DOCTORAL DISSERTATION

Essays on Internet and Network Mediated Marketing Interactions

submitted to the
David A. Tepper School of Business
in partial fulfillment for the requirements for the degree of
DOCTOR OF PHILOSOPHY in Industrial Administration

By Liye Ma
Carnegie Mellon University
Tepper School of Business
Pittsburgh, Pennsylvania 15213

DISSERTATION COMMITTEE:
Kannan Srinivasan (co-chair)
Baohong Sun (co-chair)
Kinshuk Jerath
Ramayya Krishnan
Alan Montgomery
Yaroslav Kryukov

April 2011

Acknowledgements

Contents

Chapter 1

Introduction

My dissertation focuses on marketing interactions enabled by Internet and networks. The advancement of Internet and computer electronic technology has led to ever richer interactions of consumers and firms. Such interactions often occur in and are mediated by networks, which present an exciting opportunity for marketing research. My dissertation consists of three essays, each focusing on a specific interaction in this area. In my first essay, "*A 'Position Paradox' in Sponsored Search Auctions,*" I study how quality differentiated firms bid for online advertising positions when they explicitly account for consumers' knowledge structure and search behavior. This study uses a game theoretic approach. Features such as click-through-rate and value-per-click, often taken as exogenous in extant literature, are endogenized in this study. Modeling at this lower level of primitive reveals an interesting position paradox, where a low quality firm bids higher than a high quality firm, is placed at a higher position, yet still receives fewer clicks. Analysis shows that explicitly modeling of consumer search is essential for understanding the key drivers of auction outcome, including "residual demand", i.e. the amount of clicks a firm can receive at lower position, "incremental value", i.e. the additional value a higher position offers, and "differential cost", i.e. the different cost implications of informed consumers to both firms under pay-per-click.

In my second essay, "*Homophily or Influence? An Empirical Analysis of Purchase within a Social Network,*" I study consumers' product purchase in a social network environment. Consumers who are close to one another often behave similarly, attributable to either their intrinsic similarity, i.e. an "unobserved homophily" effect, or their influence on one another

through communications. Teasing out the two effects is important for target marketing, but is also challenging. In this study, I use a dynamic model to incorporate both factors, where identification is achieved through separating static product taste from dynamic influence arising from communications, and estimate it using a dataset obtained from a large Indian telecom company, which contains information on repeated purchases and communications. I find strong homophily effect in consumers' product choice decision. In contrast, the purchase timing decision is heavily influenced by other consumers. I show that ignoring either effect will lead to an overestimation of the other. Furthermore, I show that detailed communication data is crucial for accurately identifying influence effects.

In my third and final essay, "*A Dynamic Competitive Analysis of Content Production and Link Formation of Internet Content Developers*," I study how Internet content developers compete for viewership through producing content and linking to one another. Hundreds of revenue sharing content websites have greatly contributed to the recent proliferation of social media. Content at these websites is supplied by external developers, whom the websites attract through revenue sharing. This leads to a competition for viewership among developers at a website. A feature recently introduced at many sites, namely allowing developers to link to one another, brings intriguing interactions to this competition. In this study, I develop a dynamic oligopoly model, following the framework of Ericson and Pakes (1995), to investigate the interaction of production and linking decisions of content developers, the tradeoffs they face over time, and the resulting market structure. The model is estimated using the data obtained from a popular Internet product review site, applying the two-step estimator recently developed by Bajari, Benkard, and Levin (2007). I find that reciprocal links are naturally encouraged by a promote-the-promoter effect, which then induces certain developers to strategically initiate non-

reciprocal links, in anticipation of reciprocation. I find that the prospect of receiving incoming links can either encourage or discourage content production, depending on the situation a developer is in. Furthermore, I find that although both more content and higher network position increase viewership, only the latter leads to higher net benefit once cost is accounted for. This suggests that linking may impede competition, by giving competitive advantage to a subgroup of content developers, and simulation suggests that limiting links could indeed lead to higher overall viewership at the website.

Chapter 2

A "Position Paradox" In Sponsored Search Auctions

1. Introduction

Sponsored search advertising has grown into one of the major forms of online advertising in the past decade and is expected to grow at an annual compound rate of more than 12% in the near future (Riley et al. 2007). Firms – global and local, big and small – now actively advertise in the "sponsored links" sections of search engines such as Google, Yahoo!, Bing (MSN Live Search) and Ask. When a consumer searches for a specific keyword on the search engine, she is presented with two lists of clickable links: one a list of organic search results and the other a list of sponsored links. The list of sponsored links is determined by auctioning the positions to firms which want to advertise in response to the searched keyword. Advertised links are typically ordered in decreasing order of the firms' bids and some search engines, such as Google, also augment bids by a quality metric for the associated link. Two widely used payment mechanisms are pay-per-impression, where all firms are charged whenever a consumer searches the keyword and their links are displayed, and pay-per-click, where a firm is charged only when a consumer clicks on its link. Typically, search engines use a second-price auction, i.e., a firm has to pay the bid of the firm directly below it or, if it is in the last position, it has to pay a pre-specified minimum amount.

The spectacular commercial success of sponsored search has motivated several recent academic studies on it. While this literature is in its nascent stage, a main take-away from both empirical and theoretical work is that advertisements at higher positions attract more clicks from consumers (e.g., Feng et al. 2007, Misra et al. 2006) and higher-quality firms will be placed at

higher positions (e.g., Athey and Ellison 2011). In this paper, we analytically model the bidding strategies of firms when quality differentiation exists and some consumers know the firms' qualities beforehand (the informed consumers). We analyze both pay-per-impression and pay-per-click mechanisms and obtain a host of results on the optimal bidding strategies of high-quality (superior) firms and low-quality (inferior) firms and the outcomes of these bidding strategies.

An intriguing finding revealed in our study is the "position paradox" – we find that, under certain conditions, a superior firm may bid lower than an inferior firm and obtain a position below it, but still obtain more clicks than the inferior firm does. We first show this under the pay-per-impression mechanism, in which we also find that as the quality premium for the superior firm increases, the incentive for the inferior firm to be at the top may, in fact, increase. Furthermore, we find that the position paradox is actually strengthened under the pay-per-click mechanism, i.e., the incentive for the inferior firm to be at the top is stronger under pay-per-click than under pay-per-impression case. Moreover, we also find that while under the pay-per-impression mechanism the number of informed consumers does not influence which firm wins the auction, under the pay-per-click mechanism, as this number increases, the incentive for the inferior firm to be at the top may increase.

The position paradox phenomenon is unusual at first look. Why would firms ranked lower receive more clicks than firms ranked higher? And more importantly, why would such an outcome be optimal for all firms, i.e., why would the top ranked firm pay the highest bid even though it receives fewer clicks than lower ranked ones? We explain this puzzle using an analytical model which accounts for differentiation in firms' product qualities, and differentiation in consumers, where some consumers know the firms' qualities beforehand.

Our analysis focuses on three key factors that drive auction outcomes: residual demand, incremental value, and differential cost. Intuitively, the residual demand effect means that a consumer may continue searching through sponsored links if she could not find a product of her choice at the firms she visited earlier, or if she expects to find a better one if she continues to search. Thus, even if she is processing links starting from the top, lower-ranked links can still benefit from the residual consumers from above and generate positive sales. The incremental value effect refers to the fact that, from a firm's point of view, a higher-ranked link is worth only the additional revenue it generates over the lower-ranked links, instead of the absolute revenue it generates. The differential cost effect, relevant to the pay-per-click mechanism, refers to the fact when a firm pays if a consumer clicks on its link, informed and uninformed consumers have different relative cost implications to the competing firms. The first two factors influence the revenues of firms, while the third factor influences the advertisement costs of firms.

With the interaction of these three factors, a superior firm may prefer a lower-ranked position over a higher-ranked one if it can receive only slightly fewer number of clicks at the lower-ranked position but can substantially reduce its cost there. An inferior firm, on the other hand, may want to take the higher-ranked position because it would receive substantially fewer clicks at the lower-ranked positions. Furthermore, under the pay-per-click scheme, if some consumers (the informed consumers) already know the identity of the superior firm and will click on that firm's link regardless of its position, then the superior firm will be even more unwilling to be placed at higher positions with higher per-click costs, making the inferior firm relatively more competitive.

The position paradox result is not only counter intuitive, it is also in contrast with the prevailing view in the literature, which almost takes it as a given that the higher the link position,

the more clicks it will attract from consumers and, consequently, superior firms will also emerge at the top. (We review the literature in detail later.) To validate our theory, we analyze a dataset on sponsored search auctions from a popular Korean search engine firm, and find indirect empirical support for our theory. Specifically, we find that: (i) a large proportion of auction outcomes in the data show the position paradox, and (ii) sharp predictions of data patterns from our analytical model are validated in the data.

Numerous empirical studies on sponsored search have found that, when aggregated across keywords, the number of clicks on a link decreases approximately exponentially as one proceeds down a list of sponsored links (e.g., Feng et al. 2007, Misra et al. 2006). In contrast to these widely accepted notions, we find in our data that, when analyzed at the keyword level, the number of clicks is often greater at lower positions. Other studies have looked at how position influences purchase conversion rates after a consumer has clicked on a link, and Agarwal et al. (2009) find that, interestingly, the conversion probability is often highest at the second or third position and not always the first position. Other empirical studies have focused on how keyword-level characteristics (e.g., whether the keyword being bid on is a generic keyword, a brand-specific keyword or a retailer-specific keyword) influence clicks and purchase conversion (Rutz et al. 2011, Rutz and Bucklin 2011, and Ghose and Yang 2009). Yao and Mela (2011) develop a dynamic structural model to jointly study the strategic behavior of searchers, advertisers and the search engine firm. Goldfarb and Tucker (2011) empirically find that, because sponsored search is typically well targeted, firms bid overall higher in markets where it is harder for them to find matching consumers on their own. Another line of research focuses on developing computational methods to determine, using past data, keywords that firms should bid on and the corresponding bids they should make (e.g., Kitts and Leblanc 2004, Abhishek and Hosanagar 2007).

Several analytical studies in this area focus on the optimal design of the auction mechanism and its implications. Edelman et al. (2007) and Varian (2007) study auction forms that closely resemble the auction scheme of Google and discuss the properties of the equilibria arising from this auction. Liu et al. (2010) study an auction mechanism where the search engine determines link ranking by weighting firms' bids by the quality of each link, and show that incorporating quality both increases efficiency and boosts the revenue of the search engine. Weber and Zheng (2007) argue that the optimal scheme for a search engine to rank links may be to weight bids with relevant performance metrics and a low-quality bidder will lose in such an auction. Balachander and Kannan (2009), on the other hand, find that using more information to weight links might lead to a reduction in profits for the search engine. Katona and Sarvary (2010) develop a normative model to suggest how search engines should adjust advertisers' bids taking into account consumers' clicking behavior for both organic and sponsored links. Chen and He (2006) model bid behavior of sellers with products of varying degrees of relevance to the keyword. They find that when search cost is present, higher relevance sellers bid higher and get placed higher. Athey and Ellison (2011) study an equilibrium model which accounts for both firms' bids and consumer search at the same time, and describes the equilibria under two different scenarios, one where consumers believe that firms are sorted randomly, the other where consumers believe that firms are sorted on decreasing quality. Zhu and Wilbur (2011) analyze auctions in which advertisers decide to bid on a pay-per-impression or pay-per-click basis. Shin (2009) studies the incentives firms have to purchase their own and their competitors' branded keywords. Desai and Shin (2009) find that employing advertiser-specific minimum bids can increase search engine profit. Our research finds that neither is it always true that higher positions obtain more clicks, nor is it always true that firms are sorted in descending order of

quality. These observations and an associated explanation are missing from the literature, and our attempt in this paper is to fill this gap.

The rest of this chapter is organized as follows. Section 2 describes the analytical model. In Section 3, we derive the equilibrium when consumers have homogeneous search costs, first for the pay-per-impression scheme, and then for the pay-per-click scheme. In Section 4, we extend our results to the case of heterogeneous search costs for consumers. In Section 5 discusses the empirical support of our theory using data obtained from a search engine firm in South Korea. We conclude in Section 6.

2. Model

We model two firms, S and I, competing for sponsored search advertisement positions for a specific keyword at a search engine firm. Firm S, the superior firm, has a higher-quality product than firm I, the inferior firm. The product of the inferior firm provides a consumer a net product utility (the utility of the product minus its price) of V, while that of the superior firm provides a consumer a net utility of $V+Q$. Here, $Q>0$ represents the quality premium of firm S over firm I. The per-unit margin of firm i is denoted by $m_i, i \in \{S, I\}$. We assume that the superior firm has a higher per-unit margin, i.e. $m_S > m_I$. (This assumption is only made to highlight the position paradox, as we will show that even with higher margin, the superior firm may not desire the top position, and bid to be placed at the second position. Relaxing this assumption will only make this more likely to happen, as we elaborate later.)

A group of consumers, with mass normalized to 1, search the keyword in question at the search engine. In response, the search engine returns an ordered list of sponsored links.[1] We

1Search engines typically return both sponsored and organic links in response to a query. This paper focuses on the sponsored

assume that there are two types of consumers. The first type is the informed consumers, who can tell whether a firm is the superior firm or the inferior firm upon viewing its advertisement link in the search results. The second type is the uninformed consumers, who know that there is a superior firm offering quality $V+Q$ and an inferior firm offering quality V, but cannot tell a firm's quality from the advertisement link itself. These consumers have to search for this information by clicking on a firm's link and obtaining information about the product, e.g. find out the product specifications and price, read consumer reviews, etc. After this exercise, the uninformed consumers can also determine the quality of the firm. We assume that the size of the informed consumers is $\phi \in (0,1)$, so the size of the uninformed consumers is $1-\phi$. The parameter ϕ can be interpreted as a measure of how widespread the reputation of the superior firm is in the market.

However, knowing the product quality is not sufficient for a consumer to make the purchase decision – she also has to assess her "match" with a product. This match can only be assessed by clicking on a firm's link and obtaining information about the product. Hence, before purchasing a product, both informed and uninformed consumers have to click on a firm's link to assess their match with the product being offered. In any consumer's purchase decision, product quality can be interpreted as an objective dimension and match can be interpreted as a subjective dimension. We assume that the informed consumers always start their search with the superior firm and may go to the inferior firm if they do not obtain a match with the superior firm, while the uninformed consumers start their search with the firm at the top, which could be the superior or the inferior firm, and based on the quality and the match they may then choose to search

---

links.

further or stop.[2]

Note that we assume that uninformed consumers are naive – they start their search from the top and proceed downwards link-by-link irrespective of how firms are ordered by quality. One could think of a scenario in which some uninformed consumers are sophisticated, in that they can rationally expect the order of firms in equilibrium and, therefore, directly click on the superior firm's link.[3] These sophisticated uninformed consumers, however, can then be treated simply as informed consumers and this only increases the size of the informed population (i.e., increases the value of ϕ), leading to a model that maps back to the current model. In other words, we could think of consumers as coming from three segments –informed, sophisticated uninformed and naive uninformed. All we need for our results to qualitatively hold is that the mass of the naive uninformed consumers is positive and these consumers start their search from the top.[4] Given that the population of consumers using the Internet is very diverse and heterogeneous, and considering the evidence from Behavioral Economics that shows that consumers often do not have perfect strategic foresight (Camerer, Ho and Chong 2004, Ho, Lim and Camerer 2006), the assumption that at least some consumers are naive is reasonable. Furthermore, the assumption that these naive consumers start searching from the top position is in accordance with literature on how online consumers process ordered lists (Hoque and Lohse 1999, Granka, Joachims and

2As an illustrative example, suppose the query "running shoes" returns the links BadShoes.com (the inferior firm) at the top and GoodShoes.com (the superior firm) at the bottom. An informed consumer first clicks on GoodShoes.com but her choice of color may not match and she may, therefore, also click on BadShoes.com. An uninformed consumer first clicks on BadShoes.com, realizes that it is the low quality product and, even if her color choice matches, she may choose to explore GoodShoes.com and click on it.

3This could be modeled by a Rational Expectations Equilibrium or a Perfect Bayesian Equilibrium.

4Another way to think about these consumers is that they are boundedly rational (Simon 1955). In other words, they find the cost of thinking through the bidding strategies of the firms and figuring out how they will be ordered in equilibrium to be more than the expected cost of the search effort they have to expend in searching through the links of the firms using the heuristic of starting from the top and going downwards. Hence, they choose to search rather than compute the equilibrium.

Gay 2004).

Consistent with the existing literature on search, we assume that the first search is free (this assumption does not qualitatively impact our results), while a search cost of $s > 0$ applies for subsequent searches. As several studies have shown, this search cost can be substantial and has a significant impact on how much an online consumer searches (Johnson et al. 2004, Brynjolfsson et al. 2009). We also assume that once a link has been clicked on by a consumer, she can always go back to this link without incurring any further search cost. We study both the case of homogeneous search cost, where every consumer has the same search cost, and that of heterogeneous search cost, where different consumers may have different search cost. We assume that the match probability is the same for both firms and is equal to p. (Assuming this to be different for different firms does not alter our results qualitatively.) We assume that every consumer makes her purchase or subsequent search decisions to maximize her expected utility.

The search engine can auction the position either through a pay-per-impression or a pay-per-click mechanism. We study both mechanisms in this paper. Consistent with industry practice, we assume that both firms submit their bids simultaneously, the winning firm pays the amount of the losing firm's bid and the losing firm pays the minimum bid $\underline{b}$. (Ties are broken randomly with equal probability.) In both mechanisms, the firm which bids higher is placed on top, while the other is placed at the bottom. In the pay-per-impression mechanism, both firms pay their respective fees whenever their links are displayed, i.e., whenever a consumer searches the keyword. In the pay-per-click mechanism, a firm pays only when its link is actually clicked. Note that in both cases, we assume the search engine imposes a minimum bid $\underline{b}$, which is the minimum bid amount required for either firm to participate in the auction. In real-world cases, firms regularly impose this restriction. An alternative interpretation of this minimum bid is that it

is the minimum amount a firm needs to bid to be placed in the top two positions, when there are more than two firms bidding for more than two positions. We make the assumption that both firms always place their bids, i.e., no firm exits the bidding game. This assumption does not affect any insights from the model, while it makes the analysis cleaner in some cases.

Note that we do not endogenously model the prices charged by the firms for their products and assume that margins on sale are exogenously specified, as discussed earlier. This assumption is a standard assumption in the literature on position auctions and is justified by the fact that sponsored search advertising is typically a small part of any firm's total budget and does not significantly influence the price of its product.

We model the game in two stages. In the first stage, both firms submit bids simultaneously and the firm that bids higher obtains the top position. In the second stage, each consumer conducts her search. The equilibrium concept we use is subgame perfect Nash equilibrium. Under this solution concept, in the first stage, each firm calculates the profit it will make in the second stage for different outcomes. i.e., each firm considers how consumers will react in all possibilities in the second period based on the positions the firms will obtain according to their bids in the first period. In equilibrium, each firm's bid in the first period is such that it maximizes the firm's expected profit given the other firm's bid.

3. Homogeneous Search Costs

We begin our analysis with the case of homogeneous search cost, where all consumers have the same search cost. This enables us to analyze the workings of the underlying factors in various situations. We extend our result to the case of heterogeneous search cost later.

3.1. Pay-Per-Impression

In this section, we study the pay-per-impression case, in which firms pay their respective advertisement fees whenever a consumer searches the keyword and the link is displayed. Depending on the parameters s, p, m_S, m_I, V, and Q, multiple scenarios exist, each with different optimal search behavior by consumers and, therefore, different optimal bidding behavior by firms. Behind all these scenarios, however, are two key factors that drive the bidding behavior of firms.

The first is that of residual demand, which intuitively means that the firm placed on top may not get all the sales. This first arises from the uncertainty in finding a matched product. When a consumer goes to a firm's website, she finds a matched product only with probability p. If a match is not found, the consumer may search on. Furthermore, a consumer may decide to search on if the first firm she visits is the inferior one, even if she already finds a match there, provided the quality premium, Q, is sufficiently high to outweigh her search cost. These make it possible for a losing bidder to still make sales. This factor makes a sponsored search auction qualitatively different than a standard winner-take-all auction.

The second factor, partly arising from the first, is that of incremental value. Since both the winning and losing bidders may make positive sales, it is the difference in revenue between winning and losing that decides a firm's bid (not the absolute revenue from winning). For instance, a firm may make higher sales than the other if placed on top, but if it also makes higher sales than the other if placed on bottom, then its desire to be placed on top is not necessarily higher. Taking this argument further, if one firm makes the same sales regardless of its link position, while the other makes less sales at the lower position, then the latter firm is expected to

outbid the former even if its absolute sales are lower. Hence, a firm decides its bid based on the additional profit of being at the top position.

We now discuss each scenario in detail and then consolidate the results across these scenarios.

Scenario I: $s \le p\min\{Q,V\}$

Scenario I can be described as that of "low search cost." In this case, the search cost is sufficiently low so that consumers' decisions are primarily driven by product values. Consider an uninformed consumer. Being uninformed, she will start from the top link. If the top link is that of the superior firm and she finds a match there, she will buy the product and stop. If she does not find a match, she will continue to click on the bottom link of the inferior firm, because the search cost is lower than the expected product value, pV. If the top link is that of the inferior firm, however, the consumer will continue to click on the bottom link, regardless of whether she finds a match at the top link. This is because the expected incremental value to the consumer, pQ, is higher than the search cost. If she finds a match at the superior firm, she will buy the product and stop. Otherwise, she will either buy the product from the inferior firm if a match had been found when she first searched there, or stop if not.

The case for an informed consumer is simpler, as she always starts from the superior firm, regardless of where it is placed. Such a consumer will buy from the superior firm if a match is found, and searches the inferior firm in case of no match.

Denote the profit of firm $i \in \{S,I\}$ when it is placed at position $j \in \{1,2\}$ by $\Pi_{i,j}$. Given the expected search behavior of consumers, if the superior firm is placed on top and pays b_S per search, the expected profits for the two firms are given by

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = (1-p)pm_I - \underline{b}.$$

If the inferior firm is placed on top and pays b_I per search, the expected profits for the two firms are given by

$$E[\Pi_{S,2}] = pm_S - \underline{b} \text{ and } E[\Pi_{I,1}] = (1-p)pm_I - b_I.$$

Note that, in the expressions above, b_S and b_I denote the payment per search for firm S and firm I, respectively, *not* their equilibrium bids. These equilibrium bids, denoted by b_S^* and b_I^*, will be derived subsequently.

We can observe from the above that, in this case of low search cost, each firm has the same expected revenue at both positions (equal to pm_S and $(1-p)pm_I$ for firms S and I, respectively). This is a direct outcome of the effect of residual demand – because search cost is low, every consumer will search the other firm whenever she fails to find a match at one; and because the search cost is low relative to the product quality premium, the consumer will search the superior firm even if she first visits the inferior one and finds a match there. Consequently, the effect of incremental value suggests that neither firm will compete for the top position. We see from the derivations below for the firms' bids that this is indeed the case.

We can derive the equilibrium bids for each firm for position 1 by noting that a firm in position 2 pays $\underline{b}$, and will bid an amount such that if it indeed gets the top position and has to pay this full amount, its profit from should be equal to its profit at the bottom position where it pays the minimum bid. Intuitively, a firm will bid the amount that the top position offers in addition to the bottom one. As noted in Varian (2007), these will be Nash equilibrium bids and this is a natural equilibrium to consider, because it assumes that a firm sets its bid so that it

makes a profit if it moves up in the ranking. Note that these bids also characterize exactly the locally envy-free equilibrium defined in Edelman et al. (2007), since another way of looking at this equilibrium is that if a firm in the bottom position is already making this bid then the firm in the top position must be bidding more than this value. However, the firm in the bottom position does not want to bid more than this anyway and, therefore, does not envy the firm above it. We also note that this is a weakly dominant strategy to both firms.

In accordance with this, firm S will be willing to pay up to b_S^* for position 1, where b_S^* equates its profit from position 1 and position 2, i.e.:

$$pm_S - b_S^* = pm_S - \underline{b} \Rightarrow b_S^* = \underline{b}$$

Similarly, firm I will be willing to pay up to b_I^* for position 1, where b_I^* equates its profit from position 1 and position 2, i.e.,

$$(1-p)pm_I - b_I^* = (1-p)pm_I - \underline{b} \Rightarrow b_I^* = \underline{b}$$

Hence, in this scenario, both firms will bid $b_S^* = b_I^* = \underline{b}$, the minimum required amount, with the tie broken randomly by the search engine (with equal probability). Note that we already have a case where the inferior firm might be at the top position. However, this is a somewhat uninteresting result because the positions are decided randomly. In the scenarios to follow, we will show situations which leads to the position paradox not in a random manner.

Scenario II: $Q > V$ and $pV < s \leq pQ$

Scenario II exists if the quality premium is significant, i.e., $Q > V$. In this scenario, the search cost is higher than the expected utility of visiting the inferior firm, but lower than the expected quality premium.

An uninformed consumer in this case will still start from the top link. If the superior firm has the top link and the consumer finds a match there, she will buy and stop, just as in Scenario I. Different than Scenario I, however, is that she will stop even if she does not find a match. This is because after the first search, she already knows that the second link belongs to the inferior firm, and since the search cost is higher than the expected utility from searching the inferior firm for a match, her best action is to stop searching. If the top link belongs to the inferior firm, the consumer will search on regardless of whether she finds a match, since the expected quality premium outweighs the search cost, the same as in Scenario I.

An informed consumer, in contrast, will always start from the superior firm. And in this case, she will not visit the inferior firm regardless of whether she finds a match at the superior one. Given these, the expected firm profits when the superior firm is placed on top are:

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = 0 - \underline{b}^{5}$$

If the inferior firm is placed on top, the expected firm revenues are:

$$E[\Pi_{S,2}] = pm_S - \underline{b} \text{ and } E[\Pi_{I,1}] = (1-\phi)(1-p)\,pm_I - b_I\,.$$

As in Scenario I, the superior firm has the same expected revenue regardless of where its link is placed. The inferior firm, however, makes no sales if it is placed at bottom. From a similar analysis as before, the equilibrium bids for the two firms will be $b_S^* = \underline{b}$ and $b_I^* = (1-\phi)(1-p)\,pm_I + \underline{b}$. Note that the inferior firm bids higher and, therefore, will have the top link position. Once again, we have a case that exhibits the position paradox, with the superior firm also obtaining more clicks even though it is at the bottom.

5Note that this profit is negative. However, we have assumed that firms do not exit the auction even if they make negative profit. This is a purely technical assumption that keeps the analysis cleaner; it has no effect on the insights from the model.

Scenario III: $Q < V$ and $pQ < s \leq pV$

Note that Scenario III and Scenario II are similar and mutually exclusive, and Scenario III exists if the quality premium is not high, i.e., $Q < V$. In this scenario, if a consumer first visits the superior firm and finds a match, she will still buy and stop. If she first visits the inferior firm and finds a match, she will also buy and stop, because the expected quality premium does not warrant an additional search. If a consumer does not find a match in the first firm she visits, she will continue to search the other firm. As in the first two scenarios, an uninformed consumer will first search whichever firm that is on top, while an informed one will search the superior firm first.

The expected firm profits when the superior firm is on top are:

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = (1-p)pm_I - \underline{b}$$

If the inferior firm is placed on top, the expected firm profits are:

$$E[\Pi_{S,2}] = \phi pm_S + (1-\phi)(1-p)pm_S - \underline{b} \text{ and } E[\Pi_{I,1}] = \phi(1-p)pm_I + (1-\phi)pm_I - b_I$$

The equilibrium bids can be derived as:

$$b_S^* = (1-\phi)p^2 m_S + \underline{b} \text{ and } b_I^* = (1-\phi)p^2 m_I + \underline{b}.$$

In this scenario, both firms will generate higher revenue when placed on top than when placed on bottom. However, since $m_S > m_I$, the superior firm has "more to lose" at the bottom position than does the inferior firm. Therefore, the superior firm will outbid the inferior firm and will occupy the top position while the inferior firm will be placed at the bottom.

Scenario IV: $p\max\{Q,V\} < s \leq p(V+Q)$

In Scenario IV, the search cost is higher than both the expected quality premium and the

expected utility of visiting the inferior firm, but lower than the expected utility of visiting the superior firm. In this case, if a consumer finds a match at the first firm she visits, she will buy the product and stop. If she does not find a match there, then if the other firm is the superior firm, she will search on, but if the other firm is the inferior one, she will stop. Again, an uninformed consumer searches whichever firm is on top first, while an informed consumer starts with the superior firm.

The expected firm profits when the superior firm is on top are:

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = 0 - \underline{b}$$

If the inferior firm is placed on top, the expected firm profits are:

$$E[\Pi_{S,2}] = \phi pm_S + (1-\phi)(1-p)\,pm_S - \underline{b} \text{ and } E[\Pi_{I,1}] = (1-\phi)\,pm_I - b_I$$

The equilibrium bids can be derived as:

$$b_S^* = (1-\phi)\,p^2 m_S + \underline{b} \text{ and } b_I^* = (1-\phi)\,pm_I + \underline{b}$$

In this scenario, both firms value the top position more than they value the bottom one. If the margin of the superior firm is significantly higher than that of the inferior firm, i.e., $pm_S > m_I$, then the superior firm will have bid higher and be placed on top. Otherwise, the inferior firm will be placed on top. Even if inferior firm is on top, it will obtain more clicks only if $\phi < p/(1+p)$, otherwise the superior firm will obtain more clicks.

Scenario V: $s > p(V+Q)$

In Scenario V, the search cost is higher than even the expected utility of visiting the superior firm. In this case, a consumer will stop after conducting the first search, regardless of which firm is visited or whether a match is found. An uninformed consumer again starts from the top, while

an informed one again starts from the superior firm.

The expected firm profits when the superior firm is on top are:

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = 0 - \underline{b}$$

The profits when the inferior firm is on top are:

$$E[\Pi_{S,2}] = \phi pm_S - \underline{b} \text{ and } E[\Pi_{I,1}] = (1-\phi)\, pm_I - b_I$$

The equilibrium bids can be derived as:

$$b_S^* = (1-\phi)\, pm_S + \underline{b} \text{ and } b_I^* = (1-\phi)\, pm_I + \underline{b} \ .$$

In this case, the superior firm has more to lose if it does not win the top position than does the inferior firm. Hence, it will bid to be placed on top.

Summarizing the Scenarios

The results of the five scenarios discussed above are summarized in the following proposition:

Proposition 1: *In the pay-per-impression bidding game:*

- *If the search cost is very low (Scenario I), both firms will bid an equal amount.*
- *If the search cost is moderately low (Scenarios II and III) then, if the quality premium is high (Scenario II), the inferior firm will bid higher, otherwise (Scenario III) the superior firm will bid higher.*
- *If the search cost is moderately high (Scenario IV) then, if the superior firm's margin is significantly higher than that of the inferior firm ($pm_S > m_I$), the superior firm will bid higher, otherwise the inferior firm will bid higher.*
- *If the search cost is very high (Scenario V), the superior firm will bid higher.*

For each scenario, the equilibrium bids by the two firms are presented in Table 1.

Table 1: Equilibrium Bids under Pay-Per-Impression

Scenario	**Firm S bid** **Firm I bid**
$s \le p\min\{Q,V\}$	$\underline{b}$
	$\underline{b}$
$Q > V$ and $pV < s \le pQ$	$\underline{b}$
	$(1-\phi)(1-p)pm_I + \underline{b}$
$Q < V$ and $pQ < s \le pV$	$(1-\phi)p^2 m_S + \underline{b}$
	$(1-\phi)p^2 m_I + \underline{b}$
$p\max\{Q,V\} < s \le p(V+Q)$	$(1-\phi)p^2 m_S + \underline{b}$
	$(1-\phi)pm_I + \underline{b}$
$s > p(V+Q)$	$(1-\phi)pm_S + \underline{b}$
	$(1-\phi)pm_I + \underline{b}$

In the above proposition, the cases in which the search cost is moderate are intriguing. When search cost is moderately low, if the quality premium is high, the superior firm will not lose sales when placed at the bottom, but the inferior firm will. Therefore, the inferior firm will bid higher and get the top position. But if the quality premium is low, then the superior firm will also lose sales when placed at the bottom, and more so than the inferior firm. Therefore, the superior firm will bid higher and take the top position.

When search cost is moderately high, both firms always prefer the top position. However, whereas the superior firm gets "a second chance" even if it is placed at the bottom, the inferior firm does not, and thus it needs the top position more "desperately." The result is that the superior firm will win the bid only when its margin is significantly higher than that of the inferior firm.

Two more points are worth noting. First, as the equilibrium bids show, the size of the portion of informed consumers has no effect on which firm will be the winner. This may not be

obvious at first, as informed consumers will always click on the superior firm first, so the more such consumers, the better it is for the superior firm. However, recall that one key intuition is the incremental demand. Since the informed consumers do not change their click behavior in response to different link positioning, it is only the uninformed consumers that the two firms are competing for. Furthermore, for pay-per-impression, both firms will pay for both types of consumers anyway, regardless of whether and which links the consumers click. Therefore, although the size of the informed consumers changes the expected values of the top position to both firms, it changes them in a proportional manner that has no bearing on the outcome of the auction. We state this result in the following proposition.

Proposition 2: *In the pay-per-impression auction, the number of informed consumers in the population does not impact which firm wins the auction.*

The second point is that a higher quality premium does not make the superior firm more likely to be the winning bidder. Quite to the contrary, comparing Scenarios II and III shows that a higher quality premium may, in fact, make the inferior firm more likely to win the auction. Although surprising at first look, this can again be understood from the key intuitions of incremental value and residual demand –the higher the quality premium, the more likely that the superior firm will be searched even if it is placed at the bottom. This reduces the equilibrium bid of the superior firm, therefore making the inferior firm more competitive. We state this result in the following proposition.

Proposition 3: *In the pay-per-impression auction, under certain conditions, a higher quality premium for the superior firm makes the inferior firm more likely to win the auction.*

The analysis above also shows that in Scenarios III, IV and V, the per-unit margins of the firms influence which firm wins the auction. But, as we noted earlier, if we relax the assumption

that $m_S > m_I$, it will only become more likely for the inferior firm to be placed on top. In fact, if $m_S < m_I$, then the inferior firm intuitively has the characteristic of the high-valuation bidder, given its higher per-unit profit. The assumption $m_S > m_I$ thus serves to *raise* the bar for the inferior firm to win the auction and highlights the effects discussed above, by showing that they can bring about the position paradox even in the case where other factors work in the opposite direction.

3.2 Pay-Per-Click

We now analyze the case of the pay-per-click auction mechanism, where a firm pays the advertisement fee only when its link is clicked. The two key factors discussed in the pay-per-impression case – residual demand and incremental value – continue to apply here. In addition, there is another key factor, which changes the outcome of the auction compared to the pay-per-impression case and, in general, makes the superior firm even more likely to end up at the bottom position.

We call this effect the differential cost effect. Both firms know that some consumers are informed, and these consumers will start with the superior firm no matter how links are positioned. In the pay-per-click case, the inferior firm will pay for these consumers only if they actually click on its link. This is different from the pay-per-impression case, in which both firms pay as long as the links are displayed, i.e., both firms pay for search by every consumer. In other words, under pay-per-click the inferior firm will not unnecessarily pay for the informed consumers who never click on it. This reduces the expected cost per search of the inferior firm, thereby increasing its bidding capacity. In contrast, the superior firm will see the added fee (the additional amount needed to be at the top) paid for these consumers as pure waste, since these

consumers will click on the firm's link anyway. Hence, this effect, on the margin, increases the relative bidding power of the inferior firm.

Similar to the case of pay-per-impression, there are five scenarios depending on the values of parameters s, p, m_S, m_I, V and Q, each resulting in different optimal bids and link positions. In each scenario, the search behavior of consumers and the expected revenues of firms when either is placed on top are exactly the same as they are in the corresponding scenario in the pay-per-impression case. The only difference is on the cost side – in pay-per-click a firm pays its bid weighted by the probability of click (which is less than or equal to 1), whereas in pay-per-impression the firm pays its bid for the full mass of all consumers who search (which is equal to 1).

In the following, we highlight the basic insights by discussing Scenario III in detail. We simply summarize the results for the other scenarios and discuss them in detail in Technical Appendix TA1.1.

Scenario I: $s \leq p\min\{Q,V\}$**:** Since all consumers click on both firms, this is exactly the same as the pay-per-impression case.

Scenario II: $Q > V$ and $pV < s \leq pQ$: The inferior firm always emerges on top, but the superior firm obtains more clicks.

Scenario III: $Q < V$ and $pQ < s \leq pV$

In Scenario III, in which the search cost is higher than the expected quality premium but lower than the expected utility of the inferior firm, both firms prefer the top position. The expected revenues are as in the pay-per-impression mechanism, so we now look at the expected cost. If the superior firm is placed on top, the probability that it will be clicked is 1. If it is placed at the

bottom, this probability is $\phi+(1-\phi)(1-p)$, where the first part corresponds to informed consumers and the second part corresponds to the uninformed consumers who do not find a match at the inferior firm. If the inferior firm is placed on top, its probability of being clicked is $\phi(1-p)+(1-\phi)$, where the first part corresponds to informed consumers who do not find a match at the superior firm and the second part corresponds to uninformed consumers. If the inferior firm is placed at the bottom, the probability of it being clicked is $1-p$.

If the superior firm pays b_S per click and is placed on top (so the inferior firm pays $\underline{b}$ per click), then the expected profits of the firms are:

$$E[\Pi_{S,1}] = pm_S - b_S \text{ and } E[\Pi_{I,2}] = (1-p)(pm_I - \underline{b})$$

If the inferior firm pays b_I per click and is placed on top (so the superior firm pays $\underline{b}$ per click), then the expected profits of the firms are:

$$E[\Pi_{S,2}] = (\phi+(1-\phi)(1-p))(pm_S - \underline{b}) \text{ and } E[\Pi_{I,1}] = (\phi(1-p)+(1-\phi))(pm_I - b_I)$$

If the maximum amount the superior firm is willing to pay per click is given by b_S^*, then

$$pm_S - b_S^* = (\phi+(1-\phi)(1-p))(pm_S - \underline{b}) \Rightarrow\ b_S^* = (1-\phi)p^2 m_S + (1-(1-\phi)p)\underline{b},$$

and the maximum amount the inferior firm is willing to pay per click is given by b_I^*:

$$(\phi(1-p)+(1-\phi))(pm_I - b_I^*) = (1-p)(pm_I - \underline{b}) \Rightarrow\ b_I^* = \frac{(1-\phi)p^2 m_I + (1-p)\underline{b}}{1-\phi p}.$$

Therefore, in equilibrium, the superior firm will bid b_S^* and the inferior firm will bid b_I^*. The expressions above show that, depending on the parameters, either bid can be higher, i.e., either firm can win the auction. This is different from the pay-per-impression case, in which the

superior firm always wins in this scenario. Note that the differential cost effect is driving this – since the informed consumers will only click on the superior firm and the inferior firm does not have to pay for these consumers, it can bid higher for the clicks of the uninformed consumers. The expression for b_I^* also shows that, as ϕ increases, the inferior firm bids higher, consistent with this intuition. Furthermore, if the inferior firm wins in this scenario, then comparing the expressions for clicks, we find that the inferior firm will obtain more clicks than the superior firm only if $\phi < 1/2$, otherwise the superior firm will obtain more clicks even though it is at the bottom position.

Scenario IV: $s > p\max\{Q,V\}$ and $s \le p(V+Q)$: Depending on the values of the parameters, either firm may end up at the top position and, even if the inferior firm wins the auction, the superior firm may obtain more clicks.

Scenario V: $s > p(V+Q)$: Depending on the values of the parameters, either firm may end up at the top position and, even if the inferior firm wins the auction, the superior firm may obtain more clicks.

Summarizing the Scenarios

The results of the five scenarios discussed above are summarized in the following proposition:

Proposition 4: *In the pay-per-click bidding game:*

- *If the search cost is very low (Scenario I), both firms will bid an equal amount.*
- *If the search cost is moderately low and the quality premium is high (Scenario II), the inferior firm will bid higher.*
- *In all other scenarios (Scenarios III, IV and V), either firm may bid higher.*

For each scenario, the equilibrium bids by the two firms are presented in Table 2.

Table 2: Equilibrium Bids under Pay-Per-Click

Scenario	**Firm S bid** **Firm I bid**
$s \le p\min\{Q,V\}$	$\underline{b}$ $\underline{b}$
$Q > V$ and $pV < s \le pQ$	$\underline{b}$ $(1-p)pm_I$
$Q < V$ and $pQ < s \le pV$	$(1-\phi)p^2 m_S + (1-(1-\phi)p)\underline{b}$ $((1-\phi)p^2 m_I + (1-p)\underline{b})/(1-\phi p)$
$p\max\{Q,V\} < s \le p(V+Q)$	$(1-\phi)p^2 m_S + (1-(1-\phi)p)\underline{b}$ pm_I
$s > p(V+Q)$	$(1-\phi)pm_S + \phi\underline{b}$ pm_I

The situation of low search cost and that of moderately low search cost with high quality premium are similar to that in the pay-per-impression case. The other three scenarios are qualitatively different. As we can see, in the pay-per-click case the inferior firm is more likely to be the higher bidder than in the pay-per-impression case: while in pay-per-impression the superior firm is guaranteed to be the winner in Scenarios III and V, in pay-per-click the inferior firm may be the winner in these situations as well. This result can be readily deduced from Propositions 1 and 4; we state it below as a corollary to these propositions.

Corollary 1: *The inferior firm will win the position auction for a larger range of parameter values in the pay-per-click auction as compared to the pay-per-impression auction.*

Under the pay-per-click scheme, the existence of informed consumers reduces the willingness to pay of the superior firm but does not affect that of the inferior firm, thereby increasing the relative bidding power of the latter. Furthermore, as we have already discussed above, as the number of informed consumers increases, the inferior firm bids higher and is more

likely to win the position auction. Hence, we obtain a somewhat ironical result – the more widespread the reputation of the superior firm, the more likely it is that the inferior firm will win the position auction. We state this result in the following proposition.

Proposition 5: *In the pay-per-click auction, as the fraction of informed consumers in the population increases, the inferior firm will bid higher and is more likely to win the auction and be placed on top.*

4. Extension to Heterogeneous Search Costs

In the basic model, we assumed that all consumers have the same search cost s. In reality, however, we may expect that there is heterogeneity in search costs across consumers. In this section, we extend our study to the case where consumers are differentiated in their search costs.

We assume that both the informed consumers and the uninformed consumers are distributed uniformly in their search costs on the interval $[0,1]$. Recall that each consumer, informed or uninformed, falls into one of the five scenarios discussed earlier. Hence, to make the search cost scale meaningful and ensure that all the five scenarios in the previous section fall into this scale, we normalize the other parameters accordingly by assuming that $p(V+Q) \le 1$. In this case, firms will choose the bidding strategy to maximize expected profit across all consumers. The assumption of uniform distribution is made solely for the ease of exposition. Qualitatively, the results in the following sections hold in the case of arbitrary distribution of consumers, as the intuitions remain the same.

4.1 Pay-Per-Impression

We start with pay-per-impression. To facilitate the discussion, we label each segment of the consumers using the scenario number they belong to (e.g., consumers with search cost

$s \le p\min\{Q,V\}$ are labeled as Segment I consumers since they fall under Scenario I). In equilibrium, each firm simply bids the incremental value of the top position relative to the bottom one, plus the minimum required bid. The incremental value of the top position is the weighted average of that value for each segment of the consumers. We state the result in the following proposition. The technical details are presented in the Technical Appendix TA1.2.

Proposition 6: *In the pay-per-impression auction with uniformly distributed search costs for consumers, the superior firm will bid:*

$$b_S^* = p(1-p(1-p)V - pQ)(1-\phi)m_S + \underline{b}$$

while the inferior firm will bid:

$$b_I^* = p(1-p(1-p)V - p^2Q)(1-\phi)m_I + \underline{b}$$

Depending on the parameter values, either firm's bid may be higher, so either firm may win the auction and take the top position.

We now discuss some comparative statics results. First, note that the higher m_S (m_I) is, the more likely it is that the superior firm (inferior firm) will win the auction. This is easy to understand, as the higher a firm's profit margin is, the more valuable the top position is to the firm. Second, we can see that the higher the quality premium Q is, the more likely it is that the inferior firm will win the auction. The reason is the same as explained in the earlier discussion when search cost is not differentiated. Third, we can also see that the higher V is, the more likely it is that the inferior firm will win the auction. To understand the intuition behind this, note that changing V changes the relative size of each consumer segment. If $V \ge Q$, increasing V increases the size of the Segment III of consumers while decreasing that of the Segment V. Both

firms would like to bid lower for consumers in Segment III than for those in Segment V, but the reduction in bid is higher for the superior firm. Therefore, increasing V reduces the relative bid of the superior firm and makes it more likely for the inferior firm to win. If $V < Q$, increasing V increases the size of Segments I and IV while decreasing that of Segments II and V. The superior firm would like to bid lower for Segment IV consumers than for Segment V consumers, while bid the same for Segment I and II consumers. The inferior firm would like to bid lower for Segment I consumers than for Segment II consumers, while bid the same for Segments IV and V consumers. Overall, the reduction in bid is higher for the superior firm, so increasing V again reduces the relative bid of the superior firm and makes it more likely for the inferior firm to win the auction. We state this result in the following proposition.

Proposition 7: *In the pay-per-impression auction with uniformly distributed search costs for consumers, the higher the base product value (V) or quality premium (Q), the more likely it is that the inferior firm will win the auction.*

Finally, we note that the size of the informed consumers does not change the likelihood of either firm winning the auction, as both firms are effectively competing only for the uninformed consumers, while the informed consumers add to both firms' costs proportionally. This result is similar to the result in Proposition 2. However, it is a stronger result, since it holds for the more general case of heterogeneous search costs for consumers. We highlight this result by stating it below as a remark.

Remark: *In the pay-per-impression auction with uniformly distributed search costs for consumers, the number of informed consumers in the population does not impact which firm wins the auction.*

4.2 Pay-Per-Click

We now analyze the pay-per-click case. Using a similar approach as in the pay-per-impression case, we can derive the optimal bids for both firms. We state the result in the following proposition, while the technical details can be found in Technical Appendix TA1.2.

Proposition 8: *In the pay-per-click auction with uniformly distributed search costs for consumers, the superior firm will bid* $b_S^* = p(1-p(1-p)V - pQ)(1-\phi)m_S + (\phi + p(1-p)(1-\phi)V + p(1-\phi)Q)\underline{b}$. *If* $V \geq Q$, *the inferior firm will bid* $b_I^{V*} = \frac{p(1-p(1-p)V-p^2Q)(1-\phi)m_I + p(1-p)V\underline{b}}{1-\phi+\phi p(1-p)V}$, *while if* $V < Q$, *the inferior firm will bid* $b_I^{Q*} = \frac{p(1-p(1-p)V-p^2Q)(1-\phi)m_I + p(1-p)V\underline{b}}{(1-\phi)(1+p^2V-p^2Q)+\phi p(1-p)V}$. *Depending on the parameter values, either firm's bid may be higher, so either firm may win the auction and take the top position.*

As before, the case of pay-per-click is, in general, similar to that of pay-per-impression. The significant difference concerns the size of the informed consumers – the larger the size of informed consumers, ϕ, is, the more likely it is that the inferior firm will win the auction. This is because the superior firm needs to pay for the informed consumers, while the inferior firm does not, unless these consumers click on its link. This increases the relative bidding power of the inferior firm, making it more likely to win. This result is a generalization of Proposition 3 and we highlight it by stating it below as a remark.

Remark: *In the pay-per-click auction with uniformly distributed search costs for consumers, as the fraction of informed consumers in the population increases, the inferior firm will bid higher and is more likely to win the auction and be placed on top.*

4.3 Search Engine Profit Comparison

We now compare the search engine profits under pay-per-impression and under pay-per-click. Under pay-per-impression, the search engine profit is determined only by firm bids. With the second price auction, search engine profit is simply the minimum bid $\underline{b}$ paid for the bottom link plus the lower one of the bids b_S^* and b_I^*, paid for the top link, where b_S^* and b_I^* are as stated in Proposition 6. Under pay-per-click, the probabilities of users clicking on the links also need to be accounted for. The expected profit of the search engine is $\Pi_{SE} = P_T \min\{b_S^*, b_I^*\} + P_B \underline{b}$, where P_T is the probability that the top link will be clicked by a consumer, P_B is the probability that the bottom link will be clicked by a consumer, and b_S^* and b_I^* are as stated in Proposition 8 (b_I^* corresponds to b_I^{V*} when $V \geq Q$, and b_I^{Q*} otherwise). The click probabilities in each scenario can be calculated as before.

The comparison of search engine profit between pay-per-impression and pay-per-click is stated in the proposition below. (To simplify the analysis when comparing profits, we assume that the exogenous minimum bid is very close to zero, i.e., $\underline{b} \to 0$.)

Proposition 9: *If $V \geq Q$, then if*

$m_S(1-p(1-p)V-pQ)(1-\phi+\phi(1-p)pV) > m_I(1-p(1-p)V-p^2Q)$, *search engine profit is higher under pay-per-click, otherwise the search engine profit is higher under pay-per-impression. If $V < Q$, then if*

$m_S(1-p(1-p)V-pQ)((1-\phi)(1-p^2Q+p^2V)+\phi(1-p)pV) > m_I(1-p(1-p)V-p^2Q)$, *search engine profit is higher under pay-per-click, otherwise the search engine profit is higher under pay-per-impression*.

The result can be alternatively stated, in a more intuitive way, as: if the condition is such that the superior firm will win the auction under pay-per-click, then pay-per-click brings higher profit to the search engine than pay-per-impression, while if the inferior firm will win the auction under pay-per-click, then pay-per-impression brings higher profit. The basic insight behind this result can be understood as follows. With the second price auction, search engine profit is determined by the lower bid. Hence, the mechanism that leads to a more aggressive lower bid will yield more profit for the search engine. With the presence of informed consumers, the inferior firm will bid more aggressively under pay-per-click than under pay-per-impression, while the reverse is true for the superior firm. Therefore, when the superior firm will give the lower bid, the profit is higher under pay-per-impression, while when the inferior firm gives the lower bid, the profit is higher under pay-per-click. The search engine is better off with the mechanism that intensifies the competition between the two firms for positions, and either pay-per-impression or pay-per-click can be that mechanism depending on the situation. Again, the difference between the two mechanisms lies in the presence of informed consumers. It can be shown that when there are no informed consumers (i.e. $\phi = 0$), search engine profit is the same under both pay-per-impression and pay-per-click.

5. Empirical Support

To empirically validate our model's predictions, we obtained a database of sponsored search advertisements from a leading search engine firm in Korea. For a given keyword, the search engine uses a pay-per-impression position auction to sell up to five different advertising positions in the sponsored list to potential advertisers. For a given keyword, the data consist of the daily positions of the advertisers and the corresponding daily impressions (i.e., the number of times the keyword was searched) and click counts at each position over a 15-day period in July

2008. Our dataset is unique in that we have click counts at all positions for each keyword, while most previous empirical studies have click counts only for one advertiser.

We have the exact URL of each advertiser when the keyword is searched. However, we do not have data on quality scores of advertisers. Therefore, we use a firm's reputation to impute the quality of a given web link. We hired three independent annotators from the United States and Korea to assess, for each keyword, whether each advertiser is a high-quality or a low-quality firm. Among 246 keywords provided for this research, we excluded keywords if annotators could not recognize the firms by their names, and only considered keywords for which annotators could confidently classify the qualities of the firms. We also excluded keywords when only one advertiser was displayed during the data period. We then computed the proportion of agreements on the quality of the firms across keywords between the annotators. The inter-rater reliability score ranged from 0.90 to 0.94, indicating a very high level of reliability. From the high-quality firms in a given keyword, we selected the best-ranked firm among them as the superior firm. Likewise, from the low-quality firms, we selected the best-ranked firm among the low-quality firms as the inferior firm. Finally, we ended up with a total of 102 keywords and categorized each into one of the following three different configurations based on the average number of clicks over the data period:

C1: The superior firm is above the inferior firm and the superior firm obtains more clicks.

C2: The inferior firm is above the superior firm and the inferior firm obtains more clicks.

C3: The inferior firm is above the superior firm but the superior firm obtains more clicks.

We find that all three configurations above have significant representation in our dataset. Out of 102 keywords, we find 48, 25 and 27 keywords in C1, C2 and C3, respectively. Table 3

shows the daily average number of clicks corresponding to the three different configurations. The extant literature focuses on C1 as an equilibrium configuration, while our theoretical model predicts that C2 and C3 can also arise in equilibrium. We observe that C2 and C3 occur in 25% and 27% of all cases, respectively, and this offers direct confirmation of our results. Note that the data from a pay-per-impression auction provides a conservative test of the position paradox because, as shown by our analytical model, it is less likely under the pay-per-impression auction as compared to the pay-per-click auction.

Table 3: Daily Average Number of Clicks

Quality	*C1*	*C2*	*C3*	Total
High	16.80	4.18	13.34	12.68
Low	6.37	11.58	5.42	7.42

After establishing the existence of the position paradox outcome, we now use our analytical model to derive two sharp predictions related to click-through rates (CTRs) in configurations C1, C2 and C3, and then test them on our data. Since our data are from a pay-per-impression auction, we draw upon our results in Section 3.1. Our theory based on consumer search cost provides specific predictions regarding which configuration (C1, C2 and/or C3) can occur under which search cost scenario (Scenarios I to V). Note that the search cost s increases from Scenario I to Scenario V. As the discussion in Section 3.1 shows, C1 can occur under Scenarios I, III, IV and V, C2 only under Scenario IV, and C3 under Scenarios I, II and IV. Furthermore, each keyword in our data is categorized into one of the three configurations C1, C2 or C3. Assuming that different keywords (corresponding to different products/categories) have different search costs for consumers, and given that higher search cost will lead to fewer clicks on the sponsored list, the result in Section 3.1 provides the following two predictions.

P1: Average CTR across all keywords categorized into C2 should be the smaller than the

average CTRs across all keywords categorized into C1 and into C3.

P2: The dispersion in CTRs across all keywords categorized into C2 should be the smaller than the dispersion in CTRs across all keywords categorized into C1 and into C3.

The reasoning behind P1 is that C2 can occur only in a high-search-cost scenario, while C1 and C3 can occur in low-search-cost scenarios as well. The reasoning behind P2 is that C2 can occur in only one search-cost scenario, while C1 and C3 can occur in multiple search-cost scenarios.

To test these predictions, we calculate the average CTRs and the standard deviation (dispersion) in CTRs across all keywords categorized into each configuration. The values are reported in Table 4. From the first row, we can see that prediction P1 regarding configuration C2 having the lowest average CTR holds. From the second row, we can see that prediction P2 regarding configuration C2 having the least dispersion in CTRs holds. Furthermore, from Table tab:hypotheses(a) we can also conjecture that the mean and standard deviation numbers for configurations C1 and C3 should be similar, which we can see from Table 4 is indeed the case.

Table 4: Click Statistics by Keyword Configuration

	C1	*C2*	C3
Average CTR	0.10	0.05	0.09
Std. dev. Of CTR	0.13	0.07	0.13

To summarize, the empirical analysis offers indirect support of our theory by showing that: (i) a large proportion of auction outcomes in the data show the position paradox, and (ii) two sharp predictions from our analytical model are validated in the data.

6. Discussion and Conclusion

Surfers on the World Wide Web typically rely on search engines to direct them to web sites that

contain content of immediate interest to them. In sponsored search advertising, firms bid for links of their web sites to be displayed in response to keywords that consumers search. This provides a highly targeted advertising medium for firms and, therefore, sponsored search has been hugely successful as an online advertising medium.

The unprecedented rise of sponsored search has fueled recent academic study into it. A main result from both empirical and theoretical work is that advertisements at higher positions attract more clicks from consumers (e.g., Misra et al. 2006, Feng et al. 2007) and higher-quality firms will be placed at higher positions (e.g., Athey and Ellison 2011). In this paper, we study bidding strategies for firms when they offer products of different quality levels and some consumers know the reputations of firms, i.e., some consumers know which firm offers the higher-quality product. In this scenario, among several other results, we find the existence of an important paradoxical outcome of the position auction in which a high-quality firm bids less than a low-quality firm to be placed below it, yet still obtains more clicks than the low-quality firm.

We show this position paradox in the pay-per-impression mechanism and also show that it is more likely to happen in the pay-per-click mechanism. Furthermore, we show that it is more likely to happen when the quality differential between the two firms is larger and when the high-quality firm has a better reputation (i.e., more consumers are informed of its identity). We extend our results to the case of heterogeneous search costs for consumers and also derive implications for search engine profit under different mechanisms. To support our theoretical analysis, we show strong empirical support for the basic position paradox result in a rich dataset obtained from a leading South Korean search engine.

We focus on three key insights with respect to sponsored search. The residual demand effect implies that consumers may continue searching if they do not find a product to their

satisfaction. The incremental value effect implies that, for a firm, a higher position is worth only the extra revenue it generates over the position below it, not the absolute revenue it generates. Together, these effects imply that an inferior firm may value a higher position much more than a lower one, while in contrast a superior firm may be confident that even if it is placed lower, it will get a sufficient mass of residual consumers because it has a high quality product. In other words, if the residual demand effect is sufficiently strong for a superior firm, an inferior firm may have more incremental value for the top position than the superior firm and may outbid it. The differential cost effect, which is relevant under the pay-per-click mechanism where a firm has to pay only when a consumer clicks on its link, implies that different types of consumers have different relative cost implications to firms. Specifically, the inferior firm only has to pay for a fraction of the population (i.e. only the uninformed consumers) and can bid higher as compared to the pay-per-impression auction, and win more often.

We now discuss some limitations of our analysis and some avenues for future work. First, we have analyzed two popular auction mechanisms, pay-per-impression and pay-per-click, in detail. However, as mentioned before, many search engines such as Google and Yahoo! control for the quality of the bidding firms by weighting firms' bids with weights that are usually positively correlated with their qualities. The intuition here is that a higher-quality firm can generate more clicks and, to increase revenues, the search engine should favor it for positions that generate higher revenue per click. In our model, such an adjustment of the bids will make it less likely that the inferior firm is placed at the top. However, the position paradox can still occur in this case, although for a reduced range of the parameter values, as the key drivers of the result remain intact. A main challenge in modeling this scenario realistically is that the exact quality scores used by search engines are not publically released. As more information on this aspect becomes

available, future work can look into this mechanism in more detail.

Second, mechanism design for position auctions is not in the scope of our paper. For completeness of the current study, we compare search engine profits under the pay-per-impression and pay-per-click mechanisms and find that, under different conditions, either mechanism can yield higher profit. Mechanism design with firm reputation effects can be a fruitful area for future work and such studies can also consider other practices in more detail, such as click weighting mentioned above.

Third, we consider two firms bidding for two positions, while there are usually multiple firms bidding for multiple positions in a real sponsored search auction. However, the three basic effects that drive our results (residual demand, incremental value and selective payment) are robust effects that should carry over to a multiple-firms setting. The complete analysis for such a case is extremely complicated due to combinatorial explosion in the number of cases to consider in consumer search. However, for a three-firm case, we have confirmed that a position paradox in which superior firms are placed below inferior firms but obtain more clicks can still exist (analysis not presented here). Future work can address the multiple-firms case with firm reputations and consumer search in greater detail.

Fourth, the auction is modeled as a one-shot game for simplicity, as do almost all other theoretical papers on this subject. In our case, this assumption is reasonable since the search engine we obtain data from conducts its auctions at most once every day. In fact, we find that firm positions are very stable in the time period for which we have data. However, many search engines such as Google and Yahoo! conduct position auctions on a continuous basis, i.e., any bidder can change its bid at any time. Modeling a repeated position auction presents an interesting opportunity for future research.

Finally, the position paradox is generated because consumers lack information on the quality of products offered by firms and search sequentially for their desired product. This indicates that we should observe this paradox to be stronger for keywords that are related to markets in which consumers have high uncertainty about firm offerings. Exploring the relationships between keyword and market characteristics and the outcome of sponsored search auctions can be a very fruitful area for future research.

Chapter 3

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

1. Introduction

The advent of Information Technology has enabled the collection of data with ever increasing scales at finer granularity. Detailed communication and decision data provide an exciting opportunity to further the understanding of behaviors, especially in a social context. Electronically enabled social networks allow consumers to communicate more efficiently amongst themselves, while at the same time offer the means for researchers to observe social networks and interactions. Knowledge about the structure of a social network provides marketers with a distinct opportunity to better understand their customers and improve their promotional decisions. In this study, we leverage the data from an electronic social network to investigate a question of particular importance to marketers: what drives purchase decision similarity among friends?

Consider a firm that is selling products within a social network. It makes one sale to a customer and then makes a second sale to the customer's friend. The key question is: did the second sale occur because one customer influenced the other, or because these two customers have similar tastes, since after all, they are friends? The answer to this question is critically important to the firm's marketing strategy, and is the focus of our study. We refer to the former case where a customer influences their friends to purchase as *social influence*. If social influence is responsible for the purchase, then the firm may want to incentivize customers to promote the

product to her friends using a referral bonus program. If it is the latter case of similarity in tastes then the firm could rely on the social network to help identify new potential customers by targeting the friends of the customer. Social scientists have long recognized that people with similar characteristics are more likely to form ties, an effect termed as *homophily* (McPherson et al. 2001). Consequently, people who have close ties tend to have similar traits.

Social networks, social influence, and homophily have long been a topic of interest to sociologists, economists, and marketers. The advent of information technology has enabled the gathering and processing of large scale network data, leading to a growing number of studies on social networks in various fields such as economics (Jackson and Watts 2002), marketing (Hartmann et al. 2008), information systems (Hill et al. 2006), and machine learning (Zheng et al. 2008). This literature is concerned with both the formation of social networks (Ansari et al. 2011, Braun and Bonfrer 2010) 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 effect: social interactions (Hartmann 2010), peer effects (Nair et al. 2010), contagion (Van den Bulte and Lilien 2001, Iyengar et al 2010), conformity (Bernheim 1994), imitation (Bass 1969, Choi et al 2008), and neighborhood effects (Bell and Song 2007). The different terms may have subtle differences but they all describe the dependence of one's decisions on those of others, an effect we term as social influence in our study at the general level. Recently, this influence effect has received attention from researchers in economics and marketing where it has been studied in the context of diffusion (Van den Bulte and Stremersch 2004) and word-of-mouth (Godes et al. 2005). Structural models have been used to try to uncover the detailed causal effects behind the

observed influence. Hartmann (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 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 among physicians and show that the opinions of influentials have a great impact in the prescription decisions of other physicians. However, it is unclear whether the focus should be only on the opinion leader. For instance Watts and Dodds (2007) show that although influentials can trigger large-scale "cascades" in certain situations, in many cases change is simply driven by easily influenced individuals who sway other easily influenced individuals.

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

ceteris paribus, we should observe that people with ties tend to behave similarly. Indeed, this effect has been used as the basis for improving marketing forecasts (Hill et al. 2006).

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

Dynamic data may allow one to separate these effects but care still must be taken in decomposing these effects. Nair at al. (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 to account for homophily on product taste. This approach is similar to the one that we propose in our study. However, Hartmann (2010) focuses on group coordination and

does not elaborate on the role that homophily could play in purchase behavior. Furthermore, homophily suggests that network neighbors may be similar on most decision-relevant characteristics, instead of just base-level product taste. In our study, we model homophily on all decision-relevant characteristics: product taste, purchase interval, and intrinsic susceptibility to influence on purchase incidence and product choice. Finally, Hartmann (2010) studies coordinated consumption, while the product used in our study is individual consumption goods, and the influence effect is asymmetric and comes from communication.

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

Our model, which is discussed in section 2, is framed within the context of a hierarchical Bayesian model which simultaneously incorporates both the homophily and the social influence

effect in the purchasing decision processes of consumers. It accounts for both the timing of purchase and product choice. Social influence is allowed on both purchase time and brand choice decisions, while the impact of homophily is measured for product taste, purchase interval, and susceptibility to influence parameters. We discuss these issues in more depth in section 3.

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

Our study contributes to the literature by simultaneously quantifying both the homophily and the social influence effect in product purchase decisions by consumers in a social network context, using a unique, large-scale, real word dataset. Our study is among the first to quantify the impact of homophily and influence jointly in the decision process of consumers. Social network researchers have long recognized the importance of distinguishing homophily from

social influence, and the difficulty to do so. Although observed variables can be used proxy for homophily effect (Aral et al 2009), they do not shed light on the similarities of unobserved characteristics, such as product taste. Such characteristics, although not observed, are nonetheless important decision drivers, and their similarities among friends are crucial for business managers to understand. Our study thus fills the gap in literature by jointly quantifying these similarities, i.e. homophily effect, with social influence effect that arises from communication. The study is made possible by the unique dataset which contains both communication and product purchase information over time – firms in industry are increasingly gaining access to such detailed data with the advent of technology. By developing a hierarchical model and applying it to this dataset, we are able to evaluate the relative magnitude of both factors in both purchase-timing and product-choice decisions. We show that models which ignore one of the factors result in the overestimation of the other factor. We also distinguish in-group influence from out-group influence. Finally, we demonstrate the importance of communication data by demonstrating the bias in estimation that would arise if communication is not explicitly accounted for.

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

2. Modeling Product Purchase within a Social Network

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

Our dataset also contains detailed transaction records about caller ring-back tones (CRBT) purchased by these customers. These tones are usually short snippets of musical songs.[6] To understand its functioning considers what happens when customer *A* purchases a certain ring-back tone. Once another person *B* calls *A* then *B* will hear the ring-back tone instead of the usual ringing. Notice that only customer *A*'s callers hear this tone and not customer *A*. To use the CRBT feature, a customer must pay a monthly subscription fee, select the individual tone that he wants played when he is called. Each tone a customer purchases is valid for 90 days, but a customer can change the tone by purchasing a different one at any time. About seven-hundred fifty-thousand customers purchase ring-back tones during our time frame, or roughly 20% of customers. The type of tones that are selected for purchase and when they are selected for purchase forms the basis for the consumer purchase decision in our problem.

There are two steps in our purchase decision: 1) when to buy and 2) what to buy. This dichotomy follows other that focus upon consumer purchases (Chintagunta 1993). We model the

[6] CRBT is a popular phone feature in a number of Asian countries including India, although it is not presently available in the American market.

first step, the *when-to-buy* decision, using a model of inter-purchase timing with its corresponding hazard rate. The second step, the *what-to-buy* decision, is modeled using a discrete choice model. In the following sub-section we describe our model for this two-step decision process, and then discuss in the following subsection how to captures homophily and the social influence effects.

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

2.1 Purchase Timing Model

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

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

The rate is allowed to vary through time to reflect the chance that consumers become more (or less) likely to buy if they are exposed to others that use the product. A customer is exposed to CRBT either through calling others inside his social group or from calling those outside his

social group. If a friend hears another friend's ring-back tone we would expect this to be more influential than hearing a tone for someone outside his social group. We denote $E_{gi,t,k}$ as the amount of exposure that consumer *gi* had at period *t* from either inside ($k = In$) or outside ($k = Out$) his group. (We discuss the construction of the exposure variable further in section 3.5.) We propose the following model of the purchase rate parameter as a function of cumulative product exposure from both inside and outside the group:

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

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

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

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

$\bar{\lambda}$ is the population level average base rate, and σ_λ^2 measures the dispersion of this base rate parameter in the population. We choose a lognormal hyper-prior for $\bar{\lambda}$, and an inverse-Gamma hyper-prior for σ_λ^2, both of which are conjugate priors. Both hyper-priors are chosen to be

diffuse, as we have little knowledge of them ex-ante. The correlation parameter r_λ must be within the interval $[\underline{r},1]$, where $\underline{r}$ is the smallest number to make the correlation matrix positive-definite. We choose a uniform hyper-prior for r_λ. When homophily exists we expect the values of consumers in the same group to be positively correlated, or $r_\lambda > 0$ is an indication of homophily.

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

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

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

2.2 Product Choice Model

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

product j at time period t is the sum of the product characteristics, cumulative exposure, and a random error:

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

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

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

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

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

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

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

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

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

3. Discussion Concerning the Data and Model

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

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

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

3.1 Defining Groups within the Social Network

In the model a group consists of people who have a close relationship with one another. The homophily effect within a group should remain stable over a short period of time, such as the three months that we observe in our data, therefore it is important that we identify stable relationships. We believe that people who call each other frequently are likely to have a close relationship. To ensure that we capture true relationships rather than sporadic phone calls, we consider two customers as belonging to a group only if they made at least five phone calls in the first month of the three-month period to one another. The choice of five phone calls as the threshold is a subjective one on our part, but we consider it as a way to delineate close contacts.[7] If the threshold is set to too low a value (e.g., one or two) then the network is contaminated with many contacts that are not part of the caller's social circle. If we increase the threshold (e.g., ten or more) then we substantially reduce the number of groups that are formed. Our selection of five phone calls was meant to achieve a balance between these two factors. Furthermore, it is

[7] 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.

necessary to uncover only the *existence* of connections but not their *strength*, since we can infer strength through our model's social influence effects.

This produces an undirected network or graph, where each node corresponds to a customer, and there is an edge between two nodes if the two corresponding customers have close relationship. To identify groups in this network, we then run a clique-searching algorithm to find all four-cliques in the graph.[8] A four-clique is a sub-graph with four nodes, where every node is connected to every other node in the sub-graph. Each four-clique then corresponds to a group of four customers, where each customer has a close relationship with every other customer in the group. We identify a total of 1,654 groups in our dataset. In order to shorten our estimation time we randomly select 300 of these groups.

The choice of four in defining our groups is another subjective decision on our part. Given that a clique implies everyone in the group is connected to everyone else, a group of larger size implies stronger connections between members within the groups but fewer groups of larger size. If we choose a smaller size (e.g., two) then our group effect would represent only dyadic relationships and may not capture more general social processes. Our desire is to keep the modeling framework simple by conditioning upon the network structure. The choice of phone call thresholds and clique size could be nested within our model, but at the expense of a more complex model. Note that the four-cliques may be embedded in cliques of larger size. That is, a group of four customers where everyone is connected to everyone else may be a subset of a larger group where all are pairwise connected. However, this is not an issue for the homophily effect, the condition of which is the existence of connections but it does not need to be exclusive connections. In another word, if the four customers have similar preference because they are

[8] 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.

friends, such similarity will not disappear simply because they all are friends of yet another person. This may be an issue if the strength of ties and differential degrees of similarities are accounted for, which we leave for future research. The embedding issue is a concern on the influence effect, as out-group communications can still come from a connected person, carrying influence effect of in-group magnitude. However, in our model this will bias the estimate of out-group influence towards the in-group influence, making it harder to find difference between the two. Therefore, this bias makes it harder for us to draw conclusions comparing the influence effect, thus strengthening the findings we report.

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

3.2 Product Choice

Ideally, each caller ring-back tone could be treated as an alternative in our choice model. However, there are more than eleven thousand different tones that have been purchased by customers, most of which were purchased by only a few people. This sparse dataset makes it difficult to estimate the parameters and to interpret the results at the individual tone level.

Therefore we have chosen to categorize all the tones to ten different genres. These genres were gathered from the telecom's website which provided our dataset. We were able to categorize seven thousand tones. Upon further analysis we found that only three of the ten genres had more than 5% of market share. Therefore, we chose to combine the tones from the small genres and uncategorized tones into an "Other" category. This results in four "products" or categories or genres of tones. For simplicity we refer to the choice of a tone within a music genre or category as a product. Table 1 lists these product IDs and their respective market share. Unfortunately we possess little information about the tones and music genres, so we are only able to define product dummies in the product characteristic vector.

Table 1: CRBT Market Share by Category

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

3.3 Evidence of Similarity

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

Table 2: Probability of Purchasing by Category

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

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

3.4 Understanding Social Influence with Caller Ring-Back Tones

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

The influence a customer imposes on a caller is conveniently encoded in their communication records within the telecom network. Whenever a person places a call to a customer with a certain ring-back tone, we can infer that the caller has heard the tone from the callee. This caller is automatically exposed to two things. First, the caller immediately perceives that the person is using the CRBT service. Second, this caller is exposed to the product and more

precisely the customer's chosen tone. The social influence argument suggests that both the purchase-timing decision and product-choice decision may be influenced through exposures resulting from phone calls. We thus quantify this external influence based on the phone calls made by the customer. As both the phone call records and the ring-back tone purchase records are time-stamped, we can infer how many times a customer is exposed to our products within a certain period. As stated earlier, in our study we treat social influence as the dependence of one's decision on those of others at the general level, for which such communication data is sufficient. At the detailed level of behavioral factors, we note that the influence in our study is of passive nature – the influence customer A "exerts" to customer B is not done through explicit persuasion from A to B, but through passive observation by B of A. This passive effect may arise out of either observational learning or imitation among other factors. Distinguishing these underlying factors is left for future study.

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

3.5 Quantifying Exposure

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

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

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

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

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

4. Empirical Results

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

4.1 Purchase Timing

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

Table 3: Purchase-Timing Model Parameter Estimation

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

$r_{\gamma Out}$	0.192	0.235	-0.213	0.604
$\overline{\kappa}^{pi}$	0.736	0.0448	0.677	0.850

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

4.2 Product Choice

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

someone outside the group has fairly strong positive effect on choosing the same category—the choice probability increased by about 20%. This effect is statistically significant.

Table 4: Product-Choice Model Parameter Estimation

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

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

and may cancel each other out. This may explain why our in-group influence effect is not statistically different than zero.

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

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

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

4.3 Model Comparison

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

Table 5: Purchase-Timing Parameter – Influence Only

Parameter	Posterior Mean	Posterior Standard Deviation	2.5% Posterior Quantile	97.5% Posterior Quantile
$\bar{\lambda}$	0.0325	0.00121	0.0302	0.0349
$\bar{\gamma}_{In}$	1.52	0.313	1.01	2.02
$\bar{\gamma}_{Out}$	0.682	0.162	0.377	1.01
σ^2_{λ}	0.342	0.0304	0.286	0.404
$\sigma^2_{\gamma In}$	0.663	0.334	0.222	1.38
$\sigma^2_{\gamma Out}$	0.557	0.229	0.234	1.09
$\bar{\kappa}^{pi}$	0.747	0.0525	0.645	0.827

Table 6: Product-Choice Model Parameter – Influence Only

Parameter	Posterior Mean	Posterior	2.5% Posterior	97.5% Posterior

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

Table 7: Purchase-Timing Model Parameter – Homophily Only

Parameter	Posterior Mean	Posterior Standard Deviation	2.5% Posterior Quantile	97.5% Posterior Quantile
$\overline{\lambda}$	0.0350	0.00138	0.0324	0.0377
σ^2_λ	0.360	0.0353	0.294	0.430
r_λ	0.109	0.0521	0.0156	0.220

Table 8: Product-Choice Model Parameter – Homophily Only

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

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

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

Table 9: Model Comparison – Likelihood

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

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

4.4 The Role of Communication Data

Communication is necessary for influence. For a customer A's action to influence customer B's decision, not only does A need to take the action, the knowledge of action must be conveyed to B as well. Data on communication is thus crucial for accurate assessment of influence effect. Detailed communication data, however, is rare in datasets available to researchers. Consequently, existing research usually accounts for influence by making one's decision directly dependent on the decision of others (e.g. Nair et al. 2010, Bell and Song 2007, Iyengar et al 2010), with an implicit assumption that one's action perfectly observed by others.[9]

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

Table 10: Purchase-Timing – Compare with Missing Communication

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

[9] This could be a reasonable assumption if the data is at aggregate level due to law of large numbers, or if each time period is sufficiently long so that with high confidence one's decision has been conveyed to others in the time period. The reasonableness of the assumption should be evaluated on a case-by-case basis.

Table 11: Product Choice – Compare with Missing Communication

Parameter	Full Model	"No Communication" Model I	"No Communication" Model II
$\overline{\rho}_{In}$	0.0857	0.1384	0.500
r_{β_1}	0.773	0.745	0.822
r_{β_2}	0.317	0.530	0.370
r_{β_3}	0.523	0.767	0.430

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

The estimated influence effect for the no-communication model, presented in Table 11, is larger than that in the full model. This again is because of miscounting of exposure, either an exposure is counted once every day a customer possesses a tone (model I) or once only when the customer purchases a tone (model II), whereas in reality there may be multiple phone calls on one day and none on another. The total amount of exposure can thus be either over- or

[10] When communication data is not used, out-group exposure cannot be quantified, and is thus left out of the equations for the no-communication model.

understated (this is in contrast to purchase timing decision where the exposure amount can only be overstated, as each song change is observed at most once). With mis-timed exposure which could be either over- or understated, the influence estimate could be biased either up- or downward. Also shown in Table 11 is that the homophily effect is overestimated in both model I and II, compared with the full model. This may be because the influence effect is mistakenly loaded onto the intrinsic product taste parameter, when communication is not accounted for. Both estimates clearly demonstrate the crucial importance of detailed communication data on accurately measuring influence effects.

5. Policy Simulation

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

5.1 Targeting Known Customers

The first policy simulation evaluates the effectiveness of targeting promotions to existing customers. The objective of the firm is to increase the purchase probability of its existing customer base over a certain period of time and promote the sales of a specific category.[11] To do so, it is assumed that the firm has a number of coupons to distribute to a selected subset of the customers. Each coupon is assumed to increase a customer's intrinsic purchase rate by a specific percentage point. At the same time, it will increase a customer's base utility of one specific category of tones by a certain amount.[12]

We conducted the policy simulation based on the estimates of the individual-level parameters when we discussed in sections 4.1 and 4.2. We evaluate the effect of distributing 100 coupons to selected customers in the whole group to optimize the objective. The choice of 100 is to ensure decent coverage of the 1200 consumers in the sample emanating from our random sample of 300 four-cliques, while at the same time keep it selective enough so efficiency matters. Each coupon is assumed to improve the intrinsic purchase rate by $c_p\%$, and increase the base utility of the first category of ring-back tones by c_c (we treat the first category as the target category of promotion). To utilize the influence effect, we assume that each person is influenced by a friend in the beginning of the promotion and that the firm observes this interaction. We measure the change in purchase probability and choice probability in the week that follows the promotion – we choose a week as the period for performance measurement as the influence effect will largely diminish after a week, according to our estimated result.

[11] 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.

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

Table 12: Policy Simulation 1 – Target Existing Customers

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

Figure 1: Policy Simulation 1 – Purchase Probability Enhancement

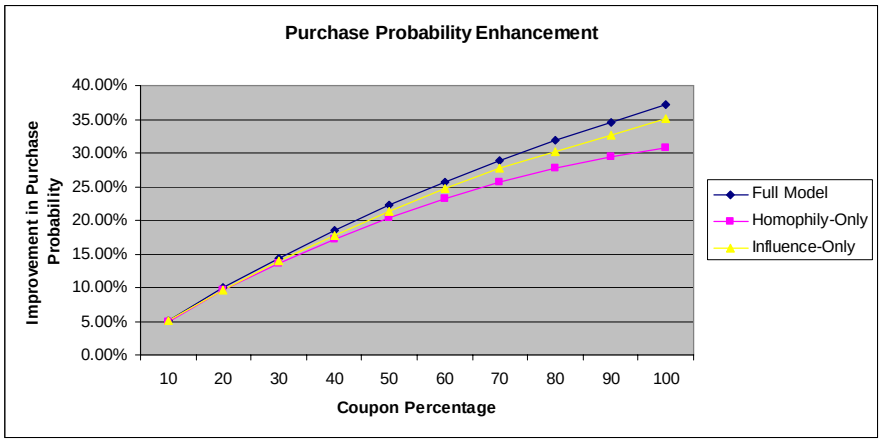
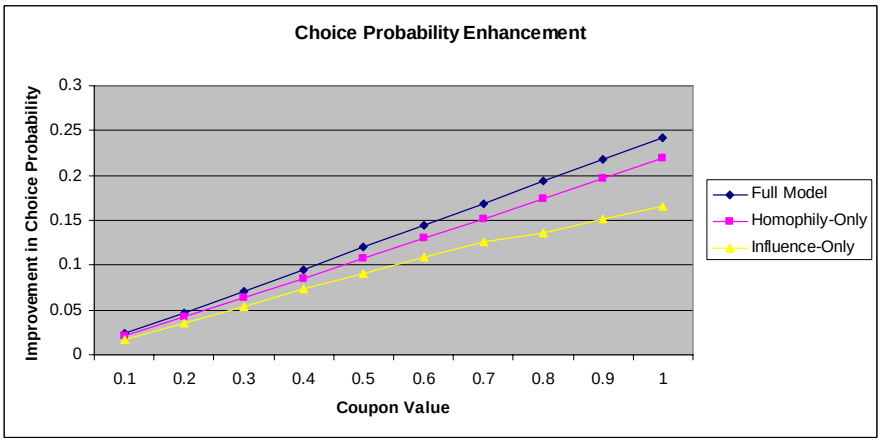


Figure 2: Policy Simulation 1 – Choice Probability Enhancement



We evaluate the purchase and choice probability of each person either with or without the coupon, and distribute the coupons to the consumers whose purchase and choice probability

increase the most. In this simulation, we consider the change in probability only for the individual customers, while the effect on other group members is considered in the next simulation. A series of values for c_p and c_c are evaluated. The resulting average enhancement in purchase probability and choice probability is plotted in Figure 1 and Figure 2. Using $c_p\% = 50\%$ and $c_c = 0.5$ as an example, the result of which is reported in Table 12. Targeting based on parameter estimates from the full model increases purchase probability by 22.18%, as compared with 20.45% of the homophily only model and 21.36% of the influence only model. This represents an 8.5% and 3.8% improvement, respectively, arising from recognizing both homophily and influence effects as compared with recognizing only one of them. The improvement in the probability of choosing the first category is 11.99% for the full model, and 10.73% and 9.12% for the homophily-only and influence-only model, respectively. This represents an 11.7% and 31.5% improvement using the full model over the homophily-only and influence-only model, respectively. On average, the full model performs 10.47% better than the homophily-only model and 4.08% better than the influence-only model on enhancing purchase probability, and it performs 11.65% and 35.37% better than the two models on enhancing product choice probability. These results show that consumer targeting can be improved by recognizing both homophily and influence effects.

5.2 The Multiplier Effect

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

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

Table 13: Policy Simulation 2 – Multiplier Effect

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

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

5.3 Targeting New Customers

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

Table 14: Policy Simulation 3 – Targeting New Customers

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

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

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

6. Conclusion

Two effects that social network researchers have long recognized are of particular interest on this topic. The first is homophily, the concept that similar people are more likely to form ties. Given

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

Enabled by a unique electronic social network dataset, we investigate the role both these factors play in consumers' decision process. Our study contributes to the literature by simultaneously quantifying both the homophily and the social influence effect in product purchase decisions by consumers in a social network context, using a unique, large-scale, real word dataset. Our study is among the first to quantify the impact of homophily and influence factors jointly in the decision process of consumers.

In our study, we estimate a purchase-timing and product choice model within the context of a hierarchical Bayesian model. Our study is made possible by our unusual dataset which

contains both communication and product purchase information over time. Applying our model to this dataset, we find strong social influence effect in both purchase-timing and product-choice decisions. We further distinguish in-group influence from out-group influence, and find that the former is more salient in the purchase-timing decision, while the latter is more salient in the product-choice decision. This is an important finding, which suggests that although customers are subject to influence by their friends, but this may be moderated by a desire not to be perceived as imitating. Furthermore, we find a strong homophily effect in the product-choice decision, where customers who are close by tend to have similar product tastes as well as similar susceptibility to influence. There is evidence of homophily in the purchase-timing decision as well, although this is not conclusive. We show that models which ignore one of the factors result in the overestimation of the other factor. Furthermore, we demonstrate the importance of communication data by showing that influence effect can be either over- or underestimated when communication is not explicitly accounted for.

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

Two limitations of our study call for further investigation in future work. First, our study identifies friends by using n-cliques, which constrain the connection structure to groups of equal number of people who are tightly connected. Network structures are in fact more versatile. For example, certain people may have many friends but the ties may be weak, while some others

may have only a few friends with very strong connections. Investigating how the number and strength of social ties impact consumer's decision, and how to best identify consumer preference in such flexible network structures will further our understanding of the implications of social networks. Second, our study shows that in-group influence is lower than out-group influence on product choice probability. Our conjecture is that the in-group influence on choosing a specific brand can have both a positive effect, where a consumer trusts his friend's selection, and a negative one, when a consumer tries to avoid imitating his friends. A more sophisticated model is needed to further isolate these two effects. Overall, we believe that understanding consumers within their social networks can lead to better marketing decisions.

Chapter 4

A Dynamic Competitive Analysis of Content Production and Link Formation of Internet Content Developers

1. Introduction

Content is the lifeblood of Internet marketing. The emergence of hundreds of revenue sharing content websites has greatly contributed to the recent proliferation of social media. A wide range of content vital for online business and consumer activities is provided at these websites: product reviews at Epinions.com facilitate online retailing; video clips at Youtube.com generate advertising revenue; articles at Fool.com attract subscribers, etc. Millions of viewers visit these websites on a monthly basis, making them a major component of Internet business (Table 1). Such websites typically generate revenue through advertising or sales referral. Consequently, their success depends crucially on the amount of viewership traffic they can attract.[13]

A key characteristic of such revenue sharing content sites is the *democratization of content*: instead of hiring employees to create content, companies operate these websites as platforms where external, independent developers come to supply content. Since the success of the websites depends crucially on the viewership their content attracts, the websites must encourage the independent content developers, or producers, to produce actively.[14] To encourage content production, website companies typically share revenue with each producer based on the viewership her content attracts. Interestingly, this creates an intra-website competition among the

[13] Display advertising fee can be charged on a pay-per-impression basis, with rates quoted in "cost per milli", which is the fee for every thousand times the advertisement is viewed, or on a pay-per-click basis, where a fee is charged every time an advertising link is clicked. Sales referral commission is often charged on a pay-per-action basis, where a content site is paid based on the sales it helps e-commerce sites generate by directing viewers to those sites. The amount of viewership traffic is the key to all these revenue models.

[14] Both "developer" and "producer" are widely accepted terms in the industry, and they are used interchangeably in this study.

independent producers, as each seeks to maximize *the viewership of her own content*, and when viewers come to the website and choose among different producers' content, producers effectively compete against each other for viewership. To attract viewership, producers naturally need to actively produce content – the more content a producer provides, the more likely a viewer will find what she needs from that producer, and the higher her viewership.[15]

Table 1: Revenue Sharing Websites with Independent Content Producers[1]

Website	*Content Type*	*Linking (Name)*	**Monthly Visitors (in millions)[3]**
about.com	Advice	No	43.9
answers.yahoo.com	Questions & Answers	Yes (Fan)	43.7
associatedcontent.com[2]	General content	Yes (Favorite)	10.7
ehow.com	How-to tip	Yes (Subscription)	30.5
epinions.com	Product review	Yes (Trust)	4.1
hubpages.com	General content	Yes (Follow)	9.9
iReport.com	News report	Yes (Follow)	1.1
seekingalpha.com	Investment advice	Yes (Follow)	1.2
squidoo.com	General content	Yes (Fan)	6.6
youtube.com	Video	Yes (Subscription)	104.1

1. A list of more than 100 revenue sharing content sites can be found at http://socialmediatrader.com/resource-list-100-revenue-sharing-sites/
2. Acquired by Yahoo! in May 2010 for about $100 million.
3. Source: Compete.com, June 2010

Making the competition more intriguing is another feature that is being increasingly introduced to such websites: *inter-producer linking*. As Table 1 shows, most such sites now allow producers to create links pointing to other producers at the site. Links may vary by name, such as trust, favorite, follow, etc, but all serve as a form of endorsement of the target by the source, and make the target's content easily accessible from the source's. Such links together

[15] Other factors also matter, such as the quality and diversity of content, and will be accounted for in this study.

form a *producer network* that evolves over time. Since Internet viewers often navigate through links to view content, and search engines also rely on the link structure to rank search results, where a producer is positioned in this network significantly influences the viewership of her content. In general, the more incoming links a producer has, and the better positions the sources of the links have, the better is her position in the network (Brin and Page 1998). This is because incoming links drive viewership traffic to a producer's content, and a producer with more and better incoming links also gets preferential placement when search engine displays search results.

The introduction of inter-producer linking leads to several intriguing questions. Marketing research on Internet content and linking is still at the early stage. Existing research has shown, in a static and analytical setting, that linking can promote the position of the target, and meanwhile enhance the content of the source – a viewer may visit a producer even if she does not have the desired content, if she can point to another producer who does (Mayzlin and Yoganarasimhan 2008, Katona and Sarvary 2008). [16] However, questions related to link formation in a dynamic context and the interaction of linking and content production decisions largely remain open. For example, how do producers form links over time, how do producers adjust their production decisions under the presence of linking, and how does one respond to others' decisions? More importantly, from the perspective of the website, would allowing producers to link encourage or discourage content production, and would it increase or decrease the overall viewership at the website? The objective of the websites introducing the linking feature is certainly to encourage production and increase traffic. But to find out whether this objective is met, we need a detailed understanding of how content producers interact with one another as they compete for viewership. Considering this, we address the following questions in our study: (1) What drives a

[16] This refers to the extension on reference links in Katona and Sarvary (2008). The main model of that paper focuses on advertising links which are price mediated, which does not apply to the situations in our study, as the links among content producers at these sites are not bought and sold but established by the sources on volition.

producer's linking decisions over time, and when and to whom would she link to? (2) Does the ability to form links encourage or discourage a producer to produce content, and how does this impact vary across producers? (3) What market structure will emerge from this competition through content production and link formation under a given policy at a website? (4) Finally, what is the overall effect of linking on the viewership at the website level, and should the website company regulate linking? Since these websites rely on the producers producing content to attract viewers, yet they can only incentivize but cannot control those producers, answers to the above questions are crucial to help the website companies understand content producers' decision process, draw implications from it, and improve their platform design.

In this study, we model the competition among content producers at a website as a dynamic game. In our model, each producer chooses her actions (produce content and link to other producers) over time to maximize her payoff – discounted viewership net of costs incurred in producing content and forming links. Producers adopt Markov strategies, and such strategies together constitute a Markov-perfect equilibrium, or MPE (Maskin and Tirole 1988, Ericson and Pakes 1995). The equilibrium characterizes the dynamic interactions among content producers and the tradeoffs they face. In making her decisions, a content producer balances the cost and benefit of her actions, both immediate and in future, and accounts for the strategic reactions from other producers, as one's actions can change the competitive positions of others. We estimate the model using the two-step estimator developed by Bajari, Benkard, and Levin (2007). Applying the model and estimation approach to a dataset obtained from a popular Internet product review website, we estimate the viewership demand and cost functions, and analyze the driving forces of producers' decisions and their implications.

Our study leads to several findings. We first demonstrate that link formation is a dynamic strategic decision. We show that the nature of the competition encourages reciprocity – linking to someone who already links back – due to a *promote-the-promoter* effect. In the dynamic context, this tendency towards reciprocity further encourages certain producers to *strategically initiate non-reciprocal links* in anticipation of the reciprocation from targets, which increases viewership in future through improved position brought about by incoming links. We find that a producer with higher content volume is more likely to strategically initiate such links to "invite" reciprocation. Next, we find the dynamic effect of linking can *either encourage or discourage content production*, depending on the situations of the producers: to obtain and in anticipation of future rewards through incoming links, a producer will produce the most content when she has high content volume but low network position. Meanwhile, the prospect of linking discourages a producer with low content volume but high network position from producing content, as she expects her relative network position to diminish over time.

Furthermore, our analysis suggests that the current linking design overall could impede competition. We find that although both more content and higher network position lead to higher viewership, only the latter leads to higher net benefit once cost is accounted for. Thus potential advantage from having more content is mostly competed away, yet significant competitive advantage is accrued to better network position. That a subgroup of producers enjoys sustainable advantage over others may soften the competition, and lead to inefficiency from the website's perspective. This is confirmed in our simulation, which suggests that alleviating the imbalance through reducing links could lead to higher overall viewership at a website.

We contribute to the literature by jointly modeling content production and link formation decisions, investigating their inter-dependence in a dynamic setting, and evaluating the impact of

linking when both decisions are determined endogenously. Existing studies have analyzed the impact of commerce network on firm profits (Stephen and Toubia 2010) without explicitly modeling the formation process of such network, and modeled the formation of content networks on the web in a static setting where content is exogenously given (Katona and Sarvary 2008). Our study extends the literature by analyzing how linking and content production decisions interact with each other, and we evaluate the impact of linking on website viewership when its effect on content production is accounted for. Furthermore, by studying the decision process and competition in a dynamic context, we show how inter-temporal tradeoffs and the strategic interactions among producers drive decisions over time, which cannot be shown in a static framework, such as the strategic invitation of reciprocal links and the content production in anticipation of incoming links from other producers. We also contribute to the literature by providing a rational economic framework for empirically analyzing the formation of links in a dynamic strategic setting. Our empirical findings provide much needed recommendations to industry managers.

The rest of the paper is organized as follows. In section 2 we review relevant literature. We then develop the dynamic game model in section 3. Following that, we discuss in section 4 the approach used for estimating this model. Section 5 discusses the empirical application, where we explain the data used in our study and discuss the result. Finally, we conclude in section 6.

2. Literature Review

Our work is related to the broad literature on Internet content and on economic networks. Marketing researchers have shown great interest in Internet content, specifically on product reviews and online word-of-mouth (WOM). Chevalier and Mayzlin (2006) investigate the effect of online book reviews on sales, and find that improvement in reviews leads to higher relative

sales. Godes and Mayzlin (2004) find that the dispersion of conversation in online communities has explanatory power on TV ratings. Chintagunta et al. (2010) find the valence of online reviews influence the box-office sales of movies. While the effect of online product reviews has been studied frequently, relatively less attention has been paid to the supply of such reviews, especially when they are supplied as information goods with profit incentive. Supply-side structural models have generally only recently gained attention in marketing (Srinivasan 2006), and our work fills in this gap in the case of Internet content.

Our work is also related to the formation of economic networks and their impacts. A rich literature exists on the formation of social and economic networks. For example, Bala and Goyal (2000) develop a non-cooperative game model to study linking decisions. Jackson (2004) gives an extensive survey on network formation literature with emphasis on stability and efficiency. Most studies use certain general value functions arising from network; while given the wide variety of networks, it is reasonable to expect that the benefit of the network, and its formation in turn, is situation specific. Two studies in marketing focus on the creation of links online. Mayzlin and Yoganarasimhan (2008) investigates why an author of an Internet blog may link to another competing blog, even though doing so effectively promotes her rival. They show that the ability to link to information is valuable to readers in addition to the ability to produce the information – if the blog does not have the information, readers will still appreciate a link to another blog that does. The borrowed content component in our study models this effect. Katona and Sarvary (2008) study the formation of links among content sites as a non-cooperative game, where links are created either for paid advertising or for reference effect in the extended model. In both studies, the content at the websites is treated as exogenous. In contrast, Stephen and Toubia (2010) study the effect of online commercial networks. They find that allowing online retailers to

link to one another creates economic value, and such value comes from improved accessibility. The study focuses on the effect of the network and does not explicitly address its formation process. Our study contributes to the literature by jointly studying both network formation and content production decisions and highlighting their interaction effect in a dynamic setting.

Our work draws from the rich literature on empirical industrial organizations from the methodology perspective. Specifically, we adopt the concept of Markov perfect equilibrium, or MPE (Maskin and Tirole 1988, Ericson and Pakes 1995, Maskin and Tirole 2001), for modeling dynamic oligopolistic competitions. Early estimation methods for MPE (Pakes and McGuire 1994, Pakes and McGuire 2001) extend the nested fixed point approach (Rust 1987) to explicitly compute equilibrium strategies. But the high dimensionality of typical dynamic competition models restricts the use of such methods to games with only few players. Recent advancement leads to several two-step estimators (Aguirregabiria and Mira 2007, Bajari, Benkard and Levin 2007, Pakes, Ostrovsky and Berry 2007) which extend the conditional choice probabilities approach (Hotz and Miller 1993). Such two step estimators bypass explicit computation of equilibrium by calculating continuation values through forward simulation, and by doing so enable the estimation of dynamic games with many players. Ackerberg et al. (2007) provides a comprehensive survey of these estimation methodologies. We implement the estimator developed in Bajari, Benkard and Levin (2007), hereafter BBL. The BBL estimator has been used for studies in industrial organizations (e.g. Ryan 2009) as well as in marketing (Yao and Mela 2011).

3. Model

We discuss the model in this section. To prepare for the model, we begin with a brief summary of the key elements of the *industry setup*. We consider a content *website* on the Internet. *Viewers*

come to the website to view content, which is produced by external, independent *content producers*, whom the website attracts through revenue sharing.

Each content producer seeks to maximize the viewership *of her own content* over time. In addition to *producing content*, a producer can *create links* pointing to other producers. Since viewers can easily follow a link to navigate to the target producer's content from the source producer's, a link benefits the *target* producer by putting her in a good position to receive viewership traffic. Furthermore, when viewers search for a specific topic and the content from multiple producers matches that search criteria, the search engine ranks the search results based on the linking structure, where producers with more incoming links and links from other producers with good positions receive preferential placement. Links thus again help the *targets* through this *positional benefit*. For the *source* of a link, the benefit is to *enhance content*, as a producer who links to other producers makes it convenient for viewers to find the content they want, and will be favored by viewers.

This industry setup leads to a *competition* among content producers, since each producer cares about her own viewership only, and viewers choose the content from multiple producers.[17] To attract viewership effectively, each producer must make her production and linking decisions while taking into account her own situation, other producers' situations, and the strategic response to her actions by other producers. She also needs to balance current and future benefits. Such considerations lead to interesting dynamic interactions. For example, more content attracts higher viewership, but producing content also incurs a cost. Depending on a producer's position, this cost-benefit tradeoff may or may not justify production. However, having more content may

[17] For example, a viewer may search for a topic, and read only the top two articles on the list retrieved by the search engine. In this case, each producer wants her content placed in the top two positions, and is competing against other producers for that.

also attract links from other producers, which improves her position later on. This additional benefit could make content production worthwhile, even if it does not attract much immediate viewership. Such dynamic interactions among maximizing agents call for a dynamic game model, which we use in this study.

In our model, there are J independent content producers competing for viewership. Time is discrete and is indexed by t, $t = 1,2,\ldots$. In each time period, each producer decides whether to produce content and whether to link to other producers. In the following subsections, we first describe the viewership demand market that clears in each time period given producers' content states and the link structure. We then discuss producers' dynamic content production and link formation decisions, and how content and link structure evolve according to such decisions. Finally, we explain the dynamic competition and the equilibrium concept, and discuss the tradeoffs faced by producers which shape their strategies.

3.1 Viewership Demand

There are M consumers, or viewers, in each period.[18] Each viewer chooses to view the content of one content producer among the J producers at the website, or chooses to go to an external website, i.e. the *outside option*. This viewership constitutes the demand for producers' content. We adopt a logit demand model, which has been widely used in modeling oligopolistic competitions (e.g. Berry 1994, Berry et al. 1995, Dube et al. 2009), to characterize viewership demand in this per-period market.[19] The discrete-choice framework of the logit demand model

[18] The terms "viewer" and "reader" are used interchangeably in this study.

[19] The logit demand model is based on a discrete-choice framework, yet it is possible that a reader may read multiple articles of a producer in a period, e.g., reading the product reviews of different products, or the content of several producers. An in-depth modeling of such behavior requires detailed clickstream data of readers which we unfortunately do not have. Instead, we treat each pageview as one single viewer in our model (that is, if a viewer reads three product review articles in the period, it is counted as three viewers in the model). This reduced-form

reflects the competitive nature of the viewership demand, i.e. viewership of one producer's content may come the cost of another's. A viewer i's latent utility from reading the content of producer j in period t is:

$$(1) \qquad u_{i,j,t} = \begin{cases} \bar{u}_{i,j,t} + \varepsilon_{i,j,t} = f(C_{j,t}, P_{j,t}, C^b_{j,t}; \beta_i) + g(Q_j, Q^b_{j,t}; \gamma) + \varepsilon_{i,j,t} & j = 1..J \\ 0 + \varepsilon_{i,0,t} & j = 0 \end{cases}$$

In equation (1), $\bar{u}_{i,j,t} = f(C_{j,t}, P_{j,t}, C^b_{j,t}; \beta_i) + g(Q_j, Q^b_{j,t}; \gamma)$ is the deterministic component of the utility. $C_{j,t}$ is the content *quantity* of producer j at time t, $P_{j,t}$ is a numeric measure of her *network position*, and Q_j is a vector of quality variables of the producer that remains constant over time.[20] Furthermore, $C^b_{j,t}$ measures the total quantity of *borrowed content*, i.e. content derived from linking to other producers. Similarly, $Q^b_{j,t}$ measures the average quality of the producers being linked to. These measures are explained in detail later when we discuss producer actions and the network structure. The function $f(.;\beta_i)$ specifies how content, network position, and borrowed content enter into the utility function, with β_i as the parameter. Since viewer navigation behavior is not explicitly modeled, we estimate multiple specifications of functional forms for $f(.;\beta_i)$, with the best specification chosen using certain model selection criteria. The function $g(.;\gamma)$ captures the quality differentiation among producers. Quality is used mainly for control purpose in our study, so we adopt a linear specification with γ as the parameter: $g(Q_j, Q^b_j; \gamma) = (Q_j, Q^b_j)^t \gamma$.

---

treatment of readership demand can be improved by explicitly modeling a viewer's navigation behavior, which we leave for future research as richer data become available.

[20] In our model, we treat quality as a characteristic of the producers instead of content. This assumes away potential variation of quality across different content produced by the same producer. This is a reasonable assumption in the context of our study, since the quality of individual content is not observed before a viewer decides to view the content.

The relative attractiveness of a producer is determined by the amount of content she has, i.e. the content quantity, the location of the producer in the network, i.e. the network position, and the quality of the producer. Furthermore, the attractiveness of a producer is also influenced by the content of the other producers she links to. Intuitively, the more content a producer has, the more viewership she would receive, as viewers are more likely to find the content they want. Similarly, the more prominent a producer's position in the network, the higher viewership demand she would receive, as her content will receive more preferential placement by the search engine, and more viewers may be directed to her content when they navigate through the links. Borrowed content should further enhance a producer's attractiveness due to the convenience benefit it affords the viewers. We expect these to be reflected from the parameter vector β_i in accordance with the specific functional form. For example, we expect all coefficients to be positive if factors enter the utility function linearly.

Finally $\varepsilon_{i,j,t}$ is an i.i.d random component which follows the type I Extreme Value distribution, resulting in the familiar logit probability of viewer i choosing producer j at time t:

$$\Pr_{i,j,t} = \frac{\exp\{\bar{u}_{i,j,t}\}}{1+\sum_{j'=1}^{J}\exp\{\bar{u}_{i,j',t}\}} \tag{2}$$

Note that this viewership model is a reduced form one, and assumes away any explicit state-dependence on viewer's side. In reality, a viewer's behavior in one period may be influenced by her past behaviors, e.g. she becomes a routine follower of a content producer. In our model, this dependence can come indirectly through the persistence of a producer's state: a product review of an obsolete product produced earlier may be of no value now, but it attracted

viewers at that time, some of whom then continues to visit the producer's page, and this is reflected in the utility function where a cumulative measure of content is used.[21]

Viewers may have different navigation patterns and content requirements, which results in different relative emphasis placed on different components in the utility function.[22] This heterogeneity is captured using a *latent class* approach (Kamakura and Russell 1989). That is, we assume there are N segments of viewers, each characterized by its own set of coefficients, $\{\beta_n\}_{n=1..N}$, and portion of each type is denoted as λ_n, so that $\sum_{n=1}^{N} \lambda_n = 1$.

3.2 Content Producer

In any time period, a content producer j is characterized by a collection of variables: $\{C_{j,t}, P_{j,t}, C^b_{j,t}, Q_j, Q^b_{j,t}\}$. Content, network position, and borrowed content all evolve over time according to the actions of both producer j and other producers. A producer can take two types of actions, *content production* and *link formation*. We discuss these actions below and how the variables evolve according to these actions.

3.2.1 Content Production

A producer's content quantity, $C_{j,t}$, is determined solely by her own production decisions over time. In each period, a producer decides whether to produce additional content to add to her webpage – write another product review, break another news story, create another analytical

[21] Since the emphasis of our study is on producer's production and linking behavior, structurally modeling viewer's persistence over time adds great complexity to the model but might not provide much added value. It also requires detailed viewer navigation data. We leave the joint structural modeling of producer and consumer behavior for future research.

[22] In the case of a sequence of page views, certain page views may be related more to the page content (e.g. following a topic search) while others may be related more to network positions (e.g., navigating through links or using a search engine that accounts for network positions). The heterogeneity also captures this effect, since a viewer in the model actually corresponds to a viewer-page view pair in the real world, as discussed earlier.

report, etc – and if yes, the amount of content to produce. We denote this action by producer j at time t as $a_{j,t}^{p}$, where the superscript p indicate it is the production decision. Specifically,

(3) $$a_{j,t}^{p} = \begin{cases} 0 & \text{do not produce content} \\ k & \text{produce } k \text{ units of content, } k \in \{1,2...\} \end{cases}$$

In the equation, k represents the number of units of content produced. Each unit of content may correspond to an article in the real world, thus the action is discrete.

Producing content increases the content quantity at a producer's webpage, $C_{j,t}$. Meanwhile, there is an opposite, *depreciation*, force at work: a product review will become less needed as the reviewed product becomes obsolete; a news story will become non-news after a few days; an analytical report will become less relevant as the situation expires, etc. Similar to existing literature modeling capacity depreciation (e.g. Besanko and Doraszelski 2004), we assume that the producer's content at a website depreciates with a certain ratio over time. Combining the effects of production and depreciation, the content quantity at a producer's webpage evolves as:

(4) $$C_{j,t} = \delta C_{j,t-1} + a_{j,t}^{p}$$

In equation (4), $\delta \in (0,1)$ is the depreciation rate of the content. The smaller the value of δ is, the faster is the depreciation.

Producing content is a costly activity. We denote the cost of producing k units of content by producer j as $c^{prod}(k, X_j; \phi)$, with $c^{prod}(0, X_j; \phi) = 0$, i.e. the producer incurs no cost if she does not produce content. X_j is a vector of characteristics of producer j that may affect cost, and ϕ is a vector of parameters for the production cost function. The production cost is expected

to be an increasing function of k, the units of content produced. The exact functional form of $c^{prod}(.)$ used in this study is specified in section 5 where we discuss the empirical application.

3.2.2 Link Formation

In each time period, a producer may also create a link pointing to another producer, assuming one to that producer does not already exist.[23] We denote this action by producer j at time t as $a^{l}_{j,t}$, where the superscript l indicate it is the linking decision. Specifically:[24]

$$a^{l}_{j,t} = \begin{cases} 0 & \text{do not create link} \\ j' & \text{create a link to producer } j',\ j' \in \{1..J\},\ j' \neq j \end{cases} \tag{5}$$

Link formation may also be a costly activity. To form a link, a producer needs to spend time specifying so at the website. We denote the cost of creating a link by producer j as $c^{link}(j', X_j; \psi)$. The cost may vary according to the target of the link. For example, if reciprocity has intrinsic value, the producer will incur higher cost creating a non-reciprocal link, i.e. links to a producer j' when j' already links back at her, than creating a non-reciprocal one. Similar to production cost, ψ is the vector of parameters for the linking cost function. The exact functional form of $c^{link}(.)$ used in this study is specified in section 5.

3.2.3 Producer Network and Network Position

The links created by all producers together form a *producer network*, which is formally represented as a directed graph. Each node in the graph corresponds to a producer, and an edge

[23] Links are at producer level instead of content level, e.g. from producer *A* to *B* instead of a specific article of producer *A* to that of producer *B*.

[24] In our model, we consider the case where only creation but not removal of links is allowed. This is consistent with the dataset used in the empirical application. In real-world settings, certain websites allow link removal, while others do not. It is straightforward to extend our model to allow link removal. Also, we assume that a producer can create only one link in a period. This assumption is also made based on the dataset used in this study, and it is also straightforward to change it to allow a producer to create multiple links in a period.

exists if the producer corresponding to the source node has a link pointing to the producer corresponding to the destination node. The network evolves as producers create links over time. The network at time period t is denoted as G_t.

From the topology of the network, a numerical measure of each producer's network position, $P_{j,t}$, can be derived. As discussed earlier, the position of a producer in the network greatly influences the amount of viewership traffic directed to her content – the more incoming links a producer's has, and from the more prominent positions those incoming links come, the more traffic will be directed to the producer. Thus, both the number of incoming links and the positions of the sources matter. The PageRank measure (Brin and Page 1998), initially adopted by Google, elegantly captures both effects. Statistically, PageRank represents the probability of reaching each web page in a network when viewers follow a random walk along the links.

PageRank is equivalent to the eigenvector centrality of a damped adjacency-graph of the network. Interestingly, a rich literature in sociology has well established the importance of eigenvector centrality in social networks (e.g. Bonacich 1987, Faust and Wasserman 1992, Wasserman and Faust 1994, Bonacich and Lloyd 2001), where higher centrality it is associated with higher prestige. Recent marketing literature (Katona & Sarvary 2008) has also adopted PageRank in characterizing the network position of players. Following these, we use the PageRank of each producer in the network as the measure of her network position:

$$P_{j,t} = PageRank_{j,t} \tag{6}$$

The computation of PageRank is explained in the Appendix. The higher the PageRank, the more prominent a producer's position is in the network. This is the network position measure that enters into the demand function as specified in equation (1).

That incoming links increase a producer's position also means a producer's own position will reduce when she creates a link pointing to another producer – an outgoing link increases the target's position, and since position is relative, it would also reduce that of the source. This constitutes a *strategic* cost of link formation, which must be balanced with the benefit of borrowed content.

3.2.4 Borrowed Content

When a producer j has a link to another producer j', the content of producer j' can be easily accessed when a reader is viewing producer j's content. This augments the source's content, making the producer's webpage more appealing (Katona and Sarvary 2008). This effect is captured in our model using *borrowed content*, $C^b_{j,t}$, which is simply the sum of the content of all other producers being linked to at the time:

$$C^b_{j,t} = \sum_{j'=1}^{J} C_{j',t} I\{j' \neq j, j \rightarrow j'\} \tag{7}$$

In the equation, $I\{.\}$ is the indicator function which equals 1 if the link exists and 0 otherwise.

Similarly, the borrowed quality $Q^b_{j,t}$ is the average of quality measures of the producers being linked to:

$$Q^b_{j,t} = \sum_{j'=1}^{J} Q_{j'} I\{j' \neq j, j \rightarrow j'\} / \sum_{j'=1}^{J} I\{j' \neq j, j \rightarrow j'\} \tag{8}$$

3.3 Dynamic Competition

The competition among content producers over time is naturally modeled as a dynamic game. The key characteristic of the competition is that actions taken by producers not only determine

the current payoff, but also affect future strategic interactions. Consequently, when a producer makes content production and link formation decisions, she needs to account for not only the current benefit, but also the future benefit according to the strategic response to her actions by other producers.

In each time period, the state of the competition is fully described by a set of commonly observed state variables. Producers take actions to maximize their respective discounted payoffs. Such actions are taken based on the current state of competition and in anticipation of the strategic response. The solution concept for producer's optimizing behavior is that of Markov-perfect equilibrium, or MPE (Ericson and Pakes 1995). In an MPE, the strategy played by each producer is a Markov strategy, where actions are fully determined by the current state, and the strategy of each producer constitutes the best response to other producers' strategies.

3.3.1 State

The state at time period t, denoted as s_t, is the collection of the individual content states of all producers and the state of the producer network: $s_t = (s_{1,t},...,s_{J,t},G_t)$, where $s_{j,t} = \{C_{j,t}, Q_j, X_j\}$ characterizes the quantity of producer j's content in period t and the characteristics of the producer related to quality and cost, and G_t contains the topology of the producer network. Note that $s_{j,t}$ does not include $P_{j,t}$, as the position of each producer in the network is fully determined by the topology of the network, which is encoded in G_t; nor does it include $C^b_{j,t}$ or $Q^b_{j,t}$, as the borrowed content is determined jointly by the topology of the network and the content of all producers. In another word, $P_{j,t}$, $C^b_{j,t}$ and $Q^b_{j,t}$ are *derived* from the state instead of the primitives of the state.

3.3.2 Action

In each time period, producer j 's action $a_{j,t} = (a^p_{j,t}, a^l_{j,t})$ is its content production and link formation decision. Let a_t denote the vector of actions taken by all producers at time t, i.e. $a_t = (a_{1,t},...,a_{J,t})$.

Consistent with extant literature (e.g. Rust 1987, BBL 2007), we assume that before choosing her action at time t, each producer j receives an action-specific private shock $\nu_{j,t}(a_{j,t})$ that is independent among producers and over time. Since in our setting the actions are discrete, this private shock is a vector where each element corresponds to a specific action that can be taken at the time. Also consistent with extant literature, we assume the private shock follows an extreme value distribution. This private shock is needed in dynamic game models to account for the variability in actions that goes beyond the observed states. The collection of action-specific private shocks across all producers is denoted as $\nu_t = (\nu_{1,t},...,\nu_{J,t})$.

3.3.3 Payoff

In each time period, according to the viewership market demand and producer actions, producer j 's current-period payoff is:

$$(9) \qquad \pi_j(a_t, s_t, \nu_{j,t}) = mr\sum_{n=1}^{N} M\lambda_n \Pr_{n,j}(s_t) - c^{prod}(a^p_{j,t}, X_j; \phi) - c^{link}(a^l_{j,t}, X_j; \psi) + \nu_{j,t}(a_{j,t})$$

In equation (9), mr is the marginal benefit associated with each viewer visit, and $M\lambda_n$ is the number of viewers in segment n. In each period, the payoff of producer j is the benefit of viewership demand net of any cost associated with the action taken by the producer.

Each producer is concerned not just with the payoff of the current period, but also the overall payoff over time. The total discounted payoff to producer j at time t, which the producer seeks to maximize, is:

(10) $$E[\sum_{\tau=t}^{\infty}\beta^{\tau-t}\pi_j(a_\tau, s_\tau, \nu_{j,\tau}) \mid s_t]$$

In equation (10), $\beta \in [0,1)$ is the discount factor. The expectation is over the private shock, producers' actions in the current period, as well as future states, actions, and private shocks. As is shown clearly in the equation, the payoff to a producer depends on not only her own actions, but also the actions of other producers. This leads to strategic interactions which are characterized using an MPE.

3.3.4 Strategy and Equilibrium

We assume all producers follow Markov strategies. A Markov strategy profile σ of the dynamic game is the collection of the strategies of all producers: $\sigma = (\sigma_1, \sigma_2, ..., \sigma_J)$ where σ_j is the strategy played by producer j which depends on the state and the private shock, $\sigma_j : S \times \nu_j \mapsto A_j$, where S is the set of all states, ν_j is the set of private shocks and A_j is the set of all actions producer j can take.

Given a strategy profile, a producer's value function is the expected discounted payoff given the state, integrated over private shocks. It can be written recursively as follows:

(11) $$V_j(s;\sigma) = E_\nu[\pi_j(\sigma(s,\nu), s, \nu_j) + \beta \int V_j(s';\sigma) dP(s' \mid \sigma(s,\nu), s) \mid s]$$

When choosing a strategy, a producer needs to take into account not only the current state, but also other producers' strategies. Following convention in literature, we use σ_{-j} to denote the strategies played by all producers other than producer j. A producer's optimization problem is:

$$(12)\quad V_j(s;\sigma_{-j}) = \max_{\sigma_j(s,\nu)} \{E_\nu[\pi_j((\sigma_j(s,\nu),\sigma_{-j}(s,\nu)),s,\nu_j) + \beta \int V_j(s';\sigma_{-j}) dP(s'|(\sigma_j(s,\nu),\sigma_{-j}(s,\nu)),s) \mid s]\}$$

The strategy which is the solution to equation (12) for producer j is the best response of the producer to others' strategies. An MPE is a strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*, ..., \sigma_J^*)$ where each producer's strategy is the best response to other producers' strategies. That is, in an MPE, when holding the strategies of other producers unchanged, no producer can increase its own expected payoff by unilaterally deviating to another strategy:

$$(13)\quad V_j(s;\sigma_j^*,\sigma_{-j}^*) \geq V_j(s;\sigma_j,\sigma_{-j}^*), \forall s, \sigma_j$$

With observations of viewership demand and producer actions according to the states over time, we can estimate the parameters for the viewership demand model and the dynamic structural parameters, i.e. cost parameters, using the optimality condition implied by the equilibrium, which we discuss in detail in section 4.

3.4 Inter-temporal Tradeoffs

We now qualitatively discuss the tradeoffs content producers face in their production and linking decisions which are incorporated in the model. When deciding whether to produce content, producers obviously face a tradeoff between the cost incurred in producing content and the viewership such content attracts over time. Furthermore, there are several tradeoffs induced by linking, which lead to interesting interactions among producers. To begin with, when linking to another producer, a producer faces the tradeoff between borrowing the content of another

producer and lower network position arising from promoting her competitor. Depending on how much the borrowed content will help and how severely the link will reduce her own network position, the producer may or may not find it worthwhile to form a link. Interestingly, when we take this tradeoff a step further, to consider not only whether to form a link but also whom to link to, we can see this tradeoff provides a simple explanation to a well known phenomenon in networks: the tendency to form *reciprocal* links. Reciprocity can be explained by social norm in sociology literature (Gouldner 1960), and through reward and punishment schemes in repeated games (Axelrod and Hamilton 1981). In the situation that we study, however, reciprocity may arise naturally from the consideration of network position. To see this, recall that the source's network position positively influences the target's. Suppose producer *A* wants to create a link, and producer *B* already has a link to producer *A* while producer *C* does not. Then if *A* links to *B*, thereby improving *B*'s position, the enhanced position of *B* will be partially carried over to *A*. Whereas if *A* links to *C*, who is not *A*'s source, then *A* will not get this indirect benefit. Other things equal, this *promote-the-promoter* effect would favor reciprocal links over non-reciprocal ones.[25] That is, it is better to promote one's own promoter instead of another producer.

More tradeoffs come into play when we consider the interactions among producers over time. When making linking decisions, since a reciprocal links are naturally encouraged, a forward-looking producer may intentionally create a non-reciprocal link, if she expects that the producer she links to would reciprocate in the near future. That is, a producer may strategically create a link to "invite" reciprocation. The tradeoff she faces in this decision is between lower network position now and higher network position later on, if and when the target reciprocates. Furthermore, the prospect of linking may also encourage or discourage content production. A

[25] A Monte-Carlo simulation using random graphs will easily show that, on average, the reduction in network position through creating a reciprocal link is less than that through creating a non-reciprocal link.

producer may be encouraged to produce more content than she otherwise would, if she expects that by producing more content, she can attract incoming links from other producers later on. The tradeoff she faces in this decision is between the cost of producing content now and better network position later on when she receives incoming links. At the same time, if a producer expects her competitors to receive incoming links, which diminishes her relative network position over time, she may produce less than she otherwise would. All these tradeoffs play a central role in determining content production and link formation decisions, and lead to the equilibrium strategy adopted by content producers.

4. Estimation

Our estimation requires that the content production and link formation decisions of all producers over a number of time periods are observed, so is the per-period viewership of each producer's content in multiple time periods. The parameters to be estimated are the segment-specific viewership demand coefficients and the sizes of the segments, the quality coefficients, the content depreciation rate, the marginal benefit to the producer per reader visit, and the cost parameters of content production and link formation, as summarized below:

$$\Theta = (\{\beta_n, \lambda_n\}_{n=1..N}, \gamma, \delta, mr, \phi, \psi)$$

The marginal benefit and the cost parameters are not jointly identified. Considering this, we normalize $mr = 0.001$ for identification, which implies that the unit of account for cost is the marginal benefit per thousand views.[26] The first half of the parameters, $(\{\beta_n, \lambda_n\}_{n=1..N}, \gamma, \delta)$, are the parameters governing the viewership market in each period. The identification rests on the cross-sectional and inter-temporal variation of the content and network of producers, together

[26] This follows the industry standard on display advertising, where fees are quoted as cost-per-mille, or CPM, which represents the amount an advertiser needs to pay for every thousand times an advertisement is displayed to viewers.

with the corresponding variation of viewership. The second half of the parameters, (ϕ,ψ), are the dynamic structural parameters that together with the viewership demand parameters govern the dynamic competition, the identification and estimation of which rest on the optimality condition of the equilibrium.

Estimating dynamic games is challenging due to the "curse of dimensionality" – the state space has high dimensionality as it incorporates the states of all players. Early estimation methods (e.g. Pakes and Mcguire 1994) rely on explicitly solving for equilibrium through value-function iteration, and have limited scalability. Recently developed two-step estimators call for estimating as many structural parameters offline as possible, and bypassing the computation of equilibrium when estimating the dynamic structural parameters. Our estimation is implemented using one such two-step estimator as specified in BBL (2007). BBL approaches the estimation task in two stages. In the first stage, we recover the equilibrium strategy of producers in reduced form, based on observed states and actions. Estimation of the equilibrium strategy, also termed the *policy function*, should strike the right balance between flexibility and data availability. A flexible functional form is desired for accurate representation of the equilibrium strategy, but it also requires more data. The second task for the first stage is to estimate the transition of states over time according to producer actions. The viewership demand will also be estimated in the first stage. In the second stage, using the knowledge of policy function, state transition, and viewership demand estimation in the first stage, we perform forward-simulation of the observed policy versus perturbed policies. As the observed policy constitutes an equilibrium, the optimality condition dictates that a producer's payoff when she plays the equilibrium strategy is no less than her payoff under an alternative perturbed strategy, while other producers still follow

the equilibrium strategy. This optimality constraint forms the basis for constructing the objective function of a GMM estimator.

As is common in research on empirical dynamic games, we focus on symmetric pure strategy equilibrium. Such restriction allows us to pool data across all producers, which reduces data requirement and improves estimation efficiency.

4.1 First Stage

In the first-stage of the estimation, we recover the policy function, the state transition process, and the viewership market demand system.

4.1.1 Policy Function

In equilibrium, each producer chooses her action based on her own state as well as the states of other producers and the producer network. In the first stage of estimation, we recover this policy function, σ^*, which maps states to actions, in reduced form. BBL recommends using flexible functional forms to approximate the equilibrium policy with precision, which needs to be balanced with data availability.

Facing this tradeoff, we first transform the state space by deriving the vectors of network positions and borrowed content of all producers from the content state of each individual producer and the network structure – these are the variables that enter the utility functions directly. We then partition the transformed state space of an individual producer into quintiles along both the content dimension and the network position dimension. For each cell in this partitioned state space, we run a separate set of regressions with producer actions as dependent variables. The independent variables include the quality and cost related characteristics of the

producer, the borrowed content and quality of the producer, the number of other producers in each cell of the partitioned state space, and the average quality of other producers.[27]

Since linking actions differ by destination, we distinguish the target on the following four dimensions: reciprocity, content, network position, and quality. We separate a reciprocal link from a non-reciprocal one, and for each of the other dimensions, we perform a median-split on the target: separate a high content producer (whose content quantity is above median) from a low content one (below median); separate a high network position producer from a low network position one; separate a high quality producer from a low quality one.[28] There are thus sixteen different types of linking targets, which combined with an action of no-link results in seventeen possible linking actions. We estimate each regression function using generalized linear models, with log link function for content production and logit link function for link formation.

The set of regression functions through this estimation fully describes the strategy played by each producer in equilibrium. These policy functions form the basis for forward-simulation that is used in the second stage of the estimation to recover dynamic structural parameters.

4.1.2 State Transition

State transition probabilities are needed for performing forward-simulations in the second stage of estimation. In our model, the transition of states given the actions of all producers is deterministic – linking actions deterministically change the network structure, while production actions together with depreciation deterministically change content state. Consequently, state

[27] Essentially we are estimating the policy function nonparametrically on a producer's own state but parametrically on other producers' states. Ideally, the policy function should be estimated nonparametrically over the entire state space, but the high dimensionality of the state space makes this impractical, as to do so requires enormous amount of data. BBL suggests using local linear regression, which is similar to what we do here.

[28] Since quality attributes are constant over time in our model, the multi-dimensional quality measure of a producer can be reduced to a single dimensional number once the viewership demand is estimated, by weighting based on the estimated coefficients.

transition does not need to be estimated once the policy function is recovered. In the second stage forward simulation, we simply simulate producer actions based on the estimated policy function, and state transition can be calculated deterministically once the actions are simulated.[29]

4.1.3 Viewership Demand

The viewership market demand in each period can be estimated rather straightforwardly with MLE. Denote $s^m_{j,t}(s_t;\{\lambda_n,\beta_n\}_{n=1}^N,\gamma)=\sum_{n=1}^N \lambda_n \Pr_{n,j}(s_t;\beta_n,\gamma)$ as the theoretical market share of producer j at time t given the state and the parameters, and the actual market share observed from data as $\hat{s}^m_{j,t}$.[30] Assuming that the difference $\eta_{j,t}=\log\hat{s}^m_{j,t}-\log s^m_{j,t}$ follows an i.i.d. normal distribution (Holmes 2011), the parameters $\{\{\lambda_n,\beta_n\}_{n=1}^N,\gamma\}$ can be estimated using maximum likelihood.[31] The market size, i.e. the total number of viewers, M, is needed for calculating market share, and is assumed to be observed.[32] The content depreciation parameter, δ, could be estimated either jointly with the other parameters of the viewership market demand equation, or separately in an offline manner.

4.2 Second Stage

We now discuss the second stage estimation of the dynamic structural parameters, i.e. cost of producing content and forming links. The key to the second stage estimation is the optimality condition of an equilibrium: given the equilibrium strategy profile $\sigma^*=(\sigma_1^*,\sigma_2^*,...,\sigma_J^*)$, for any

[29] Content production is similar to investment in empirical IO, where studies also use probabilistic state transition models (e.g. Besanko and Doraszelski 2004). The difference is minor, as a tradeoff between the precision of states and the precision of state transition. Our model allows for deterministic state transition because the exact content state and the state of the producer network are used.
[30] The superscript *m* represents "market". This is to avoid confusion with the same symbol *s* that represents producer state.
[31] For the case of one viewer segment only, this is the same as the inversion suggested in Berry (1994).
[32] Changing the market size will change only the constant term of the estimated demand parameters.

alternative strategy $\sigma_j^{'}$ for an arbitrary producer j and a randomly chosen state s, the equilibrium condition dictates that:

(14) $$V_j(s,\sigma_j^*,\sigma_{-j}^*;\phi,\psi) \geq V_j(s,\sigma_j^{'},\sigma_{-j}^*;\phi,\psi)$$

Given a specific σ^*, a tuple $x=\{j,s,\sigma_j^{'}\}$ indexes one such equilibrium condition. Following BBL's notation, define

(15) $$g(x;\phi,\psi) = V_j(s,\sigma_j^*,\sigma_{-j}^*;\phi,\psi) - V_j(s,\sigma_j^{'},\sigma_{-j}^*;\phi,\psi)$$

And define objective function

(16) $$Q(\phi,\psi) = \int (\min\{g(x;\phi,\psi),0\})^2 dH(x)$$

where H is a distribution over the set X of the equilibrium conditions. Then the true parameter (ϕ_0,ψ_0) satisfies:

(17) $$Q(\phi_0,\psi_0) = 0 = \min_{\phi\in\Phi,\psi\in\Psi} Q(\phi,\psi)$$

The estimation is the empirical counterpart of this condition: let $\{X_k\}_{k=1}^{n_I}$ be a set of n_I randomly chosen optimality conditions. For each $X_k=\{j_k,s_k,\sigma_{j_k}^{'}\}$, we calculate the payoff of the focal producer j_k when she follows the equilibrium strategy, $\hat{V}_{j_k}(s_k,\sigma_{j_k}^*,\sigma_{-j_k}^*;\phi,\psi)$, and that when she follows the alternative strategy, $\hat{V}_{j_k}(s_k,\sigma_{j_k}^{'},\sigma_{-j_k}^*;\phi,\psi)$, for a proposed parameter value (ϕ,ψ). The empirical counterpart of the objective function is then

(18)
$$\begin{aligned} Q_n(\phi,\psi) &= \frac{1}{n_I}\sum_{k=1}^{n_I}(\min\{\hat{g}(X_k;\phi,\psi),0\})^2 \\ &= \frac{1}{n_I}\sum_{k=1}^{n_I}(\min\{(\hat{V}_{j_k}(s_k,\sigma_{j_k}^*,\sigma_{-j_k}^*;\phi,\psi) - \hat{V}_{j_k}(s_k,\sigma_{j_k}^{'},\sigma_{-j_k}^*;\theta),\phi,\psi\})^2 \end{aligned}$$

BBL shows that $Q_n(.)$ can be calculated through forward simulation, and the parameter that minimizes the objective function

(19) $$(\hat{\phi},\hat{\psi}) = \arg\min_{\phi\in\Phi,\psi\in\Psi} Q_n(\phi,\psi)$$

is a consistent estimate of the true parameter under mild regularity conditions. This recovers the mean estimate of the parameter, while the standard error can be calculated using re-sampling of these equilibrium conditions.

5. Empirical Application

5.1 Data

Our data is obtained from a popular online product review website, which in recent years consistently attracts several million visitors on a monthly basis. A product reviewer can start writing product reviews once she creates an account at the website. The products reviewed at the website range from automobiles to toys, books, and movies, etc. Such *reviews* correspond to the *content* in our model. In addition to writing product reviews, a reviewer can also link to other reviewers by putting them into her list of trusted reviewers. Creating a link is solely at the discretion of the source reviewer, without the need for consent from the target reviewer. Such links together form a so-called *web of trust* among reviewers, and this corresponds to the *producer network* in our model. Viewers can easily navigate through the trust links to go from one reviewer's reviews to the reviews of another reviewer whom she trusts. Furthermore, product reviews written by reviewers who are trusted by many other reviewers, and trusted by reviewers who are themselves trusted by other reviewers, will receive preferential placement

when viewers search the website. The position of a reviewer in this web of trust thus heavily influences the likelihood of her reviews being accessed by viewers.[33]

Although there are thousands of reviewers writing reviews at the website, in this study we focus on a small group of the most active ones, known at the website as the *top reviewers*. These top reviewers write product reviews frequently and consistently over time, and they are paid by the website based on the viewership their reviews attract. This group of elite reviewers is suitable for the model we developed earlier, as they are likely dedicated producers who are driven by profit incentive and who choose their actions strategically.[34] This small group of reviewers also is responsible for a significant share of the website viewership traffic.[35] Focusing on this group also eases the estimation of the model, as the number of players is kept at a reasonable level, and a long history of content production and link formation decisions is available for these active producers.

Our data set contains the decisions of writing reviews and creating links at the daily level, from June 2008 to March 2010. It also contains the viewership information starting from November 2009: for each four-day period starting from November 2009, the number of times each reviewer's reviews is visited is recorded.[36] There are a total of 199 top reviewers at the site. Among them, 6 left the site during the period, and we exclude them from the data set.

[33] In an interview, the former CTO of the company said of the trust system "… based on anecdotal evidence, those who have started using it end up completely depending on it to navigate the site."
[34] That is, as compared to other "occasional" users who write reviews infrequently, and who may be driven by other incentives such as a spontaneous desire to express one's opinion, for which a strategic framework may not be applicable.
[35] Comparing the viewership statistics of this group of reviewers with the website level statistics suggests they are responsible for about 30% of the overall viewership.
[36] The Website displays the cumulative view count at the reviewer level and the information is updated daily. However, the update is not well synchronized for all reviewers. Thus we aggregate the information into 4-day periods to eliminate the noise created by this technical issue.

Table 2: Summary Statistics

	Mean	SD	Min	Max
Reviews Written Per Reviewer	168.14	161.07	5	704
Links Created Per Reviewer	10.31	11.19	0	67
Total View Count Per Reviewer	34375	50834	680	372188
Number of Reviewers	193	Total Reviews Written		33123
Number of Decision Days	646	Total Links Created		2039
Number of View Count Periods	28	Total View Counts		6840629
Total Non-reciprocal Links	1148	Total Reciprocal Links		891
Percent of Reciprocated Non-reciprocal Links	40.70%	Average Days Taken To Reciprocate		52.5

The summary statistics are reported in Table 2. As is shown in the table, these top reviewers are highly active in writing reviews, averaging one review article per reviewer about every three days. In addition to writing reviews, they also created over two thousand links over the period, although the frequency of creating links is lower than writing reviews, with each reviewer adding a link roughly every two months. These top reviewers together attract a large audience, totaling more than six million view counts over a period of about four months. Comparing with website level traffic information shows that this small group is responsible for about 30% of the total visit at the website, a significant share.

Based on the information available at the website, we use three variables for the *quality* factors in our model: *diversity*, *popular*, and *advisor*. These three variables and their summary statistics are described in Table 3. Together, these factors cover three important aspects which can affect viewership demand in addition to content volume and network position: diversity, popularity, and quality.

Table 3: Reviewer Quality Factors

Factor	Value	Description
Diversity	Integer	The number of product categories for which the reviewer

		writes reviews as top reviewers
Popular	Binary Indicator	The reviewer was recognized as the top 100 most popular authors before
Advisor	Binary Indicator	The reviewer is recognized as trusted source on content quality
Average Diversity		1.53
Number of "Popular" Reviewers		59
Number of "Advisor" Reviewers		112

5.2 Result – Viewership Demand

We first estimate the model of viewership demand as determined by each reviewer's content, network position, borrowed content, and quality factors. As discussed earlier, the content depreciation parameter, δ, can be estimated together with the other viewership demand parameters. In our dataset, however, the content production and link formation data covers a much longer period than the data for viewership market demand. Furthermore, the overall viewership at the website has remained fairly stable over the period for which we observe the actions. Therefore, we estimate this depreciation parameter "offline", by treating it as a discount factor and finding the value that best keeps the content quantity stable over time. We arrive at the estimate $\delta = 0.9893$ this way. The summary statistics of network positions, and those of the effective discounted content and borrowed content, both calculated according to the depreciation parameter, are reported in table 4. The market size M is set to be twice the average total visit counts at the website to allow for substitution effect among competitors' websites. Website level statistics show that there were on average 5.2 million views per month, which result in $M = 1386667$. Changing this market size will change the constant term of the utility function without affecting other parameters.[37]

[37] The website was established in 1999 and is at mature stage now. The website level viewership remained fairly stable over the observation period, thus we do not consider the growth of market size in this study.

Table 4: Summary Statistics - Content, Network Position, and Borrowed Content

	Mean	SD
Content	27.83	35.69
Network Position	5.18E-03	2.61E-03
Borrowed Content	1304.5	780.4

As discussed in section 3, multiple functional forms of the function $f(.;\beta_i)$ in equation (1) need to be estimated, with the best model chosen with certain model selection criterion. This flexibility is important because our treatment of the viewership market is reduced form, so estimating multiple functional forms can give us more robust results. We estimate the following four specifications:

$$(20)\quad \begin{cases} I(Linear) & f(C_{j,t},P_{j,t},C^b_{j,t};\beta_i)=\beta_{i,1}C_{j,t}+\beta_{i,2}P_{j,t}+\beta_{i,3}C^b_{j,t} \\ II(Linear-Quadratic) & f(C_{j,t},P_{j,t},C^b_{j,t};\beta_i)=\beta_{i,1}C_{j,t}+\beta_{i,2}{C_{j,t}}^2+\beta_{i,3}P_{j,t}+\beta_{i,4}{P_{j,t}}^2+\beta_{i,5}C^b_{j,t}+\beta_{i,6}{C^b_{j,t}}^2 \\ III(Log) & f(C_{j,t},P_{j,t},C^b_{j,t};\beta_i)=\beta_{i,1}\log(C_{j,t})+\beta_{i,2}\log(P_{j,t})+\beta_{i,3}\log(C^b_{j,t}) \\ IV(Log-Embedded) & f(C_{j,t},P_{j,t},C^b_{j,t};\beta_i)=\beta_{i,1}\log(C_{j,t}+\beta_{i,3}C^b_{j,t})+\beta_{i,2}\log(P_{j,t}) \end{cases}$$

Specification I is the simplest functional form that accounts for all three factors, and we expect each coefficient to be positive to reflect their positive impact on viewership demand. Specification II extends the first specification by including a quadratic term for each factor to account for potential diminishing rate of return. For example, although linking to other producers provides a convenience benefit to viewers, when there are too many such links, viewers could also feel annoyed, so the content borrowing effect could become saturated. Similarly, although having higher network position gives a producer's content favorable placement, this benefit may become saturated beyond a certain threshold, if the network position is high enough to distinguish the producer in most cases. The quadratic terms are used to capture such effects. Specification III explicitly accounts for such diminishing return by using log transformation.

Finally, Specification IV also uses log transformation, but adds a weighted component of the borrowed content to the original content before applying the log.

The quality factors are included in our study mainly for control purposes, and we adopt a simple linear functional form for the quality as well as borrowed quality:

$$g(Q_j, Q_j^b; \gamma) = \gamma_0 + \gamma_1 Diversity_j + \gamma_2 Popular_j + \gamma_3 Advisor_j + \gamma_4 Diversity_j^b + \gamma_5 Popular_j^b + \gamma_6 Advisor_j^b \quad (21)$$

The result of estimation is presented in Table 5 (covariates are standardized). In all four specifications I-IV, the coefficients for content, network position, and borrowed content are all positive and statistically significant. This is clear evidence that all three are important factors in determining viewership demand, where higher content volume, more prominent network position, and more borrowed content all lead to higher viewer utility and in turn higher viewership demand for the reviewer's reviews. The coefficients for the three quality factors are also all positive and statistically significant, suggesting they positively influence viewership demand. Among them, the popularity indicator has the highest impact on viewer utility.

Table 5: Viewership Demand Estimation

Model Specification	I	II	III	IV	V	
	Linear	Linear Quadratic	Log	Log-Embedded	LatentClass	
Parameter					Segment 1	Segment 2
Content	0.2679(***)	0.3001(***)	0.1026(***)	0.1189(***)	0.5923(***)	-0.2268(.)
Content^2		-0.0041			-0.0348(*)	0.0429(*)
BorrowedContent	0.1273(.)	0.2260(**)	0.0731(**)	0.0134(.)	0.1115	1.7342(***)
BorrowedContent^2		-0.0480(**)			-0.0192	-0.2986(***)
NetworkPosition	0.3505(***)	0.8268(***)	0.5922(***)	0.7991(***)	2.112(***)	4.1225(***)
NetworkPosition^2		-0.0949(***)			-0.5463(***)	-0.4509(***)
Diversity	0.0945(***)	0.0796(***)	0.1357(***)	0.1420(***)	0.0982(***)	
Popular	0.5224(***)	0.5219(***)	0.5238(***)	0.5656(***)	0.5024(***)	

Advisor	0.0818(***)	0.0491(*)	0.0716(***)	0.0432(**)	0.0463(***)	
BorrowedDiversity	-0.0303	-0.0575	-0.1188(**)	-0.0684(***)	-0.1212(*)	
BorrowedPopular	0.1853(***)	0.1759(***)	0.1730(***)	0.1562(***)	0.2578(**)	
BorrowedAdvisor	0.1265(**)	0.1221(**)	0.1366(***)	0.1819(***)	0.0832(*)	
Constant	-10.1255(***)	-10.4231(***)	-7.7108(***)	-7.8858(***)	-10.8349(***)	
Segment Size					0.932	0.068
-LL	6930.51	6734.51	6977.46	7128.42	6692.44	
BIC	6990.82	6820.67	7037.77	7188.73	6847.52	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Looking at specification II, we find that content and borrowed content have similar contributions to the viewer utility, while network position has higher impact than both content and borrowed content. Specification II shows the quadratic terms of borrowed content and network position both have negative signs, suggesting that diminishing return exists for both factors. The quadratic term of content is also negative. However, its magnitude is very small and it is not statistically significant. Thus there is no clear evidence of diminishing return on the content dimension. Specifications III and IV both use log transformation, where the coefficient magnitude corresponds to percentage change. Specification V is the latent class version of specification II with two segments. In both segments, both network position and borrowed content positively influence viewer utility and exhibit diminishing returns. The first segment has content coefficient larger than that in specification II. Interestingly, the second segment has a negative content coefficient, and the coefficients for network position and borrowed content are quite large. This seems to suggest that this portion of the demand is mainly driven by the position in the network and the borrowed content, but not by the producer's own content.

Among the five competing model specifications, specification II, the Linear-Quadratic specification, has the best model fit after adjusting for number of parameters using BIC. We

therefore adopt this specification as the per-period viewership demand equation for the estimation of dynamic structural parameters.

5.3 Result – Dynamic Competition

5.3.1 Policy Function

The policy function regression, which captures reviewers' writing and linking decisions, is only the intermediate step for estimating the dynamic model parameters, and the coefficients are not interpretable. Instead, we report a few patterns of producer actions based on their content and network position states.[38] Note that the policies, estimated in a reduced-form fashion, constitute the equilibrium play resulting from reviewers' dynamic competition, and encapsulate the concept of best response. In this section we simply present the observed patterns. In the subsequent section 5.4, we investigate in detail how incentives and strategic interactions lead to such actions.

Figure 1 shows the average daily content production, conditional on the reviewer's own state along the content and network position dimension. As shown in the figure, reviewers with higher content volume in general write more frequently, and it is more so for reviewers with low network positions. In fact, reviewers with high content volume but low network position write reviews most frequently. This could be unexpected at the first look – the viewership demand equation, which captures the payoff through immediate viewership, shows that reviewers with higher network positions have higher marginal benefit and thus should have higher propensity to produce content. In section 5.4, we show how this discrepancy is explained with the dynamic tradeoffs faced by reviewers.

[38] Actions can be summarized according to other dimensions, too, such as quality. In this study, we focus on the two dimensions, content and network position, as they are the direct results of reviewers' review writing and link formation actions. As specified in section 4, content and network positions are each partitioned into quintiles for the policy function regression, so we report the action patterns based on the quintile partitions along these two dimensions.

Figure 1: Average Content Production by Own State

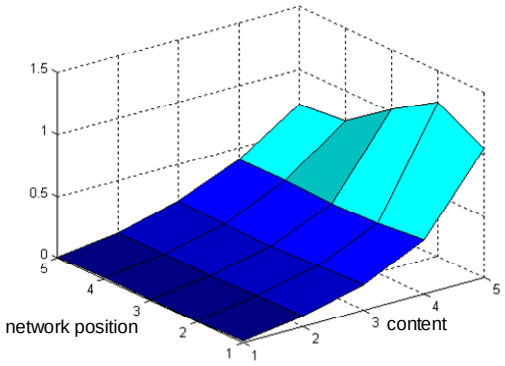


Figure 2: Probability to Form Links by Own State

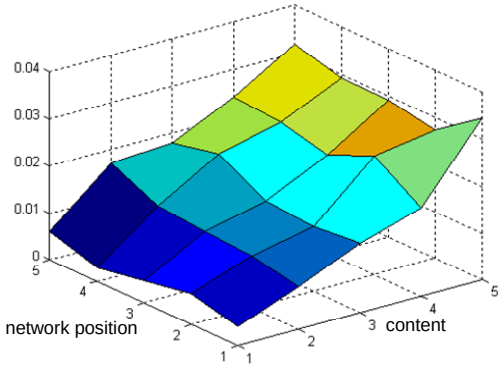


Figure 3: Relative Link Probability – Non-reciprocal over Reciprocal

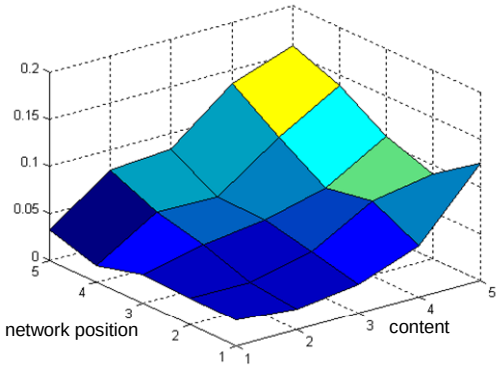


Figure 2 shows the frequency of creating an outgoing link, conditional on reviewers' own state. Reviewers with higher content volume create links more frequently. Reviewers with very

high network positions (5-th quintile) also create links with have higher frequencies, although not by much.

Since our analysis in section 3.4 indicates that reciprocal links would be favored by reviewers, we also report the relative probability of creating a non-reciprocal link over that of a reciprocal one conditional on a reviewer's own state, as shown in Figure 3. The first to note from the figure is that reviewers of all states are much more likely to create reciprocal links than non-reciprocal ones – the ratios are all much smaller than 1. Furthermore, reviewers with higher content volume have higher relative probability to create non-reciprocal links.

Other patterns are that reviewers with more content are more likely to receive incoming links, and that reviewers with different network positions have similar likelihood of receiving incoming links as long as they have similar content, with higher network positions increasing the likelihood but only slightly. Together, these patterns summarize the decisions made by reviewers as they interact with one another in the competition, each trying to maximize her own benefit. The incentives behind these actions are analyzed in detail in section 5.4.

5.3.2 Cost Estimation

We now discuss the estimate of the dynamic structural parameters, i.e. cost parameters. To operationalize the estimation, we randomly pick 500 states from the dataset. For each state, we randomly pick one reviewer and performed two forward-simulations. In the first, all reviewers follow the equilibrium strategy according to the estimated equilibrium policy, while in the second simulation, the chosen reviewer follows a perturbed strategy. Each simulation is run for 600 periods, and repeated multiple times with the average taken. We set the discount factor to 0.9995 as our observation is at daily level, which is similar to the 0.995 often set in dynamic structural studies when data is at weekly level (e.g. Erdem & Keane 1996). We then run the

minimum distance estimator to find the cost parameters which minimize the deviation from the optimality condition of equilibrium, as specified in equation (19). The standard errors of the estimates were obtained through re-sampling of the chosen state-player pairs.

For the production cost function, we adopt a linear functional form.[39] Production cost may depend on the reviewer's quality, as a reviewer needs to exert more effort to achieve higher quality. The reviewer's tenure might also influence cost, due to learning-by-doing. Considering this, we assume the unit cost of production is a linear function of the reviewer's effective quality and tenure with the website, as shown in equation (22). We also assume the cost of linking is a linear function of the reviewer's effective quality and tenure, plus an indicator of whether the link is reciprocal, as shown in equation (23). This final term is added to tease out possible intrinsic value of forming reciprocal links – an intrinsic preference for reciprocal links, aside from the consideration of how it affects viewership, would imply lower cost of forming reciprocal links that non-reciprocal links at the model primitive level, and be reflected from a negative coefficient for this final term.

(22) $$c^{prod}(a^{p}_{j,t}, X_j; \phi) = a^{p}_{j,t} \cdot (\phi_0 + \phi_1 g(Q_j, \gamma) + \phi_2 Tenure)$$ [40]

(23) $$c^{link}(a^{l}_{j,t}, X_j; \psi) = \psi_0 + \psi_1 g(Q_j, \gamma) + \psi_2 Tenure + \psi_3 I\{a^{l}_{j,t} \text{ is reciprocal}\}$$

Table 6: Dynamic Cost Parameter Estimation

	Estimate	Low 95% CI	High 95% CI

[39] A strictly convex cost function is often used in industrial organization literature. In our empirical application, however, it is reasonable to assume there is a unit cost for writing a review article, hence the linear form. Equilibrium condition holds as long as the market size is finite. We also estimated the quadratic specification of cost function, and the result when averaged for unit cost is similar to the linear specification. The result for quadratic cost function is available from the author upon request.

[40] In a slight abuse of notation, we use the same function symbol, *g*, to represent a reviewer's own quality effect: $g(Q_j; \gamma) = \gamma_0 + \gamma_1 Diversity_j + \gamma_2 Popular_j + \gamma_3 Advisor_j$, excluding the borrowed quality effect – cost should be determined by a reviewer's own characteristics.

Production			
Constant	0.1481(*)	0.1326	0.1597
Quality	0.2268(*)	0.1866	0.2536
Tenure	-0.0051	-0.1156	0.0101
Link			
Constant	0.025(*)	0.0015	0.0715
Quality	-0.0646	-0.1703	0.0556
Tenure	-0.0132	-0.0226	0.0026
Reciprocal	0.1761	-0.1315	0.4755

Unit of measure: thousand page views

The result of the estimation is reported in Table 6. The constant term for the production cost regression is 0.148 and statistically significant at .95 level. This means the cost of writing a review article is equivalent to the benefit of 148 page views, which is a reasonable number for unit cost estimate, as the summary statistics show that on average a review article is viewed a little over 200 times. The coefficient for reviewer quality is positive and statistically significant. This suggests that reviewers of higher quality put more effort in writing product review articles and thus incur higher cost per article written, consistent with expectation.[41] The estimate also shows that a reviewer's tenure at the website does not have a significant impact on her production cost.

The cost of linking is very close to zero, indicating that linking itself is not a high effort activity. Neither quality nor tenure is shown to have a significant effect on the cost of linking. More notable is the coefficient for the reciprocity term. The positive sign of the coefficient shows that the cost of forming a non-reciprocal link is less than that of forming a reciprocal link, although the result is not statistically significant. As discussed earlier, the existence of intrinsic

[41] A more general model is to assume that all reviewers are of the same type, and that when they write articles they can choose to write either a high or a low quality one, with the former entailing higher cost than the latter, similar for links. However, to estimate such a model requires quality information at the level of each review article, which we do not have. This is beyond the scope of our study and is left for future work. Our model can be considered as a restricted model in this broader context – each reviewer is restricted to choose a quality type and then follow it throughout the whole period.

value for reciprocal links would be reflected from a negative coefficient for this term, thus there is no evidence of such intrinsic value. That reciprocal links are more likely to be formed, as observed in the dataset, thus should be mainly attributed to the strategic considerations, i.e. the promote-the-promoter effect as discussed in section 3.4.

5.4 Decision Dynamics and Interdependence

Using the estimated viewership demand equations, the dynamic cost parameters, and the equilibrium policy, we now investigate the competitive dynamics in detail. To address the research questions raised for this study, we analyze three aspects of the competitive dynamics: First, we investigate the incentive to form links and how it depends on link types and reviewer states. Next, we analyze how linking influences content production decisions. Finally, we evaluate the net benefit accrued to reviewers at different states and the market structure that emerges from the competition.

5.4.1 Dynamics of Link Formation

The decision of whether to create a link and whom to link to is driven by both the tradeoff between borrowed content and network positions, and the dynamic interactions between reviewers. To understand the incentives to form links for reviewers at different states, we evaluate how such links impact reviewers' viewership demand.

To analyze the implication of forming links, we first quantify, using the dataset, the average change in network position through establishing an outgoing link and that through receiving an incoming link given a reviewer's state. Receiving an incoming link normally increases the reviewer's network position noticeably. Creating an outgoing link, however, reduces the network position, and a reciprocal link typically leads to smaller reduction than a

non-reciprocal link as discussed earlier. We then use a subset of the data, covering the three-month period from January 2009 to March 2009, to calculate the incremental benefit of creating a link for each reviewer in each day. For creating a reciprocal link, the incremental benefit is calculated as the difference in discounted viewership between two otherwise identical scenarios except that in the second scenario the focal reviewer creates a reciprocal link to another reviewer who already links to her. Other factors are held constant in this calculation. This calculation captures the effect of creating a reciprocal link, which can be considered as a "close-loop" action.[42] For creating a non-reciprocal link, however, this calculation captures only the direct effect, i.e. the tradeoff between more borrowed content and lower network position, but not the strategic aspect arising from dynamic interactions, i.e. the target reviewer may decide to reciprocate in future. To account for this dynamic interaction, we calculate the probability of a non-reciprocal link being reciprocated in future and the average days taken to receive the reciprocation, conditional on the source reviewer's state, using the equilibrium policy recovered from data. We then calculate the change in discounted viewership assuming that a reciprocal link is established with the corresponding probability and delay.

[42] We can consider that a reciprocal link *finishes* a round of dynamic interaction – the target reviewer already has a link pointing back and will not further "respond" to the reciprocal link. Thus a "loop" is closed. In contrast, a non-reciprocal link *starts* a round of dynamic strategic interaction – the target reviewer will in subsequent periods decide whether to reciprocate. Thus a loop is opened.

Figure 4: Effect of Creating Reciprocal Links

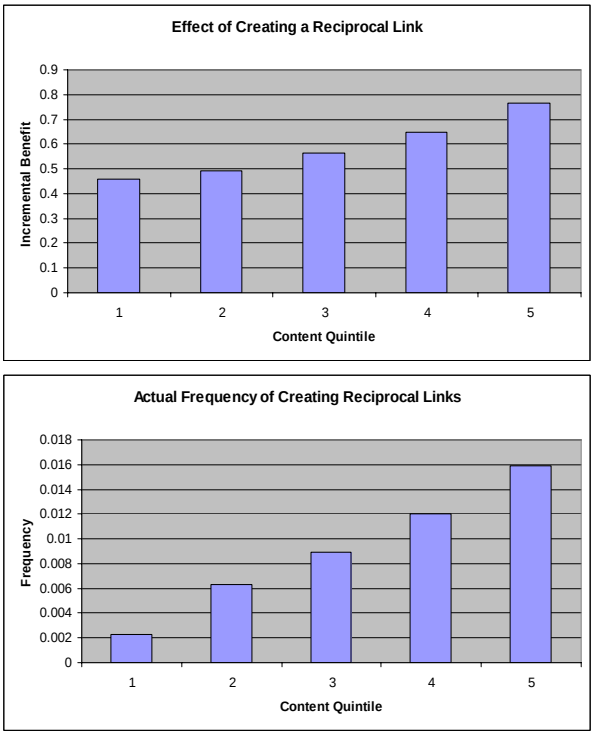
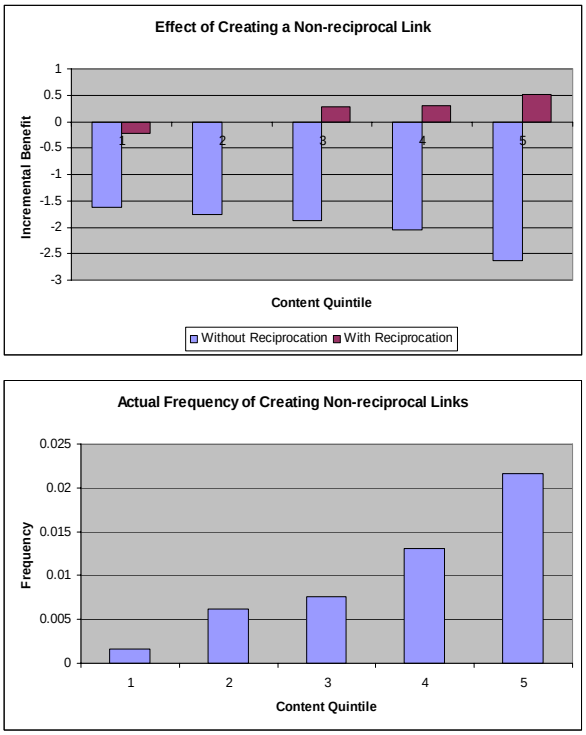


Figure 5: Effect of Creating Non-reciprocal Links



The incremental benefit of creating a reciprocal link is reported in Figure 4. The result is summarized along the content dimension in quintiles. The top figure shows positive average effects for all five quintiles, suggesting that in general the benefit of more borrowed content outweighs the cost of reduced network position through forming a reciprocal link. Also, the figure shows that reviewers with more content benefit more from a reciprocal link. This is consistent with the policy function where reviewers with more content are more likely to create reciprocal links, as shown in the bottom figure of Figure 4.

The result for creating a non-reciprocal link is reported in Figure 5, also summarized along the content dimension. Creating a non-reciprocal link typically reduces network position more than does a reciprocal one. As shown in the first series of the top figure, which includes the direct effect but does not account for future reciprocation, the average incremental benefit is

negative for all five reviewer quintiles, suggesting that the cost of reduced network position outweighs the benefit of more borrowed content. The incremental benefit is also significantly lower than that of forming reciprocal links. Recall that the cost estimate in section 5.3 shows no evidence of intrinsic value for reciprocal links, we know that in the context of this study, the tendency towards reciprocity is mainly explained by the comparatively favorable impacts of reciprocal links on viewership, due to the promote-the-promoter effect. This is a notable result. Sociology literature has long recognized the prominence of reciprocity in social networks, and statistical network models often consider that as model primitives. Our study provides an alternative explanation in a rational economic rather than social context, that reciprocity can be naturally favored by dynamic strategic considerations, without the need for a social explanation as model primitive.[43]

However, the first series in the top figure also shows that the more content a reviewer has, the lower her incremental benefit from forming a non-reciprocal link, yet the policy function shows that reviewers with more content are more likely to create non-reciprocal links (the bottom figure). Thus a static perspective alone does not explain the linking actions well. This discrepancy is resolved once the dynamic strategic perspective is taken into account. As the second series in the top figure shows, after accounting for future reciprocation, the incremental benefit of forming a non-reciprocal link increases significantly for all five quintiles, and reviewers with more content have higher incremental benefit. This is because a reviewer with more content is more confident to see the target reviewer reciprocate, and with shorter delay. After all, the target reviewer also can benefit from borrowed content, and when she decides to create a link, she would favor a reciprocal one to maintain her own network position, thus

[43] That is, an explanation such as "people tend to form reciprocal links because by nature they like reciprocity, i.e. there is an intrinsic value to reciprocate".

making the source reviewer a favorable target. This incentive to reciprocate is further enhanced when the source reviewer has more content. In essence, a reviewer is "inviting" reciprocation when creating a non-reciprocal link, in anticipation of the strategic response from the target reviewer, and the more content a reviewer has, the more effective this strategy is. Comparing the two scenarios clearly shows how the dynamic interactions drive reviewers' linking decisions.

5.4.2 The Impact of Linking on Content Production

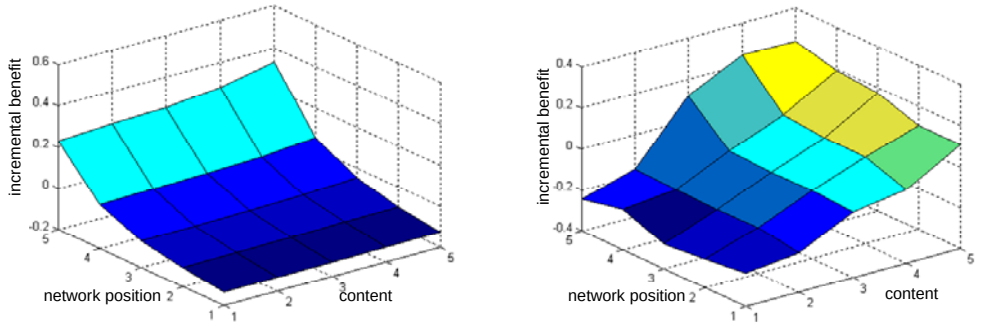
We now analyze how linking influences reviewers' content production decisions. Similar to the analysis of link formation, we evaluate the incremental effect of writing reviews by reviewers at different states, accounting for the dynamic interaction effects arising from linking.

For this analysis, we use the same subset of the data as used in analyzing link formation. We begin with analyzing the direct effect of writing reviews: for each day and each reviewer, we calculate the difference in discounted viewership between two otherwise identical scenarios: in the first, the focal reviewer does not write reviews; in the second, she writes one review article, which depreciates at the estimated depreciation rate. This difference in viewership approximates the direct incremental benefit of producing one unit of content, from which the production cost is then subtracted to arrive at the net incremental benefit. The result is shown in Figure 6(A). The direct incremental benefit is much higher for reviewers with higher network positions, and is positive only for reviewers in the top two quintiles of the network position dimension. Reviewers with more content also get higher benefit, but the difference along the content dimension is not as large as along the network position dimension.

The direct benefit is only one part of the incentive behind content production. When deciding whether to produce content, a reviewer considers not only the immediate viewership, but also the future linking actions of other reviewers. For example, if a reviewer in a high content

state anticipates other reviewers to link to her in the near future, which leads to higher network position, then her incentive to produce will be increased, as the additional benefit from higher network position later on adds to the direct benefit. Whereas if a reviewer in another state expects her competitors to receive more incoming links, which reduces her relative position in the network, then her incentive to produce will be lower than suggested by the direct benefit. Thus linking could significantly alter the incentive to write reviews depending on the states of reviewers.

Figure 6: Incremental Benefit of Content Production by State



(A) not account for change in network position

(B) account for change in network position

To analyze how linking influences the incentive to produce, we use the equilibrium policy to calculate the average change in network positions, arising from linking, corresponding to different reviewer states. The calculation shows that reviewers at high content and low network position states get highest average increase in network position over time, while reviewers in the opposite states see their network positions reduce later on. We then incorporate this state-dependent change in network position into the calculation of the incremental effect of writing one more review. The net incremental benefit calculated this way, reported in Figure 6(B), shows that once the prospect of linking is accounted for, content level, instead of network position, becomes the main determining factor of the incremental benefit. Reviewers in the top two

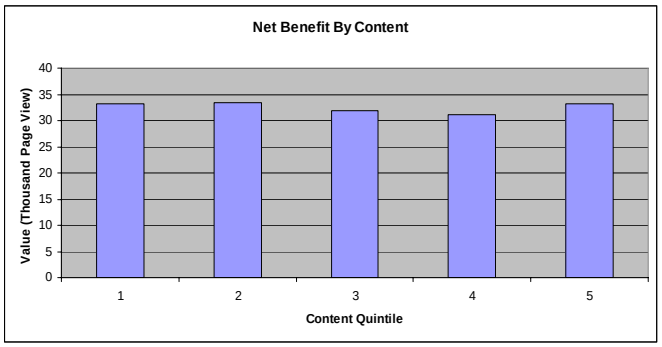
quintiles of the content dimension have positive net benefit from producing content, while other reviewers have negative benefit, even for those with high network positions. In other words, a reviewer with high content but low network position writes reviews because she expects other reviewers to link to her later on, even though the immediate viewership is not much. Meanwhile, a reviewer at the opposite state finds it not worthwhile to write reviews as she foresees lower network position ahead. This is consistent with the observed policy function, which shows that content level influences the frequency of writing reviews more than network position does.

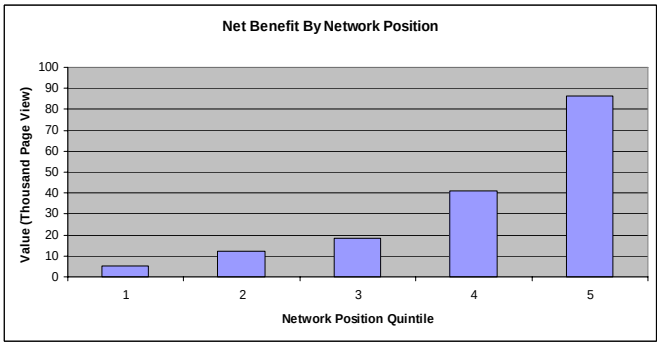
In summary, the results demonstrate a close interdependence between link formation and content production, and that linking is a major driver of reviewers' writing decisions. Interestingly, the prospect of linking *encourages* the content production of reviewers with high content and low network positions, while *discourages* the content production of reviewers with low content and high network positions.

5.4.3 Net Benefit and Market Structure

Given the viewership demand and the cost estimates, we now analyze the net benefit accrued to reviewers at different states. To do so, we calculate the discounted net benefit over rolling six-month windows over the entire period. Net benefit is simply the viewership minus the cost of production and linking. Costs are derived from reviewers' actual decisions while viewership is inferred from reviewers' states and the estimated viewership demand equation.

Figure 7: Net Benefit by Content and Network Position





The result is presented in Figure 7, where we show the average net benefit on the content dimension and network position dimension. The figure shows that reviewers with higher network positions derive significantly higher benefit than reviewers with lower network positions. In contrast, however, reviewers with more content do not derive higher benefit than reviewers with less content. This may be unexpected at first look, as the demand equation shows that more content leads to higher viewership. But it is explained by the cost side: although reviewers with more content can attract higher viewership, they also incur higher cost as they write more reviews. The result shows that additional viewership demand is mostly offset by the increased cost, making a reviewer with more content no better than one with less content in terms of net benefit. This result is also reasonable when we consider the competitive effect: since content level is determined solely by a reviewer's own production decisions, were there to be

significantly higher net benefit with higher content level, all reviewers would write more and in so doing, the potential advantage from higher content level would be largely competed away.

In contrast, the advantage coming from higher network positions cannot be competed away as easily. This is because, even though desirable, a reviewer cannot unilaterally increase her network position. Instead, it takes incoming links from *other* reviewers to increase that. Thus a reviewer may enjoy significant advantage from having a high network position, while other reviewers lack an effective way to counter that. Indeed, our calculation suggests that higher network positions offer significant competitive advantage and lead to higher net benefit. The competition thus results in a market where reviewers are differentiated along the network position dimension, while on the content dimension surplus is mostly competed away.

The contrast between content and network position should be taken note by companies operating such websites. As a website seeks to maximize its overall viewership, it should seek to encourage content production by creating a competitive environment internally. Any form of "sticky" competitive advantage enjoyed by a subset of producers, such as that led to by higher network positions in this context, may create imbalance in the system. This imbalance can potentially lead to differentiations that soften the competition and reduce the overall content level at the site. Consequently, the effect of linking to the overall website viewership is a matter of concern that is worth further investigation.

5.5 The Effect of Linking on Website Viewership – A Simulation

For marketing managers who operate those content websites, it is important to know whether the network among content producers increases the overall viewership at the website, and how the linking feature should be designed to generate optimal viewership outcome. If content is exogenously given, then we would expect a network superimposed among producers to increase

overall viewership, as linking enhances content.[44] When content production is determined endogenously in a dynamic context, however, the overall effect of network is not at all clear. Qualitative analysis of tradeoffs also reveals forces towards both directions. On one hand, since a producer with more content is more likely to receive incoming links, linking provides an incentive for certain producers to produce more content. On the other hand, however, linking could also discourage other producers from producing content, as is shown in section 5.4. At the website level, the content enhancement effect of linking is expected to increase overall viewership. However, if there are producers with high content volume sitting at obscure positions in the network, while others occupy prominent positions yet do not have much content, then the network may hinder efficiency by not effectively directing viewer traffic to content.

Given these factors with opposite effects, the network could either increase or decrease overall viewership. A sign of concern, though, is that as shown in section 5.4, reviewers with more prominent network positions enjoy significant competitive advantage, yet competitive advantage enjoyed by a small set of players in general reduces competition intensity. This suggests that the current network may impede the competition among content producers, and that alternative policies regulating link formation may help the website improve overall viewership.

To evaluate the overall effect of network, ideally we want to compare two situations which are otherwise identical, except that in the first link creation is allowed, while in the second it is not. Similarly, we want to compare situations under alternative linking regulations, such as restricting the total number of links a developer can create, to find out which link regulation leads to best viewership outcome for the website. However, current methodological restrictions

[44] This is consistent with existing literature, which shows that network increases overall sales in an online shopping center environment (Stephen and Toubia 2009).

prohibit us from making these comparisons directly, as it requires explicitly solving for the equilibria of alternative dynamic games, which are computationally infeasible.[45]

Considering this, we resort to a "second best" approach by performing simulations which alter initial states but do not alter the existing equilibrium – since it is only the initial states that changes, while the structural parameters and the game remain the same, the existing equilibrium recovered from data still applies. To analyze through this approach whether the imbalance induced by the current network reduces viewership, we pick a state from the data, and for each reviewer, we randomly remove her outgoing links until she has no more than five outgoing links remaining.[46] After this system-wide link removal, the network becomes more sparse and balanced. We then perform two forward simulations for 60 periods, with the first starting from the original state and the second from this new state after the link removal. We compare content production, link formation, and overall viewership between these two simulations to evaluate the overall effect of the network.

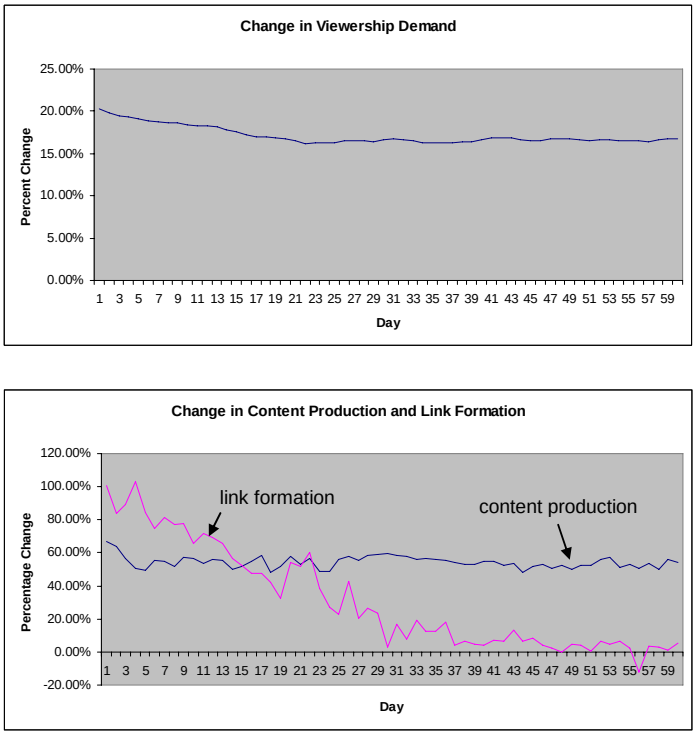
Table 7: Simulation – System-wide Link Removal

	Demand	Content Production	Link Formation
Average Increase Per Day	12.096	38.08	1.22
Percentage Increase	17.12%	54.40%	31.32%

[45] BBL recovers structural parameters without explicit computation of equilibrium, thus bypassing the "curse of dimensionality" issue. In policy simulation, however, any change in the "rule of the game" can potentially lead to a new equilibrium, so equilibrium must be explicitly computed. For example, if link formation is prohibited, all reviewers will adjust their strategy for writing reviews accordingly. To evaluate that change, we must compute the new equilibrium, the cost of which is prohibitive given the number of reviewers in our study. Recent methodological advancement, e.g. the concept of oblivious equilibrium developed in Weintraub, Benkard, and Van Roy (2008), can potentially solve this issue, with the drawback that the solution concept itself is an approximation. We leave the potential use of oblivious equilibrium for future work.

[46] We do not remove all links to avoid potential issues of boundary bias for the estimated policy functions.

Figure 8: Simulation – System-wide Link Removal



The result of the simulation is reported in Table 7. With the system wide link removal, content production, link creation, and overall viewership demand all increase significantly. The average daily viewership increases by 17.12% for the website overall, while content production and link creation both increase by more than 30%. This suggests that the current network among content producers, although providing benefit through enhancing content, also brings too much market power to certain producers – those with high network positions – and impedes efficiency. When the field of competition is leveled, competition intensifies, with reviewers collectively producing more content, and the overall viewership at the website increases.

Figure 8 shows the simulation result in further detail along the time dimension. The viewership demand jumps immediately after the link removal, likely because reviewers with

high content volumes but low network position now become more visible and attract more viewership. Over time, the demand increase moderates slightly but still holds stably above 15%. This is supported by sustained increase in content production – with a leveled playing field, the competition is intensified and reviewers collectively have higher incentive, or are forced, to produce more content. Link creation also jumps initially, but this is comparatively short-lived, as link creation falls back to the pre-removal rate after about thirty periods.

In summary, this simulation provides evidence that the current design over time leads to inefficient internal competition among reviewers. Alternative policies that regulate link formation could potentially lead to overall viewership and should be considered for experimentation at the website.

6. Discussion, Limitation and Conclusion

The advent of online social media brings about many intriguing phenomena. A prominent one is the emergence of a large number of revenue sharing content websites, which rely on external content producers to supply content and induce an internal competition for viewership among producers. The linking feature recently introduced to many websites further leads to complex and intriguing dynamic interactions among content producers. Meanwhile, the implication of linking on the overall viewership, crucial to the website platform builders, remains an open question. A detailed understanding of producers' interactions and their implications thus not only is of academic interest, but also has important managerial implications, as this phenomenon is quickly gaining momentum in the industry.

Motivated by this, we develop a dynamic oligopoly model to study the competition among content producers. In our model, producers compete against one another through producing

content and forming links, and we characterize their strategic interactions using the solution concept of Markov-perfect equilibrium. We estimate the model using the data obtained from a popular product review website, leveraging the two-step estimation approach developed in Bajari, Benkard, Levin (2007), and provide a detailed analysis of the interactions among reviewers in their decision process.

Our study contributes to the literature by investigating the interactions of link formation and content production decisions, by analyzing the inter-temporal tradeoffs that drive the interactions dynamically, and by providing a rational economic framework for empirically studying the formation of networks in a dynamic strategic setting. Our study leads to several findings with managerial implications. We find that viewership demand is positively influenced by content volume and network positions, and there is a content borrowing effect through linking. We find that reciprocal links are more likely to be formed than non-reciprocal ones, and this is encouraged by the nature of the strategic interaction – a promote-the-promoter effect. This tendency towards reciprocity further induces producers with high content volume to strategically create non-reciprocal links, in anticipation of reciprocation later on which will enhance their network positions. We find that the prospect of linking encourages producers with high content volume but low network position to produce more content, yet discourages producers at opposite states. Furthermore, we find that the producers' net benefit increases with their network positions but not with their content volume, as the higher viewership from more content is offset by the higher cost incurred in producing the content. This suggests that linking may lead to inefficiency as competitive advantage is accrued to producers with high network positions. Finally, our simulation suggests that limiting the links at the website may lead to higher content production and overall viewership demand.

Managers who operate content websites can consider several alternative linking policy designs to improve efficiency. They could prohibit linking altogether by not offering the feature. This will prevent competitive advantage from being accrued to a subgroup of producers. Between completely disabling linking and not regulating linking at all, an alternative at the middle ground is to restrict the number of links each producer can form. This could alleviate the imbalance over time, while producers would also become more selective in forming links. Another alternative is to impose a time limit on links so that they expire after some time. This could make the network structure less rigid and ease the issue of unbalanced competition. Methodological restrictions limit our ability to analyze these alternative policies in detail, while industry managers could explore these and other policies through experimentation at the websites.

A few other limitations of our study can be addressed in future work. First, our study focuses on the profit motive of content producers, and we use a group of top producers for our analysis. Although most websites have a significant share of their viewership generated by a small group of elite producers, there is also a larger group of more casual content producers. This mass group of casual producers may have incentives other than profit, and a richer model is called for to study their behaviors and contributions to the business. Second, in the social media market, the line between consumers and producers is blurred. While our focus on the small group of elite producers allow us to still follow the traditional supply-side demand-side dichotomy, an exciting opportunity exists to advance the literature by investigating the dual roles the website users may play. Finally, not all content is the same, and different content may be either complements or substitutes. For feasibility reasons, our model considers all content to be of the same type, while we leave the interactions induced by different content types for future research. We also hope that, with the rapid advancement in econometrics on dynamic game estimation

methodologies, we will be able to admit more heterogeneity among producers in future, and to explicitly evaluate the effects of alternative policies when they lead to different equilibrium situations.

Online content markets, and social media in general, bring much closer and more dynamic interactions among consumers, between consumers and producers, and among producers, than the traditional offline market does. With that, it also opens an exciting frontier for marketing research. Our work is an early step towards this direction, and we are confident that future research will bring further insights in this area and offer much needed managerial guidance.

References

Abhishek, V. and K. Hosanagar (2007), “Keyword Generation for Search Engine Advertising Using Semantic Similarity Between Terms,” Proceedings of the Ninth International Conference on Electronic Commerce, 258, 89-94.

Ackerberg, D., C.L. Benkard, S. Berry and A. Pakes, (2007), “Econometric Tools for Analyzing Market Outcomes”, Handbook of Econometrics, 6, 4171-4276

Agarwal, A., K. Hosanagar and M.D. Smith (2009),”Location, Location, Location: An Analysis of Profitability of Position in Online Advertising Markets,” Working Paper.

Aguirregabiria, V. and P. Mira, (2007), “Sequential Estimation of Dynamic Discrete Games”, Econometrica, 75, 1-53

Ansari, A., O. Koenigsberg and F. Stahl, (2011), “Modeling Multiple Relationships in Social Networks”, Journal of Marketing Research, forthcoming

Aral, S., L. Muchnik and A. Sundararajan, (2009), “Distinguishing Influence Based Contagion from Homophily Driven Diffusion in Dynamic Networks”, Proceedings of the National Academy of Sciences, Vol. 106, No. 51, 21544-21549

Athey, S. and G. Ellison (2011), “Position Auctions with Consumer Search,” Quarterly Journals of Economics, forthcoming

Axelrod, R. and W. Hamilton, (1981), "The Evolution of Cooperation," Science, 211, 1390-96

Bajari, P., C.L. Benkard and J. Levin, (2007), “Estimating Dynamic Models of Imperfect Competition”, Econometrica, 75, 1331-1370

Bala, V. and S. Goyal, (2000), “A Non-cooperative Model of Network Formation,” Econometrica, 68, 1181–1230

Balachander, S., K.N. Kannan (2009), and D.G. Schwartz, “A Theoretical and Empirical Analysis of Alternate Auction Policies for Search Advertisements,” Review of Marketing Science, 7(5)

Brynjolfsson, E., A.A. Dick and M.D. Smith (2009), “A nearly perfect market? Differentiation vs. price in consumer choice,” Quantitative Marketing and Economics, 8(1), 1-33

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

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

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

Berry, S.T., (1994), “Estimating Discrete-Choice Models of Product Differentiation”, The RAND Journal of Economics, 25, 242-262

Berry, S., J. Levinsohn and A. Pakes, (1995), “Automobile Prices in Market Equilibrium”, Econometrica, 63, 841-890

Besanko D. and U. Doraszelski, (2004), “Capacity Dynamics and Endogenous Asymmetries in Firm Size”, The RAND Journal of Economics, 35, 23-49

Bonacich, P., (1987), “Power and Centrality: A Family of Measures”, The American Journal of Sociology”, 92, 1170-1182

Bonacich P. and P. Lloyd, (2001), “Eigenvector-like Measures of Centrality for Asymmetric Relations”, Social Networks, 123, 91-201

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

Braun, M. and A. Bonfrer, (2010), “Scalable Inference of Customer Similarities from Interactions Data Using Dirichlet Processes”, Marketing Science, forthcoming

Brin, S. and L. Page, (1998), “The Anatomy of a Large-scale Hypertextual Web Search Engine”, Computer Networks and ISDN Systems, 30, 107-117

Camerer, C.F., T.H. Ho and J.K. Chong (2004), “A Cognitive Hierarchy Model of One-Shot Games,” Quarterly Journal of Economics, 119(3), 861--898.

Chen, Y. and C. He (2006), “Paid Placement: Advertising and Search on the Internet,” Working Paper.

Chevalier, J.A. and D. Mayzlin, (2006), “The Effect of Word of Mouth on Sales: Online Book Reviews”, Journal of Marketing Research, Vol. XLIII, 345-354

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

Chintagunta, P., S. Gopinath and S. Venkataraman, (2010), “The Effects of Online User-Reviews on Movie Box-Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets”, Marketing Science, 29(5)

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

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

Danaher, P (2007), "Modeling Page Views Across Multiple Websites With An Application to Internet Reach and Frequency Prediction," Marketing Science, 26(3), 422--437.

Desai, P. and W. Shin (2009), "Advertiser-Specific Minimum Bids and Advertising Budgets in Keyword Search Auctions," Working Paper.

Dube, J.P., G.J. Hitsch and P.E. Rossi, (2009), "Do Switching Costs Make Markets Less Competitive?" Journal of Marketing Research, Vol. XLVI, 435–445

Edelman, B., M. Ostrovsky and M. Schwarz (2007), "Internet Advertising and the Generalized Second Price Auction: Selling Billions of Dollars Worth of Keywords," The American Economic Review, 97(1), 242--259.

Erdem, T. and M.P. Keane, (1996), "Decision Making Under Uncertainty: Capturing Dynamic Brand Choice Processes in Turbulent Consumer Goods Markets", Marketing Science, 15, 1-20

Ericson, R. and A. Pakes, (1995), "Markov-Perfect Industry Dynamics: A Framework for Empirical Work", The Review of Economic Studies, 62, 53-82

Faust, K. and S. Wasserman, (1992), "Centrality and Prestige: A Review and Synthesis", Journal of Quantitative Anthropology, 4, 23-78

Feng, J., H.K. Bhargava and D.M. Pennock (2007), "Implementing Sponsored Search in Web Search Engines: Computational Evaluation of Alternative Mechanisms," INFORMS Journal on Computing, 19(1), 137-148.

Ghose, A. and S. Yang (2009), "An Empirical Analysis of Search Engine Advertising: Sponsored Search in Electronic Markets," Management Science, 55(10) 1605–1622

Godes, D. and Mayzlin, D., (2004), "Using Online Conversations to Study Word-of-Mouth Communications", Marketing Science, 23, 545-560

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

Goldfarb, A. and C. Tucker (2011), "Search Engine Advertising: Channel Substitution when Pricing Ads to Context," Management Science, forthcoming

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

Gouldner, A.W., (1960), “The Norm of Reciprocity: A Preliminary Statement”, American Sociological Review, 25, 161-178

Granka, L.A., T. Joachims and G. Gay (2004), “Eye-Tracking Analysis of User Behavior in WWW Search,” Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 25-29, 2004, Sheffield, United Kingdom.

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

Hartmann, W.R., (2010), “Demand Estimation with Social Interactions and the Implications for Targeted Marketing”, Marketing Science, 29(4)

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

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

Ho, T.H., N. Lim and C.F. Camerer (2006), “How ‘Psychological’ Should Economic and Marketing Models Be?” Journal of Marketing Research, 43(August), 341-344.

Holmes, T.J., (2011), “The Diffusion of Wal-Mart and Economies of Density”, Econometrica, 79(1), 253-302

Hoque, A.Y. and G.L. Lohse (1999), “An Information Search Cost Perspective for Designing Interfaces for Electronic Commerce,” Journal of Marketing Research, 36(August), 387-394.

Hotz, V.J. and R.A. Miller (1993), “Conditional Choice Probabilities and the Estimation of Dynamic Models”, The Review of Economic Studies, 60, 497-529

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

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

Jackson, M.O., (2004), “A Survey of Models of Network Formation: Stability and Efficiency,” in G. Demange and M. Wooders, eds., “Group Formation in Economics: Networks, Clubs, and Coalitions”, Cambridge University Press

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

Johnson, E.J., W.W. Moe, P.S. Fader, S. Bellman, and G.L. Lohse (2004), "On the Depth and Dynamics of World Wide Web Shopping Behavior," Management Science, 50(3), 299-308.

Kamakura, W.A. and G.J. Russell, (1989), "A Probabilistic Choice Model for Market Segmentation and Elasticity Structure", Journal of Marketing Research, 26, 379-390

Katona, Z. and M. Sarvary, (2008), "Network Formation and the Structure of the Commercial World Wide Web", Marketing Science, 27, 764-778

Katona, Z. and M. Sarvary (2010), "The Race for Sponsored Links: Bidding Patterns for Search Advertising," Marketing Science, 29(2), 199-215

Kitts, B. and B. Leblanc (2004), "Optimal Bidding on Keyword Auctions," Electronic Markets, 14(3), 186-201.

Liu, D., J. Chen, A. B. Whinston (2010), "Ex Ante Information and the Design of Keyword Auctions," Information Systems Research, 21(1) 133–153

Manchanda, P., P.E. Rossi and P.K. Chintagunta (2004), "Response Modeling with Non-Random Marketing Mix Variables," Journal of Marketing Research, 41(November), 467-478.

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

Maskin, E. and J. Tirole, (1988), "A Theory of Dynamic Oligopoly: I & II", Econometrica, 56, 549-600

Maskin, E. and J. Tirole, (2001), "Markov Perfect Equilibrium, I. Observable Actions", Journal of Economic Theory, 100, 191-219

Mayzlin, D. and H. Yoganarasimhan, (2008), "Link to Success: How Blogs Build an Audience by Promoting Rivals", Working Paper

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

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

Misra, S., E. Pinker and A. Rimm-Kaufman (2006), "An Empirical Study of Search Engine Advertising Effectiveness," Workshop on Information Systems and Economics (WISE), 2006.

Nair, H., P. Manchanda and T. Bhatia, (2010) “Asymmetric Peer Effects in Physician Prescription Behavior: The Role of Opinion Leaders”, Journal of Marketing Research, 47(5), 883-895

Pakes, A. and P. McGuire, (1994), “Computing Markov-Perfect Nash Equilibria: Numerical Implications of a Dynamic Differentiated Product Model”, The RAND Journal of Economics, 25, 555-589

Pakes, A. and P. McGuire, (2001), “Stochastic Algorithms, Symmetric Markov Perfect Equilibrium, and the 'Curse' of Dimensionality”, Econometrica, 69, 1261-1281

Pakes, A., M. Ostrovsky and S. Berry, (2007), “Simple Estimators for the Parameters of Discrete Dynamic Games (With Entry/Exit Examples)”, The RAND Journal of Economics, 38, 373-399

Riley, E., A. Peach, N. Scevak and Z.D. Wigder (2007), “US Online Advertising Forecast, 2007 to 2012,” Jupiter Research Corporation report, June 14th, 2007.

Rust, J., (1987), “Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher”, Econometrica, 55, 999-1033

Rutz, O., M. Trusov, R. E. Bucklin (2011), “Modeling Indirect Effects of Paid Search Advertising: Which Keywords Lead to More Future Visits?” Marketing Science, forthcoming

Rutz, O. and R. E. Bucklin, (2011), “From Generic to Branded: A Model of Spillover Dynamics in Paid Search Advertising,” Journal of Marketing Research, 48(1), 87–102.

Ryan, S.P., (2009), “The Costs of Environmental Regulation in a Concentrated Industry”, Working Paper

Shin, W. (2009), “Being with the Big Boys: When to Buy Competitor’s Keyword?” Working Paper.

Simon, H. (1955), “A Behavioral Model of Rational Choice,” Quarterly Journal of Economics, 69(1), 99-118.

Srinivasan, K., (2006), “Invited Commentary: Empirical Analysis of Theory-Based Models in Marketing”, Marketing Science, 25, 635-637

Stephen, A. and O. Toubia, (2010), “Deriving Value from Social Commerce Networks”, The Journal of Marketing Research, 47(2), 215-228

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

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

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

Varian, H.R. (2007), “Position Auctions,” International Journal of Industrial Organization, 25(6), 1163-1178.

Wasserman, S. and K. Faust, (1994), “Social Network Analysis: Methods and Applications”, Cambridge University Press

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

Weber, T.A. and Z. Zheng (2007), “A Model of Search Intermediaries and Paid Referrals,” Information Systems Research, 18(4), 414-436.

Weintraub, G.Y., C.L. Benkard and B. Van Roy, (2008), “Markov Perfect Industry Dynamics with Many Firms”, Econometrica, 76, 1375-1411

Yao, S. and C.F. Mela (2011), “A Dynamic Model of Sponsored Search Advertising,” *Marketing Science, forthcoming*

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

Zhu, Y. and K.C. Wilbur (2011), “Hybrid Advertising Auctions,” Marketing Science, forthcoming

Technical Appendix 1: for Chapter 2

TA1.1 Analysis for Pay-Per-Click Auction

In the pay-per-click auction, the revenues for each firm stay the same but the payments change because they depend on the actual clicks (and not just the impressions). We use the revenue expressions from the pay-per-impression section and only discuss the payments based on clicks.

Scenario I: $s \leq p\min\{Q,V\}$

The analysis of this low search cost scenario is the same as in the pay-per-impression case because payments remain the same.

Scenario II: $Q > V$ and $pV < s \leq pQ$

In this scenario, when the inferior firm is placed on top, the probability that it is clicked is $1-\phi$ (the portion of uninformed consumers) and when it is placed on bottom, no consumer clicks on it. Therefore, $E[\Pi_{S,1}] = pm_S - b_S$ and $E[\Pi_{I,2}] = 0$, and $E[\Pi_{S,2}] = pm_S - \underline{b}$ and $E[\Pi_{I,1}] = (1-\phi)(1-p)\,pm_I - (1-\phi)b_I$. The equilibrium bids are $b_S^* = \underline{b}$ and $b_I^* = (1-p)\,pm_I$. The insight is that since the inferior firm pays only when its link is clicked, it can bid the expected revenue conditional on the link being clicked, instead of the unconditional expected revenue as in the pay-per-impression case. Hence, we again have a case in which the inferior firm emerges on top, but obtains fewer clicks.

Scenario III: $Q < V$ and $pQ < s \leq pV$

Please see Section 3.2 of in the main text.

Scenario IV: $s > p\max\{Q,V\}$ and $s \leq p(V+Q)$

As explained in Section 3.1, if the superior firm is placed on top, the probability that it will be clicked is 1, while if it is placed at the bottom, this probability is $\phi + (1-\phi)(1-p)$. If the inferior firm is placed on top, its probability of being clicked is $1-\phi$, while if it is placed at the bottom, this probability is zero.

If the superior firm pays b_S per click and is placed on top (so the inferior firm pays $\underline{b}$ per click), then the expected profits of the firms are $E[\Pi_{S,1}] = pm_S - b_S$ and $E[\Pi_{I,2}] = 0$. If the inferior firm pays b_I per click and is placed on top (so the superior firm pays $\underline{b}$ per click), then the expected profits of the firms are $E[\Pi_{S,2}] = (\phi + (1-\phi)(1-p))(pm_S - \underline{b})$ and $E[\Pi_{I,1}] = (1-\phi)(pm_I - b_I)$. The equilibrium bids can be obtained as follows:

$$pm_S - b_S^* = (\phi + (1-\phi)(1-p))(pm_S - \underline{b}) \Rightarrow b_S^* = (1-\phi)p^2 m_S + (1-(1-\phi)p)\underline{b}$$
$$\text{and } (1-\phi)(pm_I - b_I) = 0 \Rightarrow b_I^* = pm_I.$$

If $(1-\phi)p^2 m_S + (1-(1-\phi)p)\underline{b} > pm_I$, then the superior firm wins the auction, otherwise the inferior firm wins the auction. If the inferior firm wins, it will obtain more clicks than the superior firm only if $\phi < p/(1+p)$, otherwise the superior firm will obtain more clicks even though it is at the bottom position.

Scenario V: $s > p(V+Q)$

In this scenario, if the superior firm is placed on top, the probability that it will be clicked is 1. If it is placed at the bottom, this probability is ϕ. If the inferior firm is placed on top, the probability that it will be clicked is $1-\phi$, while if it is at the bottom, this probability is zero. Therefore, $E[\Pi_{S,1}] = pm_S - b_S$ and $E[\Pi_{I,2}] = 0$, and $E[\Pi_{S,2}] = \phi(pm_S - \underline{b})$ and $E[\Pi_{I,1}] = (1-\phi)(pm_I - b_I)$. The equilibrium bids are

$b_S^* = (1-\phi)pm_S + \phi\underline{b}$ and $b_I^* = pm_I$. Either firm may end up at the top position and as ϕ increases, the inferior firm bids higher and will win the auction. Furthermore, if the inferior firm wins, then it will still obtain fewer clicks if $\phi \geq 1/2$.

An assumption made in the above analysis is that in all equilibria, both firms' bids as specified in the equations are larger than or equal to the minimum bid ($\underline{b}$). The assumption will hold as long as both firms' margins are considerably larger than the minimum bid, and when the match probability is not too close to 0 or 1.

TA1.2 Details of Analysis with Heterogeneous Search Costs

Pay-Per-Impression Auction

In equilibrium, each firm simply bids the incremental value of the top position relative to the bottom one, plus the minimum required bid. The incremental value of the top position is the weighted average of that value for each segment of the consumers.

Consider first the case where $V \geq Q$. With probability pQ , the customer will have a search cost between 0 and pQ . For these customers (Scenario I), the incremental value of the top position is 0 to both the superior and the inferior firm. With probability $p(V-Q)$, the customer will have a search cost between pQ and pV (Scenario III). For these customers, the incremental value of the top position is $(1-\phi)p^2m_S$ for the superior firm and $(1-\phi)p^2m_I$ for the inferior firm. With probability pQ , the customer will have a search cost between pV and $p(V+Q)$ (Scenario IV). For these customers, the incremental value of the top position is $(1-\phi)p^2m_S$ for the superior firm and $(1-\phi)pm_I$ for the inferior firm. Finally, with probability $1-p(V+Q)$ the customer has a search cost between $p(V+Q)$ and 1 (Scenario V). For these

customers, the incremental value of the top position is $(1-\phi)pm_S$ for the superior firm and $(1-\phi)pm_I$ for the inferior firm. The expected incremental value of the top position for the superior firm is given by

$$0\times pQ+(1-\phi)p^2m_S\times p(V-Q)+(1-\phi)p^2m_S\times pQ+(1-\phi)pm_S\times(1-p(V+Q))$$
$$=(1-\phi)pm_S\times(1-p(1-p)V-pQ).$$

While the expected incremental value of the top position for the inferior firm is given by:

$$0\times pQ+(1-\phi)p^2m_I\times p(V-Q)+(1-\phi)pm_I\times pQ+(1-\phi)pm_I\times(1-p(V+Q))$$
$$=(1-\phi)pm_I\times(1-p(1-p)V-p^2Q).$$

In equilibrium, both firms will bid the respective incremental value plus the minimum bid. The case where $V<Q$ is analyzed similarly and the result is the same as above. This gives us the following proposition.

Pay-Per-Click Auction

Using a similar approach as in the pay-per-impression case, we can derive the optimal bids for both firms. Consider first the case where $V\geq Q$. With probability pQ, the customer will have a search cost between 0 and pQ. For these customers (Scenario I), the incremental revenue of the top position is 0 to both the superior and the inferior firm. The incremental cost of the top position over the bottom position for the superior firm is $b_1-\underline{b}$, where b_1 is the amount the superior firm pays per click at top position, and that for the inferior firm is $(\phi(1-p)+(1-\phi))b_2-(1-p)\underline{b}$, where b_2 is the amount the inferior firm pays per click at top position. With probability $p(V-Q)$, the customer will have a search cost between pQ and pV (Scenario III). For these customers, the incremental revenue of the top position is $(1-\phi)p^2m_S$ for

the superior firm and $(1-\phi)p^2 m_I$ for the inferior firm. The corresponding incremental cost of the top position is $b_1 - (\phi + (1-p)(1-\phi))\underline{b}$ for the superior firm and $(\phi(1-p) + (1-\phi))b_2 - (1-p)\underline{b}$ for the inferior firm. With probability pQ, the customer will have a search cost between pV and $p(V+Q)$ (Scenario IV). For these customers, the incremental revenue of the top position is $(1-\phi)p^2 m_S$ for the superior firm and $(1-\phi)pm_I$ for the inferior firm. The corresponding incremental cost of the top position is $b_1 - (\phi + (1-p)(1-\phi))\underline{b}$ for the superior firm and $(1-\phi)b_2$ for the inferior firm. Finally, with probability $1-p(V+Q)$ the customer has a search cost between $p(V+Q)$ and 1 (Scenario V). For these customers, the incremental revenue of the top position is $(1-\phi)pm_S$ for the superior firm and $(1-\phi)pm_I$ for the inferior firm. The incremental cost is $b_1 - \phi\underline{b}$ for the superior firm and $(1-\phi)b_2$ for the inferior firm.

Combining the above scenarios, $b_1 = p(1-p(1-p)V - pQ)(1-\phi)m_S + (\phi + p(1-p)(1-\phi)V + p(1-\phi)Q)\underline{b}$ will make the superior firm indifferent between the top and bottom position, while $b_2 = \frac{p(1-p(1-p)V - p^2Q)(1-\phi)m_I + p(1-p)V\underline{b}}{1-\phi+\phi p(1-p)V}$ will make the inferior firm indifferent. These bids are thus the equilibrium bids.

The case where $V < Q$ is analyzed similarly. The equilibrium bid for the superior firm is the same as in the case where $V \geq Q$, while that for the inferior firm is $b_2 = \frac{p(1-p(1-p)V - p^2Q)(1-\phi)m_I + p(1-p)V\underline{b}}{(1-\phi)(1+p^2V - p^2Q)+\phi p(1-p)V}$.

Technical Appendix 2: for Chapter 3

MCMC Estimation Algorithm

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

A.1 Inter-purchase timing

A.1.1 draw λ_g :

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

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

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

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

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

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

A.1.3 draw γ_g:

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

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

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

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

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

A.1.5 draw σ_λ^2:

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

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

A.1.6 draw σ_γ^2:

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

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

A.1.7 draw r_λ:

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

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

A.1.8 draw r_γ:

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

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

A.1.9 draw κ_g^{pi}:

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

In (A9), $E_g(\tilde{\tilde{E}}_g, \kappa_g^{pi})$ represents function to calculate smoothed exposure using the raw exposure and the smoothing parameter κ_g^{pi}. This step is repeated for each group: $g = 1..G$. We use Metropolis with random walk. The random walk step is an independent draw from $MVN(\vec{0}, 0.2I)$.

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

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

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

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

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

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

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

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

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

A.2 Product-choice

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

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

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

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

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

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

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

A.2.3 draw ρ_g :

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

In (B3), $\phi(.)$ is the density of normal distribution. P_g (E_g) represents the purchases (exposures) of all persons in the group across all time periods. This step is repeated for each group, $g = 1..G$.

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

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

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

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

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

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

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

A.2.6 draw σ_ρ^2:

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

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

A.2.7 draw r_{β_j}:

$$f(r_{\beta_j} \mid \beta_{g,j} : g = 1..G, \bar{\beta}_j, \sigma_{\beta_j}^2) \propto \prod_{g=1}^{G} \phi(\begin{pmatrix}\beta_{g,1,j}\\ ...\\ \beta_{g,I,j}\end{pmatrix} \mid \begin{pmatrix}\bar{\beta}_j\\ ...\\ \bar{\beta}_j\end{pmatrix}, \sigma_{\beta_j}^2 \begin{bmatrix}1 & r_{\beta_j} & r_{\beta_j}\\ r_{\beta_j} & ... & r_{\beta_j}\\ r_{\beta_j} & r_{\beta_j} & 1\end{bmatrix}) \quad \text{(B7)}$$

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

A.2.8 draw r_ρ:

$$f(r_\rho \mid \rho_g : g = 1..G, \bar{\rho}, \sigma_\rho^2) \propto \prod_{g=1}^{G} \phi(\begin{pmatrix}\rho_{g,1}\\ ...\\ \rho_{g,I}\end{pmatrix} \mid \begin{pmatrix}\bar{\rho}\\ ...\\ \bar{\rho}\end{pmatrix}, \sigma_\rho^2 \begin{bmatrix}1 & r_\rho & r_\rho\\ r_\rho & ... & r_\rho\\ r_\rho & r_\rho & 1\end{bmatrix}) \quad \text{(B8)}$$

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

A.2.9 draw κ_g^{pc}:

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

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

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

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

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

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

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

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

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

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

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

Technical Appendix 3: for Chapter 4

PageRank

In this section we explain the detail of PageRank, the measure of network position used in our study. PageRank, first presented in Brin and Page (1998), is behind the initial Google search engine. Given a network of pages, it produces a numerical measure for each page to represent its relative importance in the network. The measure is well documented in literature, and is explained here for completeness:

Let $p_1, p_2, ... p_n$ be n nodes (web pages) which are connected by directional links. Let c_i be the out-degree of page p_i – the number of outgoing links from that page. Let d be a damping factor which value between 0 and 1. Denote A as the modified adjacency matrix for the graph of the nodes and the links, where:

(A-1) $$[A]_{ij} = \begin{cases} 1/c_i & i \to j \\ 0 & otherwise \end{cases}$$

The PageRank, denoted as PR, is a vector such that:

(A-2) $$PR = (1-d) \cdot PR + d \cdot PR \cdot A$$

The i-th element of PR is the PageRank of node p_i. A larger value indicates higher importance of the node in the network. The measure is rooted on a Markov random navigation model: assume there is a person visiting the pages; at any time, with probability d she chooses to follow an outgoing link, with link chosen randomly with equal probability when multiple outgoing links exist, and with probability $1-d$ she will jump to another page, with each page

having the same probability to be the destination. The PageRank of a page is then the steady-state probability of that page being visited.

As stated in Brin and Page (1998), "*Another intuitive justification is that a page can have a high PageRank if there are many pages that point to it, or if there are some pages that point to it and have a high PageRank.*" This insight proves crucial for the success of PageRank in capturing the relative importance of web pages on the Internet, and is instrumental in our study. The PageRank is also similar to eigenvector centrality that is widely used in social network literature, where it is shown to reflect the power or prestige of a node in the network. (Let I be an $n \times n$ matrix where $[\mathrm{I}]_{ij} = 1/n$, then A-2 can be written as $PR = PR((1-d) \cdot \mathrm{I} + d \cdot A)$, i.e. the PageRank is an eigenvector of the adjacency matrix further modified by the damping factor.)