Homophily or Influence? – Analysis of Purchase Decisions in a Social Network Context

Liye Ma, Alan Montgomery and Ramayya Krishnan

iLab, Heinz College

Carnegie Mellon University

1. Introduction

Consider a firm selling products to consumers in a social network. The firm knows that friends in the network often make similar purchases. The question is: what is the reason behind this similarity? Is it because they have similar tastes, since, after all, they are friends? Or, is it because one influences the other's decision, as they communicate frequently? Consider that the firm wants to improve the sales of a product by leveraging the knowledge that a certain consumer just purchased it. If it is the taste similarity that drives the similar decisions, the firm should directly target *friends of that customer* by offering discounts to them. If, instead, it is social influence that drives the similarity, the firm should incentivize *that customer* to promote the product or service to her friends. The answer to the original question thus clearly bears critical importance to businesses today.

People who have close relationships are known to have similar social and economic behaviors. Uncovering the reasons behind this similarity is of great interest to researchers. However, two obstacles stand in the way. The first is data availability. A large dataset which consists of information on social connections, communications and relevant decisions is usually required for this research. Only very recently have datasets like this begun to emerge. The second major challenge is that many different factors that drive this similar behavior are close to observationally equivalent. Manski (1993) studies three different effects, correlated, endogenous, and exogenous, and shows that in a static context they are impossible to be separately identified. Even in a dynamic context, care must be taken for proper identification.

Research has long recognized that people with similar characteristics are more likely to form ties, an effect termed as *homophily*. Consequently, people who have close ties tend to have similar traits. Research has also established that human decisions, often made within a social context, are subject to influence of others. Many terms are used to describe this effect, such as social interaction, peer influence, conformity, imitation, etc. In this study, we term this force category simply as *social influence*. In the example given above, the former is the homophily effect, while the latter is the social influence effect.

Identifying and measuring homophily and social influence effects is the focus of our study. We build a hierarchical Bayesian model which simultaneously accounts for both homophily and social influence effect in consumers' decision process. We model both the purchase timing and the product choice decisions. We estimate our model using a unique panel dataset obtained from an Indian telecom company, and find strong evidence of homophily – individual characteristics such as propensity to purchase, product valuation, and susceptibility to influence, all demonstrate positive within group correlation. We also find strong evidence of social influence in consumer's product choice, although influence is not evident in purchase timing decisions. To the best of our knowledge, simultaneous quantification of homophily and influence effects have not been

reported in literature. This is the contribution we seek to provide. Our work is ongoing as we are estimating alternative models for robustness check and conducting policy simulations.[1]

2. Model

Let there be G groups of consumers. Each group consists of I consumers. We index the i-th consumer of g-th group as gi. Consumers in a same group have close social relationship (e.g. are friends). There are J products. The characteristics of each product j is X_j, which is a $K \times 1$ vector. There are T time periods. In each period, each consumer makes a purchase decision. The decision consists of two steps. A consumer first decides whether to buy a product in the period. If yes, she then decides what product to buy.

We model the first step, the *when-to-buy* decision, using a hazard-rate model. Specifically, we assume that the inter-purchase time of a consumer gi follows an Erlang-2 distribution with rate $\lambda_{gi,t}$. The survivor function is:

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

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

In equation (2), $E_{gi,t}$ denotes the amount of product exposure the consumer gi had from friendsd in period t. γ_{gi} is the *susceptibility* parameter, indicating the extent to which the consumer is subject to external influence in making her decisions (a large positive value indicates the consumer positively values inputs from others in purchase timing decisions).

We model the second step, the *what-to-buy* decision, using a discrete choice model. Let β_{gi} be the valuation coefficient of consumer gi (like X_j, β_{gi} is a $K \times 1$ vector). Denote $E_{gi,j,t}$ as the amount of exposure the consumer gi received in period t on product j. The utility of consumer gi purchasing product j at time period t is

$$U_{gi,j,t} = X_j^T \beta_{gi} + \rho_{gi} E_{gi,j,t} + \varepsilon_{gi,j,t} \qquad (3)$$

Similar to the parameter γ_{gi} in purchase timing, the parameter ρ_{gi} in equation (3) indicates how much a consumer's perceived utility of a product is influenced through communication with others. Assuming $\varepsilon_{gi,j,t}$ follow the type-I extreme-value distribution, the product choice probability then follows that of a multinomial-logit model.

Homophily

To model homophily, we allow consumers of a same group to have correlated parameters:

$$\begin{pmatrix} \theta_{g1} \\ \dots \\ \theta_{gI} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \bar{\theta} \\ \dots \\ \bar{\theta} \end{pmatrix}, \sigma_\theta^2 \begin{bmatrix} 1 & r_\theta & r_\theta \\ r_\theta & \dots & r_\theta \\ r_\theta & r_\theta & 1 \end{bmatrix}\right) \qquad (3)$$

[1] We omit a thorough literature review and reference due to page limits. Broadly speaking, our work is related to the literature on social and economic networks in the fields of information system, marketing, economics, and computer science.

In equation (3), $MVN(\bar{\mu},\Sigma)$ represents the multivariate normal distribution. Homophily is reflected from the parameter r_θ: if consumers of the same group have similar characteristics, then their parameter values should be positively correlated, i.e. $r_\theta > 0$. If, however, the homophily effect does not exist, then we expect $r_\theta = 0$ (if consumers have opposing characteristics, we expect $r_\theta < 0$). In the equation, θ represents each of the aforementioned parameters: λ_{gi}, γ_{gi}, β_{gi}, and ρ_{gi} [2]. That is, we allow for the homophily effect in all parameters ex ante.

Identification and Estimation

The key to our identification strategy is the static nature of the homophily effect versus the dynamic nature of social influence: while the characteristics of consumers such as product valuation remain stable overtime, the consumers are exposed to different levels of influence over time. Therefore, the effects of social influence and homophily can be separated. In our models, the exposure variables change over time, while others remain constant. With adequate variation of time of the exposure, the parameters can be identified.

We use the MCMC method to draw parameters from their posterior distributions. The likelihood function shows that the inter-purchase timing component and the product choice component are factorized. Therefore, they can be estimated separately.

3. Data and Result

We estimate our model using a unique dataset obtained from a large Indian telecom company. The dataset contains the phone call histories of the company's over 3.7 million customers in a major city over a six-month period. The dataset also contains the detailed purchase records of caller ring-back tones (CRBT) by these customers[3]. To use CRBT, a customer must pay a monthly subscription fee, and purchase individual songs. More than 740 thousand customers purchased CRBTs in the covered period, and this is the purchase decision that we analyze in our study.

To fit the model to the data, we need to identify groups as well as quantify the influence using the social network information encoded in the phone call records. We consider two consumers as having close relationship if they made at least 5 phone calls in the first month. This produces a network, where nodes represent customers and edges represent relationships among them. To identify groups in this network, we then search for all 4-cliques in the graph, each representing a group of consumers with close social ties among them. A total of 2243 such groups are discovered, 300 of which are randomly selected for estimation.

There are over 11 thousand songs which are classified into 10 categories. Only 3 of the 10 categories have significant market shares, so we combine the rest into one category. This results in 4 products, each corresponding to a category. As there is little information about the songs and categories, we only include the product dummies in the product characteristic matrix.

---

[2] Due to positivity constraint, λ_{gi} is assumed to follow a log-normal distribution rather than a normal one.

[3] CRBT is a popular phone feature in India. The way it works is when a customer *A* purchases a certain ring-back tone, then when another person *B* calls *A*, *B* will hear the ring-back tone when the phone is connected, before *A* picks up the call.

Since the products are caller ring-back tones, the influence people impose on one another is conveniently encoded in their communication records within the network: when a customer calls another who uses caller ring-back tone, she will hear the song played. This exposes her to two things: first, this person is using caller ring-back tone; and second, this person chooses this song. The social influence argument then suggests that both her purchase timing decision and product choice decision may be influenced through this phone call.

We thus quantify this external influence based on the phone calls made by the customer. As both the phone call records and the ring-back tone purchase records are time-stamped, we can infer how many times a customer is exposed to a certain category of songs within a certain period. To account for possible delayed effects (e.g., a customer makes a phone call and is exposed to a song, and then buys another song in the same category, but three days later instead of on the same day), the exposure is exponentially smoothed across time periods.

The estimation result for product choice is reported in table 1. We present only the parameters related to homophily and social influence while suppressing the rest. The influence parameter has a population level mean of 0.144, which is positive and statistically significant. This presents clear evidence that consumer's product choice decision is influenced by others around her. The mean group level correlations of the product valuation coefficients are 0.630, 0.132, and 0.498, respectively. All are positive and the first and third are statistically significant. The mean group level correlation of the influence parameter is 0.577, also positive and statistically significant. These indicate that homophily effect is present on both product valuation and susceptibility to influence, confirming that consumers who are close by indeed have similar characteristics.

Table 1: Product Choice Model Parameter Estimation

Parameter	Posterior Mean	2.5% Posterior Quantile	97.5% Posterior Quantile
$\bar{\rho}$	0.144	0.0773	0.221
σ_{ρ}^{2}	0.0339	0.0154	0.0771
r_{β_1}	0.630	0.320	0.871
r_{β_2}	0.132	-0.168	0.653
r_{β_3}	0.498	0.211	0.816
r_{ρ}	0.577	0.0666	0.841

The estimation result for purchase timing is reported in table 2. Again, only influence and homophily related parameters are presented. The table shows that the influence parameter has a population level mean of -0.014 but is not statistically different from zero. This shows that there is no evidence that, at the population level, a consumer's timing of purchase is influenced by those around her. The group level correlation for the base purchase timing rate is 0.334, and that for influence parameter is 0.332. Both are positive and statistically significant, confirming that homophily effect exists for purchase timing related parameters as well. Thus we know that consumers who are close by also have similar characteristics regarding when to purchase a CRBT song.

Table 2: Purchase Timing Model Parameter Evaluation

Parameter	Posterior Mean	2.5% Posterior Quantile	97.5% Posterior Quantile
$\bar{\gamma}$	-0.0142	-0.0552	0.0251

σ_γ^2	0.0122	0.0102	0.0146
r_λ	0.334	0.256	0.426
r_γ	0.332	0.225	0.444

4. Conclusion and Further Research

People close to one another often have similar social and economic behaviors. Researchers have long strived to uncover the factors behind this similarity, as each factor may call for a distinct managerial or policy response. Due to difficulties arising from data availability and econometric identification, however, this remains an open and active area of study. Our work contributes to the literature by simultaneously identifying and quantifying both the homophily and the social influence effect in consumers' purchase timing and product choice decision process. We estimate our model using a unique dataset which contains both social network and product purchase information, and find clear evidence of the homophily effect in all aspects of consumer's purchase decision. We also find strong evidence of social influence effect in consumer's product choice decision, but not in purchase timing decision. To our best knowledge, results of this type have not been covered in existing literature. Currently we are conducting robustness checks and policy simulations. We will report additional findings in the near future.

Our work is ongoing and we will report additional findings in the near future. To strike the right balance between reliably identifying social relationship and ensure data availability and representativeness, we used five phone calls as the threshold for relationship and extract groups of four members. To ensure the robustness of the result, we are estimating other configurations using different thresholds and/or group sizes. We are also evaluating the profit implications by conducting simulations on various promotion policies. Network information is increasingly being leveraged for improving business performance. Our work contributes to this important body of knowledge and provides guidance to practitioners.

5. References

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