MARKETING SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0732-2399 | EISSN 1526-548X | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Predicting Individual Behavior with Social Networks

Sharad Goel, Daniel G. Goldstein
Yahoo Research, New York, New York, 10018

A basic objective of the social and economic sciences is to predict behavior. 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. While the similarity of connected individuals is well established, it is unclear if and how social data can predict behavior, and whether such predictions 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 offline department store to taking interest in an advertisement to joining a recreational league. Across all the domains, we find that social data are strongly informative in identifying individuals who are most likely to undertake actions, and that in identifying such individuals, social data generally improve the predictive accuracy of baseline models.

Key words: Computational social science, product adoption, social networks

---

Predicting individual behavior is a basic objective of the social sciences, from economics (Manski 2007) to psychology (Ajzen and Fishbein 1980), sociology (Burt 1987, Coleman et al. 1966), and beyond. For the marketer, predicting the future behavior of individuals is essential for targeting, selecting to whom to direct advertising and tailor products. A compelling contemporary question is whether recently available social network data, generated by firms such as Facebook, Google, Yahoo, Twitter and LinkedIn, can elevate the prevalent standards of prediction and targeting.

In traditional marketing practice, large markets are segmented into groups with homogeneous preferences and targeted by firms with marketing actions. When only a few television stations and magazines reached the majority of the population, marketing communication would reach both intended and unintended parties, leading to considerable waste (Iyer et al. 2005). As time progressed, technological changes led to more effective targeting. 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 multiplication in media in which a handful of television networks became hundreds, and relatively few print publishers began to compete with large numbers of websites. As a result broadcast advertising narrowed and efficiency increased. Since the rapid spread of the consumer Web beginning in the 1990s, the ability to target became even more precise. Targeting, through its history, has incorporated whatever predictors that were effective, affordable, and available. Accordingly, online marketers have, for over a decade, predicted behavior at the individual level using variables like age and sex (demographic targeting), location (geographic targeting), and website usage patterns (behavioral targeting). The broad segments of classical marketing strategy are being replaced with individual-level models, including those capable of learning in real time.

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. This paper focuses on social network data: connections (edges) between individuals (nodes) and records of individual behavior. Until now, the difficulty in observing social interactions has limited the feasibility of large-scale social prediction tasks, such as identifying the individuals in a population who are most likely to be influenced by advertising, adopt an innovative product, support a cause, abandon a service, or otherwise take action.

Though the idea that social network data may hold promise for improving targeting has received surprisingly little attention in the marketing literature, a handful of studies in other disciplines have shown that friends of adopters are themselves more likely to adopt, even after controlling for covariates (Hill et al. 2006, Bhatt et al. 2010, Provost et al. 2009). For example, Hill et al. (2006) identified a set of "network neighbors" (NNs) in the telecommunications domain who were socially connected to (i.e., had communicated with) people who had adopted a new service. By comparing these NNs to an equally sized group of non-NNs—matched on all available attributes but known not to be connected to an adopter—the authors demonstrated that having an adopting contact was a statistically significant predictor of adoption that led to improved rankings of prospects. In other 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. Using decision-tree models based on social network features (existence of adopting contacts, number of network neighbors, changes in network structure, etc.), user features (IM communication frequency, sex, age, etc.), they ranked customers according to their propensity to adopt in the next month. They find 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.

We build upon this previous work in several ways. The practical value of social network data for prediction can only be quantified relative to some baseline. By progressively adding stronger predictors—starting with basic demographics and moving on to granular, individual-level transaction data—we provide managerial insight into when it may be worthwhile to invest in social network data. Second, we study both first-time and repeat adoption, significantly expanding the range of applications for network-based predictions and enabling us to directly compare the predictive power of social data to individuals' past adoption history, a particularly strong baseline. Third, while past investigations are largely single-domain studies, we examine a dozen independent outcomes grouped into three domains. In particular, we study domains in which it is relatively costless (clicking on an ad), to moderately involved (signing up for a free service), to one with considerable monetary stakes (purchasing). While it may be the case that any given domain yields an idiosyncratic result, across multiple examples each experiment is placed in perspective, providing expectations for generalizing our findings. Lastly, we extend upon previous work by presenting and analyzing a novel theoretical framework for networked adoption that lends insight into when and by how much social data can be expected to improve predictions of individual behavior.

1. Related Work

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 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 (Christakis and Fowler 2007, Centola 2010), 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 non-causal effects (see Van den Bulte 2010, Peres et al. 2010, for useful reviews). For example, Manchanda et al. (2008) model the adoption of pharmaceuticals at the individual level as a joint consequence of contagion and marketing effects. Iyengar et al. (2011) report evidence for social influence and its moderators in the adoption of a risky drug. Trusov et al. (2010), present a technique to identify which social network members exhibit influence on the activity levels of others, and Godes and Mayzlin (2009) use 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) takes contagion as established and stresses moderators. De Bruyn and Lilien (2008) model word-of-mouth influence in viral marketing. Similarly, Aral et al. (Aral et al. 2009, Aral 2011) seek 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 can be used to improve targeting and prediction. Even when individuals influence one another, peer-to-peer transmissions can be so rare as to have little practical value. For example, Goel et al. (2012) show that in many domains of online diffusion, only a small percentage of adopters are actually influenced by a peer. And even when individuals are known *not* to influence one another, social ties may still be predictive, as observed in the sociological research on homophily (McPherson et al. 2001, Lazarsfeld and Merton 1954). For this reason, we investigate the predictive value of social data irrespective of causal influence, providing insight into the managerially-relevant question of the worth of social network data for targeting and prediction.

2. Data and Methods

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 to those individuals with at least one correspondent resulted in a symmetric network of 132 million people and 719 million edges, with a median number of 3 contacts per individual (mean 11). Figure 1 provides the degree distribution for this network. Demographic information (age and sex) was available from user profiles.

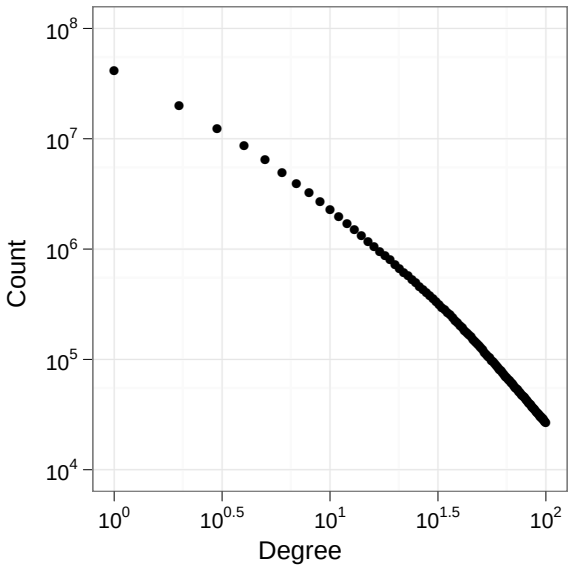

Figure 1 **Distribution of the number of social contacts per individual (degree) in the social network. To avoid counting mailing lists, a connection was only established between users if messages were exchanged in both directions during the two-month period.**

To assess the value of social predictors, we examine individual-level behaviors in three diverse domains comprising 12 distinct outcomes: purchasing (offline and online) at a national department store chain, participating in an online recreational league with millions of players, and responding to national advertisements for a variety of products and services.

Predictions of retail purchases are based on a 1.3 million member department store customer database, which includes one year's purchase data for each individual. Intersecting this dataset with the communications network results in approximately 588,000 people, for whom we have a record of both their own purchases (if any) as well as any purchases of their social contacts. To preserve anonymity, this matching was done by a specialist third party. Since not all customer accounts could be linked to the larger network, the social effects that we report are likely to be conservative. Nevertheless, 28% of the users in our dataset—a relatively large fraction—have at least one social contact with recorded transactions. The data for each customer were divided into two consecutive six-month periods, with information from the first period used to predict outcomes in the second. We used logistic regression models with purchasing as the outcome (a binary indicator of purchase in the second period), and the following predictors: demographics (age and sex, recorded for all users), past behavior (a binary indicator of purchase during the first period), and social contacts' behavior (the number who purchased during the first period). Related social predictors, such as the adoption rate among contacts, yield similar results. See the Appendix for details of the models. Adoption in this domain is related to the age and sex of the consumer, as shown in Figure 2.

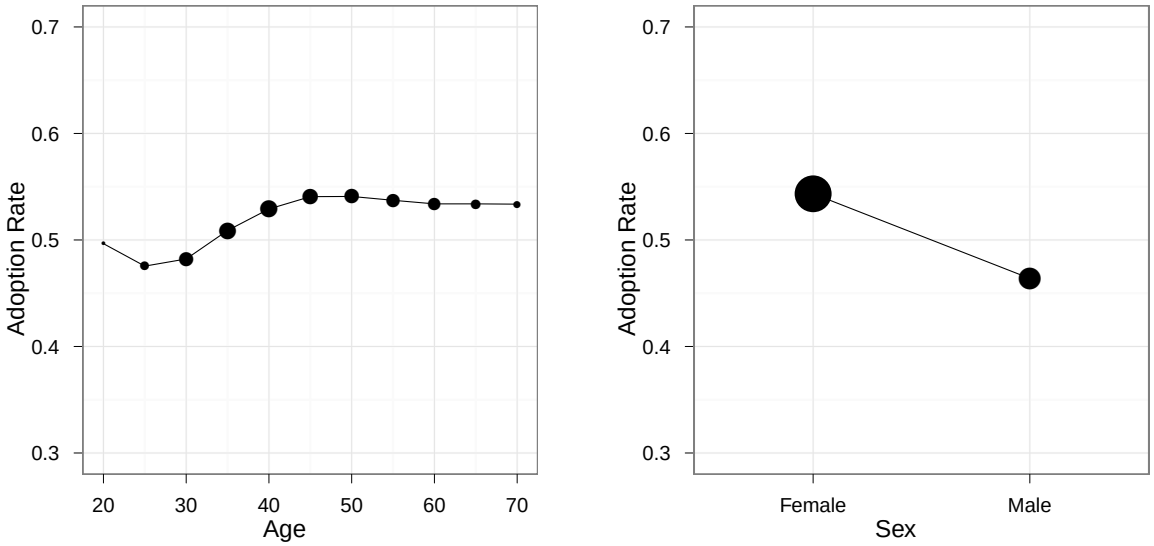

Figure 2 **Distribution of age and sex in the retail domain and how both variables relate to probability of adoption (purchase in the latter, six-month prediction period). The area of each point is proportional to the number of individuals in the corresponding category.**

To predict participation in a recreational league, we analyzed the Yahoo Sports Fantasy Football competition, which has approximately four million annual registrants. We reduce the communications network by intersecting it with the set of users who had made recent visits to the Yahoo Sports website, resulting in a population of 9.3 million, of whom approximately 6% ultimately participated in the competition. A logistic regression model was used to predict participation in the 2009 league on the basis of the individual's demographics and behavior (a binary indicator of participation in the 2008 league) and social contacts' behavior (the number who participated in the 2008 league). Though the overall probabilities of adoption are lower compared to the retail domain, age and sex are useful predictors, as seen in Figure 3. While the most-likely adopters in the retail domain are women over 40 years, those in the fantasy football domain are primarily males under 40. We note that this natural variation in domains helps make our study a robust test of the utility of social network data for prediction.

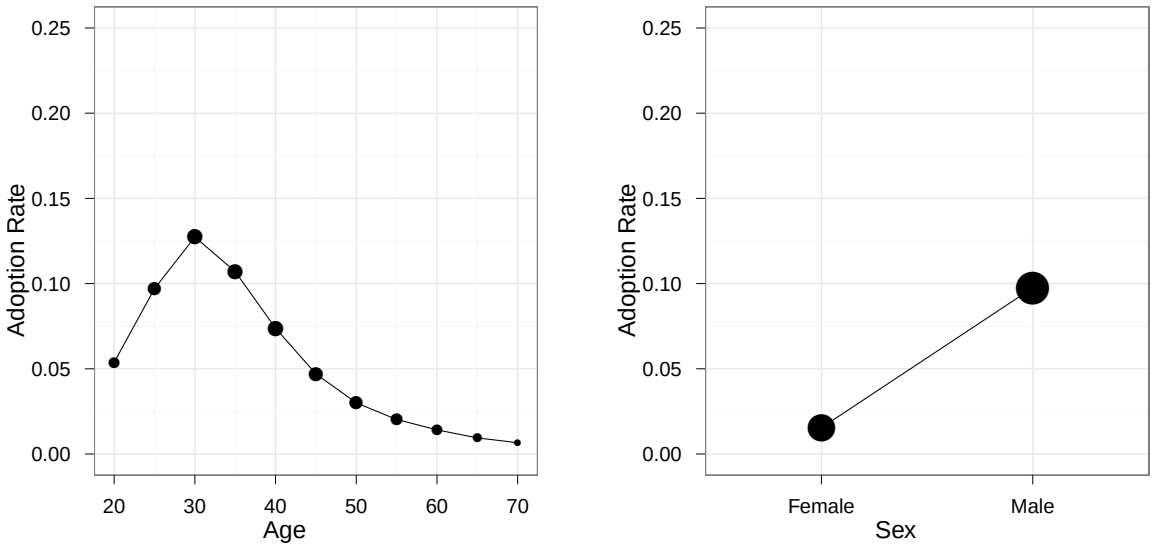

Figure 3 **Distribution of age and sex in the recreational league domain and the relationship of both variables to probability of adoption (participation in the subsequent season). The area of each point is proportional to the number of individuals in the corresponding category.**

Finally, we examine 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. As with the other domains, we intersect with the communications network, leaving 7–8 million users per campaign. As the ads varied substantially in their content, the click-through rates were correspondingly diverse, ranging from 0.01% to 0.6%. A segment-based model predicted clicks on the advertisement based on demographics and social contacts' behavior (binary indicator of whether at least one contact clicked on the ad).

3. Results

Figure 4 shows that contacts of adopters are themselves considerably more likely than average to adopt, consistent with previous investigations (Hill et al. 2006, Aral et al. 2009, Bhatt et al. 2010, Provost et al. 2009). For example, among retail consumers with four contacts who made a purchase during the first six-month observation period, 70% made a purchase themselves during the second period, substantially higher than the overall purchase rate (Fig 4A). Likewise, while 6% of all users participated in the recreational league, there is an approximate 50% adoption rate among users with four contacts who participated in the previous year's event (Fig 4B). Finally, we find similar results for advertising response, though the effect varies considerably across campaigns (Fig 4C). For five of the ten advertisements, users whose contacts clicked on a given ad had markedly higher rates of clicking themselves, with increases ranging from 20% to over 1,000%. For the remaining five campaigns, there were no significant effects, as detailed in the Appendix.

That contacts of adopters have relatively high adoption rates does not in itself imply that social data are valuable for prediction. For example, in the three domains we study, the vast majority of individuals have no adopting contacts, and social data consequently provide little differential information for most individuals. Relative to simply basing predictions on the overall adoption rate, adding social predictors does not appreciably improve average performance. In quantitative terms, social data reduce root mean squared error (RMSE) by less than 2% compared to the baseline in all cases we consider, as described in the Appendix.

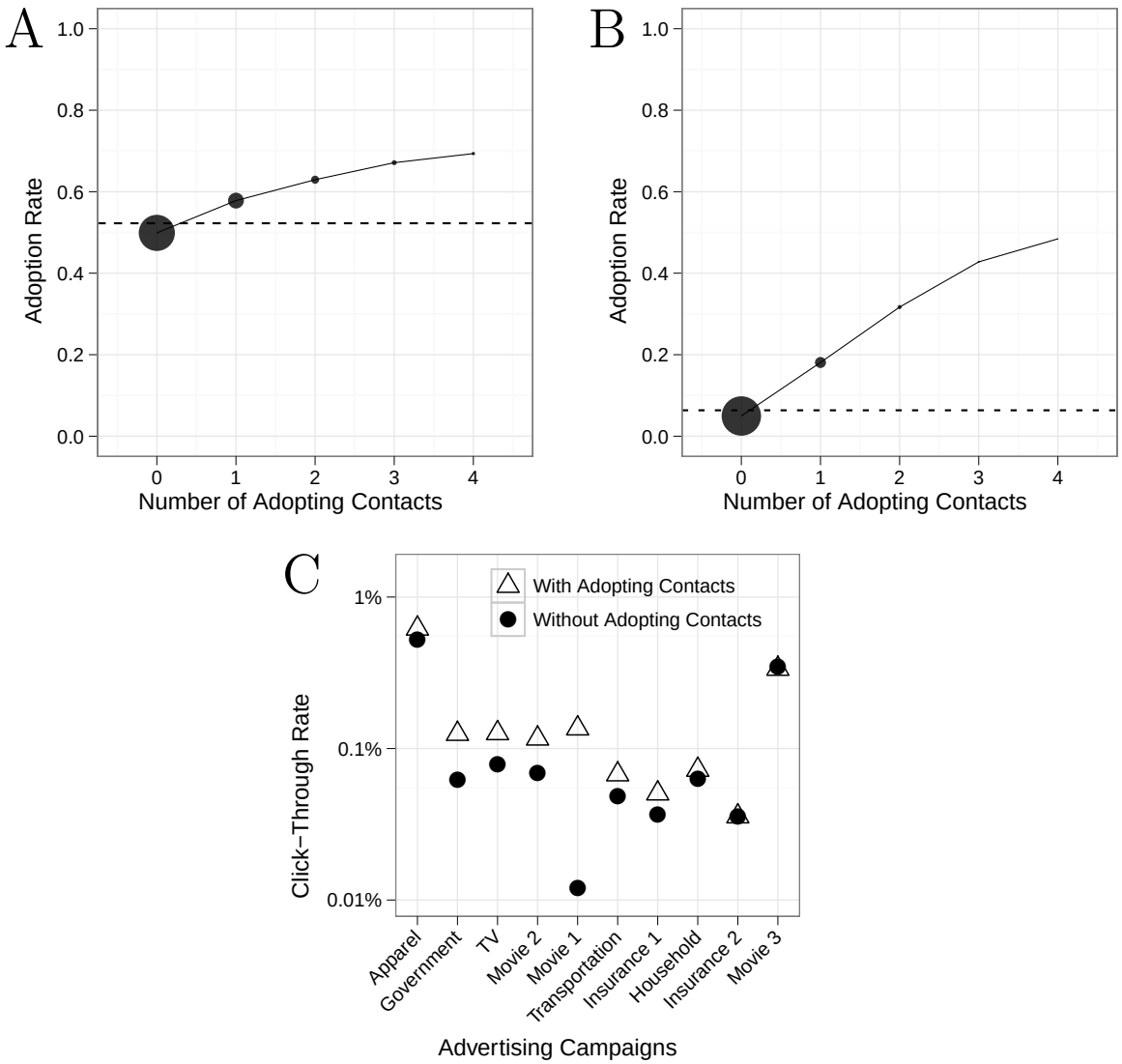

Figure 4 **Adoption rate of (A) purchasing at a department store, (B) joining a recreational league, and (C) clicking on a display advertisement as a function of having social contacts who have done the same. In the first two domains, the area of each point indicates the relative number of individuals in the respective category, and the dotted lines indicate the average adoption rate. In Panel C, campaigns are ordered by decreasing statistical significance of the social effect, and click-through rate is displayed on a log scale.**

3.1. Identifying those most likely to act

While mean squared error is a common measure of prediction performance, it is neither the only one nor always the most appropriate. When directing scarce resources either to encourage adoption (e.g., promoting energy-saving technologies) or to discourage adoption (e.g., anti-smoking campaigns), it is often useful to identify those individuals who are most likely to act. We test the predictive value of social data in such settings by examining the adoption rates of top candidates as selected by competing models. Specifically, we compare a baseline demographic model to a social model, and to a model fit on both social and demographic data. Modeling details have been placed in the Appendix for ease of exposition.

In the retail and recreation domains, Figure 5A–B shows that in a "top-k" competition, the social model substantially outperforms the demographic baseline in identifying those individuals

most likely to adopt. For example, the top retail consumers selected by the social model are approximately 20% more likely to adopt than those identified by the demographic model, and for the recreational league, the top social candidates are nearly twice as likely to adopt as the top demographic candidates. As indicated by the solid line in Figure 5, combining demographic and social data leads to further gains, a relatively modest boost in the retail domain but a substantial increase for the recreational league. While the social data provide an advantage for targeting those who are most likely to act, note that for targeting large percentages of the base (over 60% in the shopping domain and over 30% in the recreational league) the social data are not useful—using demographics alone is just as effective.

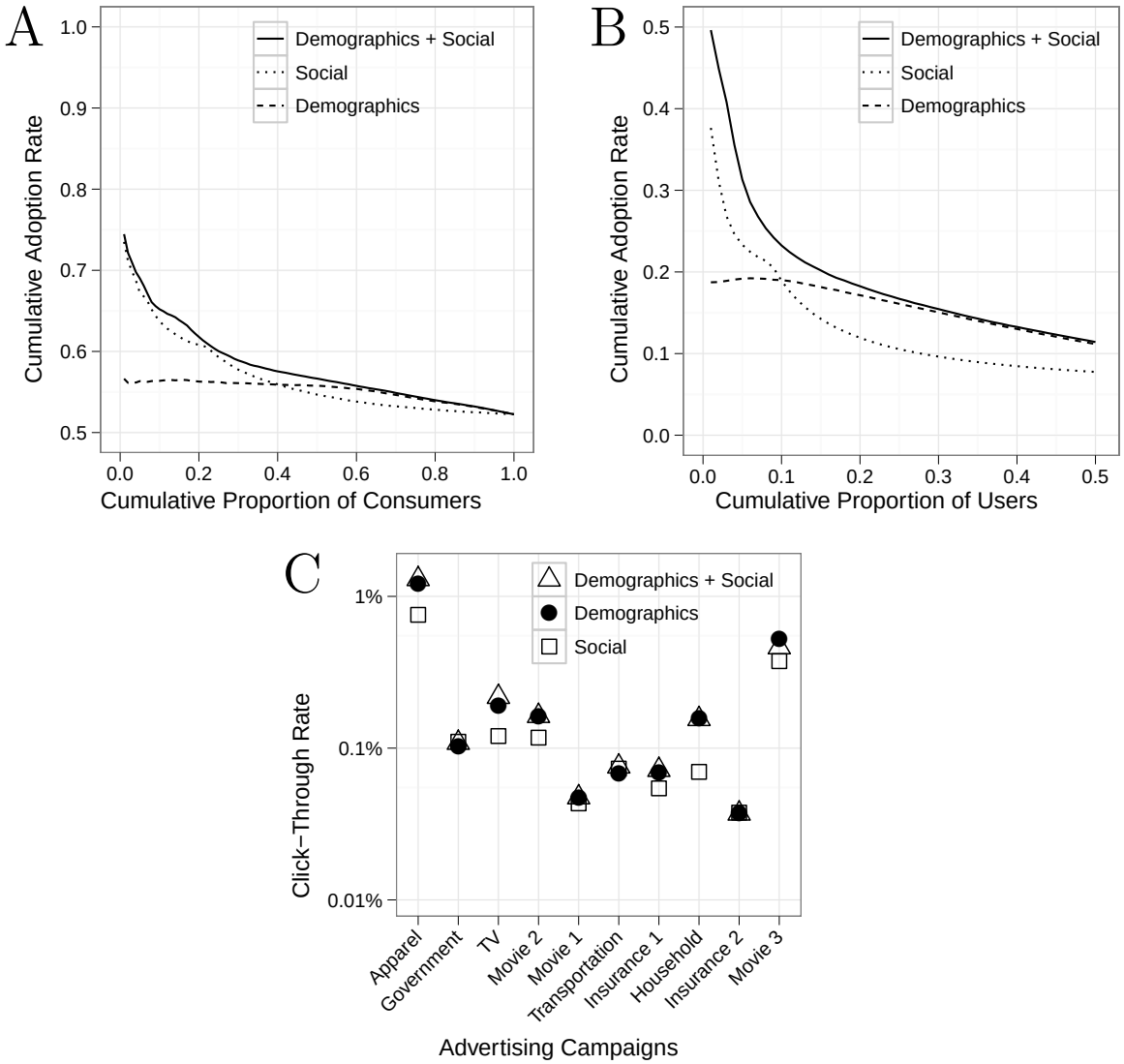

Figure 5 **Adoption rates for varying numbers of high-scoring individuals under a demographic model, a social model, and a model that includes both demographic and social attributes for the shopping (Panel A) and recreational league (Panel B) domains. The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to increasingly large proportions of the population (at the far right). Panel C: Click-through rates of the top 10,000 users as scored by the three models.**

Relative to demographics, social data were of little use for predicting clicks on the ten display advertisements. Figure 5C shows click-through rates for the top 10,000 users (top 0.1%) as predicted by demographic, social, and hybrid demographic-plus-social models. In identifying likely adopters, no campaign yields a statistically significant improvement when social data are added to the demographic baseline, a result that persists even when we examine a larger pool of candidates, as shown in the Appendix. We note that since click-through rates are low and network degrees are small, relatively few users are connected to any adopters at all. Thus, even though neighbors of adopters have relatively high adoption rates (Figure 4C), including them in an even moderately sized set of high-scoring individuals results in negligible improvement.

3.2. Initial adoption

For many applications (e.g., modeling the adoption of innovations), the relevant question is not simply who will undertake an action, but rather who will undertake that action for the first time. To address such situations, we examine the 53% of retail customers for whom no sales transaction was recorded for six months prior to the prediction period, and analogously for the recreational league, we restrict to the 94% of our sample who did not participate in the previous year's competition. Mirroring our results above, Figure 6 shows that social data substantially outperform the demographic baseline in identifying most-probable initial adopters, and moreover, additional gains are realized by combining social and demographic predictors. In the retail domain, for example, the top candidates selected by the combined social and demographic model are approximately 10% more likely to adopt than those selected by a model based solely on demographics. For the recreational league, including social data results in top candidates who are more than twice as likely to adopt than their demographically–selected counterparts. Again, when targeting larger fractions of the base (greater than 60% in the retail domain and 20% in the recreational league), social data do not improve predictions. A clear application of these data seems to be identifying those who are most likely to act.

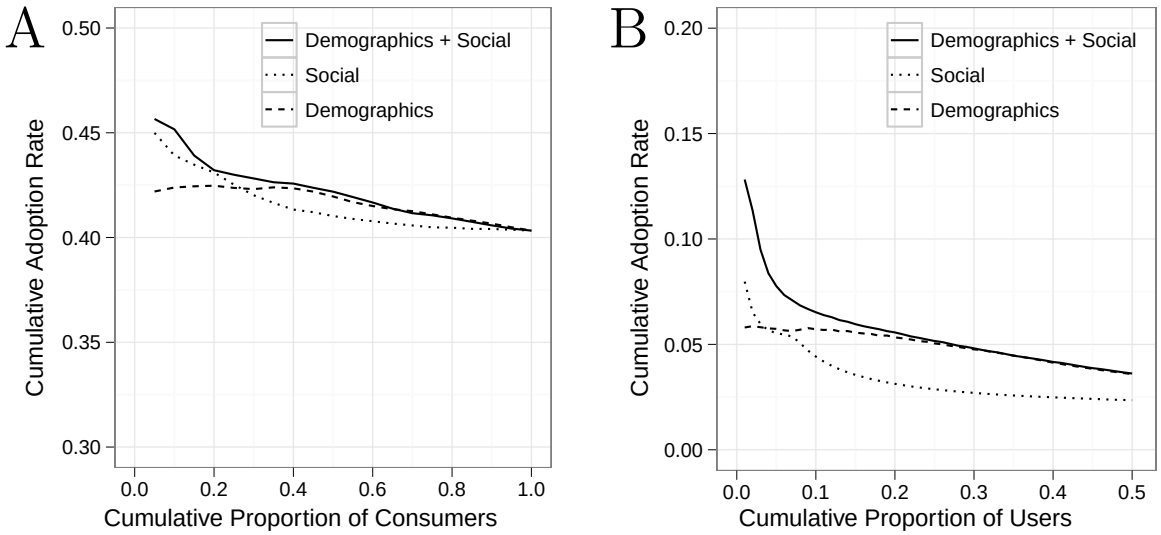

Figure 6 **Adoption rates for users with no record of past adoption, ranked by a demographic model, a social model, and a combined model in the (A) department store and (B) recreational league domains. The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to increasingly large proportions of the population (at the far right).**

3.3. Incorporating transactional data

While we have thus far evaluated the utility of social data in augmenting demographic predictors, richer baselines are available in select settings. Often, the best predictor of future behavior is past behavior. As seen in Figure 7, 65% of department store consumers who made a purchase during the first period of our dataset went on to make purchases during the latter period, and 79% of users who participated in the previous year's fantasy football competition played again the subsequent year. (In the advertising domain, past behavior is unavailable as the ads were shown only for a single day.)

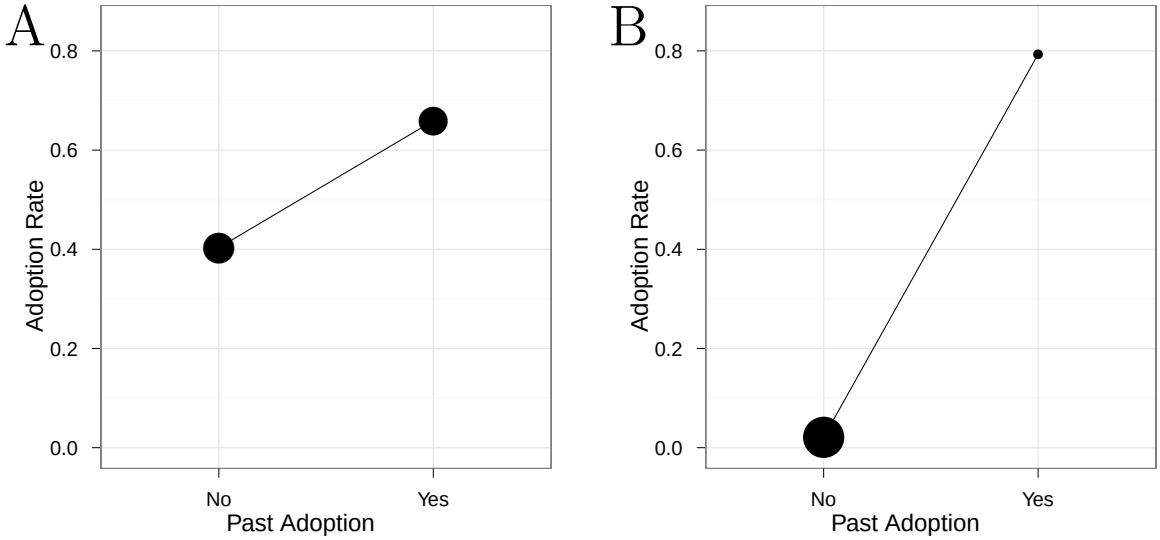

Figure 7 **Probability of adoption conditional on past adoption in the retail (panel A) and recreational league (panel B) domains. The area of each point is proportional to the number of individuals in the corresponding category.**

Figure 8A–B indicates that even relative to a strong behavioral-plus-demographic baseline that includes one's past adoption status, the marginal value of social data in identifying adopters remains substantial. Finally, in the retail domain, we include not only whether a consumer has made past purchases, but also the dollar amounts of those sales. As shown in Figure 8C, against a model built on both demographics and detailed transactional information, the marginal value of social data is negligible, illustrating that rich measures of past behavior constitute particularly strong baselines when forecasting individual-level adoption.

We have seen over multiple examples how social data were most valuable for targeting a small percentage of individuals and least valuable in domains in which alternative predictors constituted a strong baseline (in particular, transactional data). Beyond these examples, it would be useful to characterize the kinds of domains in which social data would be expected to be predictive. In the next sections, we provide a theoretical framework to assess new domains and propose a simple social marketing strategy that should be of interest to managers.

4. A Theoretical Framework for Network-Based Predictions

To better understand the value of social data in predicting individual behavior, we introduce a novel model of network formation and adoption, and in turn analytically assess the performance of network predictions across a range of parameters. Our model captures two key quantities that intuitively affect predictive performance: (1) variation across individuals in their likelihood to adopt (i.e., the degree to which some individuals are more likely to adopt than others); and (2)

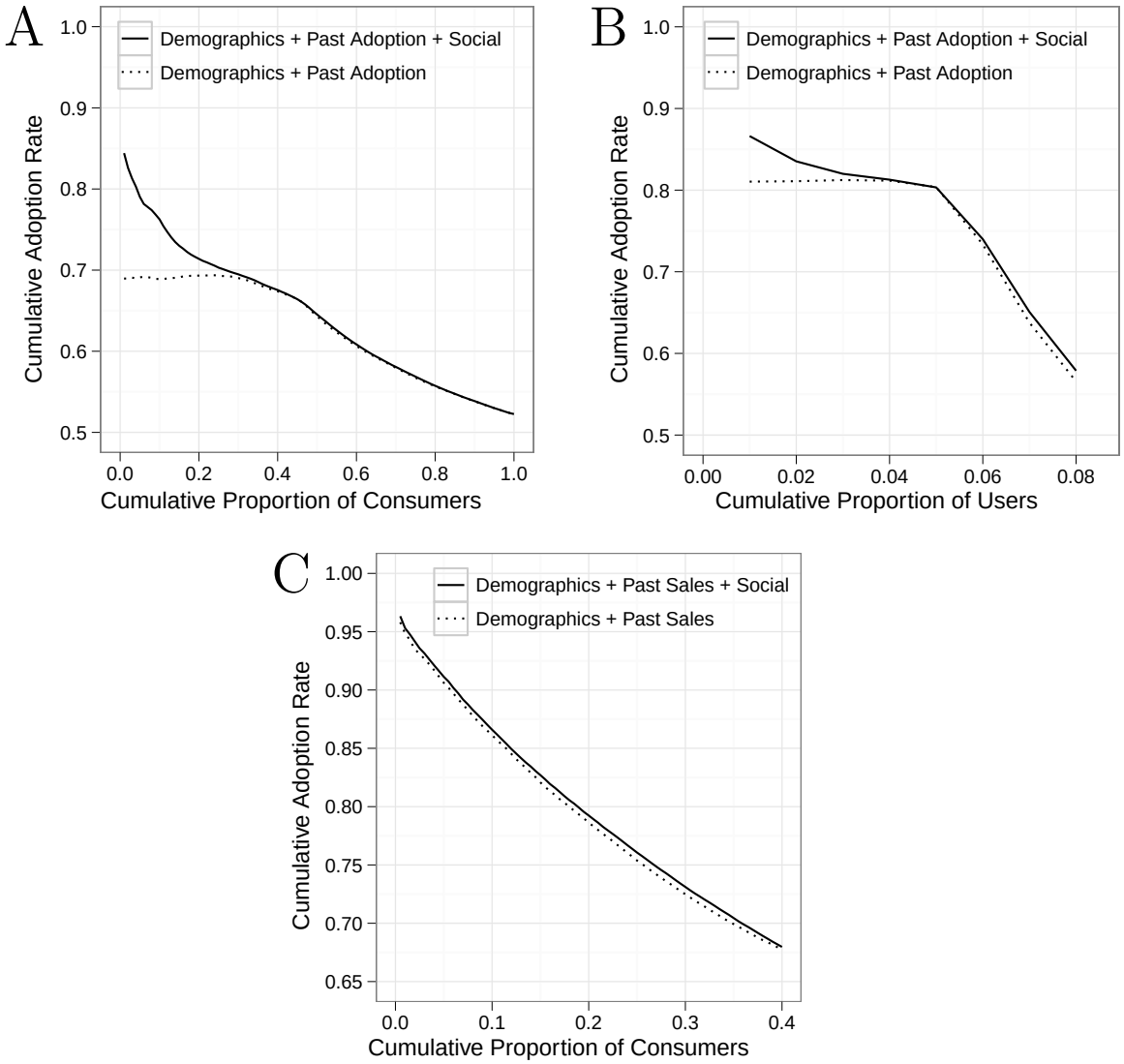

Figure 8 **Adoption rates for individuals ranked by demographic and behavioral predictors, with and without incorporating social data. Panels A and B plot results for the retail and recreational league domains for models that include an individual's (binary) past adoption status, while the retail model in panel C incorporates (continuous) dollar amounts for past purchases. The set of individuals included in the cumulative averages varies from the highest-scoring individuals according to each model (at the far left) to increasingly large proportions of the population (at the far right).**

the tendency for individuals to connect to others with adoption propensities similar to their own. Imagine, for example, that the population consists of two relatively insular subpopulations of high and low likelihood-to-adopt individuals. Then the adoption behavior of one's contacts is a strong signal for determining the subpopulation in which an individual lies, and thus their propensity to adopt. If, by contrast, individuals are equally likely to form ties with high and low likelihood candidates, network data would not improve predictions.

4.1. A model of networked adoption

Our formal analysis is based on stochastic block models (SBMs) (Holland 1976), one of the simplest and most widely used classes of network models with community structure. In these networks, nodes belong to one of K communities, or blocks, and the probability of an edge between any two nodes depends only on their corresponding block assignments. The specific generative process for

a network with n nodes and K blocks is as follows. First, for each node v_i, independently and uniformly at random assign it to one of K blocks $\{1,\ldots,K\}$. Next, for each ordered pair of nodes (v_i, v_j), flip a coin with bias θ_+ (resp. θ_-) for nodes in the same (resp. different) blocks to determine if an edge exists from v_i to v_j. Networks generated under this model may loosely be characterized as a mixture of Erdos-Renyi networks, with an edge density θ_+ within blocks and θ_- between. In the assortative case ($\theta_+ > \theta_-$), nodes tend to form more edges within than between their blocks, resulting in dense communities of interconnected individuals.

To complete our model of networked adoption, we specify that nodes independently adopt with probability that depends only on their block assignments. Specifically, for a set of parameters $p_1,\ldots,p_K$, one for each of the K blocks, a node in the k-th block adopts with probability p_k, independent of the adoption decisions of other nodes. Despite its simplicity, this model captures one of the most salient features of social networks, namely the tendency of individuals to cluster into communities of similar others. For example, the "blocks" in our model could correspond to socioeconomic segments of the population, whose constituents may both have similar adoption propensities for any given product and be disproportionately likely to form within-group ties.

As described above, stochastic block models are parameterized by four values: n, K, θ_-, θ_+. A straightforward calculation shows that each node has an expected number of $d_+ = (n-1)\theta_+/K$ within-block neighbors, and $d_- = (n-1)\theta_-(K-1)/K$ between-block neighbors. SBMs are thus equivalently parameterized by the four values n, K, d_+, d_-, where we can convert between the two specifications by noting $\theta_+ = Kd_+/(n-1)$ and $\theta_- = d_-K/[(n-1)(K-1)]$. For our purposes, it is most convenient to describe SBMs in terms of the expected number of within and between block neighbors $d_\pm$, and we continue with that notation throughout the remainder of our discussion.

In our model of network formation and adoption, the probability that a randomly selected node adopts is $p_{\text{avg}} = (1/K)\sum_{k=1}^{K} p_k$. When a node has adopting neighbors, however, it is more likely to be in a block of high-likelihood adopters, and so its conditional probability of adoption may be substantially higher than p_{avg}. This effect is not a result of social influence—all individuals make their adoption decisions independently—but rather is due to the fact that individuals with similar adoption propensities tend to cluster together. Theorem 1 below precisely quantifies the value of network information for predicting behavior in this model, deriving the likelihood of a node adopting conditional on its neighbors actions.

THEOREM 1. *For fixed constants K, d_-, d_+, consider an n-node (random) network G_n drawn from the family of stochastic block model networks SBM(n, K, d_+, d_-). Further suppose that for $1 \le k \le K$, a node in the k-th block of G_n adopts with probability $p_k > 0$. Then for*

$$w_k = p_k\left(d_+ - \frac{d_-}{K-1}\right) + \frac{d_-}{K-1}\sum_{\ell=1}^{K} p_\ell.$$

we have

$$\lim_{n\to\infty} \mathbb{P}_n(v \in G_n \text{ adopts} \mid v \text{ has } m \text{ adopting neighbors}) = \frac{\sum_{k=1}^{K} p_k w_k^m e^{-w_k}}{\sum_{k=1}^{K} w_k^m e^{-w_k}} \tag{1}$$

and

$$\lim_{n\to\infty} \mathbb{P}_n(v \in G_n \text{ adopts} \mid v \text{ has at least one adopting neighbor}) = \frac{\sum_{k=1}^{K} p_k\left(1-e^{-w_k}\right)}{\sum_{k=1}^{K} 1-e^{-w_k}}. \tag{2}$$

Before deriving the result, we note that the limiting conditional probabilities in Equations (1) and (2) can be viewed as weighted averages of the adoption propensities p_k, where the weights w_k capture the level of homophily in the network. For example, when there is no homophily (i.e., $\theta_- = \theta_+$), we have $d_+ = d_-/(K-1)$ and so the weights w_k are equal for all k. Thus, as intuition

suggests, social data do not provide any information in this case, with the conditional adoption probabilities equal to the average adoption probability p_{avg} regardless of one's number of adopting neighbors. In another extreme where there is no variance in adoption propensity between blocks (i.e., $p_k = p_{\text{avg}}$), social data again do not provide any information, even when there is significant homophily and the weights are correspondingly skewed. The value of Theorem 1 thus primarily lies in the remaining cases where there is both homophily and variance in adoption propensity.

Proof of Theorem 1. We start by proving Eq. (1). For each graph G_n and each node $v \in G_n$,

$$\begin{aligned}
&\mathbb{P}_n(v \text{ adopts} \mid v \text{ has } m \text{ adopting neighbors}) = \\
&\quad = \frac{\mathbb{P}_n(v \text{ adopts } \& \ v \text{ has } m \text{ adopting neighbors})}{\mathbb{P}_n(v \text{ has } m \text{ adopting neighbors})} \\
&\quad = \frac{\sum_{k=1}^{K} \mathbb{P}_n(v \text{ adopts } \& \ v \text{ has } m \text{ adopting neighbors} \mid v \text{ is in block } k)\mathbb{P}_n(v \text{ is in block } k)}{\sum_{i=1}^{K} \mathbb{P}_n(v \text{ has } m \text{ adopting neighbors} \mid v \text{ is in block } k)\mathbb{P}_n(v \text{ is in block } k)}.
\end{aligned}$$

Note that $\mathbb{P}_n(v \text{ is in block } k) = 1/K$, and furthermore, v adopting and v having m adopting neighbors are conditionally independent given v's block assignment. Consequently,

$$\mathbb{P}_n(v \text{ adopts} \mid v \text{ has } m \text{ adopting neighbors}) = \frac{\sum_{k=1}^{K} p_k \mu_{n,k,m}}{\sum_{k=1}^{K} \mu_{n,k,m}} \tag{3}$$

where $\mu_{n,k,m} = \mathbb{P}_n(v \text{ has } m \text{ adopting neighbors} \mid v \text{ is in block } k)$.

Given that a node is in the k-th block, its number of adopting neighbors follows a binomial distribution, where the probability that any one of the other $n-1$ nodes is an adopting neighbor of v satisfies

$$\begin{aligned}
&\mathbb{P}_n(w \text{ is an adopting neighbor of } v \mid v \text{ is in block } k) = \\
&\quad = \sum_{\ell=1}^{K} \mathbb{P}_n(w \text{ is an adopting neighbor of } v \mid v \text{ is in block } k,\ w \text{ is in block } \ell)\mathbb{P}_n(w \text{ is in block } \ell) \\
&\quad = \frac{1}{K} \sum_{\ell=1}^{K} p_\ell \mathbb{P}_n(w \text{ is a neighbor of } v \mid v \text{ is in block } k,\ w \text{ is in block } \ell) \\
&\quad = \frac{p_k \theta_+}{K} + \frac{\theta_-}{K} \sum_{\ell \neq k} p_\ell \\
&\quad = \frac{p_k d_+}{n-1} + \frac{d_-}{(n-1)(K-1)} \sum_{\ell \neq k} p_\ell.
\end{aligned}$$

Consequently, the probability v has m adopting neighbors given that it is in the k-th block is:

$$\mu_{n,k,m} = \binom{n-1}{m} \left[\frac{p_k d_+}{n-1} + \frac{d_-}{(n-1)(K-1)} \sum_{\ell \neq k} p_\ell \right]^m \left[1 - \frac{p_k d_+}{n-1} + \frac{d_-}{(n-1)(K-1)} \sum_{\ell \neq k} p_\ell \right]^{n-1-m}.$$

To conclude, we find the limiting value of $\mu_{n,k,m}$ as n tends to infinity. First note that since $\binom{n-1}{m} = n^m/m! + o(n^m)$,

$$\begin{aligned}
\lim_{n\to\infty} \binom{n-1}{m} \left[\frac{p_k d_+}{n-1} + \frac{d_-}{(n-1)(K-1)} \sum_{\ell \neq k} p_\ell \right]^m &= \frac{1}{m!} \left[p_k d_+ + \frac{d_-}{K-1} \sum_{\ell \neq k} p_\ell \right]^m \\
&= \frac{w_k^m}{m!}
\end{aligned}$$

where the final equality comes from regrouping terms in w_k. Furthermore,

$$\lim_{n\to\infty}\left[1-\frac{p_k d_+}{n-1}+\frac{d_-}{(n-1)(K-1)}\sum_{\ell\neq k}p_\ell\right]^{n-1-m} = \exp\left(-p_k d_+ - \frac{d_-}{K-1}\sum_{\ell\neq k}p_\ell\right) = e^{-w_k}.$$

Combining the above limits, we have $\lim_{n\to\infty}\mu_{n,k,m} = w_k^m e^{-w_k}/m!$. Eq. (1) now follows from Eq. (3), where we note the $m!$ terms cancel as they appear in both the numerator and denominator of of the result.

To show the second statement of the Theorem (Eq. 2) note that similar to Eq. (3),

$$\mathbb{P}_n(v \text{ adopts} \mid v \text{ has at least one adopting neighbor}) = \frac{\sum_{k=1}^K p_k\mu_{n,k}}{\sum_{k=1}^K \mu_{n,k}} \tag{4}$$

where $\mu_{n,k} = \mathbb{P}_n(v \text{ has at least one adopting neighbor} \mid v \text{ is in block } k)$. Moreover, since $\mu_{n,k} = 1-\mu_{n,k,0}$, the above shows that $\lim_{n\to\infty}\mu_{n,k} = 1-e^{-w_k}$, establishing the result.
□

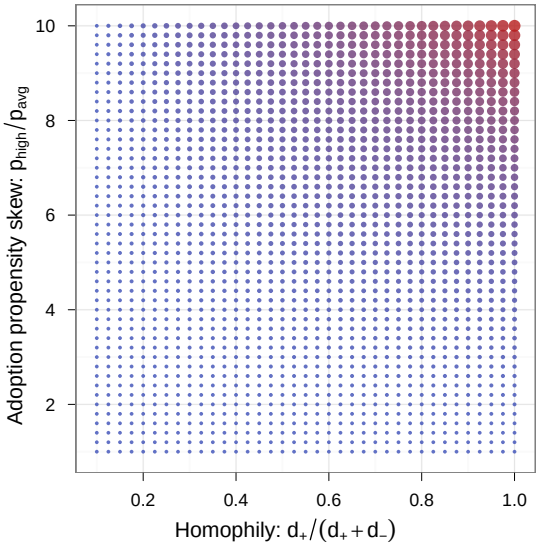

Figure 9 **The value of network information for predicting individual behavior in a stylized model of network formation and adoption. The area (and color) of each point is proportional to the probability of adoption for individuals with at least one adopting neighbor, where model parameters are given by the x and y axes.**

We illustrate the magnitude of the effects suggested by Theorem 1 by explicitly working out a simple example of networked adoption. Suppose G_n is an n-node SBM graph with $k = 10$ blocks and expected degree 20 (i.e., $d_+ + d_- = 20$). Further suppose that exactly one of the ten blocks consists of high-likelihood adopters—having adoption probability p_{high}—and that the remaining nine blocks are comprised of low-likelihood adopters, with adoption probability p_{low}. In this case, the average adoption probability is $p_{\text{avg}} = 0.1p_{\text{high}} + 0.9p_{\text{low}}$. Now, holding the expected total degree constant at 20, and the average adoption probability at 1%, Figure 9 shows the value of social data as we vary network homophily and skew in adoption probabilities. In particular, if, on average, half a node's neighbors are in its own block (i.e., $d_+ = 10$) and $p_{\text{high}} = 6\%$ (implying $p_{\text{low}} = 0.4\%$), then the

adoption probability among nodes with an adopting neighbor is 2%, twice as likely than average. Thus, as this example shows, when high propensity adopters cluster together, social network data can be quite useful in identifying those individuals most likely to adopt.

In our empirical analyses, we found that the predictive power of social data was relatively stronger in the shopping and recreational domains than in the advertising realm. Viewed in terms of our theoretical model, one possible explanation of this result is that tie formation is more closely related to shared interests such as fantasy football than to the propensity to click on advertisements. That is, those who play fantasy football may cluster together more than those who are likely to respond to advertisements. While connected individuals may have similar propensities to click due to other shared characteristics such as age and sex, this effect is presumably smaller than in domains where the relevant activity is more inherently social.

4.2. A Social Marketing Strategy

As the above empirical and theoretical analyses demonstrate, connection to an adopter is a useful predictor of adoption. For new products and campaigns, however, there are no prior adopters and thus no connections to use for prediction. A two-stage process could nonetheless solve this problem. Specifically, the first stage of an advertising campaign could use standard demographic and behavioral targeting measures to define and advertise to a set of targets. The campaign would yield a set of adopters, whose contacts would then be advertised to in the second stage. Theorem 2 below derives the effectiveness of this social marketing strategy.

THEOREM 2. *Consider a set of n nodes $v_1, \ldots, v_n$ such that each node has k neighbors, and that the sets $\Gamma_i = \{v_i\} \cup \{\text{neighbors of } v_i\}$ are disjoint (i.e, the nodes are relatively isolated from one another). Suppose the nodes v_i independently adopt with probability p_{avg}, and that the conditional probability of a node adopting given it has an adopting neighbor is p_{social}. For the two-stage targeting strategy described above, where the n nodes $\{v_i\}$ are targeted in the first stage, let ρ_n be the (random) fraction of targeted nodes that adopt over the two stages. Then,*

$$\lim_{n\to\infty} \rho_n = p_{avg}\left(\frac{1+kp_{social}}{1+kp_{avg}}\right) \qquad \textit{almost surely.} \tag{5}$$

Proof. Let Y_i indicate the number of nodes that are eventually targeted as a result of including v_i in the first stage. That is, $Y_i = 1$ if v_i does not adopt and $Y_i = k+1$ if v_i does adopt. Consequently, $\mathbb{E}Y_i = 1 + kp_{\text{avg}}$. Furthermore, let X_i be the number of nodes that eventually adopt as a result of including v_i in the first stage. Then, $X_i = 0$ if v_i does not adopt, and $X_i = 1 + \text{Binomial}(k, p_{\text{social}})$ if v_i adopts. We thus have $\mathbb{E}X_i = p_{\text{avg}}(1 + kp_{\text{social}})$. Finally, note that

$$\rho_n = \frac{X_1 + \cdots + X_n}{Y_1 + \cdots + Y_n}$$

and so the result follows from the law of large numbers.
□

To give a specific application of Theorem 2, suppose $p_{\text{avg}} = 1\%$, $p_{\text{social}} = 2\%$, and $k = 100$. Then the limiting adoption rate of the social marketing strategy is $1.5p_{\text{avg}}$, a 50% increase in adoption relative to a traditional, single-stage campaign. Leveraging social information may thus lead to substantial performance increases over strategies based solely on demographic and behavioral data. We caution, however, that when adoption rates are low and individuals have few contacts, contacts of adopters form a relatively small set and thus the reach of this strategy would be reduced accordingly.

5. Conclusion

If new data sources are to extend the limits of predictive accuracy, they must not merely substitute for traditional predictors, but should complement them. In three diverse domains, we find that social data do in fact augment traditional methods of predicting the behavior of individuals. In particular, those candidates that are both demographically and behaviorally suited to a product and who also have an adopting contact are much more likely than average to themselves adopt. While social data are particularly good at identifying such select, high-likelihood candidates, we find they have limited use when targeting large segments of the customer base, where classical predictors are difficult to improve upon.

We find social data tend to be useful in both isolation and in conjunction with traditional covariates. The magnitude of improvement, however, is contingent on the domain and the nature of existing predictors. Across several varied domains, we observe that social data contribute materially to predictive abilities when only basic demographics are known about a targeted base, and that this predictive advantage decreases as individual-level transactional data become available. Accordingly, social targeting seems particularly worthwhile in situations where a target's social network is known, but past behavior—and possibly demographic information—are absent. There are at least two common situations in which this happens. First, when new members join existing social networks, they often quickly link to their associates who are already members, for example by importing contacts from their email accounts. Second, new members on many Web sites have the option to link their site accounts to their social network accounts. In both these scenarios, 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.

To generalize beyond our empirical examples, we developed a theoretical model to think about new domains on two key dimensions: variation in the likelihood of adopting, and the tendency to have social ties to people with similar adoption propensities. Together, these dimensions help characterize when social data will be predictive. Our analytic results suggest social data are particularly valuable in domains where there is a small group of tightly knit, high-likelihood adopters. In this case, having an adopting contact is a strong signal that one will adopt. This theoretical heuristic appears consistent with our empirical findings. Moreover, our analysis shows that such domains are particularly well-suited for two-stage social targeting strategies in which candidates are first identified through classical methods, yielding adopters whose contacts can then be targeted in the second phase.

Finally, we note that nearly all the marketing literature we have reviewed has focused on the topic of proving, disentangling, or modeling causality in diffusion across social networks. It may be tempting to conclude from our results that shopping habits or leisure activities are contagious. Though there is probably some truth to that claim, establishing such is neither our objective nor justified from our analysis (Shalizi and Thomas 2010). Nevertheless, while the value of social data may concern both influence and homophily, our approach demonstrates that disentangling the two is not necessary for improving predictions of individual behavior.

Acknowledgments

We thank Jake Hofman, Randall Lewis, David Pennock, David Reiley, and Duncan Watts for their feedback.

Appendix.

We provide additional details on model fitting and evaluation for each of the three domains: retail purchases, league registrations, and response to advertising.

Retail Purchases. One year of purchase data was divided into two consecutive six-month periods, and the task was to predict adoption during the latter period based only on information available during the first.

For a given individual with k adopting contacts, the social model predicts he or she adopts with probability equal to the adoption rate among all individuals with k adopting contacts. For example, as indicated by Figure 1A, the social model predicts that those with no adopting contacts themselves adopt with probability 0.50 and that those with one adopting contact adopt with probability 0.58. We evaluate the overall prediction performance of this model by comparing with the simple baseline of always predicting an individual adopts with probability 0.52, the adoption rate over the entire population. In terms of root mean squared error (RMSE), the social model (0.497) yields negligible improvement over the baseline (0.499).

To identify those individuals most likely to adopt, the social model (Figure 2A) orders customers by number of contacts who made purchases during the first period, with ties broken at random. The demographic and demographic-plus-social models (Figure 2A) were based on logistic regression. For the demographic model, predictors included sex, age and powers of age, and their interactions; the demographic-plus-social model additionally included the number of (period 1) adopting contacts together with interactions. These same models were used to predict initial adoption (Figure 3A), where the only difference is that the models were trained and evaluated on the set of users for whom no sales transaction was observed during the first period.

Logistic regression models were also used to identify likely adopters when incorporating an individual's past behavior (Figure 4). In Figure 4A, the baseline model included sex, age and powers of age, a dichotomous variable indicating whether the individual had made a purchase during period 1, and interactions of these terms; the baseline-plus-social model additionally included each user's number of adopting contacts along with interactions. Finally, Figure 4C augmented the above baseline and baseline-plus-social models by adding as a predictor the amount spent by each individual during the first period. Five-fold cross validation was used to generate all predictions.

Recreational League Registrations. In the recreation domain, the task was to predict registration for Yahoo Fantasy Football in 2009 based on demographic, behavioral, and social predictors available in 2008. Our models take the same form as those used for retail purchases. In particular, we evaluate overall performance by comparing social predictions to a baseline, constant prediction of 0.06—the population adoption rate. As in the retail domain, we find a small difference in RMSE between the social (0.239) and the baseline (0.244) models. Similarly, to identify high-likelihood adopters, the social-only models in Figures 2B and 3B rank users by number of contacts who participated in 2008; and the demographic and demographic-plus-social models are based on logistic regression with predictors including sex, age and powers of age, number of adopting contacts (in the social variant), and interactions. The models in Figure 4 augment those described above by including a dichotomous variable indicating whether a user participated in the 2008 competition. All regression models were trained on a randomly selected set of 10,000 users and predictions were evaluated on the remaining users.

Response to Advertising. Our analysis of online advertising was based on ten different campaigns, spanning a variety of services and products. As noted in the main text, for five of the ten campaigns we did not observe statistically significant differences (at the 95% level) in click-through rates between users with and without contacts who had clicked the ads (Figure 10). Furthermore, none of the campaigns exhibited appreciable differences in RMSE between a social model and a baseline model that predicts each individual has click-through rate (CTR) equal to the population average.

As with retail purchases and fantasy football signups, we evaluated predictive models to gauge the relative value of social data compared to demographic information for identifying high-likelihood adopters (Figure 2C). For each campaign, models were trained on a randomly selected set of 1 million users and were tested on the remaining users. To identify top demographic candidates, mean CTRs were estimated for 14 demographic groups (e.g., 20–30 year-old males, 20–30 year-old females, etc.) on the training set, and users in the test set were ranked based on these estimated CTRs. For the demographic-plus-social model, users with adopting contacts were promoted to the top of their demographic group. We note that these segment-based models performed better than traditional regression models. Finally, the social-only model ranked users by number of adopting contacts. As shown in Figure 10 panels B and C, none of the ten campaigns exhibited statistically

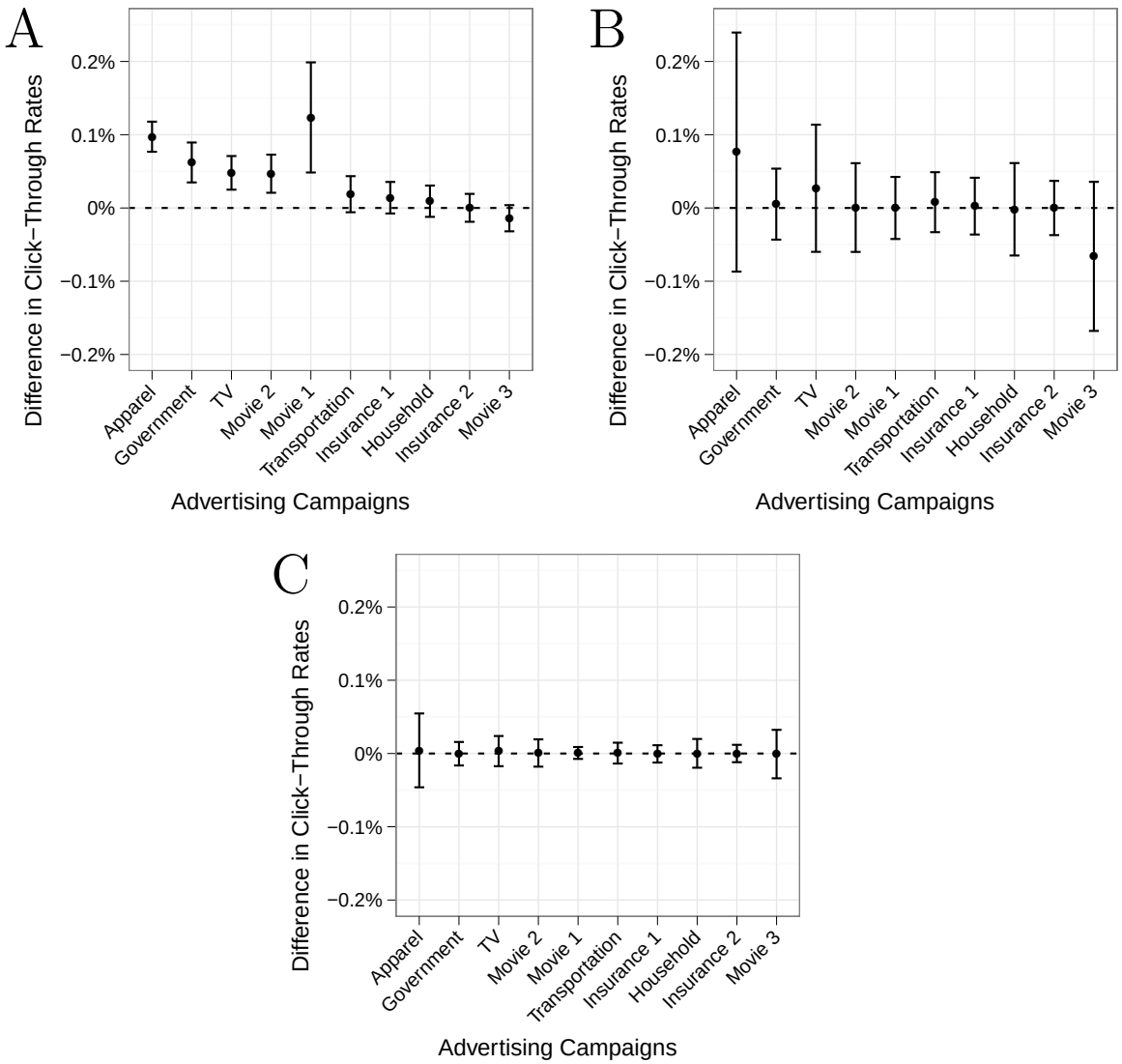

Figure 10 **Figure 10: Difference in click-through rates between users with and without adopting contacts (A), and top 10,000 (B) and top 100,000 (C) candidates selected by a demographic-plus-social model and a model based solely on demographics. Error bars indicate 95% confidence intervals.**

significant differences in CTR between top candidates identified by the demographic and the demographic-plus-social models.

References

Ajzen, Icek, Martin Fishbein. 1980. *Understanding attitudes and predicting social behavior*. Prentice Hall, Englewood Cliffs, NJ.

Aral, Sinan. 2011. Commentary–identifying social influence: A comment on opinion leadership and social contagion. *Marketing Science* **30**(March/April) 217–223.

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

Bass, Frank M. 1969. A new product growth model for consumer durables. *Management Science* **15** 215–227.

Bhatt, Rushi, Vineet Chaoji, Rajesh Parekh. 2010. Predicting product adoption in large-scale social networks. *Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM)*.

Burt, Ronald S. 1987. Social contagion and innovation: cohesion versus structural equivalence. *American Journal of Sociology* **92**(6) 1287–1335.

Centola, D. 2010. The Spread of Behavior in an Online Social Network Experiment. *Science* **329**(5996) 1194.

Christakis, N.A., J.H. Fowler. 2007. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine* **357**(4) 370.

Coleman, James S., Elihu Katz, Herbert Menzel. 1966. *Medical innovation: A diffusion study*. Bobbs-Merrill, Indianapolis.

De Bruyn, Arnaud, Gary L. Lilien. 2008. A multi-stage model of word-of-mouth influence through viral marketing. *International Journal of Research in Marketing* **25** 151–163.

Godes, David. 2011. Commentary–invited comment on 'opinion leadership and social contagion in new product diffusion'. *Marketing Science* **30**(March/April) 224–229.

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

Goel, S., Watts D.J., D.G. Goldstein. 2012. The structure of online diffusion networks. *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*.

Hill, Shawndra, Foster Provost, Chris Volinsky. 2006. Network-based marketing: Identifying likely adopters via consumer networks. *Statistical Science* **21**(2) 256–276.

Holland, P.W. 1976. Local structure in social networks. *Sociological Methodology* **7** 1–45.

Iyengar, Raghuram, Christophe Van den Bulte, Thomas W. Valente. 2011. Opinion leadership and social contagion in new product diffusion. *Marketing Science* **30**(2) 195–212.

Iyer, G., D. A. Soberman, J. M. Villas-Boas. 2005. The targeting of advertising. *Marketing Science* **24**(3) 461–476.

Lazarsfeld, P. F., R. K. Merton. 1954. Friendship as a social process: A substantiative and methodological analysis. M. Berger, ed., *Freedom and Control In Modern Society*. Van Nostrand, New York, 18–66.

Manchanda, Puneet, Ying Xie, Nara Youn. 2008. The role of targeted communication and contagion in product adoption. *Marketing Science* **27**(6) 961–976.

Manski, Charles F. 2007. *Identification for prediction and decision*. Harvard University Press, Cambridge.

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

Peres, Renana, Eitan Muller, Vijay Mahajan. 2010. Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing* **27** 91–106.

Provost, Foster, Brian Dalessandro, Rod Hook, Xiaohan Zhang, Alan Murray. 2009. Audience selection for on-line brand advertising: Privacy-friendly social network targeting. *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'09)*.

Shalizi, C.R., A.C. Thomas. 2010. Homophily and contagion are generically confounded in observational social network studies. ArXiv:1004.4704v1.

Trusov, Michael, Anand V. Bodpati, E. Bucklin, Randolph. 2010. Determining influential users in social networks. *Journal of Marketing Research* **47**(August) 643–658.

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