MARKETING SCIENCE
Vol. 33, No. 1, January–February 2014, pp. 82–93
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.2013.0817
©2014 INFORMS

# Predicting Individual Behavior with Social Networks

Sharad Goel, Daniel G. Goldstein
Microsoft Research, New York, New York 10011 {sharadg@microsoft.com, dgg@microsoft.com}

With the availability of social network data, it has become possible to relate the behavior of individuals to that of their acquaintances on a large scale. Although the similarity of connected individuals is well established, it is unclear whether behavioral predictions based on social data are more accurate than those arising from current marketing practices. We employ a communications network of over 100 million people to forecast highly diverse behaviors, from patronizing an off-line department store to responding to advertising to joining a recreational league. Across all domains, we find that social data are informative in identifying individuals who are most likely to undertake various actions, and moreover, such data improve on both demographic and behavioral models. There are, however, limits to the utility of social data. In particular, when rich transactional data were available, social data did little to improve prediction.

*Key words*: social networks; targeting; electronic commerce; homophily; product; computational social science
*History*: Received: August 16, 2012; accepted: September 15, 2013; Preyas Desai served as the editor-in-chief and Sunil Gupta served as associate editor for this article. Published online in *Articles in Advance* October 24, 2013.

## 1. Introduction

Predicting individual behavior is a basic objective of the social sciences, from economics (Hiebert 1974, Manski 2007) to psychology (Ajzen and Fishbein 1980), sociology (Burt 1987, Coleman et al. 1966), and business (Bass 1969, Mahajan et al. 1990). In marketing, the practice of targeting refers to the selection of pools of individuals to address. The targeting decision is informed by predicting which individuals are most likely to take action, for example, to adopt an innovative product, to support a cause, to switch providers, or to change in response to marketing communications.

Historically, when only a few television stations and magazines reached the majority of the population, marketing communications would reach both intended and unintended parties, a form of unfocused targeting with considerable waste (Iyer et al. 2005). With time, electronic record keeping made it possible to collect and retain information on individual customers, and third-party market intelligence firms brought about an era of direct, list-based targeting. Increased television, satellite, and Internet bandwidth led to a proliferation of media outlets by which a handful of television networks became hundreds and relatively few print publishers became thousands of websites. As a result, broadcast advertising narrowed and efficiency increased.

Over the years, targeting has incorporated whatever predictors were effective, affordable, and available. For over a decade, online marketers have predicted behavior at the individual level using variables such as age and sex (demographic targeting), location (geographic targeting), and website usage patterns (behavioral targeting). Today, ad servers can respond to the text of the page being viewed, be it a news story or personal email, and deliver ads on the fly that match page content (contextual targeting). The broad segments of classical marketing strategy are being replaced with individual-level predictions. With individualized predictions of who will adopt, firms can decide whom to engage (Rossi et al. 1996) and perform "customer lifetime value" calculations to determine how much to spend to acquire a particular customer (Gupta et al. 2004, Malthouse and Blattberg 2005).

After so many years of advances, the baseline models for predicting consumer behavior have become strong. Nonetheless, new sources of data will continually beg the question of the degree to which targeting can be improved. Accordingly, a compelling contemporary issue, and the focus of this article, is whether recently available social network data—generated by firms such as Facebook, Google, LinkedIn, Microsoft, Twitter, and Yahoo!—can elevate the prevalent standards of behavioral prediction and targeting. Until recently, the difficulty in observing social interactions has made it infeasible to conduct such investigations at scale. In what follows, we analyze large numbers of connections (edges) between individuals (nodes) and records of individual behavior. These data allow for all the geographic, behavioral, and demographic cues available for one individual to be augmented by the same variables for his or her social contacts. The main empirical and theoretical questions we address are how much information there is in the edges of a network and whether that information improves on baseline models for selecting individuals likely to undertake various actions.

Although social network-based targeting has received surprisingly little attention in the marketing literature, a handful of studies in related disciplines have shown that friends of adopters are themselves more likely to adopt, even after controlling for covariates (Bhatt et al. 2010, Hill et al. 2006, Provost et al. 2009). In the telecommunications domain, Hill et al. (2006) identified a target set comprising customers who were socially connected to (i.e., had communicated with) people who had adopted a new service, and they showed that these individuals were statistically more likely than average to themselves adopt the service. In similar work, Bhatt et al. (2010) predicted the adoption of a paid voice-over-IP service using a social network defined by instant message (IM) contacts. They constructed decision-tree models based on social network features (existence of adopting contacts, number of network neighbors, changes in network structure, etc.) and user features (IM communication frequency, sex, age, etc.) and used these models to rank the customer base by propensity to adopt in the next month. They found that user features and social features are roughly equally important for predicting adoption and that these feature sets are not redundant: combining them improves prediction considerably.

One question this previous research leaves open is whether, in practice, social network targeting is an effective strategy, as the number of people with an adopting contact may be exceedingly small. For example, in one domain we investigate, we found that fewer than 1 person in 750 is connected to an adopter, and in the work of Hill et al. (2006), the target set constituted just 0.3% of the customer base. In practical targeting applications, what is a statistically significant predictor may be practically insignificant in identifying, for example, the top 25% of targets for an advertising campaign. It could be the case that the 0.3% of customers connected to an adopter would have been included (or excluded) in the top quartile regardless of whether the social predictor was in the model. Even in the most extreme case, in which the social cue would move the entire candidate set from outside to inside the top quartile, the new set of targeted individuals would be largely (98.8%) identical to the old set, limiting the maximal change in adoption rate that could be observed.

In addition to the work described above, our investigation is broadly related to mainstream marketing research on identifying and quantifying social influence in networks. For much of the past century, the Bass model (Bass 1969) and its extensions dominated diffusion modeling. The Bass approach uses aggregate diffusion data as input and operates without knowledge of the underlying social network. In contrast, in this paper and in more contemporary marketing research, the network is known and adoption can be studied at the individual level. People who are in social contact can influence one another (Centola 2010, Christakis and Fowler 2007), and the network-based marketing literature has largely focused on identifying these causal effects in product adoption and on articulating tests to distinguish between causal and noncausal effects (for useful reviews, see Peres et al. 2010, Van den Bulte 2010). For example, Manchanda et al. (2008) modeled the adoption of pharmaceuticals at the individual level as a joint consequence of contagion and marketing effects. Iyengar et al. (2011) reported evidence for social influence and its moderators in the adoption of a risky drug. Trusov et al. (2010) presented a technique to identify which social network members exhibit influence on the activity levels of others, and Godes and Mayzlin (2009) used a field test to show that firms can create word of mouth exogenously. In recent years, focus has shifted from establishing whether influence exists to understanding its mechanics and prevalence. For instance, Godes (2011) took contagion as established and stresses moderators. Similarly, Aral and colleagues (Aral 2011, Aral et al. 2009) sought primarily to quantify, not establish, the contribution of social influence relative to other factors.

The establishment of social influence, however, does not provide an answer to the question of whether social network data will improve targeting and prediction in practice. Even when individuals influence one another, peer-to-peer transmissions can be so rare as to have no practical value. And even when individuals are known *not* to influence one another, social ties may be still be predictive, as observed in the sociological research on homophily (Lazarsfeld and Merton 1954, McPherson et al. 2001).

In our analysis we remain agnostic as to whether there is social influence in the domains we investigate. However, in one of these domains, it is highly unlikely that there could be social influence. Specifically, we measure whether the social contacts of a person who clicks on an online advertisement are themselves likely to click on the same advertisement. Since these advertisements are untargeted and run for one day only, and since people would not typically inform their social contacts of the ads on which they have clicked (or, for that matter, would even know that other people are served the same ads), social influence is highly improbable. Nonetheless, as we demonstrate, the social contacts of those who click ads do indeed show an elevated probability of clicking on the same ads. Because establishing the presence of social influence is irrelevant for our goal, we purposefully focus instead on the managerially relevant question of assessing the worth of social network data for targeting and prediction. This is not to say that causal influence is rarely of interest in network marketing. On the contrary, when nodes are known to influence each other, marketers may wish to affect diffusion patterns, for instance, by seeding influential nodes with marketing actions in order to encourage them to adopt early.

We build on past marketing and network research in several ways. First, by progressively adding stronger predictors to baseline models—starting with basic demographics and moving on to individual-level transaction data—we provide managerial insight into when it may be worthwhile to invest in social network data. Second, whereas past investigations are largely single-domain studies, we examine a dozen independent outcomes grouped into three domains. In particular, we study settings in which the relevant action is relatively costless (clicking on an ad), to moderately involved (joining a recreational league), to one with monetary stakes (purchasing at a store). Although it may be the case that any given domain yields an idiosyncratic result, across multiple examples, each analysis is placed in perspective, providing expectations for generalizing our findings. Third, we show how social network data can proxy for other predictors and discuss cases in which social data are available but ordinary predictors are not. Finally, our results highlight an often overlooked point in social network research: where there is homophily, one can, in principle, predict an individual's behavior based on the attributes and actions of his or her associates, regardless of whether that similarity is due to social contagion.

Our analysis is based on individuals within the Yahoo! communications network, where we establish an edge between all pairs of people who mutually exchanged email or instant messages during a fixed two-month period. Restricting analysis to those individuals with at least one correspondent resulted in a symmetric network of 132 million people and 719 million edges, with a mean of 11 contacts per individual. (See Figure A1 in the online appendix, available as supplemental material at http://dx.doi.org/10.1287/mksc.2013.0817, for the full-degree distribution for this network.) Mutual email exchanges were used as a criterion for establishing an edge between nodes to exclude messages between mailing lists (spam or legitimate) and their recipients. Although the Yahoo! email and IM network is one of the largest in the world, it is, of course, the case that its users take part in other electronic social networks as well. The predictive power of social network ties we observe should thus be taken as a lower bound on what could be observed in a larger network or a combination of networks. Nonetheless, the network we study represents one of the largest that a marketer could, in this day, hope to engage for the purposes of social network targeting.

To assess the value of social predictors, we examine individual-level behaviors in three diverse domains comprising 12 distinct outcomes: responding to national advertisements for 10 products and services, participating in an online recreational league with millions of players, and purchasing (off-line and online) from a national department store chain. We treat each domain in turn, and within a domain we test baseline models that range from extremely simple (using social network data or demographic data alone) to strong (adding social network data to models that include demographics as well as individual-level transactional data).

## 2. Response to Advertising

We examined individual response to online advertisements, measured by clicks on 10 display ads prominently shown on the Yahoo! front page. Advertisements ran for one day each, in random rotation with another ad, and were not targeted (i.e., were shown with equal probability to all users). In total, each advertisement was viewed by approximately 14–15 million logged-in users. To study the effectiveness of social signals, we restrict this set of individuals to those present in our communications network, leaving 7–8 million users per campaign.

Are social data useful for predicting clicks on ads? We first answer this question for the scenario in which only social data are available. Table 1 shows for 10 different advertising campaigns the clicking rates for people without and with social contacts who clicked on the same ad. (Ads are ordered by the percentage increase in probability of clicking between those with and without social contacts who clicked.) The largest increase was observed for Movie 1, where people whose social contacts clicked on this ad were more than 10 times as likely (1,140%) to click than those without contacts who clicked. The Insurance 2 and the Movie 3 campaigns had the smallest social effect, a 10% increase in probability to click. Thus, consistent with past studies (Aral et al. 2009, Bhatt et al. 2010, Hill et al. 2006, Provost et al. 2009), contacts of adopters are themselves more likely than average to adopt, and in at least some of the campaigns we examined, this social signal is quite strong. As noted

**Table 1** **Probability of Clicking on 10 Display Advertisements Related to Having At Least One Social Contact Who Also Clicked on the Ad**

| Domain | Click rates for individuals without contacts who clicked (%) | Click rates for individuals with contacts who clicked (%) | Percentage of individuals with contacts who clicked |
|---|---|---|---|
| Movie 1 | 0.038 | 0.47 | 0.036 |
| Government | 0.209 | 0.46 | 0.225 |
| Movie 2 | 0.225 | 0.44 | 0.239 |
| TV | 0.260 | 0.50 | 0.303 |
| Transportation | 0.155 | 0.25 | 0.160 |
| Insurance 1 | 0.124 | 0.19 | 0.138 |
| Apparel | 1.723 | 2.43 | 1.881 |
| Household | 0.205 | 0.27 | 0.222 |
| Insurance 2 | 0.118 | 0.13 | 0.129 |
| Movie 3 | 1.185 | 1.30 | 1.335 |

*Notes.* Because clicking is rare, most people have one or zero contacts who clicked the ad. Ads are sorted by the relative increase in probability of clicking, from 1,140% for Movie 1 to 10% for Movie 3.

above, the effects we observe are unlikely due to social influence, as people would not typically inform their social contacts of the ads on which they have clicked; rather, contact with an individual who clicked is likely a proxy for latent similarity and thus indicates an individual's greater inherent propensity to click on the ad.

That contacts of those who clicked have relatively high click rates does not in itself imply that social data are valuable for prediction, in part because few people may be connected to individuals who clicked. For example, in the Movie 1 campaign—which exhibited the largest social signal—only 1 in 2,700 users had a contact who clicked (indicated in the final column of Table 1), a result of the low overall click rate (0.04%) and the relative sparsity of the social network. The question thus arises whether the observed social effects are of any practical use.

In many contexts, the central objective is to identify pools of individuals most likely to take action. Such is the case, for example, when directing scarce resources either to encourage action (e.g., promoting energy-saving technologies) or to discourage action (e.g., anti-smoking campaigns). We thus assess the predictive value of social data via a top-$k$ analysis: given a particular prediction model, one first orders the members of the population according to their estimated probability of clicking, and then one computes the observed click rates for hypothetical segments of increasingly many candidates, starting with only the highest-scoring candidates and concluding with a segment consisting of the entire population.

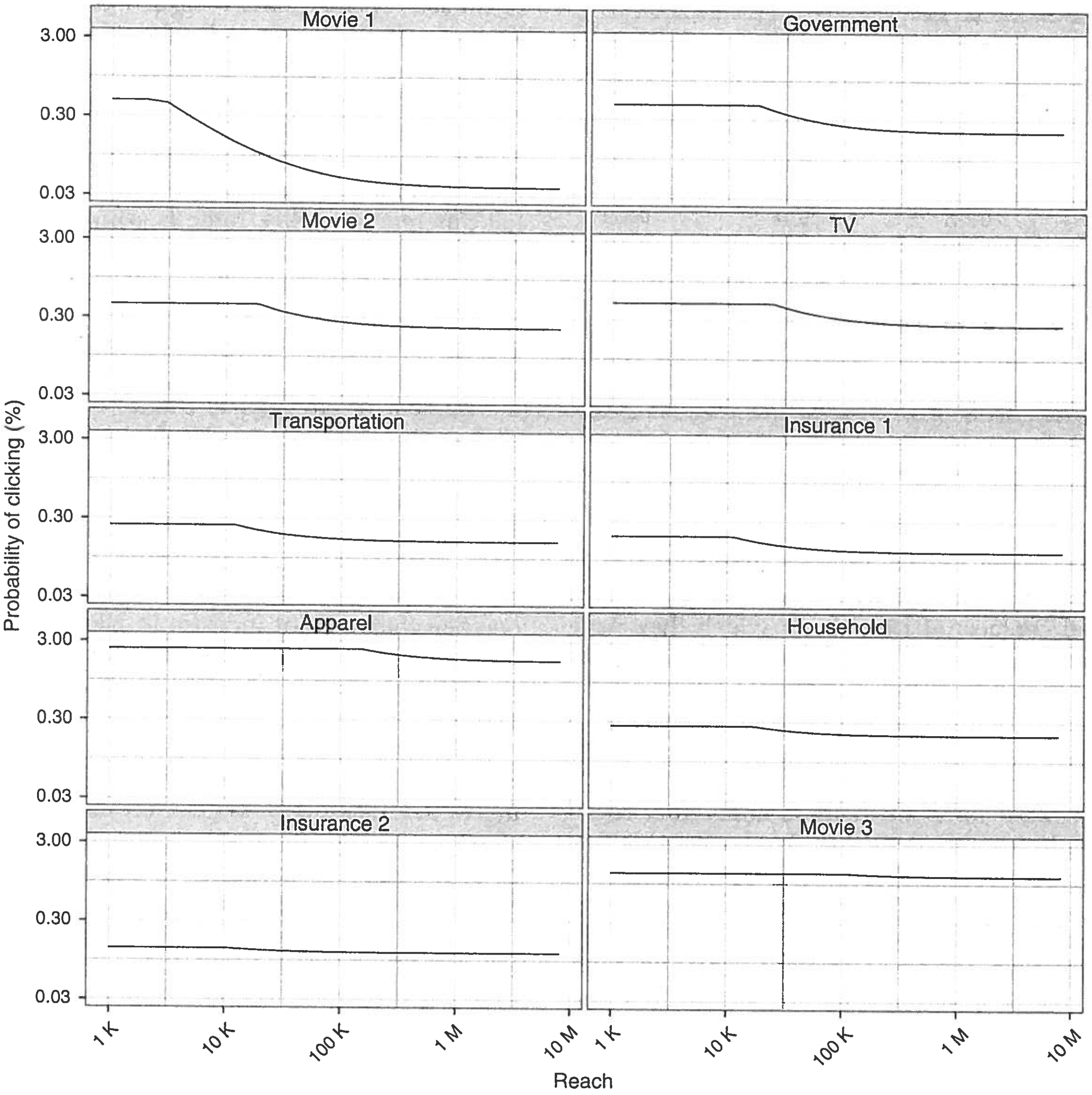
Figure 1 shows top-$k$ curves for each of the 10 advertisements based on a simple model that uses only the social signal. Specifically, the model ranks people with a contact who clicked the ad above those without one (and randomly orders individuals within each of those categories). For this reason, the curves begin as horizontal lines and then gracefully descend as increasingly many individuals without contacts who clicked on the ad are added to the pool. In particular, the rightmost points in the plots correspond to including every individual in the hypothetical candidate pool. (We note that both the $x$ and $y$ axes are displayed on a log scale to accommodate the substantial differences in click rates as a function of the size of the candidate pool.)

As Figure 1 indicates, the social signal allows one to construct pools of between 10,000 and 100,000 candidates who are much more likely than average to click, where the size of these pools effectively reflects the number of individuals connected to people who clicked. Corresponding to Table 1, the increase in click rates of these high-ranked candidates varies substantially from one campaign to the next, ranging from 10% to over 1,000%. However, these pools are small relative to the entire set of approximately eight million individuals who saw the ads. In particular, a one-percentage-point increase in click rate on 10,000 individuals results in 100 additional people on average clicking the ad, which may be of little value in practice.

Although we have thus far seen that social data on their own may not be particularly useful for identifying individuals likely to click on ads, it is often the case that social data are not the only resource available for targeting. As a case in point, Yahoo! collects age and sex for each of its registered users and routinely uses this information for ad targeting. (Whereas Yahoo! collects age as an integer, we treat it in this paper as belonging to one of five categories for ease of fitting and communicating models, with no appreciable difference noted when age is modeled as continuous.) Overall, older adults are more likely to click on ads. Sex differences tend to be minor and vary by campaign, generally in line with expectations of the intended audience of the product advertised. One unexpected finding was that the apparel ad was more often clicked on by men than women, despite it being for women's apparel (namely, lingerie). Figure A2 in the online appendix shows how age and sex relate to the probabilities of clicking on the ten ads we study.

Given the effectiveness of demographic data to predict clicking on advertisements, do social data substantively improve on demographics in constructing pools of likely adopters? It could be, for example, that social connections are simply a proxy for demographic information, in which case social data would offer little marginal value. To assess the marginal value of social data relative to demographic information, we generate two candidate rankings, one based only on demographic attributes and the other based

**Figure 1** Probability of Clicking for Varying Numbers of High-Scoring Individuals Under a Model That Only Uses Connection to a Person Who Clicked as a Predictor



*Note.* The set of individuals included in the cumulative average varies from the highest-scoring individuals according to the model (at the far left) to the entire population (at the far right).

on both demographic and social information. Then, analogous to the above analysis, we examine the average click rate among pools of top-*k* candidates under each ranking.

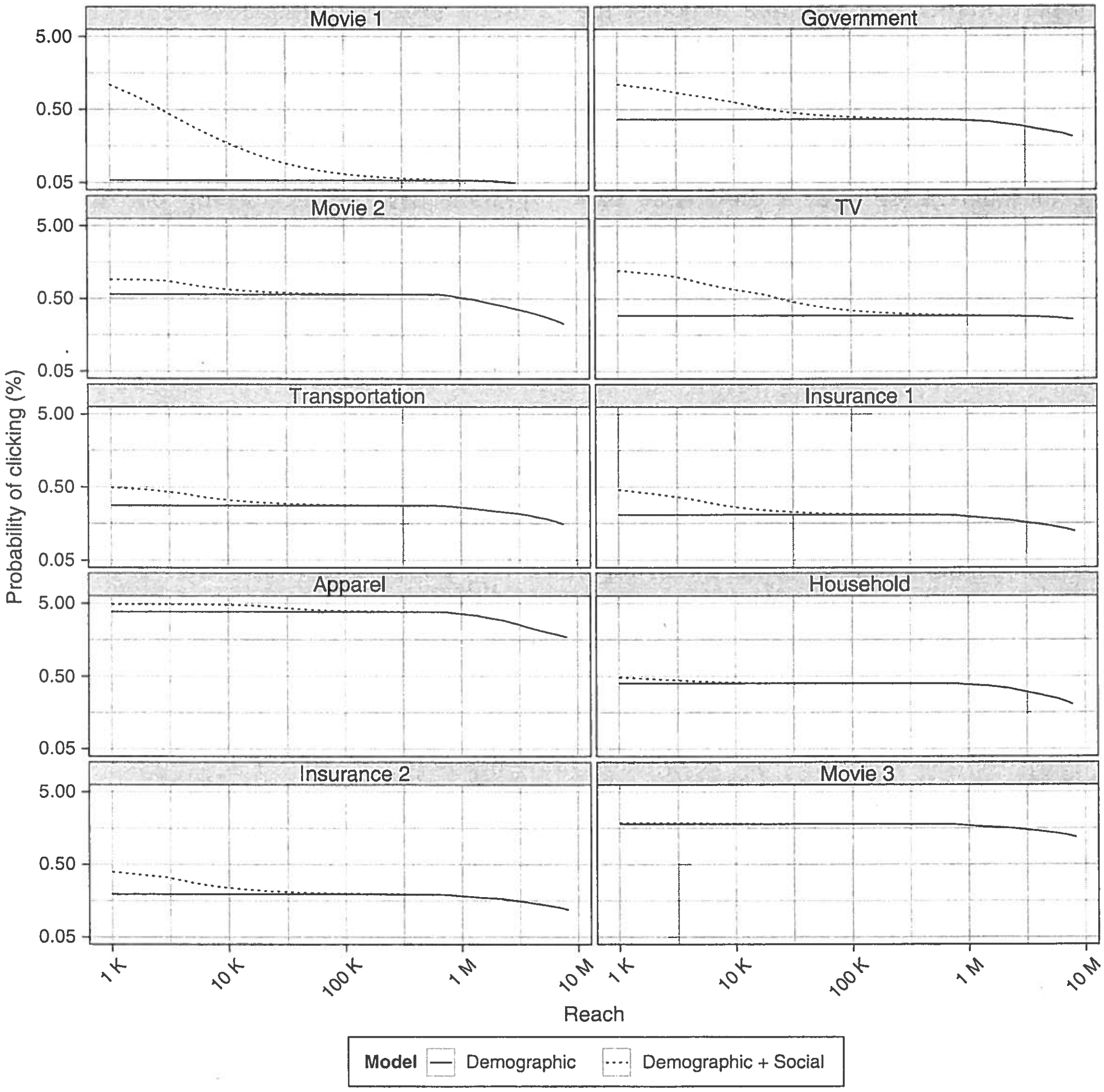
To construct the candidate rankings, we estimate the likelihood that individuals click on an advertisement, fitting separate demographic and demographic-plus-social models for each of the 10 advertisements. The basic structure of the models was the same in all cases. Specifically, likelihood to click was estimated using logistic regression. For the demographic model, the independent variables were age (expressed in five categories: 18–24, 25–34, 35–44, 45–54, and 55+), sex, and all two-way interactions; for the demographic-plus-social model, these independent variables were augmented with a binary variable indicating whether the individual had any contacts who clicked with all two- and three-way interaction terms. To avoid overfitting the models to the data, we generated cross-validated estimates. That is, for each advertisement we first randomly divided the full candidate set of approximately eight million people into five subsets; predictions for individuals in each subset were then obtained from models trained on the combined data from the other four subsets. We thus fit 10 different models for each of the 10 advertisements (resulting in 100 total models): a demographic model and a demographic-plus-social model for each of the five subsets of the population. As an illustrative example, Table A1 in the online appendix lists coefficients for the demographic and demographic-plus-social models for the "apparel" advertisement, where for simplicity details are given for models fit on the entire data set (i.e., without cross-validation).

Figure 2 shows the performance of candidate pools selected via the demographic and demographic-plus-social models for each of the 10 advertisements. In particular, the plots illustrate three points. First, of the total candidate set of approximately eight million people, demographic information is generally useful for constructing pools of up to several million individuals that are substantially more likely than average to click. Second, even on top of this baseline, augmenting demographic information with social data often helps to identify pools of between 10,000 and 100,000 individuals that are even more likely to click. For example, in the government advertisement, whereas the overall click rate is 0.21%, the top 10,000 candidates selected by the demographic model have a click rate of 0.36%, and the top 10,000 from the demographic-plus-social model click at 0.62%—a rate increase of over 70% relative to the demographic-only ranking. Third, as before, the reach of social targeting is relatively small for these 10 ads, limiting the utility of social data in such predictions.

## 3. Recreational League Registrations

We next consider the extent to which social data can help to identify participants in the Yahoo! Sports Fantasy Football competition, one of the largest such

**Figure 2** Click Rates for Varying Numbers of High-Scoring Individuals Under a Demographic Model and a Model That Includes Both Demographic and Social Attributes for the Advertising Domain



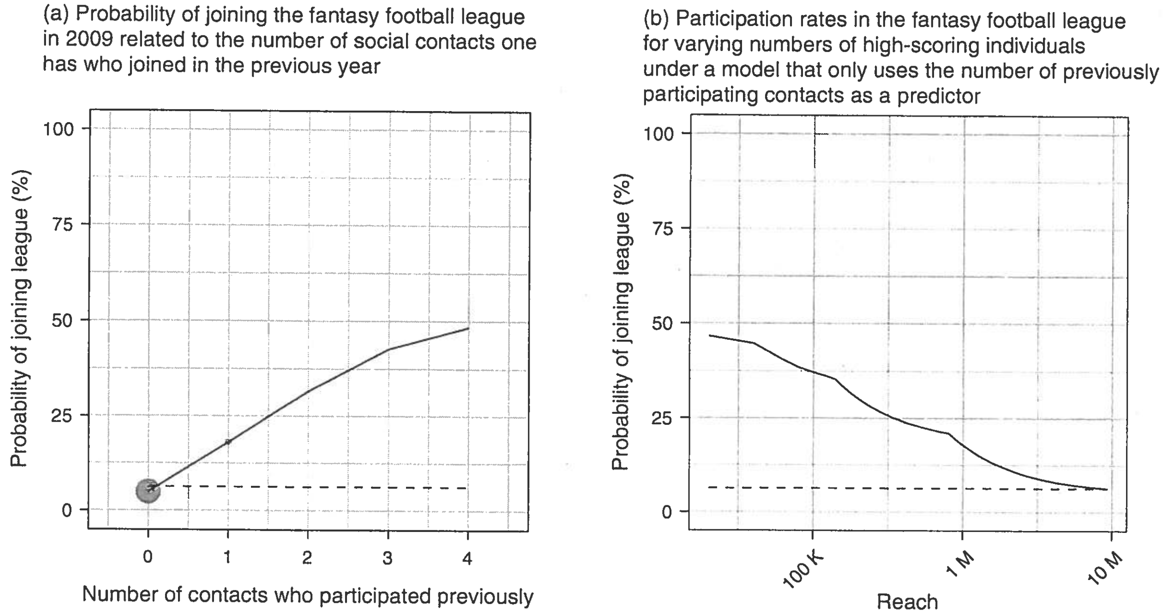
*Note.* The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right).

competitions, with approximately four million annual registrants. Specifically, our goal is to identify those individuals likely to participate in the 2009 fantasy football competition based on combinations of social, demographic, and behavioral predictors available the previous year. Our initial pool of candidate registrants comprises those users in our communications network who had also made recent visits to the Yahoo! Sports website, resulting in a population of 9.3 million, 6% of whom participated in the 2009 competition.

We begin by examining the relationship between an individual's propensity to participate and his or her number of contacts who participated the previous year. That is, we compute participation rates for the 2009 competition as a function of the number of contacts one has who participated in 2008. As shown in Figure 3, panel (a), having one or more contacts who previously participated in the competition is a strong indicator of participation. For example, although the overall participation rate is 6%, approximately 15% of those with one previously participating contact themselves participated, and nearly 50% participated among those with four contacts who participated in the previous year's event. However, as discussed for advertising, although those with a previously participating contact are themselves considerably more likely to register, is it is not immediately clear whether this signal is useful for selecting candidates because the vast majority of individuals have no previously participating contacts, as indicated by the sizes of the points in Figure 3, panel (a). To investigate this question, we again rank candidates by their number of previously participating contacts (breaking ties at random) and compute participation rates for pools of the top-*k* individuals. Figure 3, panel (b) shows that the social signal is indeed quite effective in this case, allowing us to construct pools of hundreds of thousands of individuals who participate at four to five times the base rate, and even extending into pools of millions of individuals that are substantially more likely than average to play. In contrast to the advertising setting, social data let us build much larger pools of good candidates in the fantasy football domain. This observation results from a variety of factors, but perhaps most important is simply that many more people have contacts who have played fantasy football than contacts who have clicked on a particular advertisement.

Although we have seen that social data alone are useful to select candidates, demographics are also a strong indicator of participation, with men in their 30s and 40s particularly likely to play fantasy football (see Figure A3 in the online appendix). We thus next assess the extent to which social information improves candidate selection relative to a demographic baseline. Mimicking our above analysis in the advertising domain, we construct two ranked lists of candidates based on estimated participation rates under a demographic and a demographic-plus-social model. Specifically, we use logistic regression models to estimate likelihood to participate in the 2009 fan-

**Figure 3** **Likelihood to Join a Fantasy Football League as a Function of One's Social Network**


(a) Probability of joining the fantasy football league in 2009 related to the number of social contacts one has who joined in the previous year

(b) Participation rates in the fantasy football league for varying numbers of high-scoring individuals under a model that only uses the number of previously participating contacts as a predictor

*Notes.* In panel (a), the area of each point indicates the relative number of individuals in the respective category, and the dashed line indicates the overall average participation rate. In panel (b), the set of individuals included in the cumulative average varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right). The dashed line indicates the overall average participation rate.

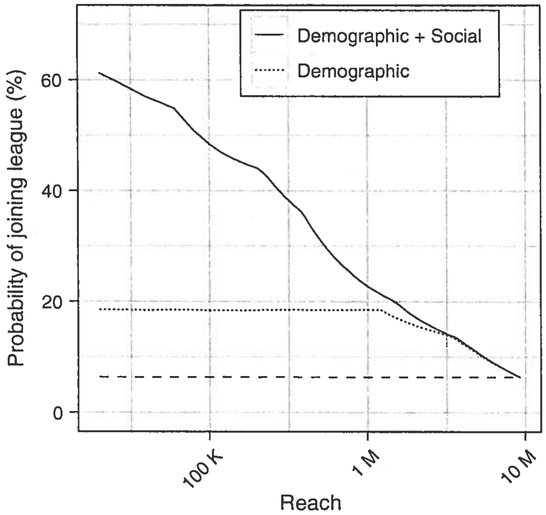
tasy football competition. For the demographic model, the independent variables were age (expressed in five categories), sex, and all two-way interactions; for the demographic-plus-social model, these independent variables were augmented with a predictor giving the number of an individual's contacts who played the previous year, with all two- and three-way interaction terms. As before, we generated cross-validated estimates to avoid over-fitting: the approximately nine million candidates were randomly partitioned into five subsets, and predictions for individuals in each subset were obtained from models trained on the combined data from the other four subsets. In total, we thus fit 10 models, a demographic model, and a demographic-plus-social model for each of the five subsets of the population. For illustrative purposes, Table A2 in the online appendix displays coefficients for demographic and demographic-plus-social models fit on the entire data set of 9.3 million (i.e., without cross-validation).

Figure 4 shows participation rates for pools of top-ranked candidates under the demographic and demographic-plus-social models. Both selection methods identify millions of candidates with more than twice the overall participation rate. However, augmenting demographic data with the social predictor enables one to identify hundreds of thousands of candidates with exceptionally high participation rates. In particular, the top 100,000 candidates selected under the demographic-plus-social model participate at a rate of 49%, more than twice the participation rate of those selected under the demographic model (19%). As an additional point of comparison, under the social-only model, 37% of the top 100,000 candidates participated, as shown in Figure 3, panel (b). Social information does indeed complement demographic data for selecting those likely to join the recreational league.

We have thus far evaluated the utility of social data in augmenting demographic predictors. Richer baselines, however, are available in select settings. Often, the best single predictor of future behavior is past behavior. Indeed, 79% of users who participated in the 2008 fantasy football competition played again the subsequent year—a fact that potentially can be exploited to identify future participants. We generated two additional ranked lists of candidates corresponding to models based on (1) demographics and past participation and (2) demographics, past participation, and the number of previously participating contacts. These models are direct analogs of the demographic and demographic-plus-social models discussed above, adding only a binary variable indicating whether an individual participated in the previous year's competition, together with the two-way interaction terms (see Table A2 in the online appendix).

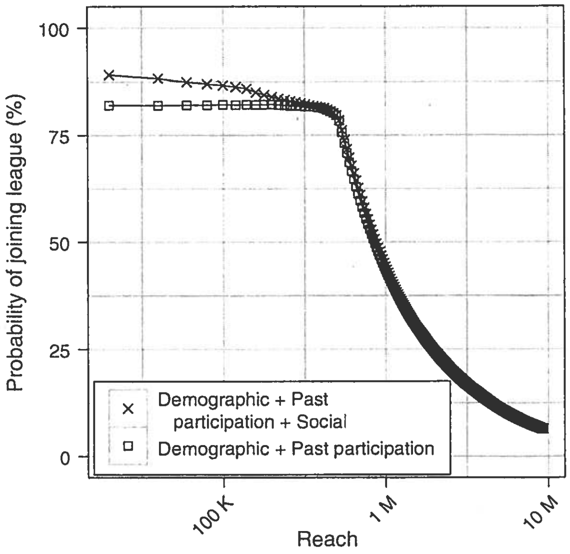
We find that social data improve even on this strong baseline that incorporates both demographics and past participation. For example, as shown in Figure 5, the top 100,000 candidates selected by considering demographics and past participation join at a rate of

**Figure 4** **Probability of Joining the Fantasy Football League for Varying Numbers of High-Scoring Individuals Under a Demographic Model and a Model That Includes Both Demographic and Social Attributes**



*Notes.* The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right). The dashed line indicates the overall average participation rate.

**Figure 5** **Probabilities of Joining the League for Varying Numbers of High-Scoring Individuals Under a Model That Incorporates Demographics and Past Participation vs. the Same Model That Adds the Number of Previously Participating Contacts**



*Note.* The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right).

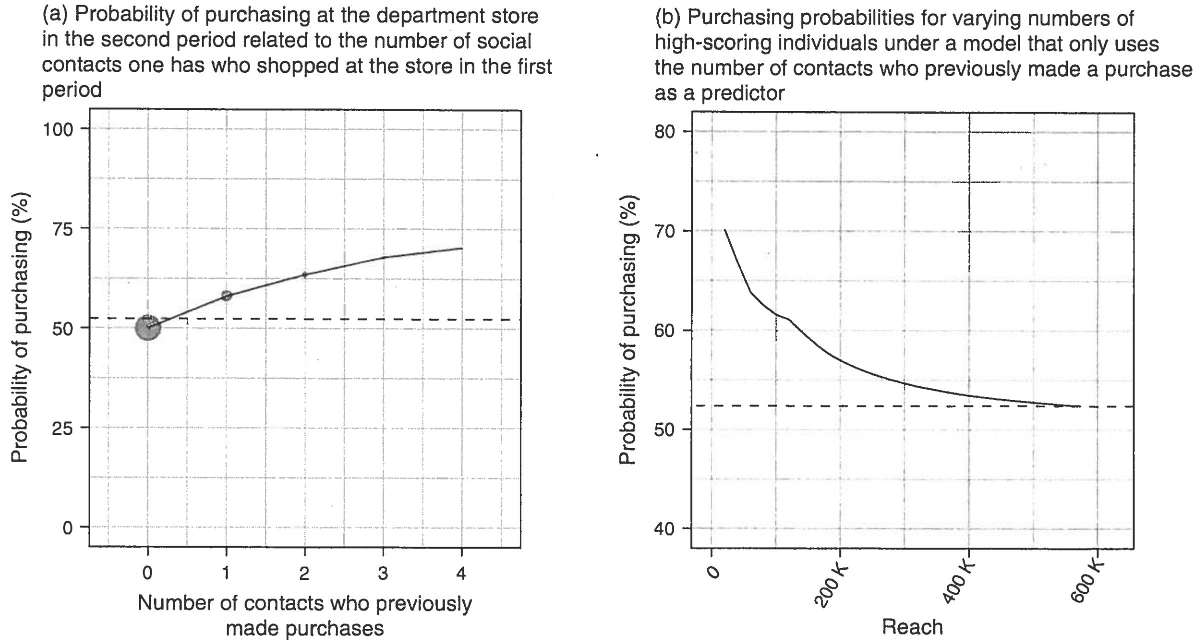
82%, whereas adding social data results in a candidate pool with an 87% participation rate. Although we saw substantially larger lifts when adding social data to a demographic-only model, it is surprising to observe any improvement at all in this setting. Social predictors, it appears, provide a signal that is at least partially orthogonal to both one's demographics and past behavior.

## 4. Retail Purchases

The third and final domain we investigate is retail purchasing at a national department store chain. As in the other domains we studied, our goal here is to construct pools of candidates that are likely to take a specific action, which in this case corresponds to making an online or off-line purchase at the retail outlet. To generate the universe of candidates, we intersect the department store's 1.3 million member customer database with our communications network, resulting in approximately 588,000 people, for whom we had a record of both their own purchases (if any) as well as any purchases of their social contacts for a 12-month time span. To preserve anonymity, this matching was done by a specialist third party. The data for each retail customer were divided into two consecutive six-month periods, with combinations of demographic, behavioral, and social information from the first period used to construct pools of candidates likely to make purchases in the second period. We note that compared with the previous two domains—advertisements and fantasy football—this example has two key distinguishing features: first, the action in question is arguably the most costly we have yet considered; and second, here we have detailed purchase histories (i.e., transaction amounts in the first period) that provide exceptionally strong baselines against which we can evaluate social data.

We begin our analysis by examining the relationship between purchase rates and the number of one's contacts who made a purchase in the first period. Consistent with the two domains we studied above, Figure 6, panel (a) shows that having contacts who previously made a purchase is indicative of substantially higher than average (second-period) purchase rates. For example, whereas the overall purchase rate is 52%, the rate is 70% among candidates having four contacts who previously made purchases. Such individuals, however, form a small subset of the population—as indicated by the size of the points in Figure 6, panel (a)—and so it is not immediately clear how useful this social signal is for generating candidate pools. Figure 6, panel (b) addresses this question, showing purchase rates for pools of the top $k$ candidates as ordered by their number of previously purchasing contacts (with ties broken at random). From that top-$k$ plot, we see that the social signal does, in fact, allow us to construct large subsets of candidates—both in absolute number and in relative terms—that are substantially more likely than

**Figure 6** Likelihood to Purchase as a Function of One's Social Network



(a) Probability of purchasing at the department store in the second period related to the number of social contacts one has who shopped at the store in the first period

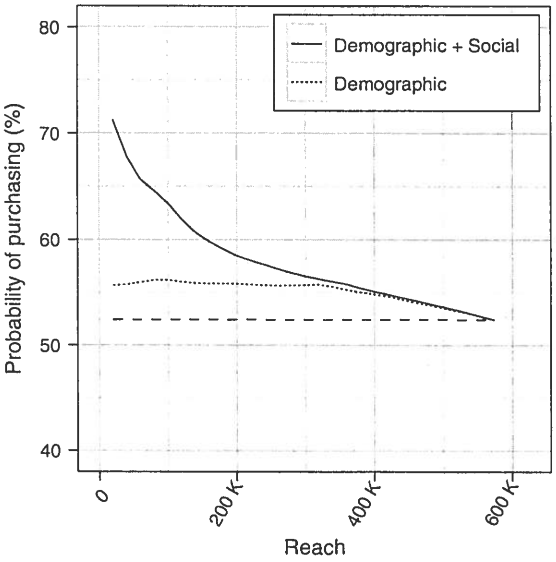
(b) Purchasing probabilities for varying numbers of high-scoring individuals under a model that only uses the number of contacts who previously made a purchase as a predictor

*Notes.* In panel (a), the area of each point indicates the relative number of individuals in the respective category, and the dashed line indicates the overall average purchasing probability. In panel (b), the set of individuals included in the cumulative average varies from the highest-scoring individuals according to the model (at the far left) to the entire population (at the far right). The dashed line indicates the overall average purchasing probability.

average to make purchases. In particular, the top 100,000 socially selected candidates have a purchase rate of 62%. Thus, at least when alternative predictors are not available, social data are effective for identifying subsets of individuals likely to make purchases, in line with our results from the fantasy football domain.

We next assess the utility of social data relative to a candidate selection strategy based on demographics. As shown in Figure A4 in the online appendix, women in their 40s and older are particularly likely to purchase at this particular retailer. Following our analysis of advertising and fantasy football, we construct a demographic baseline by first estimating each candidate's likelihood to purchase using a logistic regression model with age (represented in five categories), sex, and their two-way interactions as predictors; candidates are then ranked by these model-estimated probabilities. To measure the marginal value of social data on top of demographic information, we analogously rank candidates by a demographic-plus-social model, in which we add as predictors one's number of contacts who previously made a purchase and the corresponding two-way interactions. As before, estimates in both cases are generated via fivefold cross-validation. For simplicity, Table A3 in the online appendix shows coefficient estimates for the demographic and demographic-plus-social models fit on the entire data set (i.e., without cross-validation).

**Figure 7** **Purchase Rates for Varying Numbers of High-Scoring Individuals Under a Demographic Model and a Model That Includes Both Demographic and Social Attributes for the Shopping Domain**



*Notes.* The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right). The dashed line indicates the overall average purchasing probability.

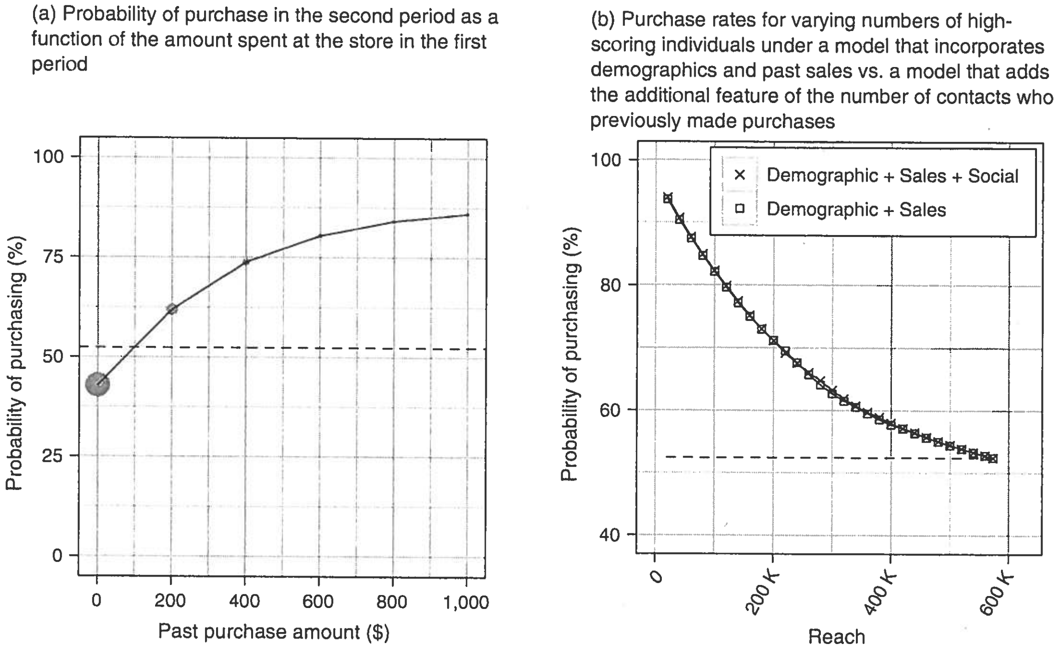
Figure 7 compares the effectiveness of augmenting demographic with social data for candidate selection. Demographic data alone yield a modest lift, but adding social predictors substantially improves the selection of top candidates. In particular, among the top 100,000 candidates selected using only demographic predictors, 56% make purchases (compared with an overall purchase rate of 52%); in comparison, 63% of the top 100,000 make purchases when candidates are ranked by both demographic and social predictors. Thus, social data do complement demographic information in identifying those candidates most likely to make purchases.

We conclude our analysis of the retail domain by evaluating the marginal value of social information relative to a baseline ranking derived from both demographic predictors and detailed transactional information about an individual's past behavior. Namely, in ranking candidates, we use the total dollar amount one spent in the first period to predict second-period purchases. We note that in contrast to the fantasy football domain, where we had a binary indicator of past behavior (i.e., whether an individual competed the previous year), here we have a continuous measure of past behavior (i.e., transaction amount), a stronger signal. In particular, Figure 8, panel (a) indicates a steady increase in purchase rate as the amount of one's past transactions increases. For example, among those who made approximately $200 in purchases during the first period, approximately 62% made purchases in the second period, compared with a purchase rate of 86% for customers who spent $1,000.

To incorporate this precise behavioral measure, we added past sales and its two-way interactions to the demographic and demographic-plus-social logistic regression models we use to rank candidates. Table A3 in the online appendix lists fitted coefficients from these additional models that include the behavioral measure. (As before, our analysis is based on cross-validated estimates, though for simplicity, the coefficients in Table A3 are for models fit on the entire data set.) The results of ranking candidates by these models are shown in Figure 8, panel (b). Against a baseline candidate selection method that includes both demographics and detailed past behavior, we find that social data offer nearly no marginal benefit. In this case, the predictive signal one gets from an individual's past transactions trumps the potential benefits of the social signal.

We note that recency-frequency-monetary (RFM) (Blattberg et al. 2008) models are standard tools for predicting repeat purchase, but we do not test them here because we did not have access to the required recency or frequency data. We do have monetary data, which we incorporate in our predictions. However,

**Figure 8** **The Value of Transaction Data for Estimating Purchase Rates**


(a) Probability of purchase in the second period as a function of the amount spent at the store in the first period

(b) Purchase rates for varying numbers of high-scoring individuals under a model that incorporates demographics and past sales vs. a model that adds the additional feature of the number of contacts who previously made purchases

*Notes.* In panel (a), the area of each point is proportional to the number of individuals in the corresponding category. The dashed line indicates the overall average purchasing probability. In panel (b), the set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to the entire population (at the far right). The dashed line indicates overall purchasing probability.

because the effect of social data is largely absent in a model with only monetary data, it should be even less likely to improve on a full RFM model.

## 5. Discussion and Conclusion

Returning to our motivating question, we find that there are a variety of circumstances in which social data are useful for identifying select groups of individuals with relatively high propensities to take various actions, from clicking on advertisements, to registering for a recreational league, to making department store purchases. Across all three domains we study, mere connection to an individual who has previously taken such an action is indicative of a higher than average propensity to act oneself. We also find that social data are not only useful in isolation but also often complement both demographic and behavioral predictors. In particular, social predictors substantially augmented demographic-based candidate selection in the shopping and fantasy football domains. Moreover, given that our results are based on only two communications networks—though quite large ones—it is likely one would find even more predictive value from incorporating social data from additional sources.

Social data have proven to be widely effective in the examples we study, but there are limits to their benefits. Specifically, when relatively few people undertake a particular action, and when individuals have few contacts, the contacts of these initial actors form a relatively small set; consequently, the reach of social targeting strategies may be small. The advertising domain is a case in point: with only one in several thousand candidates connected to an individual who has clicked on a given ad, social predictors reveal only a relatively small fraction of the candidate pool to be likely themselves to click. Furthermore, when detailed transaction data were available—as in the retail domain—we find social data provide almost no marginal benefit. More generally, it seems likely that when enough information is available at the individual level, the marginal value of network data is muted.

Accordingly, social data seem particularly valuable in situations where a potential target's social network is known but information about his or her past behavior, and possibly demographic characteristics, is limited. There are at least two common situations in which this happens. First, when new members join existing social networks, they may quickly link to their associates who are already members—for example, by importing contacts from their email accounts. A second scenario occurs when new members link their site accounts to their social network accounts—some sites require such linking. In both these cases, users who are too new to a site to have built up a behavioral or transactional profile can nonetheless be targeted on the basis of their social contacts' behavior.

In addition to the situations described above, our results also suggest the value of a two-stage social marketing strategy. In the first stage, standard demographic and behavioral targeting measures could be used to reach candidates. This first campaign would yield a set of adopters whose contacts could then be advertised to in the second stage. Since these second-stage candidates are by construction connected to adopters, they themselves should be much more likely than average to adopt.

Finally, we note that the marketing literature we have reviewed has focused on the topic of proving, disentangling, or modeling causality in social networks. It may be tempting to conclude from our results that shopping habits or leisure activities are "contagious." Although social influence likely plays a role in effects we find, establishing such is neither our objective nor justified from our analysis. Nevertheless, whereas the value of social data may concern both influence and homophily, our approach demonstrates that disentangling the two is not necessary for identifying and targeting likely adopters.

## Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/mksc.2013.0817.

## Acknowledgments

The authors thank Jake Hofman, Randall Lewis, Gary Lilien, David Pennock, Foster Provost, David Reiley, Olivier Toubia, Duncan Watts, and Russ Winer for their feedback.

## References

Ajzen I, Fishbein M (1980) *Understanding Attitudes and Predicting Social Behavior* (Prentice-Hall, Englewood Cliffs, NJ).

Aral S (2011) Identifying social influence: A comment on opinion leadership and social contagion. *Marketing Sci.* 30(23):217–223.

Aral S, Muchnika L, Sundararajana A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci. USA* 106(51): 21544–21549.

Bass FM (1969) A new product growth model for consumer durables. *Management Sci.* 15:215–227.

Bhatt R, Chaoji V, Parekh R (2010) Predicting product adoption in large-scale social networks. *Proc. 19th Internat. Conf. Inform. Knowledge Management (ICIKM 2010)* (ACM, New York), 1039–1048.

Blattberg RC, Kim B-D, Neslin SA (2008) *Database Marketing: Analyzing and Managing Customers* (Springer, New York).

Burt RS (1987) Social contagion and innovation: Cohesion versus structural equivalence. *Amer. J. Sociol.* 92(6):1287–1335.

Centola D (2010) The spread of behavior in an online social network experiment. *Science* 329(5996):1194–1197.

Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *New England J. Medicine* 357(4): 370–379.

Coleman JS, Katz E, Menzel H (1966) *Medical Innovation: A Diffusion Study* (Bobbs-Merrill, Indianapolis).

Godes D (2011) Invited comment on "Opinion leadership and social contagion in new product diffusion." *Marketing Sci.* 30(2): 224–229.

Godes D, Mayzlin D (2009) Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Sci.* 28(4):721–739.

Gupta S, Lehmann DR, Stuart JA (2004) Valuing customers. *J. Marketing Res.* 41(1):7–18.

Hiebert LD (1974) Risk, learning, and the adoption of fertilizer responsive seed varieties. *Amer. J. Agricultural Econom.* 56(4): 764–768.

Hill S, Provost F, Volinsky C (2006) Network-based marketing: Identifying likely adopters via consumer networks. *Statist. Sci.* 21(2):256–276.

Iyengar R, Van den Bulte C, Valente TW (2011) Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* 30(2):195–212.

Iyer G, Soberman DA, Villas-Boas JM (2005) The targeting of advertising. *Marketing Sci.* 24(3):461–476.

Lazarsfeld PF, Merton RK (1954) Friendship as a social process: A substantive and methodological analysis. Berger M, ed. *Freedom and Control in Modern Society* (Van Nostrand, New York), 18–66.

Mahajan V, Muller E, Bass FM (1990) New product diffusion models in marketing: A review and directions for research. *J. Marketing* 54(1):1–26.

Malthouse E, Blattberg R (2005) Can we predict customer lifetime value? *J. Interactive Marketing* 19(1):2–16.

Manchanda P, Xie Y, Youn N (2008) The role of targeted communication and contagion in product adoption. *Marketing Sci.* 27(6):961–976.

Manski CF (2007) *Identification for Prediction and Decision* (Harvard University Press, Cambridge, MA).

McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Rev. Sociol.* 27:415–444.

Peres R, Muller E, Mahajan V (2010) Innovation diffusion and new product growth models: A critical review and research directions. *Internat. J. Res. Marketing* 27(2):91–106.

Provost F, Dalessandro B, Hook R, Zhang X, Murray A (2009) Audience selection for on-line brand advertising: Privacy-friendly social network targeting. *Proc. 15th ACM SIGKDD Internat. Conf. Knowledge Discovery Data Mining (KDD'09)* (ACM, New York), 707–716.

Rossi PE, McCulloch RE, Allenby GM (1996) The value of purchase history data in target marketing. *Marketing Sci.* 15(4):321–340.

Trusov M, Bodpati AV, Bucklin RE (2010) Determining influential users in Internet social networks. *J. Marketing Res.* 47(4): 643–658.

Van den Bulte C (2010) Opportunities and challenges in studying customer networks. Wuyts S, Dekimpe MG, Gijsbrechts E, Pieters R, eds. *The Connected Customer: The Changing Nature of Consumer and Business Markets* (Routledge, London), 7–35.