Finding similar mobile consumers with a privacy-friendly geosocial design

INFORMATION SYSTEMS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 1047-7047 | EISSN 1526-5536 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Finding Mobile Consumers with a Privacy-Friendly Geo-Similarity Network*

Foster Provost
Department of Information, Operations and Management Sciences, Leonard N. Stern School of Business, New York University, NY fprovost@stern.nyu.edu

David Martens
Department of Engineering Management, Faculty of Applied Economics, University of Antwerp, Belgium
David.Martens@uantwerp.be

Alan Murray
Coriolis Labs, New York, NY

This paper focuses on finding the same and similar users based on location-visitation data in a mobile environment. We propose a new design that uses consumer location data from mobile devices (smart phones, smart pads, laptops, etc.) to build a "geo-similarity network" among users. The geo-similarity network (GSN) could be used for a variety of analytics-driven applications, such as targeting advertisements to the same user on different devices or to users with similar tastes, and improving online interactions by selecting users with similar tastes. The basic idea is that two devices are similar, and thereby connected in the geo-similarity network, when they share at least one visited location. They are more similar as they visit more shared locations, and as the locations they share are visited by fewer people. This paper first introduces the main ideas and ties them to theory and related work. Next we introduce a specific design for selecting entities with similar location distributions, the results of which are shown using real mobile location data across seven ad exchanges. We focus on two high-level questions: (1) does geo-similarity allow us to find different entities corresponding to the same individual, for example as seen through different bidding systems? And (2) do entities linked by similarities in local mobile behavior show similar interests, as measured by visits to particular publishers? The results show positive results for both. Specifically, for (1) even with the data sample's limited observability, 70-80% of the time the same individual is connected to herself in the GSN. For (2), the GSN neighbors of visitors to a wide variety of publishers are substantially more likely also to visit those same publishers. Highly similar GSN neighbors show very substantial lift.

Key words: Design science, Mobile computing, Analytical modeling, Network analysis

1. Introduction

This design-science paper is about finding the same and similar users based on location-visitation data in a mobile environment. A *user* is an *instance* of an individual as viewed through some

* This research was conducted while Foster Provost was at Coriolis Labs.

information system(s), and to avoid confusion we sometimes will explicitly refer to a *user instance*. For example, an individual would be viewed as two different user instances if the individual was interacting with the online advertising ecosystem on two different devices or was viewed on the same device through two different bidding systems.

Specifically, we investigate "geo-similarity"—the similarity of instances of consumers based on the distribution of the locations they have been observed to visit. The basic idea is that two users are similar, and thereby connected in a Geo-Similarity Network (GSN), when they share at least one visited location. Users are more similar as they visit more shared locations and locations that are less frequently visited.[1] This design is motivated by research showing that geographical co-occurrence provides a strong indication of being friends (Crandall et al. 2010, Cho et al. 2011) and influences each other's purchasing behavior (Pan et al. 2011). de Montjoye et al. (2013) found that human mobility traces are highly discriminative: based on the analysis of 15 months of location data for 1.5 million users, they find that four locations are sufficient to accurately identify 95% of the users. This not only motivates our design, it also places emphasis on the desirability of privacy-friendly designs, which we discuss further below.

This paper's motivating application is mobile advertising. As with traditional display ad targeting, mobile ad targeting could be based on context, demographics or psychographics, if such data are available. As such data are largely unavailable or unreliable for mobile consumers in the advertising ecosystem, ad targeting firms increasingly are looking to other means of finding suitable candidates for ads. An alternative method that has been gaining traction for (non-mobile) ad targeting both in research and in practice is targeting based on direct or indirect connections to specific individuals, which we loosely call "social" targeting, as we discuss in detail below. For example, brands often want to target individuals who are similar to particular existing customers of the brand or product in question, or the same existing customers on different devices.[2] Geo-similarity provides a viable alternative for such indirect social targeting for mobile advertising, where (anonymized) location data are available, and importantly anonymized location data may be much more readily available than other targeting data.

No matter how users are chosen for targeting, mobile advertisers need to deal effectively with the problem of *consumer fragmentation*: individuals are observed only through the multiple (baroque)

[1] And as we discuss in future directions, the network perspective could allow similarity judgments also between users who are indirectly connected in the network.

[2] Note that this paper is not about the ultimate effectiveness of the brand's choice of whom to target—it may be that targeting the selected users on different devices or targeting users with similar interests is not appropriate from a marketing standpoint. This is a question specific to each brand and the brand's campaign goals (e.g., direct marketing versus brand advertising). Furthermore, this question is intricately intertwined with the design of the creative being delivered, which also is outside the scope of this paper.

information systems that comprise the digital advertising ecosystem, and a particular individual may correspond to various user instances (Kerho 2012). For example, an individual may have different instances because she is observed on different devices, such as her mobile phone, tablet, laptop, and PC. Individuals may also have different instances because they appear in different advertisement bidding systems, each of which presents the individual differently. In many situations these instances are not associated with a unique identifier for the individual, unlike the cookies that are used to represent a user-browser pair in most work on individualized desktop ad targeting. Moreover, as we will discuss presently, there are important reasons why we might prefer that the instances not be associated with a personal identifier. The bottom line is that once an instance of an individual is identified as being a good target, advertisers would like to be able also to target other instances of the same individual (as well as individuals with very similar interests).

Geo-similarity could also be used for privacy-friendly "hyperlocal" targeting—meaning, targeting people in a precise location (statistically speaking), without needing to store data on the actual locations of the users. For example, consider a hyperlocal coupon campaign: a couponing company wants to target special offers for the local restaurant on the corner. The best prospects are those people who frequent this precise area. If the restaurant can provide some (anonymized) "seed" users—for example, existing clients of the particular local business (e.g., identified via an online loyalty program)—the geo-similarity neighbors of these seed users may have a high probability of also frequenting the same precise locations, and thereby be very good prospects for the hyperlocal coupon.

Looking from a different perspective, we have seen the sort of uproar that arises from the idea that our location behavior is being "tracked" by our mobile technology.[3] Therefore, marketers who dream of location-driven targeting should think carefully about what the FTC calls "privacy by design,"[4] and consider what options can provide effective advertising with minimal data collection and storage. As will be detailed in Section 4, we explicitly take this desideratum into account by anonymizing both the device and location identifiers. This method of using "doubly anonymized" data for privacy friendliness is described in more detail elsewhere (Provost et al. 2009).

In sum, geo-similarity can be used for composing a mobile audience for targeting as follows. Based on a doubly anonymized geo-similarity network, find individuals who are closely linked to individuals we know already to have the characteristic(s) that we desire, such as prior purchase history (as with "retargeting"), brand affinity (e.g., via visiting a web page, "liking" a brand or product, or clicking on an ad), key demographics, or even based on sophisticated targeting

[3] http://pogue.blogs.nytimes.com/2011/04/28/wrapping-up-the-apple-location-brouhaha/

[4] http://www.ftc.gov/opa/2010/12/privacyreport.shtm

designs (Bampo et al. 2008, Agarwal et al. 2008, Heidemann et al. 2010, Perlich et al. 2014). Once these key "seed" users have been identified, targeting close neighbors in the geo-similarity network will tend to target both users with similar interests and users who are other instances of the same individual. Below we present theoretical justification for such an approach, tying it in to related research. We then investigate empirically whether the method indeed tends to target (i) other instances of the same individual and/or (ii) individuals with similar interests. The empirical study is based on data drawn from actual real-time bidding (RTB) exchanges for mobile advertising.

Before we move on to related work, please let us note that our proposed Geo-Similarity Network is different from Geo-Social-Networking, which is defined as a type of social networking in which geographic services are used to enable additional social dynamics (Quercia et al. 2010). The latter starts from an actual network of interpersonal relationships, and adds location data to recommend other locations and events. In contrast, the GSN uses location data to create links between users, without the explicit *guarantee* that these links correspond to an actual interpersonal relationship (however this surely might be, cf. Crandall et al. (2010), as described in the next section).

The next section argues theoretically why this GSN design *should* be a good idea. After that we present the high-level GSN design in more detail, followed by specifics of the implementation we examine empirically. Then we provide results from an empirical study based on real mobile location data from real-time bidding exchanges. Section 5 describes the data and experimental setup, including several technical definitions for geo-similarity. The ability of the GSN to connect different instances of the same individual is evaluated in Section 6. The use of the GSN to select users with similar interests and tastes is assessed in Section 7.

2. Motivation and Related Work

2.1. Social Targeting

Advertising targeting has evolved substantially over the past half century. As information systems provided access to new sources and types of data, marketers added new targeting strategies designed around the new data. For example, as demographic data became available a few decades ago, contextual targeting—targeting based on inferring audience composition from the context in which the ad will be shown (e.g., a billboard location, tv show, magazine, etc.)—had to share the spotlight with data-driven demographic targeting, either based on explicit demographic profiles or based on predictive modeling. As data aggregators coalesced and integrated information such as magazine subscriptions and catalog purchases, "psychographic" data entered the mix, and broadened yet again the space of targeting designs.

Recently, we have seen the introduction of a different sort of targeting design, which we can generally call *social targeting.* Social targeting differs from the aforementioned targeting methods

because it relies on explicit linkages between specific individuals. For example, Hill et al. (2006) showed the remarkable effectiveness of *social-network targeting*: targeting consumers who are linked to known customers by a social network Sundararajan et al. (2013). Subsequently, Facebook (and others) have attempted to implement social-network targeting for online advertising, with varying degrees of success (see e.g., Oinas-Kukkonen et al. (2010)).

We explicitly generalize from social-network targeting to social targeting, in order to retain the notion that the targeting is based on linkages to specific, other individuals, but to relax the notion that the linkages need to be "true" interpersonal relationships. The "quasi-social" design of Provost et al. (2009) is an example of social targeting that is not based on "true" interpersonal relationships: the linkages between individuals are based on a bipartite content-affinity network. So the social targeting there is based on forming an audience by finding consumers who are linked by shared content visitation with other specific consumers who are known to have brand affinity (more on that later). Similarly, Martens and Provost (2011) define a network among banking customers based on payment transaction data, which is subsequently used to target marketing offers for financial products.

2.2. Geo-Similarity

The geo-similarity network can be very fine-grained, based on shared (anonymized) location data, for example, IP addresses, fine-grained latitude/longitude cells, or small geographic tracts. Why would targeting geo-similarity network neighbors of pre-selected seed users be a good idea? There are several reasons. Let's call mobile devices that are (directly) linked in the geo-similarity network "(first-degree) network neighbors."

First of all, first-degree network neighbors share at least one location, and possibly several. As the number of shared locations grows, we conjecture that the likelihood increases that the two devices actually belong to the same person. For example, who besides me is observed primarily on my home IP address and my office IP address, let alone my favorite coffee shop. We see analogous results showing that different instances of the same person call the same phone numbers (Cortes et al. 2001) and cite the same references (Hill and Provost 2003).

Second, direct, coarse-grained geographic targeting already is used widely in offline advertising (not via social targeting), because geography is a reasonable proxy for demographics and other predictive features. An important difference in the present work is that we don't choose the geographies to target explicitly, but instead use them implicitly. This allows the actual locations to be anonymized. Furthermore, it allows us to use location information that is too fine-grained even to include in most predictive modeling—for example, locations appearing only in a tiny fraction of instances (e.g., a home wifi address may only appear in 0.000001% or fewer of the users' location sets), as well as transient wifi locations that only connect two devices for a brief time.

Third, fine-grained location information is likely to contain more detailed (latent) information than standard geographic information. Not only would it link devices by approximate wealth, income, demographics, etc., it may well link by employer, educational institution, interests, community, and even shopping habits. Thus, we conjecture that geo-similarity targeting may combine the advantages of geographic targeting discussed in the previous paragraph, with advantages similar to content-affinity social targeting, which has been shown to be effective specifically for online advertising (Provost et al. 2009, Perlich et al. 2014). We might call this "locale-affinity" social targeting. The empirical results below show that indeed one's geo-similarity neighbors are much more likely to frequent the same online publishers and mobile apps.

If that weren't enough, research provides yet another reason to expect that geo-similarity targeting may be especially effective. In a 2010 article in *PNAS*, Crandall et al. show that geographic co-occurrences between individuals are very strongly predictive of the individuals being friends: "The knowledge that two people were proximate at just a few distinct locations at roughly the same times can indicate a high conditional probability that they are directly linked in the underlying social network" (Crandall et al. 2010). This means that a geo-similarity network not only would capture the advantages of geographic targeting and locale-affinity targeting, it may also incorporate actual social-network targeting—also shown to be extremely effective for marketing (Hill et al. 2006). In fact, when massive descriptive data are available, the (usually latent) similarity between social-network neighbors has been shown to explain much of the marketing advantage previously attributed to social influence (Aral et al. 2009).

Moreover, interestingly, research also has shown that the homophily (McPherson et al. 2001) that has been used to explain the effectiveness of social-network targeting, may actually be due largely to the constraints placed by opportunity (Kossinets and Watts 2009). Tie formation in social networks is biased heavily by triadic closure, and thus by structural proximity. Over many generations of tie formation, biases in the selection of structurally proximate individuals "can amplify even a modest preference for similar others, via a cumulative advantage-like process, to produce striking patterns of observed homophily" (Kossinets and Watts 2009). Why is this important for the present paper? Because with the exception of social links formed solely online (like new Facebook-only friends), constraints on opportunity are framed by constraints on physical co-location. As Kossinets and Watts (2009) observe "an individual's choice of relations is heavily constrained by other aspects of his or her life, such as geographical location, choice of occupation, place of work." These are exactly the sorts of links that would be represented in a geo-similarity network based on mobile device location data (including tablets and laptops). Thus, the geo-similarity network may in addition reveal a large driver of homophily, and thus expand the effectiveness of a social-network targeting strategy to individuals who also would have been similar "friends," but for whatever reason were not chosen.

3. Geo-similarity, geo-similarity networks, and mobile ad targeting

The basis for the targeting design is to create a network among users based on geo-similarity, and then to use the network for inference. For this paper, we will create a geo-similarity network directly among user instances.[5] The elements of the network-targeting design include: how will entities (mobile user instances) be represented? What exactly will be the geo-similarity links? And, how exactly will the predictive inference be conducted?

In this section we will discuss the design at a high level, including some further-looking elements that are not part of the experiments, but which provide completeness for the discerning reader. In order to make the discussion more concrete, we first present the motivating data scenario. The design should generalize to mobile settings beyond this specific data scenario.

3.1. The Mobile Real-Time Bidding Ecosystem

A high-level overview of how an RTB exchange works is given in Figure 1. RTB exchanges bring together advertisers and the publishers of web pages and apps. When a user visits a web page/app (hereafter, "web page") that sells ad slots through RTB exchanges, a *bid request* is created. Such a bid request will provide additional data on the user and web page visited, such as IP address, publisher and device type (see Table 1 for a list of fields and statistics on their availability across RTBs). Based on such data, advertisers can choose to bid. For example a brand may want to target users that previously visited their automobile website, browsing for a eco-friendly hybrid car. The highest bidder is allowed to load a potentially personalized ad, for example showing an ad for an eco-friendly hybrid. All this occurs in real-time, in less than 100 milliseconds. The two interrelated problems this paper addresses can be stated specifically in the context of this example: (i) How does the targeter find instances of this individual who previously visited the brand's website? (ii) Can the targeter find users with very similar interests (e.g., the focal user's husband, best friend), under the presumption that they might also be reasonable candidates for an ad for an eco-friendly hybrid?

For this paper, we address the data currently available in the digital advertising ecosystem. Specifically, via RTB exchanges we can observe a massive number of mobile users, along with the data made available for advertisers to decide whether to bid on the user for a particular campaign (see e.g. Pubmatic (2010)). Two aspects of the data are crucial: (1) each instance of a mobile user is associated with a key; when this key is observed in the future, we know that the future data item involves the same mobile device.[6] (2) Part of the data associated with each mobile device

[5] An alternative is to create a bipartite network between users and locations and draw inferences directly from the bipartite network. We do not investigate that explicitly in this paper; however, the interested reader should take a careful look at the similarity calculations we present below.

[6] The notion of an instance is important. Many instances of the same device may have different keys in the online advertising ecosystem. As discussed above, connecting these is one focus of this research.

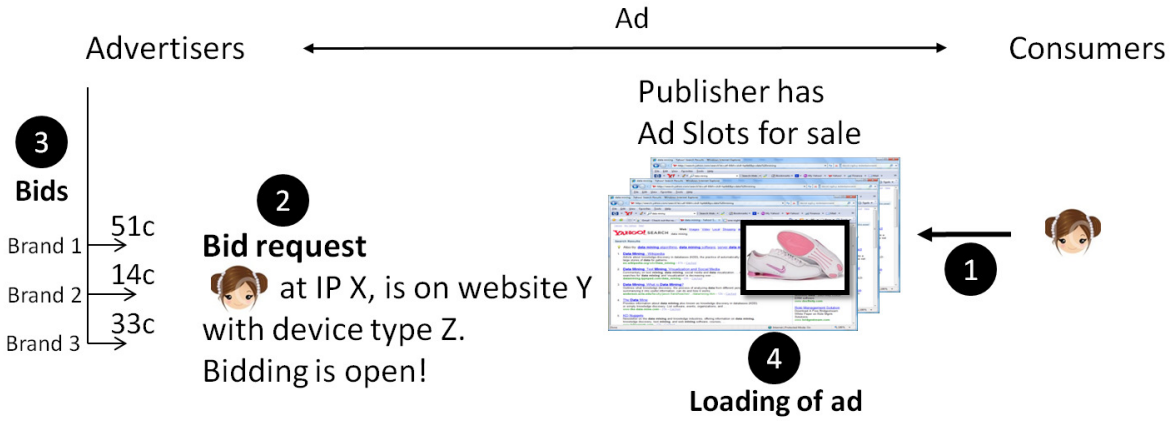


Figure 1 **Working of Real-Time Bidding: when a user visits a webpage that sells ad slots through RTB (1), a bid request is created (2), after which several advertisers bid to allow to show an ad at that moment on that webpage to that user (3). The winning bidder is allowed to load a potentially personalized ad (4).**

is a location. For this paper, we consider that location to be the (anonymized) IP address of the wifi network[7] currently in use (although there are various sorts of location data). We observe the locations visited by each user for the data records included in the research data set (described in more detail in Section 5.1).

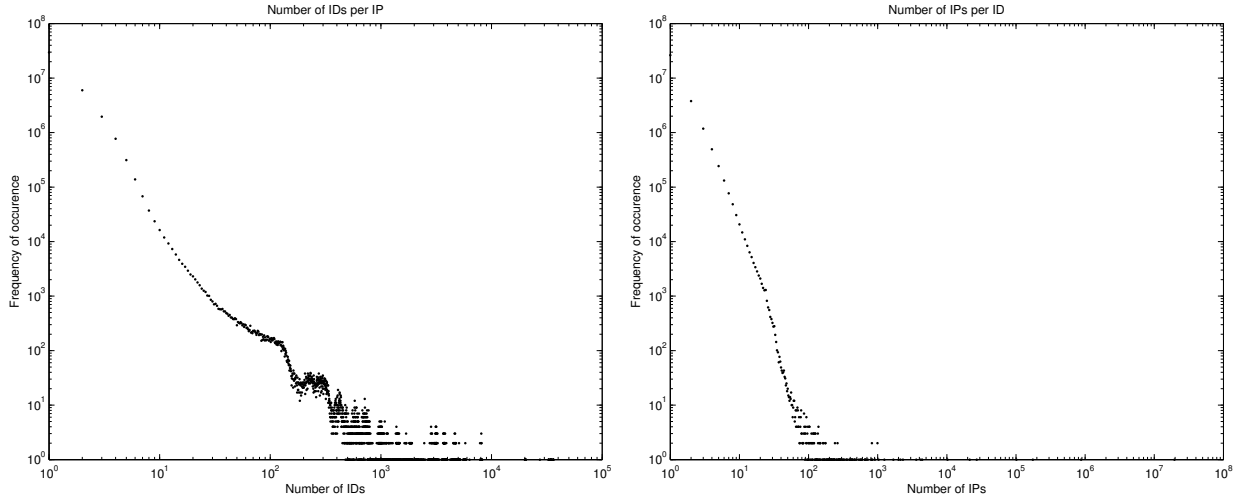
The mobile RTB ecosystem holds different information than for the typical online (desktop) world, which is important for our design. The GSN design measures similarity in location visitation behavior. One might wonder if other behavioral data can be obtained from the RTB systems, such as website visitation data (Provost et al. 2009, Raeder et al. 2012). Table 1 shows which typical fields (besides the key) are visible on mobile devices over seven different RTB exchanges. Let us consider the different data fields as possible sources of behavioral data.

First, let's have a look at the fields that are available everywhere: the time and date of the bid request (*created*), the RTB exchange (*network*), dimensions of the ad (*dims*), device type (*device*), *user agent*, *publisher* and *IP* address. The IP address (anonymized) is the location we use in our GSN design. Figures 2(a) and 2(b) show the distribution of the number of users seen at an IP, and the number of IPs seen per user respectively (more information on this dataset is provided later, in Section 5.1). On average a user instance visits 1.66 IPs; an IP is visited on average by 2 users (over a time period of 10 days, see further). Although in our dataset some known 3G IP (subranges) have already been removed, Figure 2(a) shows that high-volume IP addresses still are present. Some of

[7] For this paper, some transient IP addresses are removed based on known IP ranges (such as those belonging to 3G or other carrier networks), which are incomplete. We consider transient IP addresses that are not explicitly removed to be noise in the data that must be accommodated by the similarity/targeting design. Better identification and removal techniques could increase the strength of the results presented below.

	RTB 1	RTB 2	RTB 3	RTB 4	RTB 5	RTB 6	RTB 7
created	100%	100%	100%	100%	100%	100%	100%
network	100%	100%	100%	100%	100%	100%	100%
dims	100%	100%	100%	100%	100%	100%	100%
id	100%	100%	100%	36%	0%	19%	44%
device	100%	100%	100%	100%	100%	100%	100%
user_agent	100%	100%	100%	100%	100%	100%	100%
iabcats	0%	100%	100%	98%	100%	100%	84%
geo	100%	16%	100%	100%	100%	100%	100%
location	0%	4%	56%	42%	1%	20%	100%
referer	8%	0%	0%	17%	0%	6%	3%
publisher	100%	100%	100%	100%	100%	100%	100%
url	100%	0%	0%	5%	93%	0%	0%
ip	100%	100%	100%	100%	100%	100%	100%

Table 1 Data availability for mobile devices across different RTB systems. More specifically, how often are the following fields provided: time (created) of bid request, RTB network identifier, dimensions (dims) of the ad, identifier (id) of the device, device type, user agent fields, high-level categories for the webpage/app iabcats, ip-inferred city, state or country (geo), location in lat/long, referer from where the user comes, publisher of the webpage, url and ip.

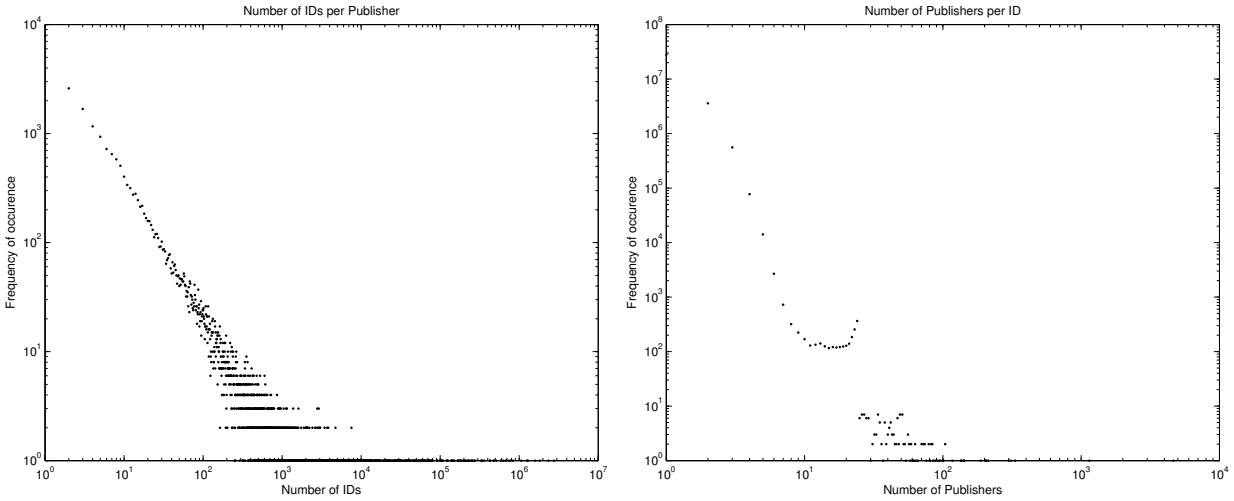


(a) Users per IP (b) IPs per user

Figure 2 Distribution of (a) number of unique user instances seen per IP, and (b) number of IPs a user instance logs into.

these IPs with a very large number of unique users seen correspond to large institutions, think for example of a university wifi address.

Although the publisher data item is available for all bid requests, we observe that most user instances visit only a single publisher (about 95%, while less than 1% visit three or more publishers, and only 0.009% are seen on 10 or more publishers). The publisher-specific categorizations (*iabcats*)



(a) Users per publisher (b) Publisher per user

Figure 3 Distribution of (a) number of unique user instances seen per publisher, and (b) number of publishers a user instance visits.

suffer from the same problem.[8] The reason is that, as mentioned above, a mobile device has different identifiers when viewed through different channels in the mobile advertising ecosystem, such as different RTB systems or even different apps. On the other hand, a publisher has on average 1,366 users that visit it, and publishers by-and-large seem to be RTB-specific. Figures 3(a) and 3(b) show more details of the distributions. As a consequence, using publisher-visitation data to connect users is not broadly helpful, as two user instances in different RTB systems rarely share a publisher. The IP address is always available, and is passed consistently through the different apps and RTB systems.

The other consistently available fields (such as user agent, device type, dimensions) are always the same for a single device. These could in principle be used to improve same-device finding, but are only marginally helpful for finding the same user on different devices or for similar-user finding (e.g., using Apple products vs. Android products); we do not take advantage of them in this paper, instead focusing specifically on assessing the ability of the geo-similarity design (as opposed to how to best engineer a same-user-finding system).

The *url* field is missing in all but two RTB exchanges. The *geo* field is often limited to the city/state and country, as inferred from the IP address. Location data in the form of lat/longs (*location*) is only available in some RTB systems and the accuracy of the lat/long data in the current mobile ecosystem is questionable,[9] again making it not suitable for broad use.

[8] As well as being too coarse-grained to be useful for identifying the same user in any event.

[9] Analyses of the lat/long data show various data problems, including flipped coordinates, coordinates set equal to each other, many lat/long points at the centroids of common geographic areas, and many impossible or highly

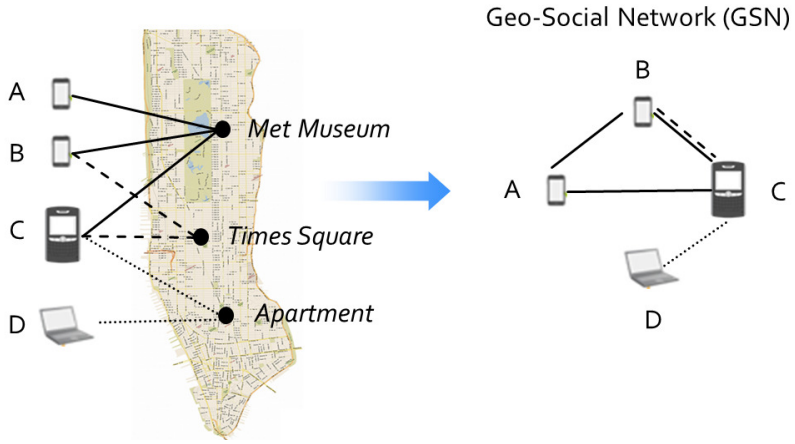


Figure 4 **Simple illustration of the creation of a Geo-Similarity Network (GSN) based on mobile location data observed for three mobile devices (A, B and C) and a laptop (D). Devices are connected in the GSN if they visited (logged into) the same location (IP address).**

3.2. Entity and Link Representation

Each mobile device is represented by a "profile" of behavior across locations, where we define a location profile as a vector of visit frequencies. This will necessarily require a sparse representation, as we observe a very large total set of locations; fortunately, each individual is seen at only a small subset of the locations. For the empirical results below, the location profile is a sparse vector of frequencies of observed visitations to locations (which can be weighted in the calculation of similarity, as described below).

The structure of the geo-similarity network is the graph with mobile devices as the nodes, and links between the devices as the (weighted) edges. The corresponding design decisions involve the selection of which devices to link at all, and the weights to place on the links. The simplest link representation places unweighted links between all pairs of devices that share at least one specific location. More sophisticated weighting criteria are based on two notions, which are discussed next and illustrated with a simple example in Figure 4. The location data can be obtained by listening to RTB ecosystems, where IP addresses of devices are broadcast (as previously shown in Figure 1).

1. Given two devices that share at least one location, the weight of the link between them can be a **function of the devices' location profiles**. Such measures can include simply the number of shared locations or could take into account the visitation percentages in the profile. Alternatively, the weighting could be a measure of similarity between the two location profiles, for example using cosine similarity.

improbable coordinates. Mobile ad industry insiders claim that the high price paid for lat/long-based hyperlocal targeting has led to a significant amount of lat/long fraud. We have found nothing written on this phenomenon, although data fraud is not uncommon in the online advertising industry (Stitelman et al. 2013); see also `http://www.adweek.com/topic/ad-fraud`.

As such, devices A, B and C are connected in the GSN of Figure 4 since they all logged into the Wifi of the Met Museum. B and C also logged into the public Wifi at Times Square, and therefore should be more strongly connected than A and B (or B and C) that share only one location.

2. Links can be **weighted according to the component locations' (lack of) popularity**. For example, it might be argued that the link formed by sharing a location with a small number of other devices (e.g., my apartment) indicates a stronger geo-similarity than sharing a location with a massive number of other devices. This intuition can be extended: if two devices spend a lot of time at such an unpopular locale (e.g. my apartment), then that should indicate a very strong similarity. And if two devices have approximately the same distributional profile across several of these sites (their "location profile"), then they're either the same person or close friends/soulmates. In our example, C and D are connected because they both logged into the Wifi of an apartment, while B and C are connected by logging into the public (popular) Wifi at Times Square. Hence, the connection between C and D (which are likely devices belonging to roommates or belonging to the same user) should be stronger than the connection between B and C (which are likely devices belonging to two different tourists).

This notion of similarity can be incorporated technically by (1) adapting notions from information retrieval: we can weight locations for a given device by their device-specific popularity (from the device's location profile), divided by the (log of) the location's overall popularity; let's call that LF×IDF (location frequency × inverse device frequency). Then (2) the strength of the link between two devices that share a location would be a function of the corresponding LF×IDF scores.

Once we have one or more geo-similarity networks, there are various ways to take advantage of them for targeting or other inference. The simplest method for inference is simply to target all of the geo-similarity network neighbors. Alternatively, one could target the "closest" neighbors based on one or more notions of geo-similarity (as discussed above). These are the methods used for the empirical results presented below.

4. Specifics of the GSN Design

As described above, in this design the strength of the link between two individuals in the GSN is based on the similarity in the distribution of the locations that they visit. For this paper's results, locations will be (anonymized) IP addresses. We now define different similarity measures of varying complexity, which will be evaluated in the following sections.

4.1. Notation

The location profile of a user is represented by a vector $\vec{x}$ of length m where element i denotes the number of visits to IP address IP_i: $\vec{x} \in \mathbb{R}^m$. Vector $\vec{x}_{bool}$ is a binary vector of size m where element i denotes whether the user has visited IP_i. For similarity computations, both the Hadamard product

($\circ$) and the inner product ($\cdot$) are used, and the minimum (min) operator is defined to operate componentwise on two vectors, as well as on a vector and a scalar. Their logic is defined in Equations (1-4). The maximum (max) and average (avg) operator are similar. The weight of an IP is defined by its popularity: similarly to the inverse document frequency used in text mining (Hotho et al. 2005), the weight w_i for IP_i (the inverse device frequency) is defined as the logarithm of the total number of users n divided by the number of (unique) users that visit that specific IP n_i.

$$\vec{z} = \vec{x} \circ \vec{y}, \quad z_i = x_i \times y_i \tag{1}$$

$$a = \vec{x} \cdot \vec{y} = \sum_{i=1:m} (x_i \times y_i) \tag{2}$$

$$\vec{z} = min(\vec{x}, \vec{y}), \quad z_i = min(x_i, y_i) \tag{3}$$

$$\vec{z} = min(\vec{x}, a), \quad z_i = min(x_i, a) \tag{4}$$

$$w_i = log_{10}(n/n_i) \tag{5}$$

4.2. Link strength metrics

Fifteen similarity metrics are defined in Table 2, and are illustrated by measuring the strength between the location profiles of two example users: $\vec{x}_1$ and $\vec{x}_2$ shown in Eq. (6). The first user visits IP_1 three times and IP_2 twice, while the second one visits IP_1 twice and IP_3 once. Several variants of the same metrics are considered. As described above, some only take into account the unique number of shared locations, while others take into account the frequency of the visits to these locations and the inverse device frequency. The latter are denoted by *freq* and *W*, respectively.

The first similarity metric, s_{count}, counts the number of shared locations. Since the two example users share only IP_1 this strength is 1. The IDF weighted version $s_{count,W}$ sums the IDF values of the shared locations, which is 1.3 for our simple example. The most basic metric is given by s_{bool} which is 1 when a location is shared and 0 otherwise, indicating whether two users are neighbors or not.

The $s_{freq,min}, s_{freq,max}$ and $s_{freq,avg}$ metrics compare the frequencies of the visits to the shared locations. They take the minimum, maximum and average respectively of the frequencies for each shared location, and sum them. The weighted versions $s_{freq,min,W}, s_{freq,max,W}$ and $s_{freq,avg,W}$ do the same, but weight each frequency with the corresponding *IDF* value.

The s_{cosine} metric measures the similarity by taking the cosine of the angle between the two vectors. The Jaccard metric is defined as the size of the intersection over the size of the union. In this case it counts the number of shared locations, but normalizes these over the total number of locations both users have visited. Users that visit many locations will have a higher chance to share some location with another user, so the Jaccard metric penalizes them. For our basic example,

$s_{Jaccard}$ is 1 (IP_1) over 3 (IP_1, IP_2, IP_3). Once more, variants taking into account the frequency and the IDF are defined as well.

$$\begin{aligned} \vec{x}_1 &= [3 \quad 2 \quad 0] \\ \vec{x}_2 &= [2 \quad 0 \quad 1] \\ \vec{w} &= [1.3 \quad 1.7 \quad 2] \end{aligned} \tag{6}$$

	Similarity metric			range	$d(\vec{x}_1, \vec{x}_2)$
1:	$s_{count}(\vec{x}_1, \vec{x}_2)$	=	$\vec{x}_{bool,1} \cdot \vec{x}_{bool,2}$	$[0, \infty)$	1
2:	$s_{bool}(\vec{x}_1, \vec{x}_2)$	=	$min(s_{count}(\vec{x}_1, \vec{x}_2), 1)$	$\{0, 1\}$	1
3:	$s_{count,W}(\vec{x}_1, \vec{x}_2)$	=	$(\vec{w} \circ \vec{x}_{bool,1}) \cdot \vec{x}_{bool,2}$	$[0, \infty)$	1.3
4:	$s_{freq,min}(\vec{x}_1, \vec{x}_2)$	=	$\|\|min(\vec{x}_1, \vec{x}_2)\|\|_1$	$[0, \infty)$	2
5:	$s_{freq,max}(\vec{x}_1, \vec{x}_2)$	=	$\|\|max(\vec{x}_{bool,2} \circ \vec{x}_1, \vec{x}_{bool,1} \circ \vec{x}_2)\|\|_1$	$[0, \infty)$	3
6:	$s_{freq,avg}(\vec{x}_1, \vec{x}_2)$	=	$\|\|avg(\vec{x}_{bool,2} \circ \vec{x}_1, \vec{x}_{bool,1} \circ \vec{x}_2)\|\|_1$	$[0, \infty)$	2.5
7:	$s_{freq,min,W}(\vec{x}_1, \vec{x}_2)$	=	$\|\|min(\vec{x}_1 \circ \vec{w}, \vec{x}_2 \circ \vec{w})\|\|_1$	$[0, \infty)$	2.6
8:	$s_{freq,max,W}(\vec{x}_1, \vec{x}_2)$	=	$\|\|max((\vec{x}_{bool,2} \circ \vec{w}) \circ \vec{x}_1, (\vec{x}_{bool,1} \circ \vec{w}) \circ \vec{x}_2)\|\|_1$	$[0, \infty)$	3.9
9:	$s_{freq,avg,W}(\vec{x}_1, \vec{x}_2)$	=	$\|\|avg((\vec{x}_{bool,2} \circ \vec{w}) \circ \vec{x}_1, (\vec{x}_{bool,1} \circ \vec{w}) \circ \vec{x}_2)\|\|_1$	$[0, \infty)$	3.25
10:	$s_{cosine}(\vec{x}_1, \vec{x}_2)$	=	$\frac{\vec{x}_1 \cdot \vec{x}_2}{\|\|\vec{x}_1\|\|_1 \cdot \|\|\vec{x}_2\|\|_1}$	$[0, 1]$	0.74
11:	$s_{cosine,W}(\vec{x}_1, \vec{x}_2)$	=	$\frac{(\vec{x}_{bool,1} \circ \vec{w}) \cdot (\vec{x}_{bool,2} \circ \vec{w})}{\|\|\vec{x}_{bool,1} \circ \vec{w}\|\|_1 \cdot \|\|\vec{x}_{bool,2} \circ \vec{w}\|\|_1}$	$[0, 1]$	0.60
12:	$s_{Jaccard}(\vec{x}_1, \vec{x}_2)$	=	$\frac{s_{count}(\vec{x}_1, \vec{x}_2)}{\|\|max(\vec{x}_{bool,1}, \vec{x}_{bool,2})\|\|_1}$	$[0, 1]$	0.33
13:	$s_{Jaccard,W}(\vec{x}_1, \vec{x}_2)$	=	$\frac{s_{count,W}(\vec{x}_1, \vec{x}_2)}{\|\|max(\vec{x}_{bool,1} \circ \vec{w}, \vec{x}_{bool,2} \circ \vec{w})\|\|_1}$	$[0, 1]$	0.26
14:	$s_{Jaccard,freq}(\vec{x}_1, \vec{x}_2)$	=	$\frac{s_{freq,min}(\vec{x}_1, \vec{x}_2)}{\|\|max(\vec{x}_1 \circ \vec{x}_{bool,2}, \vec{x}_2 \circ \vec{x}_{bool,1})\|\|_1}$	$[0, 1]$	0.67
15:	$s_{Jaccard,freq,W}(\vec{x}_1, \vec{x}_2)$	=	$\frac{s_{freq,min,W}(\vec{x}_1, \vec{x}_2)}{\|\|max((\vec{x}_1 \circ \vec{x}_{bool,2}) \circ \vec{w}, (\vec{x}_2 \circ \vec{x}_{bool,1}) \circ \vec{w})\|\|_1}$	$[0, 1]$	0.67

Table 2 Similarity metrics between two user's location visit distributions $\vec{x}_1$ and $\vec{x}_2$.

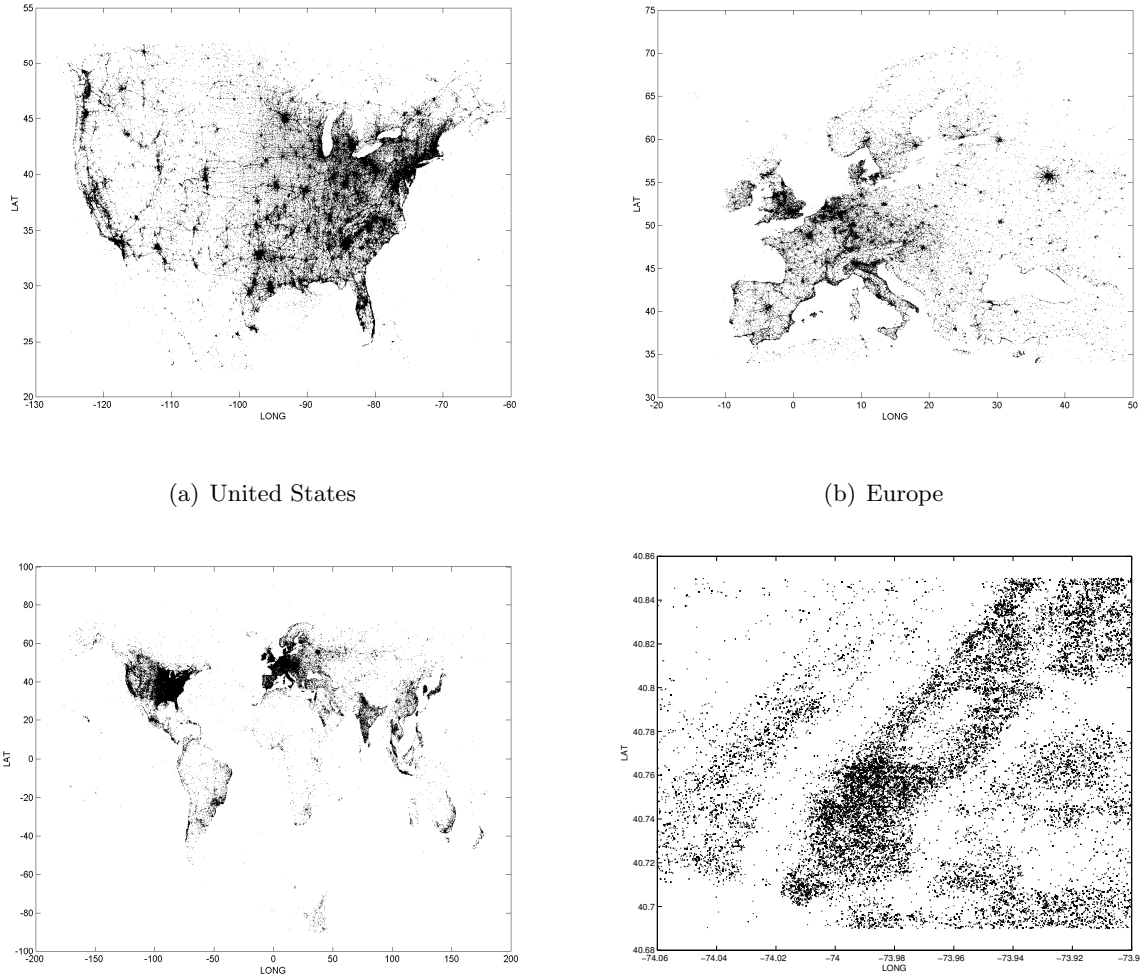
5. Experimental Setup

5.1. Dataset

The dataset $\mathcal{D}$ for the empirical results is 10 days of anonymized advertising bid requests from mobile devices (smart phones, tablets, etc.) observed across seven RTB exchanges. A total of 322,770,794 bid requests are observed, coming from 42,437,559 unique user instances. Over these 10 days, we observe 41,829,088 unique IP addresses.

The number of unique user instances per RTB exchange is given in Table 3. The distributions of the number of bid requests per device type and operating system are given in Tables 4 and 5, which show that the majority of bid requests come from smartphones and from the IOS operating system.

A user (instance) corresponds to an exchange-assigned userid. One person on several devices will hence correspond to several users. One device on several exchanges will often correspond to



(a) United States

(b) Europe

(c) Earth

(d) Manhattan

Figure 5 **A scatterplot of a sample of latitude and longitudes broadcast by mobile devices (note that no maps are drawn). It gives a striking picture of population density across the world, and makes us wonder what's going on with mobile devices in Antarctica.**

Exchange	Volume	Perc.
RTB 1	3,175,979	6%
RTB 2	2,159,686	4%
RTB 3	14,566,098	29%
RTB 4	20,501,243	41%
RTB 5	2,275,015	5%
RTB 6	5,055,815	10%
RTB 7	2,223,418	4%

Table 3 **User instances per RTB exchange.**

Type	Volume	Perc.
Phone	292,358,772	91%
Tablet	30,016,334	9%

Table 4 **Distribution of bid requests per device type.**

Type	Volume	Perc.
Android	129,315,617	40.1%
IOS	192,923,648	59.8%
Windows	531,529	0.2%

Table 5 **Distribution of bid requests per operating system.**

several users. Also, it might be that one ad exchange sometimes provides different userids to the same device, for example when the device is using different particular apps. As described above, an important task in online advertising is to be able to target advertisements to different instances of the same individual, with reasonable probability, regardless of the complexities of identification in the baroque online advertising ecosystem.

5.2. Sampling

For the results in this paper, we apply the geo-similarity network analysis to data samples. Each sample corresponds to a cohort/neighborhood and is built as follows: a single (random) "ego" user is chosen, user X, around which a neighborhood is built. Every neighbor of the ego user must share at least one IP address, so we begin by obtaining all (anonymized) IP addresses visited by user X. Next, all other users that visited at least one of these IPs (the neighbors) are added. The ego must have at least one neighbor to be included in this analysis. Finally, we obtain the complete location distribution of each of these neighbors. When the next sample is chosen, we ensure that none of the users already included in the previous samples is taken as the ego user (X). As such, we ensure that we don't have repeated ego-neighbor dyads. There may be shared neighbors of two different egos, but two egos will never be immediate neighbors in the data sample.

The result of the sampling over 500 cohorts is summarized in Figure 6: on average an ego has 71 neighbors, the median being 1. The maximum number of neighbors is 3,167: this user probably visited a public wifi that thousands of other users also logged in to. We chose 500 cohorts as we observe that the statistics become quite stable, as seen in Figure 7. Even 200 cohorts seems to be enough to obtain stable sample statistics.

The lifespan of user instances in this set of 500 cohorts is shown by the cumulative frequency in Figure 8. Observe that only about 40% of the user instances are seen at different hours. About 30% of the user instances are seen only once in the sampled week. This illustrates that it is vital to understand the notion of user *instances*. These percentages are misleading if we interpret the user instances as individuals or devices. One individual/device without stable identification in the

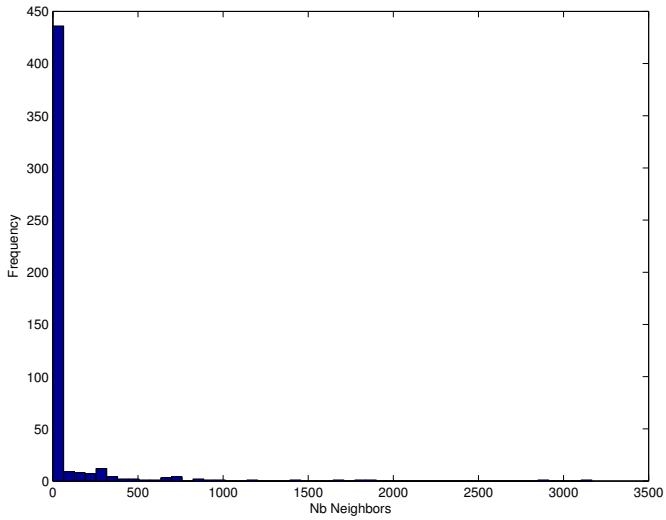


Figure 6 **The distribution of the number of neighbors for 500 randomly sampled ego users.**

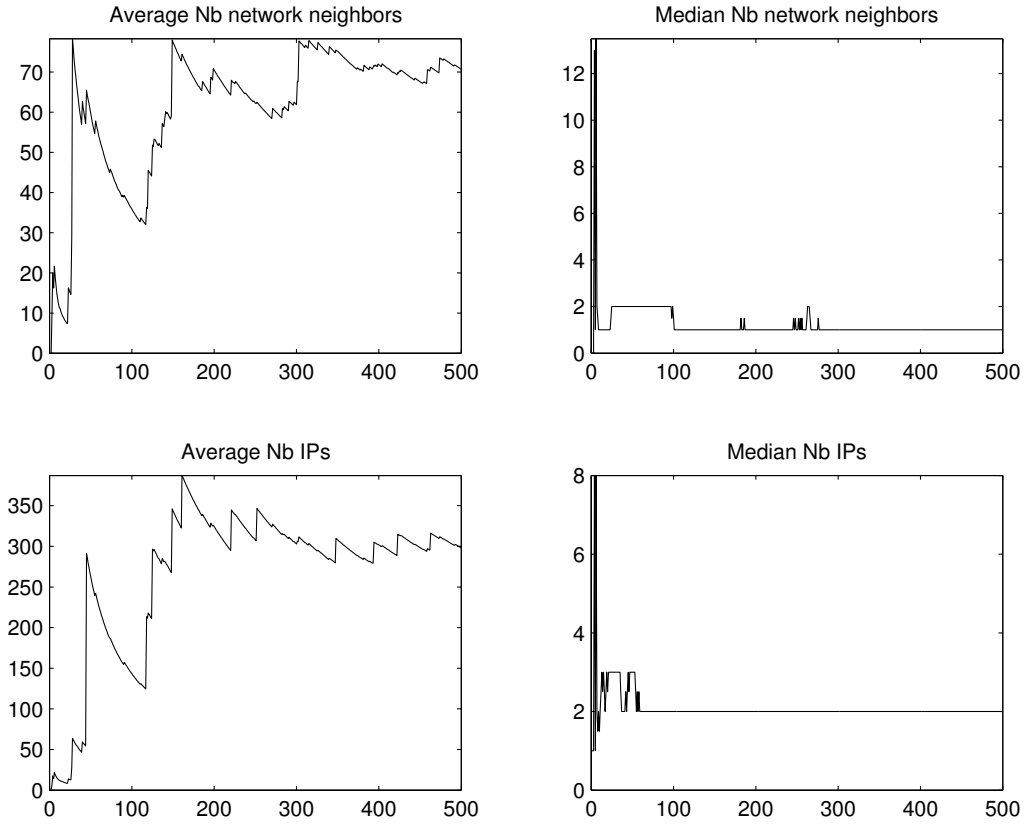


Figure 7 **Evolution of several sample statistics as an increasing number of cohorts are created. The horizonal axis denotes the number of cohorts. By 500 cohorts, the statistics have become stable.**

online advertising ecosystem will create many transient user instances, which will drive up the percentage of user instances that are seen only once, or that are seen only within a short timeframe.

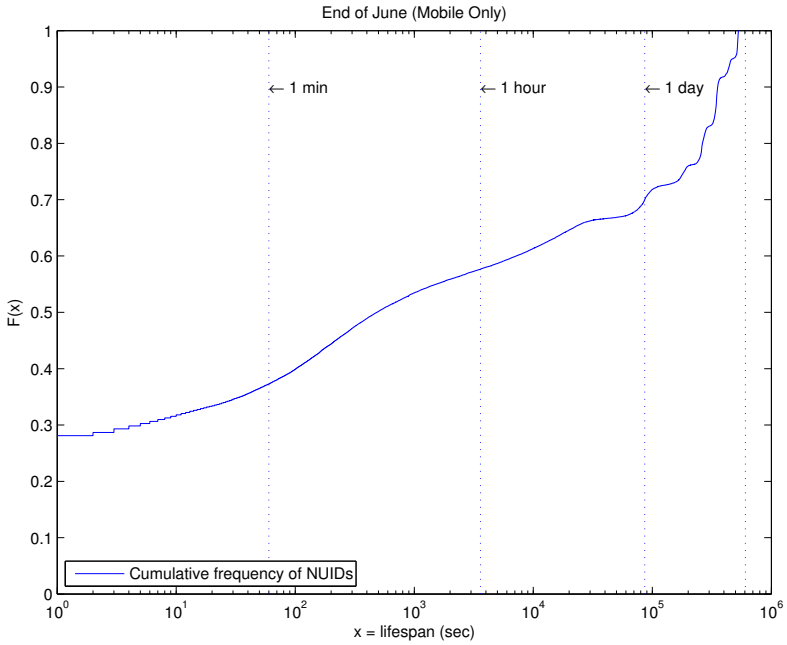


Figure 8 **Cumulative frequency of lifespan of user instances.**

We have not figured out how to determine or estimate reliably how many "unstable" users/devices are represented by the user instances in these data. We keep all the data for the analyses below (rather than filtering very short-lived user instances) to maintain realism.

6. Study 1: Connecting Instances of the Same Individual

Returning to our motivating application, as mentioned previously, an important task in mobile advertising is to be able to target a particular individual in the face of the individual's fragmentation into different user instances (for example, as seen through different bidding systems). Advertisers understand that it may be impossible to find the same user with 100% accuracy. Nonetheless, it often is important for them to reach the user's various instances without targeting too many others.[10] The GSN is designed in part based on the hypothesis that different instances of the same person would visit similar locations, and therefore they will be connected in the GSN. In this section we present results examining whether and to what extent that is indeed the case. Further, the results quantify to what extent the same user receives a high GSN-similarity to herself, as compared to her other neighbors.

Three experimental settings are used to assess judgments that two users are the same. The first two simulate different user instances by splitting up the bid requests of known user instances. The

[10] The astute reader may notice that, if the other users targeted are very similar to the target user, the advertising may be well placed despite not being shown to the exact target user. That is not the subject of this study. Study 2 shows evidence that these two notions are not independent when using the GSN.

last one is not simulated: it links users via the IFA device identifier, observed over several RTB exchanges for a subset of device instances.

	IP1	IP2	IP3
X	1	5	0
X1	0	3	0
X2	1	2	0

Table 6 Randomly splitting user X's IP visits, creating two new artificial users: X1 and X2.

1. *Random*: we divide the IP visits of a user randomly across two simulated users. Table 6 shows for example how we can divide a user's IP visits across two new users: X1 and X2, for which we know that they are actually the same individual. Results with Random should be used only as a ceiling on performance, as the results will be overly optimistic.

2. *Temporal* : a more realistic splitup of the transactions is a temporal splitup, where first the middle day of X's transactions is determined (hence dynamically chosen for each user). All IP visits up to this middle day are assigned to X1, all transactions after the middle day are assigned to X2. It will only be applied to users who are seen on multiple days.

3. *Identifier For Advertising (IFA)*: IFA is an advertising identifier from Apple that users can change or ask not to be used for advertising, and is seen as a privacy-friendlier version of the controversial, fixed device id (the "UDID"). The IFA is observed in three exchanges, with volumes shown in Table 7: for example, 1,940,855 unique IFAs are seen in RTB 1.[11] When we observe the same IFA across different exchanges, we now know that two devices in different exchanges actually are the same. The frequency of unique IFAs visible *across* exchanges is shown in Table 8. For example, a total of 3,018,223 IFAs are seen when we look at both RTB 1 and RTB 2, with 306,594 of these (10.16%) seen on both exchange 1 and exchange 2. What if there wasn't an IFA and the device would have had a different identifier in each exchange? To what extent are these two devices instances then connected in the GSN?

	Volume
RTB 1	1,940,855
RTB 2	1,383,962
RTB 3	789,292

Table 7 The volume of IFA occurrence in our dataset.

With these different "gold standard" methods to indicate whether or not two users are the same, we assess the ability of the geo-similarity metrics to link the same user, determining how often

[11] For confidentiality, RTB 1 in this setting is not necessarily the same as RTB 1 in Section 4 or 7.

	Frequency
RTB 1/RTB 2	306,594 (10.16%)
RTB 2/RTB 3	45,256 (1.69%)
RTB 3/RTB 1	18,417 (0.85%)

Table 8 The frequency of unique IFAs visible across RTB exchanges.

two instances of the same person are connected in the GSN (Section 6.1). Afterwards we shall assess the different GSN *strength* metrics' ability to rank users, determining to what extent two instances of the same person exhibit *strong* geo-similarity, as compared to the other neighbors (Section 6.2). For the random and temporal splitups, 200 samples are used. For the IFA-based method the complete population is used for measuring the degree of connectedness to the same user; to analyze the relative strength of the connected users, 200 samples are used.

6.1. Same individual, different "screens"

We now present results on the GSN connectedness, measured as how often two instances of the same user are connected in the GSN (in percentage). Note that the results, summarized in Table 9, should be interpreted in light of the limited number of days in building the research data set, which will limit the GSN connectivity. Nonetheless, the relatively high connectivities observed even with this sample provide quite promising results, which lends considerable support to the merit of the overall design.

	% connected
Random	91%
Temporal	68%
IFA – RTB 1/RTB 2	67%
IFA – RTB 2/RTB 3	82%
IFA – RTB 3/RTB 1	73%

Table 9 To what extent is the same user connected to herself.

Random: Using the random split, two users that correspond to the same individual are not connected in only 9% of the neighborhoods. Moreover, in these cases, user X (whose IP visits are split up) usually has only 1 or 2 entries, and the random splitup results in no visits for one of the two simulated users.

Temporal: The temporal splitup is more realistic than the random one, where we assume that a person uses simulated device one in the first half of the week (e.g., during the weekend), and simulated device two in the second half of the week (e.g., during the week).

Users X1 and X2 are not connected in 32% of the neighborhoods. Whereas the random splitup was too optimistic, the temporal one is too pessimistic, as in reality there likely is some overlap in

the time periods of use of two different devices. If user X visits an IP only in the first days of the week, this IP will not occur in user X2's transactions and reduces the chance to be connected to X1.

IFA: We look at all combinations of the three exchanges with IFA identifiers. For example, there are 45,256 unique IFAs seen on both RTB 2 and RTB 3. In 82% of these, the two IFAs visit at least one same IP on both exchanges, and hence are connected in our GSN.

The bottom line is that even with this limited slice of online behavior, different instances of the same individual are connected in the GSN 70–80% of the time. Encouragingly, the results from the simulated scenarios concur with the results from the real (IFA) scenarios (including the aforementioned optimism and slight pessimism of the two simulated scenarios). Consistency over these different settings gives additional confidence in the results, suggesting that the GSN indeed holds promise for a high degree of success at targeting the same user across the fragmentation of the online advertising ecosystem.

6.2. Ranking user instances

Next, given that two user instances of the same individual are indeed connected, we assess to what extent the second instance of the same user is ranked highly among its network neighbors, based on the weight of their linkage in the GSN. We first will discuss the most general results, using the best ranking method, and then will look across different ranking methods.

We decompose the results based on the number of network neighbors (NN), since there are two boundary cases that require special interpretation. Tables 10-14 show the proportions of the ego networks that fall into three scenarios. First, a user may have only one NN. In this scenario, the ranking is perfect (trivially)—the other instance of the same user is the only user instance given a non-zero score. As shown in the tables, this scenario accounts for 30%-60% of the ego networks (depending on the splitup method).

The second boundary case is when the ego has exactly two network neighbors. This scenario accounts for 10%-25% of the cases, depending on the splitup method. Here the table also reports whether the other known user instance is ranked first (including ties) among the two neighbors. The other known instance of the same user is ranked first 60%-80% of the time, depending on the splitup scenario. The IFA results are affected strongly by the particular RTB setting—this may indicate that the publishers on the different RTBs have different policies for passing the IFA. Importantly, these results are conservative—possibly very conservative. The other NNs may also be instances of the same user unbeknownst to us—for example, in the IFA case, because an app does not pass the IFA to the RTB, or because the other user is the same user on a different device, and thus does not share the IFA.

The final scenario is when more than two network neighbors are present. As shown in the tables, this accounts for 20%-50% of the cases, depending on the splitup scenario. In this case, we measure the percentile in the ranking where the instance of the known-same-user falls (see below), as well as the Area under the ROC curve (AUC) (Fawcett 2006), which is equivalent to the Mann-Whitney-Wilcoxon (MWW) statistic. The AUC/MWW measures to what extent data points with label 1 are ranked higher than those with label 0. In the results that follow (see Table 15), the AUC is measured for each sample, where any user instance among the neighbors that corresponds to the same individual is labeled as 1, all others are labeled as 0. Thus, in this setting the AUC measures how well the neighbors are ranked by the likelihood of being the same individual.[12] The average over all samples is reported. If for a sample all neighbors have the same score, that AUC is set to 0.5. In addition to the factors noted above, this also will tend to make the results conservative. In addition, this evaluation will be conservative because the AUC will be 0.5 if the GSN links an ego user to a set of neighbors that are all known instances of the same individual.

The ranking metric is the percentile in the ranking at which the known-same-user instance is ranked, with lower values being better. For example, if the same-user instance has the highest score (closest connection) among five neighbors, the rank is 20% (1/5). Note again that this is quite conservative: in this example, the user is ranked at the top of the list, but because there were only 5 neighbors, the highest possible score is 20%. In case the known-same-user instance has the same score as another neighbor, the average rank is reported.

[12] Technically, in this setting, the AUC/MWW is the probability that a randomly selected neighbor that is in fact a known instance of the same user is ranked more highly than a randomly selected neighbor that is not known to be an instance of the same user. If the known same-user instances are all ranked above the other-user instances, the AUC=1.0. If they are all below, then the AUC=0.

Ego has only 1 NN :	31.4%
Ego has exactly 2 NN :	19.5%
Same user ranked first:	80.5%
Ego has more than 2 NN :	49.1%

Table 10 In what portion of cases is the same user connected, depending on number of network neighbors (NN) - Random splitup. For scenario with 2 NNs, the table reports the percentage of cases where the other known instance of the same user is ranked first (including ties).

Ego has only 1 NN :	30.8%
Ego has exactly 2 NN :	18.7%
Same user ranked first:	83.9%
Ego has more than 2 NN :	50.5%

Table 11 In what portion of cases is the same user connected, depending on number of network neighbors (NN) - Temporal splitup.

Ego has only 1 NN :	44.1%
Ego has exactly 2 NN :	12.1%
Same user ranked first:	82.1%
Ego has more than 2 NN :	44.0%

Table 12 In what portion of cases is the same user connected, depending on number of network neighbors (NN) - IFA RTB 1/RTB 2 splitup.

Ego has only 1 NN :	61.9%
Ego has exactly 2 NN :	19.7%
Same user ranked first:	71.2%
Ego has more than 2 NN :	18.4%

Table 13 In what portion of cases is the same user connected, depending on number of network neighbors (NN) - IFA RTB 2/RTB 3 splitup.

Ego has only 1 NN :	53.5%
Ego has exactly 2 NN :	25.1%
Same user ranked first:	63.4%
Ego has more than 2 NN :	21.4%

Table 14 In what portion of cases is the same user connected, depending on number of network neighbors (NN) - IFA RTB 3/RTB 1 splitup.

	AUC (if > 2 NN)					Ave. Rank Percentile (if > 2 NN)				
Ranking Metric	R	T	RTB1/2	RTB2/3	RTB3/1	R	T	RTB1/2	RTB2/3	RTB3/1
1	78%	64%	68%	59%	55%	29%	48%	34%	43%	54%
2	59%	58%	67%	58%	54%	45%	53%	34%	44%	55%
3	79%	69%	74%	61%	61%	29%	44%	**29%**	40%	48%
4	86%	75%	69%	61%	59%	24%	**40%**	32%	41%	50%
5	83%	70%	71%	59%	63%	26%	44%	31%	43%	47%
6	85%	74%	71%	60%	65%	25%	41%	31%	42%	45%
7	**86%**	**76%**	**73%**	62%	64%	**24%**	40%	29%	**40%**	46%
8	83%	70%	72%	59%	65%	26%	44%	30%	43%	45%
9	85%	74%	71%	60%	**67%**	**25%**	41%	30%	42%	**44%**
10	79%	69%	69%	61%	63%	29%	44%	32%	41%	47%
11	79%	69%	71%	**62%**	64%	44%	44%	30%	41%	46%
12	81%	68%	68%	60%	62%	28%	46%	34%	42%	49%
13	81%	70%	71%	61%	64%	27%	44%	30%	40%	46%
14	74%	67%	66%	61%	48%	34%	47%	35%	41%	60%
15	74%	67%	66%	61%	48%	34%	47%	35%	41%	60%

Table 15 When there are more than two network neighbors, to what extent is the known-same-user strongly connected to itself? R = Random split. T = Temporal split. RTB x/y = corresponding RTB split. AUC: larger is better. Rank Percentile: smaller is better. Please see text for discussion of ranking metrics.

As shown in Table 15, all AUC values exceed 0.5, ranging up to 0.8. This means that in every case, a known-same-user instance is likely to be ranked higher than an instance not known to be the same user. The frequency-based ranking metrics (4-9) perform quite well in connecting instances of the same user strongly.

Almost all metrics perform better than a random model, with performances that even come close to the best possible solution. Consistently performing quite well are the frequency-based metrics (metrics 3-9), with metric 7 (the minimum of the IDF weights of the shared locations) getting the most wins. Not performing well is the very basic count metric, as well as the boolean metric that provides a binary score only. Interesting to notice, the Jaccard metrics do not perform well either. It seems that penalizing users that visit many IPs negatively affects the results, an issue we will return to in the next study.

7. Study 2: Does the GSN select users with similar interests?

The foregoing section addressed our the paper's first claim—that the GSN design indeed links instances of the same user, and links them more strongly than instances of other users. In this section, we turn to the paper's second claim—that the GSN links users with similar interests. We will measure similar interests by similarity in behavior accessing particular publishers on the web and using mobile apps.[13]

RTB 1		**RTB 2**	
Publisher	$\mathcal{D}$ *(%)*	*Publisher*	$\mathcal{D}$ *(%)*
Burstly	3.04	m.worldnow.com	0.32
myYearbook.com	1.62	m.tmz.com	0.24
My Fitness Pal	0.65	m.topix.com	0.14
TextMe, Inc.	0.56	babycenter.com	0.15
GameResort	0.25	meetme.com	0.06
Flixster	0.19	apps.facebook.com	0.06
247Sports	0.15	itunes.apple.com	0.39
Daily Workout Apps, LLC	0.09	app.evite.com	0.02
Mobile Deluxe	0.07	beautyandskincares.com	0.03
YoYo Games Ltd.	0.04	mmajunkie.com	0.01
RTB 3		**RTB 4**	
Publisher	$\mathcal{D}$ *(%)*	*Publisher*	$\mathcal{D}$ *(%)*
Conversion Exchange	5.53	Pinger Phone - iOS - Conversation	1.01
Pinger, Inc.	2.96	Talkatone iPhone App	0.07
Top Game Developer	4.29	DiceWithBuddies	0.13
PremiumEntertainmentApp-FamilyOriented-Android	0.59	9GAG Reader	0.12
Enflick	0.53	Video Downloader Pro Lite	0.16
Clapfoot Games	0.44	Rage Comics	0.09
FunPokes, Inc.	0.18	Spades 3D Lite	0.06
Talkatone	0.15	Apalabrados	0.03
Sevenlogics, Inc.	0.08	Relax Melodies HD	0.02
A Star Software	0.07	Celeb Me - PhotoMaker	0.00

Table 16 Distribution of publishers per RTB, in terms of the percentage of users in our dataset that visit the given publisher.

[13] From now on all user instances will be complete, real user instances; the simulated split-up scenarios (Random, Temporal) were created specifically for the purpose of the same-user study, above.

7.1. Connecting users with similar interests

To what extent are the GSN neighbors of visitors to particular publishers also visitors to those same publishers? In order to assess whether the GSN connects users with similar interests/behavior, we select a diverse collection of 40 publishers, listed in Table 16. A publisher can correspond to a mobile app, a website, or a set of apps/websites. We selected both publishers that are visited by many users and more niche publishers with fewer visitors. As can be seen from Table 16, several types of publishers can be distinguished; the main groups are games (e.g. GameResort, Top Game Developer) and social networking tools (e.g. apps.facebook.com, Text Me Inc.).

More specifically, we create 200 samples for 40 publishers, 10 publishers for each of four different RTB exchanges, listed in Table 16.[14] The percentage of unique users that visit each publisher in our complete dataset is also shown ($\mathcal{D}$). To create each sample, an "ego" user is selected that has visited the publisher in question (an "ego visitor"); this ego visitor's GSN neighborhood is created. Next, we assess how many of the ego's neighbors also visited that same publisher. This will be compared to a baseline visit rate, to produce lift and leverage values (Provost and Fawcett 2013)—specifically: how much more likely is it for the neighbors of ego visitors to visit the publisher than would be expected by chance (see below)? We consider three different setups that produce different baselines.

1. *Considering all RTB exchanges to determine the baseline population*: a user is considered a visitor for a given publisher if she visits the publisher on *any* RTB exchange (since some publishers are visible on different exchanges). Hence, the baseline visit rate is: the number of users (on any RTB) who visit that publisher, divided by the total number of users across all RTB exchanges. This will provide optimistic lifts, since most publishers appear only on one RTB exchange, while the baseline is measured over all exchanges.

2. *Considering one RTB exchange only to determine the baseline population*: a user is considered a visitor for a given publisher and RTB if she visits the publisher on the *same* RTB exchange. This changes the baseline to: the number of users (on this RTB) who visit that publisher, divided by the total number of users on this RTB exchange. The lift and leverage results will be worse than in the first setting, as the baseline will be larger. Seeing that some publishers are seen on different RTB exchanges, this setting is slightly pessimistic.

3. *Take time of location visits into account*: in the third setting, we link users only if they visit an IP within the same time period. This will be elaborated on in Section 7.3.

[14] For confidentiality, the RTB numbering here is again different from above.

Results are shown in Tables 17 and 18. The percentage of publisher visitors in the neighborhood is given as $\mathcal{N}$. This percentage is compared to the baseline percentage of users that visit that publisher in the complete data $\mathcal{D}$ using two metrics: the *lift* is the ratio of these numbers; the *leverage* is the difference. Lift shows relative improvement and is particularly useful for very small base rates (consider that an activity with a base rate of 30% cannot possibly have a lift higher than about 3). Leverage is more telling for larger base rates, as it shows the absolute improvement (here in percentage points).

When considering all the network neighbors (All NN), the lifts for all publishers are greater than 1, except for one: Burstly on RTB 1 has a lift of 0.97 (essentially, no lift), but only when we consider RTB 1 only as the baseline. This publisher already has a very high baseline, and as we will see in the next section, when we consider only the top 10 and 1 network neighbors, the lifts go up to 3 and 5 respectively. Burstly provides a generic platform for building apps, and so is a weaker indication of user interest than many of the other publishers (fitness, babies, games).

All other lifts exceed one by substantial margins and go up to even 2,391 (Apalabrados on RTB 4, a Spanish-language social Scrabble-like game). Obviously, the conclusion that the network neighbors indeed are more likely to visit the same publishers is highly significant by a simple sign test, comparing the visitation rate within the geo-similarity neighborhood to that of overall population either via lift or leverage; $p<0.01$. Thus, network neighbors of publisher visitors are indeed more likely also to visit the same websites; the actual values of the lifts and leverages show that they are substantially more likely.

The very high lifts could be attributed to two phenomena: they may provide some evidence that the GSN embeds a true social network, as pointed out by the good results for the social networking publishers (e.g. Apalabrados with a lift of 2,391, apps.evite.com with a lift of 2,212, and others). Friends visiting the same social networking sites (here, unusual ones) would increase the lift and leverage over just having different instances of the same user visiting the same websites. An alternative explanation is that some of these publishers are much more likely to be visited across different devices of the same user. It's not clear whether this is actually true. However, it argues for obtaining specific and broad data on which devices *are not* the same user, to shed further light on, for example, any broad data that includes a (anonymized) user identifier.

From these results, we can conclude that GSN neighbors indeed exhibit substantially similar publisher visitation behavior. Next, we assess to what extent more strongly connected network neighbors are more similar than less strongly connected network neighbors.

RTB 1

Considering all RTB exchanges to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Burstly	3.04	8.63	2.84	5.59	47.22	15.54	44.18	25.61	8.43	22.57
2	myYearbook.com	1.62	25.77	15.94	24.15	44.74	27.67	43.12	35.16	21.74	33.54
3	My Fitness Pal	0.65	4.15	6.38	3.5	36.9	56.77	36.25	15.55	23.91	14.9
4	TextMe. Inc.	0.56	18.5	33.06	17.94	39.02	69.75	38.46	25.18	45	24.62
5	GameResort	0.25	4.93	19.77	4.68	18.52	74.26	18.27	11.86	47.58	11.62
6	Flixster	0.19	1.55	8.15	1.36	8.14	42.85	7.95	1.79	9.42	1.6
7	247Sports	0.15	27.57	188.58	27.42	77.62	530.93	77.48	74.53	509.75	74.38
8	Daily Workout Apps. LLC	0.09	1.92	21.47	1.83	19.23	215.46	19.14	11.54	129.27	11.45
9	Mobile Deluxe	0.07	1.87	28.67	1.8	23.53	360.9	23.46	9.63	147.7	9.56
10	YoYo Games Ltd.	0.04	3.52	99.58	3.49	21.57	609.64	21.53	13	367.45	12.96

Considering this RTB exchange only to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Burstly	8.85	8.63	0.97	-0.22	47.22	5.33	38.37	25.61	2.89	16.76
2	myYearbook.com	4.71	25.77	5.47	21.06	44.74	9.5	40.03	35.16	7.46	30.45
3	My Fitness Pal	1.89	4.15	2.19	2.25	36.9	19.48	35.01	15.55	8.21	13.65
4	TextMe. Inc.	1.63	18.5	11.35	16.87	39.02	23.94	37.39	25.18	15.45	23.55
5	GameResort	0.73	4.93	6.78	4.2	18.52	25.49	17.79	11.86	16.33	11.14
6	Flixster	0.55	1.55	2.8	0.99	8.14	14.71	7.59	1.79	3.23	1.24
7	247Sports	0.43	27.57	64.73	27.14	77.62	182.23	77.2	74.53	174.97	74.1
8	Daily Workout Apps. LLC	0.26	1.92	7.37	1.66	19.23	73.95	18.97	11.54	44.37	11.28
9	Mobile Deluxe	0.19	1.87	9.84	1.68	23.53	123.88	23.34	9.63	50.7	9.44
10	YoYo Games Ltd.	0.1	3.52	34.18	3.42	21.57	209.25	21.47	13	126.12	12.9

RTB 2

Considering all RTB exchanges to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	m.worldnow.com	0.32	13.27	40.86	12.95	54.39	167.42	54.06	31.65	97.44	31.33
2	m.tmz.com	0.24	6.89	28.86	6.65	23.08	96.7	22.84	10.68	44.75	10.44
3	m.topix.com	0.14	15.07	104.79	14.93	53.45	371.6	53.3	33.75	234.65	33.61
4	babycenter.com	0.15	6.87	47.31	6.72	43.48	299.56	43.33	17.74	122.24	17.6
5	meetme.com	0.06	6.02	103.8	5.97	54.72	942.93	54.66	27.85	479.9	27.79
6	apps.facebook.com	0.06	3.1	48.39	3.03	9.09	141.97	9.03	2.76	43.08	2.69
7	itunes.apple.com	0.39	25.77	65.71	25.38	51.43	131.12	51.04	29.49	75.18	29.09
8	app.evite.com	0.02	48.28	2212.43	48.25	54.76	2509.68	54.74	48.28	2212.43	48.25
9	beautyandskincares.com	0.03	18.19	642.87	18.16	8.48	299.94	8.46	11.7	413.49	11.67
10	mmajunkie.com	0.01	1.36	201.64	1.35	26.67	3955.5	26.66	13.33	1977.74	13.33

Considering this RTB exchange only to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	m.worldnow.com	4.34	13.27	3.06	8.93	54.39	12.53	50.05	31.65	7.29	27.31
2	m.tmz.com	3.19	6.89	2.16	3.7	23.08	7.24	19.89	10.68	3.35	7.49
3	m.topix.com	1.92	15.07	7.84	13.15	53.45	27.81	51.53	33.75	17.56	31.83
4	babycenter.com	1.94	6.87	3.54	4.93	43.48	22.42	41.54	17.74	9.15	15.8
5	meetme.com	0.78	6.02	7.77	5.25	54.72	70.57	53.94	27.85	35.92	27.07
6	apps.facebook.com	0.86	3.1	3.62	2.24	9.09	10.63	8.24	2.76	3.22	1.9
7	itunes.apple.com	5.24	25.77	4.92	20.53	51.43	9.81	46.19	29.49	5.63	24.25
8	app.evite.com	0.29	48.28	165.58	47.98	54.76	187.82	54.47	48.28	165.58	47.98
9	beautyandskincares.com	0.38	18.19	48.11	17.81	8.48	22.45	8.11	11.7	30.95	11.32
10	mmajunkie.com	0.09	1.36	15.09	1.27	26.67	296.03	26.58	13.33	148.01	13.24

Table 17 RTB 1 and 2: Percentage of network neighbors that also visit the publisher, plus corresponding lift and leverage values. Results are given for the complete dataset ($\mathcal{D}$), the complete neighborhood of a user who visited the given publisher (All NN), and considering the neighborhood of closest 1 and 10 users ($\mathcal{N}_1$ and $\mathcal{N}_{10}$).

RTB 3

Considering all RTB exchanges to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Conversion Exchange	5.53	34.54	6.24	29.01	64.71	11.69	59.17	50.77	9.18	45.24
2	Pinger. Inc.	2.96	46.02	15.53	43.06	65.12	21.97	62.15	51.48	17.37	48.52
3	Top Game Developer	4.29	47.24	11.02	42.95	60	13.99	55.71	48.97	11.42	44.68
4	PremiumEntertainmentApp-FamilyOriented-Android	0.59	23.79	40.12	23.2	20.45	34.49	19.86	22.25	37.52	21.66
5	Enflick	0.53	11.71	21.95	11.17	21.57	40.44	21.04	12.27	23	11.73
6	Clapfoot Games	0.44	8.55	19.32	8.11	11.54	26.06	11.1	13.99	31.61	13.55
7	FunPokes. Inc.	0.18	11.11	61.28	10.93	50	275.77	49.82	15.79	87.08	15.61
8	Talkatone	0.15	13.13	88.51	12.98	31.11	209.76	30.96	17.91	120.78	17.77
9	Sevenlogics. Inc.	0.08	2.71	32.59	2.63	5.88	70.74	5.8	8.33	100.22	8.25
10	A Star Software	0.07	3.33	44.86	3.25	23.08	311.24	23	9.09	122.61	9.02

Considering this RTB exchange only to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Conversion Exchange	11.45	34.54	3.02	23.09	64.71	5.65	53.25	50.77	4.43	39.32
2	Pinger. Inc.	6.13	46.02	7.5	39.89	65.12	10.62	58.98	51.48	8.39	45.35
3	Top Game Developer	8.88	47.24	5.32	38.36	60	6.76	51.12	48.97	5.52	40.09
4	PremiumEntertainmentApp-FamilyOriented-Android	1.23	23.79	19.38	22.57	20.45	16.66	19.23	22.25	18.13	21.02
5	Enflick	1.1	11.71	10.6	10.6	21.57	19.54	20.46	12.27	11.11	11.16
6	Clapfoot Games	0.92	8.55	9.33	7.64	11.54	12.59	10.62	13.99	15.27	13.08
7	FunPokes. Inc.	0.38	11.11	29.6	10.74	50	133.22	49.62	15.79	42.07	15.41
8	Talkatone	0.31	13.13	42.76	12.82	31.11	101.34	30.8	17.91	58.35	17.61
9	Sevenlogics. Inc.	0.17	2.71	15.74	2.54	5.88	34.17	5.71	8.33	48.41	8.16
10	A Star Software	0.15	3.33	21.67	3.17	23.08	150.36	22.92	9.09	59.23	8.94

RTB 4

Considering all RTB exchanges to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Pinger Phone - iOS - Conversation	1.01	71.81	70.85	70.8	80.77	79.68	79.76	70.14	69.2	69.13
2	Talkatone iPhone App	0.07	11.38	172.3	11.31	29.03	439.68	28.97	17.02	257.78	16.96
3	DiceWithBuddies	0.13	7.41	59.04	7.28	18.18	144.92	18.06	7.69	61.31	7.57
4	9GAG Reader	0.12	62.13	497.99	62	49.09	393.5	48.97	49.05	393.17	48.92
5	Video Downloader Pro Lite	0.16	20.03	123.15	19.87	22.58	138.84	22.42	17.93	110.27	17.77
6	Rage Comics	0.09	6.9	80.1	6.81	13.33	154.86	13.25	6.9	80.1	6.81
7	Spades 3D Lite	0.06	12.5	225.76	12.44	27.27	492.57	27.22	12.5	225.76	12.44
8	Apalabrados	0.03	77.42	2391.71	77.39	75	2316.97	74.97	79.32	2450.43	79.29
9	Relax Melodies HD	0.02	31.58	2002.29	31.56	50	3170.29	49.98	31.58	2002.29	31.56
10	Celeb Me - PhotoMaker	0	3.23	821.7	3.22	7.14	1819.48	7.14	3.23	821.7	3.22

Considering this RTB exchange only to determine baseline

		$\mathcal{D}$	All NN			$\mathcal{N}_1$			$\mathcal{N}_{10}$		
			$\mathcal{N}$	Lift	Lev.	$\mathcal{N}_1$	Lift	Lev.	$\mathcal{N}_{10}$	Lift	Lev.
1	Pinger Phone - iOS - Conversation	8.51	71.81	8.44	63.3	80.77	9.49	72.26	70.14	8.24	61.63
2	Talkatone iPhone App	0.55	11.38	20.53	10.82	29.03	52.38	28.48	17.02	30.71	16.47
3	DiceWithBuddies	0.09	7.41	82.11	7.32	18.18	201.54	18.09	7.69	85.27	7.6
4	9GAG Reader	0.97	62.13	63.97	61.16	49.09	50.55	48.12	49.05	50.51	48.08
5	Video Downloader Pro Lite	0.44	20.03	45.36	19.59	22.58	51.14	22.14	17.93	40.62	17.49
6	Rage Comics	0.37	6.9	18.79	6.53	13.33	36.32	12.97	6.9	18.79	6.53
7	Spades 3D Lite	0.46	12.5	26.9	12.04	27.27	58.68	26.81	12.5	26.9	12.04
8	Apalabrados	0.21	77.42	371.93	77.21	75	360.31	74.79	79.32	381.06	79.11
9	Relax Melodies HD	0.13	31.58	238.54	31.45	50	377.69	49.87	31.58	238.54	31.45
10	Celeb Me - PhotoMaker	0.03	3.23	97.89	3.19	7.14	216.76	7.11	3.23	97.89	3.19

Table 18 RTB 3 and 4: Percentage of network neighbors that also visit the publisher, plus corresponding lift and leverage values. Results are given for the complete dataset ($\mathcal{D}$), the complete neighborhood of a user who visited the given publisher (All NN), and considering the neighborhood of closest 1 and 10 users ($\mathcal{N}_1$ and $\mathcal{N}_{10}$).

7.2. Ranking users with similar interests

Are the more-similar geo-similarity neighbors of a publisher's visitors even more likely also to be visitors to the same publisher?

To assess whether more-similar geo-similarity neighbors are even more likely to share interests, we compute the measures exactly as in the previous study, except instead of considering all network neighbors, only the 1 ($\mathcal{N}_1$) and 10 ($\mathcal{N}_{10}$) neighbors with the highest scores are considered (see Tables 17-18). First, we must decide which of the 15 measures of link strength to use. Table 19 reports the rank aggregations of the different metrics, as to which provide the best lifts. Specifically, for each of the 40 publishers, each scoring metric provides a lift, and for that publisher the scoring metrics can be ranked (the best being ranked 1, etc.). The rank aggregation is the average rank for a scoring metric across all 40 publishers. Thus, if one metric provided the best lift for all publishers, it would get an aggregated rank of 1.

	Top 1NN			Top 10 NN	
1	Shared Unique IPs	8.1	1	Shared Unique IPs	8.225
2	Boolean - Sharing an IP	8.625	2	Boolean - Sharing an IP	8.913
3	**Shared Unique IPs - IDF weighted**	**7.025**	3	**Shared Unique IPs - IDF weighted**	**7.975**
4	**Shared IP visits - min**	**7.063**	4	**Shared IP visits - min**	**7.413**
5	**Shared IP visits - max**	**7.95**	5	**Shared IP visits - max**	**7.25**
6	Shared IP visits - avg	8.05	6	**Shared IP visits - avg**	**7.488**
7	**Shared IP visits - min - IDF weighted**	**6.625**	7	**Shared IP visits - min - IDF weighted**	**7.375**
8	**Shared IP visits - max - IDF weighted**	**7.975**	8	**Shared IP visits - max - IDF weighted**	**7.225**
9	**Shared IP visits - avg - IDF weighted**	**7.938**	9	**Shared IP visits - avg - IDF weighted**	**7.525**
10	Cosine similarity	8.338	10	Cosine similarity	8.45
11	Cosine similarity - IDF weighted	8.213	11	Cosine similarity - IDF weighted	8.863
12	GSN Jaccard	8.925	12	GSN Jaccard	9
13	GSN Jaccard - IDF weighted	8.325	13	GSN Jaccard - IDF weighted	9.138
14	GSN Jaccard Freq	8.425	14	GSN Jaccard Freq	8.5
15	GSN Jaccard Freq - IDF weighted	8.425	15	GSN Jaccard Freq - IDF weighted	8.5

Table 19 Rank aggregation per metric over all publishers (across all RTB)

To visualize the best metrics, the metrics with an average ranking less than 8 are shown in boldface. The Jaccard metrics do not perform well in this study either, showing that penalizing the connection to users that visit many locations is not sensible. Remember that the Jaccard metric is based on the idea that users that visit many locations will have a higher chance to share some location with another user. However, we have so many locations in total that visiting a couple more locations will increase only marginally this probability of sharing a location by chance. A user that logs into more IPs and is more active should hence not be penalized and ranked lower than other less active users. Metric 3, which sums the IDF scores of the shared locations performs quite well. Given the operational efficiency advantage over the frequency based metrics, this metric is chosen to measure the strength of the links.

As seen from the results in Tables 17-18, the lift and leverage values for the strongly connected (top 1 and top 10) network neighbors are even higher, sometimes astronomical (3,170 for Relax melodies HD (RTB 4)). These results are highly significant by sign tests comparing either the lift or the leverage between the top-ranked neighbors and the entire cohort ($p<0.01$). The actual values again indicate that the users with the strongest geo-similarity are often substantially more likely to visit the same publisher.

Thus, we can conclude that not only does geo-similarity find users with similar interests, it also *ranks* users well by their likelihood of having similar interests.

An interesting follow-up question is whether the characteristics of a publisher are of importance for the results. For example, are GSN neighbors more similar in terms of using the same social network app as compared to playing the same games? To answer this question, the publishers are categorized into six classes: funny, social, news, games, communication and miscellaneous. In Figure 9, the lifts of the publishers are shown per category. Some general trends can be observed, where publishers related to funny content have the highest median lift. This could be explained by the fact that users often forward funny content to their friends. The social websites/apps follow next, which seems to further demonstrate that friends are likely linked in the GSN. However, publishers in the communication category perform not so well compared to the other categories and rather high variances are observed in the different categories. Please note that high variance is observed across the categories; therefore, it is difficult to make well-supported claims from these results.

7.3. Timing

As a final evaluation, we explore whether the inclusion of temporal similarity in location visitation is helpful for defining GSN connections. To this end we introduce a setting where two users are connected only if they visit an IP in the same time period of a day. Specifically, we divide a day into different time windows—for example two time periods of 12 hours—where users are connected only if they visit the same IP in the same time period of a day (which might be on different days). Practically, two digits are added to an IP which denote the starting moment of the corresponding time period. We report on time periods of 2 hours and of 12 hours (other periods yield similar results). We also include the setting where users are connected only when they visit the same IP on the *same day*, in light of the result of Crandall et al. (2010) that visiting the same location at around the same time is indicative of friendship.

The resulting lifts are reported in Figures 10, 11 and 12 when considering (respectively): all network neighbors, the top 1 network neighbor, and the top 10 network neighbors. The results when not considering the time period are repeated as well (All), limited to the conservative case

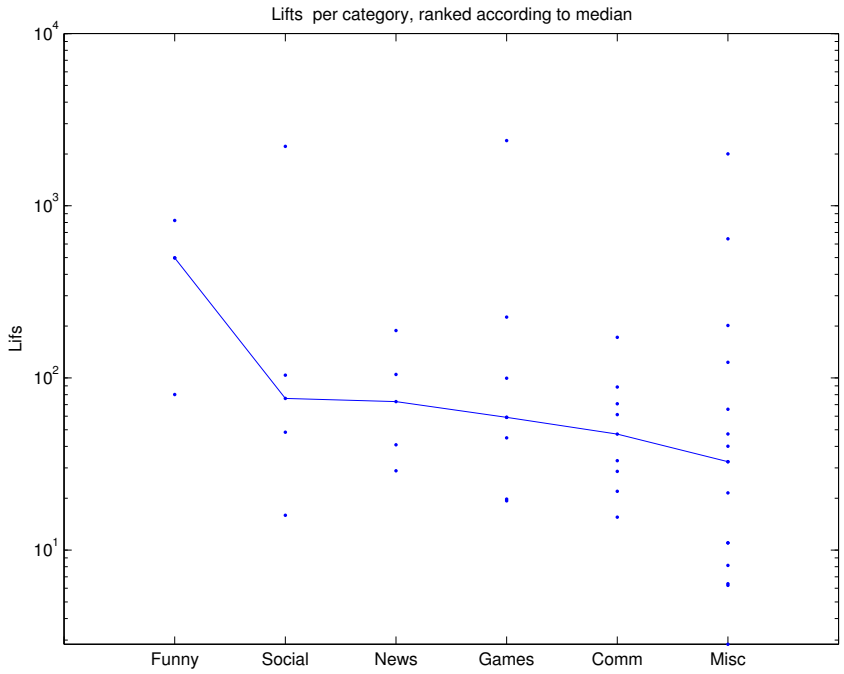


Figure 9 **Lifts of the publishers (for all four RTBs), ranked according to the median lift within the category (the medians are shown by the full line).**

	All	TP2	TP12	Day
RTB 1	**1.1**	2.9	3.9	2.1
RTB 2	**1**	2.4	3.4	3.2
RTB 3	**1.1**	3	4	1.9
RTB 4	**1.1**	2.8	3.8	2.3

Table 20 **Rank aggregations in terms of lift averaged across all publishers (per RTB). Linking based on time does not improve over linking ignoring time. All: use all data for location profile; TP2: use two-hour windows for location profile; TP12: use 12-hour windows for location profile; Day: use same-day window for location profile.**

	All	TP2	TP12	Day
RTB 1	-	*0.002*	*0.002*	*0.006*
RTB 2	-	*0.002*	*0.002*	*0.002*
RTB 3	-	*0.002*	*0.002*	0.065
RTB 4	-	*0.002*	*0.002*	*0.004*

Table 21 **Do designs that restrict similarity to specific time windows perform significantly worse than those that do not limit by time? Cells show p-values of signed rank tests. All: use all data for location profile; TP2: use two-hour windows for location profile; TP12: use 12-hour windows for location profile; Day: use same-day window for location profile.**

of only considering one exchange (see above). The baseline (lift of one) is indicated with a red horizontal line (near the horizontal axis in every case); lifts higher than 1000 are shown as 1000 to limit the range, in order to be able to visualize the results. Tables 20 and 21 show the average rank aggregation (see the discussion of rank aggregation above) in terms of lift per RTB and the

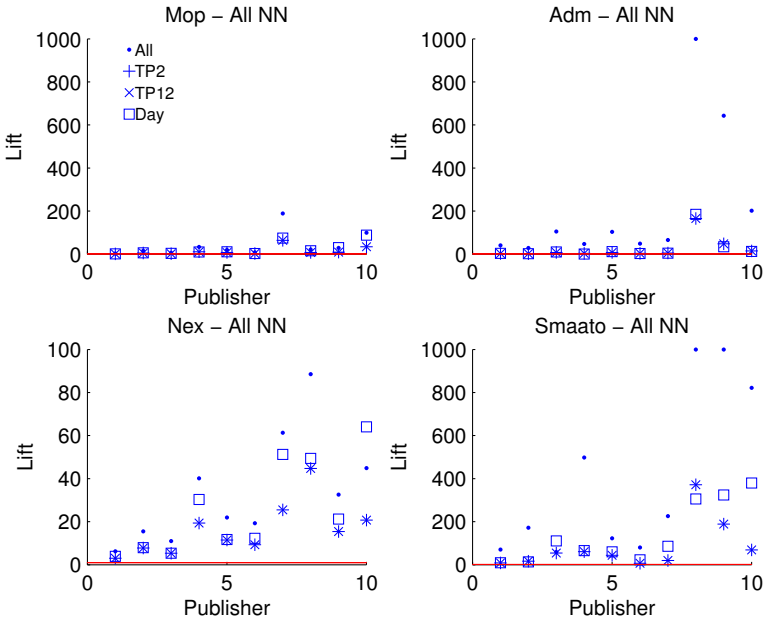


Figure 10 **Lifts, considering all network neighbors (NN), per RTB system (each panel), with and without the time constraint, limited to one network (cf. 'Considering one RTB exchange only' in Section 7.1), for the 40 publishers. Lifts higher than 1000 are shown at 1000. All: use all data for location profile; TP2: use two-hour windows for location profile; TP12: use 12-hour windows for location profile; Day: use same-day window for location profile.**

p-values of a signed rank test showing whether not limiting location profiles to a time window (All) performs significantly better, respectively.

When comparing the different temporal settings, we see that the 'Day' variant performs best, and limited differences exist between the time periods of 2 and 12 hours. More important to observe from these results is that (overall) not including time in the GSN design outperforms all three temporal variants substantially. Only for RTB 3 and the Day version, where users are connected only if they visit the same IP on the same day, are the lifts not significantly worse at the 1% level than having no time constraint (although even here they are significantly worse at the 10% level). This shows that for finding users with similar interests, knowing which locations they visit is more important than knowing when they visit them. By including the time constraints, connections between previously connected users are lost simply because they visit locations in different time periods. Although we use time to define links, metrics that include time to define the strength of a link might improve the results further and we consider that an interesting issue for future research.

8. Conclusions, Limitations & Implications

This paper presented a new design for using geo-similarity to connect user instances in a geo-similarity-weighted network. In the GSN, users are connected if they share at least one observed

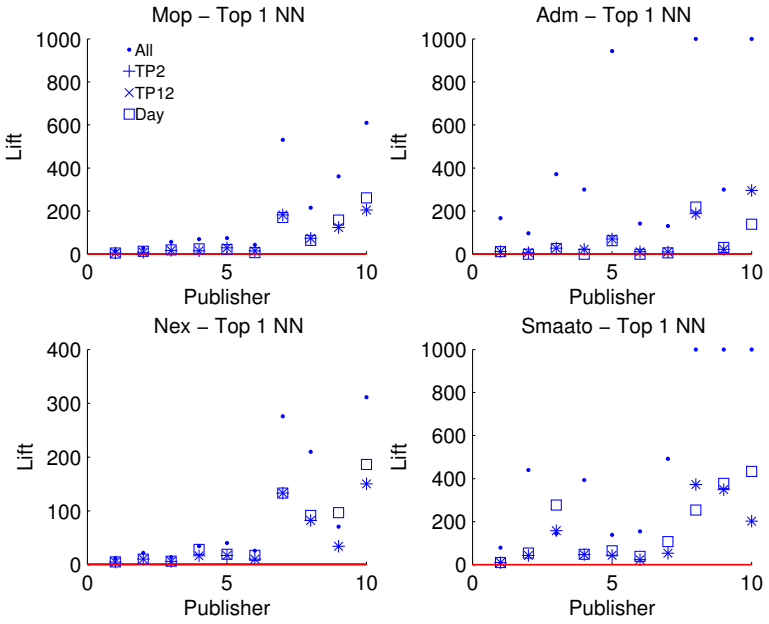


Figure 11 **Lifts, considering the closest 1 network neighbor (NN), per RTB system (each panel), with and without the time constraint, limited to one network (cf. 'Considering one RTB exchange only' in Section 7.1)), for the 40 publishers. Lifts higher than 1000 are shown at 1000. All: use all data for location profile; TP2: use two-hour windows for location profile; TP12: use 12-hour windows for location profile; Day: use same-day window for location profile.**

location. Various metrics can be used to yield degrees of similarity. We presented intuitive arguments, theory, and prior research that suggest that geo-similarity and the GSN design should link similar users. We also provided strong empirical support. Specifically, Study 1 shows that the GSN indeed links different users instances corresponding to the same individual, and the geo-similarity metrics rank user instances by their likelihood of corresponding to the same individual. Study 2 shows that the GSN links users who have similar interests, as measured by their propensity to visit the same publishers/user the same apps. It also provides strong evidence that the geo-similarity metrics rank user instances by their likelihood of having similar interests. Taken together, the results provide strong support that the geo-similarity network links users with similar interests—in many cases because they actually represent the same individuals.

The main limitation of the research presented in this paper is that the data are not sufficient to distinguish whether the similarity in interests is solely based on connecting instances of the same user. This is an important avenue for future research. This limitation notwithstanding, the very strong results from Study 2 have important implications in either case: either the network is indeed finding different individuals with substantially similar interests, or we have very strong additional

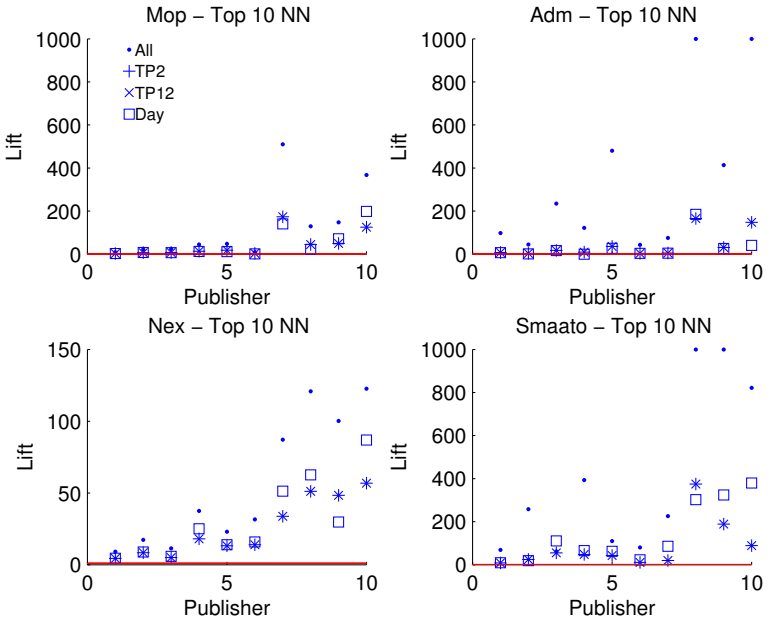


Figure 12 **Lifts, considering the closest 10 network neighbors (NN), per RTB system (each panel), with and without the time constraint, limited to one network (cf. 'Considering one RTB exchange only' in Section 7.1), for the 40 publishers. Lifts higher than 1000 are shown at 1000. All: use all data for location profile; TP2: use two-hour windows for location profile; TP12: use 12-hour windows for location profile; Day: use same-day window for location profile.**

support for the ranking results of Study 1, as the neighbors with the strongest geo-similarity have significantly higher affinity for the same publishers/apps.

Another area of limitation is the potential sparsity of the data, depending on the particular locations chosen for the implementation of the location-profile design. For example, when using (anonymized) IP addresses as the locations, we find in our data sample a large number of user instances with only one location. This results in a highly fragmented network (with many disconnnected components). This is not necessarily detrimental to the finding of user instances corresponding to the same user or similar users, because (i) multiple user instances may share this single location, and thus be similar (e.g., all the instances of the devices that never leave my apartment), (ii) these user instances may also share this location with a user instance that has more than one location (my laptop), and (iii) if so, the multi-location user instance may tie these users to other users and devices. If we want to take advantage of the third case (iii), we would need to extend the design to include indirect connections in the network (for example, using short paths between user instances rather than direct connections). A situation more troublesome is when there are many users who only have a single location *and those locations only have a single user*. Such users would not be connected to anyone else in the network either directly or indirectly, and thus would not

be similar to any other users in the network. If the number of such users is problematically large, then the choice of location data points (IPs in our implementation) is too fine-grained. The method is not limited to using IP addresses; they just form a particularly intuitive and accessible choice. Instead, actual latitude/longitude coordinates or IP addresses with known or inferrable geographic locations could be generalized into somewhat larger geographic areas, and these areas could then form the coordinates of the location profiles, providing richer connections in the geo-similarity network.

These results have broad and immediate management implications within our motivating application of mobile advertising. As discussed above, digital marketers would like to target the many fragmented instances of individuals in the digital advertising ecosystem, as well as users who have similar interests to particular "seed" users. Location data, such as IP addresses, are readily available across the advertising ecosystem. The most direct use of the GSN for mobile ad targeting is to seed a targeting campaign with users chosen to exhibit some characteristics of interest. These could be chosen through any of the myriad methods currently used in advertising. The GSN will directly allow the targeting of the other members of these users' geo-similarity cohorts, which will include other instances of the same user, other similar users, and likely both. So, for example, if a set of users has been identified to have brand affinity via visits to a brand's website, the GSN cohorts of these users would provide an attractive avenue to expand the reach of a campaign to target them (extending traditional retargeting) and also target users similar to the seeds (customer prospecting). In the research data set we see that different instances of the same users are very often connected in the GSN, and that those connected in the GSN exhibit substantially similar interests.

The GSN also could be used for (privacy-friendly) hyper-local targeting—meaning, targeting people who frequent a particular location, without needing to store data on the actual locations of the users, as described earlier. IP addresses currently are used by some marketers for coarse-grained demographic targeting: inferring geography from IP address registration data, and then inferring demographics from the geography. The geo-similarity network provides a complementary alternative: it will connect people who visit the same fine-grained locations. As described above in the couponing example, if a campaign is seeded with customers of a local establishment (e.g., via an online loyalty program), geo-similarity can target others who frequent the same locations.

An exciting potential future use of the GSN is the evaluation of marketing campaigns across different channels. This has become an important for mobile advertizing.[15] For example, I might receive an ad for a hot new product on my mobile device. I may be interested in the product,

[15] http://technorati.com/business/advertising/article/taking-on-the-top-3-cross/

but I'm unlikely to buy this product on my mobile phone. Rather, I'll buy the product using my laptop or PC. By linking the screens of the same user, we are now able to provide a broader (and possibly much more robust) estimate of campaign success (e.g. conversion rate) across different channels by aggregating the success metrics across GSN neighbors. For example, similarly to how a traditional ad campaign may run/not run an ad in a controlled study across different cities, agencies could target/not-target different (matched) users and then look at the success rates in their respective geo-similarity neighborhoods. This should capture effects of the same individual on different devices.

As discussed at the outset, advertisers need to be cognizant of privacy concerns regarding the collection, storage, and use of data. This paper follows what the FTC calls the "privacy by design" approach. By design, the technique does not need to store any direct personal information about mobile users; there is no need for non-anonymized identifiers, demographics, geographics, psychographics, etc. In addition, the storage of indirect information about users can be severely limited as well. The method does not need the actual locations—only anonymized location "keys." So, for example, IP addresses can be replaced with random numbers[16] without affecting the GSN performance. More technically, at the "outer wall" of the system or firm, each device id can be irreversibly hashed to a random key. The only requirement is that the same device be hashed to the same key if encountered again. Similarly, at the "outer wall" of the system or firm, every location also can be irreversibly hashed to a random key. The geo-similarity network can be formed just the same with the random keys as with the actual locations. If more privacy is desired, hashing can be done irreversibly many-to-one, and in that case it becomes impossible to associate definitively any particular location with any particular device/user.

Finally, an aspect of this research that is important for managers, but not typically covered in the predictive modeling literature, is the notion of increasing the *reach* of a campaign. For example, as discussed above, possibly the most straightforward use of the geo-similarity network is to select the same actor on different mobile devices. This would allow us to expand the reach of a "retargeting" campaign.[17]

Increasing reach has important subtleties in the current advertising ecosystem—it is not just the other side of the coin of increasing lift. The reason is that there are many targeters in the online/mobile advertising ecosystem all of whom are using the same data: retargeting data and

[16] See the FTC's 2010-2011 privacy-by-design report and the comments thereon, along with Provost et al. (2009) for a thorough treatment of this topic.

[17] Retargeting is the targeting of browsers who have previously purchased from the brand or who have taken some other indicative brand action, such as browsing the brand's site. Retargeting is considered by many to be one of the most effective targeting strategies. (Albeit to our knowledge these conclusions are drawn based on assessing click or conversion rate rather than the actual influence of the advertising. Please see Stitelman et al. (2011).)

demographic/geographic/psychographic data that are purchased from third-party data providers. However, these data are available only on a subset of devices. Thus all parties who are using these data are competing for the same, sometimes small, set of devices. Since large advertisers typically contract with multiple targeting firms, the effect is that the advertiser is paying the firms to compete against each other, effectively raising the price in the auction, and thereby raising their own cost of advertising![18] What's more, these will be the same devices that also will be targeted for other campaigns, because they are the devices for which the targeters have data.

However, by connecting devices in a geo-similarity network we can expand campaigns to devices for which there is no retargeting or third-party data available at all. If there is less competition for these devices, we should be able to target them for a lower price. Thus expanding reach has implication for the cost-effectiveness of achieving a certain level of predictive performance (e.g., lift).

Acknowledgements

We thank Lauren Moores, Tom O'Reilly, and the teams at Everyscreen Media and Dstillery for many discussions and debates. Foster Provost thanks Andre Meyer for a Faculty Fellowship. We also thank the Moore and Sloan Foundations for their support of the Moore-Sloan Data Science Environment at NYU. This research was conducted while Foster Provost was at Coriolis Labs.

References

Agarwal, Ritu, Anil K. Gupta, Robert E. Kraut. 2008. Editorial overview - the interplay between digital and social networks. *Information Systems Research* **19**(3) 243–252.

Aral, Sinan, Lev Muchnik, Arun Sundararajan. 2009. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences* **106**(51) 21544–21549.

Bampo, Mauro, Michael T. Ewing, Dineli R. Mather, David Stewart, Mark Wallace. 2008. The effects of the social structure of digital networks on viral marketing performance. *Information Systems Research* **19**(3) 273–290.

Cho, Eunjoon, Seth A. Myers, Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11, ACM, New York, NY, USA, 1082–1090.

Cortes, Corinna, Daryl Pregibon, Chris Volinsky. 2001. Communities of interest. Frank Hoffmann, David Hand, Niall Adams, Douglas Fisher, Gabriela Guimaraes, eds., *Advances in Intelligent Data Analysis*, *Lecture Notes in Computer Science*, vol. 2189. Springer, 105–114.

[18] The latter is our conjecture.

Crandall, David J., Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, Jon Kleinberg. 2010. Inferring social ties from geographic coincidences. *PNAS* **107**(52) 22436–22441.

de Montjoye, Yves-Alexandre, César A. Hidalgo, Michel Verleysen, Vincent D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* **3**.

Fawcett, Tom. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* **27**(8) 861–874.

Heidemann, Julia, Mathias Klier, Florian Probst. 2010. Identifying key users in online social networks: A pagerank based approach. *Information Systems Journal* **4801**(December) 157–160.

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

Hill, Shawndra, Foster Provost, Chris Volinsky. 2006. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science* **22**(2) 256–276.

Hotho, Andreas, Andreas Nürnberger, Gerhard Paass. 2005. A brief survey of text mining. *LDV Forum* **20**(1) 19–62.

Kerho, Steve. 2012. Mobile marketing - a new analytics framework, what we have & what we need. Presented at the Marketing on the Move Conference at The Wharton School, Philadelphia.

Kossinets, Gueorgi, Duncan J. Watts. 2009. Origins of homophily in an evolving social network. *AJS* **115**(2) 405–450.

Martens, D., F. Provost. 2011. Pseudo-social network targeting from consumer transaction data. NYU Stern School of Business Working Paper No. CEDER-11-05 .

McPherson, Miller, Lynn Smith-Lovin, James M. Cook. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* **27** 415–444.

Oinas-Kukkonen, Harri, Kalle Lyytinen, Youngjin Yoo. 2010. Social networks and information systems: Ongoing and future research streams. *Journal of the Association for Information Systems* **11**(2) 61–68.

Pan, Wei, Nadav Aharony, Alex Pentland. 2011. Composite social network for predicting mobile apps installation. Wolfram Burgard, Dan Roth, eds., *Proceedings of AAAI 2011*. AAAI Press.

Perlich, Claudia, Brian Dalessandro, Troy Raeder, Ori Stitelman, Foster Provost. 2014. Machine learning for targeted display advertising: Transfer learning in action. *Machine Learning* **95**(1) 103–127.

Provost, Foster, Brian Dalessandro, Rod Hook, Xiaohan Zhang, Alan Murray. 2009. Audience selection for on-line brand advertising: privacy-friendly social network targeting. *KDD '09: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 707–716.

Provost, Foster, Tom Fawcett. 2013. *Data Science for Business: What You Need to Know About Data Mining and Data-analytic Thinking*.

Pubmatic. 2010. Understanding real-time bidding (RTB) from the publisher's perspective. Tech. rep., Pubmatic.

Quercia, Daniele, Neal Lathia, Francesco Calabrese, Giusy Di Lorenzo, Jon Crowcroft. 2010. Recommending social events from mobile phone location data. *Proceedings of the 2010 IEEE International Conference on Data Mining*. ICDM '10, IEEE Computer Society, Washington, DC, USA, 971–976. doi:10.1109/ICDM.2010.152. URL `http://dx.doi.org/10.1109/ICDM.2010.152`.

Raeder, Troy, Ori Stitelman, Brian Dalessandro, Claudia Perlich, Foster Provost. 2012. Design principles of massive, robust prediction systems. *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1357–1365.

Stitelman, O., B. Dalessandro, C. Perlich, F. Provost. 2011. Estimating the effect of online display advertising on browser conversion. *Proceedings of the Fifth International Workshop on Data Mining and Audience Intelligence for Advertising (ADKDD 2011)*. 8–16.

Stitelman, Ori, Claudia Perlich, Brian Dalessandro, Rod Hook, Troy Raeder, Foster Provost. 2013. Using co-visitation networks for detecting large scale online display advertising exchange fraud. *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1240–1248.

Sundararajan, Arun, Foster Provost, Gal Oestreicher-Singer, Sinan Aral. 2013. Information in digital, economic, and social networks. *Information Systems Research* **24**(4) 883–905.