

Modeling Online Consumer Product Search*

Jun B. Kim[†]

Paulo Albuquerque[‡]

Bart J. Bronnenberg[§]

February 4, 2008

*Discussions with Andrew Ainslie, Anand Bodapati, Randy Bucklin, Minha Hwang, and Raphael Thomadsen have been very helpful and constructive throughout the research.

[†]UCLA Anderson Graduate School of Management, Phone: +1-310-825-7873, Fax: 1-310-206-7422, jun.kim.2009@anderson.ucla.edu

[‡]Simon Graduate School of Business, University of Rochester, Phone: 1-585-273-3253, Fax: 1-585 -273-1140, paulo.albuquerque@simon.rochester.edu

[§]Tilburg University, Phone: +31 13 466 8939, Fax: +31 13 466 8354, bart.bronnenberg@uvt.nl

Abstract

Using publicly available aggregate data at an online retailer, we study how consumers search for durable goods. Specifically, we investigate the search patterns when, as is common with durable products, the total set of options consumers can choose from is very large. To study search patterns, we propose to analyze our aggregate data by defining a product network, consisting of the products and links that designate if a product is searched conditional on searching other products. We model these links using a flexible linear mixed model for binary data. Our model combines fixed effects of brands, product attributes, and prices on product search patterns with random effects capturing dependencies among the searched products. The latter can be displayed and interpreted as a perceptual map of “search proximity” or as a “product search map.” The proposed Bayesian estimation approach helps a practitioner understand which products are searched and compared in the same online session, and importantly analyzes how observed information acquisition is organized in brand, attribute, and price related search strategies. Using a data set covering consumer search of digital camcorders at Amazon.com, we infer that the product search is strongly attribute based.

Keywords: product links, link dependency, information search, product search, network analysis

1 Introduction

The goal of this paper is to understand the patterns of product search for consumer durables in categories for which the number of choice options is very large. More specifically, we seek to analyze what drives the propensity to conduct a joint search of any two products. An understanding of product search that takes place prior to purchase is important for the following reasons. First, product search patterns reveal limits on consumer consideration sets. Second, pre-purchase product search reflects consumers' information acquisition, which partly explains substitution patterns and future choices. Third, knowledge of consumer information acquisition is fundamental to planning marketing communications and retail distribution (Newman and Staelin 1972).

In many product categories, the Internet has successfully established itself as an alternative sales channel to the traditional off-line channels. More importantly, it has become a significant destination to consumers looking to gather information and conduct product search prior to making purchase decisions (Peterson et al. 1997; Mendelsohn et al. 2007). According to an industry survey conducted in 2003 in the digital camera category in North America¹, more than 50% of online and offline consumers conducted online product research as a part of their pre-purchase activities. More than 50% of all online consumers conducted product research at Amazon.com, and 12% of them made purchases at Amazon.com. In our empirical study, we use publicly available data from Amazon.com, which posts various product-related data that summarize consumer online behavior. Given the expected growth in online product and information search in the foreseeable future, better understanding of such consumer behavior is important for both managers and academics.

In order to understand patterns of online product search in a consumer electronics category, we model the product search as a set of conditional search probabilities between two products. We subsequently make these conditional probabilities a function of product characteristics, such as brands (e.g., Sony), technical attributes (e.g., DVD), and price tiers. This allows us to estimate the effects of product characteristics on the likelihood that two products are jointly searched for, during one online session. While we do not explicitly model individual consumer's search decisions in this paper, describing and understanding the patterns of product search is an important first step towards a more

¹The survey was prepared by a leading market research firm for a major technology manufacturer.

theory-based model.

With this paper, we aim to make the following contributions to the marketing literature. First, in a realistic durable goods context, the choice set is often very large. Knowing how a consumer restricts her choice options is of practical and theoretical interest. We offer a feasible approach to studying product search patterns for large sets of products without incurring the curse of dimensionality.

Second, our paper offers an analysis of what factors drive the *contents* of search for consumer durables. The data set in this paper contains direct observation about products searched by consumers and relations between these products. Our analysis improves understanding of how information acquisition takes place and how product search sets are formed.

Third, a unique aspect of this paper is that we study product search patterns using publicly available data at Amazon.com. Amazon.com aggregates consumer browsing histories and provides summary information of these browsing histories, in part as an aid for future customers. We directly observe which products are jointly searched across a very large customer base. As these data are publicly available for nearly every consumer durable good sold through Amazon.com, our analysis has a broad application in the demand analysis for durables. While detailed syndicated data are available for non-durable products, there is a relative dearth of good data for consumer durable products. Thus, a method for analyzing these data is broadly useful.

Fourth, an important empirical detail of product search, and one which we expect to be general to many other browsing contexts, is that product search is subject to strong degrees of transitivity (under the searches of $[j|i]^2$ and $[k|j]$, the search of $[k|i]$ is observed with high likelihood) and reciprocity (under a search of $[j|i]$, a search $[i|j]$ is observed with high likelihood). We conduct a series of simulation studies to quantify the extent of search dependencies that exists in our data set. A contribution of our model is that it allows for these and other complex cross-product dependencies in consumer product and information search.

Lastly, and perhaps most importantly, we shed some light on substitution patterns for durable goods. In the past, brand substitution in consumer packaged goods has been inferred mostly using individual choice data (e.g., Elrod 1988; Allenby 1989). However, in durable goods, researchers

²The notation $[j|i]$ reads as “product j is searched conditional on searching product i ”.

observe only one purchase occasion for each individual, making substitution patterns harder to study. Since the pre-purchase viewing of multiple products includes search of substitutable products prior to choice, our data and analysis provide a more informative account of substitution patterns than that obtained from choice data.

Our approach to describe the search data yields a visualization that we refer to as “product search maps.” In these maps, products that are likely to be searched together in the same session are located close to each other, whereas products that are unlikely to be searched together are placed at distant positions. In these maps, local subsets of products can be interpreted as stereotypical product sets that are searched together, and presumably, compete more closely. The product search maps are an efficient way to organize the massive number of possible consideration sets.

Closer inspection of the content of local subsets reveals the cross-product structure of search. For instance, from an analysis of our product search maps, consumers seem to organize their search for a camcorder primarily and most importantly by media format (DVD, Hard Disk and Mini-DV), and subsequently by brand and price. Therefore, consumers are more likely to search multiple products that share the media format and less likely to search across different media formats. With respect to price search, we also find a stronger product search intensity within the same price tiers than across different tiers. We additionally find that cross-tier product search takes place predominantly between adjacent tiers and decreases across distant tiers.

This paper is organized as follows. Section 2 discusses relevant literature. Section 3 reports and discusses the data available at Amazon.com. The model, its operationalization and estimation details are explained in section 4. In section 5, we discuss the results of the proposed model. Section 6 discusses managerial implications of the findings and concludes with directions for future research.

2 Background

The marketing literature has extensively studied the process of consumer information search. Most research on consumer information acquisition in marketing has been behavioral in nature (Bettman and Kakkar 1977; Jacoby et al. 1978; Moorthy et al. 1997). From this literature, three dimen-

sions of consumer information acquisition have emerged - depth (quantity), sequence (order), and content (Jacoby et al. 1977). The main focus has been to characterize the rules and strategies by which consumers find and organize information, using information display boards (IDB) (Payne 1976; Bettman and Jacoby 1976) and eye movement measurements (Russo and Rosen 1975; Wedel and Pieters 2000). Other research efforts in this stream have looked at the scope and nature of external information search. Newman and Staelin (1972) studied the use of several distinct information sources - personal, neutral, advertising, and retail outlet - in consumer decision making for durable goods, while Westbrook and Fornell (1979) identified different groups of consumers based on information sources they used prior to durable goods purchase. Bettman and Kakkar (1977) reported that presentation format affected consumer information acquisition strategy. In a more recent study, Moorthy et al. (1997) found that prior brand experience affects the information search behavior. Concerning online behavior, Johnson et al. (2004) reported that the amount of shoppers' search activities across online stores is quite limited in commodity-type products such as books and CD's. Their findings are consistent with the previous finding in Beatty and Smith (1987) that consumers visited two or three retail stores for their TV purchase decisions.

The economics literature has studied consumer information search using a cost-benefit analysis framework to determine optimal sampling size (Stigler 1961) and optimal stopping rule (Nelson 1970). Further, several studies have found that the impact of consumer search on market performance can be profound (Nelson 1970; Anderson and Renault 1999). For instance, Nelson (1970) theorized that limited consumer knowledge on product quality due to limited consumer search can have a large effect on the market structure of consumer goods. Anderson and Renault (1999) study price competition under product differentiation and search cost. They argue that, in the presence of consumer search, equilibrium price should first fall and then rise as the balance shifts from price competition to taste for product diversity. Therefore, studying consumer information acquisition, as we do in this paper, is important in ultimately understanding market structure and market performance.

In marketing, the consideration set literature has followed the footsteps of economic theory of information search (Hauser and Wernerfelt 1990) since the concept of consideration is a logical outcome of information search (Roberts and Lattin 1991). Also, in consumer durable goods context, the

consideration set is created during product and information search process (Urban, Hulland, and Weinberg 1993; Roberts and Lattin 1997). In many two-stage choice models, consideration is treated as a latent construct that is inferred from choice data (Siddarth et al. 1995; Bronnenberg and Vanhonacker 1996). One exception is Moe (2006), who observes each shopper's viewed products from click stream data at an online store. She develops an individual-level, two stage choice model, treating viewed products as proxy for considered products.

We further point out recent research using online store data. These studies include online price competition (Chevalier and Goolsbee 2003), consumer welfare gain from the vast selection at online store (Brynjolfsson et al. 2003), and consumer welfare implications from a merger of two wireless carriers (Bajari et al. 2007). Common to all these studies is the use of Amazon.com's sales rank data to infer demand. One caveat of Amazon.com's sales rank data is that they are relatively volatile even for a short period of time. In our empirical study, we do not depend on sales rank data. Instead, we use binary, product search data that are stable over time.

Our paper differs from the previous research on information search and consideration set formation in two important aspects. First, while most behavioral marketing research is based on a survey-based methods, we use actual product search data from a very large customer base. Use of observation data in this research is critical. For instance, Newman and Lockeman (1975) found that survey-based methods understated consumer search activities by a wide margin.

Second, this paper aims to get insights on the *content* of consumer information search. Of the three dimensions in consumer information search (depth, sequence, and content), depth has received disproportionate amount of attention in the past (Newman and Lockeman 1975; Newman 1977; Beatty and Smith 1987; Johnson et al. 2004). In contrast, the content of information search has been studied less (Roberts and Lattin 1997). Given that a consumer's information search and the subsequent formation of consideration set affect her final choice, understanding search content is important. This is especially true in categories where the number of choice options is large and awareness and consideration are important factors (Roberts and Lattin 1991; Urban, Hulland, and Weinberg 1993).

3 Data

Amazon.com posts aggregate-level, product-related information that results from consumers' pre-purchase browsing activities at its online store. Presented as "*Customers who viewed this item . . . also viewed these items . . .*," this feature shows a consolidated list of products in the same product category that were viewed by past shoppers for an item currently being viewed. For instance, if a large percentage of consumers who viewed product A also viewed product B , B appears on the viewed product list of A . We will call this data set *pre-purchase search data*. An important aspect of this data set is that it reveals the relationships among products in the same category. According to Linden et al. (2005), the idea behind pre-purchase search data is as follows:

... relationships between products within an online catalog are determined by identifying products that are frequently viewed by users in the same browsing session (e.g. products A and B are related because a significant portion of those who viewed A also viewed B ...)

The pre-purchase search data are an outcome of an item-to-item, collaborative filtering mechanism, where the relationship between two products is determined by how frequently these two products are viewed by users within the same browsing session.³

The *pre-purchase search data* constitute a collection of *directional relations* or *links* that exist *between* products. These links span up an associative network of choice alternatives, in which a node represents a product, and an edge a link between two products. Further, we can represent this network by an $N \times N$ product search matrix Y , in which an element, y_{ij} , represents the presence or absence of a link from product i to product j and where $y_{ij} = 1$ if conditional on consumers' search of product i , product j is one of the top products also searched in that same online session by consumers, and zero otherwise. Thus, the relational information in this data set is binary.

The relationships in this data set are asymmetric. For instance, product A may be on the search list of product B , but B may not be on the search list of A . This implies that the sets of jointly searched products for A and B are different, due to different search patterns.

³Further details on Amazon.com's data preparation are found in the appendix.

In this study, we exploit the product search matrix information to learn which and to what extent inter-product characteristics influence product search. The product search matrix is a direct result of consumer information and product search decisions at the online store. Amazon.com provides a wide variety of ways for consumers to access product pages. For instance, consumers can conduct direct keyword search, use category filters on product attributes such as brands and price, or follow links to other products. The actual sets of products searched by consumers leads to the browsing history which Amazon.com collects and summarizes in the *pre-purchase search data*.

Amazon.com could use the *pre-purchase search data* in such a way to achieve its business goals, e.g., direct consumers to high margin products or clearance items. This, however, is not likely for the following reasons. First, provision of truly similar products is strongly aligned with Amazon.com's commercial interest. By offering more relevant selections at lower search cost, Amazon.com can help consumers choose products that best fit their needs, enhance consumer shopping experience, and reduce price sensitivity (Lynch and Ariely 2000). Amazon.com's heavy investment in personalization and recommendation technologies reflects such interests. Second, the pre-purchase product search data are quite stable over time and do not show a sudden inclusion or radical movement of products at the top of the list, which is otherwise highly likely if Amazon.com manipulates its list. Lastly, we have verified, through private communication, that product summary data are free from Amazon.com's strategic behavior. A former key manager at Amazon.com who was deeply involved with this data testified that it represents consumer browsing behavior and that Amazon.com puts consumer trust and long-term relationship over short-term gains.⁴

We have collected daily data in the digital camcorder category starting from mid June, 2006 and ending November, 2006. For the empirical analysis, we use an accumulation of daily product search and characteristics data for the month of August 2006. Under the accumulation process, a product A will have a relationship with a product B as long as A appears at least once on the daily search list of B during the month. Aggregating the data this way allows us to smooth out the product search list and include all available products for a more comprehensive analysis. In the digital camcorder category, Amazon.com offers more than 250 different models from leading manufacturers to a large

⁴The same point has been also stressed multiple times in the past by many senior Amazon.com executives in the press.

Table 1: Breakdown of products based on brands and media formats

Formats	Sony	Canon	JVC	Panasonic	Total
DVD	14	4	0	5	23
Hard Drive	4	0	9	0	13
MiniDV	20	17	11	10	58
Total	38	21	20	15	94

customer base. Although our approach is scalable to the full set of products, for practical illustration we narrow down the number of products used in the empirical study using the following criteria. First, we use products from top four manufacturers - Sony, Canon, JVC, and Panasonic. These four brands offer about 80% of digital camcorders available at Amazon.com. Second, we exclude discontinued products from manufacturers. Third, we limit storage media formats to top three popular formats - DVD, Hard-Drive, and Mini-DV. Flash memory is the next popular storage format but is often adopted by low price products from small manufacturers. Fourth, we exclude high end, professional grade digital camcorders since they are judged to form a separate, own sub-market. Applying this set of criteria narrows down the total number of products in the empirical analysis to 94. Table 1 shows the breakdown of these products by brand and storage media format.

The total number of links present in the product search network is 1162, which is 13 % of all possible entries in the matrix Y excluding the diagonal elements, among 94 products. The number of links “sent” by each product ranges from 3 to 20, with a mean of 12.4 and standard deviation of 3.54. The number of links “received” by each product ranges between 0 and 42 with a standard deviation of 11.8. The disparity between numbers of sent and received links per product implies that there is a group of products that receive far larger number of links. A more detailed breakdown of links among the brands, media formats, and prices is found in Tables 2, 3, and 4. The breakdowns in these tables serve to make clear how the links are distributed among product characteristics.

Our pre-purchase search data have at least two types of link dependency for which the modeling framework should account: reciprocity and transitivity in directed relations. Wasserman and Faust (1994) define reciprocity as the presence or absence of links y_{ij} and y_{ji} being positively correlated. Our data exhibit a very high level of reciprocity: the number of reciprocal dyads, $y_{ij} = y_{ji} = 1$, is 305. To illustrate just how large this number is, we generated random realizations of the network, keeping the

Table 2: Breakdown of links between brands

Brand (From)	Brand (To)	Count	Percentage (%)
Sony	Sony	256	56
	Canon	89	20
	JVC	42	9
	Panasonic	70	15
Sub total		457	100
Canon	Sony	61	21
	Canon	158	54
	JVC	20	7
	Panasonic	53	18
Sub total		292	100
JVC	Sony	52	24
	Canon	35	16
	JVC	95	44
	Panasonic	34	16
Sub total		216	100
Panasonic	Sony	54	27
	Canon	46	23
	JVC	13	7
	Panasonic	84	43
Sub total		197	100

Table 3: Breakdown of links between media formats

Format(From)	Format(To)	Count	Percentage(%)
MINIDV	MINIDV	612	90
	HD	15	2
	DVD	52	8
Subtotal		679	100
HD	MINIDV	20	12
	HD	125	79
	DVD	14	9
Subtotal		159	100
DVD	MINIDV	53	16
	HD	10	3
	DVD	261	81
Subtotal		324	100

Table 4: Breakdown of links between price tiers

Tier(From)	Tier(To)	Count	Percentage(%)
P_1^5	P_1	93	43
	P_2	102	48
	P_3	19	9
Subtotal		214	100
P_2	P_1	86	20
	P_2	223	50
	P_3	133	30
Subtotal		442	100
P_3	P_1	23	5
	P_2	152	30
	P_3	331	65
Subtotal		506	100

number of “sent” links constant. Among 100,000 replications, the largest number of “coincidental” reciprocal dyads is 116, far smaller than the number of observed reciprocal dyads.

Another common measure for dependency in network data is transitivity, in which y_{ij} (a link from product i to j) and y_{jk} implies a high probability of y_{ik} . While the number of observed transitive triples in the product network is 8573, the largest number of simulated “coincidental” transitive triples in the permutation study was 2073, far smaller than the number observed in the product network. Both permutation tests indicate that our product search network shows very high levels of reciprocity and transitivity. In the following section, we discuss a model that accounts for such dependencies.

4 Model

4.1 Model structure

To analyze the structure of the product search network, we use the generalized linear mixed model (GLMM) (McCulloch and Searle 2001), which extends the Generalized Linear Model (GLM) by including random effects in the model specification.⁶

⁶We note that Multidimensional scaling (MDS) can also account for the link dependencies that our approach deals with. A traditional MDS takes a set of pair-wise product distances as input and estimates product positions in a multi dimensional space. However, a problem with the application of MDS, which we avoid here, is that the definition of distance measures often poses a serious challenge in MDS applications (Nakanishi and Cooper 2003). We have tried two different approaches to define a distance from product search network: (1) by using each entry in the product search

Since pre-purchase search data are encoded as a matrix with binary entries, we select a logit link function to model the product network data. In this model, the links, $Y = \{y_{ij}\}$, are modeled as *conditionally* independent given variables X , model parameters α and β , and random effects γ . The likelihood of the data can be expressed as

$$P(Y|X, \alpha, \beta, \gamma) = \prod_{i \neq j} \Pr(y_{ij}|x_{ij}, \alpha, \beta, \gamma_{ij}), \text{ where } i, j = 1, \dots, N \quad (1)$$

where N is the total number of products. Covariates of Y are represented by X , while α and β are model parameters. Lastly, γ_{ij} quantifies any unobserved random effect of a link between two products i and j . Sources of random effects may include online-store features such as Amazon.com’s recommendations, special promotions, and various offerings that affect search. They would also include any higher order interaction effects of covariates which are not captured in the linear model. We exclude y_{ii} from this formulation and note that $\Pr(y_{ii}|\cdot) = 1$.

One concise way to include random effects is by using distance between products in a latent space (Elrod 1988; Hoff, Raftery, and Handcock 2002) as a measure of link dependency. In this approach, the closer the products are positioned in a latent space, the more likely they are related to each other. More specifically, we represent the dependency of a link between products i and j as a function of their positions, z_i and z_j , in a P -dimensional latent space,

$$\gamma_{ij} = f(z_i, z_j),$$

where $z_i, z_j \in \mathbb{R}^P$. In the product search network context, the latent space refers to a space of unobserved product characteristics that can lead to potential symmetry and transitivity among the links in the product network.

Using latent position-based random effects, the probability of observing a network Y is

$$P(Y|X, \alpha, \beta, Z) = \prod_{\forall i, j \ i \neq j} \Pr(y_{ij}|x_{ij}, \alpha, \beta, z_i, z_j). \quad (2)$$

network as a similarity measure between two products, (2) by computing a “shortest path” as the minimum number of transitions through existing links needed to reach one product from the other (Sarkar and Moore 2005). Both distance measures produced nonsensical results.

Using the logit link function for $\Pr(y_{ij}|x_{ij}, \alpha, \beta, z_i, z_j)$, we can express it as

$$\Pr(y_{ij} = 1|x_{ij}, \alpha, \beta, z_i, z_j) = \frac{1}{1 + \exp(f(z_i, z_j) - \alpha - \beta' x_{ij})}. \quad (3)$$

We further parameterize the between-products random effects in the link probability using Euclidean distance between product positions in a latent space

$$f(z_i, z_j) \equiv \| z_i - z_j \| \quad (4)$$

In this formulation, the coefficient of inter-product distance is normalized to 1 without loss of generality since it is not separable from the scale of distance.

4.2 Operationalization

We choose a two-dimensional latent space for ease of interpretation and visualization. This choice is based on earlier findings that the two dimensions already provide a high level of flexibility and that the marginal model fit improvement with additional dimensions is quite limited (Hoff et. al 2002).

To explain consumer search behavior, we propose three categories of *asymmetric* variables that capture inter-product characteristics: (1) brands, (2) media formats, and (3) prices. This set of variables is chosen for two reasons. First, a popular consumer magazine (Consumer Reports 2006) suggests that price and media format are the two most important aspects to consider in the digital camcorder category. Second, Amazon.com provides category filters based on these attributes that consumers can use in their search. The variables are defined as:

- $I_{(B_n|B_m)}^{brand}(j|i) = 1$, if product i 's brand is B_m and j 's brand is B_n , otherwise it is 0,
- $I_{(F_q|F_p)}^{format}(j|i) = 1$, if product i 's media format is F_p and j 's media format is F_q , otherwise it is 0,
- $I_{(P_y|P_x)}^{price}(j|i) = 1$, if product i 's price is in tier P_x and j 's price is in tier P_y , otherwise it is 0,

where i and j are the sender and the receiver of a link, respectively. Brands B_m and B_n are elements in a brand set $\mathbf{B}=\{\text{Sony, Canon, Panasonic, JVC}\}$, and media formats F_p and F_q are elements in a

media format set $\mathbf{F} = \{\text{DVD}, \text{Mini-DV}, \text{HD}\}$. The price tiers P_x and P_y are elements in a price tier set⁷ $\mathbf{P} = \{(\sim\$299.99), (\$300.00\sim\$499.99), (\$500.00\sim)\}$.

The above variables are all asymmetric and, hence, including these variables allows for the link probabilities to be asymmetric as well. The variable, $I_{(B_n|B_m)}^{brand}(j|i)$, encodes the brand switching from B_m to B_n . An identical interpretation applies to media format and price tier variables. Using these variables, we specify Equation 3 as the following log odds expression:

$$\begin{aligned} \log \text{odds}[p(y_{ij} = 1|\cdot)] &= \alpha + \beta' \cdot \mathbf{x}_{ij} - \|z_i - z_j\| \\ &= \alpha + \beta'_B \cdot \mathbf{x}_{ij}^B + \beta'_F \cdot \mathbf{x}_{ij}^F + \beta'_P \cdot \mathbf{x}_{ij}^P - \|z_i - z_j\|, \end{aligned} \quad (5)$$

where the covariates \mathbf{x}_{ij}^B are the effects coding of brands of product i and j using $I_{(B_n|B_m)}^{brand}(j|i)$, i.e., if the brands of products i and j are Sony and Canon, respectively, $\mathbf{x}_{ij}^B = [\dots, I_{(\text{CANON}|\text{SONY})}^{brand}(j|i), \dots] = [0, \dots, 0, 1, 0, \dots, 0]$. β'_B , β'_F and β'_P are the response parameters that capture the directional link strength for all possible product pairs. For instance, the coefficient of $\beta_{B,(\text{Canon}|\text{Sony})}$, an element in β'_B , measures the strength of brand switching from the Sony brand to the Canon brand. The covariates \mathbf{x}_{ij}^F and parameters β'_F represent the same for media formats while \mathbf{x}_{ij}^P and β'_P do the same for price. To compare coefficients across different variables in a straightforward manner, we use effects coding in \mathbf{x}_{ij} , rather than dummy coding. In effects coding, rather than setting one $\beta_{B,(i|j)}$ to zero, the normalization is such that the average effect of β_B is zero (and similar for normalizations in β_F and β_P). We refer to a study by Bech and Gyrd-Hansen (2004) for more details.

Our proposed model is a main effects model: we investigate the direct effects of three variables on the link formations in the product search network. An interaction model is possible but would require an estimation of much higher number of parameters.

4.3 Estimation

We use the Markov Chain Monte Carlo (MCMC) method to estimate the parameters and latent product positions. We choose the Bayesian approach over Maximum Likelihood (ML) since the log-

⁷We have used price tiers similar to what Amazon.com provides in its product pages.

likelihood as a function of latent positions, $\{z_i\}$, is non-concave (Hoff, Raftery, and Handcock 2002). In general, the non-concavity presents the usual difficulties in finding the ML parameters, but less so in MCMC algorithms (Hoff, Raftery, and Handcock 2002). For an outline of the estimation, we refer to the appendix.

We estimate three versions of the model - a pure random effects model and two versions of the mixed effects model - each with its own purpose. In the pure random effects model, the link probability is represented solely by product positions in a multi dimensional latent space. In this model, Equation (3) becomes:

$$\Pr(y_{ij} = 1|x_{ij}, \alpha, \beta, z_i, z_j) = \frac{1}{1 + \exp(f(z_i, z_j) - \alpha)}. \quad (6)$$

In this specification, we focus on estimating the latent position z_i 's, which in turn allow for an intuitive visualization of sets of commonly searched products in the latent space.

An important advantage of this model is that the inter-product distance captures both types of link dependency, i.e., reciprocity and transitivity. A potential limitation of the pure random effects model is that it does not allow for asymmetry. Therefore, in the second and third versions of the model, we combine fixed and random effects to allow for added flexibility. The first mixed effects model imposes a symmetry restriction on coefficients and the second does not. In the symmetric version, we set the off-diagonal coefficients identical ($\beta_B(n|m) = \beta_B(m|n)$, $\beta_F(q|p) = \beta_F(p|q)$, and $\beta_P(y|x) = \beta_P(x|y)$). We do not impose any restriction in the asymmetric version.

5 Results

5.1 Model fit and validation

We commence this section by verifying that our model fits and predicts well. We first look at in-sample fit by comparing the *observed* product search network for the month of August with the *predicted* product search network using the estimated model from that month. Second, for external validation, we use the actual product search network for the month of October, an out-of-sample data set.

Using the estimated parameters, we predict the probability of a link formation from product i to

j , by

$$\Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{\beta}, \hat{z}_i, \hat{z}_j),$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimated model parameters and \hat{z}_i and \hat{z}_j are estimated latent positions of i and j . Using the predicted link probabilities, we predict the product search network, \hat{Y} , as

$$\hat{Y} = [\hat{y}_{ij}]_{i,j=1,\dots,N,i \neq j} \sim [\Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{\beta}, \hat{z}_i, \hat{z}_j)], \quad (7)$$

where \hat{y}_{ij} is an element at (i, j) of the predicted product search network, \hat{Y} . Note that each element, \hat{y}_{ij} , is a binary random variable with a probability of $\hat{p}_{ij} = \Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{\beta}, \hat{z}_i, \hat{z}_j)$.

One simple way to compare the actual and predicted networks is to compute the root mean square error (RMSE) between two networks, Y and \hat{Y} , element-by-element:

$$RMSE(\hat{Y}, Y) \equiv \sqrt{\frac{1}{N \cdot (N - 1)} \sum_{j \neq i}^N \sum_i^N (\hat{p}_{ij} - y_{ij})^2}. \quad (8)$$

Another way to compare the actual and predicted networks is to compute the hit rate of the predicted network. We further break down the hit rates by computing them conditional on observed absence or presence of a link. In both RMSE and expected hit rate computations, we compare four different models: a base model, the pure random effects model, and the symmetric and asymmetric mixed effects models. We define the base model by setting all link probabilities to 0. This yields an unconditional, expected hit rate of 0.87. Without a detailed analysis of the network data, the choice of zero probability achieves the highest overall hit rate since it always correctly predicts the absence of links.

Table 5 compares the performances of different models using RMSE and various expected hit rates. The hit rates of our models are very good, with more complex models being better than simpler models.

The out-of-sample results are shown in Table 6. All three models exhibit high levels of predictive abilities compared to the base model. The performances of the proposed models in the external validation are comparable to those in the internal validation. High predictive abilities of the models

Table 5: In-sample fit (standard error)

Models	RMSE	Mean Hitrate for $\{\hat{y}_{ij} = 0 y_{ij} = 0\}$	Mean Hitrate for $\{\hat{y}_{ij} = 1 y_{ij} = 1\}$	Overall Mean Hitrate for $\{\hat{y}_{ij} y_{ij}\}$
Base Model: $p=0.00$	0.3646	1	0	0.8671
Pure Random	0.2667	0.9109(0.0027)	0.5071(0.0123)	0.8572(0.0028)
Symmetric Mixed	0.2415	0.9274(0.0024)	0.6180(0.0110)	0.8863(0.0025)
Asymmetric Mixed	0.2388	0.9291(0.0023)	0.6257(0.0109)	0.8887(0.0025)

reveal that all three models are robust in the presence of changes in the underlying product search network. In terms of expected hit rates, more sophisticated models show better predictive abilities. The asymmetric model outperforms both pure random effects and symmetric mixed effects models. In terms of RMSE, both mixed effects models outperform pure random effects model. In summary, we conclude that the proposed models all outperform the base model and exhibit high degree of robustness. We now turn to the interpretation and discussion of the estimation results.

Table 6: Predictive abilities of models (standard error)

Models	RMSE	Mean Hitrate for $\{\hat{y}_{ij} = 0 y_{ij} = 0\}$	Mean Hitrate for $\{\hat{y}_{ij} = 1 y_{ij} = 1\}$	Overall Mean Hitrate for $\{\hat{y}_{ij} y_{ij}\}$
Base Model: $p=0.00$	0.3466	1	0	0.8797
Pure Random	0.2795	0.8991(0.0030)	0.4926(0.0136)	0.8502(0.0031)
Symmetric Mixed	0.2635	0.9115(0.0026)	0.6033(0.0121)	0.8744(0.0027)
Asymmetric Mixed	0.2637	0.9124(0.0026)	0.6255(0.0120)	0.8755(0.0027)

5.2 The pure random effects model

As previously mentioned, the estimated latent product positions in the pure random effects model summarize all factors considered by consumers in their search process. Posterior means of estimated latent product positions are shown in Figures 1, 2, and 3. These figures present three symbol-coded versions of the map representing brand, media format, and price tier.

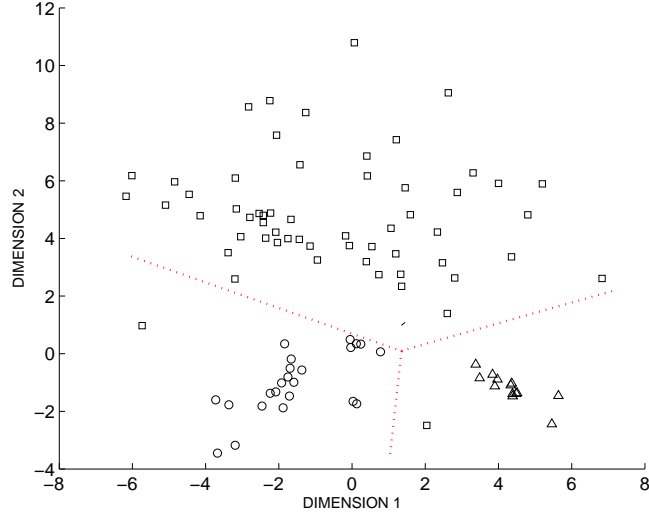


Figure 1: Estimated latent positions of products by media formats. MiniDV (\square), HD (\triangle), and DVD (\circ)

Figure 1 reveals that product search is heavily influenced by media formats and three clearly defined segments emerge. In this figure, products based on Mini-DV technology are scattered around the top of the figure while products based on DVD are scattered around bottom left, with no overlap.

In Figure 2, the three price tiers also seem to form sub-clusters within media-formats. Three price tiers seem to further segment the Mini-DV cluster (top). For instance, products in price tier P_3 (\triangle) are almost all scattered around the right side, while products in price tier P_1 (\square) appear more frequently on far left side of the cluster. In the same figure, we observe that some products in price tier P_3 (\triangle) are also scattered around the center of the product search map. In fact, price tier P_3 overarches across different media format clusters. In the consumer information search context, this may be interpreted that consumer search activities across different media formats are more likely to occur among the products of the most expensive tier of P_3 .

In terms of brands (Figure 3), we see less of a clustering. In fact, in most areas of the map, we see the coexistence of multiple products from different brands.

We can use the map to identify the neighboring products of each alternative in detail. Such a cluster can be interpreted as a stereotypical consideration set or product search set (Desarbo and Jedidi 1995). As an illustration, Figure 4 focuses on an area with product clusters in the Mini-DV

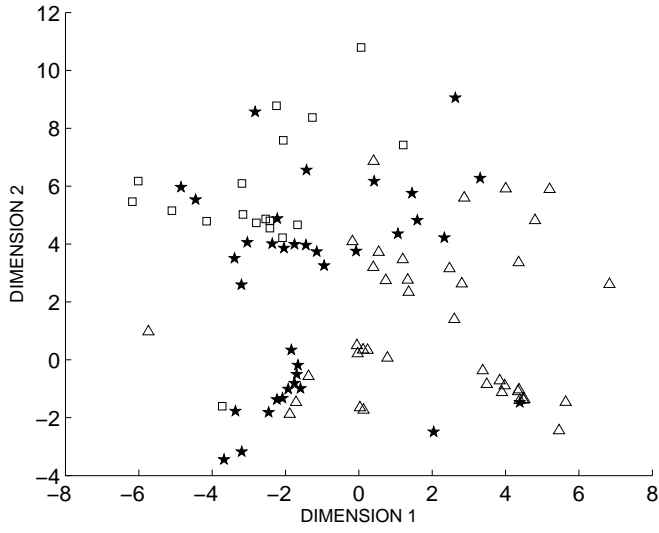


Figure 2: Estimated latent positions of products by prices. \$0~\$299.99 (□), \$300.00~\$499.99 (★), \$500.00~ (△)

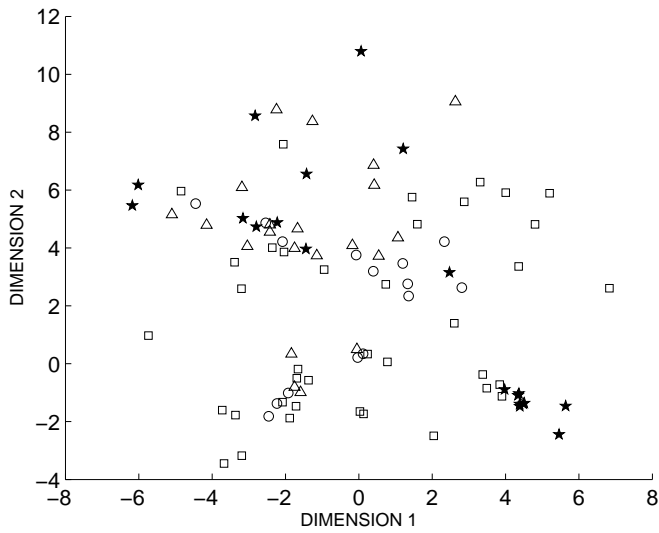


Figure 3: Estimated latent positions of products by brands. Sony (□), Canon (△), JVC (★), and Panasonic (○)

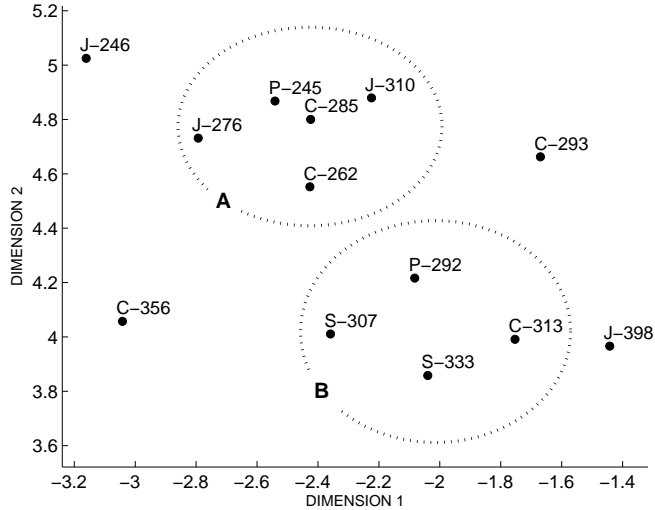


Figure 4: Detailed map of an Mini-DV area in the latent space. The encoding represents “brand-price”. The abbreviations are Sony (S), Canon (C), Panasonic (P), and JVC (J). The products in these clusters represent different brands, but are based on the same media format and in the similar price range.

segment. The products in these clusters represent different brands, but show similar prices. The majority of products at the top sub-cluster are priced below \$300 while the ones at the right bottom are priced around \$300. Note that the product coded as “C-356” (Canon, \$356) is priced substantially higher than its neighboring products and is located farther from its neighboring products in the map. From this figure, we may interpret that consumer search between “C-356” and other products in sub-clusters *A* or *B* is less likely than among products in the same sub-clusters. The clusters *A* and *B* can be interpreted as consideration sets or product search sets.

From the random effects model, we conclude that, by and large, the consumer product search is strongly guided by media format, followed by brand and price. Although intuitive and informative, this map does not quantify the effects of inter-product characteristics on the product cluster formation. We now address this limitation using a richer model.

5.3 The mixed effects models

Using asymmetric variables in the mixed effects model, we measure the effects of inter-product characteristics on link formation between two products. Estimation results for both versions of mixed

effects models are shown in Table 7.

As shown in the last row of Table 7, the asymmetric model exhibits a smaller value of DIC (Deviance Information Criterion) and is therefore judged to be better.⁸ From a consumer information search perspective, this implies that consumers conduct significant amount of asymmetric search activities across different brands, media formats, and price tiers. Therefore, we choose asymmetric model for further investigation. Recall that to facilitate interpretation of the estimated parameters we use effects coding: all parameter estimates for different levels in a variable in Table 7 sum up to zero.

First, we draw inference on the “importance” of brand, media format, and price similarity on the link formation by comparing coefficient ranges among these variables. This computation reveals that media format importance ($8.50 = \beta_{F,(H|H)} - \beta_{F,(M|D)} = 6.44 - (-2.06)$) is greater than both brand (5.06) and price (3.09) importance. This is consistent with the previous findings from the pure random effects model that media format is the most influential attribute on link formation between products. Also, brand has a greater effect on link formation than price⁹.

Second, the estimated parameters in Table 7 reveal that the link formation propensity is stronger within the same attribute levels than across different attribute levels. For instance, for the brand variable, the largest four coefficients correspond to the link formation effect between products of the same brands.

Third, we investigate the brand effects on link formation and find significant differences between brands. For example, conditional on a search for a Panasonic, the propensity to search for a Canon is far greater than the propensity to search for a Sony (0.14 vs. -0.85). Also, brand search is asymmetric. For instance, the brand coefficient from Canon to JVC ($\beta_{B,(J|C)} = -1.4445$) is smaller than its opposite ($\beta_{B,(C|J)} = -0.4537$). This implies that, at the individual product level, the consumers’ search propensity from JVC to Canon is more likely than vice-versa.

Fourth, the coefficients for link propensity across different media formats are very small, implying little or no search across different formats. We note that such findings are potentially very useful in

⁸Bayes Factor comparison of two models led to the same conclusion.

⁹We have also estimated the model with alternative price tiers in which the number of products in each price tier is well-balanced, 31(\sim \$359.99), 31(\$360.00 \sim \$525.99), 32(\$526.00 \sim), with no significant changes in results.

Table 7: Posterior means of estimated parameters (standard error)

	Mixed Effects Models ^{a b}	
	Symmetric Model	Asymmetric Model
β_0	1.1571(0.1829)	1.1242(0.1908)
$\beta_{B,(S S)}$	0.8434(0.1788)	0.8841(0.1747)
$\beta_{B,(C S)}$	-0.7155(0.1528)	-0.3227(0.1851)
$\beta_{B,(J S)}$	-1.6377(0.1884)	-1.8835(0.2687)
$\beta_{B,(P S)}$	-0.5800(0.1614)	-0.2430(0.2158)
$\beta_{B,(S C)}$	-0.7155(0.1528)	-1.1235(0.2109)
$\beta_{B,(C C)}$	2.6040(0.2635)	2.6623(0.2398)
$\beta_{B,(J C)}$	-0.9176(0.1914)	-1.4445(0.2805)
$\beta_{B,(P C)}$	0.3110(0.1909)	0.5506(0.2607)
$\beta_{B,(S J)}$	-1.6377(0.1884)	-1.3059(0.2546)
$\beta_{B,(C J)}$	-0.9176(0.1914)	-0.4537(0.2491)
$\beta_{B,(J J)}$	1.0062(0.3198)	1.0067(0.3113)
$\beta_{B,(P J)}$	-0.2241(0.1985)	0.5852(0.2598)
$\beta_{B,(S P)}$	-0.5800(0.1614)	-0.8536(0.2269)
$\beta_{B,(C P)}$	0.3110(0.1909)	0.1437(0.2567)
$\beta_{B,(J P)}$	-0.2241(0.1985)	-1.3884(0.3619)
$\beta_{B,(P P)}$	3.0741(0.3148)	3.1863(0.3293)
$\beta_{F,(M M)}$	0.8880(0.1682)	0.9414(0.1783)
$\beta_{F,(H M)}$	-1.9182(0.1973)	-1.8075(0.2983)
$\beta_{F,(D M)}$	-1.9855(0.1817)	-2.0616(0.2275)
$\beta_{F,(M H)}$	-1.9182(0.1973)	-1.9981(0.3038)
$\beta_{F,(H H)}$	6.3654(0.6568)	6.4409(0.6624)
$\beta_{F,(D H)}$	-1.3848(0.2276)	-1.5081(0.3188)
$\beta_{F,(M D)}$	-1.9855(0.1817)	-1.9553(0.2103)
$\beta_{F,(H D)}$	-1.3848(0.2276)	-1.3930(0.3816)
$\beta_{F,(D D)}$	3.3237(0.3066)	3.3413(0.3204)
$\beta_{P,(P_1 P_1)}$	0.9723(0.2767)	0.8987(0.2794)
$\beta_{P,(P_2 P_1)}$	0.1276(0.1272)	0.3301(0.1702)
$\beta_{P,(P_3 P_1)}$	-1.5019(0.2084)	-1.6977(0.2924)
$\beta_{P,(P_1 P_2)}$	0.1276(0.1272)	-0.1349(0.1752)
$\beta_{P,(P_2 P_2)}$	0.5854(0.1757)	0.6430(0.1800)
$\beta_{P,(P_3 P_2)}$	-0.0706(0.1287)	-0.2611(0.1691)
$\beta_{P,(P_1 P_3)}$	-1.5019(0.2084)	-1.2665(0.2739)
$\beta_{P,(P_2 P_3)}$	-0.0706(0.1287)	0.0947(0.1620)
$\beta_{P,(P_3 P_3)}$	1.3322(0.1993)	1.3937(0.2142)
DIC ^c	3554.9	3529.4

^aAbbreviations are as follows. Sony (S), Canon (C), JVC (J), Panasonic (P), DVD (D), miniDV (M), Hard-Drive (H), P_1 (~\$299.99), P_2 (\$300.00 ~\$499.99), and P_3 (\$500.00 ~)

^bWe use effects coding in the model.

^cDeviance Information Criterion. The constant in deviance is set to 0.

demand analysis, because they limit the set of options that are effectively substitutes for a product being searched for.

Lastly, the price effects on link formation suggest that consumers are willing to search outside a single price tier. Indeed, the importance of price – as defined above– is smaller compared with brand and media format. However, intuitively, the results suggest that adjacent price tiers promote more link formation than remote ones. For instance, there is a stronger search propensity from the low price tier (P_1) to the medium tier (P_2) than from the low price to the high price tier (P_3) ($\beta_{P,(P_2|P_1)} = 0.3301 > -1.6977 = \beta_{P,(P_3|P_1)}$). We also observe asymmetric price effects on link formation. Search propensity is stronger from the low price tier to the medium price tier than in opposite direction ($\beta_{P,(P_2|P_1)} = 0.3301 > \beta_{P,(P_1|P_2)} = -0.1349$)¹⁰.

5.4 Consolidated effects on product search

The previous section provides detailed insights about individual product level search, since the parameters in our model measure the mean effects of inter-product characteristics on directional link formations between *two products*. For instance, from Table 7, we note that Sony’s own brand coefficient ($\beta_{B,(S|S)} = 0.8841$) is smaller than that of Canon ($\beta_{B,(C|C)} = 2.6623$). This result may reflect that there are simply more Sony products in our sample and that these products are less densely linked than Canon products. In this explanation, a search link within the Sony brand may, at the aggregated level, still be more likely than that within the Canon brand. To study such issues, we investigate the consolidated effects of inter-product characteristics on link formation between *product groups*. A product group is a set of products that share a certain product characteristic. For instance, all products that are based on Mini-DV belong to “Mini-DV” group.

Our approach in estimating the effects of group descriptors (e.g., brand or media format) on product search is to make contrast realizations of the product search networks with and without these descriptors and next compute their marginal effects. The base propensity of having a link from

¹⁰This inequality holds with the alternative price tiers discussed earlier.

product i to j , excluding the effects of brands, media formats, and prices, is computed as

$$\Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{z}_i, \hat{z}_j)$$

where $\hat{\alpha}$ is the estimated intercept, and \hat{z}_i and \hat{z}_j are estimated latent positions of i and j , all from asymmetric mixed effects model. Aggregating these base probabilities over all products, we compute the *base link count* from a product group, G_m to another group, G_n :

$$C_{(G_n|G_m)}^0 = \sum_i \sum_{j \neq i} \Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{z}_i, \hat{z}_j) \cdot I_{(G_n|G_m)}(j|i),$$

where

$$I_{(G_n|G_m)}(j|i) = \begin{cases} 1 & \text{if } i \text{ belongs to product group } G_m \text{ and } j \text{ to product group } G_n \\ 0 & \text{otherwise.} \end{cases}$$

and where $C_{(G_n|G_m)}^0$ is the base link count from product group G_m to G_n . The above formulation quantifies the base link count between two *product groups* excluding the product-level characteristics. The probability term, $\Pr(y_{ij} = 1 | \cdot)$, captures the individual product effects on link formation while the summation captures the different number of products in each product group. This aggregation now accounts for the number of products available at each variable level.

The link probability from product i to j that includes both base and brand effects, but not other characteristics is computed by

$$\Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{\beta}_B, \hat{z}_i, \hat{z}_j)$$

where $\hat{\beta}_B$ is the estimated, product-level brand parameter. The link count under the base and brand effects is computed in a similar manner,

$$C_{(G_n|G_m)}^{0,B} = \sum_{j \neq i} \sum_i \Pr(y_{ij} = 1 | \mathbf{x}_{ij}, \hat{\alpha}, \hat{\beta}_B, \hat{z}_i, \hat{z}_j) \cdot I_{(G_n|G_m)}(j|i),$$

where $C_{(G_n|G_m)}^{0,B}$ is the link count under base and brand effects. The *net effect of brand* on the link

count from product group G_m to G_n is then obtained by

$$C_{(G_n|G_m)}^B = C_{(G_n|G_m)}^{0,B} - C_{(G_n|G_m)}^0,$$

where $C_{(G_n|G_m)}^B$ is the link count using the parameters guiding the coincidence of brands only. The above equation quantifies the net effect of brand on link formation between two product groups. The results are summarized in Table 8.

Among the brand-based product groups, Canon has the highest brand effect on link count increase. Its link count increase, $C_{(\text{Canon}|\text{Canon})}^B$, is 125.01. Sony follows Canon with an increase of 91.38 ($=C_{(\text{Sony}|\text{Sony})}^B$) while JVC has the lowest net brand effect.¹¹ Among the format-based product groups, the link count increase within media-format groups is very pronounced. Therefore, effects within identical media formats are influential both at the individual product level and at the product group level. Consistently, relative to the base case, we observe lower link counts across different format-based product groups. For instance, going from the Mini-DV group to the Hard-disk group, we observe a net loss in the link count ($C_{(H|M)}^F = -59.90$). Price effects are relatively small compared to other effects. The price effect on link count is greatest for the highest price tier P_3 ($C_{(P_3|P_3)}^P = 187.56$). This finding supports the finding in the pure random effects that cross-format search intensity is strongest within price tiers of P_3 .

Lastly, we would like to discuss how the product group level analysis in this section differs from the raw link breakdowns in Tables 2, 3, and 4. The link breakdown based on brand in Table 2, for instance, does not control the effects from media formats and prices. Therefore, the breakdown reflects the combined effects of “Sony brand” and any effects from accompanying media formats and prices. In contrast, the computed quantity, $C_{(\text{Sony}|\text{Sony})}^B$, measures the sole effect of brand on expected link count between Sony products while controlling for effects from media format and price. It is interesting to note that, in Table 2 Sony *products* retain a higher percentage (56%) of its outgoing links compared to Canon (54%). However, after controlling for other effects as reported in Table 8, Sony *brand’s* net effect is smaller than that of Canon ($C_{(\text{Canon}|\text{Canon})}^B = 125.01 > 91.38 = C_{(\text{Sony}|\text{Sony})}^B$).

¹¹Clearly, search is not the same as purchase. It may well be that Sony products are searched less together but bought more than Canon’s or it may be that Sony products are searched more because they do better on average on other criteria such as price or media format.

Table 8: Effects of inter-product characteristics on expected link counts between product groups

$(B_n B_m)$	Base $C^0_{(B_n B_m)}$	Base & Brand $C^{0,B}_{(B_n B_m)}$	Net Brand $C^B_{(B_n B_m)}$
$S S$	127.53	218.91	91.3849
$C S$	89.7348	72.5509	-17.1839
$J S$	77.7032	18.3894	-59.3138
$P S$	49.3743	42.0538	-7.3205
$S C$	89.7348	40.1800	-49.5548
$C C$	48.5145	173.5212	125.0067
$J C$	51.0335	17.4012	-33.6323
$P C$	29.4386	41.6007	12.1622
$S J$	77.7032	29.9858	-47.7174
$C J$	51.0335	37.5337	-13.4998
$J J$	40.0104	69.1069	29.0965
$P J$	26.8079	38.5512	11.7433
$S P$	49.3743	27.1114	-22.2629
$C P$	29.4386	32.3352	2.8966
$J P$	26.8079	9.5564	-17.2515
$P P$	19.5610	87.2661	67.7052
$(F_q F_p)$	Base $C^0_{(F_q F_p)}$	Base & Format $C^{0,F}_{(F_q F_p)}$	Net Format $C^F_{(F_q F_p)}$
$M M$	340.5983	578.6944	238.0961
$H M$	79.9091	20.0128	-59.8962
$D M$	122.7794	22.9520	-99.8274
$M H$	79.9091	16.9311	-62.9780
$H H$	20.2956	126.6124	106.3168
$D H$	28.1810	8.8025	-19.3785
$M D$	122.7794	25.2916	-97.4878
$H D$	28.1810	9.7043	-18.4767
$D D$	61.1684	245.3792	184.2108
$(P_y P_x)$	Base $C^0_{(P_y P_x)}$	Base & Price $C^{0,P}_{(P_y P_x)}$	Net Price $C^P_{(P_y P_x)}$
$P_1 P_1$	44.4241	69.7147	25.2906
$P_2 P_1$	82.6262	99.8250	17.1987
$P_3 P_1$	52.9235	13.8056	-39.1180
$P_1 P_2$	82.6262	76.1223	-6.5039
$P_2 P_2$	146.2002	214.7463	68.5461
$P_3 P_2$	129.7891	108.1783	-21.6108
$P_1 P_3$	52.9235	20.0108	-32.9128
$P_2 P_3$	129.7891	138.3367	8.5476
$P_3 P_3$	162.4992	350.0609	187.5617

In sum, this section has found that product search between attribute-based groups of camcorders takes place first and foremost within media formats. Second, at the brand level Canon has more links to itself than other brands.

6 Managerial Implication and Conclusion

To the best of our knowledge, no previous study has been carried out to analyze the structure of consumer information acquisition in an environment with many choice options, where the set of searched products is typically a small subset of all options, and therefore the contents of this subset are material to analyzing revealed preferences. Online purchasing and browsing behavior forms a natural environment to study product search among durables. Our study is also the first to use data that are publicly available across many product categories from Amazon.com to investigate pre-purchase product search patterns. We measure the effects of product characteristics on product search behavior and in doing so explain the outcome of the product search process.

We propose to model binary, search data using a random effects network model in a generalized linear mixed model with a logit link function. We estimate a pure random effects model and two versions of mixed effects models. The pure random effects model provides an intuitive visualization of products' positions in a latent product space. However, the pure random effects model is symmetric and we address this limitation with asymmetric mixed effects model. Our models predict well in and out of sample.

From a substantive perspective, the analysis of a product search network in this paper provides several important findings. First, using a “product search map” from the pure random effects model, managers can monitor, in detail, each product's neighboring products during consumers' product search stages. This was illustrated with Figure 4. This map allows managers to scrutinize the *local* relationships that exist among products and hence, to better understand substitution patterns for their own products during consumers' pre-purchase activities. Unlike many brand maps in previous marketing literature, which visualize competition among a few brands, this intuitive and informative map provides detailed information on whether a product is likely to be searched given any other

product.

Second, the results from the asymmetric mixed effects model help managers understand which inter-product characteristics influence the contents of consumer product search. Model parameters quantify the degree of influence that brands, media formats, and prices have on the formation of consumer product search. Applicable to all marketing managers in the present product category is that the intensity of consumer information search within same media formats is far greater than those across media formats. This seems an important finding with substantial implications in advertising content decisions. For instance, from Table 1, we see that JVC manufactures digital camcorders based on Mini-DV and Hard Drive technologies. Under our findings, JVC's advertisement that communicates the advantage of Hard-Drive based camcorders will be more effective than advertisement that focuses on its brand when it tries to increase their *products'* exposure in product search.

Third, the breakdowns of links based on brands, media formats, and price tiers as reported in the product group level analysis (Table 8) helps managers to separately identify the effects of brands, media formats, and prices on consumers' product search. Using this aggregate level information, brand managers can monitor how their brands are positioned with respect to other brands during consumers' product search. For instance, using our results, Panasonic managers would conclude that their information and product search volume loss is greatest to Canon and least to Sony. We can draw parallel inferences for other brands and media formats. Their marketing efforts should be aligned with these pre-purchase, brand and technology search patterns.

There are several limitations and future research opportunities in this study. Consolidated lists of search set in our data do not provide frequency information on the product links. We only use binary information in the pre-purchase search stage that indicates whether a pair of products were searched together at an aggregate level. While this presents us with some information loss, the advantage of using only binary information is that it leaves our analysis substantively unaffected by the idea that the link strength between two products is the result of native as well as Amazon.com-directed search. Another limitation is that the collected data were generated during pre-purchase browsing stages and thus the findings apply to pre-purchase stages only. However, given that a subset of searched products may constitute consideration sets, from which final choices are made, findings in this study can be

used in a cautious manner to infer mechanisms responsible for choice set generation in the digital camcorder category.

In terms of future research, we recall that the strength (but not the presence) of links is to some extent endogenous and subject to the Amazon.com's recommendation mechanism. Therefore, consumer information search and choice sets will be driven by both consumer preferences and the influence of Amazon.com's navigation tools. A next logical step using our type of data would be to introduce, in an explicit manner, how individual consumers take into account such navigation tools. We leave this for future research.

A Appendix

A.1 Amazon.com aggregate-level product search data

Amazon.com had over 69 million active customers in 2007¹² and sold over 11 billion worth of goods in 2006.¹³ According to the Amazon.com patent (Linden et al. 2005), the sequence of operations generating the pre-purchase search data is as follows:

1. User click stream or query log data that reflect products viewed by each user during an ordinary browsing session are stored for a certain period of time. A product is *viewed* by a shopper only if the corresponding product detail page is requested.
2. The degree of relationship between two products is measured based on how frequently they are viewed together by consumers. The more frequently two products are viewed together by consumers, the more closely the products are related.
3. The above measurement process is repeated on all pairs of products.
4. For each focal product, related products are sorted in the order of decreasing relationship.
5. Among the sorted products, products outside of the focal product’s category are removed from the list.¹⁴
6. The top M related products are extracted for each focal product.¹⁵

Digital camcorders, now the dominant type in camcorder category, store images and audio on a digital storage medium and offer very good picture and sound qualities. We first downloaded web pages for over 250 products, each of which contains product-related data. For each product, four web pages are downloaded. Two pages contain information on pre-purchase product search data. A

¹²Rick Dalzell, Amazon.com senior vice president, who appeared on *CIO.com* on October 17, 2007.

¹³About 55% of sales are generated in North America.

¹⁴A category is defined in many different ways at Amazon.com. During the data collection period for this paper, the categorical search data for digital camcorder only include digital camcorders and analog camcorders based on Hi8 medium.

¹⁵During the data collection period for this paper, Amazon.com had two pages each of which can list up to 9 products in the same product category. This means that every product has a maximum of 18 “neighbors” or related products. The exact number of listed “neighboring” items varies from item to item.

third web page contains detailed product information such as list price, brand, media format, number of pixels, and screen size. On the same page, Amazon.com provides product-related, store-specific information such as retail prices, sales rankings, and customer reviews. We call this product characteristics data. Finally, we also downloaded shopping cart page for each product since Amazon.com sometimes does not show its discounts until the product is placed in the shopping cart. Therefore, we downloaded shopping cart page for each item to find out Amazon.com’s actual sales price. Both *pre-purchase product search data* and *product characteristics data* are used in the empirical study in this paper. Once the web pages are downloaded, relevant pieces of data are parsed from the web pages and assembled into a data set for one day. We repeated this process over time and constructed a longitudinal database.

A.2 MCMC estimation

1. Choose random starting points for model parameters, α_0 , β_0 , and latent positions Z_0 .
2. Construct an MCMC chain as
 - (a) Propose a candidate \check{Z} from a proposal distribution, $\pi(\check{Z}|Z_k)$
 - (b) Accept $Z_{k+1} = \check{Z}$ with $\min(1, \frac{p(Y|\check{Z}, \alpha_k, \beta_k)}{p(Y|Z_k, \alpha_k, \beta_k)} \frac{\pi(\check{Z}|Z_k)}{\pi(Z_k|\check{Z})})$
 - (c) Otherwise set $Z_{k+1} = Z_k$
3. Run a procrustean transformation¹⁶ on Z_{k+1} with Z_k as a reference coordinate.
4. Sample α and β in a similar manner as in step 2.

In the second and fourth steps, we use a normal distribution as a proposal function, $\pi(\cdot)$. We also use diffuse priors, $f(\alpha, \beta, Z) \propto \text{const.}$, which cancel out in the acceptance probability computation. For model estimation, we run the MCMC chain for 20,000 iterations. We save a sample every 5

¹⁶Since Euclidean distances among a set of positions are invariant under rotation, translation and reflection, potentially there are an infinite number of configurations of positions that give same distance structure. A procrustean transformation on Z_{k+1} makes these coordinates comparable to the Z_k , i.e., chooses among many possible configurations with distances implied by Z_{k+1} that one which is most comparable to Z_k . Please refer to Sarkar and Moore (2005).

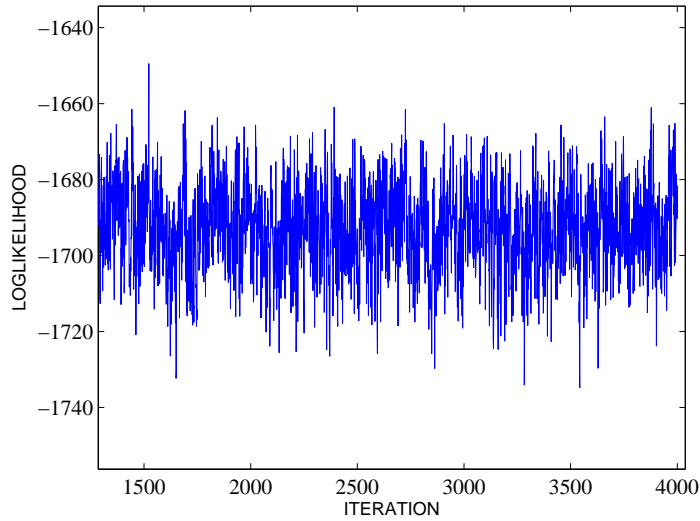


Figure 5: MCMC trace of loglikelihood in asymmetric mixed effects model

iterations to avoid autocorrelation among the samples. In the sampling process, we dynamically adjust the variance of the proposal density functions to achieve an acceptance ratio of 30% (Kao, Jen, and Allenby 2005). The MCMC chains mix relatively quickly in both estimations and we confirm the chains' convergence by visually inspecting traces of both log-likelihood and several model parameters. Please refer to Figures 5 and 6 for sample chain traces in mixed effects model estimation. We use the last 1000 samples for inference.

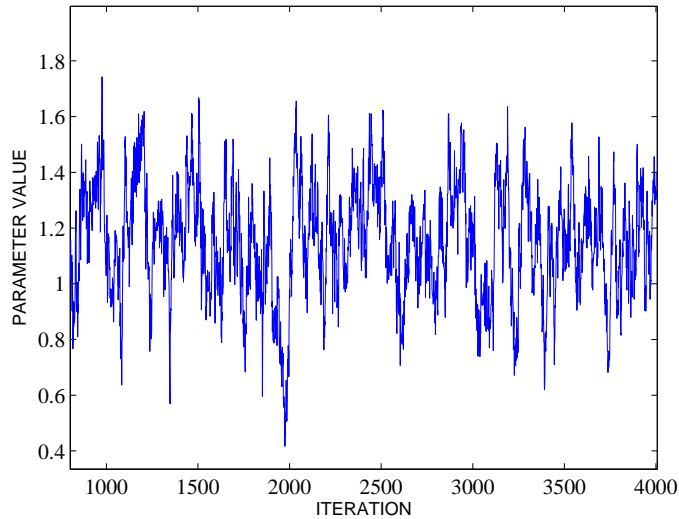


Figure 6: MCMC trace of α in asymmetric mixed effects model

References

- [1] Allenby, Greg M. (1989), “A Unified Approach to Identifying, Estimating and Testing Demand Structures with Aggregate Scanner Data,” *Marketing Science*, 8 (summer), 265-80.
- [2] Anderson, Simon P. and Régis Renault (1999), “Pricing, Product Diversity, and Search Costs: A Bertrand-Chamberlin-Diamond Model,” *The RAND Journal of Economics*, 30 (4), 719-35.
- [3] Bajari, Patrick, Jeremy T. Fox, and Stephen P. Ryan (2007), “Evaluating Wireless Carrier Consolidation Using Semiparametric Demand Estimation,” working paper, University of Chicago.
- [4] Beatty, Sharon E. and Scott M. Smith (1987), “External Search Effort: An Investigation Across Several Product Categories,” *The Journal of Consumer Research*, 14 (1), 83-95.
- [5] Bech, Mickael, and Dorte Gyrd-Hansen (2004), “Effects coding in discrete choice experiments,” *Health Economics*, 14 (10), 1079-83.
- [6] Bettman, James R. and Jacoby, Jacob (1976), “Patterns of Processing in Consumer Information Acquisition ” *Advances in Consumer Research*, 3 (1), 315-20.

- [7] ——— and Pradeep Kakkar (1977), “Effects of Information Presentation Format on Consumer Information Acquisition Strategies,” *The Journal of Consumer Research*, 3 (March), 233-40.
- [8] Bronnenberg, Bart J. and Wilfried R. Vanhonacker (1996), “Limited Choice Sets, Local Price Response and Implied Measures of Price Competition,” *Journal of Marketing Research*, 33 (2), 163-73.
- [9] Brynjolfsson, Erik, Yu Hu and Michael D. Smith (2003), “Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers,” *Management Science*, 49 (11), 1580-96.
- [10] Chevalier, Judith and Austan Goolsbee (2003), “Measuring Prices and Price Competition Online: Amazon.com and BarnesandNoble.com,” *Quantitative Marketing and Economics*, 1 (2), 203-22.
- [11] Consumer Report (2007), “Buying advice”(accessed April, 2007), [available at <http://www.consumerreports.org/cro/electronics-computers/audio-video/video/camcorders/reports/how-to-choose/index.htm>].
- [12] Desarbo, Wayne, S. and K. Jedidi (1995), “The Spatial Representation of Heterogeneous Consideration Sets,” *Marketing Science*, 14 (3), 326-42.
- [13] Elrod, Terry (1988), “Choice Map: Inferring a Product-Market Map from Panel Data,” *Marketing Science*, 7 (1), 21-40.
- [14] Hauser, John R. and Birger Wernerfelt (1990), “An Evaluation Cost Model of Consideration Sets,” *The Journal of Consumer Research*, 16 (4), 393-408.
- [15] Hoff, Peter D., Adrian E. Raftery and Mark S. Handcock (2002), “Latent Space Approaches to Social Network Analysis,” *Journal of the American Statistical Association*, 97 (Winter), 1090-98.
- [16] Jacoby, Jacob, Robert W. Chestnut and William A. Fisher (1978), “A Behavioral Process Approach to Information Acquisition in Nondurable Purchasing,” *Journal of Marketing Research*, 15 (4), 532-44.

- [17] Johnson, Eric, J., Wendy W. Moe, Peter S. Fader, Steven Bellman and Gerald L. Lohse (2004), “On the Depth and Dynamics of Online Search Behavior,” *Management Science*, 50 (3), 299-308.
- [18] Kao, Ling-Jing, Lichung Jen and Greg M. Allenby (2005), “A State-Space Model of Purchase Timing for Direct Marketing”, working paper. Ohio State University.
- [19] Levinson, Meridith (2007), “Amazon.com’s IT Leader Leaving Huge Customer Service Infrastructure as Legacy”, *CIO.com*.
- [20] Linden, Greg D., Brent R. Smith and Nida K. Zada (2005), “Use of Product Viewing Histories of Users to Identify Related Products,” *US Patent Number: 6,912,505 B2*.
- [21] Lynch, Jr., John G. and Dan Ariely (2000), “Wine Online: Search Costs Affect Competition on Price, Quality, and Distribution,” *Marketing Science*, 19 (1), 83-103.
- [22] McCulloch, Charles E. and Shayle R. Searle (2001), *Generalized, Linear, and Mixed Models*, Wiley-series in probability and statistics.
- [23] Mendelsohn, Tamara, Carrie Johnson, and Brian Tesch (2007), “The Web’s Impact On In-Store Sales: US Cross-Channel Sales Forecast, 2006 To 2012”, *Forrester Research*.
- [24] Moe, Wendy (2003), “Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-store Navigational Clickstream,” *Journal of Consumer Psychology*, 13 (1 & 2), 29-39.
- [25] Moe, Wendy (2006), “An Empirical Two-Stage Choice Model with Varying Decision Rules Applied to Internet Clickstream Data,” *Journal of Marketing Research*, 43 (4), 680-92.
- [26] Moorthy, Sridhar, Brian T. Ratchford and Debabrata Talukdar (1997), “Consumer Information Search Revisited: Theory and Empirical Analysis,” *The Journal of Consumer Research*, 23 (4), 263-77.
- [27] Nakanishi, Masao and Lee G. Cooper (2003), “Metric Unfolding Revisited: Straight Answers to Basic Questions”, Paper 2003010112, Department of Statistics, UCLA.

- [28] Nelson, Philp (1970), "Information and Consumer Behavior," *The Journal of Political Economy*, 78 (2), 311-29.
- [29] Newman, John W. (1977) "Consumer External Search: Amount and Determinants," in *Consumer and Industrial Buying Behavior*, ed. Arch Woodsided et al., New York: Elsevier, 79-84.
- [30] Newman, Joseph W. and Bradley D. Lockeman (1975) "Measuring Prepurchase Information Seeking," *The Journal of Consumer Research*, 2 (3), 216-22.
- [31] Newman, Joseph W. and Richard Staelin (1972), "Prepurchase Information Seeking for New Cars and Major Household Appliances," *Journal of Marketing Research*, 9 (August), 249-57.
- [32] Payne, John W. (1976), "Task complexity and contingent processing in decision making: An information search and protocol analysis", *Organizational Behavior and Human Performance*, 16 (August), 366-87.
- [33] Peterson, Robert A., Sridhar Balasubramanian and Bart J. Bronnenberg (1997), "Exploring the Implications of the Internet for Consumer Marketing ", *Journal of the Academy of Marketing Science*, 25 (4), 329-46.
- [34] Roberts, John H. and James M. Lattin (1991), "Development and Testing of a Model of Consideration Set Composition," *Journal of Marketing Research*, 28 (4), 429-40.
- [35] ————— (1997), "Consideration: Review of Research and Prospects for Future Insights," *Journal of Marketing Research*, 34 (3), 406-10.
- [36] Russo, J.E. and Larry D. Rosen (1975), "An Eye Fixation Analysis of Multialternative Choice," *Memory and Cognition*, 3 (May), 267-76.
- [37] Sarkar, Purnamrita and Andrew W. Moore (2005), "Dynamic social network analysis using latent space models," *ACM SIGKDD Explorations Newsletter*, 7 (2), 31-40.
- [38] Siddarth, S., Randolph E. Bucklin, and Donald G. Morrison (1995), "Making the Cut: Modeling and Analyzing Choice Set Restriction in Scanner Panel Data," *Journal of Marketing Research*, 32 (August), 255-66.

- [39] Stigler, George J. (1961), "The Economics of Information," *The Journal of Political Economy*, 69 (3), 213-25.
- [40] Urban, Glen L., John S. Hulland and Bruce D. Weinberg (1993), "Premarket Forecasting for New Consumer Durable Goods: Modeling Categorization, Elimination, and Consideration Phenomena," *Journal of Marketing*, 57 (2), 47-63.
- [41] Wasserman, Stanley and Katherine Faust (1994), *Social Network Analysis: Methods and Applications*, Cambridge, UK: Cambridge University Press.
- [42] Wedel, Michel and Rik Pieters (2000), "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science*, 19 (4), 297-314.
- [43] Westbrook, Robert A. and Claes Fornell (1979), "Patterns of Information Source Usage among Durable Goods Buyers," *Journal of Marketing Research*, 16 (3), 303-12.