Homophily or Influence?
An Empirical Analysis of Purchase within a Social Network

Liye Ma, Ramayya Krishnan, and Alan Montgomery

Carnegie Mellon University
iLab, Heinz College and Tepper School of Business
5000 Forbes Ave
Pittsburgh, Pennsylvania

June 2010

Liye Ma (email: liyem@andrew.cmu.edu) is a doctoral candidate, and Alan Montgomery (email: alanmontgomery@cmu.edu) is an Associate Professor, both at the Carnegie Mellon University's Tepper School of Business, and Ramayya Krishnan (email: rk2x@cmu.edu) is Dean of Carnegie Mellon's Heinz College and William W. and Ruth F. Cooper Professor of Management Science and Information Systems. The authors are affiliated with the iLab at the Heinz College and would like to acknowledge support of the iLab at Heinz College and our anonymous project sponsor for providing the data used in this project. All opinions expressed in this paper are our own and do not reflect those of our project sponsor.

Homophily or Influence?
An Empirical Analysis of Purchase within a Social Network

Abstract

Consumers that are close to one another in a social network are known to have similar behaviors. The focus of this study is the extent to which such observed similarity is driven by homophily or social influence. Homophily refers to the similarity in product preferences between individuals who are connected. Social influence is the dependence of consumers' purchase decisions on their communication with others. We construct a hierarchical Bayesian model to study both the timing and choice of consumer purchases within a social network. Our model is estimated using a unique social network dataset obtained from a large Indian telecom operator for the purchase of caller ringer-back tones. We find strong social influence effects in both the purchase-timing and product-choice decisions of consumers. In the purchase-timing decision, we find that consumers are three times more likely to be influenced by network neighbors than by other people. In the product-choice decision we find a strong homophily effect. We show that ignoring either homophily or social influence will result in overestimated effects of the other factor. Furthermore, we show that detailed communication data is crucial for measuring influence effect, and influence effect can be either over- or underestimated when such data is not available. Finally, we conduct policy simulations on a variety of target marketing schemes to show that promotions targeted using network information is superior. For example, we find a 4-21% improvement on purchase probability, and an 11-35% improvement for promoting a specific product.

1. Introduction

Electronically enabled social networks allow consumers to communicate more efficiently amongst themselves. In addition electronic media offer the means to observe social networks. Knowledge about the structure of a social network provides marketers with a distinct opportunity to better understand their customers and improve their promotional decisions. Consider a firm that is selling products within a social network. It makes one sale to a customer and then makes a second sale to the customer's friend. The key question in our study is: did the second sale occur because one customer influenced the other or because these two customers have similar tastes, since after all, they are friends?

The answer to this question is critically important to the firm's marketing strategy, and is the focus of our study. We refer to the former case where a customer influences their friends to purchase as *social influence*. If social influence is responsible for the purchase then the firm may want to incentivize customers to promote the product to her friends using a referral bonus program. If it is the latter case of similarity in tastes then the firm could rely on the social network to help identify new potential customers by targeting the friends of the customer. Social scientists have long recognized that people with similar characteristics are more likely to form ties, an effect termed as *homophily*. Consequently, people who have close ties tend to have similar traits.

Social networks, social influence, and homophily have long been a topic of interest to sociologists, economists, and marketers. The advent of information technology has enabled the gathering and processing of large scale network data. The result is a growing number of studies on social networks in various fields such as economics (Jackson and Watts 2002), marketing (Hartmann et al. 2008), information systems (Hill et al. 2006), and machine learning (Zheng et al.

2008). This literature is concerned with both the formation of social networks (Ansari et al. 2008, Braun and Bonfrer 2009) and the implication of the network on consumer behavior (Nair et al. 2006, see Jackson 2003 for a comprehensive survey). Our research focuses on the implication that the network has on consumer behavior.

It is well known that human decision making is influenced through social contact with other people. Many terms have been used in literature to describe this widely recognized effect: social interactions (Hartmann 2008), peer effects (Nair et al. 2006), social contagion (Van den Bulte and Lilien 2001, Iyengar et al 2010), conformity (Bernheim 1994), imitation (Bass 1969, Choi et al 2008), and neighborhood effects (Bell and Song 2007). The different terms may have subtle differences but they all describe the dependence of one's decisions on those of others, an effect we term as social influence in our study at the general level. Recently, this influence effect has received attention from researchers in economics and marketing where it has been studied in the context of diffusion (Van den Bulte and Stremersch 2004) and word-of-mouth (Godes et al. 2005). Structural models have been used to try to uncover the detailed causal effects behind the observed influence. Hartmann (2008) models social interaction as the equilibrium outcome of a discrete choice coordination game, where individuals in groups take the decisions of other group members into account, and applies the model to a data set of a group of golfers. Nair et al. (2006) quantify the impact of social interactions and peer effects in the context of prescription choices by physicians, and demonstrate the significant impact of opinion leaders.

Research on the influence effect has paid much attention to uncovering "influentials" or "opinion leaders" in a group environment. The motivation is that certain individuals in a group of people may have a disproportionally large influence over other members in the group and this should be taken advantage of in target marketing. The significant impact of opinion leaders has been used to explain patterns of product diffusion, as discussed in Van den Bulte and Joshi (2007). Also, Nair et al. (2006) confirm the existence of opinion leaders in networks of physicians and show that the opinions of influentials have a great impact in the prescription decisions of other physicians. However, it is unclear whether the focus should be only on the opinion leader. For instance Watts and Dodds (2007) show that although influentials can trigger large-scale “cascades” in certain situations, in many cases change is simply driven by easily influenced individuals who sway other easily influenced individuals.

The phenomenon of homophily, which states that people with similar characteristics are likely to establish ties, has been recognized in the sociology literature for at least eighty years (Bott 1928). A rich literature exists in sociology which discusses various aspects of this effect (McPherson and Smith-Lovin 1987). A thorough survey of homophily can be found in McPherson et al. (2001). Although originally developed to explain the formation of networks — homophily clearly plays an important role in understanding human behavior within a network context. If people with like characteristics tend to behave similarly and also tend to establish ties, ceteris paribus, we should observe that people with ties tend to behave in the same way. Indeed, this effect has been used as the basis for improving marketing forecasts (Hill et al. 2006).

Both homophily and social interaction induce correlated behaviors of people who are closely connected. Separating these two effects empirically, however, is difficult. Manski (1993) proves that with a static model it is difficult or theoretically impossible to distinguish between endogenous and exogenous effects. Endogenous effects refer to an individual’s behavior that is influenced by that of others in the group. Exogenous effects occur when an individual’s behavior covaries with exogenous group characteristics, and this correlation means individuals in a group tend to have similar characteristics. As marketing research has been mainly focusing on various forms of influence effects, exogenous effects are usually controlled for by including many observed characteristics (e.g. Iyengar et al. 2010, Choi et al. 2008). Aral et al. (2009) investigate homophily through propensity-score matching based on observed characteristics, and provid bounds on influence effect. They show that ignoring homophily results in significant overestimation of influence effects. Observed characteristics, however, cannot address unobserved heterogeneity among consumers (Gonul and Srinivasan 1993) and the correlation of such unobserved preferences among friends. Our study addresses this issue by explicitly modeling the correlation of preference parameters. That is, we separate the "unobserved homophily" effect from influence effects.

Dynamic data may allow one to separate these effects but care still must be taken in decomposing these effects. Nair at al. (2006) uses an individual fixed effect as a control for effects other than peer-influence. Hartmann (2008) jointly estimates group-level correlation with other parameters to account for homophily on product taste. This approach is similar to the one that we propose in our study. However, Hartmann (2008) focuses on group coordination and does not elaborate on the role that homophily could play in purchase behavior. Furthermore, homophily suggests that network neighbors may be similar on most decision-relevant characteristics, instead of just base-level product taste. In our study, we model homophily on all decision-relevant characteristics: product taste, purchase interval, and intrinsic susceptibility to influence on purchase incidence and product choice. Finally, Hartmann (2008) studies coordinated consumption, while the product used in our study is individual consumption goods, and the influence effect is asymmetric and comes from communication.

Identifying and measuring the homophily and the social influence effect jointly is the focus of this study. Our ability to overcome previous problems is the result of a unique dataset and compatible statistical model. The dataset contains both detailed communication information among a large network of consumers and the purchase history of caller ring-back tones (CRBT). By their very nature caller ring-back tones, which are tunes heard when one member of the network calls another, lend themselves to network analysis. The number of exposures to a CRBT is accurately captured by the call data records. Identification in our problem is due to the dynamic nature of this data. Namely, the characteristics of consumers such as product preferences or susceptibility to influence which encodes the homophily effect remains stable over time, while the communications and associated exposures to CRBT which encodes the social influence effect varies through time. Finally, a person calls another person in the telecommunications network due to an intrinsic desire to communicate with them. The exposure to the CRBT happens as a side effect of that call. The communication does not happen due to a person wanting to listen to a CRBT.

Our model, which is discussed in section 2, is framed within the context of a hierarchical Bayesian model which simultaneously incorporates both the homophily and the social influence effect in the purchasing decision processes of consumers. It accounts for both the timing of purchase and product choice. Social influence is allowed on both purchase time and brand choice decisions, while the impact of homophily is measured for product taste, purchase interval, and susceptibility to influence parameters. We discuss these issues in more depth in section 3.

Our estimation results are given in section 4 and show strong social influence effects in both the purchase-timing and product-choice decisions of consumers. In the purchase-timing decision, we find that influence by network neighbors can increase a consumer’s purchase probability by almost five-fold, while the influence of people outside the group when calls are placed to them can also increase the probability by about 100%, i.e., doubling the purchase probability. In contrast, when we focus on the product-choice decision we find that consumers are more likely influenced by people outside their close networks. Influence of people outside a customer's close network on average increases the probability of choosing a specific product by about 20%, two times that of someone who is close to the customer, where the probability increases by about 9% only. We show that ignoring either homophily or influence effect resuls in overestimation of the other effect. Furthermore, we show that when communication is not explicitly accounted for but approximated using decision data, which is often the case in existing literature due to data availability issue, social influence effect can be either over or under-estimated, and homophily effect may also be affected.

The similarity in different characteristics may call for different policy responses, which we analyze in our policy simulation in section 5. Using this knowledge, we conduct policy simulations on a variety of target marketing schemes and find a 4-21% improvement on purchase probability, and an 11-35% improvement on promoting a specific product. We conclude the paper in section 6 with a discussion of our findings, limitations of our study, and future research directions.

2. Modeling Product Purchase within a Social Network

Our ability to separate homophily and social influence is driven by our unique dataset provided by a large Indian telecom company. The dataset consists of detailed phone call histories of all of the company's customers in a major Indian city over a three-month period. For each phone call record, the caller phone number, callee phone number, date and time of the call, and length of the conversation are recorded. There are over 3.7 million customers in the dataset and over 300 million phone calls over the covered period. This call data becomes the basis for inferences about the social network.

Our dataset also contains detailed transaction records about caller ring-back tones (CRBT) purchased by these customers. These tones are usually short snippets of musical songs. CRBT is a popular phone feature in a number of Asian countries including India, although it is not presently available in the American market. To understand its functioning considers what happens when customer *A* purchases a certain ring-back tone. Once another person *B* calls *A* then *B* will hear the ring-back tone instead of the usual ringing. Notice that only customer *A*'s callers hear this tone and not customer *A*. To use the CRBT feature, a customer must pay a monthly subscription fee, select the individual tone that he wants played when he is called. Each tone a customer purchases is valid for 90 days, but a customer can change the tone by purchasing a different one at any time. About seven-hundred fifty-thousand customers purchase ring-back tones during our time frame or roughly 20% of customers. The type of tones that are selected for purchase and when they are selected for purchase forms the basis for the consumer purchase decision in our problem.

There are two steps in our purchase decision: 1) when to buy and 2) what to buy. This dichotomy follows other that focus upon consumer purchases (Chintagunta 1993). We model the first step, the *when-to-buy* decision, using a model of inter-purchase timing with its corresponding hazard rate. The second step, the *what-to-buy* decision, is modeled using a discrete choice model. In the following sub-section we describe our model for this two-step decision process, and then discuss in the following subsection how to captures homophily and the social influence effects.

We assume that consumers belong to one of G groups. Each group consists of I consumers. The i-th consumer of g-th group is indexed as gi. Consumers belonging to the same group are assumed to have a strong social relationship. Human decision and behavior in general are subject to influence of the surrounding environment. A consumer in our model regularly communicates with people who are close to her. Both her purchase-timing decision and her product-choice decision are subject to influence arising from communication with other consumers. As noted this is particularly relevant in the case of CRBT since each communication to a consumer who has adopted a ring-back tone results in the caller being exposed to the tone.

2.1 Purchase Timing Model

Consumers may choose to purchase a product at any time, and we model the when-to-buy decision using its hazard rate (Gupta 1991). Time is discrete and indexed by t which ranges from 1 to T. We assume that the inter-purchase time of a consumer gi follows an Erlang-2 distribution with a time varying rate parameter $\lambda_{gi,t}$:

$$S_{gi}(t) = (1 + \lambda_{gi,t} \cdot t)\exp(-\lambda_{gi,t} \cdot t) \tag{1}$$

The rate is allowed to vary through time to reflect the chance that consumers become more (or less) likely to buy if they are exposed to others that use the product. A customer is exposed to CRBT either through others inside his social group or from those outside his social group. If a friend hears another friend's ring-back tone we would expect this to be more influential than hearing a tone for someone outside his social group. We denote $E_{gi,t,k}$ as the amount of exposure that consumer gi had at period t from either inside ($k = In$) or outside ($k = Out$) his group. (We discuss the construction of the exposure variable further in section

3.5.) We propose the following model of the purchase rate parameter as a function of cumulative product exposure from both inside and outside the group:

$$\lambda_{gi,t} = \lambda_{gi} \exp(\gamma_{gi,In} E_{gi,t,In} + \gamma_{gi,Out} E_{gi,t,Out}) \qquad (2)$$

In equation (2), $\gamma_{gi,k}$ can be considered as a *susceptibility* parameter, which indicates the extent to which the consumer is subject to external influence in making her decisions. A large magnitude means the consumer in general values the input of others, while a small magnitude indicates that the consumer is quite opinionated makes her own decisions. A positive sign indicates the consumer positively accounts for external influence, while a negative sign shows that the consumer handles external influence negatively.

The rate parameter must be positive, therefore we assume it follows a multivariate log-normal distribution:

$$\begin{bmatrix} \ln(\lambda_{g1}) \\ \ln(\lambda_{g2}) \\ \vdots \\ \ln(\lambda_{gI}) \end{bmatrix} \sim MVN\left(\begin{bmatrix} \ln(\bar{\lambda}) \\ \ln(\bar{\lambda}) \\ \vdots \\ \ln(\bar{\lambda}) \end{bmatrix}, \sigma_\lambda^2 \begin{bmatrix} 1 & r_\lambda & \cdots & r_\lambda \\ r_\lambda & 1 & \cdots & r_\lambda \\ \vdots & \vdots & \ddots & \vdots \\ r_\lambda & r_\lambda & \cdots & 1 \end{bmatrix} \right) \qquad (3)$$

$\bar{\lambda}$ is the population level average base rate, and σ_λ^2 measures the dispersion of this base rate parameter in the population. We choose a lognormal hyper-prior for $\bar{\lambda}$, and an inverse-Gamma hyper-prior for σ_λ^2, both of which are conjugate priors. Both hyper-priors are chosen to be diffuse, as we have little knowledge of them ex-ante. The correlation parameter r_λ must be within the interval $[\underline{r},1]$, where $\underline{r}$ is the smallest number to make the correlation matrix positive-definite. We choose a uniform hyper-prior for r_λ. When homophily exists we expect the values of consumers in the same group to be positively correlated, or $r_\lambda > 0$ is a sign of homophily.

The social influence parameters of purchase timing of group g, $\gamma_{g,k} = (\gamma_{g1,k},...,\gamma_{gI,k})^T$, are assumed to follow a multivariate normal distribution:

$$\begin{bmatrix} \gamma_{g1,k} \\ \gamma_{g2,k} \\ \vdots \\ \gamma_{gI,k} \end{bmatrix} \sim MVN(\begin{bmatrix} \bar{\gamma}_k \\ \bar{\gamma}_k \\ \vdots \\ \bar{\gamma}_k \end{bmatrix}, \sigma^2_{\gamma_k} \begin{bmatrix} 1 & r_{\gamma_k} & \cdots & r_{\gamma_k} \\ r_{\gamma_k} & 1 & \cdots & r_{\gamma_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\gamma_k} & r_{\gamma_k} & \cdots & 1 \end{bmatrix} \tag{4}$$

The specification for the parameters $\bar{\gamma}_k$, $\sigma^2_{\gamma_k}$, and r_{γ_k} is similar to that of their counterparts for the base rate parameters. Again we expect that $r_{\gamma_k} > 0$, which would be evidence of homophily.

2.2 Product Choice Model

The what-to-buy decision step is modeled using a discrete multinomial choice model. There are J products and the $K \times 1$ vector of product characteristics associated with product j is $\mathbf{X}_j$. Besides the product characteristics we also believe that the amount of exposure that the consumer has received at time *t* could influence his choice, and as in equation (2) we allow for differential effects from inside versus outside the group. We denote $E_{gi,j,t,k}$ as the amount of cumulative exposure that consumer gi has received for product j at period t from either inside ($k = 1 = In$) or outside ($k = 2 = Out$) his group. The utility of consumer gi from purchasing product j at time period t is the sum of the product characteristics, cumulative exposure, and a random error:

$$U_{gi,j,t} = \mathbf{X}_j{}'\boldsymbol{\beta}_{gi} + \rho_{gi,In}E_{gi,j,t,In} + \rho_{gi,Out}E_{gi,j,t,Out} + \varepsilon_{gi,j,t} \tag{5}$$

Let $\boldsymbol{\beta}_{gi}$ be the $K \times 1$ valuation coefficient vector for consumer gi. Similar to $\gamma_{gi,k}$ in the purchasing timing equation, the parameter $\rho_{gi,k}$ in equation (5) indicates how much a consumer's perceived utility of a product is influenced through communication with others. The interpretation of the sign and magnitude of $\rho_{gi,k}$ is the same as that of $\gamma_{gi,k}$.

Assuming $\varepsilon_{gi,j,t}$ follows the type-I extreme-value distribution, the product-choice probability then follows that of a standard multinomial-logit model:

$$P(gi \text{ choose } j \text{ at period } t) = \frac{\exp\left\{X_j^T \beta_{gi} + \rho_{gi,In} E_{gi,j,t,In} + \rho_{gi,Out} E_{gi,j,t,Out}\right\}}{\sum_{l=1}^{J} \exp\left\{X_l^T \beta_{gi} + \rho_{gi,In} E_{gi,j,t,In} + \rho_{gi,Out} E_{gi,j,t,Out}\right\}} \tag{6}$$

We introduce a hierarchical specification for the $\beta_{g,k}$ parameter across the groups to allow for heterogeneity:

$$\begin{bmatrix} \beta_{g1,k} \\ \beta_{g2,k} \\ \vdots \\ \beta_{gI,k} \end{bmatrix} \sim MVN\left(\begin{bmatrix} \bar{\beta}_k \\ \bar{\beta}_k \\ \vdots \\ \bar{\beta}_k \end{bmatrix}, \sigma^2_{\beta_k} \begin{bmatrix} 1 & r_{\beta_k} & \cdots & r_{\beta_k} \\ r_{\beta_k} & 1 & \cdots & r_{\beta_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\beta_k} & r_{\beta_k} & \cdots & 1 \end{bmatrix} \right) \tag{7}$$

Similarly the social influence coefficient of group gi for either within or outside the group also follows a hierarchical specification:

$$\begin{bmatrix} \rho_{g1,k} \\ \rho_{g2,k} \\ \vdots \\ \rho_{gI,k} \end{bmatrix} \sim N\left(\begin{bmatrix} \bar{\rho}_k \\ \bar{\rho}_k \\ \vdots \\ \bar{\rho}_k \end{bmatrix}, \sigma^2_{\rho,k} \begin{bmatrix} 1 & r_{\rho,k} & \cdots & r_{\rho,k} \\ r_{\rho,k} & 1 & \cdots & r_{\rho,k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\rho,k} & r_{\rho,k} & \cdots & 1 \end{bmatrix} \right) \tag{8}$$

We choose diffuse normal conjugate hyper-priors for $\bar{\beta}_k$ and $\bar{\rho}_k$, and diffuse inverse-Gamma conjugate hyper-priors for $\sigma^2_{\beta_k}$ and $\sigma^2_{\rho_k}$. Finally, we choose a uniform hyper-prior for r_{β_k} and r_{ρ_k}. This specification is similar to those of the hyper-parameters for λ and γ.

3. Discussion concerning our data and model

Getting quality data and properly leveraging it has been a major challenge in social network research. That is not to imply that our data or model is perfect, since we have made many simplifying assumptions in creating our model and organizing our data. But we believe our dataset offers many advantages in addressing several common concerns related to homophily and social influence within a social network. In this section we explain these challenges and how we attempted to organize our data to mitigate potential problems.

The key to our identification strategy is to take advantage of the static nature of the homophily effect versus the dynamic nature of social influence effect. While the characteristics of consumers such as product valuation remain stable overtime, the consumers are exposed to different levels of influence over time. Therefore, the effects of social influence and homophily can be separated. In our model and data the exposure variables, which capture how peer selection of ring-back tones change over time while other effects remain constant. Adequate time series variation of exposure allows the identification and estimation of the parameters.

The usual identification restrictions apply for our multinomial logit model, namely that the latent utilities are identified up to a constant. Therefore, the product characteristic of one of the products will be normalized to zero. If product fixed effect is included in the product characteristics, then the corresponding parameter will not be identified and has to be normalized to zero as well. Another parameter that cannot be identified is the price coefficient. This is because all ring-back tones are sold at the same price, so it is impossible to identify price consideration based on consumer's product choices. The tradeoff between the purchase price and usage value will be encoded in the intrinsic purchase frequency parameter.

3.1 Defining Groups within our Social Network

In our model a group consists of people who have a close relationship with one another. The homophily effect within a group should remain stable over a short period of time, such as the three months that we observe in our data, therefore it is important that we identify stable relationships. We believe that people who call each other frequently are likely to have a close relationship. To ensure that we capture true relationships rather than sporadic phone calls, we consider two customers as belonging to a group only if they made at least five phone calls in the first month of the three-month period to one another. The choice of five phone calls as the threshold is a subjective one on our part, but we thought it was a commonsense one to delineate close contacts[1]. If the threshold is set to too low a value (e.g., one or two) then the network is contaminated with many contacts that are not part of the caller's social circle. If we increase the threshold (e.g., ten or more) then we substantially reduce the number of groups that are formed. Our selection of five phone calls was meant to achieve a balance between these two factors. Furthermore, it is only necessary to uncover only the *existence* of connections and not their *strength*, since we can infer strength through our model's social influence effects.

This produces an undirected network or graph, where each node corresponds to a customer, and there is an edge between two nodes if the two corresponding customers have close relationship. To identify groups in this network, we then run a clique-searching algorithm to find all four-cliques in the graph[2]. A four-clique is a sub-graph with four nodes, where every node is connected to every other node in the sub-graph. Each four-clique then corresponds to a group of four customers, where each customer has a close relationship with every other customer in the

---

[1] We repeated this analysis with other thresholds and other clique sizes using simplified versions of our model and found that the results are similar. This earlier form conditioned upon the smoothing parameters and exposures with the number of tones instead of the changes in tones.

[2] Our algorithm for enumerating n-cliques randomizes the ordering of the nodes and then searches for a clique of size n and then removes the nodes, so that these nodes will not be included in the search again.

group. We identify a total of 1,654 groups in our dataset. In order to shorten our estimation time we randomly select 300 of these groups.

The choice of four in defining our groups is another subjective decision on our part. Given that a clique implies everyone in the group is connected to everyone else, a group of larger size implies stronger connections between members within the groups but fewer groups of larger size. If we choose a smaller size (e.g., two) then our group effect would represent only dyadic relationships and may not capture more general social processes. Our desire is to keep the modeling framework simple by conditioning upon the network structure. The choice of phone call thresholds and clique size could be nested within our model, but at the expense of a more complex model. Note that the four-cliques may be embedded in cliques of larger size. That is, a group of four customers where everyone is connected to everyone else may be a subset of a larger group where all are pairwise connected. However, this is not an issue for the homophily effect, the condition of which is the existence of connections but it does not need to be exclusive connections. In another word, if the four customers have similar preference because they are friends, such similarity will not disappear simply because they all are friends of yet another person. This may be an issue if the strength of ties and differential degrees of similarities are accounted for, which we leave for future research. The embedding issue is a concern on the influence effect, as out-group communications can still come from a connected person, carrying influence effect of in-group magnitude. However, in our model this will bias the estimate of out-group influence towards the in-group influence, making it harder to find difference between the two. Therefore, this bias makes it harder for us to draw conclusions comparing the influence effect, thus strengthening the findings we report.

A possible concern with our group definition is the potential for endogenous group formation. If a group is formed with the objective of conducting a certain activity, then it is hard to draw causal inference based on observations of that activity and other related activities performed by the group. However, we believe it would be extremely unlikely that people would call one other and form a social tie just so that they can hear each other's ring-back tones. Therefore, we argue that it is unlikely that endogeneity in group formation is a concern. Certainly correlation in preferences between friends may exist, as people who form a group may have similar taste to tones, but this is the homophily effect that we wish to uncover and distinct from endogenous group formation. Although endogenous group formation is not a concern in our problem, we think that this is an interesting problem for future research.

3.2 Product Choice

Ideally, each caller ring-back tone could be treated as an alternative in our choice model. However, there are more than eleven thousand different tones that have been purchased by customers, most of which were purchased by only a few people. This sparse dataset makes it difficult to estimate the parameters and to interpret the results at the individual tone level. Therefore we have chosen to categorize all the tones to ten different genres. These genres were gathered from the telecom's website which provided our dataset. We were able to categorize seven thousand tones. Upon further analysis we found that only three of the ten genres had more than 5% of market share. Therefore, we chose to combine the tones from the small genres and uncategorized tones into an "Other" category. This results in four "products" or categories or genres of tones. For simplicity we refer to the choice of a tone within a music genre or category as a product. Table 1 lists these product IDs and their respective market share. Unfortunately we possess little information about the tones and music genres, so we are only able to define product dummies in the product characteristic vector.

[Insert Table 1 About Here]

3.3 Evidence of Similarity

Behavior similarity among people who are connected is well established in literature. Nonetheless, we would like to verify that it exists in our dataset, in order to assess the fit of the dataset for our research question and modeling framework. For this purpose, we calculated the unconditional probabilities that a person in one of our extracted groups purchases each category of tones, and those probabilities conditional on some group members purchasing that category. If consumers do make similar purchases, then we expect the conditional probability to be higher than the unconditional ones.

[Insert Table 2 About Here]

Table 2 reports these probabilities. As is shown in the table, for all four categories, the probability of purchase conditional on a group member purchasing is significantly higher than the overall unconditional probability. This is clear evidence that in-group similarity exists in the dataset, making it potentially a good fit to apply our model.

3.4 Understanding Social Influence with Caller Ring-Back Tones

Identifying and quantifying social influence is usually an extremely challenging task, since detailed communication history among people is rarely available to researchers. Even in the case where it is known who contacted whom at what time, the content of the communication is still generally unavailable. Our dataset is different because we observe each individual phone call (or communication within the network) and its timestamp. However, we do not have any information about the content of the conversations.

The influence a customer imposes on a caller is conveniently encoded in their communication records within the telecom network. Whenever a person places a call to a customer with a certain ring-back tone, we can infer that the caller has heard the tone from the callee. This caller is automatically exposed to two things. First, the caller immediately perceives that the person is using the CRBT service. Second, this caller is exposed to the product or more precisely the customer's chosen tone. The social influence argument suggests that both the purchase-timing decision and product-choice decision may be influenced through exposures resulting from phone calls. We thus quantify this external influence based on the phone calls made by the customer. As both the phone call records and the ring-back tone purchase records are time-stamped, we can infer how many times a customer is exposed to our products within a certain period. As stated earlier, in our study we treat social influence as the dependence of one's decision on those of others at the general level, for which such communication data is sufficient. At the detailed level of behavioral factors, we note that the influence in our study is of passive nature – the influence customer A "exerts" to customer B is not done through explicit persuasion from A to B, but through passive observation by B of A. This passive effect may arise out of either observational learning or imitation among other factors, distinguishing which is left for future study.

Another problem is the potential for exogenous influence or non-social influence. Two people who are connected may purchase the same product not because they have similar tastes or because one is influenced by the other, but instead because they are subject to the same exogenous shock. For example they may be exposed to a common promotional activity (e.g., radio airplay, concert, media report, etc.) and these network neighbors could be simultaneously influenced. We argue that such exogenous shocks are unlikely to persist given that our data is observed over several months and the group identities are only known by the phone company which does not presently use this network information for marketing purposes. Therefore, we believe that any group-specific exogenous shocks, even if they exist, should be transitory and well captured by our error process.

3.5 Quantifying Exposure

From phone call records we can infer the exposure received by a consumer from sources inside as well as outside her group. The quantification of such exposure is different for the purchase-timing and product-choice decisions. For purchase-timing, the decision a consumer makes at each time period is *whether-to-buy*. Consequently, it is appropriate to use the information on *tone purchase* by others as an exposure. Such an exposure event occurs when consumer *A* calls consumer *B* and is exposed to a new ring-back tone. The new tone that *A* is exposed to can arise due to two reasons. In the first case, it could be that *B* is a first time adopter of ring-back tone. In this case, previous calls from *A* to *B* would have had the default ring tone (i.e., the traditional ringing sound on the phone). The second case is that B has purchased a new tone to replace the previously heard one. In both these cases we consider consumer *A* to have been exposed to a tone purchase. For product-choice, the decision a consumer makes is *what-to-buy*. Thus it is appropriate to use the information on tone choice by others as exposure. Inferring such an exposure is straightforward: if consumer *A* calls consumer *B* and is exposed to a tone in a certain genre, then we consider consumer *A* to be exposed to that genre. Indentifying instances of exposure this way, we then count the total number of such exposures relevant to purchase incidence and product choice, respectively, to arrive at a "raw" exposure measure.

Furthermore, it is possible that a consumer is influenced by others even though she does not act on it immediately. For example, a customer makes a phone call on a given day and is exposed to a tone. The customer's propensity to purchase in the genre is increased, and then buys a tone in the same genre. However, the consumer may purchase the tone several days later instead of on the same day. To better capture these types of delays in purchase we exponentially smooth the exposure over time:

$$E_{gi,t,k} = \kappa_{gi}^{pi} E_{gi,t-1,k} + (1-\kappa_{gi}^{pi})\tilde{E}_{gi,t,k} \tag{9}$$

$$E_{gi,j,t,k} = \kappa_{gi}^{pc} E_{gi,j,t-1,k} + (1-\kappa_{gi}^{pc})\tilde{E}_{gi,j,t,k} \tag{10}$$

Where $E_{gi,t,k}$ and $E_{gi,j,t,k}$ are the exposure measures mentioned in section 2, notice that these measures are smoothed. The raw or actual exposures inferred from the phone call records are $\tilde{E}_{gi,t,k}$ and $\tilde{E}_{gi,j,t,k}$. κ_{gi}^{pi} and κ_{gi}^{pc} are the smoothing parameters of the consumer for purchase-incidence and product-choice exposures, respectively. We jointly estimate these two parameters together with the other parameters of interest, and use a logit-normal prior to ensure the parameters lies between 0 and 1. We allow for heterogeneity in these parameters using the same hierarchical structure as we did with the other parameters.

4. Empirical Results

There are 91 days in the entire dataset. We use the first 10 days to initialize the exponentially smoothed exposures, then the next 60 days for estimation, and the final 21 days as a holdout sample for predictive evaluation. We use a Markov Chain Monte Carlo (MCMC) method to draw parameters from their posterior distributions. The likelihood function discussed in Appendix A shows that the inter-purchase timing component and the product choice component are independent and therefore can be estimated separately. The details of the MCMC draws and the corresponding likelihood functions can be found in the Appendix A. For the estimation, we took 10,000 MCMC draws, with the first 5,000 discarded as burn-in draws and the remaining 5,000 used for evaluation.

4.1 Purchase Timing

The posterior mean, standard deviation, and 95% confidence interval of parameters for the purchase timing model are reported in Table 3. As is shown in the table, the population level mean purchase parameter is 0.0323, which suggests a mean purchase frequency of once every 61.9 days. The in-group influence parameter is 1.86, while the out-group influence parameter is 0.724. Both are positive and statistically significant. This shows that a strong influence effect exists in purchase-timing decisions. This estimate shows exposure to a tone change by someone outside the group can increase the purchase probability by about 105%, which is quite a sizable increase in purchase probability. But even more substantial is the effect of an exposure to tone purchase by someone inside the group which increases the purchase probability by more than five-fold (542%), making the purchase more than six times as likely. The higher magnitude of in-group influence than out-group influence is also reasonable, suggesting that customers are much more susceptible to people close to them than to others.

[Insert Table 3 About Here]

The in-group correlation on the purchase rate parameter is 0.0124, which is statistically indistinguishable from 0. This suggests that customers in the same group are not more likely to have similar intrinsic purchase frequencies. The in-group correlation on the in-group influence parameter is 0.0993, while that on the out-group influence parameter is 0.192. This shows that group members have somewhat similar levels of susceptibility to in-group as well as out-group influence. However, neither parameter is statistically significant from zero, as the 2.5% posterior quantile is negative for both parameters. The parameter estimates thus show that homophily effect is not evident in the intrinsic purchase rate. Although the homophily effect does seem to exist on the susceptibility to influence, the evidence is inconclusive. These results suggest that any observed purchase-timing similarity among customers who belong to the same group is likely the result of social influence, instead of the intrinsic similarity in their purchase timing.

4.2 Product Choice

The posterior mean, standard deviation, and 95% posterior interval of parameters for the product choice model are reported in Table 4. The mean valuation parameters for the first three music genres are -1.40, 0.236, and -1.35, respectively. These values are broadly consistent with the market shares of these music genres. The in-group influence parameter is 0.0857, which suggests that being exposed to a tone when calling someone in the group has slightly positive effect on choosing a tone in the same category (the corresponding choice probability increases by about 9%). This effect, however, is not statistically significant, as the 2.5% posterior quantile is negative. The out-group influence parameter is 0.181, meaning that exposure to a tone from someone outside the group has fairly strong positive effect on choosing the same category—the choice probability increased by about 20%. This effect is statistically significant.

[Insert Table 4 About Here]

The result may be surprising that the out-group influence is higher and more strongly evident than the in-group influence. One explanation is that influence on product choice may have two competing effects. On one hand, a person's product choice may be positively influenced by their friends because they trust their friend's selection. On the other hand, consumers may wish to be distinct and exhibit some variety seeking behavior. Specifically upon knowing their friend's choice they may not want to choose the same product and be perceived as merely imitating her friend. Unfortunately these two effects cannot be separated in our model and may cancel each other out. This may explain why our in-group influence effect is not statistically different than zero.

The variation of the parameter estimates may also lend support to our explanation. If some consumers want to imitate while others seek variety, we should expect a wide dispersion of the in-group influence parameter, and thus a high estimate of the variance. Indeed, the result in Table 4 shows that the estimated variance of the in-group influence parameter, 0.221, is more than four times larger than that of the out-group influence parameter, 0.053. Again this provides evidence in favor of our explanation.

The in-group correlations for the three product taste parameters are 0.773, 0.317, and 0.523, respectively. All are positive and statistically significant, which provides strong evidence that significant similarity exists on product tastes of customers who are close to one another. This confirms the expectation that the homophily effect exists and influences product tastes. Furthermore, the in-group correlations for the in-group influence and out-group influence parameters are 0.598 and 0.739, respectively. Both are statistically significant. This suggests similarity also exists within group members on their susceptibility to influence. In other words if a consumer is likely influenced by others in his product choices, then his friend is also likely influenced by others. This is further evidence of homophily in the product choice decision.

In summary, we find strong influence effects in both purchase-timing and product-choice decisions, and a strong homophily effect in product-choice decision as well. On purchase-timing effects we find that the in-group influence is much higher than the out-group influence, while the reverse is true on product-choice. The homophily effect may also exist on purchase-timing effects, although the evidence is inconclusive.

4.3 Model Comparison

We also estimate two special cases of the model we proposed in section 2. The first model assumes that there is no homophily effect in the intrinsic purchase frequency, product tastes and susceptibility to influence, and we refer to this model as "influence only". In this model consumers may still influence one another through communications, but the in-group correlation parameters are assumed to be zero. The second model which we call "homophily-only", assumes that there is no influence effect through communications. In this model the homophily effect may still exist on the intrinsic purchase frequency and product tastes, but the influence parameters are all assumed to be zero. Our proposed model which was presented in section 2 and includes both homophily and influence effects is called the "full" model in the comparison given in this section.

[Insert Table 5-8 About Here]

The alternative models are estimated using an MCMC algorithm similar to the one used for our full model. The results are reported in Tables 5-8, respectively. For the influence-only model, the estimated values of the purchase timing parameters – intrinsic purchase frequency and susceptibility to influence – are very close to those in the full model. This is as expected since the homophily parameters are estimated to be close to zero in the full model. The estimated values of the in-group and out-group influence parameters of the influence-only model are both higher than their counterparts in the full model (0.270/0.255 vs. 0.0857/0.181), which shows that when homophily effect is overlooked, a model will overestimate the effect of influence one exerts on another. This is exactly as expected and highlights the importance of simultaneously quantifying these two effects.

For the homophily-only model we find that the estimates of the intrinsic purchase frequency parameter to be 0.035. This estimate is slightly higher than the estimate in the full model (0.0323) and the influence-only model (0.0325). This shows that when influence effect is not accounted for the model overestimates the intrinsic purchase frequency of consumers, because influence-induced purchase is now considered to be spontaneous. Furthermore, most of the homophily parameters are estimated to be higher than their counterparts in the full-model (0.109 versus 0.0124 for intrinsic purchase frequency and 0.689/0.913/0.625 versus 0.773/0.317/0.523 for product tastes). This shows that when the peer effect that captures influences of one peer on another in purchase decisions is ignored then the model overestimates the similarity among consumers who are connected. Again, this highlights the importance of separating out the two effects of homophily and social influence.

[Insert Table 9 About Here]

The in-sample and out-of-sample likelihood of the three models is reported in Table 9. As expected the full model achieves higher likelihood in the estimation period than either of the homophily-only and the influence-only models. More importantly we notice that the full model also has a higher likelihood in the holdout period, which shows that the additional parameters are improving our forecasting performance. When applying a more formal statistical approach using Bayes factors, we find that the full model is strongly favored over both the influence-only model and the homophily-only model.

4.4 The Importance of Communication Data

Communication is necessary for influence. For a customer A's action to influence customer B's decision, not only does A need to take the action, the knowledge of action must be conveyed to B as well. Data on communication is thus crucial for accurate assessment of influence effect.

Detailed communication data, however, is rare in datasets available to researchers. Consequently, existing research usually accounts for influence by making one's decision directly dependent on the decision of others (e.g. Nair et al. 2006, Bell and Song 2007, Iyengar et al 2010), with an implicit assumption that one's action perfectly observed by others[3].

With the detailed communication data, we are able to evaluate its importance in measuring influence effect. To do so, we estimated alternative "no-communication" models. In a no-communication model, we ignore the phone calls which expose one to another's tone, and simply have one's decision enter into others' decision equations directly, an approach similar to existing studies. For purchase timing decision, whenever a consumer purchases a tone, we consider all others in her group as being exposed. For product choice decision, we evaluate two alternative models. In the first model, we consider there is one count of exposure every day to others in the group as long as a consumer possesses a tone, while in the second model we consider there is exposure only when the consumer newly purchases a tone.

[Insert Table 10 & 11 About Here]

The estimation result is reported in Tables 10 and 11 in comparison with the result for the full model[4]. As shown in Table 10, the influence parameter for purchase timing is much smaller in the no-communication model than in the full model: 0.675 vs 1.86. This is understandable: without communication data, a song change is registered as exposure whether or not it is communicated to others. This results in overstated amount of exposure and lowers the per communication influence effect. Meanwhile, the timing of exposure is not accurate. For example if customer A changes a tone on day 1, and customer B calls A on day 3, the true exposure happens on day 3, but without

[3] This may be a reasonable assumption if the data is at aggregate level due to law of large numbers, or if each time period is sufficiently long so that with high confidence one's decision has been conveyed to others in the time period. However, it is still an approximation, the seriousness of which should be evaluated on a case-by-case basis.
[4] When communication data is not used, out-group exposure cannot be quantified, and is thus left out of the equations for the no-communication model.

communication data, it is counted on day 1 in the no-communication model. This further dampens the influence effect. Ignoring communication data, therefore, leads to significantly underestimated influence effect for purchasing timing decision.

The estimated influence effect for the no-communication model, presented in Table 11, is larger than that in the full model. This again is because of miscounting of exposure, either an exposure is counted once every day a customer possesses a tone (model I) or once only when the customer purchases a tone (model II), whereas in reality there may be multiple phone calls on one day and none on another. The total amount of exposure can thus be either over- or understated (this is in contrast to purchase timing decision where the exposure amount can only be overstated, as each song change is observed at most once). With mis-timed exposure which could be either over- or understated, the influence estimate could be biased either up- or downward. Also shown in Table 11 is that the homophily effect is overestimated in both model I and II, compared with the full model. This may be because the influence effect is mistakenly loaded onto the intrinsic product taste parameter, when communication is not accounted for. Both estimates clearly demonstrate the crucial importance of detailed communication data on accurately measuring influence effects.

5. Policy Simulation

Up to this point we have focused on measuring the impact of social influence and homophily on purchase timing and choice. However, as we pointed out in the introduction if managers have access to the network then they can use the knowledge of its structure to improve their decision making. Understanding the relative importance of the various factors in consumers' purchase decision enables us to evaluate the effectiveness of target-marketing policies. Unfortunately, we do not observe any targeted marketing campaigns in our dataset that would allow us to quantify the importance of the network on targeted promotions. Additionally, we do not observe any price variation in CRBT. Hence the difficulty of our task is compounded once again by an inability to estimate price response. Therefore, we propose a series of simulations that revolve around a hypothetical coupon drop to understand how consumer targeting can be improved over those promotions which fail to exploit network structure. This coupon drop is meant to provide us a marketing vehicle that hypothetically varies price or purchase intention of the coupon's recipient. The key idea that underlies our simulations is that network information can make targeting more effective.

5.1 Targeting Known Customers

The first policy simulation evaluates the effectiveness of targeting promotions to existing customers. The objective of the firm is to increase the purchase probability of its existing customer base over a certain period of time and promote the sales of a specific category[5]. To do so, it is assumed that the firm has a number of coupons to distribute to a selected subset of the customers. Each coupon is assumed to increase a customer's intrinsic purchase rate by a specific percentage point. At the same time, it will increase a customer's base utility of one specific category of tones by a certain amount[6].

We conducted the policy simulation based on the estimates of the individual-level parameters when we discussed in sections 4.1 and 4.2. We evaluate the effect of distributing 100 coupons to selected customers in the whole group to optimize the objective. The choice of 100 is to ensure decent coverage of the 1200 consumers in the sample emanating from our random sample of 300 four-cliques, while at the same time keep it selective enough so efficiency matters.

---

[5] Typically coupons are used to increase sales. The firm can also use coupons to increase the relative share of a specific brand or products. If different brands or products have different gross margins, for example, promoting a high-margin product is profit enhancing even if the overall sales remain the same.

[6] Note that this is similar to assuming a certain price coefficient and a coupon of specific dollar value. But since we cannot estimate price coefficient given the uniform price, we have to evaluate the coupon in this "reduced form" manner.

Each coupon is assumed to improve the intrinsic purchase rate by $c_p\%$, and increase the base utility of the first category of ring-back tones by c_c (we treat the first category as the target category of promotion). To utilize the influence effect, we assume that each person is influenced by a friend in the beginning of the promotion and that the firm observes this interaction. We measure the change in purchase probability and choice probability in the week that follows the promotion – we choose a week as the period for performance measurement as the influence effect will largely diminish after a week, according to our estimated result.

[Insert Table 12 About Here]

[Insert Figure 1 & 2 About Here]

We evaluate the purchase and choice probability of each person either with or without the coupon, and distribute the coupons to the consumers whose purchase and choice probability increase the most. In this simulation, we consider the change in probability only for the individual customers, while the effect on other group members is considered in the next simulation. A series of values for c_p and c_c are evaluated. The resulting average enhancement in purchase probability and choice probability is plotted in Figure 1 and Figure 2. Using $c_p\% = 50\%$ and $c_c = 0.5$ as an example, the result of which is reported in Table 12. Targeting based on parameter estimates from the full model increases purchase probability by 22.18%, as compared with 20.45% of the homophily only model and 21.36% of the influence only model. This represents an 8.5% and 3.8% improvement, respectively, arising from recognizing both homophily and influence effects as compared with recognizing only one of them. The improvement in the probability of choosing the first category is 11.99% for the full model, and 10.73% and 9.12% for the homophily-only and influence-only model, respectively. This represents an 11.7% and 31.5% improvement using the full model over the homophily-only and influence-only model, respectively. On average, the full model performs 10.47% better than the homophily-only model and 4.08% better than the influence-only model on enhancing purchase probability, and it performs 11.65% and 35.37% better than the two models on enhancing product choice probability. These results are show that consumer targeting can be improved by recognizing both homophily and influence effects.

5.2 The Multiplier Effect

In our second policy simulation we again are interested in evaluating the effectiveness of target promotion to existing customers. Instead of evaluating the improvement that comes directly from the targeted customers, we are interested in observing the secondary improvement that arises from their communication with their friends, i.e. a "multiplier effect". The objective of the firm is to increase the purchase probability in a certain time period.

We again evaluate the effect of distributing 100 coupons. We look at the increase in purchase probability by other members of the group that customers who are targeted belong to conditional on the purchase by the targeted customer, and distribute the coupons to the customers with the highest group effects. The results are reported in Table 13.

[Insert Table 13 About Here]

As the table shows, the improvement in purchase probability is 22.40% for the true model, and 22.15% for the influence only model. This represents a 1.1% improvement in performance from the multiplier effect. Note that the homophily only model is not evaluated for this simulation, since the model by design assumes away peer influence.

5.3 Targeting New Customers

In the final policy simulation, we evaluate the effectiveness of target promotion to new customers. These customers are "new" in the sense that their purchase history is not known to the firm. However, the firm observes the communication between these existing customers and their friends (or the "new" customers). The objective of the firm is to increase the purchase probability in a certain time period by targeting these new customers.

[Insert Table 14 About Here]

We again distribute 100 coupons, each of which is assumed to increase the intrinsic purchase rate by $c_p\% = 50\%$ for a period of seven days. Similar to the first simulation, we pick the customers who are expected to have the highest change in purchase probabilities. The result is reported in Table 14. The average increase in purchase probability for the full model is 10.86%, while those for the homophily-only and influence-only models are 9.92% and 8.98%, respectively.

We first note that the overall effectiveness of the coupon is lower than when applied to known existing customers–the purchase probability improvement here is around 10%, compared with a better than 20% improvement when applied to existing customers, as shown in the first policy simulation. This is as expected, since the firm has much less information about these new customers than the existing ones, and thus cannot infer their preferences with as much precision. However, when targeting these new customers we find that the full model (which accounts for both the homophily and influence effect) also has the best relative performance, with 9.5% and 20.9% better performance than the homophily-only model and the influence-only model, respectively. This suggests that the biggest potential of a model with both homophily and influence effects is to target potential customers with which little information is known directly, but their social networks allow the firm to make inferences based upon their peers who are known to the firm. Potentially the social network provides a powerful device for targeting customers since it is self-organizing and helps the firm sort through its targets most cost effectively than targeting everyone.

6. Conclusion

Our study contributes to the literature by simultaneously quantifying both the homophily and the social influence effect in product purchase decisions by consumers in a social network context, using a unique, large-scale, real word dataset. Our study is among the first to quantify the impact of various homophily and influence factors jointly in the decision process of consumers. This is an important finding, which suggests that although customers are subject to influence by their friends, but this may be moderated by a desire not to be perceived as imitating. Furthermore, we find a strong homophily effect in the product-choice decision, where customers who are close by tend to have similar product tastes as well as similar susceptibility to influence. There is evidence of homophily in the purchase-timing decision as well, although this is not conclusive.

Social network researchers have long recognized the importance of both homophily and social influence. Both the homophily effect and the social influence effect can explain the phenomenon that consumers who are close by tend to make similar purchase decisions. However, they prescribe different target schemes: if homophily effect is the reason of the similarity, then the firm should target an existing customer’s friends directly, knowing that they likely have similar product tastes as the existing customer. But if social influence is the reason, then the firm should target the existing customers, relying on them to promote to their peers, or at least time the direct targeting to their peers so that it is enhanced by timely influence effect from the existing customers. Separating these two effects is crucial for effective marketing strategies, but it is also challenging, as the existing stream of work reveals the near observational equivalence of these effects.

In our study, we estimate a purchase-timing and product choice model within the context of a hierarchical Bayesian model. Our study is made possible by our unusual dataset which contains both communication and product purchase information over time. Applying our model to this dataset, we find strong social influence effect in both purchase-timing and product-choice decisions, and strong homophily effect in product-choice decision as well. We further distinguish in-group influence from out-group influence, and find that the former is more salient in the purchase-timing decision, while the latter is more salient in the product-choice decision. We show that models which ignore one of the factors result in the overestimation of the other factor. Furthermore, we demonstrate the importance of communication data by showing that influence effect can be either over- or underestimated when communication is not explicitly accounted for.

Using our estimates, we conduct several policy simulations to evaluate different promotional schemes. Our simulation shows that accounting for both homophily and influence effects increases the target effectiveness on purchase probability by 4-21% depending on the situation, while increases the effectiveness on product choice probability by about 11-35%. The performance can be further improved once the multiplier effect is taken into account. This demonstrates the importance for promotion of an improved understanding of the social network.

Two limitations of our study call for further investigation in future work. First, our study identifies friends by using n-cliques, which constrain the connection structure to groups of equal number of people who are tightly connected. Network structures are in fact more versatile. For example, certain people may have many friends but the ties may be weak, while some others may have only a few friends with very strong connections. Investigating how the number and strength of social ties impact consumer's decision, and how to best identify consumer preference in such flexible network structures will further our understanding of the implications of social networks. Second, our study shows that in-group influence is lower than out-group influence on product choice probability. Our conjecture is that the in-group influence on choosing a specific brand can have both a positive effect, where a consumer trusts his friend's selection, and a negative one, when a consumer tries to avoid imitating his friends. A more sophisticated model is needed to further isolate these two effects. Overall, we believe that understanding consumers within their social networks can lead to better marketing decisions.

References

Ansari, A., O. Koenigsberg and F. Stahl, (2008), “Modeling Connectivity in Online Social Networks”, Working Paper.

Aral, S., L. Muchnik and A. Sundararajan, (2009), “Distinguishing Influence Based Contagion from Homophily Driven Diffusion in Dynamic Networks”, Proceedings of the National Academy of Sciences, forthcoming

Bass, F.M., (1969), "A New Product Growth Model for Consumer Durables," Management Science, 15 (January), 215-27

Bell, D.R. and S. Song, (2007), “Neighborhood Effects and Trial on the Internet: Evidence from Online Grocery Retailing”, Quantitative Marketing and Economics, 5, 361-400

Bernheim, B.D., (1994), “A Theory of Conformity”, Journal of Political Economy, Vol. 102, No. 5, 841-877

Bott, H., (1928) “Observation of Play Activities in a Nursery School”, Genetic Psychology Monographs, 4:44-88

Braun, M. and A. Bonfrer, (2009), “Censoring, Interdependence and Scalability for Dyadic Social Media Data”, Working Paper

Chintagunta, P.K. (1993), “Investigating Purchase Incidence, Brand Choice and Purchase Quantity Decisions of Households”, Marketing Science, Vol. 12, No. 2, 184-208

Choi, J., S.K. Hui and D.R. Bell, (2008), “Bayesian Spatio-Temporal Analysis of Imitation Behavior Across New Buyers at an Online Grocery Retailer”, Journal of Marketing Research, Volume XLV

Christakis, N.A. and J.H. Fowler, (2007) “The Spread of Obesity in a Large Social Network over 32 Years”, New England Journal of Medicine, 357:370-379

Godes, D., D. Mayzlin, Y. Chen, S. Das, C. Dellarocas, B. Pfeiffer, B. Libai, S. Sen, M. Shi and P. Verlegh, (2005) “The Firm’s Management of Social Interactions”, Marketing Letters, 415-428

Gonul, F. and K. Srinivasan, (1993), “Modeling Multiple Sources of Heterogeneity in Multinomial Logit Models, Methodological and Managerial Issues”, Marketing Science, 12, 213-229

Gupta, S. (1991), "Stochastic Models of Interpurchase Time with Time-Dependent Covariates", Journal of Marketing Research, Vol. 28, No. 1, 1-15

Hartmann, W.R., (2008), "Demand Estimation with Social Interactions and the Implications for Targeted Marketing", Working Paper

Hartmann, W.R., H. Nair, P. Manchanda, M. Bothner, P. Dodds, D. Godes, K. Hosanagar and C. Tucker, (2008), "Modeling Social Interactions: Identification, Empirical Methods and Policy Implications", Marketing Letters, Vo. 19, No. 3-4, 287-304

Hill, S., F. Provost and C. Volinsky, (2006), "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks", Statistical Science, Vol. 21, No. 2, 256-276

Iyengar, R., C. Van den Bulte and T.W. Valente, (2010), "Opinion Leadership and Social Contagion in New Product Diffusion", Marketing Science, forthcoming

Jackson, M.O, (2003), "A Survey of Models of Network Formation: Stability and Efficiency", Group Formation in Economics: Networks, Clubs, and Coalitions, Cambridge University Press

Jackson, M.O and A. Watts, (2002), "The Evolution of Social and Economic Networks", Journal of Economic Theory, 106, 265-295

Manski, C.F., (1993), "Identification of Endogenous Social Effects: The Reflection Problem", Review of Economic Studies, 60, 531-542

McPherson, J.M. and L. Smith-Lovin, (1987), "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-face Groups", American Sociological Review, Vol. 52, No. 3, 370-379

Mcpherson, M., L. Smith-Lovin and J.M. Cook, (2001), "Birds of a Feather: Homophily in Social Networks", Annual Review of Sociology, Vol. 27, 415-444

Nair, H., P. Manchanda and T. Bhatia, (2006) "Asymmetric Peer Effects in Physician Prescription Behavior: The Role of Opinion Leaders", Working Paper

Van den Bulte, C. and G. Lilien, (2001), "Medical Innovation Revisited: Social Contagion versus Marketing Effort", American Journal of Sociology, 106(5), 1409-1435

Van den Bulte, C. and Y.V. Joshi, (2007), "New Product Diffusion with Influentials and Imitators", Marketing Science, Vol. 26, No. 3, 400-421

Van den Bulte, C. and S. Stremersch, (2004), "Social Contagion and Income Heterogeneity in New Product Diffusion: A Meta-Analytic Test", Marketing Science, 23(4), 530-544

Watts, D.J. and P.S. Dodds, (2007), “Influentials, Networks, and Public Opinion Formation”, Journal of Consumer Research, Vol. 34, No. 4, 441-458

Zheng, R., D. Wilkinson and F. Provost, (2008), “Social Network Collaborative Filtering”, Working Paper

Tables and Figures, to be inserted into the main text

Product/Category ID	*Market Share*
1	10.88%
2	54.88%
3	7.08%
4	27.16%

Table 1: CRBT Market Share by Category

Product/Category ID	*Unconditional Purchase Probability*	*Conditional Purchase Probability*
1	0.0822	0.150
2	0.236	0.291
3	0.0718	0.106
4	0.190	0.260

Table 2: Probability of Purchasing by Category

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\overline{\lambda}$	0.0323	0.0012	0.0301	0.0348
$\overline{\gamma}_{In}$	1.86	0.154	1.52	2.1
$\overline{\gamma}_{Out}$	0.724	0.14	0.457	0.956
σ^2_{λ}	0.337	0.0319	0.276	0.403
$\sigma^2_{\gamma In}$	0.896	0.372	0.462	1.99
$\sigma^2_{\gamma Out}$	0.584	0.14	0.333	0.883
r_{λ}	0.0124	0.056	-0.0958	0.126
$r_{\gamma In}$	0.0993	0.175	-0.177	0.431
$r_{\gamma Out}$	0.192	0.235	-0.213	0.604
$\overline{\kappa}^{pi}$	0.736	0.0448	0.677	0.850

Table 3: Purchase-Timing Model Parameter Evaluation

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\overline{\beta}_1$	-1.40	0.119	-1.68	-1.23
$\overline{\beta}_2$	0.235	0.0693	0.0833	0.362
$\overline{\beta}_3$	-1.35	0.0748	-1.482	-1.183
$\overline{\rho}_{In}$	0.0857	0.067	-0.0428	0.209
$\overline{\rho}_{Out}$	0.181	0.0414	0.101	0.259
$\sigma^2_{\beta_1}$	1.10	0.306	0.571	1.67
$\sigma^2_{\beta_2}$	0.286	0.141	0.0874	0.597
$\sigma^2_{\beta_3}$	0.458	0.29	0.137	1.06
$\sigma^2_{\rho In}$	0.221	0.106	0.0728	0.465
$\sigma^2_{\rho Out}$	0.053	0.014	0.029	0.0823
r_{β_1}	0.773	0.0877	0.568	0.900
r_{β_2}	0.317	0.179	0.0579	0.612
r_{β_3}	0.523	0.261	0.0557	0.883
$r_{\rho In}$	0.598	0.205	0.137	0.882
$r_{\rho Out}$	0.739	0.124	0.476	0.850
$\overline{\kappa}^{pc}$	0.670	0.0426	0.590	0.746

Table 4: Product-Choice Model Parameter Estimation

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\overline{\lambda}$	0.0325	0.00121	0.0302	0.0349
$\overline{\gamma}_{In}$	1.52	0.313	1.01	2.02
$\overline{\gamma}_{Out}$	0.682	0.162	0.377	1.01
σ^2_{λ}	0.342	0.0304	0.286	0.404

$\sigma^2_{\gamma In}$	0.663	0.334	0.222	1.38
$\sigma^2_{\gamma Out}$	0.557	0.229	0.234	1.09
$\bar{\kappa}^{pi}$	0.747	0.0525	0.645	0.827

Table 5: Purchase-Timing Parameter – Influence Only

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\bar{\beta}_1$	-1.25	0.103	-1.46	-1.08
$\bar{\beta}_2$	0.217	0.0732	0.0812	0.359
$\bar{\beta}_3$	-1.49	0.154	-1.74	-1.18
$\bar{\rho}_{In}$	0.27	0.0697	0.135	0.404
$\bar{\rho}_{Out}$	0.255	0.0584	0.137	0.368
$\sigma^2_{\beta_1}$	0.741	0.147	0.491	1.06
$\sigma^2_{\beta_2}$	0.342	0.0859	0.178	0.512
$\sigma^2_{\beta_3}$	0.647	0.238	0.331	1.09
$\sigma^2_{\rho In}$	0.345	0.149	0.135	0.746
$\sigma^2_{\rho Out}$	0.155	0.0458	0.0920	0.260
$\bar{\kappa}^{pc}$	0.887	0.0224	0.838	0.930

Table 6: Product-Choice Model Parameter – Influence Only

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\bar{\lambda}$	0.0350	0.00138	0.0324	0.0377
σ^2_{λ}	0.360	0.0353	0.294	0.430
r_{λ}	0.109	0.0521	0.0156	0.220

Table 7: Purchase-Timing Model Parameter – Homophily Only

Parameter	*Posterior Mean*	*Posterior Standard Deviation*	*2.5% Posterior Quantile*	*97.5% Posterior Quantile*
$\overline{\beta}_1$	-1.45	0.282	-1.91	-1.05
$\overline{\beta}_2$	0.266	0.0747	0.127	0.399
$\overline{\beta}_3$	-1.52	0.119	-1.74	-1.30
$\sigma^2_{\beta_1}$	1.02	0.687	0.208	2.18
$\sigma^2_{\beta_2}$	0.329	0.0771	0.196	0.498
$\sigma^2_{\beta_3}$	0.488	0.168	0.168	0.786
r_{β_1}	0.683	0.137	0.372	0.904
r_{β_2}	0.913	0.0586	0.698	0.917
r_{β_3}	0.625	0.214	0.241	0.896

Table 8: Product-Choice Model Parameter – Homophily Only

LL	*Full Model*	*Homophily-Only*	*Influence-Only*
Calibration	-9414.0	-9457.1	-9537.1
Holdout	-3372.2	-3922.1	-3442.1

Table 9: Model Comparison – Likelihood

Parameter	*Full Model*	*"No Communication" Model*
$\overline{\gamma}_{In}$	1.86	0.675
r_{λ}	0.0124	-0.0073

Table 10: Purchase-Timing – Compare with Missing Communication

Parameter	*Full Model*	*"No Communication" Model I*	*"No Communication" Model II*
$\overline{\rho}_{In}$	0.0857	0.1384	0.500

r_{β_1}	0.773	0.745	0.822
r_{β_2}	0.317	0.530	0.370
r_{β_3}	0.523	0.767	0.430

Table 11: Product Choice – Compare with Missing Communication

Measure	*Full Model*	*Homophily-Only*	*Influence-Only*
Purchase Probability Improvement	22.18%	20.45%	21.36%
Product-Choice Probability Improvement	11.99%	10.73%	9.12%

Table 12: Policy Simulation 1 – Target Existing Customers

Measure	*True Model*	*Homophily Only*	*Influence Only*
Purchase Probability Improvement	22.40%	NA	22.15%

Table 13: Policy Simulation 2 – Multiplier Effect

Measure	*True Model*	*Homophily Only*	*Influence Only*
Purchase Probability Improvement	10.86%	9.92%	8.98%

Table 14: Policy Simulation 3 – Targeting New Customers

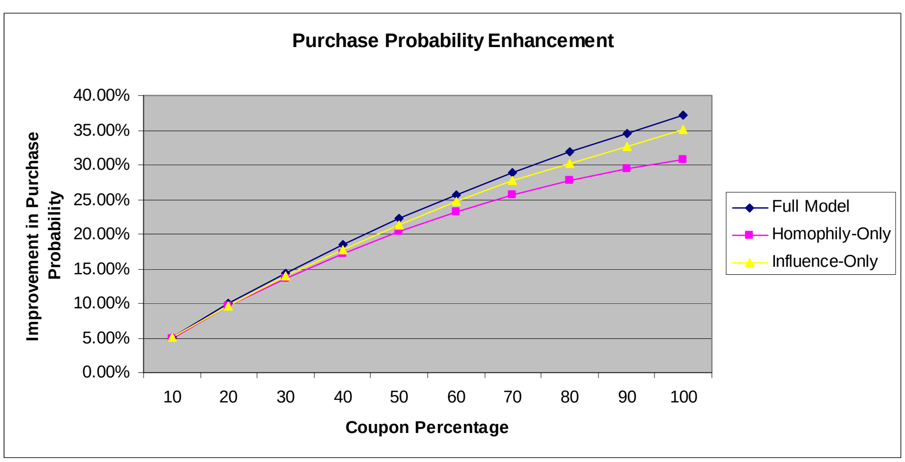

Figure 1: Policy Simulation 1 – Purchase Probability Enhancement

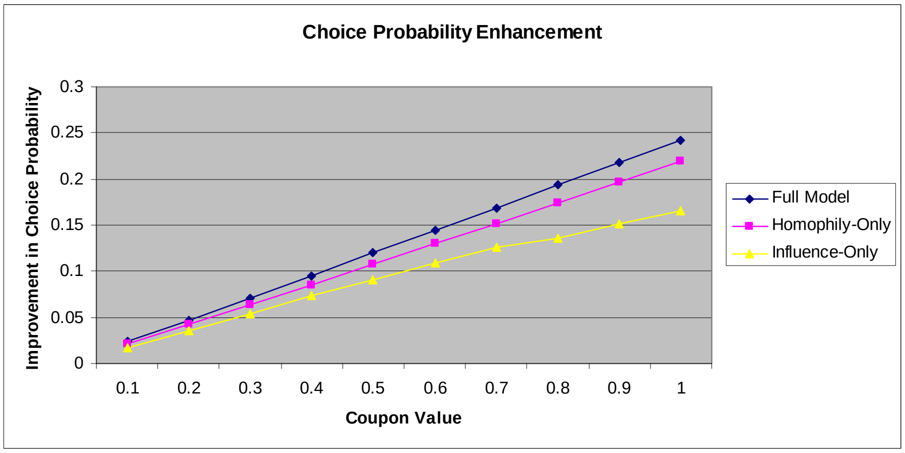

Figure 2: Policy Simulation 1 – Choice Probability Enhancement

Appendix A: MCMC Estimation Algorithm

We estimate parameters by taking MCMC draws. Draws are taken using a hybrid Metropolis-Gibbs strategy, where parameters are drawn individually conditional on others (the "Gibbs" strategy), while random-walk Metropolis is used to take individual parameter draws where the posterior cannot be sampled directly.

A.1 Inter-purchase timing

A.1.1 draw λ_g :

$$f(\lambda_g \mid P_g, E_g, \gamma_g, \bar{\lambda}, \sigma_\lambda^2, r_\lambda) \propto \psi(\begin{pmatrix} \lambda_{g,1} \\ \dots \\ \lambda_{g,I} \end{pmatrix} \mid \begin{pmatrix} \bar{\lambda} \\ \dots \\ \bar{\lambda} \end{pmatrix}, \sigma_\lambda^2 \begin{bmatrix} 1 & r_\lambda & r_\lambda \\ r_\lambda & \dots & r_\lambda \\ r_\lambda & r_\lambda & 1 \end{bmatrix}) f_{Erlang-2}(P_g \mid \lambda_g, E_g, \gamma_g) \quad \text{(A1)}$$

In (A1), $\psi(.)$ is the density of log-normal distribution. P_g (E_g) represents the purchases (exposures) of all persons in the group across all time periods. This step is repeated for each group: $g = 1..G$.

As the posterior cannot be sampled easily, we use Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0}, 0.2I)$ (here I represents the identity matrix).

A.1.2 draw $\bar{\lambda}$:

$$f(\bar{\lambda} \mid \lambda_g : g = 1..G) \propto \phi((GI + V_\lambda)^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I} \log(\lambda_{gi}) + V_\lambda \bar{\bar{\lambda}}), (GI + V_\lambda)^{-1}) \quad \text{(A2)}$$

In (A2), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_\lambda = 10000$ and $\bar{\bar{\lambda}} = 0$, so the posterior is normal.

A.1.3 draw γ_g :

$$f(\gamma_g \mid P_g, E_g, \lambda_g, \bar{\gamma}, \sigma_\gamma^2, r_\gamma) \propto \phi(\begin{pmatrix} \gamma_{g,1} \\ \dots \\ \gamma_{g,I} \end{pmatrix} \mid \begin{pmatrix} \bar{\gamma} \\ \dots \\ \bar{\gamma} \end{pmatrix}, \sigma_\gamma^2 \begin{bmatrix} 1 & r_\gamma & r_\gamma \\ r_\gamma & \dots & r_\gamma \\ r_\gamma & r_\gamma & 1 \end{bmatrix}) f_{Erlang-2}(P_g \mid \lambda_g, E_g, \gamma_g) \quad \text{(A3)}$$

In (A3), $\phi(.)$ is the density of the normal distribution. This step is repeated for each group: $g = 1..G$. Again, we use Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0}, 0.2I)$

A.1.4 draw $\bar{\gamma}$:

$$f(\bar{\gamma} \mid \gamma_g : g = 1..G) \propto \phi((GI + V_\gamma)^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I} \gamma_{gi} + V_\gamma \bar{\bar{\gamma}}), (GI + V_\gamma)^{-1}) \quad \text{(A4)}$$

In (A4), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_\gamma = 10000$ and $\overline{\overline{\gamma}} = 0$. The posterior is normal.

A.1.5 draw σ_λ^2:

$$f(\sigma_\lambda^2 \mid \lambda_g : g = 1..G, \overline{\lambda}, r_\lambda) \propto Inv-Gamma(v_{0,\lambda} + GI/2, s_{0,\lambda} + \frac{1}{2}\sum_{g=1}^{G}((\log(\lambda_g) - \begin{pmatrix} \overline{\lambda} \\ ... \\ \overline{\lambda} \end{pmatrix})^T \begin{bmatrix} 1 & r_\lambda & r_\lambda \\ r_\lambda & ... & r_\lambda \\ r_\lambda & r_\lambda & 1 \end{bmatrix}^{-1} ((\log(\lambda_g) - \begin{pmatrix} \overline{\lambda} \\ ... \\ \overline{\lambda} \end{pmatrix})) \quad \text{(A5)}$$

We choose the conjugate inverse-gamma prior with $v_{0,\lambda} = 0$ and $s_{0,\lambda} = 0$.

A.1.6 draw σ_γ^2:

$$f(\sigma_\gamma^2 \mid \gamma_g : g = 1..G, \overline{\gamma}, r_\gamma) \propto Inv-Gamma(v_{0,\gamma} + GI/2, s_{0,\gamma} + \frac{1}{2}\sum_{g=1}^{G}((\gamma_g - \begin{pmatrix} \overline{\gamma} \\ ... \\ \overline{\gamma} \end{pmatrix})^T \begin{bmatrix} 1 & r_\gamma & r_\gamma \\ r_\gamma & ... & r_\gamma \\ r_\gamma & r_\gamma & 1 \end{bmatrix}^{-1} (\gamma_g - \begin{pmatrix} \overline{\gamma} \\ ... \\ \overline{\gamma} \end{pmatrix})) \quad \text{(A6)}$$

We choose the conjugate inverse-gamma prior with $v_{0,\gamma} = 0$ and $s_{0,\gamma} = 0$.

A.1.7 draw r_λ:

$$f(r_\lambda \mid \lambda_g : g = 1..G, \overline{\lambda}, \sigma_\lambda^2) \propto \prod_{g=1}^{G} \psi(\begin{pmatrix} \lambda_{g,1} \\ ... \\ \lambda_{g,I} \end{pmatrix} \mid \begin{pmatrix} \overline{\lambda} \\ ... \\ \overline{\lambda} \end{pmatrix}, \sigma_\lambda^2 \begin{bmatrix} 1 & r_\lambda & r_\lambda \\ r_\lambda & ... & r_\lambda \\ r_\lambda & r_\lambda & 1 \end{bmatrix}) \quad \text{(A7)}$$

We use Metropolis with random walk, where the random walk step is an independent draw from $Normal(0,0.03)$.

A.1.8 draw r_γ:

$$f(r_\gamma \mid \gamma_g : g = 1..G, \overline{\gamma}, \sigma_\gamma^2) \propto \prod_{g=1}^{G} \phi(\begin{pmatrix} \gamma_{g,1} \\ ... \\ \gamma_{g,I} \end{pmatrix} \mid \begin{pmatrix} \overline{\gamma} \\ ... \\ \overline{\gamma} \end{pmatrix}, \sigma_\gamma^2 \begin{bmatrix} 1 & r_\gamma & r_\gamma \\ r_\gamma & ... & r_\gamma \\ r_\gamma & r_\gamma & 1 \end{bmatrix}) \quad \text{(A8)}$$

We use Metropolis with random walk, where the random walk step is an independent draw from $Normal(0,0.03)$.

A.1.9 draw κ_g^{pi}:

$$f(\kappa_g^{pi} \mid P_g, E_g, \gamma_g, \overline{\kappa}^{pi}, \sigma_{\kappa^{pi}}^2, r_{\kappa^{pi}}, \lambda_g) \propto$$

$$\varphi\left(\begin{pmatrix} \log it(\kappa_{g1}^{pi}) \\ ... \\ \log it(\kappa_{gI}^{pi}) \end{pmatrix} \mid \begin{pmatrix} \overline{\kappa}^{pi} \\ ... \\ \overline{\kappa}^{pi} \end{pmatrix}, \sigma_{\kappa^{pi}}^2 \begin{bmatrix} 1 & r_{\kappa^{pi}} & r_{\kappa^{pi}} \\ r_{\kappa^{pi}} & ... & r_{\kappa^{pi}} \\ r_{\kappa^{pi}} & r_{\kappa^{pi}} & 1 \end{bmatrix}\right) f_{Erlang-2}\left(P_g \mid \lambda_g, E_g(\tilde{E}_g, \kappa_g^{pi}), \gamma_g\right) \quad \text{(A9)}$$

In (A9), $E_g(\tilde{E}_g, \kappa_g^{pi})$ represents function to calculate smoothed exposure using the raw exposure and the smoothing parameter κ_g^{pi}. This step is repeated for each group: $g = 1..G$. We use

Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0},0.2I)$.

A.1.10 draw $\bar{\kappa}^{pi}$:

$$f(\bar{\kappa}^{pi} \mid \kappa_g^{pi} : g = 1..G) \propto \phi((GI + V_{\kappa^{pi}})^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I}\log it(\kappa_{gi}^{pi}) + V_{\kappa^{pi}}\overset{=}{\kappa}^{pi}), (GI + V_{\kappa^{pi}})^{-1}) \qquad \text{(A10)}$$

In (A10), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_{\kappa^{pi}} = 10000$ and $\overset{=}{\kappa}^{pi} = 0$. The posterior is normal.

A.1.11 draw $\sigma_{k^{pi}}^2$:

$$f(\sigma_{\kappa^{pi}}^2 \mid \kappa_g^{pi} : g = 1..G, \bar{\kappa}^{pi}, r_{\kappa^{pi}}) \propto$$
$$Inv-Gamma\left(v_{0,\kappa^{pi}} + GI/2,\ s_{0,\kappa^{pi}} + \frac{1}{2}\sum_{g=1}^{G}\left(\log it(\kappa_g^{pi}) - \begin{pmatrix}\bar{\kappa}^{pi}\\ \vdots \\ \bar{\kappa}^{pi}\end{pmatrix}\right)^T \begin{bmatrix}1 & r_{\kappa^{pi}} & r_{\kappa^{pi}}\\ r_{\kappa^{pi}} & \dots & r_{\kappa^{pi}}\\ r_{\kappa^{pi}} & r_{\kappa^{pi}} & 1\end{bmatrix}^{-1}\left(\log it(\kappa_g^{pi}) - \begin{pmatrix}\bar{\kappa}^{pi}\\ \vdots \\ \bar{\kappa}^{pi}\end{pmatrix}\right)\right) \qquad \text{(A11)}$$

We choose the conjugate inverse-gamma prior with $v_{0,\kappa^{pi}} = 0$ and $s_{0,\kappa^{pi}} = 0$.

A.1.12 draw $r_{\kappa^{pi}}$:

$$f(r_{\kappa^{pi}} \mid \kappa_g^{pi} : g = 1..G, \bar{\kappa}^{pi}, \sigma_{\kappa^{pi}}^2) \propto \prod_{g=1}^{G}\phi(\begin{pmatrix}\log it(\kappa_{g1}^{pi})\\ \dots \\ \log it(\kappa_{gI}^{pi})\end{pmatrix} \mid \begin{pmatrix}\bar{\kappa}^{pi}\\ \dots \\ \bar{\kappa}^{pi}\end{pmatrix}, \sigma_{\kappa^{pi}}^2\begin{bmatrix}1 & r_{\kappa^{pi}} & r_{\kappa^{pi}}\\ r_{\kappa^{pi}} & \dots & r_{\kappa^{pi}}\\ r_{\kappa^{pi}} & r_{\kappa^{pi}} & 1\end{bmatrix}) \qquad \text{(A12)}$$

We use a Metropolis algorithm with random walk, where the random walk step is an independent draw from $Normal(0,0.03)$.

A.2 Product-choice

A.2.1 draw $\beta_{g,j}$:

$$f(\beta_{g,j} \mid P_g, E_g, \beta_{g,-j}, \rho_g, \bar{\beta}_j, \sigma_{\beta_j}^2, r_{\beta_j}) \propto \phi(\begin{pmatrix}\beta_{g,1,j}\\ \dots \\ \beta_{g,I,j}\end{pmatrix} \mid \begin{pmatrix}\bar{\beta}_j\\ \dots \\ \bar{\beta}_j\end{pmatrix}, \sigma_{\beta_j}^2\begin{bmatrix}1 & r_{\beta_j} & r_{\beta_j}\\ r_{\beta_j} & \dots & r_{\beta_j}\\ r_{\beta_j} & r_{\beta_j} & 1\end{bmatrix}) f_{MNL}(P_g \mid E_g, \rho_g, \beta_g) \qquad \text{(B1)}$$

In (B1), $\phi(.)$ is the density of normal distribution. P_g (E_g) represents the purchases (exposures) of all persons in the group across all time periods. $\beta_{g,-j}$ represents all other product valuation coefficients. This step is repeated for each group, $g = 1..G$, and each product characteristic. As the product characteristic matrix consists of only product dummies, we repeat it for each product except the last one: $j = 1..J-1$ (the last one is normalized to 0 due to the identification issue of multinomial logit).

As the posterior cannot be sampled easily, we use a Metropolis algorithm with random walk. The random walk step is an independent draw from $MVN(\vec{0},0.3I)$ (here I represents the identity matrix).

A.2.2 draw $\overline{\beta}_j$:

$$f(\overline{\beta}_j \mid \beta_{g,j} : g=1..G) \propto \phi((GI+V_\beta)^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I}\beta_{g,i,j} + V_\beta \overline{\overline{\beta}}),(GI+V_\beta)^{-1}) \tag{B2}$$

In (B2), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_\beta = 10000$ and $\overline{\overline{\beta}} = 0$, so the posterior is normal. This step is repeated for each coefficient j.

A.2.3 draw ρ_g:

$$f(\rho_{g,j} \mid P_g, E_g, \beta_g, \overline{\rho}, \sigma_\rho^2, r_\rho) \propto \phi(\begin{pmatrix}\rho_{g,1}\\ \dots \\ \rho_{g,I}\end{pmatrix} \mid \begin{pmatrix}\overline{\rho}\\ \dots \\ \overline{\rho}\end{pmatrix}, \sigma_\rho^2 \begin{bmatrix}1 & r_\rho & r_\rho\\ r_\rho & \dots & r_\rho\\ r_\rho & r_\rho & 1\end{bmatrix}) f_{MNL}(P_g \mid E_g, \rho_g, \beta_g) \tag{B3}$$

In (B3), $\phi(.)$ is the density of normal distribution. P_g (E_g) represents the purchases (exposures) of all persons in the group across all time periods. This step is repeated for each group, $g = 1..G$. As the posterior cannot be sampled easily, we use Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0},0.3I)$ (here I represents the identity matrix).

A.2.4 draw $\overline{\rho}$:

$$f(\overline{\rho} \mid \rho_g : g=1..G) \propto \phi((GI+V_\rho)^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I}\rho_{g,i} + V_\rho \overline{\overline{\rho}}),(GI+V_\rho)^{-1}) \tag{B4}$$

In (B4), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_\rho = 10000$ and $\overline{\overline{\rho}} = 0$, so the posterior is normal.

A.2.5 draw $\sigma_{\beta_j}^2$:

$$f(\sigma_{\beta_j}^2 \mid \beta_{g,j} : g=1..G, \overline{\beta}_j, r_{\beta_j}) \propto Inv-Gamma(v_{0,\beta} + GI/2, s_{0,\beta} + \frac{1}{2}\sum_{g=1}^{G}((\beta_{g,j} - \begin{pmatrix}\overline{\beta}_j\\ \dots \\ \overline{\beta}_j\end{pmatrix})^T \begin{bmatrix}1 & r_{\beta_j} & r_{\beta_j}\\ r_{\beta_j} & \dots & r_{\beta_j}\\ r_{\beta_j} & r_{\beta_j} & 1\end{bmatrix}^{-1} (\beta_{g,j} - \begin{pmatrix}\overline{\beta}_j\\ \dots \\ \overline{\beta}_j\end{pmatrix})) \tag{A5}$$

We choose the conjugate inverse-gamma prior with $v_{0,\beta} = 0$ and $s_{0,\beta} = 0$. This step is repeated for each product coefficient except the last one: $j = 1..J-1$.

A.2.6 draw σ_ρ^2:

$$f(\sigma_\rho^2 \mid \rho_g : g=1..G, \overline{\rho}, r_\rho) \propto Inv-Gamma(v_{0,\rho} + GI/2, s_{0,\rho} + \frac{1}{2}\sum_{g=1}^{G}((\rho_g - \begin{pmatrix}\overline{\rho}\\ \dots \\ \overline{\rho}\end{pmatrix})^T \begin{bmatrix}1 & r_\rho & r_\rho\\ r_\rho & \dots & r_\rho\\ r_\rho & r_\rho & 1\end{bmatrix}^{-1} (\rho_g - \begin{pmatrix}\overline{\rho}\\ \dots \\ \overline{\rho}\end{pmatrix})) \tag{A6}$$

We choose the conjugate inverse-gamma prior with $v_{0,\rho} = 0$ and $s_{0,\rho} = 0$.

A.2.7 draw r_{β_j} :

$$f(r_{\beta_j} \mid \beta_{g,j} : g=1..G, \overline{\beta}_j, \sigma^2_{\beta_j}) \propto \prod_{g=1}^{G} \phi(\begin{pmatrix} \beta_{g,1,j} \\ \ldots \\ \beta_{g,I,j} \end{pmatrix} \mid \begin{pmatrix} \overline{\beta}_j \\ \ldots \\ \overline{\beta}_j \end{pmatrix}, \sigma^2_{\beta_j} \begin{bmatrix} 1 & r_{\beta_j} & r_{\beta_j} \\ r_{\beta_j} & \ldots & r_{\beta_j} \\ r_{\beta_j} & r_{\beta_j} & 1 \end{bmatrix}) \tag{B7}$$

We use Metropolis with random walk, where the random walk step is an independent draw from $Normal(0,0.1)$. This step is repeated for each coefficient except the last one: $j=1..J-1$.

A.2.8 draw r_ρ:

$$f(r_\rho \mid \rho_g : g=1..G, \overline{\rho}, \sigma^2_\rho) \propto \prod_{g=1}^{G} \phi(\begin{pmatrix} \rho_{g,1} \\ \ldots \\ \rho_{g,I} \end{pmatrix} \mid \begin{pmatrix} \overline{\rho} \\ \ldots \\ \overline{\rho} \end{pmatrix}, \sigma^2_\rho \begin{bmatrix} 1 & r_\rho & r_\rho \\ r_\rho & \ldots & r_\rho \\ r_\rho & r_\rho & 1 \end{bmatrix}) \tag{B8}$$

We use Metropolis with random walk, where the random walk step is an independent draw from $Normal(0,0.1)$.

A.2.9 draw κ_g^{pc}:

$$f(\kappa_g^{pc} \mid P_g, E_g, \rho_g, \overline{\kappa}^{pc}, \sigma^2_{\kappa^{pci}}, r_{\kappa^{pc}}, \beta_g) \propto$$
$$\varphi\left(\begin{pmatrix} \log it(\kappa_{g1}^{pc}) \\ \ldots \\ \log it(\kappa_{gI}^{pc}) \end{pmatrix} \middle| \begin{pmatrix} \overline{\kappa}^{pc} \\ \ldots \\ \overline{\kappa}^{pc} \end{pmatrix}, \sigma^2_{\kappa^{pc}} \begin{bmatrix} 1 & r_{\kappa^{pc}} & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & \ldots & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & r_{\kappa^{pc}} & 1 \end{bmatrix} \right) f_{MNL}(P_g \mid E_g(\tilde{E}_g, \kappa_g^{pc}), \rho_g, \beta_g) \tag{B9}$$

In (B9), $E_g(\tilde{E}_g, \kappa_g^{pc})$ represents function to calculate smoothed exposure using the raw exposure and the smoothing parameter κ_g^{pc}. This step is repeated for each group: $g=1..G$. We use Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0}, 0.2I)$.

A.2.10 draw $\overline{\kappa}^{pc}$:

$$f(\overline{\kappa}^{pc} \mid \kappa_g^{pc} : g=1..G) \propto$$
$$\varphi((GI + V_{\kappa^{pc}})^{-1}(\sum_{g=1}^{G}\sum_{i=1}^{I} \log it(\kappa_{gi}^{pc}) + V_{\kappa^{pc}} \overset{=}{\kappa}^{pc}), (GI + V_{\kappa^{pc}})^{-1}) \tag{B10}$$

In (B10), $\phi(.)$ is the density of the normal distribution. We choose the conjugate hyper-priors $V_{\kappa^{pc}} = 10000$ and $\overset{=}{\kappa}^{pc} = 0$. The posterior is normal.

A.2.11 draw $\sigma^2_{k^{pc}}$:

$$f(\sigma^2_{\kappa^{pc}} \mid \kappa_g^{pc} : g=1..G, \overline{\kappa}^{pc}, r_{\kappa^{pc}}) \propto$$
$$Inv-Gamma\left(v_{0,\kappa^{pc}} + GI/2, s_{0,\kappa^{pc}} + \frac{1}{2}\sum_{g=1}^{G}((\log it(\kappa_g^{pc}) - \begin{pmatrix} \overline{\kappa}^{pc} \\ \ldots \\ \overline{\kappa}^{pc} \end{pmatrix})^T \begin{bmatrix} 1 & r_{\kappa^{pc}} & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & \ldots & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & r_{\kappa^{pc}} & 1 \end{bmatrix}^{-1} (\log it(\kappa_g^{pc}) - \begin{pmatrix} \overline{\kappa}^{pc} \\ \ldots \\ \overline{\kappa}^{pc} \end{pmatrix} \right) \tag{B11}$$

We choose the conjugate inverse-gamma prior with $v_{0,\kappa^{pc}} = 0$ and $s_{0,\kappa^{pc}} = 0$.

A.2.12 draw $r_{\kappa^{pc}}$:

$$f(r_{\kappa^{pc}} \mid \kappa_g^{pc} : g = 1..G, \bar{\kappa}^{pc}, \sigma^2_{\kappa^{pc}}) \propto \prod_{g=1}^{G} \phi\left(\begin{pmatrix} \log it(\kappa_{g1}^{pc}) \\ \dots \\ \log it(\kappa_{gI}^{pc}) \end{pmatrix} \middle| \begin{pmatrix} \bar{\kappa}^{pc} \\ \dots \\ \bar{\kappa}^{pc} \end{pmatrix}, \sigma^2_{\kappa^{pc}} \begin{bmatrix} 1 & r_{\kappa^{pc}} & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & \dots & r_{\kappa^{pc}} \\ r_{\kappa^{pc}} & r_{\kappa^{pc}} & 1 \end{bmatrix} \right) \qquad \text{(B12)}$$

We use Metropolis with random walk, where the random walk step is an independent draw from a $N(0, 0.03)$.

Appendix B: Simulation Study

Note: This is a technical appendix that is to be provided as an online supplement and not to be published with the paper.

To validate the model identification as well as the estimation procedure, we apply the model and estimation code to a simulated dataset. The simulation for product choice decision and that for purchase timing decision are conducted separately. For the product choice decision, we simulated 300 groups over 30 purchase incidences. Each group consists of 4 consumers. There are 4 products to choose from. In each purchase incidence, each member in each group chooses one product based on the utility function in our model. Exposures to each product at each purchase incidence are drawn randomly.

We take 30,000 MCMC draws from the parameter posteriors. The first 10,000 are considered as burn-in draws and discarded, while the remaining 20,000 are used for parameter evaluation. The true parameter value, posterior mean, and 95% centered posterior interval are reported in Table B1 below. The MCMC plots are shown in Figures 1-3 below.

<<Insert Table B1 About Here>>

<<Insert Figures B1-B3 About Here>>

The convergence of the estimation procedure can be verified from the three figures. As Table 1 shows, the posterior mean estimate is very close to the true value for almost all parameters (with the only exception being $\bar{\beta}_3$, where the true value is -3 and the posterior mean is -2.89). For all parameters, the true parameter value falls within the 95% posterior interval centered at the posterior mean. This verifies the correctness of the model and the estimation procedure.

The simulation for the inter-purchase timing decision is conducted over 200 groups of 4 consumers each, over 300 periods. Exposures are drawn randomly for each consumer over each period. In each period, consumer makes the purchase decision according to our purchase timing model. The true parameter value and simulation results are reported in Table 2 below. The MCMC plots are shown in Figures B4-B5 below.

<<Insert Table B2 About Here>>

<<Insert Figures B4-B5 About Here>>

Again, the convergence of the estimation procedure is evident from the figures. As Table B2 shows, for all the parameters, the true parameter value is included in the 95% posterior interval. The parameter estimation for λ-related parameters is quite accurate, while that for γ-related ones is less so. The simulation is consistent with our model and code being correct.

Table B1: Simulation of Product Choice Model				
Parameter	True Value	Posterior Mean	2.5% Posterior Quantile	97.5% Posterior Quantile
$\overline{\beta}_1$	1	1.02	0.958	1.09
$\overline{\beta}_2$	-2	-2.02	-2.11	-1.93
$\overline{\beta}_3$	-3	-2.89	-3.01	-2.78
$\overline{\rho}$	0.3	0.30	0.245	0.360
$\sigma^2_{\beta_1}$	0.5	0.517	0.441	0.605
$\sigma^2_{\beta_2}$	1	0.983	0.826	1.16
$\sigma^2_{\beta_3}$	3	3.03	2.62	3.50
σ^2_{ρ}	0.2	0.210	0.186	0.237
r_{β_1}	0.6	0.619	0.535	0.697
r_{β_2}	-0.2	-0.214	-0.270	-0.141
r_{β_3}	0.3	0.280	0.197	0.365
r_{ρ}	0.5	0.503	0.439	0.564

Table B2: Simulation of Purchase Timing Model				
Parameter	True Value	Posterior Mean	2.5% Posterior Quantile	97.5% Posterior Quantile
$\overline{\lambda}$	0.1	0.100	0.0931	0.108
$\overline{\gamma}$	0.05	0.0370	-0.0424	0.117
σ^2_{λ}	0.5	0.467	0.415	0.524
σ^2_{γ}	0.02	0.0327	0.0198	0.0544
r_{λ}	0.2	0.158	0.0841	0.238
r_{γ}	0.4	0.655	0.387	0.880

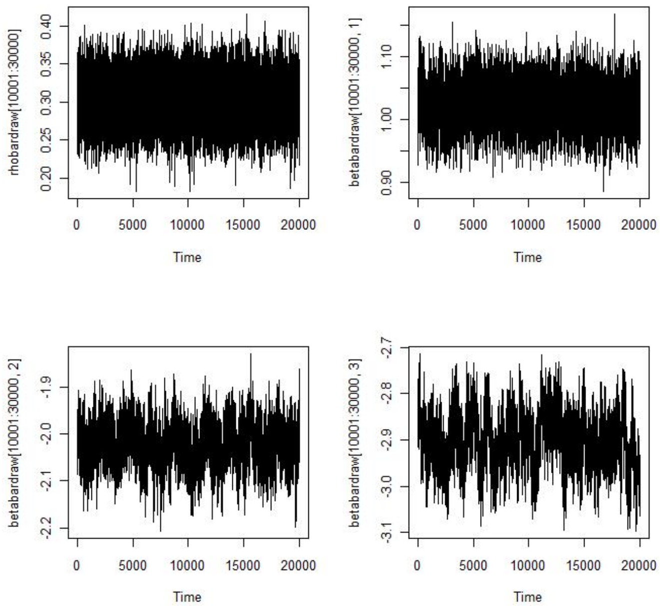

Figure B1: Product Choice Simulation – Population Mean

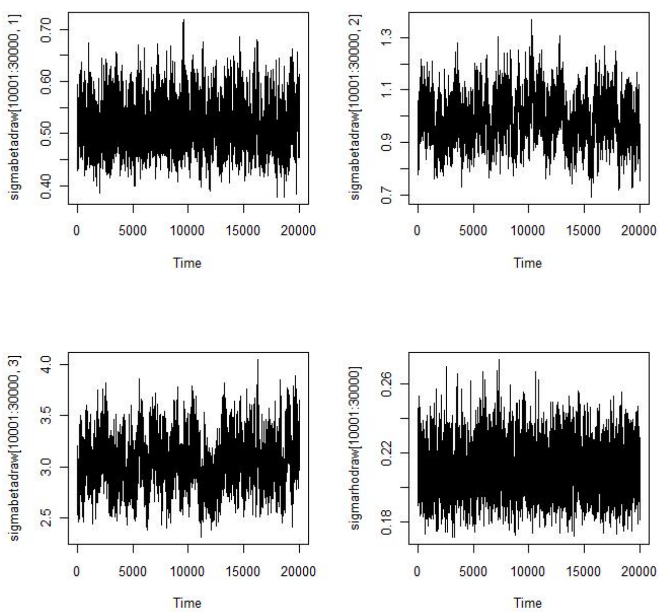

Figure B2: Product Choice Simulation – Population Variance

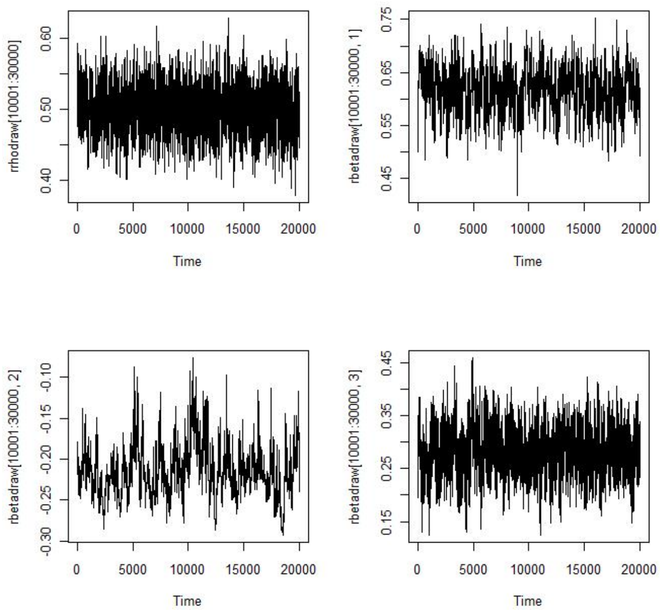

Figure B3: Product Choice Simulation – Group-Level Correlation

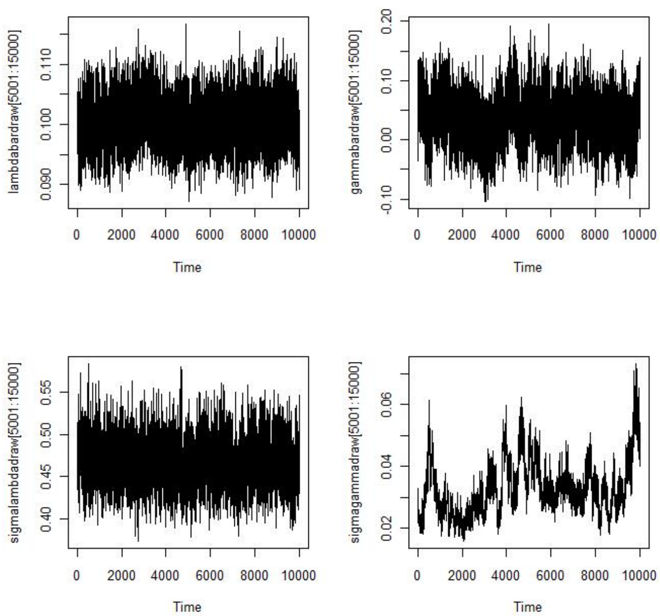

Figure B4: Purchase Timing Simulation – Population Mean and Variance

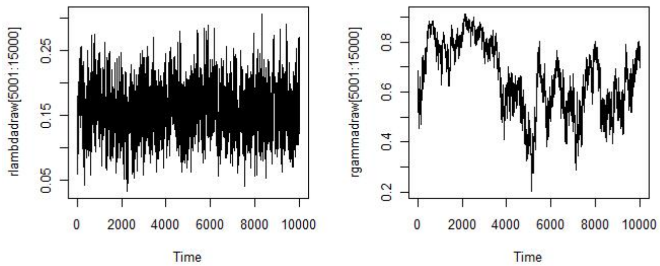

Figure B5: Purchase Timing Simulation – Group-Level Correlation