topics even if they are quite cheap. For example, if a topic has 10 new users but only costs a cent, CP might still ignore it while RP might retain it. This behavior is exacerbated for high pruning fractions. Figure 3(c) depicts this behavior in more detail. The alternative topics that RP chose, when a high pruning fraction is utilized, have an incredibly high coverage to cost ratio of 3000 while that of CP is not significant. This high coverage to cost ratio results in significant cost savings for the advertiser. As Figure 3(d) shows RP saves money for advertisers by significantly minimizing the bidding cost.

These sets of experiments clearly show the trade-offs made by CP and RP. If the objective is to maximize the coverage then the best choice is CP. However, if the objective is to also maximize the cost-benefit ratio of the campaign, then RP is the technique of choice. While it might reach a potentially smaller audience, the cost per user reached is significantly smaller compared to CP.



(a) (b) (c) (d)

Figure 3: Comparative analysis of pruning techniques

We also evaluated our algorithms over a *power law* based cost model. The behavior of both the pruning algorithms were very similar to that of the *uniform* cost model and hence we did not include these charts to conserve space. We also observed that RP runs much faster than CP for *uniform* and *power law* cost models. For example with a pruning fraction of 0.01 on the uniform cost model, it takes about 50000 seconds to run TG-NP with CP while it takes about 5000 seconds to run it with RP, a substantial speedup.

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. Figure 4 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 4(a)). The total cost also increases linearly, according to Figure 4(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 4(b).

Impact of penalty cost over time, cost and coverage: We also 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 5(a)-5(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 6 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.

Impact of approximation ratio α on run time: Recall that one of the techniques to achieve speed up is to perform approximate calculations using algorithm *APXCAL*. We evaluate how the approximation ratio α in the APXCAL algorithm affects the run time, coverage, and cost. The value of α varies between 0 and 1 with higher values implying a higher precision. Figure 7 presents the results of an experiment when we vary α. Our experiments show that coverage and cost remain unchanged when α varies from 0.2 to 1 while the run time increases. For example, the run time of algorithm TG3-NP when $\alpha = 1$ is twice as much as the run time when $\alpha = 0.2$ (20000 seconds for $\alpha = 1$ compared to 10000 seconds for $\alpha = 0.2$). We speedup the algorithms by choosing lower values of α without sacrificing coverage.

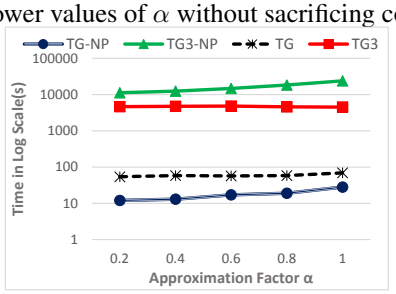
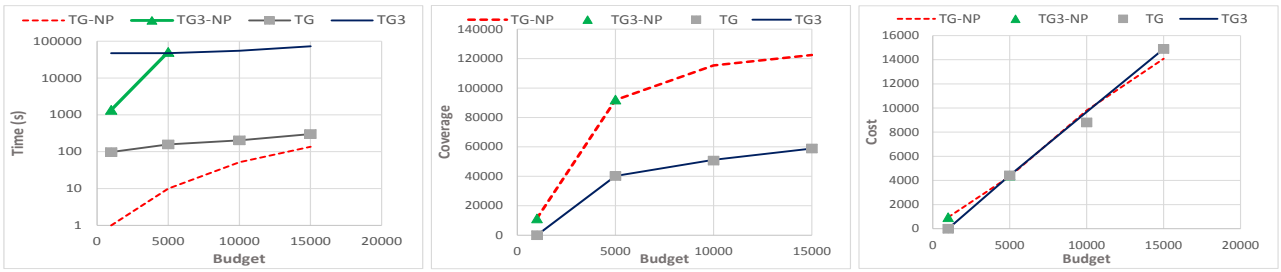


Figure 7: Impact of approximation ratio α on run time

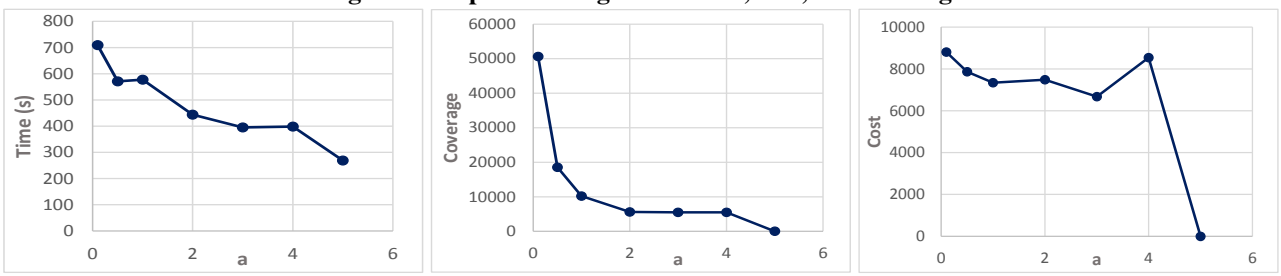
5.3 Qualitative Results

In this section, we show that the output of our algorithms are quite realistic using three sample keywords. For this purpose, we choose three diverse keywords - `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



(a) (b) (c)

Figure 4: Impact of budget over time, cost, and coverage



(a) (b) (c)

Figure 5: Impact of linear penalty cost over time, coverage, and bidding cost

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		

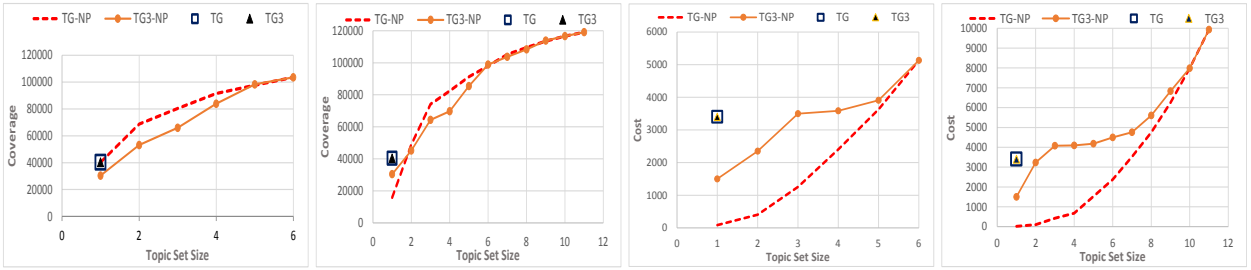
algorithms identify that users who produces content in topics such as `strata` (a data analysis language), `ai`, `ml`, etc. are also producing content in the topic `machine learning`. Also, topics such as `apparel`, `shoe`, `fashion blog` are good proxies if you wish to target producers in `Fashion`. Moreover, topics such as `online market`, `seo`, `blog` target similar users as `social media`.

6. RELATED WORKS

Social based analytics: Many works have been done on micro-blogging platforms in recent years. Sankaranarayanan et al [29] use these platforms to identify breaking news as well as to consume news [22]. Micro-blogging platforms have also been used to monitor trends with novel applications such as predicting stock prices [28]. They have also been used to detect communities based on interests [18] or bursts [15] and to rank users based on their influence [32] within their community or based on their topical expertise [26]. Behavior of users on the social platforms and communities has also been studied [1, 24].

Advertising: Twitter has joined the likes of Google and Facebook to start an online advertising platform [30]. Recent research has shown that Twitter users respond favorably to advertising [8]. Broadly, existing work on social networks have studied three different types of advertising. The first is behavioral targeting [2, 34] 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, 9, 23, 33] 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 [10, 16, 21, 27, 32]. 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 [25]. Refer to [31] for a discussion on efficient approximation algorithms for set cover. Khuller et al., [20] 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 [12] studied adoptions of the approximation algorithm for set cover to very



(a) coverage - pruning: 0.5 (b) coverage - pruning: 0.3 (c) cost - pruning: 0.5 (d) cost - pruning: 0.5

Figure 6: Impact of topic set size on coverage and cost for different pruning fractions

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.

Finding Covers Efficiently: Many important problems could be formulated as a coverage problem and hence there has been tremendous amount of prior work to identify covers efficiently by designing efficient set representations or by faster algorithms for set union and intersection operations. Efficient set representations include inverted indexes, hierarchical representations such as balanced trees, treaps and skip-lists [11]. Sets can also be represented by probabilistic data structures such as Bloom filters [6]. Using bloom filters might result in a slight reduction in covered items but combined with pruning techniques, the impact would be minimal.

Algorithms for efficient set operations operate over the preprocessed sets and can be typically categorized as adaptive [13], hashing based [6] and score based [7]. There exist a number of efficient algorithms to find covers in main memory efficiently that exploits number of properties such as asymmetry of set sizes, parallel scans, etc. Very few work tackle the problem of identifying covers in external memory, [12] being an exception. The DFG algorithm in [12] could be adapted to implement the third primitive efficiently. See [14] for an extensive related work on efficient set operations. Our algorithms are oblivious to the set representation and any of the algorithms from the related work could be used to achieve dramatic performance improvements for set union and intersections. We consider this aspect of research to be orthogonal to our work as ideas from the aforementioned papers could readily be used to increase the efficiency of our algorithms.

7. CONCLUSION

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, and propose two approximation algorithms. Further we propose two complementary pruning techniques and an algorithm to do approximate calculations to speedup the algorithms; we also studied their impact on the approximation bounds. Finally, we conduct a comprehensive set of experiments that demonstrate the efficacy of our algorithms and the quality of the results.

8. 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. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey. *Internet mathematics*, 1(4):485–509, 2004.

[7] A. Z. Broder, D. Carmel, M. Herscovici, A. Soffer, and J. Zien. Efficient query evaluation using a two-level retrieval process. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 426–434. ACM, 2003.

[8] A. L. Brooks and C. Cheshire. Ad-itudes: Twitter users and advertising. CSCW, pages 63–66. ACM, 2012.

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

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

[11] T. H. Cormen, C. E. Leiserson, R. L. Rivest, C. Stein, et al. *Introduction to algorithms*, volume 2. MIT press Cambridge, 2001.

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

[13] E. D. Demaine, A. López-Ortiz, and J. I. Munro.

Adaptive set intersections, unions, and differences. In *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*, pages 743–752. Society for Industrial and Applied Mathematics, 2000.
[14] B. Ding and A. C. König. Fast set intersection in memory. *Proceedings of the VLDB Endowment*, 4(4):255–266, 2011.
[15] M. Eftekhar, N. Koudas, and Y. Ganjali. Bursty subgraphs in social networks. In *WSDM*, pages 213–222. ACM, 2013.
[16] D. Ferreira, M. Freitas, J. Rodrigues, and V. Ferreira. Twitviz-exploring twitter network for your interests. *UMA*, pages 1–8, 2009.
[17] D. S. Hochbaum and A. Pathria. Analysis of the greedy approach in covering problems. *Naval Research Quarterly*, 45:615–627, 1998.
[18] 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.
[19] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*, pages 137–146. ACM, 2003.
[20] S. Khuller, A. Moss, and J. Naor. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39 – 45, 1999.
[21] D. Kim, Y. Jo, and I.-C. Moon. Analysis of twitter lists as a potential source for discovering latent characteristics of users. 2010.
[22] 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.
[23] 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.
[24] 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.
[25] R. G. Michael and D. S. Johnson. Computers and intractability: A guide to the theory of np-completeness. *WH Freeman & Co., San Francisco*, 1979.
[26] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM*, pages 45–54. ACM, 2011.
[27] 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.
[28] 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.
[29] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *GIS*, pages 42–51. ACM, 2009.
[30] Twitter. Start Advertising — Twitter for Business. `https://business.twitter.com/start-advertising`.
[31] V. V. Vazirani. *Approximation algorithms*. springer, 2001.
[32] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270. ACM, 2010.
[33] 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.
[34] 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.

APPENDIX

A. THE PROOF OF THE APPROXIMATION BOUND FOR ALGORITHM TG

Let W_q be the coverage of topic q, and W_A is the coverage of set A (i.e., $W_A = |(\bigcup_{q \in A} S_q) \cap S_t|$). Consider the following Lemma that holds for both algorithms TG and TG3 and is utilized in several places in this paper.

LEMMA 2. *Suppose MW_q^i (MC_q^i) is the marginal coverage (the marginal cost) of topic $q \in T$ at iteration i of algorithm TG or TG3. For any $j > i$, $\frac{MW_q^i}{MC_q^i} \geq \frac{MW_q^j}{MC_q^j}$.*

PROOF. Coverage is a submodular function that means inserting a topic q to a set A results in a bigger or an equivalent marginal increase than inserting the same topic q to a superset B of A; i.e., for any set $A \subseteq B$:

$$W_{A \cup \{q\}} - W_A \geq W_{B \cup \{q\}} - W_B \tag{1}$$

Assume R^i is the alternative topic set identified by TG (or TG3) at iteration i. Then for any $j > i$

$$R^i \subseteq R^j \tag{2}$$

Utilizing Equations 1 and 2, we conclude that

$$MW_q^i = W_{R^i \cup \{q\}} - W_{R^i} \geq W_{R^j \cup \{q\}} - W_{R^j} = MW_q^j$$

On the other hand, $MC_q^i = (C_{R^i \cup \{q\}} + C'_{R^i \cup \{q\}}) - (C_{R^i} + C'_{R^i}) = C_q + C'_{R^i \cup \{q\}} - C'_{R^i} = C_q + \sum_{u \in S_q - S_t} (f(x_u^{\textcircled{2}}) - f(x_u^{\textcircled{1}}))$

where t is the original topic, f is the penalty cost function, $x_u^{\textcircled{1}}$ is the number of times u is targeted by topics in R^i, and $x_u^{\textcircled{2}}$ is the number of times u is targeted by topics in $R^i \cup \{q\}$. Assume $x_u'^{\textcircled{1}}$ and $x_u'^{\textcircled{2}}$ are, respectively, the number of times u is targeted by topics in R^j and $R^j \cup \{q\}$. Since $R^i \subseteq R^j$:

$x_u^{①} \leq x_u'^{①}$. Since f is a non-decreasing convex function:

$$f(x_u'^{②}) - f(x_u'^{①}) \geq f(x_u^{②}) - f(x_u^{①})$$

Thus, $MC_q^i = C_q + \sum_{u \in S_q - S_t} (f(x_u^{②}) - f(x_u^{①})) \leq C_q + \sum_{u \in S_q - S_t} (f(x_u'^{①}) - f(x_u'^{①})) = MC_q^j$. Therefore, $MW_q^i \geq MW_q^j$ and $MC_q^i \leq MC_q^j$. In conclusion, for any $j > i$, $\frac{MW_q^i}{MC_q^i} \geq \frac{MW_q^j}{MC_q^j}$ □

Let O represent the optimal alternative topic set and R represent the alternative topic set identified by TG (or TG and APXCAL). Moreover let q_i be the i^{th} topic added to R by TG, R_i contain the first i topics added to R ($R_i = \{q_1, ..., q_i\}$), and MC_{q_i} be the marginal cost q_i adds, at the iteration it is added to R. Assume iteration $l+1$ is the first iteration in which a topic from O (say topic h) is considered by TG but it is not added to R because the total cost exceeds the budget. We assume that in the first l iterations, d topics are added to R. Hence after iteration l, $R = R_d$. Assume $R_{d+1} = R_d \cup \{h\}$. Lemmas 3 and 4 and the discussions following them are a generalization of the discussions on the unit cost and fixed cost versions of the problem presented in [17] and [20].

LEMMA 3. *Utilizing TG (Section 2.1) and APXCAL with an approximation parameter α (Section 3.2), for any $1 \leq i \leq d+1$, the following equation holds:*

$$W_{R_i} \geq (1 - \prod_{j=1}^{i} (1 - \frac{\alpha MC_{q_j}}{B}))W_O$$

Clearly we can set $\alpha = 1$ to consider the case that TG is utilized without APXCAL.

PROOF. First, we calculate the difference between W_O and W_{R_i} for any $1 \leq i \leq d$: $W_O - W_{R_i} \leq W_{O \cup R_i} - W_{R_i} \leq \sum_{q \in O - R_i} (W_{R_i \cup \{q\}} - W_{R_i})$.[1] Since TG (or TG and APXCAL) chooses q_{i+1} at iteration $i+1$, for all topics in $O - R_i$:

$$\alpha \frac{W_{R_i \cup \{q\}} - W_{R_i}}{C_q + C'_{R_i \cup \{q\}} - C'_{R_i}} \leq \frac{MW_{q_{i+1}}}{MC_{q_{i+1}}} \quad (3)$$

Therefore, $W_O - W_{R_i} \leq \frac{MW_{q_{i+1}}}{\alpha MC_{q_{i+1}}} \sum_{q \in O - R_i} (C_q + C'_{R_i \cup \{q\}} - C'_{R_i})$. Note that $\sum_{q \in O - R_i} (C_q + C'_{R_i \cup \{q\}} - C'_{R_i}) \leq C_O + C'_O \leq B$. Hence, $W_O - W_{R_i} \leq \frac{W_{R_{i+1}} - W_{R_i}}{\alpha MC_{q_{i+1}}} B$ that is equivalent to:

$$(W_O - W_{R_i}) \frac{\alpha MC_{q_{i+1}}}{B} \leq (W_{R_{i+1}} - W_{R_i}) \quad (4)$$

Utilizing Equation 4, we can conclude Lemma 3 by induction. The basis is straightforward. Assuming the statement for iterations R_1 to R_{i-1}, we show it for R_i. $W_{R_i} = W_{R_{i-1}} - (W_{R_i} - W_{R_{i-1}}) \geq W_{R_{i-1}} + (W_O - W_{R_{i-1}}) \frac{\alpha MC_{q_i}}{B} \geq (1 - \prod_{j=1}^{i} (1 - \frac{\alpha MC_{q_j}}{B}))W_O$. □

The proof of Theorems 1 and 7 follows. If there exists a topic q such that $W_q \geq W_O/2$, then TG reports q or some set with higher coverage hence the theorem follows. Else, for all topics the coverage is less than $W_O/2$. Consider two cases:

- $C_R + C'_R < B/2$: For all topics $q \notin R$, $C_q + C'_{\{q\} \cup R} - C'_R > B/2$; hence $O - R$ contains at most one topic q and according to our assumption $W_q < W_O/2$. Thus, $W_O - W_R < W_O/2 \Rightarrow W_R > W_O/2$.
- $C_R + C'_R \geq B/2$: Utilizing Lemma 3, we get $W_{R_d} \geq (1 - \prod_{j=1}^{d} (1 - \frac{\alpha MC_{q_j}}{B}))W_O \geq (1 - (1 - \frac{\alpha}{2d})^d)W_O \geq (1 - 1/\sqrt{e^\alpha})W_O$. Note that in these inequalities we utilize the facts that $\sum_{j=1}^{d} MC_{q_j} = C_{R_d} + C'_{R_d}$, and the fact that this expression achieves its minimum value when all MC_{q_j} values are equal.

B. THE PROOF OF THE APPROXIMATION BOUND FOR ALGORITHM TG3

Let O be the optimal solution, R be the solution identified by TG3 (TG3 and APXCAL), and R_i is the alternative topic set after iteration i. If the number of topics in O is 3 or less, TG3 identifies the optimal solution. Hence, we assume $|O| > 3$. Order the topics in O non-decreasingly according to their marginal coverage and name them $q_1, q_2, \cdots$. Let $Q = \{q_1, q_2, q_3\}$. Consider the iteration in TG3 that expands Q.

Similar to the discussions in Section A, assume h is the first topic belonging to O that is considered by TG3 but not added to the solution due to budget constraints. Moreover, assume h is evaluated in iteration $d+1$ and let $R_{d+1} = R_d \cup \{h\}$. Also let $\tilde{W}_X$ be the number of users targeted by X but not targeted by Q. The following lemma holds.

LEMMA 4. *Utilizing TG3 (Section 2.2) and APXCAL with an approximation parameter α (Section 3.2):*

$$\tilde{W}_{R_d - Q} + MW_h \geq (1 - 1/e^\alpha)\tilde{W}_{O-Q}$$

PROOF. Note that the process of expanding Q to the final solution by TG3 can be considered as an application of TG. Utilizing Lemma 3, we conclude that $W_{\tilde{R}_{d+1} - Q} \geq (1 - \prod_{j=1}^{d+1} (1 - \frac{\alpha MC_{q_j}}{B}))\tilde{W}_{O-Q} \geq (1 - (1 - \frac{\alpha}{d+1})^{d+1})\tilde{W}_{O-Q}$[2] $\geq (1 - 1/e^\alpha)\tilde{W}_{O-Q}$. Thus, $\tilde{W}_{R_{d+1}-Q} = \tilde{W}_{R_d - Q} + MW_h \geq (1 - 1/e^\alpha)\tilde{W}_{O-Q}$. □

[1] The marginal coverage of each topic q is higher at iteration $i+1$ than any later iteration, according to Lemma 2.

[2] Here we utilize the facts that $C_{R_{d+1}} + C'_{R_{d+1}} \geq B$ (recall that R_{d+1} exceeds the budget), $\sum_{j=1}^{d+1} MC_{q_j} = C_{R_{d+1}} + C'_{R_{d+1}}$, and the fact that this expression achieves its minimum value when all MC_{q_j} values are equal.

Also, $MW_h \leq MW_{q_3} \leq MW_{q_2} \leq MW_{q_1}$ (recall that the topics in O are ordered) that means $MW_h \leq 1/3W_Q$. Therefore: $W_R \geq W_Q + \tilde{W}_{R_d - Q} \geq W_Q + (1-1/e^{\alpha})\tilde{W}_{O-Q} - 1/3W_Q \geq (1 - 1/e^{\alpha})\tilde{W}_{O-Q} + (1 - 1/3)W_Q \geq (1 - 1/e^{\alpha})W_O$ thus Theorems 3 and 8 follow.

Finally we show that Theorem 9 holds. Utilizing subsets of 2 instead of 3, we get $MW_h \leq 1/2W_Q$. Hence $W_R \geq (1 - 1/e^{\alpha})\tilde{W}_{O-Q} + (1 - 1/2)W_Q$ that is greater than $(1 - 1/e^{\alpha})W_O$ if $\alpha \leq ln(2)$.

C. THE PROOF OF THE APPROXIMATION BOUND FOR THE CP PRUNING TECHNIQUE

Let T be the full set of topics and $\hat{T}$ be the set of topics after applying the CP pruning technique. Let O be the optimal alternative topic set, R be the alternative topic set identified by TG (TG3) utilizing the set T as the full topic set, and $\hat{R}$ be the alternative topic set identified by TG (TG3) utilizing the set $\hat{T}$ as the full topic set. Moreover, assume θ is the pruning fraction in the CP technique, and W_{max} is the maximum weight among all topics in T.

C.1 CP with TG

Consider the set R and $\hat{R}$. Assume iteration $i^* + 1$ is the first iteration in that the algorithm inserts different topics in R and $\hat{R}$. Let topic t^* be the topic inserted to R at iteration $i^* + 1$, and R^{i^*} be the set R or $\hat{R}$ at the end of iteration i^* (recall that both of these sets are equivalent before iteration $i^* + 1$).

LEMMA 5. *Topic t^* is in $T - \hat{T}$.*

PROOF. Assume $q^* \neq t^*$ is the topic (can be null) that is added to $\hat{R}$ at iteration $i^* + 1$. There are 2 reasons that t^* is not selected by TG to be added to $\hat{R}$. Either $t^* \notin \hat{T}$ that proves the theorem; or the marginal coverage to value for q^* is higher than t^*. However, the fact that t^* is added to R in iteration $i^* + 1$ shows that t^* has the maximum value of marginal coverage to cost among all topics in $T - R^{i^*}$. i.e., for all topics $q \in T - R^{i^*}$: $t^* = \arg \max_{q \in T - R^{i^*}} \frac{W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}}}{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}}$. This results in a contradiction. Thus, $t^* \in T - \hat{T}$. $\square$

For all topics $q \in R - \hat{R}$:

$$\frac{W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}}}{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}} \leq \frac{W_{R^{i^*} \cup \{t^*\}} - W_{R^{i^*}}}{C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}}} \quad (5)$$

Moreover since $t^* \in T - \hat{T}$ (Lemma 5), $W_{t^*} \leq W_{max}\theta$. Thus, $\frac{W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}}}{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}}(C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}}) \leq W_{R^{i^*} \cup \{t^*\}} - W_{R^{i^*}} \leq W_{t^*} \leq W_{max}\theta$. Therefore,

$$W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}} \leq W_{max}\theta \frac{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}}{C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}}}$$

Utilizing the aforementioned inequality, we compute an upper bound for the difference of the coverage between R and $\hat{R}$.

$$W_R - W_{\hat{R}} \leq \sum_{q \in R - \hat{R}} W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}},$$

$$W_R - W_{\hat{R}} \leq W_{max}\theta \sum_{q \in R - \hat{R}} \frac{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}}{C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}}}.$$

Note that $C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}} \geq C_{t^*} + C'_{t^*} \geq \tilde{C}_{min}$, $W_{max} \leq W_R$, and $\sum_{q \in R - \hat{R}} C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}} \leq B$. Thus,

$$W_R - W_{\hat{R}} \leq W_R \theta B / \tilde{C}_{min}$$

and finally

$$W_{\hat{R}} \geq W_R(1 - \theta \frac{B}{\tilde{C}_{min}})$$

Recall that $W_R \geq (1 - 1/\sqrt{e^{\alpha}})W_{OPT}$ in TG utilizing the APXCAL technique, hence the CP technique provides a $(1-1/\sqrt{e^{\alpha}})(1-\theta\frac{B}{\tilde{C}_{min}})$-approximation factor when applied together with TG and APXCAL.

C.2 CP with TG3

If O contains less than 3 topics, the difference between R and $\hat{R}$ would be less than $JW_{max}\theta$ (the worst case happens when all topics in O belong to pruned topics). Here, J is the number of topics in O that is not more than $B/\tilde{C}_{min}$. Hence $W_R - W_{\hat{R}} \leq JW_{max}\theta \leq JW_R\theta$. Thus, $W_{\hat{R}} \geq (1 - J\theta)W_R \geq (1 - B/\tilde{C}_{min}\theta)W_R$ and the theorem follows.

Assume $|O| > 3$. We order the topics in O based on their decreasing value of marginal coverage. Let $A = \{q_1, q_2, q_3\}$ be the set containing the first three topics in O. Hence $MW_{q_1} \geq MW_{q_2} \geq MW_{q_3} \geq MW_q$ for any topic $q \in O - A$. Consider the following two cases: (1) $A \subseteq \hat{T}$, and (2) $A \not\subseteq \hat{T}$.

If $A \not\subseteq \hat{T}$, then there exist a topic $p \in T - \hat{T}$ such that $p \in A$ and $W_p < \theta W_{max}$. Since A contains three sets with maximum marginal coverage, for any topic $q \in O - A$, $MW_q \leq MW_p \leq W_p \leq \theta W_{max}$. Thus, $W(O) - W(\hat{R}) \leq \sum_{q \in O - \hat{R}} MW_q \leq J\theta W_{max} \leq B/\tilde{C}_{min}\theta W_{max}$. Hence the theorem follows.

If $A \subseteq \hat{T}$, consider the iteration in TG3 that starts with A. Expanding set A to set $\hat{R}$ can be considered as an instance of utilizing TG to expand an empty set and all discussions in Appendix C.1 holds. Hence the theorem follows.

D. THE PROOF OF THE APPROXIMATION BOUND FOR THE RP PRUNING TECHNIQUE

D.1 RP with TG

Similar to the discussions in Section C, Lemma 5, and Equation 5, $t^* \in T - \hat{T}$, and for any topic $q \in R - \hat{R}$, $\frac{W_{R^{i^*} \cup \{q\}} - W_{R^{i^*}}}{C_q + C'_{R^{i^*} \cup \{q\}} - C'_{R^{i^*}}} \leq \frac{W_{R^{i^*} \cup \{t^*\}} - W_{R^{i^*}}}{C_{t^*} + C'_{R^{i^*} \cup \{t^*\}} - C'_{R^{i^*}}}$. Note that for pruned topics (topics in $T - \hat{T}$) including t^*, the ratio of initial coverage over bidding cost is less than $r\theta$. Thus, the

marginal increase in coverage over cost that each topic in $R - \hat{R}$ provides is not greater than the threshold for pruning the topics (i.e., $r\theta$). Let's assume $R - \hat{R} = \{q_1, q_2, \cdots, q_i\}$. Thus:
$W_R - W_{\hat{R}} \leq W_{q_1} + W_{q_2} + ... + W_{q_i} \leq r\theta C_{q_1} + r\theta C_{q_2} + ... + r\theta C_{q_i} \leq r\theta(C_{q_1} + ... + C_{q_i}) \leq r\theta B$

Let W_r and C_r, respectively, denote the coverage and cost of the topic with the ratio $W_r/C_r = r$. The aforementioned inequality can be written as:

$$W_R - W_{\hat{R}} \leq \theta \frac{W_{max}}{C_r} B \frac{W_r}{W_{max}}$$

Note that $W_R \geq W_{max}$. Thus,

$$W_R - W_{\hat{R}} \leq \theta \frac{W_R}{C_r} B \frac{W_r}{W_{max}}$$

Thus:

$$W_{\hat{R}} \geq W_R \times (1 - \theta \frac{B}{C_r} \frac{W_r}{W_{max}})$$

D.2 RP with TG3

We order topics in O based on their values of coverage over cost. First, assume $|O| > 3$. Let A contains the three topics with the maximum values. Consider the iteration in TG3 that starts with A. If $A \subseteq \hat{T}$, then expanding A to $\hat{R}$ is an application of utilizing TG on an empty set and a similar discussion as Section D.1 concludes the theorem.

If $A \not\subseteq \hat{T}$, we assume there exists a topic $p \in A$ such that $p \in T - \hat{T}$. Hence for any topic $q \in O - A$, $\frac{W_q}{C_q} \leq \frac{W_p}{C_p}$. Moreover as p is a pruned topic, $\frac{W_p}{C_p} < r\theta$. Thus, $W_O - W_{\hat{R}} \leq \sum_{q \in O - \hat{R}} W_q < \sum_{q \in O - \hat{R}} \theta r C_q < \theta r B$. Hence the theorem follows in both cases.

If $|Q| \leq 3$, then $W_O - W_{\hat{R}} < r\theta \sum_{q \in O} C_q < r\theta B$; the maximum difference happens when all topics in O are among the pruned topics.