Geo-Social Targeting for Privacy-friendly Mobile Advertising: Position Paper

Foster Provost
Stern School of Business, NYU and Coriolis Labs
fprovost@stern.nyu.edu

-
NYU/Stern Working Paper CeDER-11-06a
June 9, 2011

1. INTRODUCTION

This position paper is about methods for effective, privacy-friendly mobile advertising. Specifically, we propose a new social-targeting design for using consumer location data from mobile devices (smart phones, smart pads, laptops, etc.) to target advertisements in a manner that is both effective and privacy friendly. Both of these attributes are important. In traditional online advertising we see evidence that targeting based on recent advances in data modeling is much more effective than traditional methods for online targeting. On the other hand, just recently we have seen the sort of uproar that arises from the idea that our location behavior is being "tracked" by our mobile technology.[1] Therefore, marketers who dream of location-driven targeting should think carefully about what the FTC is calling "privacy by design,"[2] and consider what options can provide effective advertising with minimal data collection and storage.[3]

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. [6] showed the remarkable effectiveness of *social-network targeting*: targeting consumers who are linked to known customers by a social network. Subsequently, Facebook (and others) have attempted to implement social-network targeting for online advertising, with varying degrees of success.

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" social-network relationships. The design of Provost et al. [12] is an example of social targeting that is not "true" social-network targeting: 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).

This position paper addresses a new social-targeting design focused on mobile advertising. It will identify audiences for targeting based on their proximity in a *geo-social network* induced from aggregated, anonymized location information. The exact methods for choosing good targets will be a subject of this research, and will depend on the sort of audience desired, but the methods will apply to targeting audiences for maximizing conversions, for maximizing the increase in conversions due to the advertising, or for increasing (observed) brand affinity [12].

The expected contributions of this research are:

- To introduce designs for geo-social targeting. We will present initial design ideas below. A key element of the designs is that they can be surprisingly privacy-friendly, a topic to which we devote a section of this paper.
- To present theoretical motivation for why geo-social targeting is likely to be effective for advertising targeting. We begin this argument in the next section.
- To build actual geo-social networks among (anonymized) consumers, and analyse these networks, reporting on their structure and properties.
- To conduct a careful evaluation of whether and to what extent these geo-social targeting techniques are effective for selecting audiences for one or more advertising tasks, such as finding consumers more likely to convert or finding audiences with elevated brand affinity.

We do not have preliminary results on mobile data at this point to provide support that this proposed line of research is a good idea. Instead, we will argue (next) theoretically that

[1] http://pogue.blogs.nytimes.com/2011/04/28/wrapping-up-the-apple-location-brouhaha/
[2] http://www.ftc.gov/opa/2010/12/privacyreport.shtm
[3] In this position paper, I am taking no ethical stance on what is the right level of data storage and use. I am proposing that we broaden the set of choices that firms have by creating novel, effective, privacy-friendly designs.

Research conducted at Coriolis Labs in collaboration with Alan Murray.

it *should* be a good idea. The main contribution of the proposed work is the design science: introducing a new, general design, tying it into the existing literature, and providing one or more solid, convincing evaluations. And hopefully providing interesting auxiliary insights.

2. MOTIVATION & RELATED WORK

In short, the idea for composing a mobile audience for targeting is: to find individuals who are closely linked in a geo-social network to individuals we know already to have the characteristic that we desire. For example, target the set of devices most closely linked to devices whose users are known to exhibit brand affinity or to be responsive to ads. The geo-social network will be very fine-grained, for example based on shared (anonymized) IP addresses, small GPS "cells", etc. For mobile devices, this would mean linking two "individuals" (devices) based on having observed that they both have been active on the same IP addresses (for example).

Why would this be a good idea? There are several reasons. Let's call mobile devices that are (directly) linked in the geo-social network "(first-degree) network neighbors."

First of all, first-degree network neighbors share at least one IP address, and possibly several. As the number of shared IP addresses grows, we conjecture that the likelihood increases that the two devices actually belong to the same person—drawing an analogy to results showing that different instances of the same person call the same phone numbers [2], cite the same references [5], and visit the same web sites [12]. It seems intuitive that similar techniques to those used for linking different instances of the same person in these domains would apply to fine-grained location data. For example, who besides me is observed primarily on my home IP address and my office IP address, let alone my favorite coffee shop.

Second, geographic targeting already is used widedly, albeit not via social targeting, because it is a proxy for demographics and other predictive features. One difference in this proposed work is that we don't choose the geographies to target, but instead use them implicitly. This allows the actual locations to be anonymized. It also allows the use of location information that is too fine-grained to specify explicitly. Furthermore, it allows us to use location information that is too fine-grained even to include in normal 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 less of device profiles), as well as transient wifi locations that only connect two devices for a brief time. We should be able to handle arbitrarily large numbers of locations (e.g., IP addresses), limited only by our ability to store them, rather than our ability to model with a huge, sparse feature space. For example, we foresee no major problem with geo-social networks based on a billion or more unique locations.

Third, fine-grained location information is likely to contain more detailed (latent) information than standard geographic information. Not only would it link devices by wealth, income, demographics, etc., it may well link by employer, educational institution, interests, community, and even shopping habits. Thus, we conjecture that geo-social 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 so effective specifically for on-line advertising [12]. We might call this "locale-affinity" social targeting.

If that weren't enough, recent research provides yet another reason to expect that geo-social targeting may be especially effective. In an article a few months ago in *PNAS*, Crandall et al. show that geographic coocurrences 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" [3]. This means that a geo-social network not only would capture the advantages of geographic targeting and locale-affinity targeting, it would also incorporate actual social-network targeting—also shown to be extremely effective for marketing [6]. In fact, when massive descriptive data is available, the (usually latent) similarity between social-network neighbors has been shown to explain much of the marketing advantage previously attributed to social influence [1].

Moreover, interestingly, recent research also has shown that the homophily [10] that has been used to explain the effectiveness of social-network targeting, may actually be due largely to the constraints placed by opportunity [8]. 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" [8]. Why is this important for the present proposal? 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 observe [8], " 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 the proposed geo-social network. Thus, the geo-social 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. PRELIMINARY DESIGN

Being design-science research, the design of the geo-social targeting methods will take shape as part of the research project. However, we can outline a preliminary design. The basis for the design will be to create a geo-social network, and then to use the network for predictive inference. The elements of the design include: how will entities (mobile devices) be represented? What will be the geo-social links? And how will the predictive inference be conducted? First we will present our proposed data scenario, in order to make the discussion more concrete. The design should generalize beyond the specific data setting envisioned for this proposed project.

3.1 Proposed data setting for mobile advertising

For this project, we propose to take advantage of the data available currently in the online advertising ecosystem. Specifically, via ad exchanges and online bidding systems we

can observe a massive number of mobile devices, along with the data made available for advertisers to decide whether to bid on the device for a particular campaign. Two aspects of the data are crucial: (1) each mobile device is associated with a key; when this key is observed in the future, we know that the subsequent action involves the same mobile device. (2) Part of the data associated with each mobile device is a location. For this writeup, we can consider that location to be the IP address of the wifi network currently in use (although there are various sorts of location data). We will observe the locations visited by each device, along with their recency and frequency.

For the research we also will observe data on brand actions (conversions, website visitations, etc.) for a subset of the mobile devices. These brand actions will be employed in part to inform the inference (e.g., to propagate through the network or to train predictive models), and in part to evaluate the techniques (were devices that were predicted to be more likely to take certain actions actually more likely to take the actions?).

3.2 Entity representation

Each mobile device will be represented by a distribution of behavior across locations. This will necessarily be a sparse representation, as we expect a very large total set of locations, but we also expect each individual to be observed at only a small subset of the locations. Furthermore, we are prepared to restrict the location set to be even sparser, as necessary. For example, a device can be represented by the distribution across the current "top-k" locations, plus a catch-all "other" location. There are some technical approximations involved with maintaining the distributional picture over time if we do not want to save the complete history, but it seems unlikely that they will affect the outcome of the research. With this representation, we can enforce any degree of sparseness desired, while at the same time maintaining a picture of the devices' most representative locations (within the error of the approximation).

3.3 Link representation

The structure of the geo-social network is the graph with the devices as the nodes, and links between the devices as the edges. The corresponding design decisions involve the selection of which devices to link at all, and the weights to place on the links. We will experiment with various designs.

The simplest link representation would be unweighted links between all pairs of devices that share a specific location. More sophisticated link representations would comprise variations on both the weighting criteria and the link selection criteria:

- Given two devices that share at least one location, the weight of the link between them can be a function of the devices' location distributions. Simple measures include simply the number of shared locations, and the number of shared locations weighted by their visitation percentage in the profile. A more sophisticated measure of similarity would be to compute a measure of similarity between the two distributions, such as the Kullback-Leibler divergence, the Jensen-Shannon divergence, cosine similarity, or other similar measures.

- Links can be weighted according to their (lack of) popularity. For example, it might be argued that the link formed by sharing a location with a small number of other devices (my apartment) indicates a stronger "location affinity" than sharing a location with a massive number of other devices (the Starbucks on the corner of Washington Square Park). This intuition can be extended: if two devices spend a lot of time at such an unpopular locale (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, they they're either the same person or close friends (or soulmates). 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 DFILF (device frequency * inverse location frequency). Then (2) the strength of the link between two devices that share a location would be a function of the corresponding DFILF scores.

- For link selection (and also to aid in deciding which links to use in a device's distribution), we could create campaign-specific geo-social weightings by creating "supervised" location scores. For example, we could weight locations by a likelihood ratio—how likely they are to have been visited by positive rather than negative devices, where positive devices are those who have been observed in the past to have taken an action of interest (e.g., purchasing).

- For any weighting, possibly only the links with the largest weights would be selected, as the low weight links may simply incorporate noise in the inference procedure. This will be a subject of research.

- In building the device-specific location distributions, we may want to factor in recency in addition to frequency. This is important both to reflect the inevitable changes in mobile behavior, and also for practical reasons, as we would prefer not to have to manage location data indefinitely. An elegant way to incorporate recency without having to explicitly manage timestamps and time windows is to update a device's location distribution by an additive process with exponential decay [4]: a new observed location is formulated as a distribution d_n with all the mass on the single new location. This is then added to a decayed version of the existing distribution d_t to yield:

  $d_{t+1} = \lambda * d_t \oplus (1-\lambda) * d_n$, with $\oplus$ denoting component-wise addition (with renormalization if we want it to remain a true distribution).

3.4 Geo-social inference

Once we have one or more geo-social networks, there are various ways to take advantage of them for inference (e.g., for predictive targeting).

The simplest method for inference is simply to target the geo-social network neighbors: all of them, or the "closest" based on one or more notions of network proximity (as discussed above).

More sophisticated inference can be accomplished by using the geo-social network to create predictive features, that then are incorporated in a (higher-level) predictive model.

For example, different notions of proximity to known "brand actors" can be incorporated in a logistic regression model, which then would learn how best to weight the different measures for different campaigns. This would be similar to the techniques used for learning brand proximity in content-affiliation networks, trained based on actual brand affinity data [12]. Furthermore, in such a model, the network proximity measures could be combined with other network measures. For example, measures of local network structure, connections to "influencers," etc., could improve the ranking of the network neighbors (as seen for social-network targeting [6]).

Possibly most interestingly for the geo-social network, we can extend our inference out beyond the first-degree network neighbors. This may be critical if the locale-affinity network is sparse (either naturally or because we enforce sparseness, as discussed above). There are a variety of methods for propagating predictive information through network data, which have come to fall under the general rubric of "collective inference" (named because generally they collectively infer about a set of nodes simultaneously, using the *inferences* about nodes to affect each other). Various techniques have been used for collective inference, include random field models, belief propagation, relaxation labeling, MCMC techniques, iterative classification, and others. The description of these methods is beyond the scope of this position paper, but can be found (along with a wealth of references) elsewhere [9, 11]. To our knowledge, the only published work describing the application of such methods to a real, large-scale marketing problem was by Hill, Provost and Volinsky [7] (we would be happy to learn about other applications, if such exist).

For this project, we will explore how to design network inference methods that go beyond the immediate network neighbors. We will start with two techniques (1) simple belief-propagation-inspired methods, and (2) an approximate Gaussian random field. Given the dearth of work on applying collective inference to very large-scale consumer data, this choice is based on our intuition as to probable effectiveness having studied these methods for a decade, and also that these methods are likely to be the most efficient to run—not a trivial issue for massive networks (like our geo-social network). We foresee that we will need to make various approximations to optimize the collective inference. This will be a major focus of research, in the case where looking only at first-degree neighbors gives less reach than we would like.

4. PROPOSED PLAN AND STATUS

Our plan is straightforward:

1. We will gather data on mobile devices' location visitation histories and construct one or more geo-social networks.
2. We will further develop and implement the geo-social targeting techniques.
3. We will conduct a rigorous evaluation across several advertising campaigns, to assess the effectiveness of geo-social targeting. We hope to be able to assess effectiveness both for direct marketing-style advertising and for brand advertising. The evaluation is discussed further below.
4. We will conduct followup analyses to work to understand the results more deeply. It is difficult to predict what these analyses will be before seeing the main results. However, we can compare with a prior social-targeting study [12] to see examples of these sorts of followup analyses. There after showing the content-affinity targeting resulted in significant lift, and more interestingly, that closer brand proximity led to higher lift, we also (i) showed that browsers that were close by the brand proximity actually also were friends, (ii) showed that the close network neighbors of a brand's customers had very similar demographic profiles, and (iii) conducted controlled experiments to show that browsers with higher brand proximity have higher "organic" brand affinity than randomly targeted browsers (by targeting both with public service announcements, PSAs, rather than with ads, in a reversal of the usual controlled PSA tests to show the influence of the advertising).

4.1 Current status

The current status is: as demonstrated by this position paper, we have worked through the preliminary theoretical motivation and ties to the literature, and have developed a preliminary technical design. On the data front, we have reached agreements for procuring the data that we will need. We have partnered with one of the largest online bidding systems to obtain data on (anonymized) mobile device location history.

Separately, we have partnered with one of the largest online ad targeting firms to obtain brand action data on (a subset of) these mobile devices. This firm has data on brand actions useful for evaluating both direct-marketing-style online advertising (viz., conversions), as well as brand actions useful for evaluating online brand advertising (e.g., visits to the brand's web site, such as the home page, loyalty club page, etc.). Provost et al. [12] discuss in detail the use of such data for evaluating brand advertising, via differences in audiences' "brand action densities."

These data should be sufficient for the project, with the exception of running controlled experiments, discussed more below.

4.2 Evaluation plan

Currently we plan to evaluate our results using design science methodologies, and in particular, using the methodologies that are well accepted in the predictive modeling literature: careful evaluations on holdout data, using measures such as lift and area under the ROC curve (the Mann-Whitney-Wilcoxon statistic). I think the paper resulting from our prior study would be the illustration most analogous to this study [12].

There are some specific aspects of the evaluation here that are not typical. We would like to be able to evaluate not only standard measures like conversion lift, but also lift in the "potential" for brand advertising. The only measurable quantity of which we are aware for this (from data such as ours) is the brand actor density [12]. We would be happy to include additional measures as we become aware of them.

Also, an aspect of this research that is not typical of the predictive modeling literature (although not completely absent) is the notion of increasing the *reach* of a campaign. For example, as discussed above, possibly the most straight-

forward use of the geo-social network is to identify the same actor on different mobile devices. This would allow us to expand the reach of a "retargeting" campaign.[4]

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 ecosystem all of whom are using the same data: retargeting data and 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, who take different strategies, 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![5] 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-social network we can expand campaigns to devices for which there is no retargeting or third-party data available. 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). We could test this by comparing the winning bids on labeled "ego" nodes with the winning bids on their close network neighbors, assuming that the former have more data (e.g., the label), controlling based on whatever other data we have, and noting that network assortativity may do a reasonable job of controlling for other factors (since by the arguments outlined in Section 2, we believe closely linked devices will have close similarity along a variety of dimensions).

For the sake of evaluation, we also can compare the geo-social targeting with pure geographic targeting, and with socio-demographic targeting based on appended or inferred socio-demographic data. A key research question will be: does the geo-social inference provide all the targeting value we would see from the provided/inferred sociodemographic data (e.g., via traditional predictive modeling)? Does combining the two give a substantive advantage? Note that if the geo-social network gives equivalent or better targeting on its own, it may provide considerable advantage due to the potential for privacy-friendly design, discussed next.

We also plan to run controlled experiments, actually targeting devices using the geo-social network, and assessing whether we actually see significant lift over a baseline approach.

5. ON PRIVACY-FRIENDLINESS

This project will address explicitly "privacy by design" for targeted advertising. As an appendix we provide a comment to the FTC's recent report on privacy and online advertising. There we draw a distinction between several different methods for targeting online advertising, that have very different implications in light of the present privacy debate. Most relevant to this position paper, and simplifying here (please see the full comment), one type of targeting in particular—automated targeting based on predictive modeling—can be made surprisingly privacy-friendly, because the statistical modeling techniques have no need for the data to retain its semantic meaning. Therefore, the actual data can be "doubly de-identified." Specifically, not only can the individual's identity be removed, also the semantic meaning of the rest of the data can be removed.

For this position paper, that means that not only do we have no need to use or to store the identity of the user of a particular device (or a true identifier for the device itself), we also have no need to use or to store the actual location information. 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-social 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 any particular location with any particular device/user. An as-of-yet unexplored (to my knowledge) research question is what is the effect of such increased privacy protection on targeting efficacy.

[4]Retargeting is to target 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 my knowledge these conclusions are drawn based on assessing conversion rate rather than the influence of the advertising.)

[5]The latter is my conjecture; I have no solid evidence to support it.

6. REFERENCES

[1] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106(51):21544–21549, 2009.

[2] C. Cortes, D. Pregibon, and C. T. Volinsky. Communities of interest. pages 105–114, 2001.

[3] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, December 2010.

[4] S. Hill, D. K. Agarwal, R. Bell, and C. Volinsky. Building an effective representation for dynamic networks. *Journal of Computational & Graphical Statistics*, 15(3):584–608, 2006.

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

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

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

[8] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network. *AJS*, 115(2):405–450, September 2009.

[9] S. A. Macskassy and F. Provost. Classification in networked data: A toolkit and a univariate case study.

Journal of Machine Learning Research, 8:935–983, 2007.

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

[11] J. Neville and F. Provost. Predictive modeling with social networks. Tutorial at the ACM SIGKDD, 2008.

[12] F. Provost, B. Dalessandro, R. Hook, X. Zhang, and A. Murray. Audience selection for on-line brand advertising: privacy-friendly social network targeting. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 707–716, New York, NY, USA, 2009. ACM.