Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Liye Ma, Ramayya Krishnan, Alan L. Montgomery

To cite this article:
Liye Ma, Ramayya Krishnan, Alan L. Montgomery (2015) Latent Homophily or Social Influence? An Empirical Analysis of Purchase Within a Social Network. Management Science 61(2):454-473. https://doi.org/10.1287/mnsc.2014.1928

MANAGEMENT SCIENCE
Vol. 61, No. 2, February 2015, pp. 454–473
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2014.1928

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

Liye Ma
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742, liyema@rhsmith.umd.edu

Ramayya Krishnan
H. J. Heinz III College, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, rk2x@cmu.edu

Alan L. Montgomery
Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, alanmontgomery@cmu.edu

Consumers who are close to one another in a social network often make similar purchase decisions. This similarity can result from latent homophily or social influence, as well as common exogenous factors. Latent homophily means consumers who are connected to one another are likely to have similar characteristics and product preferences. Social influence refers to the ability of one consumer to directly influence another consumer's decision based upon their communication. We present an empirical study of purchases of caller ring-back tones using data from an Asian mobile network that predicts consumers' purchase timing and choice decisions. We simultaneously measure latent homophily and social influence, while also accounting for exogenous factors. Identification is achieved due to our dynamic, panel data structure and the availability of detailed communication data. We find strong influence effects and latent homophily effects in both the purchase timing and product choice decisions of consumers.

Keywords: social network; latent homophily; social influence; purchase timing; product choice; hierarchical Bayesian model; marketing
History: Received June 8, 2011; accepted January 30, 2014, by Sandra Slaughter, information systems. Published online in *Articles in Advance* June 23, 2014.

1. Introduction

Mobile and Internet technologies have revolutionized the way consumers interact with one another, with communication data in this arena providing insights into a consumer's social environment. For example, telephone call records provide a basis for inferring an individual's network neighbors as well as the strength of these ties. Additionally, newer mobile telecommunications technologies not only enable more efficient electronic communication but also enable commerce, since customers can now purchase and consume a wide variety of information goods such as news, traffic reports, movies, music, and applications directly with their mobile phones. Such an electronically enabled environment is a perfect context in which to study how a consumer's purchase decisions are both shaped by and a reflection of her social environment.

Consumers who are socially connected are known to have similar behaviors. The key question confronting business managers, however, is the reason behind this similarity: If two friends purchase similar products, why? Two competing explanations are readily available. The first is that a consumer can directly influence her friends to purchase similar products, a factor that is given several different names in the literature and which we refer to generally as *social influence*. The second is that two consumers who are connected are likely to have similar characteristics. Social scientists have long recognized that people tend to form ties with others who are similar to them, an effect known as *homophily*. Consequently, people who have close ties tend to have similar product tastes, hence they purchase similar products. Separating these two effects is crucial to business managers because they call for different promotion strategies. If homophily is the reason purchase decisions are similar among friends, then a firm should promote its products directly to the friends of existing customers. However, if social influence is the underlying factor, then a firm should incentivize existing customers to promote to their friends—for example, through a referral bonus. Thus, uncovering the underlying driver of the similarity in purchase decisions is crucially important to understanding the role of social networks in commerce and to formulating an appropriate marketing strategy.

However, separating homophily from influence is challenging, both in the context of product purchase and in human decisions in general, since the two effects usually lead to similar outcomes. Further complicating the issue is the possibility of a third factor—exogenous shocks. Using the above example, friends may have purchased similar products due neither to homophily nor to influence, but because of common exogenous shocks. For example, friends may have gone shopping together and been exposed to the same advertisement, which triggered their purchases. Distinguishing among all three effects is difficult. In fact, Manski (1993) proves that it is theoretically impossible to distinguish them in a static context.

A rich literature has subsequently focused on social influence, while providing different levels of control of homophily or exogenous factors. The impact of social influence is usually treated as an exogenous effect and changes in observed characteristics are usually included as covariates (Iyengar et al. 2011, Choi et al. 2010). Nair et al. (2010) analyze peer effects in prescription decisions while controlling for homophily using individual fixed effects. Aral et al. (2009) investigate homophily through propensity-score matching based on observed characteristics, and they provide bounds on the influence effect in product adoption decisions. They show that ignoring the observed homophily results in a significant over-estimation of influence effects.

Although extensive research exists in this area, many questions remain open. Three are the focus of our research. First, how can we separate the effect of latent homophily from that of social influence in purchase decisions, while also accounting for exogenous factors? Existing work has measured the influence effect while controlling for homophily using observed factors (Aral et al. 2009), but the issue of latent homophily is more challenging (Shalizi and Thomas 2011). For product purchase decisions, unobserved heterogeneity has long been recognized as a key factor (Gönül and Srinivasan 1993). Existing studies on purchase decisions, however, can identify influence only while treating latent homophily and exogenous factors as a combined effect (Anagnostopoulos et al. 2008). Explicit modeling of latent homophily is thus an important open question. Second, how do we quantify the two effects simultaneously? Existing studies usually focus on social influence effects, while mostly treating homophily as a factor to be controlled for. This leads to models that can quantify only influence but not homophily. However, an accurate measure of homophily is equally important, since it has direct implications for targeting strategies. Finally, what is the role of communication data in measuring social influence effects? Social influence is by its nature dependent on communication. Communication data, however, are often unavailable to researchers. Existing research often uses the decisions of one person as a factor in explaining the decision of another, with the implicit assumption that the former's decision is always made known to the latter instantaneously. Observing the actual communication allows us to assess the implications of such assumptions.

We address these questions by leveraging a unique data set of consumer choices over time, which includes information about purchase, network structure, and communication. We develop a compatible hierarchical Bayesian model to evaluate the effect of latent homophily and that of social influence on consumers' purchase decisions in a social network environment, while also accounting for exogenous factors. We model both purchase incidence (i.e., the timing of purchase) and product choice decisions. Social influence is allowed to influence both decisions and is measured from the dependence of these decisions on exposures to other consumers who possess the product. Latent homophily is accounted for in unobserved personal traits including product tastes, intrinsic purchase frequencies and the susceptibility to influence, and it is operationalized as the correlation of these traits among connected consumers. In addition, the model also accounts for common exogenous factors at both the population and group level, using both fixed effects and random effects. Our model thus incorporates all three major factors recognized in literature as leading to decision similarity (Manski 1993). Finally, we adopt a latent instrumental variable (LIV) approach to control for potential endogeneity concerns with respect to the influence effect (Ebbes et al. 2005).

Our estimation results show strong social influence effects in both the purchase incidence and product choice decisions of consumers, and strong latent homophily in both product tastes and susceptibility to influence. Network neighbors influence the purchase incidence decisions more than people outside the group do, while the latter are more influential on product choices. We also find that exogenous factors play an important role in the purchase and choice decisions. All these effects are robust to data sampling and selection. Furthermore, we show that ignoring either the homophily or influence effect leads to overestimation of the other effect. More importantly, we show that ignoring the communication information leads to biased estimates of the influence effects. Finally, we demonstrate through simulation that understanding these underlying drivers enable effective targeting of consumers.

Our study contributes to the literature in the following ways. First, we separate latent homophily from social influence. Existing studies have separated influence from observed homophily (i.e., similarities that can be attributed to similar consumer characteristics like age, location, religious affiliation, etc.).

Latent homophily explains similarity in consumer preferences due to unobserved characteristics, and its decomposition from social influence in product purchase decisions has not been considered. Marketing research has consistently shown that product purchase decisions are driven by unobserved preferences, which we can consider correlated among consumers due to latent homophily. Second, we simultaneously quantify latent homophily and influence effect while providing quantification of exogenous time-varying factors as well. Existing studies usually focus on the influence effect; they only control for homophily but do not measure its strength. Both effects, however, have important managerial implications because they provide the basis for targeting strategies, a point also validated in our policy simulation. Finally, we investigate the role of communication data in evaluating these drivers of purchase decisions. Decision variables have often been used in lieu of communication by extant studies. We show that communication data are necessary for the accurate evaluation of the influence effect, and using decision information to proxy for communication leads to biased estimates of influence as well as homophily.

2. Literature Review

The ubiquitous adoption of information technology has enabled the gathering and processing of large-scale network data, leading to a growing number of studies on social networks in various fields such as economics (Jackson and Watts 2002), information systems (Hill et al. 2006), machine learning (Zheng et al. 2008), and marketing (Hartmann et al. 2008). This literature is concerned with both the formation of social networks (Ansari et al. 2011, Braun and Bonfrer 2011) and the implication of the network on consumer behavior (Nair et al. 2010; see Jackson 2003 for a comprehensive survey).

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 2010), peer effects (Nair et al. 2010), social contagion (Van den Bulte and Lilien 2001, Iyengar et al. 2011), conformity (Bernheim 1994), imitation (Bass 1969, Choi et al. 2010), 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 the interaction with others, an effect we term *social influence*.

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 observed influence. Hartmann (2010) 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. (2010) quantify the impact of social interactions and peer effects in the context of prescription choices by physicians, and they 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. (2010) 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 over 80 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 in a network environment. 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).

Since at least Manski (1993), studies on social influence introduced various approaches to control for homophily, exogenous factors, or anything other than influence in general. Nair et al. (2010) uses an individual fixed effect as a control for effects other than peer influence. Hartmann (2010) jointly estimates group-level correlation with other parameters that govern a coordination game to account for homophily in product taste. This approach is similar to the one we propose in our study. However, Hartmann (2010) focuses on group coordination instead of on influence induced by communication and does not elaborate on the

role that homophily could play in purchase behavior. Furthermore, latent homophily may exist on other traits such as the susceptibility to influence, instead of just base-level product tastes. 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. Studies on social influence certainly are not restricted to purchase and adoption decisions. For example, Christakis and Fowler (2007) study the spread of obesity in social networks. They controlled for homophily using lagged independent variables, an approach which was subsequently critiqued (Lyons 2011).

Studies also seek to explicitly address the separation of homohily from influence in various contexts. Different approaches have been proposed, including experiments (Judd et al. 2010), specification of weight structures (Leenders 2002), and explicit modeling of coevolution of network and action (Snijders et al. 2010, Steglich et al. 2010). Statistical tests have also been developed to test hypotheses on homophily and influence (Steeg and Galstyan 2010, La Fond and Neville 2010). Shalizi and Thomas (2011) show that in a general nonparametric model, latent homophily is indistinguishable from influence, although purchase and choice models in the literature are generally parametric (e.g., Guadagni and Little 1983, Gupta 1991). These existing studies either address specific situations (e.g., Judd et al. 2010) or explicitly model observed homophily (e.g., Snijders et al. 2010, Steglich et al. 2010), which make them not applicable to product purchase decisions. The statistical tests also do not quantify both effects. For product purchase and adoption, Aral et al. (2009) separated influence effect from observed homophily in the adoption of a mobile service application, while Anagnostopoulos et al. (2008) proposed tests that identify influence from homophily in adoption decisions. In the latter study, however, latent homophily is not distinguished from exogenous factors and is not measured. Compared with the existing literature, our study provides explicit separation of latent homophily from influence, where both effects are quantified and exogenous factors are also accounted for and quantified.

3. Modeling Product Purchase Within a Social Network

We now discuss our model, which simultaneously accounts for latent homophily, social influence, and exogenous factors in consumer purchase decisions. Since the separation of these effects depends on the data, we begin with a description of our data set, which was provided by a large Asian telecom company and consists of detailed phone call histories of all the company's customers in a major city over a three-month period. There are over 3.7 million customers in the data set and over 300 million phone calls. For each phone call, we know the caller and callee phone numbers, date and time of the call, and length of the conversation.

Also included are purchase records of a product known as caller ring-back tones (CRBT). CRBT is a popular phone feature in a number of Asian countries such as India and China, although it is not widespread in the American market. A CRBT is usually a short snippet of a musical song played in lieu of the ordinary ringtone to the caller. For example, if customer A purchases a certain ring-back tone, then when person B calls A, B will hear the ring-back tone played over the phone instead of the usual ringing, before A picks up the call. Notice that only customer A's callers hear this tone, not customer A herself. To purchase a CRBT, a customer pays a monthly subscription fee and subsequently selects the individual tone that she wants played when she 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 750,000 customers purchased CRBT in our data set. The type of tones that are purchased and when they are purchased are the decisions we study.

The nature of CRBT provides a unique opportunity to uncover the underlying drivers of consumer purchase decisions in social networks, especially as it helps evaluate social influence. For social influence to enter purchase decisions, there must be relevant communications between friends, either actively as one talks to the other about the product, or passively as one observes that the other purchased the product. Communication data, however, are often missing in data sets available to researchers. Even when communication information is available, one usually knows only the *occurrence* of communication, but not its *content*. The nature of CRBT, however, means that from the purchase and phone call records, we can pinpoint the relevant communication between friends—if consumer A purchased a tone, then consumer B called A, then we know B was exposed to and became aware that A purchased the tone. Furthermore, CRBT is a type of cheap, casual product, the purchase of which is unlikely subject to much deliberation. The phone call records thus provide an accurate proxy for communications on the product. This, as we discuss in the next subsection, enables our identification.

3.1. Identification Strategy

A rich literature has elaborated on the challenge of uncovering the reason behind similar decisions made by connected people. The similarity may result from latent homophily—friends may purchase similar products because they have similar tastes, as people tend to associate with other people who are alike. It may result from social influence—friends may purchase similar

products because one convinced the other to do the same. Alternatively, they may result from exogenous environmental factors that friends are exposed to. For example, friends may purchase similar products because they heard the same song on the radio. That all these factors may coexist further complicates our task.

Identification in our framework is achieved because the three factors that cause purchase similarity—latent homophily, influence, and exogenous factors—lead to different purchase patterns in a dynamic environment where the product is purchased repeatedly and communication information is available. Conceptually, the key to our identification is the static nature of the latent homophily effect versus the dynamic nature of the social influence effect and the exogenous effect, and the key to identifying social influence versus the exogenous effect is the dependence of the former on communication versus the independence of the latter.

Consider two friends who repeatedly make similar purchases. If social influence is the driver, then we expect to see that one person makes the purchase first, followed by her communication with the other person, which is then followed by the similar purchases by the other person. If a common exogenous shock is the reason, then we expect to see that the two persons make the purchase around the same time regardless of communication. If latent homophily is the reason, however, then we would see the two persons make the purchases independently without communicating with each other, and not necessarily in a synchronous manner. Note that latent homophily does not preclude synchronous purchases. However, that will be out of coincidence, as the intrinsic preference similarity does not prescribe the timing of purchase. In summary, the dynamic nature of our data set with repeated choices for an individual and the availability of communication information make the separation of these effects possible.

Identifying and quantifying social influence is usually an extremely challenging task, since detailed communication history among people is rarely available to researchers. Even when the event of communication is known, the content of the communication is still generally unavailable. Our data set is different because we observe each individual phone call (or communication within the network) and its timestamp. Although we do not know the content of the conversations, the influence a customer imposes on a caller is conveniently encoded in their call records, due to the nature of CRBT. 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 knows that the callee has purchased the tone. 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. Since 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 the decisions of others at the general level, for which such communication data is sufficient. At a more detailed level, we note that the influence in our study is of a passive nature—the influence customer A "exerts" on 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.

3.2. Model

We now formally explain our model. Following the literature (Chintagunta 1993) there are two steps in our purchase decision: (1) when to buy and (2) what to buy. We model the first step, the purchase incidence decision, using a model of interpurchase timing with its corresponding hazard rate. The second step, the product choice decision, is modeled using a discrete choice model. We set up our model using a hierarchical Bayesian framework, which allows for rich heterogeneity specifications and straightforward estimation.

We assume that consumers belong to one of G groups. Each group consists of I consumers. The ith consumer of gth group is indexed as gi. Consumers who belong to the same group are assumed to have a strong social relationship and communicate regularly with those close to themselves. We define a group as being made up of callers who have called one another frequently. In §4.1 we discuss the choices of the size of the group and the call threshold. We conjecture that both the purchase incidence decisions and the product choice decisions 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.

3.2.1. Purchase Incidence. Consumers may purchase a CRBT at any time, and we model the when-to-buy decision using its hazard rate. Time is continuous and denoted by τ, and purchase occasions are discrete and indexed by t which ranges from 1 to T. We assume that the interpurchase time of a consumer gi follows an Erlang-2 distribution with a time varying rate parameter $\lambda_{gi,t}$ (Gupta 1991). The density (f) and survivor functions (S) of the interpurchase time are

$$\begin{aligned} f_{gi}(\tau) &= \lambda_{gi,t}^2 \tau \exp(-\lambda_{gi,t}\tau), \\ S_{gi}(\tau) &= (1+\lambda_{gi,t}\tau)\exp(-\lambda_{gi,t}\tau). \end{aligned} \tag{1}$$

The rate is allowed to vary over time to reflect the chance that consumers become more (or less) likely to buy if they are exposed to others that use the product, or to exogenous factors. A customer is exposed to CRBT through calling either others inside his social group or those outside his social group. Exposure from inside the group and that from outside may differ in how influential it is on the consumer's decision. We denote $E_{gi,t,k}$ as a measure of the exposure to CRBT that consumer gi had at period t from either inside ($k = \text{In}$) or outside ($k = \text{Out}$) his group.[1] We propose the following model of the purchase rate parameter as a function of cumulative product exposure from both inside and outside the group:

$$\lambda_{gi,t} = \lambda_{gi} \exp(\gamma_{gi,\text{In}} E_{gi,t,\text{In}} + \gamma_{gi,\text{Out}} E_{gi,t,\text{Out}} + \xi_t + \nu_{g,t} + \varepsilon_{gi,t}). \quad (2)$$

In Equation (2), $\gamma_{gi,k}$ (for $k \in \{\text{In}, \text{Out}\}$) can be considered as a *susceptibility* parameter, which indicates the extent to which the consumer is subject to 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 and makes her own decisions. A positive sign indicates the consumer positively responds to influence, while a negative sign shows that the consumer handles external influence negatively. To control for potential common exogenous shocks, we introduce both a fixed effect and a random effect. First, ξ_t measures a population level trend that varies over time. It captures exogenous shocks that apply to all consumers. Second, $\nu_{g,t}$ measures possible group- and time-specific random shocks, e.g., if the group goes to watch a movie together, which increases the purchase probability of the movie's theme for all members. Finally, $\varepsilon_{gi,t}$ is an error term for the consumer at time t, which we introduce to control for potential endogeneity in the influence effect using the latent instrumental variables approach, which is explained in detail in §3.2.4.

Both influence and exogenous effects enter Equation (2) directly. Latent homophily, in contrast, is modeled as group-level correlation that enters indirectly. If latent homophily is present, then members in the same group should have similar characteristics, such as purchase rate and susceptibility to influence. In a hierarchical Bayesian framework, we account for this by assuming that the individual-level parameter of all members of a group is a random vector, where individual-level elements can be correlated. First, we assume that the base purchase rate of all members of a group follows a multivariate log-normal distribution:

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

In Equation (3), $\bar{\lambda}$ the population-level base rate parameter, and σ_λ^2 measures the dispersion of this base rate parameter in the population. We choose a diffuse log-normal hyperprior for $\bar{\lambda}$, and an inverse-Gamma hyperprior for σ_λ^2, both of which are conjugate priors. 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 hyperprior for r_λ. If latent homophily exists, the parameter values of consumers in the same group should be positively correlated. That is, $r_\lambda > 0$ indicates the presence of latent homophily.

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

$$\begin{bmatrix} \gamma_{g1,k} \\ \gamma_{g2,k} \\ \vdots \\ \gamma_{gI,k} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} \bar{\gamma}_k \\ \bar{\gamma}_k \\ \vdots \\ \bar{\gamma}_k \end{bmatrix}, \sigma_{\gamma_k}^2 \begin{bmatrix} 1 & r_{\gamma_k} & \cdots & r_{\gamma_k} \\ r_{\gamma_k} & 1 & \cdots & r_{\gamma_k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{\gamma_k} & r_{\gamma_k} & \cdots & 1 \end{bmatrix} \right). \quad (4)$$

The specifications for $\bar{\gamma}_k$, $\sigma_{\gamma_k}^2$, and r_{γ_k} are similar to those for the base rate parameters. Again we expect $r_{\gamma_k} > 0$, which would be evidence of latent homophily. By allowing the susceptibility-to-influence parameter to be correlated among members, we account for the possibility that latent homophily is not limited to direct factors such as product taste and purchase frequency but can exist for all relevant consumer traits.

We choose a diffuse normal prior for each element of the fixed effect ξ_t, and we fix their sum across time to be zero for normalization. We account for group-level exogenous shocks as random effects. Specifically, we assume each $\nu_{g,t}$ is an independent draw from a population-level normal distribution: $\nu_{g,t} \sim \text{N}(0, \sigma_\nu^2)$. We choose a diffuse inverse-Gamma prior for σ_ν^2, which is a conjugate prior.

The fixed effect ξ_t accounts for factors that apply to all the consumers, i.e., consumers may be more likely to purchase during a holiday period. The random effect $\nu_{g,t}$, which enters the purchase equation of every member of the same group, controls for group-level exogenous shocks, i.e., a group of friends may go watch a movie together and subsequently purchase the movie

[1] In this case we can think of a group as those individuals to which a consumer is directly connected. We think of exposures from these direct connections as coming from inside the group. Those communications that come from those individuals who are not directly connected are treated as coming from outside the group. We discuss group construction in §4.1. Our exposure variable is an exponentially smoothed count of the number of actual exposures to CRBT. The construction of the exposure variable is discussed further in §4.5.

tune. Note that although we assume each $\nu_{g,t}$ is an independent draw, ex post correlation may exist in these variables and may exist between them and the other variables in the purchase equation, such as the exposure variables. Such correlation affects estimation efficiency but not its consistency. This random effect can account for both instantaneous group-level exogenous shocks (e.g., friends are exposed to the same music and purchase it immediately), and their delayed effect to the extent the effect depreciates with the same rate for group members (e.g., friends are exposed to the same music, which affects their subsequent purchase decisions to the same extent). This setup, however, cannot account for heterogeneous depreciation rates of group members to exogenous shocks. This is a limitation of our model, which we leave for future research.

3.2.2. Product Choice. The what-to-buy decision step is modeled using a discrete multinomial choice model. There are J products, and the $L \times 1$ vector of product characteristics associated with product j is $\mathbf{X}_j$. The products in our problem are the genre of the ringtone and are listed in Table 1 along with their market share. We chose to focus on the genre of the music instead of the individual ringtone because most tones are purchased only a few times. This is similar to an aggregation from UPC level to brand level that is routinely seen in the choice literature (e.g., Guadagni and Little 1983). Alternatively, one could consider this as the first stage of a nested choice model, where in the second stage the consumer would choose the actual ringtone. This nested choice structure is a natural one since the telecom organizes songs by these genres and guides consumers to choose genre before selecting the ringtone.

We allow the amount of exposure that the consumer has received at time t to influence his choice, and as in Equation (2), we allow for differential effects from inside and outside the group. We denote $E_{gi,j,t,k}$ as the cumulative exposures that consumer gi has received for product j at period t from either inside or outside his group ($k \in \{\text{In, Out}\}$). The utility of consumer gi from purchasing product j at time period t is

$$\begin{aligned} U_{gi,j,t} &= \bar{U}_{gi,j,t} + \varepsilon_{gi,j,t} \\ &= \mathbf{X}_j'\boldsymbol{\beta}_{gi} + \rho_{gi,\text{In}} E_{gi,j,t,\text{In}} + \rho_{gi,\text{Out}} E_{gi,j,t,\text{Out}} \\ &\quad + \delta_{j,t} + \eta_{g,j,t} + \zeta_{gi,j,t} + \varepsilon_{gi,j,t}, \end{aligned} \tag{5}$$

where $\boldsymbol{\beta}_{gi}$ is an $L \times 1$ valuation coefficient vector for consumer gi. Similar to $\gamma_{gi,k}$ in the purchasing timing equation, the parameter $\rho_{gi,k}$ 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 $\gamma_{gi,k}$. Also similar to the purchase incidence decision, in the product choice decision we account for possible exogenous shocks using both a fixed effect and a random effect; $\delta_{j,t}$ represents time-varying exogenous shocks that apply to all consumers that alter the relative attractiveness of product j, whereas $\eta_{g,j,t}$ represents group- and time-specific shocks. We denote $\zeta_{gi,j,t}$ as the individual and time-specific random term, which is introduced to control for remaining endogeneity concern using LIV, explained in §3.2.4.

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

$$\Pr(gi \text{ chooses } j \text{ at time } t) = \frac{\exp(\bar{U}_{gi,j,t})}{\sum_{l=1}^{J} \exp(\bar{U}_{gi,l,t})}. \tag{6}$$

To account for latent homophily, we again introduce a hierarchical specification for each $\beta_{g,l}$ parameter and each $\rho_{g,k}$ parameter for all members of a group as follows:

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

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

The hyperpriors for $\bar{\beta}_l$, $\bar{\rho}_k$, $\sigma^2_{\beta_l}$, $\sigma^2_{\rho_k}$, r_{β_l}, and r_{ρ_k} are chosen similarly to their counterparts in the purchase incidence decision. Again, the presence of latent homophily will be reflected from positive correlation among group members, i.e., $r_{\beta_l} > 0$ and $r_{\rho_k} > 0$ are evidences of latent homophily.

Similar to the corresponding parameters in the purchase timing model, we choose a diffuse normal prior for each element of the fixed effect $\delta_{j,t}$. We set the sum of these fixed effects across time to be zero for each product for normalization. We account for group-level random shocks as random effects, by assuming that they are independent draws from a population level distribution: $\eta_{g,j,t} \sim N(0, \sigma^2_{\eta,j})$. The setup and implications of $\eta_{g,j,t}$ are similar to those of $\nu_{g,t}$, which is discussed in §3.2.1. Note that the variance of the random effect differs across products. This is to account for the possibility that consumers may be subject to different levels of exogenous shocks for different products. For example, friends might watch a movie together

Table 1 Genres and Market Share

Genre ID	Genre name	Market share (%)
1	Regional	12.06
2	Bollywood	67.50
3	Devotional	7.39
4	Other	13.05

more frequently than going to a concert together. If that is true, movie tunes will be more subject to random exogenous shocks than classic music tunes.

3.2.3. Quantifying Exposure. From phone call records we can infer the exposure received by a consumer from sources inside and outside her group. The quantification of such exposure is different for the purchase timing and product choice decisions. For purchase incidence, 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 from phone call records. Identifying 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 the customer buys a tone in the same genre. However, the consumer may purchase the tone several days later instead of on the same day. To capture these types of delays in purchase we exponentially smooth the exposure over time:

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

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

In Equations (9) and (10), $E_{gi,t,k}$ and $E_{gi,j,t,k}$ are the exposure measures in the purchase timing and product choice models, while the raw or actual exposures based on 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 choose a logit-normal prior and jointly estimate these two parameters together with the other parameters of interest. We allow for heterogeneity in these parameters using the same hierarchical structure as we did with the other parameters.

3.2.4. Controlling for Remaining Endogeneity Using LIV. Homophily, influence, and exogenous shocks are the three key factors that determine decisions in social environments, as recognized in the literature since Manski (1993). All three factors are explicitly accounted for in our model. Nonetheless, to further control for a potent endogeneity concern on the influence effect, i.e., that the exposure variable may be correlated with the error term, we adopt a latent instrumental variable (LIV) setup (Ebbes et al. 2005).[2] LIV is an approach that is increasingly used in situations—like ours—where no observed instruments are available (Zhang et al. 2009, Rutz et al. 2012). It separates the observed covariates into two components; one is a systematic component that is not correlated with the residuals, while the other may be correlated with the residual. The former component is used as the instrument for the observed covariates. Specifically, we instrument the raw exposure measure for purchase and choice decisions in Equations (9) and (10) as follows:

$$\tilde{E}_{gi,t,k} = \varphi_k e_{gi,t,k} + \omega_{gi,t,k}, \tag{11}$$

$$\tilde{E}_{gi,j,t,k} = \phi_k e_{gi,j,t,k} + \omega_{gi,j,t,k}. \tag{12}$$

In Equation (11), $e_{gi,t,k}$, $k \in \{\text{In}, \text{Out}\}$ are the latent IVs for the raw exposure. As suggested in Ebbes et al. (2005), $e_{gi,t,k}$ is a binary discrete variable which is estimated through data augmentation. The error $\omega_{gi,t,k}$ is correlated with $\varepsilon_{gi,t}$ in Equation (2) to account for endogeneity:

$$\begin{pmatrix} \omega_{gi,t,\text{In}} \\ \omega_{gi,t,\text{Out}} \\ \varepsilon_{gi,t} \end{pmatrix} \sim \text{MVN}(0, \Sigma_\varepsilon). \tag{13}$$

In Equation (13), Σ_ε is a full-rank covariance matrix. The extent of endogeneity that leads to the correlation between exposure and the unobservable is reflected in the correlation terms of this matrix. We denote $r_{\text{In},pi}^{\text{LIV}}$ as the correlation parameter between $\omega_{gi,t,\text{In}}$ and $\varepsilon_{gi,t}$, and $r_{\text{Out},pi}^{\text{LIV}}$ between $\omega_{gi,t,\text{Out}}$ and $\varepsilon_{gi,t}$ in the covariance matrix.

Similarly, in Equation (12), $e_{gi,j,t,k}$, $k \in \{\text{In}, \text{Out}\}$ are the latent instrument variables for the raw exposures, which is operationalized as a binary discrete variable, while the error terms are correlated with $\zeta_{gi,j,t}$ in Equation (5) to account for potential endogeneity:

$$\begin{pmatrix} \omega_{gi,j,t,\text{In}} \\ \omega_{gi,j,t,\text{Out}} \\ \zeta_{gi,j,t} \end{pmatrix} \sim \text{MVN}(0, \Sigma_\zeta). \tag{14}$$

We denote $r_{\text{In},pc}^{\text{LIV}}$ as the correlation parameter between $\omega_{gi,j,t,\text{In}}$ and $\zeta_{gi,j,t}$, and $r_{\text{Out},pc}^{\text{LIV}}$ between $\omega_{gi,j,t,\text{Out}}$ and $\zeta_{gi,j,t}$ in the covariance matrix.

[2] We thank the review team for the suggestion of using the LIV approach.

4. Discussion of Our Data and Model

Getting quality data and properly leveraging it has been a major challenge in social network research. Our data set offers many advantages in addressing several common concerns related to homophily and social influence within a social network. Specifically, the presence of relevant communication data enables their separation. Given the complex nature of purchase decisions in a social network environment, however, we still have to make several simplifying assumptions in operationalizing our model with the available data. In this section we explain these challenges and how we have tried to mitigate these potential problems.

As discussed in §3, 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 and exogenous effect. While the characteristics of consumers such as product valuation remain stable overtime, the consumers are exposed to different levels of influence and exogenous shocks over time. Therefore, the effects of social influence and exogenous factors can be separated from homophily. Furthermore, social influence and exogenous shocks can be separated since the former is conditional on relevant communication, which we observe in our data set. Adequate time series variation of exposure and purchase allows the identification and estimation of the parameters.

The usual identification restrictions apply to the product choice model, i.e., the latent utilities are identified up to a constant. The product characteristic of one product is thus normalized to zero. For the time-specific fixed effect, the sum is restricted to zero for identification. 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 consumers' product choices. The trade-off between the purchase price and usage value will be encoded in the intrinsic purchase frequency parameter.

4.1. Defining Groups Within Our Social Network

In our model a group consists of people who have a close relationship with one another. We conjecture that 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 to one another in the first month of the three-month period. The choice of five phone calls as the threshold is a subjective one on our part and is chosen to achieve a balance.[3] 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., 10 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 necessary only to uncover the *existence* of connections, since we infer *strength* through our model's social influence effects induced by communication.

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 a close relationship. To identify groups in this network, we then run a clique-searching algorithm to find all four-cliques in the graph.[4] A four-clique is a subgraph with four nodes, where every node is connected to every other node in the subgraph. Each four-clique then corresponds to a group of four customers, where each customer has a close relationship with every other customer in the group.

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 more connections between members within the groups but fewer groups of larger size. If we choose a smaller size such as 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. Our algorithm extracted a total of 1,095 groups, i.e., four-cliques, from the data set. These 1,095 groups are used for estimation. Since the threshold and clique size are subjective decisions, we performed extensive analysis using alternative threshold levels and clique sizes. The results of analysis for those alternative specifications are reported in §5.5.[5]

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

[3] 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.

[4] 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.

[5] The choice of phone call thresholds and clique size could be nested within our model. However, this will lead to a more complex model that does not offer much additional insight on the main focus of our study, namely dissecting the source of similarity. Therefore, we chose this sensitivity analysis approach instead of pursuing an extended model.

be a subset of a larger group where all are pairwise connected. Hence, our cliques of size four can be considered samples of larger clique sizes. Notice that this subclique does not raise concern for evaluating the homophily effect, which requires the existence of connections but does not require exclusive connections. If 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. This embedding issue, however, can be 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 toward the in-group influence, making it harder to find a 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 another 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 latent homophily effect that we wish to uncover and distinct from endogenous group formation.

4.2. An Exploratory Analysis of Decision Similarity Amongst Group Members

Behavior similarity among people who are connected is well documented in the literature (e.g., Aral et al. 2009, Hill et al. 2006). Nonetheless, we would like to verify that it exists in our data set through an exploratory data analysis in order to assess the adequacy of our data set for our research task. Tables 2 and 3 demonstrate such similarity in both purchase incidence decisions and product choice decisions. First we calculated the average number of tones purchased, which is 1.76. Next we calculated the average number of tones a consumer purchased conditional on having another member of the group purchase a certain number of tones. This result is reported in Table 2 and shows that as the number of purchases by a connected consumer goes up, the average purchase of the consumer also goes up. In other words, consumers who are connected to frequent purchasers are more likely to be frequent purchasers themselves. This is clear evidence of similarity in the purchase incidence decisions.

Table 2 Decision Similarity—Purchase Incidence

	Average purchase
Overall	1.76
If a group member purchased	
0 tone	1.33
1 tone	1.69
2 tones	2.00
3 tones	2.17
4 tones	2.26
5+ tones	2.83

The evidence of similarity in product choice decisions is presented in Table 3. Here, we calculated the probabilities of consumers choosing each product, i.e., genre. We calculated both the unconditional probability, i.e., the overall average, and the probability conditional on another group member choosing the same genre. As shown in the table, for all four genres the conditional probability is higher than the unconditional one. For genres Regional, Devotional, and Other, the conditional probability is almost twice the unconditional probability. The increase is smaller for the genre Bollywood, but that is likely because the unconditional probability is high to begin with for that genre. This is clear evidence of similarity in choice decisions. In summary, the exploratory analysis suggests that this data set is a good source for analyzing the underlying drivers of similar purchase decisions.

4.3. Discussion on Generalizability

Our data set is a good fit for separating homophily and influence effects, largely because the nature of the CRBT product makes a phone call or, more generally a communication record, a close proxy of product exposure among peers, and because the product causes the recipient to hear the message that removes concerns about alternative exposures to the product. Additionally, the product is inconsequential to the underlying relationships. We think it is reasonable to assume that calls are not made simply to hear someone else's ringtone, which minimizes the concern for endogenous group formation. Although situations like this, which help achieve clean identification for academic research, are special cases, our modeling framework can be generalized to a wider range of settings.

Table 3 Decision Similarity—Product Choice

Genre	No purchase	Purchase
Regional	0.13	0.22
Bollywood	0.33	0.47
Devotional	0.09	0.17
Other	0.14	0.22

Our model assumes that we observe communications about the product between consumers. With the proliferation of social media, such as Facebook and Twitter, data of this type are increasingly commonplace. A great amount of communication among consumers now takes place in social network websites and online discussion forums. Monitoring and analyzing such chatter has spawned a new industry, as firms are eager to find out what consumers say about their products. Although the communication that is observed may be only a part of the overall communication, it still can be used to evaluate the influence effect. In addition to revealing the actions of communication, social media also contains the content of the communication that can be analyzed, e.g., through sentiment analysis of the text. This provides the opportunity for more detailed analysis, such as distinguishing the informative and persuasive effect of peer communication, which is not in the scope of our study where both are accounted for as influence. A challenge in generalizing our model to other products is that product purchase and use is often not recorded with the communications data. This forces the analyst to collect two sets of data: communication and purchase, which may be difficult to gather or combine due to privacy concerns. Additionally, other products may experience endogenous group formation. Communication-related products such as Skype or products for "group consumption" such as social events may be more prone to endogeneous group formation, e.g., links are made because of the product.

5. Empirical Results

There are 91 days in the entire data set. We use the first 10 days to initialize the exponentially smoothed exposures, the next 60 days for estimation, and the final 21 days as a holdout sample for predictive evaluation. We use the Markov chain Monte Carlo (MCMC) method to draw parameters from their posterior distributions. The conditional distributions are available in a technical report available from the authors. For the estimation, we took 40,000 MCMC draws, with the first 20,000 discarded as burn-in draws and the remaining 20,000 used for evaluation.

5.1. Purchase Incidence

The posterior mean, standard deviation, and 95% confidence interval of parameters for the purchase incidence decisions are reported in Table 4. The population level mean purchase parameter is 0.0129, corresponding to a mean baseline purchase frequency of once every 155 days. The in-group influence parameter is 1.660, while the out-group influence parameter is 1.333. Both are positive and statistically significant.[6] This shows that a strong influence effect exists in purchase timing decisions. The estimates show that exposure to a tone change by someone outside the group can increase the purchase probability by almost three times (279%), a quite significant increase. Even more substantial is the effect of an exposure to tone purchase by someone inside the group, which, as the estimate suggests, increases the purchase probability by about four-fold (426%). The higher in-group influence than out-group influence is reasonable, suggesting that customers are more susceptible to people close to them, although the difference is not statistically significant.

Table 4 Parameter Estimate—Purchase Incidence

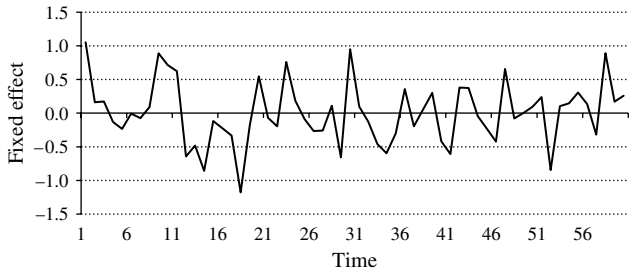
	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\lambda}$	0.0129	0.0016	0.0119	0.0153
$\bar{\gamma}_{\text{In}}$	1.660	0.111	1.411	1.839
$\bar{\gamma}_{\text{Out}}$	1.333	0.114	1.055	1.460
σ^2_{λ}	0.690	0.054	0.540	0.782
$\sigma^2_{\gamma_{\text{In}}}$	0.323	0.165	0.192	0.922
$\sigma^2_{\gamma_{\text{Out}}}$	0.521	0.270	0.214	1.094
r_{λ}	−0.005	0.029	−0.061	0.057
$r_{\gamma_{\text{In}}}$	0.039	0.164	−0.190	0.349
$r_{\gamma_{\text{Out}}}$	0.475	0.196	0.139	0.754
σ^2_{ν}	0.438	0.055	0.345	0.554
$\bar{\kappa}^{pi}$	0.773	0.006	0.762	0.782
$r^{\text{LIV}}_{\text{In},\,pi}$	−0.101	0.055	−0.173	0.002
$r^{\text{LIV}}_{\text{Out},\,pi}$	−0.180	0.105	−0.332	0.001

The in-group correlation on the purchase rate parameter is −0.005, which is statistically indistinguishable from zero. There is thus no evidence that customers in the same group tend to have similar intrinsic purchase frequencies. The correlation on the in-group influence parameter is 0.039, while that on the out-group influence parameter is 0.475, the latter of which is statistically significant. This shows that group members have somewhat similar levels of susceptibility to influence. The parameter estimates thus show that homophily effect is not evident in the intrinsic purchase rate, but it does exist on the susceptibility to influence. In other words, the observed similarity in purchase frequency among group members result first from their influence on one another, and second from their similarity in susceptibility to influence, instead of from intrinsic similarity in their purchase frequencies.

Finally, the estimate of the time-specific fixed effect is shown in Figure 1, and the variance of the group and time-specific random effect is estimated to be 0.438. The time series of the estimated fixed effects and the time series of the total purchase count by time have a correlation of 0.847. Meanwhile, the estimates imply that a group-specific random shock at the size of one standard deviation will almost double the instantaneous purchase probability, suggesting that groups are also

[6] By statistically significant, we mean that the parameter's mean-centered 95% posterior interval does not include zero. The same terminology is used for the rest of the discussion in this paper.

Figure 1 Time-Specific Fixed Effect for Purchase Incidence



subject to sizeable exogenous shocks in their purchase decisions. Taken together, these estimates show that exogenous factors play an important role in consumers' purchase incidence decisions, alongside social influence and latent homophily. Finally, the two LIV correlation parameters are both slightly negative but show only borderline statistical significance. This suggests that after the three major factors (social influence, latent homophily, and time effects) are accounted for, there is little remaining concern about endogeneity in our data set.

5.2. Product Choice

The parameters for the product choice model are reported in Table 5. The mean product taste intercept parameters for the music genres are −0.565, 1.727, and −1.020 for Regional, Bollywood, and Devotional, respectively. These values are consistent with the market shares of these music genres. The in-group influence parameter is 0.034 but is statistically indistinguishable from zero, so there is no conclusive evidence of in-group influence on music genre choices. The out-group influence parameter is 0.581 and statistically significant, meaning that exposure to a tone from someone outside the group has fairly strong positive effect on choosing the same category.

The result that the out-group influence is higher and more strongly evident than the in-group influence is surprising. One explanation is that influence on product choice may have two competing effects. A person's product choice may be positively influenced by a friend because the person trusts the friend's selection. On the other hand, consumers may wish to be distinct and exhibit some variety seeking behavior. Upon hearing a devotional tone from a friend, for example, a consumer might decide to choose a movie tone just to be different, even though she may like the tone that her friend has. 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 also lends support to this explanation. If some consumers want to imitate while others seek variety, then a wider dispersion of the in-group influence parameter is expected,

Table 5 Parameter Estimate—Product Choice

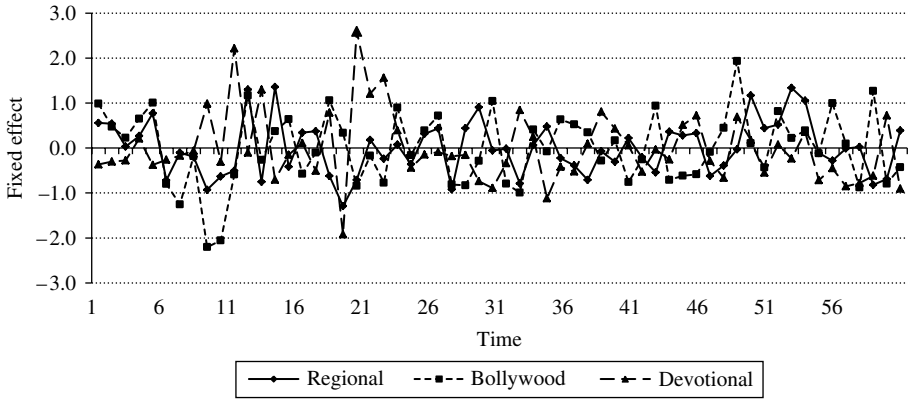
Parameter	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\beta}_1$	−0.565	0.090	−0.724	−0.392
$\bar{\beta}_2$	1.727	0.113	1.546	1.969
$\bar{\beta}_3$	−1.020	0.102	−1.217	−0.849
$\bar{\rho}_{\text{In}}$	0.034	0.076	−0.115	0.179
$\bar{\rho}_{\text{Out}}$	0.581	0.063	0.448	0.687
$\sigma^2_{\beta_1}$	1.621	0.293	1.098	2.201
$\sigma^2_{\beta_2}$	1.167	0.129	0.942	1.419
$\sigma^2_{\beta_3}$	0.620	0.081	0.478	0.866
$\sigma^2_{\rho_{\text{In}}}$	0.977	0.224	0.596	1.376
$\sigma^2_{\rho_{\text{Out}}}$	0.527	0.068	0.399	0.679
r_{β_1}	0.612	0.071	0.461	0.720
r_{β_2}	0.247	0.107	0.011	0.444
r_{β_3}	0.358	0.165	0.071	0.598
$r_{\rho_{\text{In}}}$	0.301	0.111	0.033	0.458
$r_{\rho_{\text{Out}}}$	0.183	0.183	−0.075	0.519
$\sigma^2_{\eta 1}$	0.633	0.348	0.261	1.384
$\sigma^2_{\eta 2}$	1.619	0.723	0.502	3.060
$\sigma^2_{\eta 3}$	0.671	0.196	0.268	1.001
$\bar{\kappa}^{pc}$	0.596	0.016	0.571	0.627
$r^{\text{LIV}}_{\text{In},pc}$	0.098	0.059	−0.029	0.176
$r^{\text{LIV}}_{\text{Out},pc}$	−0.030	0.037	−0.109	0.027

thus a higher estimate of the variance. Indeed, the result in Table 5 shows that the estimated variance of the in-group influence parameter, 0.977, is much higher than that of the out-group influence parameter, 0.527. However, we want to emphasize that this is only indirect evidence. Explicit separation of the influence effects to the opposite directions is more desirable, a topic we leave for future research.

The in-group correlations for the three product taste intercept parameters are 0.612, 0.247, and 0.358, respectively, all statistically significant. This provides evidence that significant similarity exists on product tastes of customers who are close to one another, consistent with the expectation that the latent homophily effect exists and plays a major role in product choice decisions. Furthermore, the in-group correlations for the in-group influence and out-group influence parameters are 0.301 and 0.183, respectively, with the former being 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 friends are also likely influenced by others. This is further evidence of homophily in the product choice decision.

The time-specific fixed effects for each genre are plotted in Figure 2. The time series of the fixed effects for each genre and the time series of the total sales by genre have correlations of 0.492, 0.395, and 0.795, for

Figure 2 Time-Specific Fixed Effect for Product Choice



Regional, Bollywood, and Devotional tones, respectively, thus population-level exogenous factors play an important role in choice decisions. Meanwhile, the variances of the group and time-specific random effects are estimated to be 0.633, 1.619, and 0.671 for Regional, Bollywood, and Devotional tones, respectively. This shows significant exogenous shocks exist at the group level. Delving into more details of these estimates, we can see that Bollywood tones are subject to the highest amount of group-level exogenous shocks. This is understandable, as going to a movie together may be a significant source of exogenous stimulants at group level, and radio exposure to movie songs may also be larger than to the other genres. In contrast, devotional tones are influenced more by the time-specific fixed effect at the population level than do tones in the other genres, since the purchase and the fixed effect have the highest correlation, and the figure shows that the fixed effects have larger spikes for this genre. This is also reasonable, as religious dates of significance may play an important role in the purchase of tones in this genre. Weeks 11 and 21 both have larger spikes that correspond to significant religious holidays. Finally, the two LIV correlation estimates are close to zero, suggesting there is little evidence of endogeneity after all three factors are accounted for.

Taking the estimates together, we can see that the observed choice similarity is mostly due to latent homophily for the regional genre, due to a combination of latent homophily and group-level exogenous shocks for the Bollywood genre, and mainly due to population-level exogenous factors and to less extent latent homophily for the devotional genre. An understanding at this level of richness illustrates the value of the model developed in this study.

In summary, we find influence effects in both purchase incidence and product choice decisions, and a strong homophily effect in product choice decision as well. On purchase incidence decisions we find that the in-group influence is higher than the out-group influence, while the reverse is true on product choice decisions. Latent homophily exists on product tastes and susceptibility to influence, but is not evident on the intrinsic purchase rate. In both purchase incidence and product choice decisions, exogenous factors exist at both the population level and the group level. These exogenous factors combined with homophily and influence effects give us a rich understanding of the nuanced underlying drivers of choice similarity. Separating the latent homophily, social influence, and exogenous effects is the primary focus of this study. The results discussed here validate the significance of all these factors and confirm the importance of integrating them into a single model for comprehensive and accurate assessment, as we do in this study.

5.3. Model Comparison

We estimate two special cases of the model we proposed in §3. The first model assumes there is no latent homophily in the baseline purchase rates, product taste intercepts, and the susceptibility to influence parameters. We refer to this model as the "influence-only" model. In this model, consumers may still influence one another through communications, but the in-group correlations are assumed to be zero for all parameters. 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 taste intercepts, but the influence parameters are all assumed to be zero. In these two models, exogenous factors are modeled in the same way as in our proposed model.

The alternative models are estimated using an MCMC algorithm similar to the one used for our full model. The results are reported in Tables 6–9, respectively. For the influence-only model, the estimates of the purchase incidence parameters are close to those in the full model. This is as expected, since in our proposed model the estimates show little evidence of

Table 6 Parameter Estimate—Purchase Incidence—Homophily Only

	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\lambda}$	0.0166	0.0016	0.0155	0.0191
σ^2_λ	0.803	0.056	0.671	0.894
r_λ	0.112	0.027	0.057	0.162
σ^2_ν	0.427	0.057	0.346	0.529

homophily on baseline purchase frequency. For the product choice decision, we can see that the in-group influence parameter has a posterior mean of 0.317 and is statistically significant in the influence-only model. This is biased, as the same parameter is close to zero in the proposed model. This upward bias, however, is exactly as expected: The estimates of the proposed model show strong homophily effects on product tastes, which result in choice similarity among connected consumers. When this latent homophily is assumed away, however, the effect will be mistakenly attributed to social influence, hence the much larger estimate for the in-group influence parameter.

Turning to the homophily-only model, we find that for the purchase incidence decision, the estimate of the intrinsic purchase frequency parameter is 0.0166. This estimate is slightly higher than the estimate in the full model (0.0129) and that in the influence-only model (0.0114). This shows that when the influence effect is not accounted for, the model overestimates the intrinsic purchase frequency of consumers because some influence-induced purchases are now considered spontaneous. More importantly, we find that for the homophily-only model, the correlation of the baseline purchase frequency is estimated to be 0.112 and statistically significant. Contrast this with the estimate of close to zero for the same parameter in the proposed model, and we can see that ignoring influence effect will result in overestimation of homophily effect in the purchase incidence decisions. For the product choice decisions, the estimates for all the homophily parameters, i.e., the in-group correlation of the product taste intercepts, are higher in the homophily-only model than in the full model (0.710/0.626/0.374 for the homophily-only model, versus 0.612/0.247/0.358 for the proposed model). This shows that if the influence effect among friends in choice decisions is ignored, the model overestimates the taste similarity among consumers who are connected. Again, this highlights the importance of separating the two effects of homophily and social influence.

We also compare the model fit of the proposed model with that of the homophily-only and influence-only models. We use both the in-sample log-marginal density (Newton and Raftery 1994, Chib 1995) and the log-likelihood for the hold-out sample. The measures are reported in Table 10. As expected, on both measures the full model outperforms both the homophily-only and the influence-only models.

Table 7 Parameter Estimate—Purchase Incidence—Influence Only

	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\lambda}$	0.0114	0.0016	0.0107	0.0126
$\bar{\gamma}_{\text{In}}$	1.571	0.114	1.346	1.757
$\bar{\gamma}_{\text{Out}}$	1.155	0.079	0.993	1.302
σ^2_λ	0.691	0.052	0.587	0.769
$\sigma^2_{\gamma_{\text{In}}}$	0.400	0.178	0.220	0.783
$\sigma^2_{\gamma_{\text{Out}}}$	0.454	0.161	0.247	0.960
σ^2_ν	0.483	0.068	0.343	0.578
$\bar{\kappa}^{pi}$	0.749	0.010	0.733	0.773

Table 8 Parameter Estimate—Product Choice—Homophily Only

Parameter	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\beta}_1$	−0.583	0.116	−0.788	−0.390
$\bar{\beta}_2$	1.900	0.078	1.748	2.046
$\bar{\beta}_3$	−1.107	0.085	−1.268	−0.946
$\sigma^2_{\beta_1}$	1.226	0.157	0.933	1.539
$\sigma^2_{\beta_2}$	1.174	0.145	0.942	1.475
$\sigma^2_{\beta_3}$	1.520	0.302	0.978	2.026
r_{β_1}	0.710	0.105	0.506	0.852
r_{β_2}	0.626	0.079	0.483	0.757
r_{β_3}	0.374	0.152	0.179	0.650
$\sigma^2_{\eta 1}$	0.533	0.242	0.288	1.211
$\sigma^2_{\eta 2}$	2.834	0.948	1.239	5.114
$\sigma^2_{\eta 3}$	0.384	0.082	0.232	0.556

We also compare the model fit of the proposed model with that of the homophily-only and influence-only models. We use both the in-sample log-marginal density (Newton and Raftery 1994, Chib 1995) and the log-likelihood for the hold-out sample. The measures are reported in Table 10. As expected, on both measures the full model outperforms both the homophily-only and the influence-only models.

Table 9 Parameter Estimate—Product Choice—Influence Only

Parameter	Posterior mean	Posterior std. dev.	2.5% posterior quantile	97.5% posterior quantile
$\bar{\beta}_1$	−0.329	0.067	−0.493	−0.218
$\bar{\beta}_2$	1.764	0.123	1.540	1.997
$\bar{\beta}_3$	−1.169	0.075	−1.309	−1.027
$\bar{\rho}_{\text{In}}$	0.317	0.070	0.166	0.436
$\bar{\rho}_{\text{Out}}$	0.688	0.065	0.544	0.785
$\sigma^2_{\beta_1}$	0.286	0.099	0.124	0.483
$\sigma^2_{\beta_2}$	2.641	0.505	1.950	3.681
$\sigma^2_{\beta_3}$	1.169	0.241	0.831	1.649
$\sigma^2_{\rho_{\text{In}}}$	0.805	0.144	0.593	1.054
$\sigma^2_{\rho_{\text{Out}}}$	0.891	0.123	0.630	1.124
$\sigma^2_{\eta 1}$	0.392	0.108	0.235	0.650
$\sigma^2_{\eta 2}$	7.899	1.417	5.403	10.290
$\sigma^2_{\eta 3}$	0.514	0.125	0.326	0.802
$\bar{\kappa}^{pc}$	0.533	0.051	0.438	0.618

Table 10 Model Fit Comparison

	Proposed model	Homophily-only	Influence-only
Calibration: Log-marginal density	−30,166.1	−32,263.4	−31,107.8
Hold-out: Log-likelihood	−8,564.7	−8,584.0	−8,907.0

5.4. 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 are thus crucial for accurate assessment of influence effect. Detailed communication data, however, are not commonly 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. 2010, Bell and Song 2007, Iyengar et al. 2011), with an implicit assumption that one's action is perfectly observed by others.[7]

Since our data set contains detailed communication data, we are able to evaluate its importance in accurately measuring the influence effect. This is done using an alternative "no-communication" model. In this case, we ignore the phone calls that 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 the purchase incidence decision, whenever a consumer purchases a tone, we consider that all others in her group are exposed to it. For the product choice decision, we evaluate two alternative specifications. In the first, denoted as no-communication model I, we consider there is exposure only when the consumer newly purchases a tone. In the second, denoted as no-communication model II, we consider there is one count of exposure every day to others in the group as long as a consumer possesses a tone. The former specification likely leads to understated exposure because each purchase is counted as exposure only once, even though in reality the friends likely have multiple phone calls over time. The latter specification likely leads to overstated exposure, as exposure is counted every day as long as a tone is in use, yet friends may not call each other that frequently. Both models feature mistimed exposures, again because the purchase or possession of a tone need not coincide with phone calls. Notice that we can no longer evaluate out-group influence in these two alternative models, since the only knowledge source for that is the communication data.

Table 11 "No-Communication" Model—Purchase Incidence

Parameter	Proposed model	"No-communication" model
$\bar{\lambda}$	0.0129 (∗)	0.0148 (∗)
$\bar{\gamma}_{\ln}$	1.660 (∗)	0.180 (∗)
r_{λ}	−0.005	0.060 (∗)
$r_{\gamma_{\ln}}$	0.039	0.772 (∗)

Table 12 "No-Communication" Model—Product Choice

Parameter	Proposed model	"No-communication" model I	"No-communication" model II
$\bar{\beta}_1$	−0.565 (∗)	−0.475 (∗)	−0.458 (∗)
$\bar{\beta}_2$	1.727 (∗)	2.042 (∗)	1.977 (∗)
$\bar{\beta}_3$	−1.020 (∗)	−0.818 (∗)	−1.286 (∗)
$\bar{\rho}_{\ln}$	0.034	0.320 (∗)	0.163 (∗)
r_{β_1}	0.612 (∗)	0.803 (∗)	0.769 (∗)
r_{β_2}	0.247 (∗)	0.553 (∗)	0.495 (∗)
r_{β_3}	0.358 (∗)	0.280 (∗)	0.037

The estimation results for purchase incidence and product choice decisions are reported in Tables 11 and 12, respectively, in comparison with the corresponding results of the proposed model. The in-group influence parameter for purchase incidence is much smaller in the no-communication model than in the proposed model: 0.180 versus 1.660. This is as expected: Without communication data, a tone change is considered an exposure whether or not it is communicated to others. This results in overstating the amount of exposure and lowers the per-communication influence effect. Furthermore, the timing of exposure is not accurate. For example, if consumer A purchased a tone on day 1, and consumer B called A on day 3, the true exposure happens on day 3, but without communication data, it is counted on day 1 in the no-communication model. This further dampens the influence effect. Ignoring the communication data thus leads to significantly underestimated influence effect for the purchase incidence decision. Meanwhile, the estimate of the homophily parameter on the baseline purchase frequency is positive and statistically significant in the no-communication model, whereas it is close to zero in the proposed model: 0.060 versus −0.005. This is again as expected: without communication data, exposure is mistimed, so influence-induced purchase can be mistakenly treated as intrinsic and considered as the result of homophily.[8]

For the no-communication model I of the product choice decision, where an exposure is counted at the

[7] One may consider this a reasonable assumption if the data are at aggregate level, due to the law of large numbers. Alternatively, if each time period is sufficiently long, so that with high confidence one's decision has been conveyed to others in the time period, this assumption can also be considered reasonable. In any case, this should be considered a strong assumption by default unless specific circumstances make it reasonable.

[8] Recall that in the discussion for identification, asynchronous purchase that is independent of communication indicates homophily, while purchase that depends on communication indicates influence.

time of tone purchase, the estimated influence effect is 0.320 and statistically significant, as reported in Table 12. This shows that having the purchase decision directly enter into the utility functions leads to overestimated influence effect. This is understandable—since exposure is understated in this model, the unit effect of one exposure is likely overestimated. The estimated homophily parameters are also higher than in the proposed model (0.803/0.553/0.280 in the no-communication model I versus 0.612/0.247/0.358 in the proposed model). This is also expected; since communication data are not available, decisions that are actually driven by influence will be mistakenly treated as driven by the intrinsic preference, which will lead to overestimated similarity, hence overestimated homophily effect. The estimated influence effect for the no-communication model II is 0.163 and statistically significant. This suggests that treating the possession of a tone as exposure also leads to overestimated influence effect (0.163 here versus 0.034 in the proposed model where it is not statistically significant). To understand this, note that in the no-communication model II, one count of exposure is assumed for every day after a tone is purchased until it is changed. This increases the chance of picking up the effects of latent homophily or group-level exogenous factors as the result of influence.

The results of these analyses clearly show that, in order to accurately attribute similarity to latent homophily, social influence, and exogenous factors, it is crucial to have communication data. Using decision variables as proxies for true exposures can lead to seriously biased results, since the former differ from the latter in both quantity and timing. Care must be taken to make such assumptions in studies, and their reasonableness should be scrutinized in the specific contexts.

5.5. Robustness Check

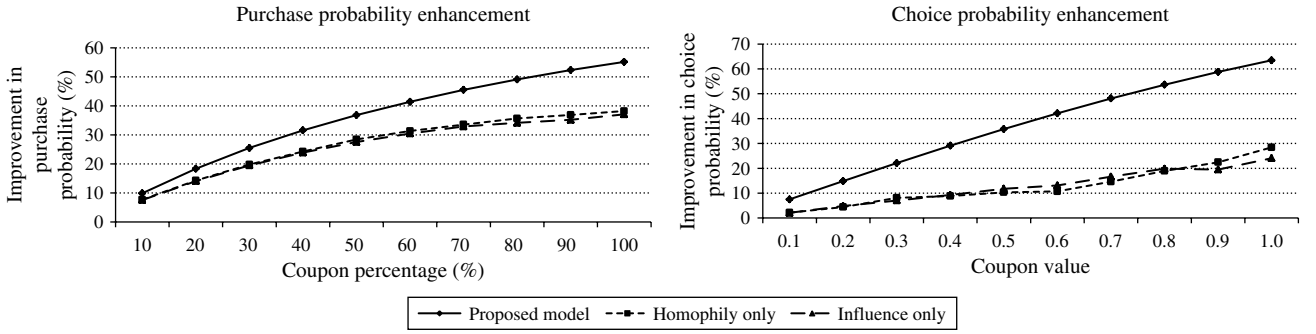
As discussed in §4, in preparing the data we used a threshold of five calls to find groups of four consumers. The selection of call threshold and group size is meant to strike a balance between detecting real relationships balanced against a large enough sample. Care must be taken, however, to ensure that the effects shown in the estimates are robust and not driven by our choice of the call threshold and group size. Considering this, we repeated the estimation using a series of other values of call threshold and group size and provide results in a technical report that is available from the authors. Our findings show that across all these specifications the purchase incidence decision is subject to a strong influence effect, that there is no evidence of homophily on the intrinsic purchase frequency, but the homophily effect does exist on the susceptibility to influence parameters. For the product choice decision, we find across all specifications that there is a strong homophily effect on product tastes, that there is a strong out-group influence effect but the in-group influence effect is not evident, that there is latent homophily on the susceptibility to influence parameters also, and that a certain genre, namely Bollywood, is subject to more group-level exogenous shock than others. In summary, our results are robust to different values of call threshold and group size.

6. Policy Simulation

Understanding of latent homophily and social influence effects has important managerial implications because managers can use such knowledge to improve their decision making as they promote products to consumers. Multiple opportunities exist. First, by accounting for homophily, influence, and exogenous effects, we arrive at a more accurate understanding of the drivers of consumers' purchase decisions than existing models do. Our hierarchical Bayesian framework allows for recovering parameters at individual consumer level. Accordingly, we can determine which consumers are best suited for targeting. Intuitively, a consumer who is intrinsically interested in a product and who is susceptible to influence is a good target. Second, when targeting a consumer, her friends may also become more likely to purchase due to influence. This creates a "multiplier effect," and our parameter estimates allow us to evaluate the extent of such effect, and find out the consumers with whom such effect is maximized. Finally, through understanding of the homophily effect, we can gain a better understanding of new consumers who have never purchased from the firm, as long as we observe their connections with existing consumers.

To assess these opportunities, we perform three policy simulations: targeting existing consumers, evaluating multiplier effects, and targeting new consumers. We perform the simulations using the same 1,095 group of consumers used for model estimation, as we rely on the individual consumer level parameter estimates for targeting. Since we do not observe any price variation in CRBT and are unable to estimate price response, our simulations revolve around a hypothetical coupon drop, which serves as a marketing vehicle that hypothetically influences purchase intention through price reductions targeted to the coupon's recipient. Specifically, we evaluate the effect of distributing 100 coupons to selected customers in those 1,095 groups to maximize coupon effectiveness. The choice of 100 is to ensure decent coverage of the sample consumers, while at the same time to keep it selective enough so efficiency matters. Promotion through social networks has long been of interest to firms and has drawn attention of recent research (e.g., Bapna and Umyarov 2012). Our policy simulation here also adds to this discussion.

Figure 3 Policy Simulation for Purchase and Choice Probability Enhancement



6.1. 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.[9] To do so, we evaluate the effect of distributing a total of 100 coupons to selected consumers in groups used for our model estimation. Each coupon is assumed to increase a customer's intrinsic purchase rate by c_p%. Meanwhile, it will increase a customer's base utility of one specific category of tones by c_c.[10] Without loss of generality, we treat the first category as the target category of promotion.

We conducted the policy simulation based on the individual consumer-level parameter estimates of the proposed model, the homophily-only model, and the influence-only model, of the consumers in our data set. For each consumer, we simulated two scenarios. In the first, a coupon is delivered to the consumer on day one; in the second, no coupon is delivered. For each scenario, we simulated the subsequent purchase and choice probabilities. The coupons are delivered to the 100 consumers for whom the increases in purchase probability when they receive the coupon are the largest. Note it is the *difference* in probabilities between these two scenarios that represents the coupon effectiveness, not the purchase probability when a coupon is delivered, since the consumer might have purchased even without the coupon. 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 simulate the purchase probability for a week as the influence effect will largely diminish after that according to our estimates.

Table 13 Policy Simulation 1—Targeting Existing Customers

Measure	Proposed model (%)	Homophily only (%)	Influence only (%)
Purchase probability improvement	36.79	28.41	27.56
Product choice probability improvement	35.78	10.33	11.78

Note. Coupon configuration: $c_p = 50\%$, $c_c = 0.5$.

In this simulation, we consider the change in purchase probabilities only for the focal consumers. 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 3. Using $c_p = 50\%$ and $c_c = 0.5$ as an example, the result of which is reported in Table 13, targeting based on parameter estimates from the proposed model increases purchase probability by 36.79%, as compared with 28.41% of the homophily only model and 27.56% of the influence only model. This represents a 29.5% and 33.5% improvement through recognizing both homophily and influence effects, compared with recognizing only one of them, respectively. The improvement in the probability of choosing the first category is 35.78% for the proposed model, and 10.33% and 11.78% for the homophily-only and influence-only model, respectively. On average, the full model performs 33.7% better than the homophily-only model and 37.3% better than the influence-only model on enhancing purchase probability, and it performs more than twice as well as the two models on enhancing product choice probability. These results show that consumer targeting can be improved by recognizing both homophily and influence effects.

6.2. The Multiplier Effect

In our second simulation, instead of evaluating the improvement that comes directly from the targeted consumers, we are interested in the effect that arises from their communication with their friends, i.e., a

[9] 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.

[10] 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 manner as a surrogate.

Table 14 Policy Simulation 2—Multiplier Effect

Measure	Proposed model (%)	Homophily only	Influence only (%)
Purchase probability improvement	18.85	NA	12.12

Note. Coupon configuration: $c_p = 50\%$, $c_c = 0.5$.

"multiplier effect." We again evaluate the effect of distributing 100 coupons, 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 14. As the table shows, the improvement in purchase probability is 18.85% for the proposed model, and 12.12% for the influence only model. This represents a 55.5% improvement in performance from the multiplier effect when recognizing both effects over recognizing only the influence effect. Note that the homophily-only model is not evaluated for this simulation, since the model by design assumes away peer influence.

6.3. Targeting New Customers

Finally, we evaluate the effectiveness of targeted 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 (the new customers). The objective of the firm is to increase the purchase probability in a certain time period by targeting these new customers.

We again distribute 100 coupons with $c_p = 50\%$. The result is reported in Table 15. The average increase in purchase probability for the full model is 5.43%, while those for the homophily-only and influence-only models are 5.22% and 4.79%, 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 5%, compared with a better than 30% improvement when applied to existing customers. 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. When targeting these new customers, we note that the proposed model also has the best relative performance, with 4.0% and 13.4% better performance than the homophily-only model and the influence-only model, respectively. This suggests that the 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 more cost effectively than targeting everyone.

Table 15 Policy Simulation 3—Targeting New Customers

Measure	Proposed model (%)	Homophily only (%)	Influence only (%)
Purchase probability improvement	5.43	5.22	4.79

Note. Coupon configuration: $c_p = 50\%$, $c_c = 0.5$.

7. Conclusion

Social network researchers have long recognized the importance of both homophily and social influence. Homophily is the concept that similar people are more likely to form ties, and thus people who are connected are likely to have similar product tastes and other traits. A variety of terms have been used to characterize social influence, such as peer influence, interaction effect, imitation, conformity, contagion, etc. All share the key feature that one consumer's decision is potentially altered through her communications with others in her social network. 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 the 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 customer, 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 necessary for effective marketing strategies, but it is also challenging. Although extant research has controlled for observed homophily while evaluating influence effect in purchase decisions, it does not address the separation of latent homophily and social influence. In addition, existing methods usually enable the quantification of only the influence effect but not homophily. Finally, due to the lack of communication information, extant research can only evaluate influence as the dependence of one's decision on another's decision instead of on their communication. This is a strong assumption, the implication of which has not been explored.

Enabled by a unique electronic social network data set, we investigate the role of both latent homophily and social influence in consumers' purchase decision process, while at the same time accounting for exogenous factors. Our study contributes to the literature by simultaneously evaluating both latent homophily and social influence effects in product purchase decisions, by doing so in a manner that both effects can be

quantified and so can directly advise managers on targeting strategies, and by investigating the importance of communication data in measuring social influence 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 data set, which contains both communication and product purchase information over time. With the advent of online and social media, these data are likely to become more common.

We find a strong social influence effect in purchase timing and product choice decisions. We further find that in-group influence is stronger than out-group influence for the purchase incidence decision, while the latter is more salient in the product choice decision. We find a strong homophily effect on product tastes and the susceptibility to influence. We show that models that ignore one of the factors result in the overestimation of the other factor, and that ignoring communication data leads to biased influence estimates. Furthermore, through policy simulation we show that understanding these factors improves target marketing performance.

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 the 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 with a compatible data set is needed to further isolate these two effects. Overall, understanding consumers within their social networks can lead to better managerial decision making in electronically enabled social networks.

Acknowledgments

The authors are affiliated with the iLab at Carnegie Mellon University's Heinz College and acknowledge the support of the iLab at Heinz College and an anonymous project sponsor for providing the data used in this project. The authors also acknowledge the support of the Wharton Customer Analytics Initiative. All opinions expressed in this paper are the authors' own and do not reflect those of the project sponsor.

References

Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. *Proc. 14th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (ACM, New York), 7–15.

Ansari A, Koenigsberg O, Stahl F (2011) Modeling multiple relationships in social networks. *J. Marketing Res.* 48(4):713–728.

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

Bapna R, Umyarov A (2012) Are paid subscriptions on music social networks contagious? A randomized field experiment. Working paper, National Bureau of Economic Research, Cambridge, MA.

Bass FM (1969) A new product growth model for consumer durables. *Management Sci.* 15(1):215–27.

Bell DR, Song S (2007) Neighborhood effects and trial on the Internet: Evidence from online grocery retailing. *Quant. Marketing Econom.* 5(4):361–400.

Bernheim BD (1994) A theory of conformity. *J. Political Econom.* 102(5):841–877.

Bott H (1928) Observation of play activities in a nursery school. *Genetic Psych. Monographs* 4(1):44–88.

Braun M, Bonfrer A (2011) Scalable inference of customer similarities from interactions data using dirichlet processes. *Marketing Sci.* 30(3):513–531.

Chib S (1995) Marginal likelihood from the Gibbs output. *J. Amer. Statist. Assoc.* 90(432):1313-1321.

Chintagunta PK (1993) Investigating purchase incidence, brand choice and purchase quantity decisions of households. *Marketing Sci.* 12(2):184–208.

Choi J, Hui SK, Bell DR (2010) Spatiotemporal analysis of imitation behavior across new buyers at an online grocery retailer. *J. Marketing Res.* 47(1):75–89.

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

Ebbes P, Wedel M, Böckenholt U, Steerneman T (2005) Solving and testing for regressor-error (in)dependence when no instrumental variables are available: With new evidence for the effect of education on income. *Quant. Marketing Econom.* 3(4):365–92.

Godes D, Mayzlin D, Chen Y, Das S, Dellarocas C, Pfeiffer B, Libai B, Sen S, Shi M, Verlegh P (2005) The firm's management of social interactions. *Marketing Lett.* 16(3–4):415–428.

Gönül F, Srinivasan K (1993) Modeling multiple sources of heterogeneity in multinomial logit models, methodological and managerial issues. *Marketing Sci.* 12(3):213–229.

Guadagni PM, Little JDC (1983) A logit model of brand choice calibrated on scanner data. *Marketing Sci.* 2(3):203–208.

Gupta S (1991) Stochastic models of interpurchase time with time-dependent covariates. *J. Marketing Res.* 28(1):1–15.

Hartmann WR (2010) Demand estimation with social interactions and the implications for targeted marketing. *Marketing Sci.* 29(4):585–601.

Hartmann WR, Nair H, Manchanda P, Bothner M, Dodds P, Godes D, Hosanagar K, Tucker C (2008) Modeling social interactions: Identification, empirical methods and policy implications. *Marketing Lett.* 19(3–4):287–304.

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

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

Jackson MO (2003) A survey of models of network formation: Stability and efficiency. *Group Formation in Economics: Networks, Clubs, and Coalitions* (Cambridge University Press, Cambridge, UK).

Jackson MO, Watts A (2002) The evolution of social and economic networks. *J. Econom. Theory* 106(2):265–295.

Judd S, Kearns M, Vorobeychik Y (2010) Behavioral dynamics and influence in networked coloring and consensus. *Proc. Natl. Acad. Sci. USA* 107(34):14978–14982.

La Fond T, Neville J (2010) Randomization tests for distinguishing social influence and homophily effects. *Proc. 19th Internat. Conf. World Wide Web* (ACM, New York), 601–610.

Leenders RTAJ (2002) Modeling social influence through network autocorrelation: Constructing the weight matrix. *Soc. Networks* 24(1):21–47.

Lyons R (2011) The spread of evidence-poor medicine via flawed social-network analysis. *Statistics, Politics, Policy* 2(1):Article 2, http://www.bepress.com/spp/vol2/iss1/2.

Manski CF (1993) Identification of endogenous social effects: The reflection problem. *Rev. Econom. Stud.* 60(3):531–542.

McPherson JM, Smith-Lovin L (1987) Homophily in voluntary organizations: Status distance and the composition of face-to-face groups. *Amer. Sociol. Rev.* 52(3):370–379.

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

Nair H, Manchanda P, Bhatia T (2010) Asymmetric peer effects in physician prescription behavior: The role of opinion leaders. *J. Marketing Res.* 47(5):883–895.

Newton MA, Raftery AE (1994) Approximate Bayesian inference by the weighted likelihood bootstrap. *J. Roy. Statist. Soc. Ser. B* 56(1):3–48.

Rutz OJ, Bucklin RE, Sonnier GP (2012) A latent instrumental variables approach to modeling keyword conversion in paid search advertising. *J. Marketing Res.* 49(3):306–319.

Shalizi CR, Thomas AC (2011) Homophily and contagion are generically confounded in observational social network studies. *Sociol. Methods Res.* 40(2):211–239.

Snijders TAB, Bunt GGV, Steglich C (2010) Introduction to stochastic actor-based models for network dynamics. *Soc. Networks* 32(1): 44–60.

Steeg GV, Galstyan A (2010) Ruling out latent homophily in social networks. *NIPS Workshop on Social Computing*. http://mlg.cs.purdue.edu/lib/exe/fetch.php?id=schedule&cache=cache&media=machine_learning_group:projects:paper19.pdf.

Steglich C, Snijders TAB, Pearson M (2010) Dynamic networks and behavior: Separating selection from influence. *Sociol. Methodology* 40(1):329–393.

Van den Bulte C, Joshi YV (2007) New product diffusion with influentials and imitators. *Marketing Sci.* 26(3):400–421.

Van den Bulte C, Lilien G (2001) Medical innovation revisited: Social contagion versus marketing effort. *Amer. J. Sociology* 106(5):1409–1435.

Van den Bulte C, Stremersch S (2004) Social contagion and income heterogeneity in new product diffusion: A meta-analytic test. *Marketing Sci.* 23(4):530–544.

Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. *J. Consumer Res.* 34(4):441–458.

Zhang J, Wedel M, Pieters R (2009) Sales effects of attention to feature advertisements: A Bayesian mediation analysis. *J. Marketing Res.* 46(5):669–681.

Zheng R, Wilkinson D, Provost F (2008) Social network collaborative filtering. Working Paper CeDER-8-08, Center for Digital Economy Research, Stern School of Business, New York University, New York. https://archive.nyu.edu/handle/2451/27735.