Audience Expansion for Online Social Network Advertising

Haishan Liu
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
haliu@linkedin.com

David Pardoe
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
dpardoe@linkedin.com

Kun Liu
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
kliu@linkedin.com

Manoj Thakur
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
mthakur@linkedin.com

Frank Cao
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
fcao@linkedin.com

Chongzhe Li
LinkedIn Corporation
605 W. Maude Avenue
Sunnyvale, CA 94085
ckli@linkedin.com

ABSTRACT

Online social network advertising platforms, such as that provided by LinkedIn, generally allow marketers to specify targeting options so that their ads appear to a desired demographic. Audience Expansion is a technique developed at LinkedIn to simplify targeting and identify new audiences with similar attributes to the original target audience. We developed two methods to achieve Audience Expansion: campaign-agnostic expansion and campaign-aware expansion. In this paper, we describe the details of these methods, present in-depth analysis of their trade-offs, and demonstrate a hybrid strategy that possesses the combined strength of both methods. Through large scale online experiments, we show the effectiveness of the proposed approach, and as a result, the benefits it brings to the whole marketplace including both LinkedIn and advertisers. The achieved benefits can be characterized as: 1) simplified targeting process and increased reach for advertisers, and 2) better utilization of LinkedIn's ads inventory and higher and more efficient market participation.

Keywords

Online advertising; Audience expansion; Lookalike modeling

1. INTRODUCTION

Advertising on social network platforms such as LinkedIn involves interactions between three groups. Users browse and interact with content on the website; in the process they express their intention and preferences. Advertisers seek to show ads to users who are valuable to their marketing campaigns, which usually consist of a set of ads with the goal of promoting products or services. The social network provides "real estate" for placing ads, and matches the ads and the user so that they are relevant to each other.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD '16, August 13 - 17, 2016, San Francisco, CA, USA

© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ISBN 978-1-4503-4232-2/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2939672.2939680

There are two main categories of channels in which ads appear on social network platforms. Sponsored content appears in users' newsfeeds, a sequence of social news items which has become the center of gravity for users' daily traffic. Text or display ads appear in ad slots available on many other parts of the website, such as the top of the page or right column. For both channels, advertisers are able to target specific audiences they want to reach. Unlike the well known sponsored search model for advertising on search engines, where advertisers target a list of keywords that they deem relevant to their campaign, advertisers on social networks are given comprehensive demographic targeting options to precisely define the desired audience. For example, on LinkedIn an advertiser can reach all software engineers having the Machine Learning and Java skills who work in a company located in the US with fewer than 500 employees. However, even with this pinpoint targeting ability, a similar challenge to the coverage problem in search advertising remains, as advertisers usually cannot cover all relevant demographic attributes related to the desired product or service. Indeed, the cardinality of most targeting attributes is well beyond tens of thouands (e.g., titles and skills), and some even rise into multiple millions (e.g., company and group). This makes it costly for advertisers to identify the attributes they wish to target with their campaigns.

To mitigate this problem, we have developed a system, named *Audience Expansion*, to help advertisers on LinkedIn increase the reach of their campaigns. This is achieved by automatically enlarging the original target audience (the exact audience) to include similar, like-minded users. For example, if a campaign targets members with the skill "Online Advertising," the campaign might also be expanded to members who list the skill "Interactive Marketing" on their profiles. This means advertisers can reach the desired target audience with less effort setting up campaigns.

Some advertising platforms have provided similar functionality; however, most commonly the expansion is conversion-oriented. That is, advertisers must provide a list of known users who have prior positive feedback, and this list is then expanded. This requirement limits the usability of the feature to big, savy advertisers who can afford the effort of continuous marketing research and monitoring. Instead, we decided to make LinkedIn's Audience Expansion feature extrememly accessible, as easy as simply checking an option to

opt-in at the time advertisers specify the targeting criteria.

Audience Expansion aims to optimize for the joint benefit of advertisers and LinkedIn with minimal impact to the user experience. It benefits advertisers and LinkedIn in the following ways:

- Advertisers – by finding, with only limited advertiser effort, a large audience segment that is likely to engage with advertiser campaigns.
- LinkedIn – by encouraging higher marketplace participation, thus incresing competition and even providing ads for users who were previously untargeted.

In this paper, we propose an efficient system to achieve the above goals. Our main contributions are as follows:

- We present a campaign-agnostic approach to expand user profiles. The expansion is achieved by employing the *Similar-X* recommender system built in-house to generate enriched user profiles with attributes similar to those in a user's original profile. This mechanism is member-oriented, and is thus immediately available for new campaigns with no time required for generating expanded audiences.
- We develop a campaign-aware approach to expand the exact target audience of a campaign to lookalike members. The expansion is based on a nearest neighbor approach where the similarity measure is calculated by *Similar-Profiles*, a custom algorithm in the Similar-*X* family.
- We present a hybrid system that combines the strength of both the campaign-agnostic and campaign-aware approaches.

This system is now in production for all traffic and has provided significant improvements in key metrics for both advertisers and LinkedIn, such as double-digit percentage increases in impressions served across LinkedIn, and triple-digit percentage increases in value achieved for those advertisers who take advantage of Audience Expansion.

The remainder of this paper is laid out as follows. We describe related work and background in Section 2. Next, Section 3 explains the general architecture of our system and its various components. In Section 4 we describe the modeling aspect of the Similar-*X* recommender systems and their application in both the campaign-agnostic and campaign-aware expansion approaches. Section 5 details our empirical evaluation over large real-world experiments. Finally, Section 6 concludes the paper and discusses possible future directions.

2. BACKGROUND AND RELATED WORK

Previous work on audience expansion for online advertising has typically fallen into one of two categories: *broad match* in keyword advertising, and *look-alike modeling* in user-targeted advertising. Keyword advertising includes sponsored search advertising, in which ads are placed alongside search results for a relevant query, and contextual display advertising, in which advertisements are placed on a webpage with relevant content. In each case, an advertiser targets its ads by specifying a set of keywords for which it would like to bid. If a keyword occurs in the apropriate context (the search query or the page content), the advertiser then enters an auction to show its ad. The keyword must match exactly, or possibly with minor modification (e.g., different spelling or verb tense). As a result, it can be difficult for an advertiser to enumerate the entire set of keywords for which it would like to bid. In response, many advertising platforms offer broad match, in which each keyword provided by the advertiser is expanded to a larger set of keywords. For example, the keyword "bike repair" might expand to related keywords such as "bicycle repair" and "where to fix my bike".

There are several documented implementations of broad match. Often, these involve doing offline processing and storing the result in a lookup table to be used online. Broder et al. [4] address the issue of broad matches for uncommon "tail" queries that are not present in the table. A system of extracting features from a query is presented; features for common queries are extracted and stored in an inverted index, and at run time the common queries most similar to a given query can be identified using this index. Gupta et al. [9] present a system that performs supervised learning on past ad impressions to learn click probabilities for potential broad matches. The online learning approach described uses a form of max-margin voted perceptron with time decay to rapidly adjust to changes in user and advertiser behavior.

While broad match helps advertisers with the problem of choosing keywords, it does introduce new challenges relating to bid optimization and mechanism design. Even Dar et al. [8] explore the problem of optimal sponsored search bidding in the presence of broad match, where the bid for a single keyword may be applied to many queries of varying value to the advertiser. Amaldoss et al. [1] perform a game-theoretic analysis of broad match for search using a model that incorporates advertisers' bidding cost (i.e., effort expended choosing bids) and identifies the conditions under which advertisers and search engines benefit. Dhangwatnotai [7] studies advertiser welfare under the generalized second price (GSP) auction when broad matches are introduced, while Chen et al. [6] describe a probabilistic mechanism for multi-slot auctions with broad matches that generates larger welfare than the standard GSP.

In user-targeted advertising, advertisers target specific users based on properties such as demographics or interests. These properties may be explicitly specified by the user, or they may be inferred from the user's behavior (this is referred to as *behavioral targeting* [12][15]). As with keyword advertising, enumerating all possible target segments can be difficult for an advertiser. In addition, ideal target segments may contain too few users, while other segments may be too broad. In response, some advertising platforms offer look-alike modeling, which identifies users similar to a given user set. Look-alike modeling can also be used in cases where an advertiser can provide a precise list of users to target, such as in retargeting of website visitors.

Approaches to look-alike modeling include k-means clustering [13] and frequent pattern mining [3]. Mangalampalli et al. [11] show that associative classification (a rule-based form of frequent pattern mining) can be more effective than other common classifiers when training look-alike models for campaigns with few conversions. Bagherjeiran et al. [2] consider the problem of look-alike modeling when both a targeting segment and advertiser conversion data are available. Here expansion is posed as an ensemble learning problem

where the goal is to complement the existing segment while minimizing overlap.

Another approach that is related to look-alike modeling is collaborative filtering. A key challenge in applying collaborative filtering to advertising is the extreme sparsity of interaction between users and campaigns. Kanagal et al. [10] address this challenge by using a product taxonomy to identify relationships between campaigns.

Perhaps the most similar method of audience expansion to the one presented in this paper is described in [14]. As at LinkedIn, advertisers can specify a targeting segment as a propositional logic formula based on user properties. This segment is then expanded by finding a new formula describing an expanded set of users. The formula is chosen based on three criteria: 1) similarity, defined as the proportion of the original segment in the expanded segment; 2) novelty, defined as the proportion of the expanded segment not in the original segment; and 3) quality, defined by a performance metric such as CTR or conversion rate.

3. SYSTEM

In this section we give a high-level overview of LinkedIn ads serving and the Audience Expansion system. The next section then provides more details of the proposed methodology from a modeling perspective. Figure 1 shows a diagram of the system overview of Audience Expansion in the ads serving flow.

3.1 Overview

LinkedIn offers multiple ad formats, including *Text Ads* that may appear at the top or side of the page, and *Sponsored Updates* that appear as native content in a user's feed. As ad targeting works the same regardless of format, our Audience Expansion system does not make a distinction between formats. For each format there may be multiple ad slots per page request; for instance, as a user scrolls through their feed they may see several Sponsored Updates.

When an advertiser creates a campaign, they specify the ad format, the ad content (also known as the *creative*), a daily and/or lifetime budget, a bid, and the targeting to use. Bids may either be per thousand impressions (cost per mille or CPM) or per click (CPC). To specify targeting, the advertiser is given choices within a number of categories, such as location, age, company name, and skills. Within each category, the advertiser is presented with a set of standardized choices, and can select options to include and exclude. The included selections in each category are ORed together, and everything is then ANDed together, producing a targeting string that represents a logical formula in conjunctive normal form. For example, a targeting string might be:

(*location* == "USA" OR *location* == "Canada") AND (*location* != "California") AND (*age* == "18-24" OR *age* == "25-34") AND (*seniority* != "unpaid") AND (*seniority* != "training")

Once the targeting has been specified, the estimated size of the audience and a suggested bid are shown to assist the advertiser. They are also given an option to enable Audience Expansion, and may change this setting at any time.

The ads serving flow starts when a member visits a LinkedIn webpage with available ad slots. Together with the page view, an ad request is issued to the backend. The member's profile attributes are fetched, and then matched with the targeting criteria of active ad campaigns to find those that target this member. The matched campaigns then compete in a generalized second price auction, where their predicted CTR and bid jointly determine a rank order, and each campaign's cost is determined by the next-ranked campaign. The winning creatives are sent to the frontend to serve. This workflow is illustrated in the *online processes* (colored in dark grey) in Figure 1.

There are two places where Audience Expansion affects ad targeting in the above workflow. First, when fetching a user's profile attributes, Audience Expansion injects a set of expanded attributes in the return, effectively making the user targetable by more ad campaigns. Second, Audience Expansion explicitly adds a set of campaigns—those not directly targeted at this user, but with an audience that looks alike—to the eligible campaigns of the user. These two different applications correspond to the two frameworks we developed for Audience Expansion. The first kind of expansion is only based upon user attributes, hence we call it *campaign-agnostic*. The second kind of expansion applies *lookalike* modeling on top of campaigns' original target audience, hence we call it *campaign-aware*. Both of these frameworks make use of batch processing carried out in an offline fashion, as illustrated in the *offline processes* (colored in white) in Figure 1. We note that CTR prediction and auction ranking apply as usual to expanded campaigns, and in fact the CTR prediction model uses a feature indicating whether a campaign matches as a result of Audience Expansion. CTR prediction can thus act as a safeguard against showing low-relevance ads due to expansion.

3.2 Campaign-Agnostic Expansion

In the campaign-agnostic framework, profile attribute expansion is achieved by applying LinkedIn's Similar-X algorithm to a set of targetable profile attributes. The "X" refers to any entity of interest, such as company, LinkedIn Group, skill, job title, etc., making Similar-X a collective term for a group of related entity-to-entity recommender systems. For example, from Similar-Skills, we know "Data Mining" is similar to "Big Data" and "Machine Learning." Applying this in profile attribute expansion, now anyone with an explicit "Data Mining" skill will be eligible to see ads targeting "Big Data" and "Machine Learning." The heart of Similar-X is a logistic regression model that takes as input features representing involved entities, and outputs a score aimed at capturing the similarity between the entities learned from historical interaction data, preferably in the ads context.

Campaign-agnostic expansion provides an always-available expansion mechanism that does not require warm-up from the time a campaign is created or expansion is enabled. The downside is the relatively coarse quality of expansion carried out on a per-attribute basis. The campaign-aware expansion described in the next section takes the whole member profile into consideration and is hence more precise. However, the tradeoff is that it takes some time for the offline process to generate the expansion, so the expansion is not available in the short period after a campaign becomes active.

3.3 Campaign-Aware Expansion

In the campaign-aware framework, we view targeting as labeling users; i.e., an advertiser's campaign emits labels for users belonging to the desired audience. We can then pose Audience Expansion as a classification problem to pre-

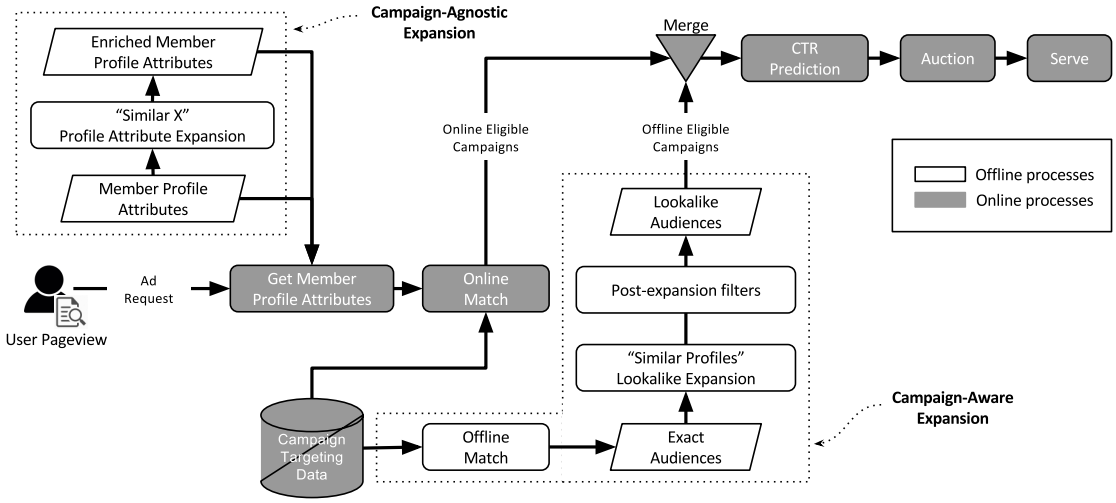

Figure 1: System overview of lookalike expansion in the ads serving workflow.

dict whether or not an untargeted user should be included in the audience for that campaign. The difficulty lies in choosing positive and negative training examples. Labeling targeted users as positive and non-targeted users as negative would simply cause us to learn a model of the original targeting. We could wait until the campaign has run for a while, then collect labels based on users' interactions. However, this takes a prolonged warm-up time for model training, and may require some careful setup to effectively explore-exploit the search space. Therefore we decided to employ a nonparametric method and reduce the problem to nearest neighbor search. This not only simplifies the label gathering process, but also avoids any assumptions that the exact targeted audience is drawn from any distribution or mixture of distributions.

Central to nearest neighbor search is the definition of a similarity measure. We adopt the Similar-Profiles algorithm, a member of the Similar-X family, to model user similarity and retrieve similar users for a given one. Details of the Similar-Profiles algorithm are discussed in Section 4. The workflow of Audience Expansion with Similar-Profiles is depicted in the box labeled "Campaign-Aware Expansion" in Figure 1. The key components of the workflow are:

1. Offline Match: A Hadoop job that calculates exact target audiences for all active campaigns that have enabled expansion by applying their targeting criteria in batch mode; the output is the campaigns' exact audiences.

2. Filter campaigns: Removal of campaigns that are unlikely to benefit from lookalike modeling, such as campaigns that target very broadly or that can easily spend their budget without expansion.

3. Expand audiences to similar profiles: The core step in the expansion, where the exact audiences are expanded to similar users, with the help of the Similar-Profiles algorithm (described in detail below).

4. Filter expansion: Additional filters on top of the expanded audience, such as those to make the new audience comply to campaigns' negative targeting criteria, or to restrict expansion on certain attributes (e.g., location and gender).

The output of the above processes is the lookalike audiences for selected campaigns, grouped by member ID in order to merge easily with the online process, which is pushed to a key-value store accessible by the services in production.

3.4 Hybrid Expansion

As mentioned earlier, campaign-agnostic expansion provides a more readily available mechanism but is relatively less precise. The campaign-aware expansion requires time to process offline but is much finer grained due to the utilization of all profile information. Due to the complementary design of these expansion methods, we can create a hybrid method that includes both the campaign-agnostic and campaign-aware methods, combining the strengths and offsetting the weaknesses of each.

The synergy in the hybrid method is achieved by the fact that both campaign-agnostic and campaign-aware expansion frameworks generate outputs in the same member-oriented fashion. Specifically, the output of campaign-agnostic expansion is an enriched member profile, $U_p = \{A_1^*, A_2^*, \ldots\}$, where A_i^* is an expanded profile attribute. The output of campaign-aware expansion is a list of predicted campaigns, $U_l = \{C_1, C_2, \ldots\}$, where each C_i is a campaign that does not target the member, but should. The enriched profile is used in the online target matching process, the output of which is then directly merged with the list of predicted campaigns. (See Figure 1 for the merge process.) This design provides a straightforward way to conduct A/B tests by selectively making either U_p or U_l available in the ads serving flow. If both U_p and U_l are made available and utilized, this effectively achieves the hybrid method for expansion.

4. MODELING

We now present details of the Similar-X models with their application in both the campaign-agnostic expansion and the campaign-aware expansion. We discuss important factors that need to be considered to bring together the various model components in the practical system.

4.1 Similar-*X* Framework

We model entity similarity through a content-based filtering approach. By treating each entity as a document to be compared against a collection of other entities, we developed a search-based system to find the best-matching entities.

Denote as $\mathcal{X}$ the value space for a given entity type. The problem can be defined as a mapping $\mathcal{X} \rightarrow \mathcal{P}_\kappa(\mathcal{X})$, where $\mathcal{P}_\kappa(\mathcal{X})$ denotes the powerset of $\mathcal{X}$ of cardinality less than κ. Specifically, the Similar-X framework outputs a list of κ target entities $t_1, t_2, \ldots, t_\kappa$ rank-ordered by the similarity scores with regard to the source entity.

4.1.1 Feature Extraction

In order to obtain an expressive representation, we model each entity as a structured (multi-fielded) document. We extract four types of fields for each entity, including n-grams / phrases, standardized named data types (e.g., industry), derived data types, and network proximities.

Consider the problem of finding similar companies as an example, where the entities of interest are companies. N-gram/phrase fields are extracted from free text contents of the company meta data, such as description or headline. N-grams are stemmed and a small set of stop words are dropped. Phrases are identified using a dictionary. Standardized named data types are extracted by dedicated classifiers for various predefined types, such as industry type and company size. Derived data types for a company may include additional attributes we infer about the company based on the commonality of its employees or followers. For example, an Internet company may have derived skills such as Network Development or Software Engineering because of the prevalence of those skills among its employees or followers. Network proximities include other related companies determined through various user-company interactions. For example, people who viewed/followed/worked for this company also viewed/followed/worked for other companies.

Table 1 shows an example of the features that may be extracted for LinkedIn as a company.

4.1.2 The Model

The fielded document representation of an entity is treated as a query and run against an inverted index of documents generated offline to retrieve similar documents. We employ the vector space model (VSM) to represent fields in documents and queries as weighted vectors in a multi-dimensional space, where each distinct term is a dimension, and we use tf-idf values as the term weights. The VSM similarity between a field f_s in the source entity (query) and a field f_t in the target entity (document) is the Cosine Similarity

$$s(f_s, f_t) = \frac{V(f_s) \cdot V(f_t)}{|V(f_s)||V(f_t)|},$$

where $V(\cdot)$ denotes the VSM representatoin of a field, and will be omitted where there is no confusion.

Note that f_s and f_t can be different fields from the two entities. For example, comparing the past job title from a source member to the current job title of a target member does tell us something useful about the similarity between these two members. In general, it is permissible to compare two fields as long as the terms of the fields (values in the vector space) are of the same type. Specifically, denote as $\mathcal{F}$ the field space of an entity. We can categorize fields based on the type of their term values, which can be either texts (mostly found in n-gram/phrase fields), or IDs (mostly found in standardized, derived, and proximity fields). Denote textual fields as $\mathcal{T}$ and ID fields as $\mathcal{I}$. It follows that $\mathcal{F} = \mathcal{T} \cup \mathcal{I}$. Denote as $G = \{\mathcal{F}_s, \mathcal{F}_t, \mathcal{E}\}$ the bipartite graph between a source and a target entity, in which an edge $(f_s, f_t) \in \mathcal{E}$ if and only if both f_s and f_t are of the same term types, i.e.:

$$\mathcal{E} = \{(f_s, f_t) : (f_s, f_t) \in \mathcal{T} \times \mathcal{T} \vee (f_s, f_t) \in \mathcal{I} \times \mathcal{I}\}\ .$$

We use $\mathbf{s} = \{s(f_s, f_t) : \forall (f_s, f_t) \in \mathcal{E}\}$ to denote the field based similarities when $s(\cdot)$ is applied to all edges in $\mathcal{E}$.

Given the field similarities, it is natural to characterize the final entity similarity so that it matches the following intuitions: 1) two entities are similar if there are a large number of similar fields; and 2) different fields contribute differently to the final entity similarity. It is then natural to define the entity similarity as a weighted linear combination:

$$S(s, t) = \mathbf{w}^T \mathbf{s}\,. \tag{1}$$

The coefficients $\mathbf{w}$ can be learned from the historical user-entity interaction log. For example, we take pairs of companies that have been historically co-targeted frequently in ads as positive examples, and companies frequently ignored when recommended to advertisers for inclusion given their existing company targetings as negative examples. We then fit a logistic regression model with elastic net regularization to the training data.

4.1.3 Personalization

One potential downside of the naive application of Similar-X attribute expansion is the lack of personalization. For example, if company A is deemed to be similar to company B from Similar-Companies, then A is added to the enriched profile of users who work at B, effectively making everyone from B eligible to see ads targeted at A, which may not be optimal for either users or advertisers. Therefore, we introduce a personalization scheme to rerank each potentially expandabe entity with regard to a given user. We achieve the personalization by employing a learned propensity model to score user-entity pairs. For each user and each entity type $x \in X$, we select the top k_x results from the available Similar-X results.

Taking companies as an example, to build the user-company propensity model we first extract features for users and companies in the same way as described in Section 4.1.1. We gather training examples from historical user-company intractions, for example, a user following a company as a positive example and un-following as negative. A logistic regression model is then trained from these examples.

4.1.4 Similar-Profiles

Similar-Profiles is the pinnacle of the Similar-X family of algorithms, with a problem size the square of the number of LinkedIn users (more than 400 million and growing).

Given the large size of the problem, we employ a Locality Sensitive Hashing (LSH) technique, named Arcos [5], to assist in finding members with high cosine similarity. Each member is mapped to one of 2^n clusters, where n is chosen to make our nearest neighbor search manageable. This cluster is built into the member index; this speeds up the subsequent nearest neighbor search because we can restrict our search to members in the same cluster. A member's cluster is specified by n bits, where each bit is determined by the output of a particular hash function. To

Type	Field	Term Values
n-gram/phrase	headline	Internet, professional, Social Network
	description	connection, productive, Talent Solution
standardized	industry	Internet
	type	public company
	company size	5001-10,000 employees
derived	skills	Software Engineering, Management, Marketing
	interests	professional identity, jobs, software development
proximities	view-browsemap	Facebook, Twitter, Pinterest
	occupation-browsemap	Google, Yahoo, Facebook

Table 1: Example features extracted for LinkedIn as a company

obtain each hash function we first choose a random vector $\mathbf{r} \in \mathbb{R}^{|\mathcal{F}|}$ with each component drawn from a Gaussian distribution $\mathcal{N}(0,1)$. The hash function corresponding to the vector $\mathbf{r}$ is defined as $h_{\mathbf{r}}(\mathbf{u}) = sign(\mathbf{r} \cdot \mathbf{u})$, which effectively partitions the space into two half-spaces by a randomly chosen hyperplane. The probability of collision is $Pr_{\mathbf{r}}[h_{\mathbf{r}}(\mathbf{u}) = h_{\mathbf{r}}(\mathbf{v})] = 1 - \theta(\mathbf{u}, \mathbf{v})/\pi$, where θ measures the angle between two vectors. It can be shown that $1 - \theta/\pi$ is closely related to the function $\cos(\theta)$. Thus, members with high cosine similarity are likely to be assigned to the same cluster.

4.2 A Note on Implementation

As mentioned earlier, the computation of Similar-X and member-entity propensity is carried out by treating a source entity as a query and performing nearest neighbor search against an index of target documents. The search takes a model configuration that specifies the coefficients ($\mathbf{w}$ in Equation 1) used in the entity similarity function. We implement the feature extraction, indexing, and search process on Hadoop. Indexing is based on Lucene with many features built in house, such as real-time indexing, faceting, etc.

For campaign-agnostic expansion, we use faceted search to achieve personalized Similar-X reranking based on member-entity propensity score, in which case the member profile in the VSM representation acts as a query against the entity index, with the pool of entities to rank being a facet in the search.

For campaign-aware expansion, we first search, for each member in the original targeted audience, for the top k_p similar members using Similar-Profiles. The found members are then considered candidates to be included in the campaign's expanded audience. We set the number of expanded campaigns for each member to be less than a threshold to prevent over-competition on a single member and to control scalability (longer lists incur non-trivial inter-process communication cost). To achieve this we derive a heuristical member-campaign fitness F between a member m and a campaign c, as shown in Equation 2.

$$F(m,c) = \frac{\sum_{m' \in T(c)} S(m,m')}{\sqrt{|T(c)|}}, \tag{2}$$

where $T(c)$ denotes the original targeted audience for the campaign c, and $S(m,m')$ is the Similar-Profiles score (see Equation 1). The intuition for this equation is that the more similar members there are in the targeted audience to an untargeted member, the more fit the untargeted member is to be included in the expanded audience. To make the fitness comparable across campaigns we have to account for the individual campaign characteristics. The numerator in the summation is positively correlated with the size of the targeted audience, hence we normalize away the size in the denominator. We empirically find that adding a square root damping to make the normalization penalize large audiences less works the best in practice.

4.3 Other Considerations

4.3.1 Campaign Selection

Not all campaigns would benefit from Audience Expansion equally. For example, if an advertiser has only a limited budget (e.g., only enough to buy a few clicks in a day), it makes sense to have them prioritize only auctions involving the exact audience, since the value of this audience is explicitly known. As another example, if an advertiser targets a very broad audience (e.g., everyone in the US), then expanding this audience is likely unnecessary, and there may not even be a sensible expansion. Therefore we devise a campaign selection strategy for Audience Expansion in order to identify a subset of the opt-in campaigns that would benefit the most from Audience Expansion. The strategy involves a series of heuristic rules evaluating the suitability of Audience Expansion for a campaign. These rules are checked each time we perform the offline workflow, and for any campaigns that do not satisfy the rules, we do not perform Audience Expansion. In particular, these rules select campaigns that meet the following intuitive requirements:

- They must target a relatively specific audience using fine-grained attributes (e.g., not simply everyone with the same location, gender, or language).
- The size of the audience should not be extremely large, as measured by both the number of matching members and the number of requests matched each day.
- The budget should be enough to afford additional impressions from the expanded audience.

The logic for evaluating the budget is as follows. Each campaign may optionally set a daily budget, and once the budget is spent the campaign no longer participates in the remaining auctions that day. In addition, to prevent campaigns from spending their entire budget early in the day, the ad serving system implements a *pacing* algorithm that randomly removes the campaign from some auctions if it is predicted that the campaign is on pace to exceed its budget. Thus, in a given period a campaign may match M requests, but only participate in $N < M$ auctions due to pacing or having spent its budget. For a campaign that has not yet enabled Audience Expansion, if the ratio of the paced request $r = N/M$ is significantly less than 1 over the past week we

can conclude that the campaign has no trouble spending its budget using its original targeting and would not benefit from expansion. For a campaign that has enabled expansion, however, we need to be more careful. We divide M into M_{exact} and $M_{expanded}$, the number of requests matched due to exact and expanded matching, respectively. We then check if $r = N/M_{exact}$ is significantly less than 1 over the past week. The reason we ignore expanded matches is that if a campaign is only able to hit its budget due to expansion, we want to continue to perform expansion, rather than oscillate between turning expansion on and off.

4.3.2 Post-Expansion Filters

After the expanded audience is generated offline, we also apply additional filtering to make sure we do not include any audience that is negatively targeted by the advertiser. We also guarantee certain targeting attributes are not violated in the expanded audience, such as location and gender. The rationale is that attributes such as location and gender are highly specific to a marketing campaign, and if they are specified, we view them as stronger and more exclusive preferences than other targeting attributes (e.g., skill or title).

In addition, as an extra measure to protect advertisers from increased costs due to Audience Expansion, we remove the most expensive members from the expanded audiences. Empirically, the more information we know about a user, the more targetable the user is, resulting in higher competition and cost. Given this intuition, we fit a regression model to predict the bid distribution of a user based on his/her targeting attributes, where the response variable is the log bid, and the explanatory variables are the member's targeting attributes in X:

$$Y = \log Bid = X\beta + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2).$$

After we fit the linear regression with coefficient β^*, we estimate the mean log bid μ for member i as $\mu_i = x_i\beta^*$. Given this, we can rank order an expanded audience by μ and remove the top p% most expensive members.

4.3.3 Online/Offline Sync

Campaign-aware expansion is an offline process that runs multiple times per day and takes a snapshot of our production database (including campaign definitions) as input. If an advertiser updates a campaign's targeting after the snapshot is taken, the expanded audience generated may not be appropriate for that campaign, at least until the next time the expansion process is run. To mitigate the potential descrepancy, we use a mechanism to achieve online/offline synchronization. Specifically, the offline process generates a timestamp along with each output record indicating the time when the input snapshot was taken. Meanwhile, the online serving application maintains a cache of campaign change timestamps. Whenever serving receives an offline-generated member-campaign pair for expansion, it checks the two timestamps, and invalidates the offline result if the offline timestamp predates the online.

5. EXPERIMENTS

5.1 Parameter Setting

Given the many components and algorithms involved in the Audience Expansion system, there are a large number of hyper-parameters to tune. We list a summary of the major parameters in Table 2.

Component	Algorithm	Parameter	Note
campaign-agnostic	personalized Similar-X	$\{k_x : \forall x \in X\}$	top-k_x elements to include from each Similar-X
campaign-aware	Similar-Profiles	k_p	top-k_p elements to include from Similar-Profiles
	campaign selection	t	threshold size of campaign exact audience
		r	ratio of paced requests
	post-expansion filter	p	percentage of most expensive members to remove

Table 2: Summary of parameters in the Audience Expansion system.

For some parameters there is a tradeoff between short term gain and long term platform health. For example, with a larger $\{k_x\}$ we would obtain a larger expansion rate, resulting in increased auction competition, but with potential harm to user experience and advertiser ROI due to the increased possibility of including irrelevant results. Therefore, we conducted many pilot A/B tests on small fractions of members to determine the hyper-parameters, carefully balancing the positive and negative outcomes produced by the system. For example, we arrived at a conservative choice of $\{k_x\}$ for Similar-X algorithms in campaign-agnostic expansion, whereas we chose a larger k_s for Similar-Profiles in campaign-aware expansion. Similarly, we tuned the post-expansion filter threshold p until the increase in campaigns' CPC due to expansion was held to an acceptable level.

5.2 Similar-Profiles in Campaign-Aware Expansion

To evaluate the performance of Similar-Profiles in campaign-aware expansion, we want to compare the quality of expanded audiences with the original targeted audiences. To accurately predict how an advertiser would value individual members in the expanded audience is a challenging task. Instead, we come up with a metric based on connection density to measure how uniform an audience is. The intuition is that the more similar two members are to one another, the more likely they are to connect on LinkedIn. Advertisers are usually interested in members with shared characteristics, therefore the targeted audiences generally have a higher-than-average connection density. If we treat the connection graph for a given audience as an undirected simple graph, then the density is defined as:

$$D = \frac{2|C|}{|M|(|M|-1)},$$

where $|C|$ is the number of connections, and $|M|$ is the number of members in the audience.

In this experiment, we randomly sampled 200 campaigns with campaign-aware expansion enabled. We are interested in two quantities for each campaign: first, the expansion ratio, i.e., the size of the expanded audience over that of the exact audience; second, the density ratio, i.e., the density of the expanded audience over that of the exact audience.

For comparison we also carry out a baseline test by expanding campaigns to a random subset of all members who have actively intracted with ads in a one-month window. We call this method *Active Clickers*-based expansion. We make sure for each campaign that the size of the expanded audience generated by Active Clickers (AC-audience hereafter

for simplicity) is roughly the same as that by Similar-Profiles (SP-audience).

To learn the distributions of density ratios for these two methods and how they vary with regard to the expansion ratio, we plot the density ratio against the expansion ratio in Figure 2. From the figure we can see that the SP-audience is on average as dense as the corresponding exact audience (density ratio close to 0 on the log scale). On the other hand, the AC-audience is much more loosely connected. Interestingly, when the expansion ratio is low the SP-audience is even denser than the exact audience.

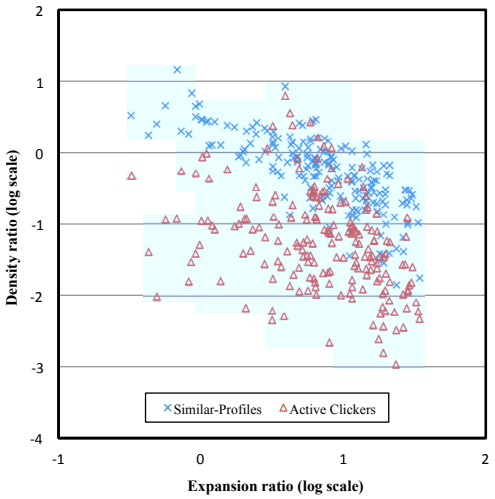

Figure 2: Network density ratio against expansion ratio for Similar-Profiles and Active Clickers.

5.3 Online A/B Test

To test the performance of our Audience Expansion implementations, we ran an experiment for two weeks on live traffic. Traffic was split by user ID across four treatments: control (no expansion, 5% of traffic), campaign-agnostic (5%), campaign-aware (10%), and hybrid (70%). We allocated a large share of traffic to the hybrid treatment because it had performed well in previous smaller tests. The remaining 10% of traffic was allocated to other treatments not described in this paper. We evaluated the results by comparing the following metrics:

- **Impressions:** As previously described, for each request we generate a ranked list of ads to show the user and fill the available ad slots in this order, generating impressions. If too few advertisers target the user, we may not always be able to generate an impression at every opportunity. Thus, we are interested in the number of impressions generated per request, and we hope that expansion would increase this number.
- **Reaches:** We define a reach as a (user, advertiser, date) tuple where the user saw an ad from the advertiser on that date. If the user sees multiple ads from the advertiser on that date, it is unly counted as one reach. By expanding advertisers' targeting, we hope to increase the number of reaches, meaning the average user sees ads from more distinct advertisers and the average advertiser shows ads to more distinct users.
- **Matches:** For each request, we record the number of advertisers whose targeting matches. Clearly, Audience Expansion should increase this number. Having more ads to choose from for a request means that we can hopefully show more relevant ads to the user.
- **Revenue:** Audience Expansion can increase revenue through three means. First, there are impressions where the winning advertiser targeted the user due to audience expansion and where there would othersiwe have been no impression. These new impressions are the ones responsible for increasing the *impressions* metric. Second, there are impressions where the winner targeted the user due to audience expansion and displaced another advertiser by bidding higher. Third, as this is a second-price auction, even if an advertiser targeting the user due to audience expansion does not win the auction, it may place second and increase the cost paid by the winner.
- **Value:** While revenue is naturally important to LinkedIn, a better measure of how well we are meeting the needs of advertisers is their total value (also known as social welfare). We make the assumption, common to second-price auctions, that a bidder's value is equal to its bid. By summing the bid for each impression (counting CPC bids only when the ad is clicked), we obtain the total value. Delivering more value to advertisers is the ultimate goal of audience expansion.
- **CTR (Clickthrough Rate):** Audience Expansion aims to find new users who will be as interested in an advertisement as the originally targeted users. Thus, we hope to see little if any decrease in CTR due to expansion.
- **CPC (Cost per Click):** An advertiser would naturally prefer not to pay more per click when its audience is expanded.
- **Dwell Time:** In addition to clicking at a similar rate, we hope that expanded users will also be similarly interested in the advertiser's landing page. Dwell time is our estimate of the amount of time a user spends on the landing page (capped at 5 minutes), and we hope to see the average dwell time remain steady when audience expansion is applied.

Table 3 shows the relative change in each metric for each expansion method when compared to the control. Due to the proprietary nature of this data, we do not report absolute numbers. Numbers in bold represent statistically significant differences ($p < 0.05$) from the control. (We use a chi-square test of equal proportions for CTR and a Wilcoxon rank sum test for the other metrics. We exclude reaches as it involves counting globally distinct tuples, and thus does not have a clear meaning at the per-request level.)

As expected, all expansion methods increase impressions, reaches, matches, revenue, and value. Campaign-aware expansion produces much larger increases than campaign-agnostic expansion, while combining the two provides some increase beyond campaign-aware expansion alone. It is reassuring that revenue and value show increases of similar magnitude, since this implies that the additional spending by advertisers who enable expansion offers a similar return on investment.

Treatment	Impressions	Reaches	Matches	Revenue	Value	CTR	CPC	Dwell Time
campaign-agnostic	**+1.31%**	+1.35%	**+1.27%**	**+3.08%**	**+3.09%**	**+1.07%**	**+0.67%**	+1.30%
campaign-aware	**+9.76%**	+9.78%	**+11.54%**	**+15.49%**	**+13.98%**	**-1.15%**	**+6.45%**	+0.36%
hybrid	**+10.36%**	+10.40%	**+12.86%**	**+17.47%**	**+15.84%**	-0.44%	**+6.92%**	+0.11%

Table 3: Relative changes in metrics compared to control. Statistically significant differences are in bold.

Treatment	Impressions	Reaches	Revenue	Value	Adj. CTR	Adj. CPC	Adj. Dwell Time
campaign-agnostic	**+7.67%**	+7.77%	**+9.33%**	**+9.72%**	-0.79%	+0.46%	+2.95%
campaign-aware	**+93.85%**	+94.29%	**+105.97%**	**+100.80%**	**-8.56%**	**+6.00%**	-0.39%
hybrid	**+96.97%**	+97.47%	**+111.60%**	**+106.00%**	**-8.08%**	**+6.86%**	-0.44%

Table 4: Relative changes in metrics when only considering impressions won by expansion-enabled campaigns.

CTR and dwell time show only small changes, as hoped, providing evidence that expansion does not hurt the relevance of the advertisements we are showing to users. CPC does increase, but further analysis shows the increase is roughly the same for advertisers who do and do not enable expansion. As a result, we can conclude that the increase is a result of more competitive auctions, and not simply because expansion causes advertisers to target more expensive users.

In addition to looking at global metrics, we can specifically examine how campaigns that enable expansion are affected by computing the same metrics using only the impressions of these campaigns. Table 4 shows the relative changes in metrics (compared to the control) when considering only the impressions of campaigns that did not disable expansion at any time during the two weeks of the experiment. Again, statistically significant changes are shown in bold. Here we use the Mantel-Haenszel adjustment for CTR, CPC, and dwell time metrics, using campaigns as strata. These metrics can differ greatly from one campaign to another, and this adjustment corrects for the different distributions over campaigns in each treatment to give a clearer picture of the change a typical campaign would see. Statistical significance for these metrics is also measured with the appropriate stratified test (Mantel-Haenszel for CTR and stratified Wilcoxon for CPC and dwell time). We omit the matches metric as it is less meaningful when considering a subset of impressions.

From Table 4 we can see that impressions for expansion-enabled campaigns nearly double under the campaign-aware or hybrid treatment, with similarly large increases for reaches, revenue, and value. As before, the change in value roughly keeps up with the change in revenue. Dwell time again shows no statistically significant change. As discussed previously, the increase in CPC for expansion essentially matches the global increase due to increased competition. The relative drop in CTR for the campaign-aware and hybrid treatments is around 8%. Although we would ideally not see any difference in CTR, this drop is not overly large, as users reached through expanded targeting are still clicking on ads at a reasonable rate. Still, improving CTR is the one obvious area for improvement, and we will later present some ideas as part of our future work. It is interesting to note that CTR remains nearly unchanged globally; the explanation is that while a user shown an impression due to expansion may be somewhat less likely to click than the campaign's originally-targeted users, the user is not less likely to click on this impression than on the impression that would otherwise have been shown without expansion.

Overall, our live traffic experiment confirms that our approaches to audience expansion (particularly the campaign-aware and hybrid approaches) are able to greatly increase the reach of campaigns while maintaining reasonable costs and user engagement levels.

Encouraged by the result, we gradually ramped the Audience Expansion system to a full enablement with the hybrid approach being the dominant strategy. The impact of the system to the marketplace has been significant, as can be seen from the change over time in desktop coverage in Figure 3. Coverage is defined as the number of delivered impressions as a percentage of the number of available impressions, which measures how effective the ads inventory is utilized. It is evident that the coverage increases in response to the Audience Expansion ramp-up.

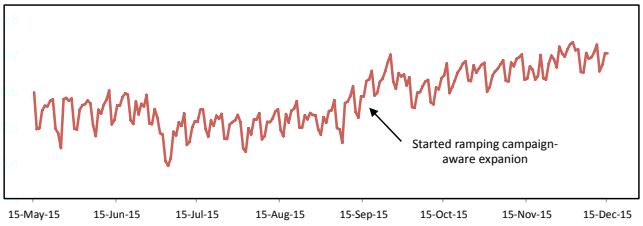

Figure 3: Sponsored Updates coverage on desktop.

5.4 User Similarity and CTR

As described in Section 4.2, campaign-aware expansion computes a similarity score for each user in the expanded audience by summing the Similar-Profile score between that user and all users in the original audience. To determine whether similarity to the original audience is a reasonable predictor of user interest, we can investigate the relationship between this score and clickthrough rates. For each impression shown to a user in a campaign's expanded audience, we look up the similarity score given to that user and then compute the quantile of this score among all of this campaign's impressions. Then, we consider only impressions where the quantile exceeds some threshold and determine the CTR for this subset of expanded impressions. In other words, we can estimate what the CTR of all expanded impressions in our live experiment would have been if, for each campaign, we had only expanded to users with scores above a given quantile for that campaign. Note that this is only possible for impressions where the user was targeted due to campaign-

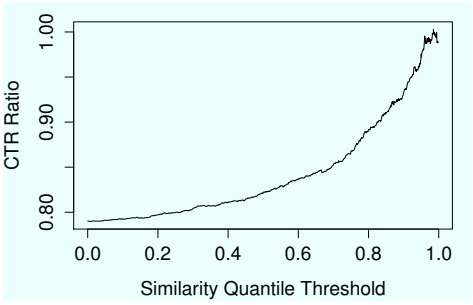

Figure 4: Relative CTR of expanded impressions, thresholded by per-campaign similarity quantile

aware expansion, as campaign-agnostic expansion does not produce a similarity score.

Figure 4 shows the relative CTR (compared to the global CTR of the control treatment) for different quantile thresholds. For a threshold of zero, meaning that all expanded impressions are included, we see a ratio of 0.79. Thus a typical expanded impression would be clicked 79% as often as an impression due to the original targeting. As the threshold increases, the CTR ratio increases, until at thresholds above 0.95 we see a CTR ratio near 1. This means that if we had only expanded each campaign to the most promising 5% of new users, we would likely have seen no significant drop in overall CTR. The clear relationship between similarity score and CTR demonstrates that our similarity scores are indeed useful for identifying promising users for expansion. In addition, if in the future we wish to tune the degree of expansion for individual campaigns, using similarity score thresholds would be a sound approach.

6. CONCLUSION AND FUTURE WORK

We present the Audience Expansion system as a way to optimize ads targeting on social network platforms. We describe two methods for Audience Expansion: campaign-agnostic user attribute expansion, and campaign-aware lookalike expansion. We also demonstrate that due to their complementary nature, we can create a hybrid method to include both approaches, combining the strengths and offsetting the weakness of each. We show the effectiveness of the proposed methods through the large-scale experiment at LinkedIn. Our results indicate significant benefits for both advertisers and LinkedIn achieved by a more accessible and efficient market ecosystem resulted from Audience Expansion.

In future work, we plan to extend the method for campaign-aware expansion to include more behavior-based signals (e.g., ad clicks and social actions). We would like to explore more advanced quality control of the expanded audience. We also want to allow advertisers to optionally include a seed audience, such as known converters, so that we can learn campaign-specific performance-based models.

7. REFERENCES

[1] W. Amaldoss, K. Jerath, and A. Sayedi. Keyword management costs and "broad match" in sponsored search advertising. *Marketing Science*, 2015.

[2] A. Bagherjeiran, A. Hatch, A. Ratnaparkhi, and R. Parekh. Large-scale customized models for advertisers. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pages 1029–1036. IEEE, 2010.

[3] A. Bindra, S. Pokuri, K. Uppala, and A. Teredesai. Distributed big advertiser data mining. In *Proceedings of the 2012 IEEE International Conference on Data Mining Workshops*, pages 914–914, Dec 2012.

[4] A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, D. Metzler, L. Riedel, and J. Yuan. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th WWW*, pages 511–520. ACM, 2009.

[5] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Thiry-fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pages 380–388. ACM, 2002.

[6] W. Chen, D. He, T.-Y. Liu, T. Qin, Y. Tao, and L. Wang. Generalized second price auction with probabilistic broad match. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, EC '14, pages 39–56. ACM, 2014.

[7] P. Dhangwatnotai. Multi-keyword sponsored search. In *Proceedings of the 12th ACM Conference on Electronic Commerce*, EC '11, pages 91–100. ACM, 2011.

[8] E. Even Dar, V. S. Mirrokni, S. Muthukrishnan, Y. Mansour, and U. Nadav. Bid optimization for broad match ad auctions. In *Proceedings of the 18th WWW*, pages 231–240. ACM, 2009.

[9] S. Gupta, M. Bilenko, and M. Richardson. Catching the drift: Learning broad matches from clickthrough data. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 1165–1174. ACM, 2009.

[10] B. Kanagal, A. Ahmed, S. Pandey, V. Josifovski, L. Garcia-Pueyo, and J. Yuan. Focused matrix factorization for audience selection in display advertising. In *29th ICDE*, pages 386–397, April 2013.

[11] A. Mangalampalli, A. Ratnaparkhi, A. O. Hatch, A. Bagherjeiran, R. Parekh, and V. Pudi. A feature-pair-based associative classification approach to look-alike modeling for conversion-oriented user-targeting in tail campaigns. In *Proceedings of the 20th WWW*, pages 85–86. ACM, 2011.

[12] S. Pandey, M. Aly, A. Bagherjeiran, A. Hatch, P. Ciccolo, A. Ratnaparkhi, and M. Zinkevich. Learning to target: What works for behavioral targeting. In *Proceedings of the 20th CIKM*, pages 1805–1814, New York, NY, USA, 2011. ACM.

[13] A. Ramesh, A. Teredesai, A. Bindra, S. Pokuri, and K. Uppala. Audience segment expansion using distributed in-database k-means clustering. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising*, ADKDD '13, pages 5:1–5:9. ACM, 2013.

[14] J. Shen, S. C. Geyik, and A. Dasdan. Effective audience extension in online advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 2099–2108. ACM, 2015.

[15] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 261–270. ACM, 2009.