Portrait of an Online Shopper: Understanding and Predicting Consumer Behavior

Farshad Kooti, Kristina Lerman
USC Information Sciences Institute
Marina del Rey, CA
{kooti, lerman}@isi.edu

Luca Maria Aiello
Yahoo Labs
London, UK
alucca@yahoo-inc.com

Mihajlo Grbovic, Nemanja Djuric, Vladan Radosavljevic
Yahoo Labs
Sunnyvale, CA
{mihajlo, nemanja, vladan}@yahoo-inc.com

ABSTRACT

Consumer spending accounts for a large fraction of economic footprint of modern countries. Increasingly, consumer activity is moving to the web, where digital receipts of online purchases provide valuable data sources detailing consumer behavior. We consider such data extracted from emails and combined with with consumers' demographic information, which we use to characterize, model, and predict purchasing behavior. We analyze such behavior of consumers in different age and gender groups, and find interesting, actionable patterns that can be used to improve ad targeting systems. For example, we found that the amount of money spent on online purchases grows sharply with age, peaking in the late 30s, while shoppers from wealthy areas tend to purchase more expensive items and buy them more frequently. Furthermore, we look at the influence of social connections on purchasing habits, as well as at the temporal dynamics of online shopping where we discovered daily and weekly behavioral patterns. Finally, we build a model to predict when shoppers are most likely to make a purchase and how much will they spend, showing improvement over baseline approaches. The presented results paint a clear picture of a modern online shopper, and allow better understanding of consumer behavior that can help improve marketing efforts and make shopping more pleasant and efficient experience for online customers.

Categories and Subject Descriptors

H.4.3 [**Information Systems**]: Information systems applications

Keywords

Online shopping, demographics, prediction

1. INTRODUCTION

Consumer spending is an integral component of economic activity. In 2013, it accounted for 71% of the US gross domestic product (GDP)[1], a measure often used to quantify economic output and general prosperity of a country. Given its importance, many studies focused on understanding and characterizing consumer behavior. Researchers examined gender differences and motivations in shopping [10, 22], as well as spending patterns across urban areas [37].

In recent years, shopping has increasingly transitioned from in-store to an online experience. Consumers use internet to research product features, compare prices, and then purchase products from online merchants, such as Amazon or Walmart. Moreover, platforms like eBay allow people to directly sell products to one another. While there exist concerns about the risks and security of online shopping [7, 32, 42], large numbers of people, especially younger and wealthier [23, 39], choose online shopping even when similar products can be purchased offline [15]. The new habits of customers have had a tremendous economic impact on online market, with an estimated $1,471 billion dollars spent by 191 million shoppers in 2014 in the United States alone[2].

Most of the online purchases result in a confirmation or shipment email sent to the shopper by the merchant. These emails provide a rich source of evidence to study online consumer behavior across different shopping websites. Unlike previous studies [31], which were based on surveys and thus limited to relatively small populations, we used large-scale email data to perform an in-depth study of online shopping. More specifically, we extracted online receipts from 20.1M Yahoo Mail users, amounting to 121 million purchases worth 5.1B dollars. The extracted information included names of purchased products, their prices, and purchase timestamps. We used email user profile to link this information to demographic and geolocation data, such as gender, age, and zip code. This information enabled us to characterize patterns of online shopping activity and their dependence on demographic and socio-economic factors. We found that, for example, men on average make more purchases and spend more money on online purchases. Moreover, online shopping appears to be widely adopted by all ages and economic classes, although shoppers from high-income areas buy more expensive products than less wealthy shoppers.

Looking at temporal factors affecting online shopping, we found patterns common to other online activities as well [26]. Not surprisingly, online shopping has daily and weekly cycles showing that people fit online shopping routines into their everyday life. Furthermore, purchasing decisions appear to be correlated. The more expensive a previous purchase was, the longer the shopper has to wait until the next purchase. This can be explained by the fact that

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.

DOI: **http://dx.doi.org/10.1145/2835776.2835831**

[1] **https://research.stlouisfed.org/fred2/series/PCE/**

[2] **http://www.statista.com/topics/871/online-shopping/**

most shoppers have a finite budget and need to wait longer between purchases to buy more expensive items.

In addition to temporal and demographic factors, social networks are believed to play an important role in shaping consumer behavior (e.g., by spreading information about products through the word-of-mouth [33]). Previous studies examined how consumers use their online social networks to gather product information and recommendations [19, 20], although the direct effect of recommendations on purchases was found to be weak [27]. In addition, people who are socially connected are generally more similar to one another than unconnected people [29], and hence are more likely to be interested in similar products. Our analysis confirmed that shoppers who are socially connected and e-mail each other tend to purchase more similar products than unconnected shoppers.

Once we understand the factors affecting consumer behavior, we can use this knowledge to predict future purchases. Given users' purchase history and demographic data, we address a problem of predicting the time and price of their next purchase. Our method attains a relative improvement of 108.7% over the random baseline for predicting the price of the next purchase, and 36.4% relative improvement over the random baseline for predicting the time of the next purchase. Interestingly, demographic features were shown to be the least useful in these prediction tasks, while temporal features carried the most discriminative information.

The contributions of the paper are summarized below:

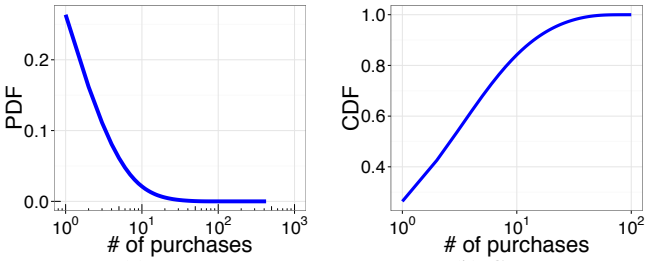
- In-depth analysis of a unique and very rich data set describing consumer behavior, extracted from purchase confirmations merchants send to shoppers (Section 2);
- A quantitative analysis of an impact of demographic, temporal, and social factors on consumer behavior (Section 3);
- Prediction of consumer behavior, where we predict the time of the next purchase and how much money will be spent (Section 4).

Better understanding of consumer behavior can benefit both consumers and advertisers. Knowing when consumers are ready to make a purchase and how much they are willing to spend can improve the effectiveness of advertising campaigns, in terms of optimizing ad impressions and budget spend. Understanding these patterns can also make online shopping experience more efficient for consumers. Considering that consumer spending presents such a large portion of the economy, even a small efficiency gain can have significant impact on the overall economic activity.

2. DATA SET

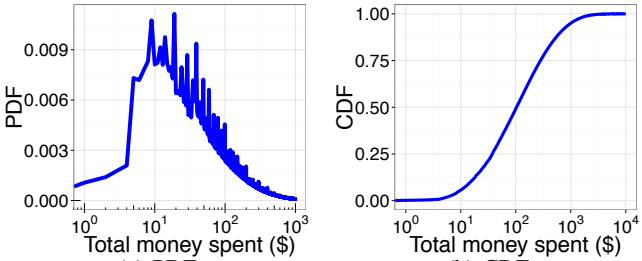
Most online purchases result in a confirmation email being sent by the merchant to the shopper. These emails provide a unique opportunity to study the shopping behavior of people across different online retail stores, such as Amazon, eBay, or Walmart.

Yahoo Mail is one of the world's largest email providers with more than 300M users[3], and many online shoppers use this email service for receiving purchase confirmations. We selected these emails by using a precompiled list of email addresses of popular merchants. Applying a set of manually specified extraction rules to the email body, we extracted the list of purchased *item names* and the *price* of each item. In case of multiple items purchased in a single order, we considered them as individual purchases occurring at the same time. Therefore, throughout the paper the expression "purchase" will refer to a purchase of a single item. In order to be able to analyze purchases by category (e.g., electronics, books,



(a) PDF (b) CDF

Figure 1: Distribution of number of purchases made by users



(a) PDF (b) CDF

Figure 2: Distribution of total money spent by users

handbags), we developed a categorization module based on the purchased item names. Specifically, an item name was used as an input to a classifier that predicts the *item category*. We used a three levels deep, 1,733 node Pricegrabber taxonomy[4] to categorize the items. The details of categorization are beyond the scope of this paper.

We limited our study to a random subset of Yahoo Mail users in the US. Our data set contains information on 20.1M users, who collectively made 121M purchases from several top retailers between February and September 2014, amounting to total spending of 5.1B dollars. For each user we extracted *age*, *gender*, and *zip code* information from the Yahoo user database. We excluded users who made more than 1,000 purchases (amounting to less than 0.01% of the sample), as these accounts are likely to belong to stores and not individual users. The analysis of the data set was performed in aggregate and on anonymized data.

In order to examine social aspects of the shopping behavior, we used the Yahoo email network data set in addition to the data set of purchases. The email network can be represented as a directed graph $\mathcal{G}$, with edges denoted by (i, j, N_{ij}) signifying that user i sent N_{ij} emails to user j. For our analysis we retained only edges with a minimum of 5 exchanged messages, and considered only a subgraph $\mathcal{C}$ of $\mathcal{G}$ induced by the two-hop neighborhood of the users who made purchases (i.e., their immediate contacts and contacts of their contacts). The subgraph $\mathcal{C}$ was used to construct a list of 1^{st}-level contacts and 2^{nd}-level contacts for each online shopper.

Let us take a closer look at the data set characteristics. In Figure 1(a) we show the distribution of number of purchases per user, which reveals the expected heavy-tailed characteristic. Figure 1(b) shows that only 5% of the users made more than 20 purchases. In contrast, the distribution of total money spent per user peaks at around 10 dollars, sharply decreasing for smaller amounts (Figure 2(a)). Also, there is a non-negligible minority of people who spend a substantial amount of money for online shopping, such as 5% of the users spent more than 1,000 dollars (Figure 2(b)).

Items being purchased also have drastically different levels of popularity, as shown by the distribution in Figure 3. Disney's Frozen

[3] `http://www.comscore.com/`

[4] `http://www.pricegrabber.com`

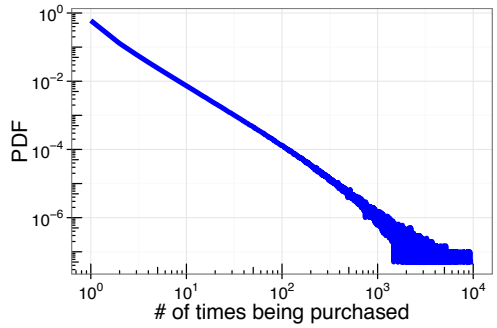


Figure 3: Number of times different items have been purchased

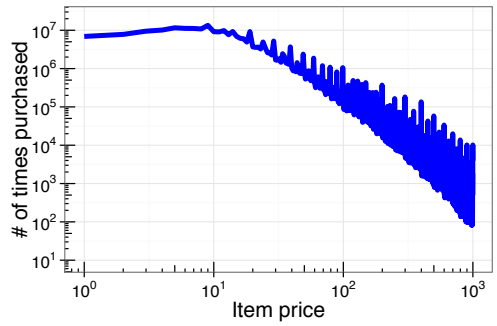


Figure 4: Number of item purchases as a function of price

DVD, the most popular item in the data set, has been purchased more than 200,000 times, whereas the vast majority of the items has been purchased less than 10 times. Table 1 lists the top 5 most frequently purchased items. Intuitively, the set of items that users have spent the most money on is a different set, because a single purchase of an expensive item would account for the same amount of money of several purchases of cheaper items (Table 2). In fact, the number of times an item is purchased negatively correlates with the price of that item (Figure 4). This is in line with previous survey-based studies [7] that found the vast majority of items purchased online are worth at most few tens of dollars.

3. PURCHASE PATTERN ANALYSIS

In this section we present a quantitative analysis of factors affecting online purchases. We examine the role of demographic, temporal, and social factors that include gender and age, daily and weekly patterns, frequency of shopping, tendency towards recurring purchases, and budget constraints.

3.1 Demographic Factors

Let us consider how gender, age, and location (i.e., zip code) affect purchasing behavior of customers. First, we measured fraction of all email users that made an online purchase. We found that higher fraction of women make online purchases compared to men (Figure 5(a)), albeit men make slightly more purchases (Figure 5(b)) and spend more money on average (Figure 5(c)). It is interesting that, as a result, men spend much more money in total (Figure 5(d)). The same patterns hold across different age groups. The presented results back up findings from earlier consumer surveys which revealed that man have a higher perceived advantage of online shopping [2], while women havie a higher concern for negative consequences of online purchasing [17], resulting in a higher number of purchases made by men.

Table 1: Top 5 most purchased products

Rank	Product name	# of purchases
1	Frozen (DVD)	202,103
2	Cards Against Humanity (Cards)	110,032
3	Google Chromecast	59,548
4	HDMI cable	54,402
5	Pampers	51,044

Table 2: Top 5 products with the most money spent on them

Rank	Product name	Money spent on product
1	Play Station 4	$ 7.0M
2	Frozen (DVD)	$ 3.8M
3	Kindle	$ 3.2M
4	Samsung Galaxy Tab	$ 2.7M
5	Cards Against Humanity	$ 2.5M

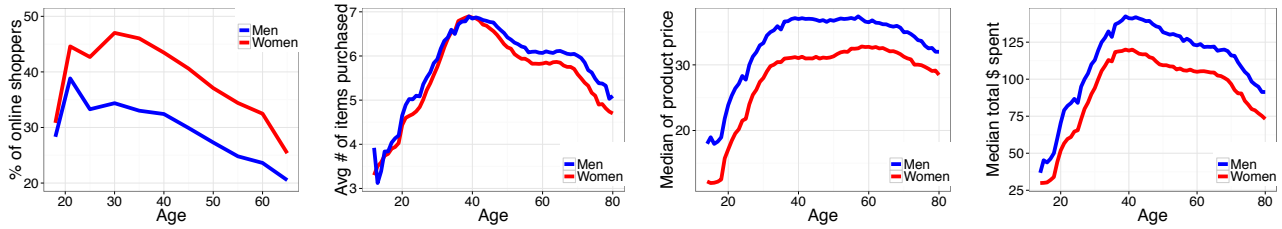
Table 3: Top product categories purchased by women and men

Rank	Top categories	Distinctive women	Distinctive men
1	Android	Books	Games
2	Accessories	Dresses	Flash memory
3	Books	Diapering	Light bulbs
4	Vitamins	Wallets	Accessories
5	Shirts	Bracelets	Batteries

With respect to the age, spending ability increases as people get older, peaking for the population between age 30 to 50 and then declining afterwards. The same pattern holds for number of purchases made, average item price, and total money spent (Figures 5(b), 5(d), 5(c)). Different generations also purchase different types of products online (Table 5). Younger shoppers (18-22 years old) purchase more phone accessories and games, whereas older shoppers (60-70 years old) are much more interested in buying TV shows. Also, blood sugar medicine is purchased more by the older users, which is expected.

Differences exist across genders as well. Table 3 shows the top five categories of purchased products for male and female customers. Even though the ranking of the top products is the same, each product accounts for different fraction of all purchases within the same gender. To find the most distinctive categories, we compare the fraction of all the items bought by both genders, and consider the categories that have the largest differences. Books, dresses, and diapering are the categories that were more disproportionately bought by women, whereas games, flash memory sticks, and accessories (e.g., headphones) are the categories purchased more by men. The largest differences range from only 0.5% to 1%, but are statistically significant. This result is aligned with previous research on offline shopping that found men more keen in buying electronics and entertainment products, while women more inclined to buy clothes [11, 22]. We repeated the same gender analysis at a product level (Table 4). Consistently, there is a high overlap between most purchased products.

In the following, we measure the impact of economic factors on online shopping behavior. We use the US census data to retrieve median income associated with each zip code, making an inferred income for a user an aggregated estimate. Nevertheless, given the large size of this data set this coarse approach was enough to observe clear trends. The number of purchases, average product



(a) Percentage of online shoppers (b) Number of purchases (c) Average price (d) Total money spent

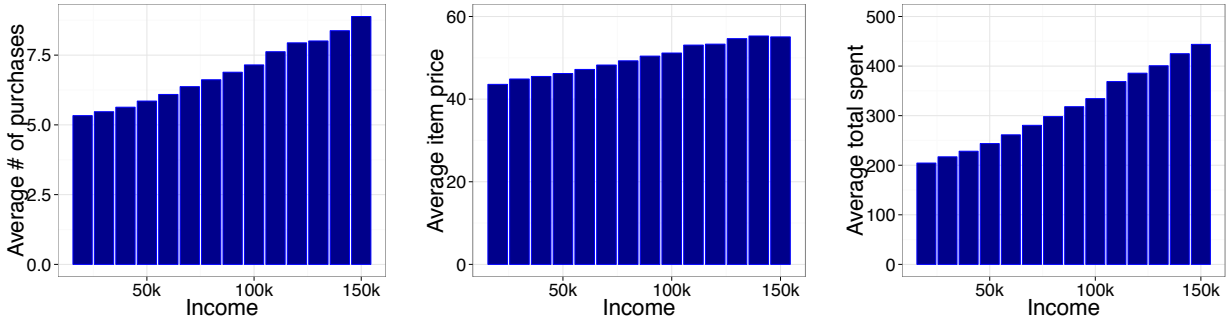
Figure 5: Demographic analysis broken down by age: (a) Percentage of online shoppers; (b) number of items purchased; (c) average price of products purchased; and (d) total spent by men and women

Table 4: Top products purchased by women and men

Rank	Top women	Top men	Distinctive women	Distinctive men
1	Frozen	Frozen	Frozen	Chromecast
2	iPhone screen protector	Game of Thrones	iPhone screen protector	Game of Thrones
3	Cards Against Humanity	Chromecast	iPhone screen protector	Titanfall Xbox One
4	iPhone screen protector (another brand)	Cards Against Humanity	iPhone case	Playstation 4
5	Game of Thrones	iPhone screen protector	iPhone case	Godfather collection

Table 5: Differences in the products purchased by younger (18-22 yo) and older (60-70 yo) users

Rank	Top younger users	Top older users	Distinctive younger users	Distinctive older users
1	iPhone screen protector	Frozen	iPhone screen protector	Frozen
2	iPhone screen protector	Game of Thrones	iPhone screen protector (another brand)	Game of Thrones
3	Cards Against Humanity	Chromecast	Cards Against Humanity	Downton Abbey
4	iPhone case	Downton Abbey	iPhone case	Blood sugar medicine
5	Frozen	Hunger Games	iPhone case	TurboTax Package



(a) Number of purchases (b) Average item price (c) Total money spent

Figure 6: Effect of income on purchasing behavior

price, and total money spent are all positively correlated with income (Figures 6(a), 6(b), and 6(c) respectively). While users living in high-income zip codes do not buy substantially more expensive products, they do make more purchases, spending more money in total than users from lower-income zip codes. Although the factors that lead to lower-income households spending less online are multiple and complex, part of this effect can be explained by the reluctance of people who are concerned with their financial safety to trust and make full use of online shopping, as pointed out by previous studies [23].

3.2 Temporal Factors

Our data set spans a period of eight months, giving us opportunity to investigate temporal dynamics of purchasing behavior and factors affecting it. Besides daily and weekly cycles and periodic purchasing, we observed temporal variations that we associate with financial depletion: users wait longer to buy more expensive items, waiting for the budget to recover from the previous purchases.

3.2.1 Daily and Weekly Cycles

Figure 7 shows the daily number of purchases over a period of two months. The figure indicates a clear weekly shopping pattern with more purchases taking place in the first days of the week and fewer purchases on the weekends. On average there are 32.6% more purchases on Mondays than Sundays. There also exists strong diurnal patterns in shopping behavior (Figure 8). Interestingly, most of the purchases occur during the working hours (i.e., in the morning and early afternoon). Note that for this analysis we in-

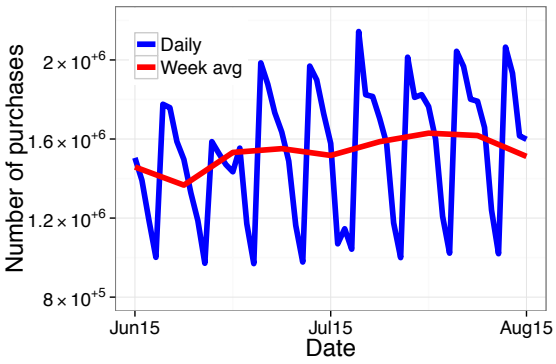


Figure 7: Number of purchases in a day and average weekly

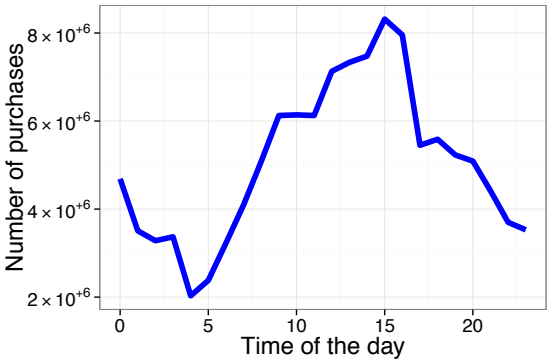


Figure 8: Number of purchases in each hour of the day

Table 6: Top 5 items with the most number of repurchases

Rank	Item name	Median purchase delay
1	Pampers 448 count	42
2	Bath tissue	62
3	Pampers 162 count	30
4	Pampers 152 count	31
5	Frozen	12

ferred the time zone from the user's zip code, which might be different from a shipping zip code for a purchase.

Researchers have also reported monthly effects, where people spend more money at the beginning of the month when they receive their paychecks, compared to the end of the month [14]. To test the *first-of-the-month* phenomenon, we compared spending in the first Monday of the month with the last Monday of the month. We considered the first and the last Mondays and not the first and the last days, because the strong weekly patterns would result in an unfair comparison if the first and the last day of the month are not the same day of the week. Our data does not support the earlier findings and there are months in which the last Monday of the month includes more activity compared to the first Monday of the month.

3.2.2 *Recurring Purchases*

Some products are purchased periodically, such as printer cartridges, water filters, or toilet paper. Finding these items and their typical cycle would help predicting purchasing behavior. We do this by counting the number of times each item has been purchased by each user, then from each user's count we eliminate those products purchased only once, and lastly we aggregate the number of

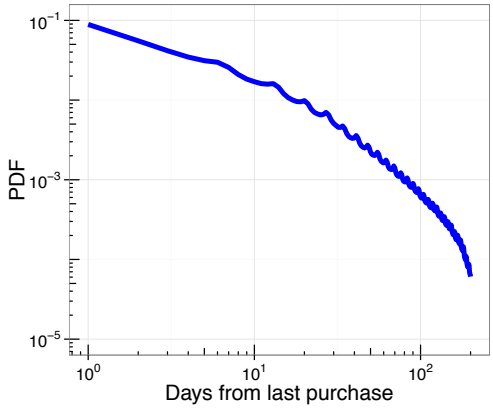


Figure 9: Distribution of number of days between purchases

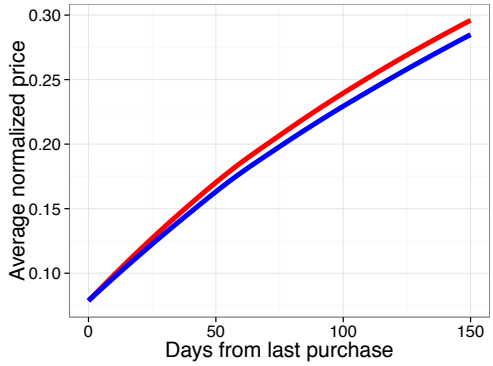


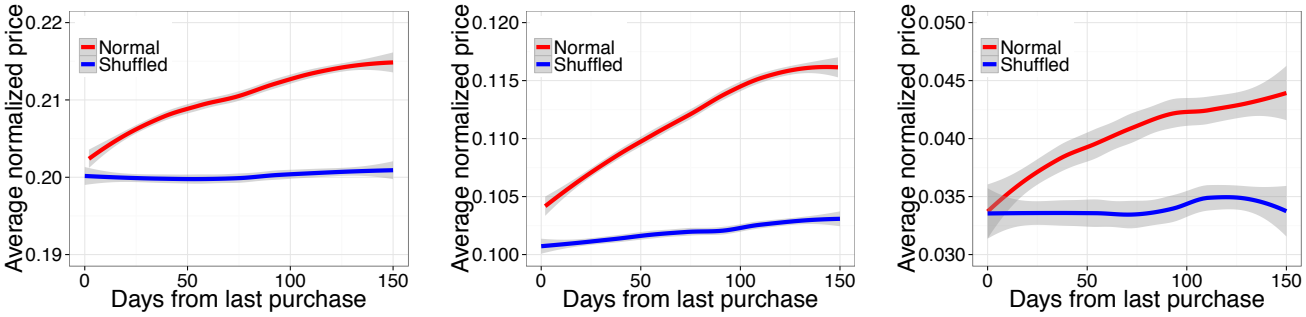
Figure 10: Relationship between purchase price and time to next purchase (0.95 confidence interval are shown yet too small to be observed)

purchases per each item. Table 6 shows the top five such products, along with the median number of days between purchases. Out of the top 20 products only four are neither toilet paper nor tissue (Frozen, Amazon gift card, chocolate chip cookie dough, and single serve coffees). In the top 20 list, the only unexpected item is the Frozen DVD, which probably made the list due to users buying additional copies as gifts or due to purchases by small stores that were not eliminated by our removal criterion of maximum 1,000 purchases. Interestingly, the number of days between purchases for most of the top 20 items is close to 1 or 2 months, which might be due to automatic purchasing that users can set up.

3.2.3 *Finite Budget Effects*

Finally, we study the dynamics of individual purchasing behavior. Figure 9 shows the distribution of number of days between purchases. The distribution is heavy-tailed, indicating bursty dynamics. The most likely time between purchases is one day and there are local maxima around multiples of 7 days, consistent with weekly cycles we observed.

An individual's purchasing decisions are not independent, but constrained by their finances. Budgetary constraints introduce temporal dependencies between purchases made by the user: After buying a product, the user has to wait some time to accumulate money to make another purchase. Previous work in economics studied models of budget allocation of households across different types of goods to maximize an utility function [13] and analyzed



(a) Users with 5 purchases (b) Users with 9-11 purchases (c) Users with 28-32 purchases

Figure 11: Relationship between purchase price and time to next purchase with 0.95 confidence interval

conditions under which people are willing to break their budget cap [8]. However, we are not aware of any study aimed to support the hypothesis of the time of purchase being partly driven by an underlying cyclic process of budget depletion and replenishment.

To test this hypothesis, we examined the relationship between purchase price and the time period since last purchase. Since different users have different spending power, we considered the normalized change in the price given the number of days from the last purchase. In other words, we computed how users divide their personal spending across different purchases, given the time delay between purchases. We then averaged the normalized values for all users, and report the change for each time delay. Figure 10 shows that as the time delay gets longer, users spend higher fraction of their budgets, which supports our hypothesis. To test that our analysis does not have any bias in the way the users are grouped, we perform a shuffle test by randomly swapping the prices of products purchased by users. This destroys the correlation between the time delay and product price. We then do the same analysis with the shuffled data and expect to see a flat line. However, the same increase also exists in the shuffled data, indicating a bias in the methodology. This is due to the heterogeneity of the underlying population: we are mixing users with different number of purchases. Users making more purchases have lower normalized prices and also shorter time delays, and are systematically overrepresented in the left side of the plot, even in the shuffled data.

To partially account for heterogeneity, we grouped users by the number of purchases (i.e., those who made exactly 5 purchases, those with 9-11 purchases, and 28-32 purchases). Even within each group there is variation as the total spending differs significantly across users, which we address by normalizing the product price by the total amount of money spent by the user, as explained above. If our hypothesis is correct, there should be a refractory period after a purchase, with users waiting longer to make a larger purchase. We clearly observe a positive relationship between (normalized) purchase price and the time (in days) since last purchase (Figure 11), but not in the shuffled data, which produces a horizontal line. We conclude that the relationship between time delay and purchase price arises due to behavioral factors, stemming from the limited budget of customers.

3.3 Social Factors

An individual's behavior is often correlated with that of his or her social contacts (or friends). In online shopping, this would result in users purchasing products that are similar to those purchased by their friends. Distinct social mechanisms give rise to this correlation [5]. First, a friend could influence the user to buy the same product by highly recommending it. This is the basis for social contagions in general, and "word-of-mouth" marketing in particular, although empirical evidence suggests that influence has a limited effect on shopping behavior [27]. Alternatively, users could have bought the same product as their friends purchased, because people tend to be similar to their friends, and therefore, have similar needs. The tendency of socially connected individuals to be similar is called homophily, and it is a strong organizing principle of social networks. Studies have found that people tend to interact with others who belong to a similar socio-economic class [16, 29] or share similar interests [25, 4]. Finally, a user's and their friend's behavior may be spuriously correlated because both depend on some other external factor, such as geographic proximity. In reality, all these effects are interconnected [9, 3] and are difficult to disentangle from observational data [35]. For example, homophily often results in selective exposure that may amplify social influence and word-of-mouth effects.

We investigate whether social correlations exist, although we do not resolve the source of the correlation. Specifically, we study whether users who are connected to each other via email interactions tend to purchase similar products in contrast to users who are not connected. To measure similarity of purchases between two users, we first describe the purchases made by each user with a vector of products, each entry containing the frequency of purchase. This approach results in large and sparse vectors due to the large number of unique products in our data set. To address this challenge, we use vectors of product categories, instead of product names. There are three levels of product categories, and we perform our experiments at all levels.

We compare similarity of category vectors of pairs of users who are directly connected in the email network (104K pairs of users) with the same number of pairs of randomly chosen users (who are not directly connected). We use cosine similarity to measure similarity of two vectors. Using top-level categories to describe user purchases gives average similarity of 0.420 between connected pairs of users, whereas random pairs have similarity of 0.377 on average (+11% relative change as compared to connected pairs). Using the more detailed level-2 categories to describe purchases gives average similarity of 0.215 for connected versus 0.170 for random pairs of users (+26% relative increase). Finally using the most detailed, level 3, categories results in average similarity of 0.188 for connected vs. 0.145 for random pairs of users (+30% relative increase). Although the absolute similarity decreases as a more detailed product vector is used, shoppers who communicate by email are always more similar than random shoppers who are not directly connected.

Gender also plays an important role when measuring purchase similarity between user pairs. To quantify this effect, we calculate

the cosine similarity between the vectors of number of purchases from the detailed category (level 3), and take the average of the cosine similarity. Instead of taking the average for all the connected pairs, we separate the pairs based on the gender of the users in the pair: woman-woman, man-man, and woman-man. The woman-woman pairs have the highest average cosine similarity with 0.192, next followed by man-man pairs with average similarity of 0.186. Heterogeneous pairs are the least similar ones, with average cosine similarity of 0.182. The similarity measures are still greater than measures for random pairs of users, which have similarity of 0.145. Woman-man pair having the smallest similarity primarily supports our earlier finding about a sensible difference in the type of goods that attract interests of the two genders. Previous work also found that receiving a shopping recommendation from a friend will have a greater positive effect on willingness to purchase online among women than among men [17]. The highest similarity of female-female pairs might be partly explained by that effect.

4. PREDICTING PURCHASES

Predicting the behavior of online shoppers can help e-commerce sites to improve the shopping experience by the means of personalized recommendations on one hand, and on the other to better meet merchants' needs by delivering targeted advertisements. In a recent study, Grbovic et al. addressed the problem of predicting the next item a user is going to purchase using a variety of features [50]. In this work, we consider the complementary problems of predicting *i)* the *time* of the next purchase, and *ii)* the *amount* that will be spent on that purchase.

Predicting the exact time and price of a purchase (e.g., using regression) is a hard problem, therefore we focused on the simpler classification task of predicting the class of the purchase among a finite number of predefined price or time intervals. We experimented with different classification algorithms and Bayesian Network Classification yielded the highest accuracy. To estimate the conditional probability distributions we used direct estimates of the conditional probability with $\alpha = 0.5$. The classifier was trained on the first six months of purchase data and evaluated on the last two. From each entry we extracted 55 features belonging to a variety of categories:

- *Demographics of online shoppers* (4 features): Gender, age, location (zip code), and income (based on zip code).
- *Purchase price history* (19 features): Price of the last three purchases, price category of the last three purchases, number of purchases, mean price of purchased item, median price of purchased items, total amount of money spent, standard deviation in item prices, number of earlier purchases in each price group (5 groups), price group with the most number of purchases and the count for it, and total number of purchases until that point.
- *Purchase time history* (13 features): Time of last three purchases, mean time between purchases, median time between purchased, standard deviation in times between purchases, number of earlier purchases in each time group, and time group with the most number of purchases and the count for it.
- *Purchase history of products* (4): Last three categories of products purchased, most purchased category.
- *Time or price of the next purchase* (1 feature): We also assume that we know when the next purchase is going to happen. This seems unrealistic at first, but we include this feature because the system is going to make recommendations at a given time, and we assume that the shopper is going to make the decision at that time. For having a symmetrical problem we also consider the price of the next purchase, which would be similar as knowing the budget of the user.
- *Contacts* (14 features): Mean, median, standard deviation, minimum, maximum, and 10th and 90th percentile of price and time of the purchases of the contacts of the users.

Table 7: Top predictive features for prediction of the price of the next item and their χ^2 values

Rank	Feature	χ^2 value
1	Most used class earlier	214,996
2	Number of under $6 purchases	115,560
3	Median price of earlier purchases	106,876
4	Mean price of earlier purchases	91,409
5	Number of over $40 purchases	84,743

For the aggregated features such as average price of item purchased, we used only purchases in the training period and did not consider future information. To evaluate the proposed approach, we compared results of our classifier to three baselines:

- Random prediction;
- Price or purchase delay of the previous purchase;
- Most popular price or purchase delay of the purchases a target user made in the past.

4.1 Price of the Next Purchase

We partition prices in five classes using $6, $12, $20, and $40 as price thresholds to obtain equally-sized partitions. These thresholds represent (a) very cheap products that cost less than $6 (20.7% of the data); (b) cheap products between $6 and $12 (20.3%); (c) medium-priced products between $12 and $20 (19.3%); (d) expensive products that cost more than $20, but less than $40 (19.9%); and finally (e) very expensive products worth more than $40 (19.8%). Our classifier achieves an accuracy of 43.2% with a +108.7% relative improvement over the 20.7% accuracy of the random classifier (i.e., relative size of the largest class).

A category of the last purchase and the most frequent purchase category turn out to be quite strong predictors, achieving accuracy of 29.3% and 29.8% by themselves, respectively. The supervised approach outperforms them with a +47.4% and +45.0% relative improvement, respectively. When measuring the predictive power of the features with the χ^2 statistics (Table 7) we find that the highest predictive power is the most frequent class of earlier purchases, by far. This suggests that users tend to buy mostly items in the same price bracket. The second feature in the ranking is the number of purchases from the very cheap category, followed by median and mean of earlier prices. In general, all the top 16 most informative features are related to the price of earlier purchases. After those, median time between purchases and time delay before the last purchase are the most predictive features. The relatively high position of the last time delay in the feature rank suggests that the recommender system should consider the time that has passed from the last purchase of the user, and change the suggestions dynamically. In other words, if the user has made a purchase recently, cheaper products should be favored over more expensive products to the user, whereas if a long period of time has passed since the last purchase, more expensive products should be advertised to the user, as they are more likely to be purchased. All of the demographics features have limited predictive power and are ranked last (though the demographics might affect the purchase history), with income being the most important among them.

Table 8: Top predictive features for prediction of time of next purchase and their χ^2 values

Rank	Feature	χ^2 value
1	Number of earlier purchases	48,719
2	Median time between purchases	35,558
3	Time since the first purchase	30,741
4	Previous time delay	30,692
5	Class of previous time delay	22,710

4.2 Time of the Next Purchase

Similarly to purchase price, prediction of purchase time could be leveraged to make a better use of the advertisement space. If the user is likely not to purchase anything for a certain period of time, ads can be momentarily suspended or replaced with ads that are not related to consumer goods.

For creating the categories, we choose thresholds of 1, 5, 14, and 33 days. Very short delays are within a day (22.8% of our data), short delays between 1 and 5 days (20.9%), medium delays between 5 and 14 (19.6%), long delays between 14 and 33 (18.2%) and the very long delays exceed 33 days (18.5%). Training a Bayesian Network on all the features yields an accuracy of 31.1%, a +36.4% relative improvement over the 22.8% accuracy of the random prediction baseline. The accuracy of our classifier is also +24.9% relatively higher than the baseline of predicting as the last purchase delay, which has accuracy of 24.9%. Finally, the most occurred purchase has an accuracy of 22.2%, which is outperformed by our classifier by +40.1% relatively.

Ranking features by their χ^2 (Table 8), we find that the most informative feature is the number of earlier purchases that the user has made so far, followed by median time delay, previous purchase delay, time since the first purchase, and the class of the previous purchase delay.

To summarize, we trained two classifiers for predicting the price and the time of the next purchase. Our algorithm outperformed the baselines in both prediction tasks, by a higher margin in case of predicting the price. Table 9 summarizes all of our results showing a relative improvement of 108.7% for predicting the price of the next item purchased and 36.4% for predicting the time of the next purchase over the majority vote baseline. Interestingly, user demographics were not particularly helpful for making any prediction, and the observed correlations in earlier sections of the paper are masked by other features such as the history of prior purchases.

5. RELATED WORK

Most of previous research on shopping behavior and characterization of shoppers has been conducted through interviews and questionnaires administered to groups of volunteers composed by at most few hundred members.

Offline shopping in physical stores has been studied in terms of the role of demographic factors on the attitude towards shopping. The customer's gender predicts to some extent the type of purchased goods, with men shopping more for groceries and electronics, while women more for clothing [11, 22]. Gender is also a discriminant factor with respect to the attitude towards financial practices, financial stress, and credit, and it can be a quite good predictor of spending [22]. Many shoppers express the need of alternating the experience of online and offline shopping [46, 41], and it has been found that there is an engagement spiral between online and offline shopping: searching for products online positively affects the frequency of shopping trips, which in its turn positively influences buying online [15].

Online shopping has been investigated since the early stages of the Web. Many studies tried to draw the profile of the typical online shopper. Online shoppers are younger, wealthier, more educated than the average Internet user. In addition, they tend to be computer literate and to live in the urban areas [47, 39, 40, 15]. Their trust of e-commerce sites and their understanding of the security risks associated with online payments positively correlate with their household income and education level [23, 24], and it tends to be stronger in males [17]. The perception of risk of online transactions influences shoppers to purchase small, cheap items rather than expensive objects [7]. Customers of online stores tend to value the convenience of online shopping in terms of ease of use, usefulness, enjoyment, and saving of time and effort [32]. Their shopping experience is deeply influenced by their personal traits (e.g., previous online shopping experiences, trust in online shopping) as well as other exogenous factors such as situational factors or product characteristics [32].

Demographic factors can influence the shopping behavior and the perception of the shopping experience online. Men value the practical advantages of online shopping more and consider a detailed product description and fair pricing significantly more important than women do. In contrast, some surveys have found that women, despite the ease of use of e-commerce sites, dislike more than men the lack of a physical experience of the shop and value more the visibility of wide selections of items rather than accurate product specifications [45, 2, 43, 24]. Unlike gender, the effect of age on the purchase behavior seems to be minor, with older people searching less for online items to buy but not exhibiting lower purchase frequency [38]. With extensive evidence from a large-scale dataset we find that age greatly impacts the amount of money spent online and the number of items purchased.

The role of the social network is also a crucial factor that steers customer behavior during online shopping. Often, social media is used to communicate purchase intents, which can be automatically detected with text analysis [20]. Also, social ties allow for the propagation of information about effective shopping practices, such as finding the most convenient store to buy from [19] or recommending what to buy next [27]. Survey-based studies have found that shopping recommendations can increase the willingness of buying among women rather than men [17].

Factors leading to purchases in offline stores have been extensively investigated as they have direct consequences on the revenue potential of retailers and advertisers. Survey-based studies attempted to isolate the factors that lead a customer to buy an item or, in other words, to understand what the strongest predictors of a purchase are. Although the mere amount of online activity of a customer can predict to some extent the occurrence of a future purchase [6], multifaceted predictive models have been proposed in the past. Features related to the phase of information gathering (access to search features, prior trust of the website) and to the purchase potential (monetary resources, product value) can often predict whether a given item will be purchased or not [21, 31].

Prediction of purchases in online shopping is a task that has been addressed through data-driven studies, mostly on click and activity logs. User purchase history is extensively used by e-commerce websites to recommend relevant products to their users [28]. Features derived from user events collected by publishers and shopping sites are often used in predicting the user's propensity to click or purchase [12]. For example, clickstream data have been used to predict the next purchased item [44, 34]; click models predict online buying by linking the purchase decision to what users are exposed to while on the site and what actions they perform while on the site [30, 36]. Besides user click and purchase events, one can

Table 9: Summary of the prediction results. Accuracy: percentage of correctly classified samples. Majority vote: always predicting the largest group, or predicting randomly. Most used: the group the user had the most in earlier purchases. AUC: Weighted average of Area Under the Curve for classes. RMSE: Root Mean Square Error. The improvements are reported over the majority vote.

Prediction	Majority vote (random classifier)	Last used	Most used	Our classifier	Absolute improvement	Relative improvement	AUC	RMSE
Item price	20.7%	29.3%	29.8%	43.2%	22.5%	108.7%	0.676	0.3806
Purchase time	22.8%	24.9%	22.2%	31.1%	8.3%	36.4%	0.634	0.4272

leverage product reviews and ratings to find relationships between different products [49]. Email is also a valuable source of information to analyze and predict user shopping behavior [50]. Click and browsing features represent only a weak proxy of user's purchase intent, while email purchase receipts convey a much stronger intent signal that can enable advertisers to reach their audience. The value of commercial email data has been recently explored for the task of clustering commercial domains [18]. Signals to predict purchases can be strengthened by demographic features [1]. Also, information extracted from customers' profiles in social media, in combination with the information of their social circles, can help with predicting the product category that will be purchased next [48].

6. CONCLUSION

Studying the online consumer behavior as recorded by email traces allows to overcome the limitations of previous studies that focused either on small-scale surveys or on purchases' logs from individual vendors. In this work, we provide the first very large scale analysis of user shopping profiles across several vendors and over a long time span. We measured the effect of age and gender, finding that the spending ability goes up with age till the age of 30, stabilizes in the early 60s, and then starts dropping afterwards. Regarding the gender, a female email user is more likely to be an online shopper than an average male email user. On the other hand, men make more purchases, buy more expensive products on average, and spend more money. Younger users tend to buy more phone accessories compared to older users, whereas older users buy TV shows and vitamins & supplements more frequently. Using the user location, we show clear correlation between income and the number of purchases users make, average price of products purchased, and total money spent. Moreover, we study the cyclic behavior of users, finding weekly patters where purchases are more likely to occur early in the week and much less frequently in the weekends. Also, most of the purchases happen during the work hours, morning till early afternoon.

We complement the purchase activity with the network of email communication between users. Using the network, we test if users that communicate with each other have more similar purchases compared to a random set of users, and we find indeed that is the case. We also consider gender of the users and find that woman-woman pairs are more similar than man-man pairs that are also more similar to each other than the woman-man pairs. Finally, we use our findings to build a classifier to predict the price and the time of the next purchase. Our classifier outperforms the baselines, especially for the prediction of the price of the next purchase. This classifier can be used to make better recommendations to the users.

Our study comes with a few limitations. First, we can only capture purchases for which a confirmation email has been delivered. We believe this is the case for most of online purchases nowadays. Second, if users use different email addresses for their purchases, we would not have their full purchase history. Similarly, people can share a purchasing account to enjoy some benefits (e.g., an Amazon Prime account between multiple people). However, as suggested by the fact that less than 0.01% of the users have goods shipped to more than one zip-code, that occurs rarely in our data set. Third, the social network that we considered, albeit big, is not complete. However, the network is large enough to observe statistically significant results. Lastly, we considered the items that were purchased together as separate purchases; it would be interesting to see which items are usually bought together in the same transaction.

Acknowledgements

This work was supported in part by AFOSR under contract FA9550-10-1-0569, and by DARPA under contract W911NF-12-1-0034.

7. REFERENCES

[1] Combination of multiple classifiers for the customer's purchase behavior prediction. *Decision Support Systems*, 34(2):167 – 175, 2003.

[2] Understanding gender-based differences in consumer e-commerce adoption. *Commun. Association for Information Systems*, 26, 2005.

[3] L. M. Aiello, A. Barrat, C. Cattuto, G. Ruffo, and R. Schifanella. Link creation and profile alignment in the aNobii social network. In *SocialCom*, 2010.

[4] L. M. Aiello, A. Barrat, R. Schifanella, C. Cattuto, B. Markines, and F. Menczer. Friendship prediction and homophily in social media. *ACM Trans. Web*, 6(2):9:1–9:33, Jun 2012.

[5] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, pp. 7–15, 2008.

[6] S. Bellman, G. L. Lohse, and E. J. Johnson. Predictors of online buying behavior. *Commun. ACM*, 42(12):32–38, 1999.

[7] A. Bhatnagar, S. Misra, and H. R. Rao. On risk, convenience, and internet shopping behavior. *Commun. ACM*, 43(11):98–105, Nov. 2000.

[8] J.-S. Chiou and C.-C. Ting. Will you spend more money and time on internet shopping when the product and situation are right? *Computers in Human Behavior*, 27(1):203 – 208, 2011.

[9] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD'08,* pp. 160–168, 2008.

[10] R. R. Dholakia. Going shopping: key determinants of shopping behaviors and motivations. *Int. J. Retail & Distribution Management*, 27(4):154–165, 1999.

[11] R. R. Dholakia. Going shopping: key determinants of shopping behaviors and motivations. *Int. J. Retail and Distribution Management*, 27:154–165, 1999.

[12] N. Djuric, V. Radosavljevic, M. Grbovic, and N. Bhamidipati. Hidden conditional random fields with deep user embeddings for ad targeting. In *ICDM*, pp. 779–784, 2014.

[13] R. Y. Du and W. A. Kamakura. Where did all that money go? understanding how consumers allocate their consumption

budget. *J. Marketing*, 72(6):109–131, 2008.

[14] W. N. Evans and T. J. Moore. Liquidity, economic activity, and mortality. *Review of Economics and Statistics*, 94(2):400–418, Jan. 2011.

[15] S. Farag, T. Schwanen, M. Dijst, and J. Faber. Shopping online and/or in-store? a structural equation model of the relationships between e-shopping and in-store shopping. *Transportation Research Part A: Policy and Practice*, 41(2):125–141, 2007.

[16] S. L. Feld. The focused organization of social ties. *The American Journal of Sociology*, 86(5):1015–1035, 1981.

[17] E. Garbarino and M. Strahilevitz. Gender differences in the perceived risk of buying online and the effects of receiving a site recommendation. *J. Business Research*, 57(7):768 – 775, 2004.

[18] M. Grbovic and S. Vucetic. Generating ad targeting rules using sparse principal component analysis with constraints. In *WWW Companion*, pp. 283–284, 2014.

[19] S. Guo, M. Wang, and J. Leskovec. The role of social networks in online shopping: Information passing, price of trust, and consumer choice. In *EC*, pp. 157–166, 2011.

[20] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa. Identifying purchase intent from social posts. In *ICWSM*, 2014.

[21] T. Hansen, J. Moller Jensen, and H. Stubbe Solgaard. Predicting online grocery buying intention: a comparison of the theory of reasoned action and the theory of planned behavior. *Int. J. Information Management*, 24(6):539–550, 2004.

[22] C. R. Hayhoe, L. J. Leach, P. R. Turner, M. J. Bruin, and F. C. Lawrence. Differences in spending habits and credit use of college students. *J. Consumer Affairs*, 34(1):113–133, 2000.

[23] J. A. Horrigan. Online shopping. *Pew Internet & American Life Project Report*, 36, 2008.

[24] T.-K. Hui and D. Wan. Factors affecting internet shopping behaviour in singapore: gender and educational issues. *Int. J. Consumer Studies*, 31(3):310–316, 2007.

[25] J.-H. Kang and K. Lerman. Using lists to measure homophily on twitter. In *AAAI workshop on Intelligent Techniques for Web Personalization and Recommendation*, July 2012.

[26] F. Kooti, L. M. Aiello, M. Grbovic, K. Lerman, and A. Mantrach. Evolution of conversations in the age of email overload. In *WWW*, pp. 603–613, 2015.

[27] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), May 2007.

[28] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, Jan. 2003.

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[30] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Predicting online purchase conversion using web path analysis. Technical report, 2002.

[31] P. A. Pavlou and M. Fygenson. Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS quarterly*, pp. 115–143, 2006.

[32] T. Perea y Monsuwé, B. G. Dellaert, and K. De Ruyter. What drives consumers to shop online? a literature review. *Int. J. Service Industry Management*, 15(1):102–121, 2004.

[33] T. Rodrigues, F. Benevenuto, M. Cha, K. Gummadi, and V. Almeida. On word-of-mouth based discovery of the web. In *IMC*, pp. 381–396, 2011.

[34] S. Senecal, P. J. Kalczynski, and J. Nantel. Consumers' decision-making process and their online shopping behavior: a clickstream analysis. *J. Business Research*, 58(11):1599–1608, 2005.

[35] C. R. Shalizi and A. C. Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

[36] C. Sismeiro and R. E. Bucklin. Modeling purchase behavior at an e-commerce web site: A task-completion approach. *J. Marketing Research*, 41(3):306–323, 2004.

[37] S. Sobolevsky, I. Sitko, R. Tachet des Combes, B. Hawelka, J. Murillo Arias, and C. Ratti. Cities through the prism of people's spending behavior. *arXiv*, 2015.

[38] P. Sorce, V. Perotti, and S. Widrick. Attitude and age differences in online buying. *Int. J. Retail & Distribution Management*, 33(2):122–132, 2005.

[39] W. R. Swinyard and S. M. Smith. Why people (don't) shop online: A lifestyle study of the internet consumer. *Psychology & Marketing*, 20(7):567, 2003.

[40] W. R. Swinyard and S. M. Smith. Activities, interests, and opinions of online shoppers and non-shoppers. *International Business and Economics Research Journal*, 3(4):37–48, 2011.

[41] M. Tabatabaei. Online shopping perceptions of offline shoppers. *Issues in Information Systems*, 10(2):22–26, 2009.

[42] T. S. Teo. Attitudes toward online shopping and the internet. *Behaviour & Information Technology*, 21(4):259–271, 2002.

[43] F. Ulbrich, T. Christensen, and L. Stankus. Gender-specific on-line shopping preferences. *Electronic Commerce Research*, 11(2):181–199, 2011.

[44] D. Van den Poel and W. Buckinx. Predicting online-purchasing behaviour. *European Journal of Operational Research*, 166(2):557 – 575, 2005.

[45] C. Van Slyke, C. L. Comunale, and F. Belanger. Gender differences in perceptions of web-based shopping. *Commun. ACM*, 45(8):82–86, Aug. 2002.

[46] M. Wolfinbarger and M. C. Gilly. Shopping online for freedom, control, and fun. *California Management Review*, 43(2):34–55, 2001.

[47] M. Zaman and M. Y. W. Meng. Internet shopping adoption: A comparative study on city and regional consumers. In *ANZMAC*, pp. 2421–2428. Deakin University, 2002.

[48] Y. Zhang and M. Pennacchiotti. Predicting purchase behaviors from social media. In *WWW*, pp. 1521–1532, 2013.

[49] J. McAuley, R. Pandey, J. Leskovec. Inferring networks of substitutable and complementary products. In *ACM SIGKDD*, pp. 785–794, 2015.

[50] M. Grbovic, V. Radosavljevic, N. Djuric, N. Bhamidipati, J. Savla, V. Bhagwan, and D. Sharp E-commerce in Your Inbox: Product Recommendations at Scale. In *ACM SIGKDD*, pp. 1809–1818, 2015.