Identifying consumer consideration set at the purchase time from aggregate purchase data in online retailing

Bin Gu, Prabhudev Konana, Hsuan-Wei Michelle Chen

Abstract

Online retailers provide a substantial amount of product information to their customers. The information includes not only product features and customer reviews, but also information on alternative products that may better fit a consumer's needs. The systematic provision of information on alternative products could have a significant impact on consumers' purchase decision process at online retailers. In this study, we analyze one aspect of the impact - the degree to which consumers consider multiple products at the purchase time. We leverage a popular feature provided by online retailers - "What Do Customers Ultimately Buy after Viewing This Item?" We show that information contained in this feature can be used to identify consumers' product consideration set and choice at the purchase time when combined with product sales data. The identification is exact in analyzing competition between two products. For competition involving three products, the identification is exact under the assumption that consumer choice follows a discrete choice model. For competition involving more than three products, the information provides a lower bound of the percentage of consumers that consider only one product at the purchase time. We apply the model to 38,400 unique products from Amazon's Electronics category. The results show that more than 78% of consumers purchase a product without considering any other products on Amazon at the purchase time.

1. Introduction

Research on electronic commerce has shown that information provided by online retailers, such as product recommendations and consumer reviews, has a significant impact on consumer product choice [8,10,12,23]. In particular, information provided by online retailers includes not only that pertaining to the product under consideration, but also information on alternative products that may better fit a consumer's needs. This systematic provision of information on alternative products at the time of purchase could have an impact on consumers' purchase decision process. In this study, we analyze one aspect of the impact - the degree to which consumers consider multiple products at the purchase time.

Our approach differentiates from prior studies on consumer consideration set in two ways. First, we develop a new methodology to identify consumer consideration set at the purchase time from aggregate data provided by online retailers. Prior studies on consumer consideration set either require detailed individual level clickstream data or leverage individual level purchase data with potentially unrealistic assumptions on the underlying consumer information search behavior [2,24,25,27]. Our approach leverages a new type of aggregate data provided by online retailers and shows that it is feasible to identify consumer consideration sets using the aggregate data under fewer assumptions. Second, our analysis focuses on consumer consideration sets and choices at the purchase time. While it is well known that consumers make purchase decisions in multiple stages, few studies have analyzed the degree to which consumers make product comparisons at the last minute before making purchase decisions. This question is particularly relevant in online retailing environment as consumers are constantly exposed to information on alternative products, such as those provided by context-sensitive display ads. Understanding the consumer choice process at the moment of purchase in such an environment could help businesses and retailers develop better real-time marketing strategies.

Our research is facilitated by the availability of an increasingly popular feature among online retailers - "What Do Customers Ultimately Buy After Viewing This Item?" For example, two of the largest online retailers, Amazon.com and Buy.com, offer this feature prominently on their product pages. We show that, using information

© 2012 Elsevier B.V. All rights reserved.
from this feature and product sales data, we can exactly identify the size of consumer groups with different consideration sets when analyzing competition between two or three products, and obtain a lower bound on the percentage of consumers that consider no alternative products at the purchase time when analyzing competition involving more than three products.

We apply our model to 38,400 unique products collected from Amazon’s Electronics category. We choose Amazon as our research context because it is one of the largest online retailers and has been studied extensively in prior literature [13,18,21]. The results show that more than 78% of the consumers do not consider any other products at the purchase time. The finding suggests that a majority of the consumers have made decisions before their visits for the purchase and they are not often affected by alternative product information at the purchase time.

The remainder of this paper proceeds as follows. In Section 2, we review the prior literature that relates to our research topic. In Section 3, we present the theories and methodologies to identify consumer consideration sets. We present our data and modeling results along with discussions in Sections 4 and 5, respectively. We conclude our paper in Section 6.

2. Literature review

This study is closely related to marketing literature on consumer consideration formation [17,25,27]. These studies have shown that product consideration set significantly influences consumer choice. In particular, faced with cognitive limitations, complex choice tasks, and evaluation costs, consumers often resort to a phased decision process [14]. The phased consumer decision-making process involves two stages—the consideration stage and choice stage. In the consideration stage, consumers choose a small set of products for evaluation. In the choice stage, consumers evaluate every product in the consideration set and purchase the one with the highest utility. A key challenge in research on the consideration stage is that product consideration is often not observable to researchers, especially in offline settings. As a result, a significant amount of effort has been focused on identifying “whether the consideration stage ... corresponds to a cognitive stage of consideration in the consumer's decision process, or if it is just a statistical artifact of the data.” [28]. The availability of online clickstream data helps answer the aforementioned question [24]. By analyzing what consumers have viewed and what they ultimately purchase, Moe (2006) [24] finds that consumers use different decision criteria for the consideration stage and the choice stage.

While clickstream data offers rich information on consumer behavior, such information is often not available to product manufacturers or competing retailers. Further, extracting useful information from clickstream data requires extensive effort in data collection and analysis. A number of studies have thus focused on inferring consumer consideration sets from non-clickstream data (e.g., scanner data). One challenge of such an approach is that researchers must impose potentially unrealistic assumptions to make inference on the underlying consumer consideration process. For example, Roberts and Lattin (1991) [27] and Andrews Srinivasan (1995) [2] assume that a product’s probability of being considered is independent of a consumer’s consideration of other products, an assumption is unlikely to hold in reality as products with similar characteristics are more likely to be considered together.

Our study extends the extant research on consumer consideration sets on two fronts. First, we show that consumer consideration sets at the purchase time can be identified using a new type of aggregate data provided by the online retailers with few assumptions on the underlying consumer consideration process. Second, prior studies on consumer consideration set have mainly focused on the initial consideration set formed by the consumers. Our analysis instead focuses on consumer consideration sets at the purchase time and identifies the degree to which consumers make product comparison at the last minute.

This study also contributes to a growing interest in electronic commerce investigating product variety and consumer purchase decisions [7]. The first stream of related literature is concerned with the increased product variety introduced by online retailers [1,4,5,26]. These studies suggest that online retailers are able to carry a greater variety of products than their physical counterparts [4,5], thus, expanding the number of products considered by a consumer. In addition, the lower search costs facilitated by product information and consumer recommendation in digital commerce further help consumers discover niche products and expand product consideration [1,7]. Recent studies also find that consumers often conduct information search through third-party infomediaries (e.g., CNET) before making purchase at an online retailer [16]. We complement this research stream by identifying the degree to which consumers compare multiple products at the time of purchase.

Our research also draws on the literature relating to the impact of growing electronic commerce on consumer behavior [6,30,31]. With the increase of marketing channels facilitated by the advancement of technologies, media, and advertising activities, consumers have often been overloaded with information. The limitation of human cognitive and perceptual capability, however, restricts the number of products that can be considered by a consumer. In particular, literature on consumer behavior finds that consumers can only hold seven, plus or minus two, chunks of product information, and sometimes variably less, due to short-term memory limitation [3]. The limits in consumers’ memory capability force consumers to be more selective in product consideration and information processing [9,20,29]. It also motivates the development of a wide array of consumer decision aids and recommendation systems [19,22,32].

3. Theories and methodologies

3.1. Revealed preferences in online retailers

Online retailers provide a variety of product sales statistics to help consumers make better decisions. For example, Amazon provides the following statistics to their customers: 1) products “Frequently Bought Together;” 2) “What Do Customers Ultimately Buy After Viewing This Item;” 3) product sales ranks; 4) “Customers Who Bought This Item Also Bought;” and 5) customer reviews. Such statistics reveal consumer preferences in purchasing decisions, enabling researchers and practitioners to infer the underlying purchase process [15,26]. In this study, we show that two statistics — “What Do Customers Ultimately Buy After Viewing This Item” and product sales ranks — can be used to identify consumer consideration and choice at the purchase time.

Fig. 1 provides a screenshot of “What Do Customers Ultimately Buy After Viewing This Item.” The figure shows that Amazon provides the top 4 products ultimately bought by consumers after viewing a given product and the corresponding percentages of consumers who have done so. While Amazon never reveals the exact calculation of the percentages, discussions with industry insiders suggest that the percentages are calculated from consumer web sessions that result in purchases. Web sessions that do not result in purchases are not included in the calculation. Given this design, the percentages capture consumer decision making at the purchase time, i.e. the last stage of the decision process.

The percentages in Fig. 1 represent conditional probabilities of consumers purchasing product Y after they have viewed X. We show later that consumer consideration sets and their choices can be identified using these conditional probabilities and the sales data when analyzing competition between two or three products. Further, we show that the information provides a lower bound of the percentage of consumers who view only one product when analyzing competition among more than three products.
3.2. Competition between two products

Product competition often centers between two products that are similar in nature. In such cases, retailers and product manufacturers often focus on analyzing the competition between the two products while ignoring consumers who bought other products.

Consider a market with two products, X and Y. Each consumer purchases one and only one product. Consumers make purchase decisions in two stages. In the Consideration Stage, they form a consideration set that could include one of the two products, or both. In the Choice Stage, they make purchase decisions. Given the assumption that the market contains two products, there exist three customer groups at the purchase time: (i) those who have considered only X, (ii) those who have considered only Y, and (iii) those who have considered both X and Y. We use \( P(x|X) \), \( P(y|Y) \), \( P(x|Y) \), and \( P(y|X) \) to denote the total number of consumers who have purchased product X while viewing Y. The number of consumers who have purchased Y after viewing X can be expressed as a ratio of product sales over probabilities carry unique information.

Given the market only has two products and consumers purchase one and only one product, \( P(x|X) + P(y|X) = 1 \) and \( P(x|Y) + P(y|Y) = 1 \). This indicates that only two of the four conditional probabilities carry unique information.

We further note that the probability of purchasing a product after viewing the product can be expressed as a ratio of product sales over the sum of number of customers that have viewed the product. In particular,

\[
P(x|X) = \frac{x}{N(XY) + N(X|Y)}.
\]

The equation suggests that the probability of consumers purchasing X after viewing X can be calculated as the sales of X divided by the sum of number of customers who have viewed only X and the number of customers who have viewed both X and Y. Similarly, we have

\[
P(y|Y) = \frac{y}{N(X|Y) + N(Y|X)}
\]

Finally, we note that, since consumers purchase one and only one product, the total number of viewers must equal to the total number of customers.

\[
x + y = N(X|Y) + N(Y|X) + N(XY)
\]

Eqs. (1)–(3) allow us to identify three different groups of consumers: those who viewed both X and Y, those who viewed X only and those who viewed Y only. It is also useful to note that \( P(y|X) \) and \( P(x|Y) \) are not included in Eqs. (1)–(3) because they are linear functions of \( P(x|X) \) and \( P(y|Y) \) respectively.

Given Eqs. (1)–(3), we can identify the size of each customer group:

\[
N(X|Y) = \frac{x}{p(x|X)} + \frac{y}{p(y|Y)} - x - y = \frac{p(y|X)}{p(x|X)} + \frac{p(x|Y)}{p(y|Y)} y
\]

\[
N(Y|X) = x + y - \frac{y}{p(x|X)}
\]

Intuitively, \( \frac{x}{p(x|X)} \) identifies the total number of consumers who have viewed product X while \( \frac{y}{p(y|Y)} \) identifies the total number of consumers who have viewed product Y. The number of consumers who have viewed both products is counted twice in the foregoing calculation. We can, thus, identify these consumers using the difference between the sum of the two and the total number of consumers. In the extreme case where every consumer views only one product, \( P(x|X) \) and \( P(y|Y) \) will be equal to 1 and \( N(X|Y) \). \( N(Y|X) \) \( N(XY) \) will be equal to 0, x, and y, respectively.

We can also identify the consumer choice decision when they consider both products. To identify the choice process of those who viewed both products, we note that x customers who bought product X can be divided into two groups: those who viewed only product X and those who viewed both products X and Y. Since there are \( N(X|Y) \) customers in the first group, the second group contains \( x - N(X|Y) \) customers. The choice probability of x among customers that viewed both products is thus

\[
P(x|XY) = \frac{x - N(X|Y)}{N(XY)} = \frac{P(x|Y)P(y|X)}{P(y|Y)P(X|Y) + P(x|Y)P(Y|X)}
\]

Fig. 2 illustrates our approach using sales rank and conditional probability data on two popular software products: Adobe Photoshop Elements 9 and Premiere Elements 9 (Win/Mac) by Adobe Windows Vista / 7 / XP, Mac OS X ```

\[
\text{• 1% buy} \quad \text{Adobe Photoshop Elements 9 (Win/Mac) by} \quad \text{Adobe} \quad \text{Windows Vista / 7 / XP, Mac OS X} \quad \text{☆☆☆☆☆ (129)}
\]

\[
\text{• 7% buy} \quad \text{Adobe Photoshop Elements 9 & Premiere Elements 9 (Win/Mac) by} \quad \text{Adobe} \quad \text{Windows Vista / 7 / XP, Mac OS X} \quad \text{☆☆☆☆☆ (120)}
\]

\[
\text{• 1% buy} \quad \text{Adobe Premiere Elements 9 (Win/Mac) by} \quad \text{Adobe} \quad \text{Windows Vista / 7 / XP, Mac OS X} \quad \text{☆☆☆☆☆ (64)}
\]

\[
\text{• 1% Buy} \quad \text{Photoshop Elements 9: The Missing Manual by Barbara Brundage} \quad \text{Paperback} \quad \text{☆☆☆☆☆ (20)}
\]

Fig. 1. An example of “What Do Customers Ultimately Buy After Viewing This Item?” on Amazon.com.
I. Information from Amazon

<table>
<thead>
<tr>
<th>Viewed</th>
<th>Bought</th>
<th>Photoshop Element 9</th>
<th>Bundle 9</th>
<th>Sales Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photoshop Element 9</td>
<td>91%</td>
<td>7%</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Bundle 9</td>
<td>23%</td>
<td>77%</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

II. Derivation of Consumer Groups with Different Consideration Sets

i. Viewed both products

\[ \frac{x}{P(x|X)} + \frac{y}{P(y|Y)} = x + y = 264.03 \]

ii. Viewed only Adobe Element 9

\[ x + y = 1.447.46 \]

iii. Viewed only Adobe Element 9 + Adobe Premiere Element 9

\[ \frac{x}{P(x|X)} + \frac{y}{P(y|Y)} = 363.05 \]

Probability of viewing both products before purchase = 12.73%.

III. Derivation of Consumer Choice

Probability of buying Adobe Photoshop Element 9 after viewing both products

\[ \frac{x - N(x|X)}{N(x|X)} = 54.63\% \]

We assume the following power law relationship between product sales and sales rank:

\[ \log \text{Sales} = 8.046 - 0.613 \log \text{Sales Rank} \]

Fig. 2. Illustration of consumer consideration and choice of two products.

Element 9 (PS 9) and Adobe Photoshop and Premiere Element 9 (Bundle 9). The former is a popular photo editing software for amateurs, while the latter is a bundled product that includes one copy of Adobe Photoshop Element 9 and one copy of Adobe Premiere Element 9, a popular video editing software. Data from Amazon indicates that fewer than 2% of consumers who viewed either of the products end up purchasing something else. So the competition is mainly between the two products at the purchase time. We, thus, rescale the data to remove consumers who ultimately bought other products and focus on consumers who bought either of the two products. For illustration, we assume that the relationship between sales rank and sales is known. The analysis shows that only 13% of the consumers consider both products. The remaining 87% of consumers consider only one product at the purchase time. Among those who consider both products, 55% choose to purchase PS 9.

3.3. Competition among three products

To identify consumer choice and consideration set for more products, we note that there exist \(2^n-1\) consumer groups for a market of \(n\) products given all the possible combinations of products for consideration sets. In addition, to exactly identify consumer choice within each consumer group, we need identification of \(\sum_{k=1}^{n-1}(k-1)\) choice probabilities. In total, \(\sum_{k=1}^{n-1}k\) variables need to be identified. Since Amazon provides \(n^2\) statistics, the identification is feasible for \(n=3\).

To identify consumer consideration and choice process for three products, we need the identification of 9 parameters — 7 parameters for consumer groups based on their consideration sets and 2 parameters for the utility of the three products. Amazon provides 9 conditional purchase probabilities for three products and 3 data points on product sales. The 9 conditional purchase probabilities, however, are not fully independent. As in the case of two-product market, the sum of conditional purchase probabilities after viewing a given product always equals 1. As such, only 6 of the conditional purchase probabilities carry unique information. Combining the 6 conditional purchase probabilities with the 3 data points on product sales, we can exactly identify the 9 parameters for the consumer consideration and choice process of three products.

Specifically, we use the following nine equations. The first set of three equations identifies the conditional probability of purchasing a product after viewing the product. The second set of three equations identifies the conditional probability of purchasing a different product after viewing a given product. The final set of three equations identifies the ultimate sales for each product. It is useful to note that the second and third sets of equations are non-linear and thus require numeric solution.

\[
P(x|X) = \frac{x}{N(XYZ) + N(XYZ) + N(XYZ) + N(XYZ)}
\]

\[
P(y|Y) = \frac{y}{N(XYZ) + N(XYZ) + N(XYZ) + N(XYZ)}
\]

\[
P(z|Z) = \frac{z}{N(XYZ) + N(XYZ) + N(XYZ) + N(XYZ)}
\]

\[
x = N(XYZ) + \frac{u_x}{u_x + u_y + u_z} N(XYZ) + \frac{u_x}{u_x + u_y + u_z} N(XYZ)
\]

\[
y = N(XYZ) + \frac{u_x}{u_x + u_y + u_z} N(XYZ) + \frac{u_y}{u_x + u_y + u_z} N(XYZ)
\]

\[
z = N(XYZ) + \frac{u_x}{u_x + u_y + u_z} N(XYZ) + \frac{u_z}{u_x + u_y + u_z} N(XYZ)
\]

Fig. 3 illustrates our approach by extending the earlier two-product example to three software products: Adobe Photoshop Element 9 (PS 9), Adobe Photoshop and Premiere Element 9 (Bundle 9), and Adobe Premiere Element (PR 9). The data from Amazon

3. See Appendix A for the rescaling procedure.
4. The total number of consumer groups is calculated as the sum of the number of consumer groups with \(k\) products in their consideration sets, with \(k = 1, ..., n\), i.e. \(\sum_{k=1}^{n}k = 2^n-1\).
5. For each consideration set with \(k\) products, we need \(k-1\) variables to identify the choice probability for each product.
6. Amazon provides \(n\) statistics on product sales, and \(n(n-1)\) statistics on the conditional probabilities. In total, we have \(n^2\) statistics.

\[8\text{ We only need 2 parameters to identify the utility of three products because utility is a relative measure. If we multiple the utility of each product by a constant, consumers will make exactly the same choice. As such, we only need to identify the utility of the first two products while normalizing the total utility of the three products to a constant.} \]
I. Information from Amazon

<table>
<thead>
<tr>
<th>Viewed</th>
<th>Bought</th>
<th>Photoshop Element 9</th>
<th>Photoshop/Premiere Element Bundle 9</th>
<th>Premiere Element 9</th>
<th>Sales Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>92%</td>
<td>23%</td>
<td>14%</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7%</td>
<td>76%</td>
<td>16%</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1%</td>
<td>1%</td>
<td>50%</td>
<td>135</td>
</tr>
</tbody>
</table>

II. Derivation of Consumer Groups with Different Consideration Sets

The derivation is based on numeric solutions to the nine non-linear equations outlined in Section 3.3.

i. Viewed Only Adobe Photoshop Element 9 = 1.34673
ii. Viewed Only Adobe Photoshop / Premiere Element Bundle 9 = 312.36
iii. Viewed Only Adobe Premiere Element 9 = 136.78
iv. Viewed Both Adobe Photoshop Element 9 and Adobe Photoshop/Premier Element Bundle 9 = 267.23
v. Viewed Both Adobe Photoshop Element 9 and Adobe Premiere Element 9 = 116.13
vi. Viewed Both Adobe Photoshop/Premiere Element Bundle 9 and Adobe Premiere Element 9 = 55.73
vii. Viewed All Three Products = 0

The probability of PS 9 being considered is 78%.
ix. The probability of Bundle 9 being considered is 29%
x. The probability of PR 9 being considered is 14%.

III. Derivation of Consumer Choice

i. Utility of PS 9 = 0.517
ii. Utility of Bundle 9 = 0.428
iii. Utility of PR 9 = 0.055

The above estimated utilities suggest the following choice probability:

1) Probability of buying PS 9 after viewing both PS 9 and Bundle 9 is 54.68%
2) Probability of buying PS 9 after viewing both PS 9 and PR 9 is 90.36%.
3) Probability of buying Bundle 9 after viewing both Bundle 9 and PR 9 is 88.60%.

II.4. Competition among multiple products

To consider competition among more than three products, we use $D_i$ to denote the action of viewing product $i$ and $d_i$ to denote the number of consumers who purchase product $i$. Amazon provides the following information:

a. $d_i$ — the number of consumers who purchased product $i$, for all $i$.
b. $P(d_i|D_i)$ — the probability of consumers purchasing $j$ after viewing product $i$, for all $i$ and $j$.

We again note that $\sum_i P(d_i|D_i) = 1$, for all $i$. This condition indicates that while (b) provides a total number of $n^2$ statistics, only $n(n-1)$ of them contain unique information.

Given that the conditional probability is between each pair of products, but consumers may consider more than two products in this setting, we cannot exactly identify the size of all possible consideration sets. However, we show now that the aforementioned information is sufficient to provide a lower bound of the percentage of consumers who consider no alternative products at the purchase time.

Note that the conditional probability of purchasing a product after viewing the product can be expressed as follows:

$$P(d_i|D_i) = \frac{\sum m \in I}{}$$

In Eq. (7), $N(\cap m \in [i] D_m \cap m \neq i [j] D_j)$ refers to the consumers who only view product $i$ at the purchase time with $j$ referring to the entire product set. $N(\cap m \in [i] D_m \cap m \neq j [i] D_i)$ refers to the consumers who only view products $i$ and $j$ at the purchase time and $N(\cap m \in [i] D_m)$ refers to the consumers who view all the products at the purchase time. Swapping the LHS with the denominator in the RHS, we have

$$N(\cap m \in [i] D_m \cap m \neq i [j] D_j) + \sum m \neq i \sum j \neq i N(\cap m \in [j] D_m \cap m \neq j [i] D_i) + \cdots + N(\cap m \in [i] D_m)$$

(7)

Summing Eq. (8) over all $i$, we have

$$\sum_i N(\cap m \in [i] D_m \cap m \neq i [j] D_j) + \sum m \neq i \sum j \neq i N(\cap m \in [j] D_m \cap m \neq j [i] D_i) + \cdots + N(\cap m \in [i] D_m)$$

(8)

Note that

$$\sum_i N(\cap m \in [i] D_m) = n N(\cap m \in [i] D_m)$$

(10)

$$\sum i \sum j \neq i N(\cap m \in [j] D_m \cap m \neq j [i] D_i) = 2 \sum \sum i j N(\cap m \in [i] D_m \cap m \neq j [i] D_i)$$

(11)

We have

$$\sum i \sum j \neq i N(\cap m \in [i] D_m \cap m \neq i [j] D_j) + 2 \sum \sum i j N(\cap m \in [i] D_m \cap m \neq j [i] D_i) + \cdots + n N(\cap m \in [i] D_m) = \sum i \frac{d_i}{P(d_i|D_i)}$$

(12)

We further note that the sum of all consumer groups with different consideration sets equals the number of total consumers, i.e.

$$\sum i N(\cap m \in [i] D_m \cap m \neq i [j] D_j) + \sum \sum i j N(\cap m \in [j] D_m \cap m \neq j [i] D_i) + \cdots + n N(\cap m \in [i] D_m)$$

(13)

Subtracting Eq. (13) from Eq. (12), we have

$$\sum \sum i j N(\cap m \in [j] D_m \cap m \neq j [i] D_i) + \cdots + (n-1) N(\cap m \in [i] D_m) = \sum \left(\frac{d_i}{P(d_i|D_i)} - d_i\right)$$

(14)
Note that the total number of customers who consider more than one product is:

\[
\sum_{\text{all } i} \left( \left( \sum_{\text{all } j} \frac{N(\{i \in \{j\} \cap \{\text{D}_m\})}{N(\{\text{D}_m\})} \right) \frac{1}{i-1} + \frac{1}{n-1} \right) \frac{1}{n-1} \sum_{\text{all } i} \frac{d_i}{P(d, \{\text{D}_m\})}^2 - \frac{d_i}{P(d, \{\text{D}_m\})^2} \right) \]

\[
\sum_{\text{all } i} \left( \frac{d_i}{P(d, \{\text{D}_m\})} - d_i \right),
\]

thus, identifies the upper bound of the number of customers who consider more than one product. We can therefore express the lower bound of consumers who consider no alternative product as \(\sum_{\text{all } i} \left(2d_i - \frac{d_i}{P(d, \{\text{D}_m\})}\right)\). This bound is tight if few customers consider three products or more at the purchase time. In addition, one advantage of this bound is that it is derived without any assumption on the consideration process.

Fig. 4 uses our earlier example of three Adobe products to identify the lower bound of customers who consider no alternative products at purchase time. The result shows that the bound is 80.03%, very close to the percentage (80.35%) identified in Fig. 3.

4. Data and empirical results

4.1. Data

We collect data from Amazon’s “What Do Customers Ultimately Buy After Viewing This Item?” (see Fig. 1). The data are extracted using automated scripts to access and parse HTML pages from the retailer. For each product, Amazon provides a list of top four products that consumers purchase after viewing the focal product along with their respective purchase probability. In most cases, the same set of four products appears on each other’s ultimate purchase list. In a few cases, the products in the ultimate purchase lists do not overlap perfectly, in which case we construct the four-product group based on the procedure outlined in Appendix A. In total, we collected 38,400 unique products under Electronics category in 9600 product groups.

Our data were collected on May 15th, 2008. Table 1 lists summary statistics for our data. In Table 1, C1 refers to the conditional purchase percentage for the most purchased product after viewing the product page. In most cases, the most purchased product is the product being viewed. Similarly, C2 refers to the conditional purchase percentage of the second most purchased product after viewing the product page. C3 and C4 follow the same rationale for the third and the fourth most purchased products, respectively. The table shows that over 69% of consumers purchase the first product and less than 10% of customers purchase the third or fourth product.

4.2. Consideration set

We start our analysis by estimating the lower bound of the number of customers who view no alternative products. We conduct the aforementioned analysis for each group of four products. Fig. 5 shows the mean, standard deviation, and the histogram of the estimation. The results indicate that the average lower bound is 78% across all product groups. Given that the statistic is the lower bound, the actual percentage of customers who viewed no alternative products is likely to be higher. The histogram further shows significant variations across product groups. For some product groups, over 90% of consumers made purchase without viewing any other products; while for other product groups, most consumers view alternative products at the purchase time.

4.3. Consumer choice

Besides estimating consumer consideration sets at the purchase time, it is also useful to understand the consumer choice process when they consider more than one product. Prior research has found that consumers use different criteria for the consideration stage and for the choice stage [24]. The finding indicates that products offering high utilities might not be considered by many consumers. Our approach provides identification of consumer choice when the

### Table 1
Summary statistics of data.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs.</th>
<th>Mean</th>
<th>Std dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales rank</td>
<td>38,400</td>
<td>345.49</td>
<td>337.75</td>
<td>1</td>
<td>249,242</td>
</tr>
<tr>
<td>Rank 1 purchase propensity (C1)</td>
<td>38,400</td>
<td>69.90%</td>
<td>0.14</td>
<td>25%</td>
<td>99%</td>
</tr>
<tr>
<td>Rank 2 purchase propensity (C2)</td>
<td>38,400</td>
<td>12.28%</td>
<td>0.11</td>
<td>0.5%</td>
<td>45%</td>
</tr>
<tr>
<td>Rank 3 purchase propensity (C3)</td>
<td>38,400</td>
<td>5.42%</td>
<td>0.05</td>
<td>0.5%</td>
<td>32%</td>
</tr>
<tr>
<td>Rank 4 purchase propensity (C4)</td>
<td>38,400</td>
<td>3.28%</td>
<td>0.03</td>
<td>0.5%</td>
<td>17%</td>
</tr>
</tbody>
</table>

![Image](image_url)

A: Summary Statistics

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std dev</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>78.21%</td>
<td>21.73%</td>
<td>0.00%</td>
<td>98.46%</td>
</tr>
</tbody>
</table>

B: Histogram

![Image](image_url)

Fig. 4. Illustration of calculation of lower bound of percentage of consumers who consider no alternative products at purchase time.

Fig. 5. Descriptive statistics on lower bound of consumers who viewed no alternative products at purchase time.
identification of consideration sets is exact (i.e. in the case of two or three products). Given that Table 1 shows that most of the consideration is between the two leading products in each group, we choose to analyze consumer choices between the two leading products using the approach outlined in Section 3.2.

Panel A in Fig. 6 shows the summary statistics of the percentage of consumers who considered a given product in each group and the percentage of consumers who purchase a given product after viewing both products. The average consideration probability is 57%. Note that if each consumer considers only one product, the average consideration probability in a two-product group will be 50%, while if every consumer considers both products, the probability will be 100%. The result provides another indication that few consumers consider more than one product at the purchase time. Panel A also shows that there is a negative correlation between a product’s consideration probability and its choice probability when being considered along with an alternative product. The negative correlation suggests firms may differ in their business strategies — some firms focus on increasing the probability of their products being considered, while other firms focus on improving product value proposition relatively to its competitors [24]. Panel B of Fig. 6 shows the scatterplot of the relationship between consideration probability and choice probability. The x-axis represents the probability of a product being considered and the y-axis represents the probability of the product being chosen if compared side by side against the alternative product. The plot indicates that the relationship between the two probabilities vary substantially. For some products, their consideration probability is significantly higher than the choice probability, indicating that the product sales are mainly driven by them being frequently considered by consumers. For other products, their choice probability is significantly higher than the consideration probability, indicating that, while these products are less known among consumers, they offers higher values compared to the competitors.

4.4. The effects of consideration probability and choice probability on purchase

To compare the relative importance of consideration and choice on product sales, we note that there is a non-linear relationship among consideration, choice, and product sales. As such, we cannot use linear regressions or ANOVA for the analysis. Instead, to demonstrate the effects of consideration and choice on product purchase, we do so separately for the two effects. We first assess the influence of consideration by removing the influence of choice with the assumption that consumers have equal probability of choosing either product in the consideration set. We then calculate the predicted product sales and report the summary statistics of predicted product sales and its correlation with actual sales in Table 2. Since the percentage of variation expected in product sales equals to the square of the correlation, the result suggests that consideration alone explains 98% of the variation. We conduct the same analysis for the choice probability by assuming consumers give equal consideration to products. The result in Table 2 suggests that choice alone explains only 8.4% of the variation.

5. Conclusions

In this study, we developed a methodology to identify consumer consideration and choice at the product purchase time. We show that most consumers consider only one product at the purchase time. There are two possible explanations of the result: consumers either do not conduct product search before purchase, or they engage in pre-purchase information search and product comparison well before making the final purchase. Given that our analysis is conducted in the electronic category where products are relatively expensive, we believe the second explanation is likely to be true. This finding is also consistent with recent studies that suggest consumers conduct information search on third-party informediaries before making product purchase at online retailers [15].

This research shows that most consumers narrowed down their consideration set to one product at the time of final purchase and they are not influenced by information on alternative products in the purchase process. Our analysis also reveals that the majority of variations in product sales can be explained by heterogeneity in consumer consideration. Product utility has a limited impact.

Our analysis has a number of implications for product, price, and marketing strategies for retailers and manufacturers. The finding that many consumers consider only one product when making the final product purchase suggests that the provision of information on alternative products has limited influence on consumer purchase decision in late stages. Our findings also highlight the value of marketing effort to ensure products being considered by consumers.

By developing a model to measure product consideration and choice using only publicly available aggregate purchase statistics provided by online retailers, we also contribute to the literature from a methodological perspective. While prior two-stage consideration choice models require detailed individual level data and often impose significant restrictions on the underlying consumer behavior, we develop a methodology that shows consumer consideration and choice can be identified using a new type of aggregated information from online retailers with few assumptions on the underlying consumer behavior. This is particularly important for electronic commerce research given that clickstream data are not generally accessible.
Our study also presents a number of future research opportunities. First, we note that information on conditional product sales after viewing the focal product not only provides researchers an opportunity to identify the underlying consumer consideration set at the purchase time, but also provides information that could influence the consumer decision process itself. For example, if a retailer shows that most consumers purchase other products after viewing the focal product, it could have a significant impact on consumers’ purchase decisions. Therefore, consumer consideration process is a dynamic process influenced by prior consumer decisions. It will be valuable for future studies to model the dynamic aspect of the process [11]. Second, our methodology can only be used to identify consumer consideration of substitute products at the purchase time. In reality, consumer purchases not only substitute products but also complementary products. For example, consumers who bought cameras may also need battery, lenses, and memory card to complement the camera purchase. Online retailers such as Amazon provides a substantial amount of information on complementary products as well and it will be valuable to study how consumers form consideration set and make product choice for complementary products [12] and how online information influences the process.

Acknowledgement

We thank seminar participants at the 2011 International Conference on Information Systems (ICIS) for their valuable comments and suggestions. Prabhudev Konana acknowledges support from the National Science Foundation’s Information Technology Research Grant IIS-0218988. All errors remain ours.

Appendix A. Construction of four-product groups

1. For each product, Amazon identifies a list of top four products that consumers purchase after viewing the focal product (for parsimony, the list is called “ultimate purchase list” thereafter). Technically, it is possible for the focal product not to make to the list. In practice, the focal product is always one of the four products on the list.

2. For each of the other three products, we visit its product page and obtain its ultimate purchase list.

3. If the product sets of the four ultimate purchase lists are the same, the four products on the lists form a four-product group.

4. If the product sets of the four ultimate purchase lists are not the same, we identify the four products that appear most frequently on the four lists as the four-product group. We then take the following two steps to adjust the conditional purchase probability:

   a. If a product (say A) does not appear on the ultimate purchase list of another product (say B), we estimate the conditional purchase probability of purchasing A after viewing B as:

   \[ P(A|B) = 1 - \sum_{x \neq A} P(x|B) \]

   Please note that our underlying assumption is that all the unreported sales after viewing B are attributed to product A. This assumption is deliberately biased towards a larger consideration set since our goal is to identify the lower bound. It is also useful to note that P(A|B) is typically very small, thus our assumption has relatively little impact on the estimation of the lower bound. To test the robustness of our approach, we also conduct analysis with the assumption of P(A|B) = 0. The results are qualitatively the same.

   b. If a product (say E) appears on the ultimate purchase list of another product (say B) but is not in the four-product group, it is necessary to rescale purchase probabilities to fit our model, which assumes that consumers have purchased one of the four products. The rescaling is straightforward. To remove all the consumers who purchase E after viewing B and focus on only consumers who purchase products in the product group, we divide each conditional product purchase probability P(x|B) by (1 – P(E|B)).

References


Dr. Prabhudev Konana is Professor of Information Management, Distinguished Teaching Professor, William H. Sequy Centennial Professors in Business, and Assistant Director for Center for Research in Electronic Commerce (CREC) at the McCombs School of Business, the University of Texas at Austin. He received his MBA and Ph.D. in management information systems from the University of Arizona, Tucson in 1991 and 1995, respectively. He has an undergraduate degree in chemical engineering from Karnataka Regional Engineering College, India. His research interests are in business value of IT, virtual communities, outsourcing and offshoring. He is also interested in understanding the impact of IT on developing countries.

Dr. Michelle Chen is an assistant professor at the San Jose State University. She received her Ph.D. from the University of Texas at Austin. Her research focuses on virtual communities and social network, networked economy, electronic marketplaces and Internet marketing, data mining and business intelligence. Her work has appeared in Information Systems Research and others IS journals. Michelle’s research won the ISR Best Paper award in 2008.

Dr. Michelle Chen is an assistant professor at the San Jose State University. She received her Ph.D. from the University of Texas at Austin. Her research focuses on virtual communities and social network, networked economy, electronic marketplaces and Internet marketing, data mining and business intelligence. Her work has appeared in Information Systems Research and others IS journals. Michelle’s research won the ISR Best Paper award in 2008.