Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting

Foster Provost
NYU & Media6°, NYC, USA
provost@acm.org

Brian Dalessandro
Media6°, NYC, USA
briand@media6degrees.com

Rod Hook
Coriolis Ventures, NYC, USA
rod@coriolisventures.com

Xiaohan Zhang
NYU & Media6°, NYC, USA
zazahan@gmail.com

Alan Murray
Coriolis Ventures, NYC, USA
alan@coriolisventures.com

ABSTRACT

This paper describes and evaluates privacy-friendly methods for extracting quasi-social networks from browser behavior on user-generated content sites, for the purpose of finding good audiences for brand advertising (as opposed to click maximizing, for example). Targeting social-network neighbors resonates well with advertisers, and on-line browsing behavior data counterintuitively can allow the identification of good audiences anonymously. Besides being one of the first papers to our knowledge on data mining for on-line brand advertising, this paper makes several important contributions. We introduce a framework for evaluating brand audiences, in analogy to predictive-modeling holdout evaluation. We introduce methods for extracting quasi-social networks from data on visitations to social networking pages, without collecting any information on the identities of the browsers or the content of the social-network pages. We introduce measures of brand proximity in the network, and show that audiences with high brand proximity indeed show substantially higher brand affinity. Finally, we provide evidence that the quasi-social network embeds a true social network, which along with results from social theory offers one explanation for the increases in audience brand affinity.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications—data mining; I.2.6 [**Artificial Intelligence**]: Learning—induction; I.5.1 [**Pattern Recognition**]: Models—statistics; J.4 [**Computer Applications**]: Social and Behavioral Sciences

General Terms: Algorithms, Design, Experimentation

Keywords: on-line advertising, predictive modeling, social networks, user-generated content, privacy

This work was conducted while the authors were at Media6°. Foster Provost thanks NEC for a Faculty Fellowship, Daniel P. Paduano for a Fellowship in Business Ethics, and the Marketing Science Institute for a research award.

KDD'09, June 28–July 1, 2009, Paris, France.
Copyright 2009 ACM 978-1-60558-495-9/09/06 ...$5.00.

1. INTRODUCTION

This paper introduces a privacy-friendly method for taking advantage of user-generated content on social networking sites (and beyond) to improve audience identification for on-line brand advertising. Unlike direct-marketing-style on-line advertising, the goal of on-line brand advertising is not only to generate clicks or near-term on-line purchases. On-line brand advertising focuses on getting a brand-oriented message to an audience of interest. This introduces opportunities as well as challenges for developing a data mining solution. For example, one challenge is that there are no true "negative examples" with which to train classifiers. An opportunity is that advertisers tend to believe that brand affinity is likely to cluster in social networks [21], similar to product affinity [16]. Therefore, the social-network neighbors of those already exhibiting brand affinity are an attractive brand audience. The techniques we describe and evaluate attempt to take advantage of this clustering, in a privacy-friendly fashion.[1]

On-line brand advertising has a huge opportunity for growth. Even though the majority of on-line ads are display ads, sponsored search advertising is responsible for the majority of advertising revenue and profit [8]. Well-designed brand advertising may be more appropriate for much display advertising, since unlike for sponsored search the user has not come for the express purpose of clicking on a returned link.[2] ComScore [7] recently reported a clear correlation between seeing on-line brand advertising and increasing both on-line and off-line purchases, well into the future (beyond the reach of current view-through conversion measurement[3] technology), which echoed prior industry results (e.g., [1]). Furthermore, due in large part to the stabilization of the ad technology business landscape, with a few large ad exchanges auctioning massive numbers of non-premium display slots, industry analysts forecast that the growth of the non-premium display market will significantly outpace the overall online ad market [8].

[1] The techniques we introduce may be beneficial for traditional on-line advertising as well, but that is not the focus of this paper.

[2] Improving on-line brand advertising may also have a substantial impact on social welfare: the access to a large amount of free content on-line is due largely to (the hope for) sponsorship by display advertising.

[3] Large ad exchanges collaborate with advertisers to try to determine whether a browser undertook a conversion action within a prespecified time period of viewing an ad, even without having clicked on the ad.

One contribution of this paper is to address the key question: How can on-line brand audiences be assessed? We present a framework for assessing on-line brand advertising audiences. It is meant to complement rather than supplant traditional brand-advertising evaluations, and to offer an on-line alternative to click-maximizing. The key is that certain brand actions become visible and measurable on-line, which helps to circumvent the traditional difficulty of evaluating brand advertising. In short, we adapt a predictive-modeling-style hold-out evaluation based on a selected audience's density of *brand actors*—those browsers who take certain observable actions indicative of brand affinity, e.g., visiting a brand loyalty club page or a purchase thank-you page. More specifically, to evaluate whether a technique identifies a "better" audience for a brand, we compare the density of brand actors in the identified (holdout) audience to the baseline density of brand actors in the population as a whole. The evaluation is based on an inference that resonates well with advertisers and marketers: if the audience has a higher density of brand actors, then the non-actors in the audience (the vast majority) will be better candidates for brand advertising.

The second contribution of this work is the introduction of a method for identifying good brand audiences. We extract a (quasi-)social network from browsing data, select the social-network neighbors of previous brand actors, and then calculate a measure of "brand proximity" to rank the social-network neighbors. A brand audience of a desired size can be chosen by selecting the top of the resultant ranking.

Existing on-line brand advertising usually follows an on-line adaptation of off-line brand advertising (the vertical television/magazine model): associate brand advertisements with top-notch, brand-relevant content. Unfortunately, consumers spend most of their time on-line away from such "premium slots." Our technique is complementary: we identify an audience of browsers of interest and target them anywhere on the web, e.g., through ad networks or via ad exchange auctions of non-premium display slots.

For this study we use data on visits to social networking sites, which allows us to reach (potentially) up to 75% of Internet users [20], and also allows us to infer the structure of social networks among Internet users. One of the most influential results of social theory is the notion of homophily [21]: that social relationships tend to be made between people with similar characteristics. This has been shown to be directly useful for targeted direct marketing: Hill et al. [16] show that social-network neighbors of existing customers are substantially more likely to respond to an offer for a telecommunications product than consumers who do not have an existing customer as a social-network neighbor. We evaluate this technique on brand-affinity data for more than a dozen well-known national and international brands, showing that the identified audiences indeed exhibit markedly higher brand affinity. We also show evidence that a robust measure of brand affinity can be learned.

The third main contribution of this paper is to provide suggestive evidence that the extracted quasi-social network actually embeds a true social network. Thus, more generally, this paper offers an approximate method of identifying "friends" anonymously. An ad network implementing such a technique can engage in social-network targeting without collecting or saving any data on browsers' identities or the content of the social-network pages they visit.

2. NETWORK NEIGHBORS IN MICRO-CONTENT AFFINITY NETWORKS

The method we introduce for creating brand audiences is based on two assumptions. First, *micro-content affinity*—co-visitation of the same user-generated micro-content—leads to brand affinity. By user-generated micro-content (UGC), we mean pages created by individuals outside the scope of a professional engagement, such as pages on social networking sites (the focus of this paper), photograph sites, non-professional blogs, etc. For social networking sites, this assumption is supported by the results presented below.

The second assumption is that micro-content affinity acts in important ways like true interpersonal relationships, and in fact may indicate actual interpersonal relationships depending on the UGC. For example, people who visit the same social-network pages may well be true friends or relatives. If so, such affinity networks embed actual social networks. Network neighbors being true friends is not critical for the techniques to work, but provides an important motivation for brand advertising, due to the wealth of results showing that people with social relationships are more likely to be similar along many different dimensions [21], including likelihood of purchasing a particular product [16].

2.1 Brand audiences via network neighbors

To select the audience for a brand we first use *visits* to UGC to define an anonymous, quasi-social network among web browsers. A *browser* is an anonymous visitor to one or more web pages. Advertising networks serve massive numbers of ads to massive numbers of browsers, and via cookies keep track of which browsers visit what content. Each time two browsers are observed to visit the same UGC page, an affinity-network link is placed between the browsers. Technically, this network is induced from the bipartite affinity graph between users and UGC. Frequencies of visitation can become strengths for the individual links. This method for audience selection has the advantage that it can operate on doubly de-identified data: browsers are represented by random numbers, *and* content pages are represented by random numbers. Targeting the audience can be done through normal ad network procedures, which require only that the ad network tell the ad exchange to target the browsers in a given set based on their cookies—the ad network need transfer no data about the browser besides the (otherwise random) cookie id.

For brevity, clarity and emphasis, since there are two different affinity graphs, we will refer to the network induced *among browsers* as the "quasi-social" network. We add the prefix "quasi-" here because technically at this point we do not know who are true friends and who just share a strong content affinity but don't actually know each other. We return to this in section 4.5. Keeping this in mind, we will drop the "quasi" except where necessary to distinguish from a real social network.

In order to assemble proposed high-quality brand audiences, we select the subset of the social network neighbors closest to a set of *seed nodes*. The seed nodes are browsers in the network identified (ex ante) or estimated to exhibit brand affinity. To instantiate the method, we must define (i) a precise sort of seed node to use, and (ii) what it means to be close to the set of seed nodes.

2.2 Defining seed nodes

How to define seed nodes depends on the information available to the advertiser and to the ad network implementing the method. For example, seed nodes could represent existing customers, or consumers who have exhibited interest in the company's product, or consumers estimated to belong to a desired demographic or psychographic group.[4] For this paper we will consider observable brand-associated actions of interest. More specifically, define *brand actors* to be those browsers observed to have visited a particular brand-oriented page selected by the advertiser, e.g., a brand loyalty page, a customer login landing page, a purchase thank-you page, or simply the company's home page. The *seed nodes* are browsers known at the time of audience selection to be brand actors (future brand actors will be used for evaluation). Specifically, let $\mathcal{B}$ be a set of M web browsers under consideration. We will consider the brand audience for each brand separately. Let the seed nodes compose a subset of the browser set $\mathcal{B}^+ \subseteq \mathcal{B}$. Let all the other non-seed-node browsers (*candidate nodes*) belong to the set $\mathcal{B}^0 = \mathcal{B} - \mathcal{B}^+$.

2.3 Brand proximity

Our goal is to compose a brand audience of interest $\mathcal{A} \subseteq \mathcal{B}^0$ based on browsers' proximity to $\mathcal{B}^+$, such that a larger-than-baseline proportion of the browsers in $\mathcal{A}$ are likely to be as-of-yet unobserved brand actors. *Brand proximity* is a distance/similarity measure between candidate nodes and the *set* of seed nodes. Brand proximity can be calculated as an aggregation over proximity measures for individual nodes, or based on the set as a whole; we experiment with both below. There are countless ways to define similarity or distance between individual nodes in a network [19]; here we use straightforward measures for simplicity and efficiency. If more sophisticated techniques [14, 18, 19, 25, 26] can scale, they may provide the basis for improving the results we present below.

Assume that there are N user-generated micro-content pages in total that the browsers have visited. Browsers and content form a bipartite graph which can be represented by a $M \times N$ browser-content matrix as:

$$\Gamma = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1N} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \cdots & \gamma_{MN} \end{bmatrix},$$

where each browser $b_i \in \mathcal{B}$ is represented by a row in Γ—a content vector $\overrightarrow{\gamma_i} = [\gamma_{i1}, \gamma_{i1}, \cdots, \gamma_{iN}]$. In this generalized representation, each γ_{ij} represents the weights of the links in the bipartite graph. In the simplest case, γ_{ij} is simply one or zero, a binary value indicating whether browser b_i has visited content piece c_j and Γ is the biadjacency matrix for the bipartite graph. Non-binary weights can be computed in various ways, e.g., as the frequency with which browser b_i has visited content piece c_j (what we do for this paper), penalizing for popularity such as with tfidf, and/or damping older counts [9].

The *network neighbor audience* for a brand is $\mathcal{N} = \{b_i : \overrightarrow{\gamma_i} \cdot \overrightarrow{\gamma_k}' \neq 0 \text{ for some } b_k \in \mathcal{B}^+\}$. For this paper, the largest high-quality audience we will select is $\mathcal{N}$, and every audience we choose will be a subset of $\mathcal{N}$ (these will be the only browsers with non-zero brand proximity). This will allow efficient computation of $\mathcal{A}$ over massive social networks, since the identification of $\mathcal{N}$ can be done very efficiently. Our conjecture is that immediate network neighbors give substantially more lift than neighbors further in the network, as with prior work on network-based marketing [16, 17]. Furthermore, for brand proximity measures that aggregate over individual node proximities, we will be most interested in the proximity of a browser to those in the set $\mathcal{B}^+$, rather than arbitrary inter-browser proximities (although it may be useful to create a baseline for comparison).

Since we do not have a theoretically optimal brand proximity measure, we can represent brand proximity for b_i by a vector $\overrightarrow{\phi_{b_i}} = [\phi^1_{b_i}, \phi^2_{b_i}, \cdots, \phi^P_{b_i}]$, where each $\phi^p_{b_i}$ is one of P different proximity measures. We use $\overrightarrow{\phi_{b_i}}$ as the basis for selecting $\mathcal{A}$. For our main results, we rank the candidate nodes $b_i \in \mathcal{B}^0$ based on some monotonic function of the projection of $\overrightarrow{\phi_{b_i}}$ onto one of the proximity dimensions:

$$score(b_i) = f_i(\overrightarrow{\phi_{b_i}} \cdot \overrightarrow{I_q}), \tag{1}$$

where $\overrightarrow{I_q} = [0, \cdots, 1, \cdots, 0]'$ is a selection vector with 1 on its *qth* row, and f_i is a monotonic function to map the single proximity measure selected by $\overrightarrow{I_q}$ to a ranking score for b_i. Alternatively, if desired, we can treat $\overrightarrow{\phi_{b_i}}$ as a feature vector to learn a brand-specific proximity measure to rank the candidates. In either case, the brand audience would comprise the top-ranked browsers in $\mathcal{B}^0$.

2.4 Some brand proximity measures

For the results presented below, we use five brand proximity measures. For clarity, we will use b^0_i to denote a candidate node $b_i \in \mathcal{B}^0$ and b^+_k to denote a seed node $b_k \in \mathcal{B}^+$. For a browser b_i (either a seed node or a candidate node), let $\mathcal{C}_{b_i}$ be the set of content pieces to which the browser is linked in the bipartite graph; i.e., $\mathcal{C}_{b_i}$ corresponds to all nonzero entries in b_i's content vector $\overrightarrow{\gamma_i}$.

1. **POSCNT:** the number of unique content pieces through which paths in the bipartite graph connect b^0_i and any $b^+_k \in \mathcal{B}^+$:

$$\text{POSCNT}(b^0_i) = |\mathcal{C}_{b^0_i} \cap (\bigcup_{b^+_k \in \mathcal{B}^+} \mathcal{C}_{b^+_k})|. \tag{2}$$

2. **MATL:** the maximum number of unique content pieces through which paths connect a candidate browser to any single seed node ("maximum action taker linkage"):

$$\text{MATL}(b^0_i) = \max_{b^+_k \in \mathcal{B}^+} (|\mathcal{C}_{b^0_i} \cap \mathcal{C}_{b^+_k}|). \tag{3}$$

3. **maxCos:** the maximum cosine similarity of the candidate node's content vector to that of any seed node. The cosine similarity between a candidate node b^0_i and a seed node b^+_k is:

$$COS(b^0_i, b^+_k) = \frac{\overrightarrow{\gamma_i} \cdot \overrightarrow{\gamma_k}'}{\|\overrightarrow{\gamma_i}\| \|\overrightarrow{\gamma_k}\|}, \tag{4}$$

and thus the maxCos for a candidate node b^0_i is:

$$\text{maxCos}(b^0_i) = \max_{b^+_k \in \mathcal{B}^+} (COS(b^0_i, b^+_k)). \tag{5}$$

[4] See for example http://www.mindset-media.com/.

4. **minEUD:** the minimum Euclidean distance between the normalized content vector of a candidate node and that of any seed node. Specifically, for browser b_i, let $\gamma_{tot} = \sum_{j=1}^{N} \gamma_{ij}$ be the sum of weights across all content pieces that b_i is linked to. We normalize b_i's content vector as:

$$\overrightarrow{\gamma_i}^n = \frac{1}{\gamma_{tot}} [\gamma_{i1}, \gamma_{i1}, \cdots, \gamma_{iN}]. \qquad (6)$$

The minumum Euclidean distance is calculated in the obvious manner.

5. **ATODD:** the ratio of the number of a browser's neighbors that are seed nodes to the number of its neighbors that are not seed nodes ("action taker odds"). Specifically, let $\deg^+(b_i)$ and $\deg^0(b_i)$ be the number of links incident to b_i from seed nodes and candidate nodes, respectively.

$$\text{ATODD}(b_i^0) = \frac{\deg^+(b_i^0)}{\deg^0(b_i^0)}. \qquad (7)$$

3. EVALUATION FRAMEWORK

One contribution of this paper is a framework for evaluating on-line brand audiences. It is an adaptation of predictive modeling holdout testing, but to our knowledge there has been no prior application to on-line brand advertising: Choose two non-overlapping, ordered time periods, t_1 and t_2. Consider a set of browsers $\mathcal{B}$ known in time t_1. The seed nodes $\mathcal{B}^+$ are those elements of $\mathcal{B}$ for which a brand action is observed in t_1. Let us call the seed node set $\mathcal{B}_1^+$ to clarify that the seed nodes are the brand actors in time t_1. For evaluation, the candidate nodes $\mathcal{B}^0$ are those elements of $\mathcal{B}$ that are observed in time t_2. The *future brand actors*, $\mathcal{B}_2^+$, are those elements of $\mathcal{B}^0$ who are observed to take a brand action in t_2. As with a predictive modeling holdout evaluation, information about action taking of the holdout set is not used in selecting the audience.

To evaluate any audience $\mathcal{A}$, we can compute the future density of brand actors as:

$$\frac{|\mathcal{A} \cap \mathcal{B}_2^+|}{|\mathcal{A}|}. \qquad (8)$$

Audiences can be compared based on their future brand actor densities. The important twist from standard response modeling is that this can be done either with or without advertising. To advertisers, a larger proportion of an audience showing brand affinity "organically" (i.e., without advertising) is highly indicative that the audience is a good audience for brand advertising. Furthermore, unlike click-based evaluations, it can be used to judge brand affinity separately from someone just being a "clicker" [6].

Evaluation and comparison can be done based on any measure of density of a binary attribute over a set of data. We are interested in how well the different proximity measures rank the candidates, and we presume that a particular campaign will target some upper portion of the ranking depending on the advertising budget and other considerations. Thus, we report the area under the ROC curve (AUC, equivalent to the Mann-Whitney-Wilcoxon statistic), which measures how well a scoring system can rank members of one class above the other [13]. In this application, a higher AUC means that an audience selected from the top of the ranking will have a higher density of brand actors. Human brand actions are fundamentally difficult to predict, and as with targeted marketing reponse modeling we would expect low but hopefully better than random AUCs.[5] To illustrate the relative increase in brand-actor density over a baseline audience ("lift" in brand-actor density) we also report results for the top-10% of $\mathcal{N}$ for each ranking, which is reasonable for the application.

4. RESULTS

We now present results assessing the effectiveness of targeting close social network neighbors for identifying brand audiences on-line.

4.1 Data

The results are based on anonymized browsing and action-taking data from a working ad network. Specifically, Γ is built from a sample of page visits to several of the largest social networking sites over a 90-day period. The sampling begins with a quasi-random (convenience) sample of browsers from every server log file every 10 minutes across the server farm, resulting in about 10 million unique browsers who have visited social network (SN) content over a 90-day period. As far as the ad network experts can tell, there is no systematic bias to this sampling. For each of these browsers we query for all the observed page visits over SN content over the time period. On average, a browser has about 25 visits recorded to unique SN pages. The resulting Γ has approximate dimensionality $10^7 \times 10^8$, with about 250 million non-zero entries.

A set of major brands is divided into two groups: (1) four brands for which in the experimental period no advertising was done by the ad network (Hotel A, Modeling Agency, Credit Report, Auto Insurance), plus a fifth "brand" that consists of a demographic group of intense interest to certain large-scale adverisers, for which no amount of advertising will change one's membership status in the short term (Parenting); (2) ten brands for which some advertising was done (Apparel Hiphop, Apparel Athletic, Apparel Women's, Voip A, Voip B, Airline, Hotel B, Electronics A, Electronics B, Cell Phone). In group 2, advertising was done more or less uniformly across $\mathcal{N}$, so being able to rank within $\mathcal{N}$ would be indicative of some combination of brand affinity and differences in "response" to the advertisements. The latter plays more or less of a role for different brand actions. In no case does clicking on the ad lead directly to an observed brand action. However, browsing the site or purchasing may. For example, for Hotel B, the action is reserving a room. It may be that a browser is indeed influenced by the advertising to reserve a room in this particular hotel chain during the testing period, in which case brand affinity is exhibited, but it would not be purely organic.

In these data we have on average about 100,000 "seed" action takers per brand, with the actual number varying between 5000 and 1 million. In the experimental data, each seed action taker has on average 20-40 social-network neighbors, with the resultant network-neighbor audiences being up to 20-40 times the number of seeds. Choosing the "closest" in brand proximity involves sub-selecting from these.

[5] For example, good models for the KDDCUP 1998 targeted marketing data yield AUCs of around 0.6.

Table 1: Areas under ROC curves (AUCs) for different brand proximity measures and brands. Block to left of double line is over all candidate browsers; right block is only on $\mathcal{N}$. Details are in the text. Bold is max for each brand in each row and block. A * indicates a value is statistically significantly better than the next best in the left block. maxUNI shows the AUC for the best univariate measure.

Brand	**MATL**	**maxCos**	**POSCNT**	**minEUD**	**ATODD**	**maxCos($\mathcal{N}$)**	**maxUNI($\mathcal{N}$)**
Hotel A	0.617	0.617	**0.628***	0.612	0.604	0.4994	**0.5561**
Modeling Agency	0.629	**0.636**	0.630	0.618	0.634	0.5746	**0.6347**
Credit Report	0.631	**0.656***	0.630	0.643	0.597	**0.5633**	**0.5633**
Auto Insurance	**0.604**	0.584	0.603	0.593	0.600	0.4622	**0.5726**
Parenting	0.679	**0.692***	0.678	0.633	0.623	**0.5863**	**0.5863**
Apparel: Hiphop	**0.662**	0.659	0.658	0.640	0.629	0.607	**0.7343**
Voip A	0.627	**0.668***	0.617	0.636	0.567	**0.6514**	**0.6514**
Voip B	0.605	**0.610**	0.596	0.609	0.590	0.7086	**0.7122**
Airline	0.611	**0.615**	0.611	0.598	0.592	**0.6287**	**0.6287**
Hotel B	0.596	**0.611**	0.598	0.606	0.583	**0.6619**	**0.6619**
Electronics A	0.682	0.677	**0.687***	0.638	0.639	0.5543	**0.5762**
Electronics B	0.604	**0.610**	0.607	0.609	0.608	0.5343	**0.6141**
Apparel: Athletic	0.599	0.599	**0.607***	0.568	0.596	0.5535	**0.5571**
Cell Phone	0.777	**0.790***	0.778	0.743	0.699	**0.6601**	**0.6601**
Apparel: Women's	0.607	**0.618**	0.607	0.616	0.578	**0.6208**	**0.6208**

For example, in some experiments below we choose the top-10% of $\mathcal{N}$. Technically, we would expect to see different future brand actor densities even for the same brand for different brand actions; here for simplicity we just refer to each as "the brand". In one case (Voip A & B), the two "brands" are two different actions for the same brand.

4.2 Results: Univariate brand proximity

The left side of Table 1 shows AUC results for the various techniques over the entire candidate set ($\mathcal{B}^0$). The uppermost five rows correspond to brand group 1, and the bottom ten rows to brand group 2. Consider brand group 1. For all five brands, all five brand proximity measures give AUCs which show statistically significantly[6] better-than-random separation of future brand actors, indicating that the proximity measures indeed do rank audiences with respect to their brand affinity. Absolute AUC values are difficult to assess out of context, and as far as we know, this is the first attempt to evaluate brand audiences in this way. They compare favorably to the AUC values one typically gets for targeted marketing, even though here the audiences have not been advertised to. The results are even stronger for group 2; all are statistically significantly better than random, and one (Cell Phone) is remarkable (AUC=0.79).

As noted previously, ranking results for some group 2 brands may be affected by advertising. Column "maxCos($\mathcal{N}$)" provides an evaluation using maxCos on only $\mathcal{N}$, a set that received uniform advertising during the experimental period. Note that this already is a set that we believe to have excellent brand affinity. Column "maxUNI($\mathcal{N}$)" shows the AUC for the best univariate measure on $\mathcal{N}$. Note that when maxCos is statistically significantly better than the others on $\mathcal{B}^0$, it's also the best on $\mathcal{N}$. These results show one view of whether *ranking* among the network neighbors themselves provides additional gain over just *identifying* $\mathcal{N}$ and advertising to it indiscriminately. Recall that for those brands where response might result in an observed brand action, this gain is a combination of pure brand affinity and response likelihood. The results here are solidly positive as well. The restriction of the data results in many fewer positive testing examples, and increases the variance in the AUCs substantially, so detecting statistical significance becomes more difficult (and the seeming increases in AUC over the former setting may be illusory). Nonetheless, for group 2 for maxCos, 8 out of the 10 individual AUCs are statistically significantly better than random (not Apparel Hiphop or Electronics B); 10 out of 10 are greater than 0.5 (strongly significant by a sign test), and the overall average AUC is 0.61, which is quite respectable, especially when considering that the baseline here ($\mathcal{N}$) already is expected to be a high-performing group. The even stronger results for maxUNI suggest trying to select or combine measures.

An alternative evaluation is to assess the increase in brand actor density (hereafter, *density*) directly on a selection of browsers from the top of the ranking. The results are qualitatively similar to those presented in the next section (when comparing the univariate measures to the multivariate model), which are plotted there. We selected audiences comprising the most highly ranked 10% of $\mathcal{N}$ for each brand and proximity measure. Overall, the results echo the AUC results with the exception that ATODD is the dominant technique. For group 1 we see increases in density ranging from an 80% increase to a 500% increase ("lifts" of 1.8 to 6.0). In group 2 we see even larger increases. As with the AUC results, we also generally see lifts in density as compared to the density of brand actors in $\mathcal{N}$. The top-10% network neighbors clearly show an increase in density over $\mathcal{N}$ (increases in 13 of 15 cases when ranked by ATODD; highly statistically significant by a sign test); the average increase in density over the network neighbors is 100% (a lift of 2.0).

[6] All statistical significance results for AUCs are computed using the procedure described by [10], which is the procedure implemented by the commercial statistical packages SAS and Stata. Qualitatively similar results were obtained via t-tests over mean AUCs from randomly resampled test sets.

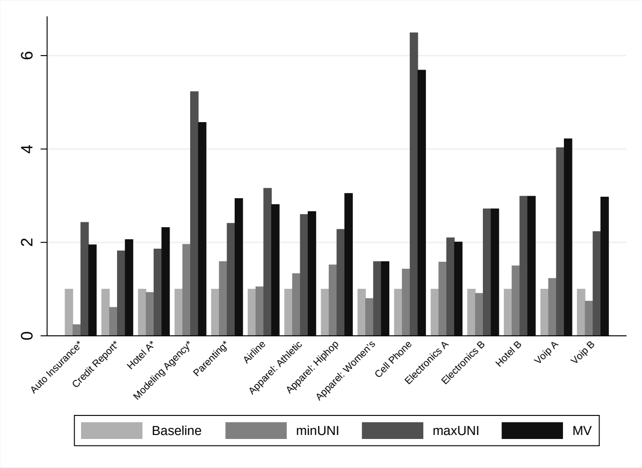

Figure 1: Brand actor densities for the top-10% of $\mathcal{N}$ selected by the best and worst univariate proximity measure and the trained multivariate measure, compared to the density over the entire candidate set $\mathcal{B}^0$ ("Baseline"; normalized to one).

4.3 Results: Multivariate brand proximity

No univariate measure is consistently the best, and in fact for each method that is statistically significantly best for some brand, there is another brand for which it is statistically significantly beaten by another measure. As introduced briefly above, since we represent each candidate browser by a vector $\overrightarrow{\phi_{b_i}}$ of proximity measures, we can directly combine them into a multivariate measure designed specifically to identify high-brand-affinity browsers. Ideally, such a measure would be more robust across brands.

For this paper, the rank of a candidate browser $b_i \in \mathcal{B}^0$ is calculated by a multivariate logistic function of the elements of $\overrightarrow{\phi_{b_i}}$:

$$rank(b_i) = \frac{\exp(\sum_{p=1}^{P} w_p \phi_{b_i}^p)}{1 + \exp(\sum_{p=1}^{P} w_p \phi_{b_i}^p)}. \tag{9}$$

where w_p are brand-specific weights. The weights are computed with standard MLE logistic regression, using an extension of the holdout framework presented in section 3. Specifically, we extract from the set of candidate nodes an additional training set, comprising brand-actors and non-brand-actors in t_2. Any evaluation, of course, will be conducted on a disjoint (hold-out) data set from t_2.[7] In order for the comparison to be as fair as possible, for this comparison we also added these training brand-actors to the seed set for the univariate measures. The technique that we report as MV chooses the best of the univariate and multivariate models, based on estimated AUC using cross-validation on the training set.

The results (not depicted) show that in terms of AUC, MV is always at least as good as the best UNI. For ranking all of $\mathcal{B}^0$ MV posts a win-tie-loss record of 14–1–0, and for ranking $\mathcal{N}$, a win-tie-loss record of 9–6–0; both are significant by a sign test and in both cases there are 5 individually significant wins. Thus, with training we can learn to do at least as well as the best univariate brand proximity measure, and often better.

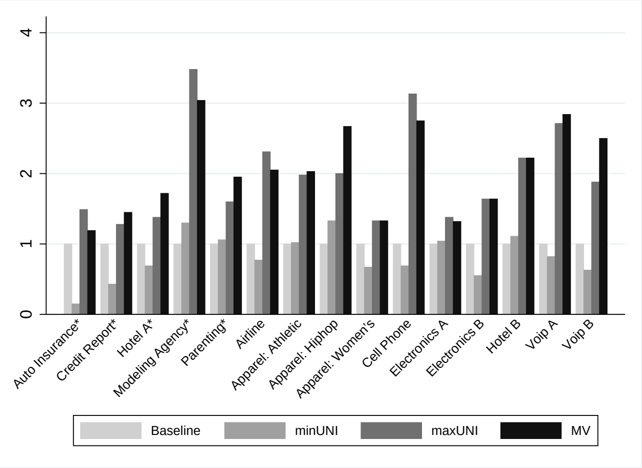

Figure 2: Brand actor densities as in Figure 1, except with the normalization baseline being the brand density of the network neighbor set ($\mathcal{N}$), showing the additional lift (if any) over targeting all $\mathcal{N}$.

Figures 1 and 2 show the brand actor densities of MV and the univariate measures with the best and worst performance for each brand, normalized so that each figure's baseline ($\mathcal{B}^0$ and $\mathcal{N}$, respectively) is one for each brand. For the top of the rankings, MV adds consistency but on average does not improve the top-of-the-list densities significantly over the best univariate measure. This may be showing that MV's advantage is in rescoring the browsers below the top-10%; however, the logistic regression training does not specifically focus on the top of the ranking, so the comparison may not demonstrate the full power of MV. Together, the MV results provide strong evidence that better or more robust models can be learned.[8]

4.4 PSA Tests

It would have been nice to have shown lifts in purely organic brand affinity for all brands. Of course, getting a working ad network to stop advertising to its largest customers is infeasible. As an alternative, we conducted an "in vivo" evaluation, by designing and running experiments in production. Specifically, for three selected group-2 brands, we identified a small audience of close network neighbors. To assess organic brand affinity, we targeted them only with a public service announcement (PSA) across the web, by bidding for them on a major ad exchange. These browsers do not get any brand advertising from the ad network. Simultaneously, we issue a quasi-randomly targeted (RON[9]) campaign with the same PSA and campaign parameters. Since we've targeted these browsers with "ads" (albeit PSAs), we now can obtain from the ad exchange statistics on "view through conversions", specifically, the number of targeted browsers in each campaign who subsequently (e.g., within 7 days) take a predefined brand action (as described previously). Of course the PSA ad should have no effect on a browser's propensity to take the action.

Table 2 shows the number of PSAs shown to top-ranked neighbors and the number shown to RONs, and the "or-

[7] Although we did not do so for this study, to match the use scenario most closely we would like to segment the evaluation data into three time periods.

[8] We have found that other learning-based methods can improve significantly both the rankings overall and the top-of-the-ranking performance.

[9] The ad exchanges allow one to bid on every display slot, called "run of network."

Table 2: Organic action lifts for three group-2 brands.

Brand	**Impressions of PSAs to top ranked**	**Impressions of PSAs to RON**	**Organic conversion lift**
Electronics A	67	53,347	5.89
Apparel: Athletic	26,161	266,661	6.06
Apparel: Hiphop	5,757	223,509	64.65

ganic conversion" lift: the ratio of the percentage of the PSA-targeted *neighbors* who took the action to the percentage of the PSA-targeted *baseline* who took the action. The results demonstrate remarkable organic lifts in brand actor density for the audiences of close network neighbors for these three group-2 brands. Apparel: Hiphop is an interesting case. The brand previously had done the majority of its advertising through word of mouth, and here we indeed see a tremendous lift in brand action density (65 times) for the close neighbors. This provides some initial evidence that the close quasi-social network neighbors are actual social network neighbors. One caveat is that here we need to believe the statistics reported by the ad exchange, and to our knowledge there is no way to verify them directly.

4.5 Social vs. Quasi-social

The notion of targeting social network neighbors resonates with brand advertisers because they believe that the personal networks of those already exhibiting brand affinity should be good targets for brand advertising. A long line of research in sociology supports this, as described by McPherson et al. [21]. In particular, they note "People's personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics."

Our quasi-social network being "only" a content-affinity network may not matter to bottom-line-oriented advertisers, if indeed the networks are identifying audiences with high brand affinity. Nonetheless, if the quasi-social network is defined across visits to pages on social networking sites, it seems that it ought to embed a true social network. Users of social networking sites generally visit their own home pages and their friends' pages (among others), and thus friends should be connected in our quasi-social network. Of course non-friends also will be connected, so one interesting question is whether strong brand proximity selects audiences comprising the actual friends of the brand actors.

One possibility is to map the explicit "friends" network from a social networking site to our network, and examine the overlap. However, this would require that we acquire personally identifying information, which we would prefer not to do. More importantly for drawing conclusions, the veracity of friend links is highly dubious [5].

Instead, based on the content visitation data, we estimate which piece of UGC is most likely to be authored by each browser, following ideas for identifying authors in citation networks [15]. Specifically, we estimate that the content piece a browser visits most, normalized by the overall popularity of the content, is owned by the browser (e.g., is her own social network page); let's call this the browser's home page.

Table 3: F-AUCs showing that friends are very likely to be ranked (by maxCos) higher than those not known to be friends.

Brand	**F-AUC on all $\mathcal{B}^0$**	**F-AUC on $\mathcal{N}$ only**
Hotel A	0.96	0.79
Modeling Agency	0.98	0.84
Credit Report	0.93	0.79
Parenting	0.94	0.80
Auto Insurance	0.97	0.81
15 Brand Average[10]	0.96	0.81

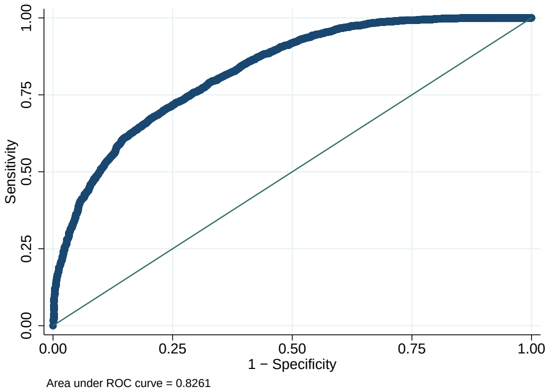

Figure 3: Friends ROC curve for Airline, showing that friends are very likely to be ranked higher than those-not-known-to-be-friends. The top of the ranking is very dense with friends (steep initial rise) and the bottom of the ranking is almost devoid of friends (flat finish).

We evaluated home-page identification accuracy on a separate data set of social-network browsing data which, while still completely anonymous as to user and content, indicated which page belonged to whom. The results showed high accuracy generally (65% correct at choosing a browser's page) and very high (80% correct) for browsers where the method is confident in its prediction (about 60% of the browsers). Then we estimate that two browsers are "friends" only if one visits the other's home page—rather than just having visited similar content.

Now, with this approximate notion of "friends" we can ask: does the audience comprising *close* neighbors of the seed brand actors in the quasi-social network (as measured by brand proximity) actually seem to include the friends of the brand actors? Of course, even if our estimation is accurate, the selection of friends that we observe in a data sample is only a small subset of all the friends. Therefore, we would like to measure whether brand proximity on average tends to rank brand actors' friends higher than those-we-don't-know-are-friends. This is measured by the Mann-Whitney-Wilcoxon test statistic, which is equivalent to the area under the ROC curve when the class is taken to be *known-friend* or *not*. Let's call that F-AUC.

The results are striking. Table 3 shows the F-AUCs for each brand for maxCos, both over all candidate browsers ($\mathcal{B}^0$) and just over $\mathcal{N}$. The $\mathcal{B}^0$ results are striking but ultimately less interesting: In every case they are greater than 0.9, with a mean of 0.96. However, we know that with no

[10] This is the average across all 15 brands.

connection to seed nodes, a browser will be ranked low. The $\mathcal{N}$ results are quite interesting and encouraging: even among the network neighbors, there is an 80% chance that a friend will be ranked higher than a browser who we don't know is a friend. Figure 3 shows the ROC curve (over $\mathcal{N}$) for Airline: the top of the ranking is very dense with friends and the bottom almost devoid of friends.

5. DISCUSSION AND LIMITATIONS

In summary, our main results show unambiguously that we can build high brand-affinity audiences by selecting the social-network neighbors of existing brand actors identified via co-visitation of social-networking pages, without saving any information about the identities of the browsers or content of the pages. These network neighbors tend to take brand actions at a higher rate organically, as well as after being targeted with ads. We also show that it is possible to learn better models, by using the individual univariate proximity measures as features in a higher-level model. And we provide evidence that the quasi-social network likely embeds a true social network (which makes sense if the visitations are over social networking pages).

Table 4: Demographic profiles for CellPhone seeds and their social-network neighbors.

Demographic	Seeds	Neighbors
Gender	Female	Female
Ethnicity	Hispanic	Hispanic
Age	Young	Young
Income	Low	Low
Education	No college	No college

Among other things, brand advertisers would like their audiences to be similar along important dimensions of interest, which is why targeting social-network neighbors resonates well [21]. As a final demonstration, for one brand (Cell Phone) we submitted a set of seed nodes and a set of close network neighbors to the internet analytics firm Quantcast (`http://www.quantcast.com`), which gives statistically estimated demographic profiles for sets of browsers. For the particular brand action selected by the advertiser, the seed nodes and the network neighbors returned with exactly the same profile along all dimensions (see Table 4).

We call this method "privacy friendly" because here we offer no formal proof of the power of the anonymization scheme, but nonetheless assert that it has several attractive properties from a privacy standpoint. First, it can be implemented without ever collecting direct PII (personally identifying information); thus it addresses a primary privacy concern for firms dealing with personal information, i.e., that someone internally can directly look up information about particular individuals. Second, in contrast to other attempts at social-network-based on-line advertising, this method does not use user-posted personal (profile) information. This is important not only from a base-level privacy standpoint, but also as a deterrent to reidentification [27].

A secondary constellation of privacy issues revolves around vulnerability to data breaches and anonymization attacks. We believe that this anonymization is relatively robust to both active and passive attacks (see e.g., [27, 22]), especially as compared to existing practice which normally ignores the danger of reidentification. (1) The data contain no public information at all. (2) The data are *sampled* via a process opaque to the browser, and thus lack many true social network connections. (3) They contain a high degree of additional "noise" links with respect to the true social network, since they are based on content visitations. (4) It is difficult to envision how an attacker would "seed" a passive attack on the anonymized quasi-social network with knowledge of certain members of the network (but see below), and (5) There is little information to "reveal" that would not be more easily found elsewhere. One exception would be the data on brand activity of browsers, which also could be anonymized for storage (to minimize the risk of a data breach), but is necessary for targeting and is also a possible entry point for reidentification. If the advertising firm's data security was breached, *and* a reidentification attack was successful, *and* the content and/or brand-action anonymization scheme were broken (which could begin with the homepage identification discussed above), it may be possible to discern who visited which social-network or brand pages. This is not trivial, but it is information that is collected routinely and widely now without any anonymization, by advertising networks, ad exchanges, search engines, and so on. In addition, the potential harm is arguably mild with respect to that associated with much personal information that is used routinely (unanonymized) for data mining, personalization, and targeting. Thus our claim that this method is privacy friendly. Nonetheless, a privacy-sensitive firm using these techniques may want to consider not even saving anonymized data on certain content, engage only well-respected brands, and beware of saving potentially sensitive details of brand activity.

The greater question of privacy and on-line advertising involves where we as citizens and consumers, collectively, would want on-line businesses to operate on the spectrum between two unacceptable extremes: (a) doing absolutely anything with consumer data regardless of any ethical questions, and (z) being unable to increase business and consumer welfare via data modeling. The answer obviously is beyond the scope of a paper like this. We hope we have illustrated that there are interesting and viable points between the extremes (cf., [11]), that promise increased privacy as well as increased business and consumer welfare (here, better-targeted advertisements). It would be valuable to develop privacy-preserving techniques to augment this privacy friendliness without reducing effectiveness too much, for example by introducing additional randomness into the content visitation network.

We are not aware of prior research on data mining for on-line brand advertising. Prior work tends to focus on sponsored search advertising [4], contextual advertising [3], and display advertising optimized for some action in response to the ad—usually clicks and sometimes more sophisticated conversions. Provost et al. [24] describe in detail a broad set of different sorts of data that can be useful for on-line ad targeting. Unlike this prior work and most actual on-line advertising [4, 3] (especially when measured by current advertising spending), the focus of our research is on finding and evaluating audiences for brand advertising. It makes sense that better brand audiences also will be more likely to click or convert, which would be a natural and attractive by-product (but we have not shown this).

Our problem has some similarity to collaborative filtering (CF), but differs in important ways. The scale of the data is different, which has implications for scalability. Just in our research data set, we have 100 million "items", as compared to at least an order of magnitude fewer even for very large CF systems. More importantly, rather than recommending from among the item set itself ("matrix completion"), our task more closely resembles traditional predictive modeling of a specific target variable, but with a massive number of variables, and technically only positive and unlabeled examples [12]. Nonetheless, it may be that CF-style dimensionality reduction [2] can further improve audience selection.

Although we have tried to design the experiments carefully, there may be some residual bias in data collection. We have investigated this both by using regression analysis including variables that may indicate bias, as well as by trying to seemingly improve the audiences by using bias-related variables, but have not found anything to lead us to question the results. For example, including the number of pages a browser has visited does not reduce the significance of the proximity variables in a regression on brand affinity, and does not systematically increase accuracy.

Any technique operating in the framework we provide will be limited by browser cookie deletion. This is a well-known problem for brand advertisers, who often "retarget" ads to known brand actors (since they've already displayed brand affinity). Over time, deletion of cookies will cause some of the chosen audience members never to present themselves (again) for advertising. Our methods are less sensitive to cookie deletion because in effect our techniques will naturally retarget lost brand actors, as long as they visit the same social networking pages. Moreover, our techniques will naturally, anonymously target brand actors on other computers, if they "act like their own best friends," which may account for some of the effect shown above.

The use of "friends" links from SN sites for direct marketing has received much criticism from the popular press due to privacy concerns [23], and from the academy [5]. In particular, Clemons et al. [5] provide a well-worth-reading argument that advertising on a SN site to SN neighbors is unlikely to be successful. However, they do not discuss brand advertising—which may be successful on SN sites, as brand advertising is different from click-inducing advertising. Nor do they consider that the value of SN sites for delivering ads and their value as a vehicle for collecting data are independent. This paper shows the substantial value of social networking sites as a mine for data on brand affinity (the advertising may take place elsewhere on the web).

6. REFERENCES

[1] Atlas_Institute. Where can you find your customer? Try the intersection of search and display. Digital Marketing Insight, 2007. http://www.AtlasSolutions.com/insights.

[2] R. Bell, Y. Koren, and C. Volinsky. Chasing $1,000,000: How we won the Netflix progress prize. *ASA Statistical and Computing Graphics Newsletter*, 18(2):4–12, 2007.

[3] A. Broder, M. Fontoura, V. Josifovski, and L. Riedel. A semantic approach to contextual advertising. In *SIGIR '07*, pages 559–566, 2007.

[4] M. Ciaramita, V. Murdock, and V. Plachouras. Online learning from click data for sponsored search. In *WWW '08*, pages 227–236, 2008.

[5] E. K. Clemons, S. Barnett, and A. Appadurai. The future of advertising and the value of social network websites: some preliminary examinations. In *ICEC '07*, pages 267–276, 2007.

[6] comScore. New study shows that heavy clickers distort reality of display advertising click-through metrics. `http://www.comscore.com/press/release.asp?press=2060`.

[7] comScore. Whither the Click? comScore brand metrix norms prove "view-thru" value of on-line advertising. `http://www.comscore.com/press/release.asp?press=2587`.

[8] R. Coolbrith. On-line advertising 2.0: The opportunity in non-premium display. ThinkEquity Partners LLC Industry Report, September 2007.

[9] C. Cortes, D. Pregibon, and C. Volinsky. Communities of interest. *Intell. Data Analysis*, 6(3):211–219, 2002.

[10] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3):837–845, 1988.

[11] G. T. Duncan, S. A. Keller-McNulty, and S. L. Stokes. Disclosure risk vs. data utility: The r-u confidentiality map. Technical Report Technical Report Number 121, National Institute of Statistical Sciences, 2001.

[12] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *ACM SIGKDD '08*, pages 213–220, 2008.

[13] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *HP Laboratories technical report*, 2003.

[14] F. Fouss, A. Pirotte, J. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE TKDE*, 19(3):355–369, 2007.

[15] S. Hill and F. Provost. The myth of the double-blind review? Author identification using only citations. *ACM SIGKDD Explorations*, 5(2):179–184, 2003.

[16] S. Hill, F. Provost, and C. Volinsky. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science*, 22(2):256–276, 2006.

[17] S. Hill, F. Provost, and C. Volinsky. Learning and Inference in Massive Social Networks. In *The 5th Intl. Wkshp. on Mining and Learning with Graphs*, 2007.

[18] G. Jeh and J. Widom. Simrank: a measure of structural-context similarity. In *Proc. of ACM SIGKDD '02*, pages 538–543, 2002.

[19] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity in networks. In *Proc. of ACM SIGKDD '06*, pages 245–255, 2006.

[20] Lotame. Moms in social media. Lotame I.D. Reports, November 2008. http://www.Lotame.com.

[21] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[22] A. Narayanan and V. Shmatikov. De-anonymizing social networks. 2009. To appear at IEEE Security and Privacy '09.

[23] J. C. Perez. Facebook's beacon more intrusive than previously thought. 2007.

[24] F. Provost, P. Melville, and M. Saar-Tsechansky. Data acquisition and cost-effective predictive modeling: targeting offers for electronic commerce. In *Proceedings of ICEC-07*, pages 389–398, 2007.

[25] J. Sun, H. Qu, D. Chakrabarti, and C. Faloutsos. Neighborhood formation and anomaly detection in bipartite graphs. In *Proc. ICDM '05*, 2005.

[26] S. White and P. Smyth. Algorithms for estimating relative importance in networks. In *Proc. of ACM SIGKDD '03*, pages 266–275, 2003.

[27] B. Zhou, J. Pei, and W.-S. Luk. A brief survey on anonymization techniques for privacy preserving publishing of social network data. *ACM SIGKDD Explorations*, 10(2):12–22, 2008.