Display Advertising Impact: Search Lift and Social Influence

P. Papadimitriou[1]   P. Krishnamurthy[2]   R. Lewis[2]   D. Reiley[2]   H. Garcia-Molina[1]

[1]Stanford University
Stanford, CA, USA
{papadimitriou@,hector@cs.}stanford.edu

[2]Yahoo! Labs
Santa Clara, CA, USA
{pkmurthy,ralewis,reiley}@yahoo-inc.com

ABSTRACT

We study the impact of display advertising on user search behavior using a field experiment. In such an experiment, the treatment group users are exposed to some display advertising campaign, while the control group users are not. During the campaign and the post-campaign period we monitor the user search queries and we label them as relevant or irrelevant to the campaign using techniques that leverage the bipartite query-URL click graph. Our results indicate that users who are exposed to the advertising campaign submit 5% to 25% more queries that are relevant to it compared to the unexposed users.

Using the social graph of the experiment users, we also explore how users are affected by their friends who are exposed to ads. Our results indicate that a user with exposed friends is more likely to submit queries relevant to the campaign, as compared to a user without exposed friends. The result is surprising given that the display advertising campaign that we study does not include any incentive for social action, e.g., discount for recommending friends.

1. INTRODUCTION

Display advertising - showing graphical, often interactive, advertisements (ads) on regular web pages - is approximately a $24 billion business [7]. In the simplest form of display advertising an *advertiser* purchases *ad impressions* on a *publisher*'s web page. An ad impression occurs each time a *user* loads the publisher's web page and sees the advertiser's ad as part of the page. To assess the success of display advertising campaigns, advertisers need metrics to estimate the user response. One such metric is the *search lift*, i.e., the ratio of the proportions of search queries related to a campaign submitted by users who were *exposed* as opposed to users who were not exposed to the campaign ads. Apart from direct user response, advertisers also expect their campaigns to impact users through *social influence*. That is, a user who is exposed to the campaign may make other users who were not directly exposed interested in the advertised products, e.g., through word-of-mouth spreading of the ad message. Thus, to get a full picture one needs to measure the response of exposed users as well as the response of friends of the exposed users.

Without reliable estimates of the display advertising impact, it is hard to assess the value of display advertising and set fair prices for both advertisers and publishers (web page owners). For example, if the advertisers over-estimate the possible social influence of their campaigns, they may purchase overpriced ad impressions and lose money. On the contrary, if the advertisers under-estimate the impact of display advertising, then the publishers may lose money by selling under-priced ad impressions. Taking into account the impact on friends also helps advertisers assess the value of an ad impression on a per user basis. For example, if 1 out of 10 friends of an exposed user responds to a display advertising campaign, a single ad impression is expected to have the same impact as an impression to two users without friends.

The key challenge for evaluating the influence of an ad campaign is in linking an exposed user to the desired actions of that or other users. For example, say Alice is exposed to an ad for the online retailer Zappos.com offering a discount on Nike shoes. The ad is displayed while Alice is browsing a Yahoo.com site. After seeing the ad, Alice communicates with Bob say via email, and then Bob performs an action related to the campaign. Ideally, we would like Bob to purchase Nike shoes at Zappos, but having Bob buy anything else at Zappos would be good, or having Bob buy Nike shoes elsewhere would also be good. The problem is that Bob's purchase will probably be made at a site unrelated to the Yahoo! site that displayed the ad, so it is very hard to connect Bob's purchase with Alice's identity in order to "credit" the ad displayed to Alice for influencing the purchase.

For these reasons, it is very hard to use a traditional metric like *purchase lift*, which measures the increase in actual product purchases due to a campaign. Another traditional metric is the click-through rate (CTR) of clickable ads. In this case, the desired action is that a user clicks on the URL presented in the ad. Here again, it is very difficult to connect a click through by Bob to the original ad seen by Alice. Bob is unlikely to see the ad for Nike shoes at Zappos even if he logged on to a Yahoo! site.

For search lift it is possible, though still challenging, to connect Alice's ad impression with Bob's submission of a search query related to Nike or Zappos. To make the connections, Alice and Bob's actions must be performed on systems that have consistent user identities. For instance, we must know that the user submitting the query related to the ad is the same user who communicates via email with Alice, who in turn is the user who saw the ad in the first place. Fortunately, for our work we are able to use "traces" of displayed ads, submitted search queries, and user relationships with consistent user identities, which makes the evaluation of search lift feasible. Of course, our results will be "conservative", in the sense that Alice may influence other users that we do not know about.

As we will discuss, another challenge is the actual definition of

search lift due to social influence. In particular, Bob may have several friends or connections. Some of the friends, like Alice, may have been exposed to an ad from the campaign, and others may not have. If Bob has 4 friends, two which have seen the ad, how do we account for influence? Is this case the same as if Bob has 4 exposed friends out of a total of 8? And how long after Alice saw an ad, do we credit that ad for actions taken by Bob? We address such questions as we develop our evaluation metrics.

Previous works have studied the impact of display advertising on exposed users [12] and independently studied influence (not related to ads) in social networks [2]. To the best of our knowledge our work is the first study of social influence effects of display advertising. There are also several works (e.g., [4]) that study social influence for viral marketing. However in most of these works social influence is the result of some explicit and observable social action, e.g., sending a movie suggestion to 5 friends. In our case the social influence is implicit and, consequently, it is sparse and harder to observe. Despite of the sparsity, such a study is possible nowadays because of the large amount of data that is available to advertising networks.

In this paper we study search lift and social influence in display advertising using a controlled experiment. In such an experiment, real users that visit a publisher's web site are randomly partitioned into a group that is exposed to an ad campaign (treatment group) and a group that is not exposed (control group), and we compare the search behavior of users in the two groups. Prior to measuring search lift, we describe how to determine the queries related to a given advertising campaign. Our methods leverage user clicks in the search results and they yield queries that are relevant to different aspects of the campaign, so that we can measure different aspects of the campaign impact. In our social influence study we compare the search behavior of users in the control group with a varying number of friends in the treatment group. Our results show that the more user friends that are exposed to ads, the more likely the user is to submit queries related to the ad.

The rest of the paper is organized as follows: in Section 2 we present some notation. In Section 3 we present our approach to estimating the search lift in the context of a field experiment. In Section 4 we present our approach to estimating the social influence in the context of a field experiment. In Section 5 we show how to determine which queries are related to a given advertising campaign. In Section 6 we present the experimental results, in Section 7 we present the related work and we conclude in Section 8.

2. NOTATION

A display advertisement (ad) a is typically a static or an interactive image that conveys some advertiser message.

An ad *impression* $V(u, a, w, t)$ is the event that the *user* u views the ad a as part of the web page w at time t.

A *search event* $S(u, q, t)$ denotes that user u submits query q to a search engine at time t.

A *display advertising campaign* is defined as $D = (A, U, W, P)$: a set of ads A, a set of targeted users U, a set of serving web pages W and a campaign period P. Sets U, W and the period P reflect the advertiser's preferences with respect to the impressions of the campaign ads. For example, the advertiser's preferences for the "buy Nike shoes at Zappos" campaign with $A = \{a\}$ can be:

- U: set of all users located in US;
- W: set of web pages under shopping.yahoo.com domain; and
- P: hours 9am-5pm during the days Dec 13 through Dec 17.

Given a campaign D an ad impression $V(u, a, w, t)$ may occur when a user $u \in U$ visits a web page $w \in W$ at time $t \in P$.

3. SEARCH LIFT

In this section we present our approach to the calculation of search lift as the result of a display advertising campaign. In Section 3.1 we briefly discuss the setup of a field experiment. In Section 3.2 we discuss what the queries of interest are for users that participate in the field experiment. Then, we define the *search lift* in Section 3.3 and we discuss some variations in Section 3.4.

3.1 Field experiment

We study the impact of a campaign D on users using a *field experiment*. In such an experiment we partition the targeted users of the campaign U into a *control* group C and a *treatment* group T and we substitute campaign D with the following two campaigns:

- the treatment campaign $D_T = (A, T, W, P)$ that we obtain from D by replacing the targeted users U with T; and
- the control campaign $D_C = (A_C, C, W, P)$ that we obtain from D by replacing the targeted U with C and the campaign ads A with A_C. Each ad $a \in A_C$ should convey no advertiser message, so that it has negligible impact on the users who view it. For example, such an ad on a shopping.yahoo.com web page could have a message about the web page itself.

Then we measure the impact of ads A by comparing the search behaviors of users in the control group C as opposed to users in the treatment group T.

3.2 Susceptible Queries

We focus on user queries that are submitted within some period δ after an ad impression. We refer to the parameter δ as the *influence period* and to the queries the are submitted in the time interval $(t, t + \delta]$ after an ad impression $V(u, a, w, t)$ as the *susceptible queries*.

The influence period δ captures the fact that the impact of an ad impression diminishes as time elapses. In our experiments we present results for different values of δ that range from 10 minutes up to 2 weeks.

Given the influence period δ and the ad impressions of a user u we define the boolean function:

$$\text{susc}\,(S(u,q,t),\delta,D) = \begin{cases} true & \exists t', a, w : a \in A \wedge \\ & w \in W \wedge V(u,a,w,t') \wedge \\ & t \in (t', t' + \delta] \\ false & \text{otherwise.} \end{cases}$$
(1)

The query q is susceptible if there is at least one ad impression $V(u, a, w, t')$ such that the query q is submitted within time δ after the time t'.

Given the susc function we let

$$B(u, \delta, D) = \{q : S(u, q, t) \wedge \text{susc}\,(S(u, q, t), \delta, D)\} \quad (2)$$

denote the *bag*[1] containing all of u's susceptible queries. We denote the bag of queries submitted by any C user as $B(C, \delta, D) = \bigcup_{u \in C} B(u, \delta, D)$ and the bag of queries submitted by any T user as $B(T, \delta, D)$.

Note that the definition of susceptible queries is the same for both the treatment and the control group users. Although the control group users do not see ads from campaign D, we still use Eq. 2 to determine the influenced queries, so that we do not introduce any bias in the selection of queries for the two groups.

---

[1] A *bag* or *multiset* is an object collection whose members need not be distinct (unlike a set, whose members are distinct).

3.3 Search Lift Definition

We estimate the impact of a campaign D by comparing the queries in the bag $B(C, \delta, D)$ with the queries in the bag $B(T, \delta, D)$. To perform such a comparison we introduce the notion of query relevance to a given advertising campaign. Then we extend the relevance notion to bags of queries and we finally use this notion to compare the two bags.

For each query q the value of function $\mathrm{rel}\,(q, D) \in [0, 1]$ shows the relevance of q to the campaign D. If the query is irrelevant then $\mathrm{rel}\,(q, D) = 0$. To illustrate, in the example of the "buy Nike shoes at Zappos" campaign, the query "zappos nike shoes" may have relevance score 1, the query "shoes" may have relevance 0.5 and the query "news" may have relevance 0. Later in Section 5 we present an algorithm for calculating the values of the rel function.

Using the rel function as a parameter we define the *effective proportion* of a query bag B that is relevant to the campaign D as:

$$\mathrm{eff}\,(B, D; \mathrm{rel}) = \frac{\sum_{q \in B} \mathrm{rel}\,(q, D)}{|B|}. \qquad (3)$$

If all of the B queries are relevant to D, then $\mathrm{eff}\,(B, D; rel) = 1$. If half of the B queries are relevant to D then $\mathrm{eff}\,(B, D; rel) = 0.5$.

Finally, we compare the relevance of queries submitted by C and T users using the *search lift* metric that is defined as follows:

$$\mathrm{lift}\,(C, T, \delta, D; \mathrm{rel}) = \frac{\mathrm{eff}\,(B(T, \delta, D), D; \mathrm{rel})}{\mathrm{eff}\,(B(C, \delta, D), D; \mathrm{rel})}. \qquad (4)$$

In other words, the search lift is the ratio of the proportion of the campaign related queries submitted by T users to the proportion of campaign related queries submitted by C users. The lift is expected to be greater than 1, since the T users are expected to submit more queries relevant to the campaign than the C users.

To draw any statistically significant conclusions about the search lift on real data, it is critical to estimate reliable confidence intervals. We elaborate on this issue in Section 6.3.1 of the experiments.

3.4 Impact of Different Campaign Aspects

Campaign ads may convey more than one message to the users. For example, most retail display advertisements include a certain *brand* product or a series of brand products that are available from some *retailer*. Examples of such ads are "buy Nike shoes at Zappos" or "buy Seagate hard drive at Newegg." The advertiser in such cases can be either the brand company or the retailer company or both. In case of the first ad, three possible user searches after an ad view are the following: "nike shoes zappos" or "nike shoes" or "zappos". Note that the second query implies some user intent that will not probably benefit the retailer, while the last query implies some intent that will not probably benefit the brand. Measuring campaign impact with respect to either of the parties is important, since it allows a fair allocation of the campaign cost or it can provide the advertiser with useful insight to improve the campaign. For example, if the advertiser is Zappos and the advertising campaign seems to raise interest mostly in Nike shoes rather than the Zappos itself, then Zappos can redesign the ads of the campaign to decrease the emphasis on Nike shoes.

Since search queries provide rich information about the user intent, search lift can be used to estimate the impact of different aspects of a campaign. Without loss of generality, in this paper we focus on retail advertising. Thus, we define three different types of search lift using different rel function definitions to measure the impact of the different aspects of a campaign:

- *Brand lift* measures the search lift between groups C and T with respect to queries that are relevant to the campaign brand, e.g., Nike or Seagate. In this case rel quantifies how relevant the query is to the brand.
- *Retailer lift* measures the search lift between groups C and T with respect to queries that are relevant to the campaign retailer, e.g., Zappos or Newegg. Similarly, the rel function quantifies how relevant the query is to the retailer.
- *Total lift* measures the search lift between groups C and T with respect to queries that are relevant either to the campaign brand or the campaign retailer. Here the rel function quantifies how relevant the query is to the either the brand or the retailer.

Note that it is hard to define analogous types for other display advertising impact metrics such as the CTR or purchase lift. CTR measures only views of the ad landing page and ignores views of the brand web site or views of retailer pages beyond the landing page. Moreover, metrics like the purchase lift would require monitoring user purchases not only in the campaign retailer store, but also in all retailer stores that carry products of the advertised brand. The cost of such monitoring would be high.

4. SOCIAL INFLUENCE

We study the impact of display advertising taking into account that users are part of a social network. We represent the social network with an undirected graph $G = (U, E)$. Each node $u \in U$ represents a user and each edge $e = (u, u')$ represents that users u and u' are friends. Our goal is to study how user search queries vary depending on the number of friends that have seen ads from campaign D. The motivation for this study is that users may convey advertisement messages to their friends. For example, say that Alice sees ads about Zappos and she thinks of buying a pair of shoes from the online store. Alice may ask the opinion of her friend Bob about the potential purchase and Bob will end up searching online for Zappos and browsing the online catalog to find items of interest.

To quantify the influence by friends exposed to ads we define the *social search lift* similarly to the search lift in Section 4.2. Prior to this definition we redefine the notion of susceptible queries in the context of social influence in Section 4.1

4.1 Susceptible Queries under Social Influence

In case of social influence, we focus on the queries of user u after his friends have been exposed to ads and within some influence period δ. Thus, if u' is a friend of u and there is an ad impression $V(u', a, w, t)$, then we refer to the queries of u during the time interval $(t, t + \delta]$ as the *susceptible queries under social influence* or simply as the susceptible queries. Similarly to Eq. 1, we define a boolean function that determines whether a query is susceptible:

$$\mathrm{susc}_s\,(S(u, q, t), \delta, D, G) = \begin{cases} true & \exists t', a, w, u' : a \in A \wedge \\ & w \in W \wedge \\ & u' \in U \wedge (u, u') \in E \wedge \\ & V(u', a, w, t') \wedge \\ & t \in (t', t' + \delta] \\ false & \text{otherwise} \end{cases}$$

$$(5)$$

The subscript in susc_s denotes that this definition is used in the context of social influence.

Thus, in case of social influence the bag that contains all of the susceptible queries of user u is defined as:

$$B_s(u, \delta, D, G) = \{q : S(u, q, t) \wedge \mathrm{susc}_s\,(S(u, q, t), \delta, D, G)\}$$

$$(6)$$

similarly to Eq. 2 (See Section 4.3 for a discussion of why chose these particular definitions).

4.2 Social Search Lift

In the case of social search lift we focus on users of the control group and divide them into subgroups based on their total number of friends and the number of friends that belong to the treatment group. We denote with $C(d_T/d)$ the set of users who have d_T friends in the treatment group out of a total of d friends:

$$C(d_T/d) = \{u \in C : d_T = |\{(u, u') \in E \wedge u' \in T\}|,$$
$$d = |\{(u, u') \in E\}|\} \quad (7)$$

For example, in Figure 1(a) we show three different control group users (the users on the left of each connected triple) that have two friends. The user on top belongs to $C(0/2)$, because he has two friends belonging to the control group. The user in the middle belongs to the $C(1/2)$ group, because he has one friend in the control and one in the treatment group. Finally, the user at the bottom belongs to $C(2/2)$, because both of his friends belong to the treatment group. Similarly, in Figure 1(b) we present examples of the user subgroups with users that have 3 friends in total.

As we argue in the Section 4.3, in this setting it is interesting to study user searches among groups with the same total of d friends, but a varying number d_T of treatment friends. In particular, we compare the queries of groups with $d_T > 0$ with the queries of users with $d_T = 0$. The social search lift is defined as follows:

$$\text{lift}_s\left(C(d_T/d), C(0, d), \delta, D; \text{rel}\right) = \frac{\text{eff}\left(B_s(C(d_T/d), .), D; \text{rel}\right)}{\text{eff}\left(B_s(C(0/d), .), D; \text{rel}\right)}, \quad (8)$$

where $B_s(C(d_T/d), .) = B_s(C(d_T/d), \delta, D, G)$ is the bag that contains the susceptible queries of all users in the $C(d_T/d)$ group.

4.3 Discussion

In this section we discuss our choices for the definition of the social search lift and discuss the reasoning behind them.

Why do we compare users with the same number of friends? Considering users with the same number of friends d is important to avoid introducing any bias in the selection of groups that we compare. For example, if we compare users with no friends in T versus users that have at least one friend in T, then we end up comparing users that may have zero friends total versus users with at least one friend. Users in such groups are different, and if we observe different search behavior, we cannot attribute these differences only to advertising.

Why are susceptible queries under social influence defined as in Eq. 5? What are the shortcomings of alternative options? We use the following example to illustrate the alternatives and discuss their shortcomings.

EXAMPLE 1. *Say that D is a campaign with a two-day period $P = \{Sunday, Monday\}$, and we wish to calculate the social search lift* lift_s *between $C(0/1) = \{Alice\}$ and $C(1/1) = \{Bob\}$ with $\delta = 1$ day. Alice's only friend X is in the control group, while Bob's only friend Y is in the treatment group. Say that Y sees a treatment ad on Monday morning and, consequently, Bob's Monday queries are considered susceptible. Say also that X sees a control ad Monday morning at the same time as Y. Which of Alice's queries should we consider susceptible? Those that we designate as susceptible will be compared with Bob's susceptible queries to obtain the* lift_s. *Since Alice's friend does not see any treatment ads that may have an impact on her, one option would be to consider all of Alice's queries during P as susceptible. However, this way will compare Bob's Monday queries with Alice's Sunday and Monday queries. The differences between such bags cannot be attributed solely to D, because Sunday queries are different from Monday queries. For instance, the Sunday queries may be more shopping*
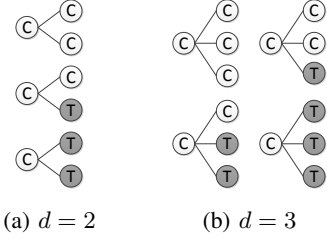

(a) $d = 2$ (b) $d = 3$

Figure 1: Control user subgroups

related compared to the Monday queries. So, with such a definition we would get a biased estimate of the social influence. However, if we use the definition of Eq. 5, we will only consider Alice's Monday queries as susceptible, since they follow X's control ad impression. That comparison will reflect the impact of D.

Note that in a real scenario Alice's and Bob's friends do not see ads exactly at the same time. However, this example illustrates the comparison we will make in our analysis knowing that, in expectation, ad delivery patterns should be statistically identical for randomly assigned control and treatment groups.

In Appendix A we provide another example to justify our definitions for users with friends in both the T and the C group.

5. QUERY RELEVANCE FUNCTION

Calculating search lift requires defining the relevance function rel. Different definitions of this function are required to measure the impact of different aspects of a campaign. There are different ways to define such a function, and in our preliminary experiments we tried some alternative approaches that all yielded similar results. In the following subsections we present a method that leverages user query logs and uses the the random walk algorithm ARW that was introduced by Fuxman et al. [6] for keyword generation in sponsored search. We picked this method because it requires minimum domain specific knowledge for the campaign that we study.

ARW Algorithm. The method uses the *click graph* $G_c(Q, W, E_c)$ which is a bipartite graph whose vertexes are user queries Q and web URLs W. The click graph has an edge (q, w) if there is at least one user who submitted query q and clicked URL w in the search results. The weight $\beta(q, w)$ of every edge is equal to the total number of clicks to w by the users that submitted q.

In addition to the bipartite click graph G_c, a seed of URLs S is required as input to the ARW algorithm. In our case, these URLs must be relevant to the search lift type that we wish to calculate. For example, in our case measuring retailer lift, the seed set may consist of all URLs in the retailer's online domain. The algorithm's output is a vector containing the probability that each query is relevant to the seed set S. We consider the value of $\text{rel}(q, D)$ as equal to the corresponding output probability for the query node q.

At a high level, $\text{rel}(q, D)$ is the probability that a random walker that starts from q in the click graph will end up at a URL in S, assuming that the S URLs are absorbing states in the walk. To prevent long walks, the algorithm assumes that there is a transition probability α that the random walker jumps to an absorbing *null* node from any node in the graph. Finally, the algorithm uses a threshold parameter γ to set all the node probabilities to zero if they are below γ.

Binarization. In Section 6.3.1 we show how we can obtain confidence intervals for the lift, assuming that the rel function is binary. To obtain a binary relevance function $\text{rel}_\theta(q, D) \in \{0, 1\}$ we can use a relevance threshold θ as follows:

$$\text{rel}_\theta(q, D) = \begin{cases} 1, & \text{if } \text{rel}(q, D) \geq \theta \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

6. EXPERIMENTS

In this section we evaluate the impact of display ads on real users using the methodology we have presented so far. The goals of the experiments are the following:

- To illustrate the queries that are relevant to a campaign as returned by the algorithm of Section 5,
- To quantify the search lift between the control and the treatment group,
- To understand how the search lift decreases as the time δ from the ad impressions increases, and
- To quantify the social search lift between users in the control group that have different numbers of friends in the treatment group.

We address these goals in Sections 6.2 to 6.4, after we present the experimental setup in Section 6.1.

6.1 Experimental Setup

6.1.1 Field Experiment

Our data comes from a field experiment with two display advertising campaigns D_1 and D_2 that ran on Yahoo! web pages. We withhold the names of the advertiser, brands, and products in compliance with contractual agreements. The unnamed advertiser in both campaigns is the same popular department store chain; we refer to it as the Retailer. Campaign D_1 advertises clothes and shoes from a luxury brand Brand1 and campaign D_2 advertises clothes and shoes from a popular designer brand Brand2. Both campaigns' advertisements look similar and feature models wearing the advertised clothes along with a message such as "Buy Brand1 clothes at the Retailer" occupying one quadrant of the ad. The campaign period P_1 for D_1 was a week of Spring 2010, and the campaign period P_2 for D_2 was the following week. The sets of serving pages, W_1 and W_2, for both campaigns were identical and they and spanned the Yahoo! network, with a supermajority of the impressions being shown on Yahoo! Mail. Finally, user sets U_1 and U_2 targeted by the ads were also identical and they composed of the same 3 million users.

For each of the two campaigns, we created a control group campaign using *house ads* with a call to action to try one of Yahoo!'s products. The house ads and the retailer's ads were identical in other respects such ad size, position on page, a call-to-action message.

We divided the set of 3 million users into three subsets with 1 million users each: the control group C, the 1st treatment group $T1$, and the 2nd treatment group $T2$. The users in the control group C saw only the house ads in both of the campaigns. The users in the first treatment group $T1$ saw only the retailer ads in both of the campaigns. Finally, users in the second treatment group $T2$ saw the same number of ads as the other two groups, but this group's ads were split evenly between house ads and the retailer's in both campaigns. In the search lift experiments, we will only focus on groups C and $T1$. However, in the social search lift experiments, we take into account both $T1$ and $T2$ friends of the control group users.

6.1.2 User Logs and Social Graph

For all of the 3 million targeted users we have 70,435,009 click-log records that span a period of 8 weeks, starting 3 weeks before the beginning of D_1 and ending 3 weeks after the end of campaign D_2. We refer to each of these 8 weeks as the 1st week, the 2nd week and so on. Campaign D_1 took place during the 4th week, and

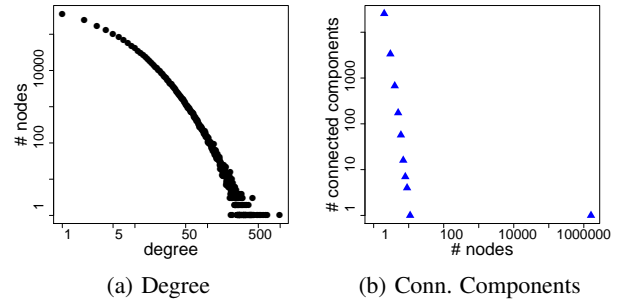(a) Degree   (b) Conn. Components

Figure 2: Social graph distributions.

campaign D_2 took place during the 5th week. Each log record corresponds to one search event $S(u, q, t)$. We also have 62,042,992 impression-log records that span the two weeks of the campaigns D_1 and D_2. Each impression-log record corresponds to an ad-view event $V(u, a, w, t)$.

We constructed a social graph for the 3 million users based on explicit connections that users created via Yahoo!'s social applications such as:

- Messenger, which allows users to exchange Instant Messages with their friends;
- Pulse, which allows users to share their current status and view the status of their friends; and
- Digu, which allows users to post blogs, likes, videos, pictures, etc.

The graph also contains implicit connections between pairs of users that have exchanged email messages through Yahoo! Mail.

The constructed graph has 1,600,962 nodes with at least one connection and 6,694,245 edges in total. The degree distribution follows a power law distribution as shown in the double logarithmic plot in Figure 2(a). The median degree in the graph is 4. The graph is not connected and has 29,793 connected components after removing the singleton nodes. In Figure 2(b) we present the distribution of nodes in these components. Note that despite the big number of components, there is a giant component that contains approximately 3/4 of the graph nodes.

Note that joining the ad impression logs, the query logs and the social graph is possible, because all 3 million users have Yahoo! accounts. If such a user is logged in at some Yahoo! property such as Yahoo! Mail, then a user identifier is passed to Yahoo! servers along with any request for a page from a Yahoo! property. This is a common practice for the web sites[2] with multiple properties and authenticated users. In this way we can record the queries and the ad impression of a particular user. We can also join these records with the social graph, since users participate in the Yahoo! social properties with their Yahoo! accounts.

6.1.3 Implementation

We implemented the ARW algorithm and function eff in Python. In the eff implementation, we first read the ad impression log and we maintain in memory the time intervals for every user when a susceptible query may occur. Then, we read the query log and keep counters for the relevant susceptible queries and all susceptible queries of every user group. The running time of our implementation was dominated by the time to read the log files. We ran our experiments on a Linux Server with Intel Xeon Quad Core with 16 GB RAM.

---

[2]E.g., google.com, bing.com, aol.com

6.2 Relevant queries

We use the algorithm that we presented in Section 5 to determine the queries that are relevant to Brand1, Brand2 and the Retailer. The algorithm input consists of all of the click log records and a seed of relevant URLs for each of the two brands and the retailer. To create these seed for Brand1, we just selected all the URLs under the Brand1 domain. We selected the seeds for Brand2 and the Retailer in a similar fashion. After experimenting with several values, we set the transition probability to the null node $a = 0.01$ and the threshold $\gamma = 0.01$. A manual inspection of the relevance scores computed for the search queries showed that these parameter values provided the most informative results.

In Figure 3 we plot the histograms of query relevance scores with respect to Brand1 (Figure 3(a)), Brand2 (Figure 3(b)) and Retailer (Figure 3(c)). In all of the three figures we have omitted the count of queries with $\text{rel}(q, D) = 0$. In each plot the total area of the bars shows the total number of queries that are relevant to the campaign, i.e., their relevance scores are greater than zero. There are 751 queries that are relevant to Brand1, 541 queries that are relevant to Brand2 and 16,549 queries that are relevant to Retailer.

Note that in all of the plots the number of queries that have relevance probability higher than 0.95 is high. These queries are mostly rare queries that appear once or twice in the click log and have clicks only to the brand or the retailer URLs. Most of the queries that have a relevance score higher that 0.3 usually include a URL from the retailer or the brand web site as the their top result.

In case of brands, the queries with relevance scores that are smaller than 0.2 refer to popular brand products that are carried by many online retailers such as amazon.com or macys.com. The users that search for these products first see the search listings for these retailers and rarely click on the listings for the brand web site. There are more such queries for Brand2, since Brand1 is a luxury brand and few retailers carry its products. In case of the retailer, the queries with relevance scores below 0.2 are not usually strongly related to the retailer. For example, some of these queries were related to products carried by many different stores, including the Retailer.

Based on these results we set the value of θ for the binary version of the relevance function that we introduced in Section 5. In particular we set $\theta = 0^+$ for the relevance functions of the two brands and $\theta = 0.2$ for the relevance function of the retailer. The relevance function for the total lift in either of the campaigns is the maximum of the two corresponding brand and retailer relevance function values. Although we use fixed θ values throughout the remainder of the paper, additional results in Appendix B show that our estimates are robust for a wide range of θ values.

6.3 Search Lift

We evaluate the search lift of the two campaigns using the relevance functions that we presented in Section 6.2. In Section 6.3.1 we present a formula for computing the confidence interval of the search lift. We use this formula throughout the experiments section. In Section 6.3.2 we compare the search behavior of the control and treatment in the pre-campaign weeks to confirm that the the two groups are similar. In Section 6.3.3 we evaluate the different types of lift for the two campaigns and present how the lift changes as the time δ increases. Throughout this section, we refer to group C as the control group and to group $T1$ as the treatment group.

6.3.1 Confidence Intervals

We defined the search lift in Eq. 4 as the ratio of the proportions $\text{eff}(B(T, \delta, D), D; \text{rel})$ and $\text{eff}(B(C, \delta, D), D; \text{rel})$. Search lift is defined similarly to the *relative risk* which is used in statistics and in mathematical epidemiology to assess the risk of an event (or of
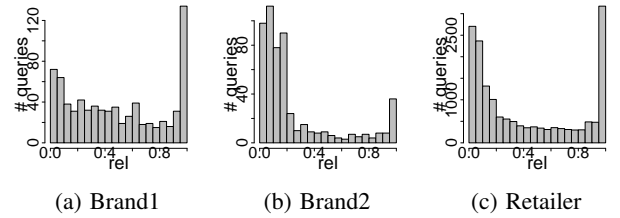

(a) Brand1   (b) Brand2   (c) Retailer

Figure 3: Relevance score histograms

developing a disease) relative to some exposure. The advantage of such a definition is that we can use the approaches developed for the relative risk to obtain a confidence interval for the search lift.

In this paper we use the formula presented in Altman et al. [1], where they show that $\log(\text{lift}(C, T, \delta, D; \text{rel}))$ approximately follows a normal distribution. This formula requires that we use the binary relevance function rel_θ that we discussed in Section 5. If $n_T = |\{q : q \in B(T, \delta, D) \wedge \text{rel}_\theta(q, D) = 1\}|$ is the number of relevant queries in the treatment group bag and n_C is similarly the number of relevant queries in the control group bag, then the standard error of the lift logarithm is:

$$SE = \sqrt{\frac{1}{n_T} - \frac{1}{|B(T, \delta, D)|} + \frac{1}{n_C} - \frac{1}{|B(C, \delta, D)|}}.$$

Thus, the $100(1 - \alpha)\%$ confidence interval for the lift is:

$$[e^{\log(\text{lift}(C,T,\delta,D)) - z^* SE}, e^{\log(\text{lift}(C,T,\delta,D)) + z^* SE}] =$$
$$[\text{lift}(C, T, \delta, D)) e^{-z^* SE}, \text{lift}(C, T, \delta, D)) e^{+z^* SE}], \quad (10)$$

where z^* is the upper $(1 - \alpha)/2$ critical value for the standard normal distribution. In the rest of the experiments, we use $\alpha = 0.05$ to obtain 95% confidence intervals.

6.3.2 Pre-campaign Searches

In this section we explore the search behavior of the control and the treatment group prior to the beginning of the ad campaigns. For this study we use the relevance function for the total search lift that we discussed in Section 6.2. Note that the presented proportions are calculated with respect to all queries, since during the pre-campaign period there are no susceptible queries (there are no ad impressions from D_1 and D_2).

In Figure 4 we present the percentage of relevant queries submitted by both the control and the treatment group during the two weeks before the beginning of the first campaign. The plot on the left focuses on queries that are relevant to the first campaign, i.e., queries relevant to Brand1 and the Retailer, while the plot on the right focuses on queries that are relevant to the second campaign, i.e., queries that are relevant to Brand2 and the Retailer. The red bars in the plots represent the control group and the blue bars represent the treatment group. The bars are indexed by the week (x-
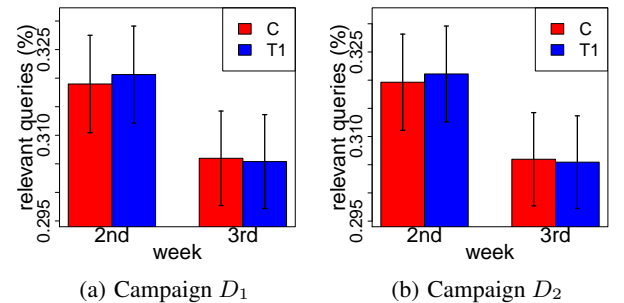

(a) Campaign D_1 (b) Campaign D_2

Figure 4: Pre-campaign searches (D_1 and D_2 took place in the 4th and 5th week, respectively).

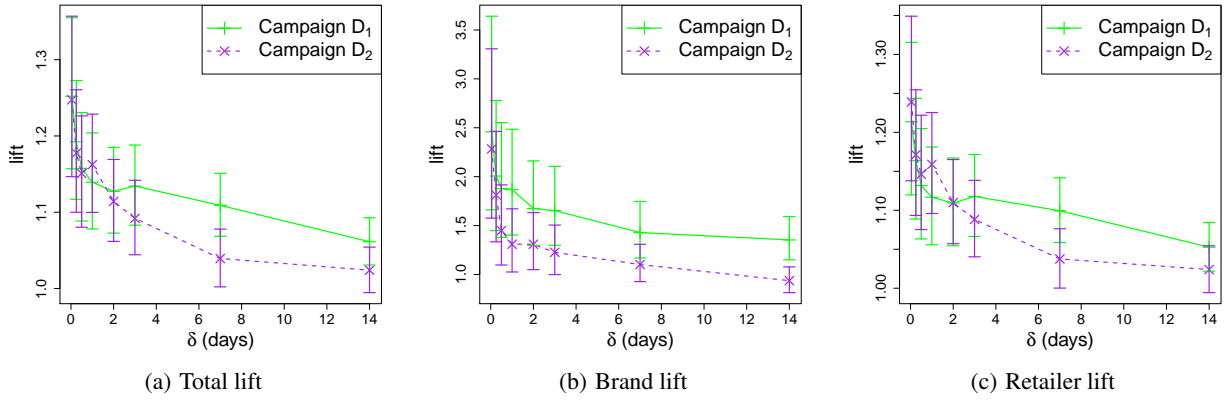|		
(a) Total lift	(b) Brand lift	(c) Retailer lift

Figure 5: Lift as a function of δ.

axis), and the height of each bar shows the percentage of queries that are relevant to the campaign. For example, the left-most red bar on the left plot shows that approximately 0.32% of the queries submitted during the 3rd week by control-group users were relevant to either Brand1 or the Retailer. In each bar we have plotted the 95% confidence interval.

Note that the percentages of the relevant queries submitted by both groups at the same week are very similar and the corresponding 95% confidence intervals significantly overlap. Difference-in-means t-tests fail to reject for each comparison, validating that the randomization of users into the treatment and control groups was successful. This fact shows that in the absence of an advertising campaign the search behavior of the two groups would be similar.

6.3.3 Campaign Effect

In this section we study the search lift as a function of the influence time δ. Our study has two goals: (a) to help estimate the advertising impact by studying the actual values of the search lift and (b) to show the persistence of the advertising effects as the time from the ad impressions grows.

We calculated the search lift for campaigns D_1 and D_2 using the following values for the parameter δ: 1 hour, 12 hours, 1 day, 2 days, 3 days, 1 week and 2 weeks. To calculate the brand lift, the retailer lift and the total lift we used the binary relevance functions that we presented in Section 6.2. In Figure 5(a) we plot the total lift, in Figure 5(b) we plot the brand lift and in Figure 5(c) we plot the retailer lift. There are two lines in each plot, one for each campaign. The x-axis in each plot shows the value of δ. The y-axis shows the lift. The error bars in the plots show the 95% confidence intervals. For example, the right-most point in the solid line of the Figure 5(a) shows that for queries submitted within 14 days after a D_1 ad impression, the total lift is approximately 1.07, i.e., the users in the treatment group $T1$ submitted 7% more queries related to D_1 compared to the users in the control group C. In general, a point that is above the line lift $= 1$ indicates that the advertising campaign has a positive impact on the treatment group.

Let us focus first on the lifts of campaign D_2. Note that all of the three lift curves (total lift, brand lift and retailer lift) are decreasing. This trend was expected since the impact of the advertising campaign should diminish over time. Note also that all of the three curves are convex which means that the lift decrease becomes slower as δ increases. The total lift is as high as 1.25, if we focus on user queries within the first hour and remains above 1 with confidence 95% even whithin a week (7 days). The brand lift drops from 2.3 in the first hour to 1.3 in the first three days. After three days the lift is not greater than 1 with 95% confidence. Finally, the retailer lift drops from 1.12 in the first hour to 1.05 in the first

week. In all three cases note that confidence intervals for small δ values are wide because for such values the number of examined queries is small.

Regarding campaign D_1, the results are similar to those for D_2, but there are some interesting differences. While the D_1 brand lift curve is decreasing and convex, similar to the D_2 curves, there is an uptick in the total lift and retailer lift curves at $\delta = 3$ days. This uptick is due to the fact that the retailer is the same in both campaigns, and there is an overlap of the D_1 post campaign period with the D_2 campaign period. Thus, the susceptible queries that correspond to D_1 ad impressions at the end of the 4th week contain retailer-related queries that are submitted after D_2 ad impressions at the beginning of the 5th week. In these plots, it is noteworthy that the uptick appears only in the total and the retailer lift curves, while the brand lift curve is smooth. This fact confirms that the lift breakdown into retailer and brand lift can successfully illustrate different aspects of the campaign impact.

To compare the two campaigns we can focus on $\delta = 2$ days where the overlap between the campaign influence periods is small. Note that the total lifts have similar values, while the D_1 retailer lift is slightly smaller and the Brand1 lift is higher than the Brand2 lift. The brand lift difference becomes more statistically significant as δ increases and has confidence higher than 95% for $\delta = 2$ weeks (when both campaigns are over). Given that the ads for both the campaigns were very similar, these differences are surprising. However, they clearly indicate that in the context of these campaigns a luxury brand such as Brand1 could raise more user interest compared to a popular brand such as Brand2. Thus, if the advertiser is the retailer, these results can help him choose the brand products that he is going to illustrate in his campaign.

To summarize, our results show that the display advertising campaigns have a statistically significant impact on user search behavior. The impact decreases as the time from the last ad impression increases. Finally, the different lift types may provide interesting insights about the impact of display advertising.

6.4 Social Influence

In this section we study the social search lift between groups of users in the control group. To compute the social search lift, recall that we need to compare the queries of the groups $C(d_T/d)$ and $C(0/d)$. The users in these groups have the same number of friends d, but the users in $C(d_T/d)$ have d_T friends in the treatment group, while the the users $C(0/d)$ have no friends in the treatment group.

Because of the power-law degree distribution of the social graph, we can obtain reliable estimates of the social search lift only for small values of d such as 1,2 and 3. As d increases, not only the number of nodes that have degree d decreases, but also these
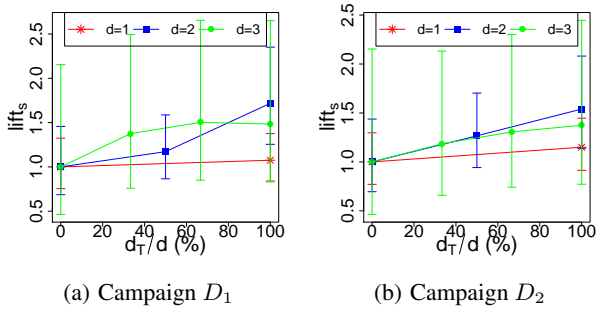
(a) Campaign D_1 (b) Campaign D_2

Figure 6: Social search lift vs T friends percentage ($\delta = 1$ day).

nodes are further divided into $d + 1$ groups with different number of friends d_T in the treatment group. Thus, $C(d_T/d)$ for larger d values is too small to obtain a tight confidence interval. However, note that due to the power law the the users with 1,2 and 3 friends account for more than half of total number of users.

In the results that we present below we refer to users that belong to both $T1$ and $T2$ as the treatment group users. Recall that the users in $T2$ are exposed to the retailer ads, but they see on average half of the retailer ads that the $T1$ users see. Unlike our study on the plain search lift, in this section we study interactions between users in the social graph and we cannot ignore the impact of $T2$ users on their friends.

We calculate the social lift using Eq. 4.2 and the binary relevance functions presented in Section 6.2. In Figure 6 we present the social search lift results for campaigns D_1 (Figure 6(a)) and D_2 (Figure 6(b)) using $\delta = 1$ day as the influence period. In each plot there are three curves and each curve looks at a different number d of total friends. The x-axis show the number of friends in the treatment group as a percentage of the total friends and the y-axis shows the value of the social search lift. The error bars in the curves show the 95% confidence intervals. For example, in Figure 6(a) the middle point of the $d = 2$ line shows that the social search lift is approximately 1.2. In other words, users in the $C(1/2)$ group submit 20% more queries relevant to the campaign compared to $C(0/2)$ users.

Since the confidence intervals are wide in all of the cases, we will not try to interpret the absolute values of the plot, but we will focus instead on some qualitative observations. Note that all curves are increasing. This shows that the number of friends in the treatment group is positively correlated with the percentage of queries that are relevant to the ad campaign. This result becomes more statistically significant as the number of friends in the treatment group increases. For example, the confidence intervals for the social search lift of $C(2/2)$ lies above the lift $= 1$ line in both plots. Note also that, as we discussed above, the confidence intervals become wider as the number of total friends increases.

In other experiments that we do not report in this paper, we estimated the social search lift for varying values of δ. In many cases, the results indicate that the social search lift reaches its maximum for $\delta = 12$ hours, before it starts decreasing. Although this result follows intuition, the evidence that we have from the data is not statistically significant. Thus, this claim requires further exploration.

7. RELATED WORK

Work on advertising effectiveness has blossomed with the advent of internet marketing. Recent work has made use of randomized field experiments [12, 13, 11, 8]. Much of the work has focused on measuring the impact of display advertising on clicks [11], account sign-ups [11, 13] and sales [12] using large-scale field experiments. Goldgard et al. [8] focus on identifying ad properties that increase advertising impact through traditional user surveys.

Unlike previous works [3, 5], in our efforts to measure the impact of display advertising on online search behavior, we have made use of a randomized field experiment in order to avoid spurious conclusions caused by activity bias. Fulgoni et al. [5] present data about the search lift and purchase lift as collected by Comscore[3]. However, their observations should be treated with caution, since they are using observational data to simulate a field experiment. Chan et al. [3] use search lift to estimate the advertising impact, but their focus is on how to use observational data coupled with statistical modeling to simulate field experiments. While Lewis et al. [14] show the impact of a display advertising campaign on search in one of their examples in the context of a field experiment, their analysis was limited to the overall impact of a single-day campaign on a list of advertiser-selected, brand-relevant keywords. Their estimate of a 6% search lift is similar to our estimates (in this limited setting).

We have advanced the aforementioned literature in search lift [3, 5, 14] by introducing a computational method of determining search query relevance in combination with ad effectiveness. The contributions of the method are found in its following abilities: (a) it can assess relevancy, even for queries in the long tail generated in response to exposure to an advertiser's campaign, (b) it can assess instantaneous or persistent effects of display advertising through the influence period δ, and (c) it can identify different aspects of the advertising impact. Additionally, this paper represents a first step in measuring the social influence of display advertising. As such, this paper represents a unique contribution to the advertising effectiveness literature regarding the impact of display advertising on retail consumer search. Finally, our study on display advertising social impact is also unique compared to studies of social influence in other application domains [4, 9, 15, 10], since in the case of display advertising there is not an obervable user action that can trigger social influence.

8. CONCLUSIONS

Our study of display advertising and search lift shows that ad impressions do influence user search behavior. The impact is stronger during the first hour after an ad impression and it diminishes as time elapses. However, even a week after an ad impression exposed users search more about the advertised products compared to unexposed users.

Search lift proved also very effective in identifying the impact of different aspects of the campaign. Using search lift we were able to identify that in two very similar retail advertising campaigns there were different aspects that elicited user interest. In particular, in the first campaign users showed more interest in the advertised products brand, while in the second campaign users got more interested in the retailer store that carries the advertised products. Such insights about the performance of display advertising are very important for the design and the evaluation of advertising campaigns.

Finally, our results show that exposed users do convey advertising messages to their friends who end up submitting queries related to the display ads. This result can help advertisers and publishers to obtain a more thorough picture about the impact of display advertising.

A promising direction for future work is the design of a system that delivers ad impressions taking into account the social influence. In such a system ads will be more likely to be shown to users that do not have exposed friends rather than users who may have already been influenced by their social contacts. Given that publishers typically have a budget for ad impressions, such a system could improve the effectiveness of display advertising.

---

[3]http://www.comscore.com

9. REFERENCES

[1] D. Altman, D. Machin, T. Bryant, and S. Gardner. *Statistics with Confidence: Confidence Intervals and Statistical Guidelines*. BMJ Books, 2nd edition edition, 2000.

[2] A. Bagherjeiran and R. Parekh. Combining behavioral and social network data for online advertising. In *ICDM Workshops*, pages 837–846, 2008.

[3] D. Chan, R. Ge, O. Gershony, T. Hesterberg, and D. Lambert. Evaluating online ad campaigns in a pipeline: causal models at scale. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 7–16, New York, NY, USA, 2010. ACM.

[4] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66, New York, NY, USA, 2001. ACM.

[5] G. M. Fulgoni and M. P. Morn. How online advertising works: Whither the click?, 2008. Online report.

[6] A. Fuxman, P. Tsaparas, K. Achan, and R. Agrawal. Using the wisdom of the crowds for keyword generation. In *WWW*, pages 61–70, 2008.

[7] A. Ghosh, P. McAfee, K. Papineni, and S. Vassilvitskii. Bidding for representative allocations for display advertising. In S. Leonardi, editor, *Internet and Network Economics*, volume 5929 of *Lecture Notes in Computer Science*, pages 208–219. Springer Berlin / Heidelberg, 2009.

[8] A. Goldfarb and C. Tucker. Online display advertising: Targeting and obtrusiveness. *Forthcoming MARKETING SCIENCE*, 2011.

[9] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, New York, NY, USA, 2003. ACM.

[10] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1, May 2007.

[11] R. Lewis. *Where's the "Wear-Out?": Online Display Ads and the Impact of Frequency*. PhD thesis, MIT Dept of Economics, 2010.

[12] R. Lewis and D. Reiley. Does retail advertising work: Measuring the effects of advertising on sales via a controlled experiment on Yahoo!, 2010. Working paper.

[13] R. Lewis and T. Schreiner. *Can Online Display Advertising Attract New Customers?* PhD thesis, MIT Dept of Economics, 2010.

[14] R. A. Lewis, J. M. Rao, and D. H. Reiley. Here, There, Everywhere: Correlated Online Behaviors Can Lead to Overestimates of the Effects of Advertising. In *To appear in WWW'11*, 2011.

[15] J. E. Phelps, R. Lewis, L. Mobilio, D. Perry, and N. Raman. Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email. *Journal of Advertising Research*, 44(04):333–348, 2004.

APPENDIX

A. ADDITIONAL DISCUSSION ON SOCIAL SEARCH LIFT

EXAMPLE 2. *Using the same campaign D as in the previous example, say that Alice has two friends $X_1, X_2 \in C$ and Bob has two friends $Y_1 \in C$ and $Y_2 \in T$. Say that X_1 and Y_1 see ads at the same time on Sunday morning and users X_2 and Y_2 see ads at the same time on Monday morning. Note that X_1, X_2 and Y_1 see control ads, while Y_2 sees a treatment ad. In this setting we wish to calculate the social search lift $lift_s$ between $C(0/2) = \{Alice\}$ and $C(1/2) = \{Bob\}$. Using similar arguments as in Example 1, ideally we should consider Alice's queries after X_2's ad impression as susceptible and Bob's queries after Y_2's ad impression as susceptible. We picked Y_2's ad impression for Bob, because the displayed ad is a treatment ad that may have some impact on Bob. Then, using a symmetry argument we picked X_2's ad impression for Alice, to make Bob's and Alice's susceptible query bags comparable. Although such selection of susceptible query bags is ideal, it is not feasible in practice, because it rare to find similar symmetrical cases in real data. Thus, in real data it is hard to pick which of the users X_1 or X_2 is the analog of user Y_2.*

Using the definition of Eq. 5, we overcome this selection problem by considering Alice's queries after the ad impressions of both X_1 and X_2 as susceptible. Similarly, we consider Bob's queries after the ad impressions of both Y_1 and Y_2 as susceptible. This approach is feasible in a real setting, because it does not rely on symmetrical pairs of users. However, it does include queries in Bob's susceptible query bag that are not affected by the campaign, i.e., queries submitted after a control user ad impression. So our feasible approach provides a conservative estimate for the actual social influence; if were able to measure the social influence in the ideal setting, where only socially susceptible searches across the treatment and control group are compared, we would expect to find larger search lifts in percentage terms. Yet, in terms of the absolute number of searches, our estimates will still be unbiased.
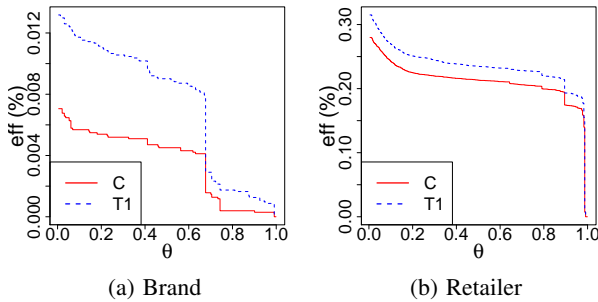
B. LIFT SENSITIVITY TO QUERY RELEVANCE

(a) Brand   (b) Retailer

Figure 7: Effective proportions for campaign D_1 as a function of θ.

In this section we study how the effective proportions of relevant queries are affected by different choices for the θ of the binary relevance function rel_θ (Section 5).

In particular, we vary θ from 0^+ to 1 and we use the rel_θ function that arises to compute the eff proportions that correspond to the brand- and the retailer-related queries. Note that we excluded the

eff values for $\theta = 0$, since for $\theta = 0$ every query is considered as relevant and the eff values are 1 for both groups.

In Figure 7 we present the percentages of relevant queries as a function of θ for campaign D_1. For all of the calculations we used $\delta = 1$ day. The plots for campaign D_2 and different values of δ are similar and we do not present them here. Each plot has two lines. The red solid line looks at the control group C and the blue dashed line looks at the treatment group $T1$. The point $(0.7, 0.008)$ of the dashed line in Figure 7(a) shows that if $\theta = 0.7$, then 0.008% of the queries submitted by the control group are relevant to Brand1.

From these plots we can draw the following two conclusions. First, the difference between the treatment and the control group search behavior can be observed for a wide range of θ values. Thus, the results that we have presented are not specific to only some particular relevance functions. Second, both plots support our choices for the θ values that we presented in Section 6.2. In particular, in the brand plot the difference between the solid and dashed curve is increasing as θ moves leftwards to smaller values. In other words in any interval $[\theta', \theta' + \Delta\theta']$ with $\theta' > 0$, the treatment group users submit more queries with $\mathrm{rel}(q, D) \in [\theta', \theta' + \Delta\theta']$ compared to the control group. This fact indicates that in the brand relevance function all queries with $\mathrm{rel}(q, D) > 0$ should be considered as relevant to the campaign brand. However, in the retailer plot note that the distance between the curves is decreasing for values of θ lower than 0.2. Assuming the the ads do have an impact on the users who see them, this is an indication that the queries with $\mathrm{rel}(q, D) < 0.2$ (here rel here stands for the retailer relevance function) should be considered as irrelevant to the campaign retailer.