A Measurement Study on Amazon Wishlist and Its Privacy Exposure

***Abstract*—Online user behavior has been largely studied in various fields due to the prosperity of Internet. In this paper, we investigate Amazon wishlists, where users keep their desired products for easier access or providing guidelines to gift givers. We collected complete Wishlists of over 30,000 users, by analyzing which we are able to reveal interesting observations regarding to user online shopping behaviors in multiple dimensions. Specifically, we show user behavior from different demographical groups, including gender and geo-locations. We found that males and females have different shopping pattern in terms of price and product types. Besides, we also take timing factors into consideration. Surprisingly we found that unlike traditional walk-in-shop type of shopping, user may not always like to shop online during holidays or weekends. Then we investigate user information exposure in Amazon wishlists. By analyzing human languages in list-descriptions, we illustrate what and to what extent are user personal information exposed to the public. Finally, we demonstrate that information in Wishlists has potential to leak personal information that users did not choose to put on Amazon. We use Support-Vector-Machine (SVM) to train and test the data. Our result indicates that based solely on the items in users' Wishlists, We can predict user gender with over 80% accuracy.**

I. Introduction

Entering the era of big data, Internet has been instantly extended to be data-rich. User data can be exploited to generate large benefit in many fields. Electronic commerce is one of the most data-driven market. There have been many works studying how users behave or user reactions in E-commercials [6], [9], [17]. Studying user behaviors in e-commercials help understanding the market and thus generate better economic outcomes. For example, user data can be monetized by feeding targeted advertising or performing price discrimination [13]. Although there are various kinds of user data on Amazon, user shopping history and preference is one of the most sensitive and valuable type of data. However, shopping history is considered private information that most users do not wish to publicize. To cope with user expectations, most e-commerce companies keep user shopping history private. Therefore, a major challenge of studying shopping preference is to collect data from users. Nonetheless, users may not always hide their purchase intentions. Wishlist, a list-type data in Amazon, is made publicly available by default. Users add items in their wishlists to record desired products for their own reference or for gift givers. As users use wishlists to record desired products, wishlists largely reflect the shopping preference of users – If a user adds a item in his/her wishlist, the user is likely to buy the item or make other people buy the item. Besides, wishlists are also indicators of shopping history since the items will not be removed after the item is bought unless manually deleted.

As the world's leading e-commerce company, Amazon is a particularly important source for understanding user shopping behaviors. In this paper, we analyze wishlists in Amazon to study user shopping patterns as well as the privacy implication of wishlists. By doing so we shed light on general user preference on e-commerce and help companies to refine both their marketing strategies and privacy policy.

We first collected complete profile and wishlists information of over 30,000 users and approximately 2 million items in Amazon by web scraping. Based on the data, we conduct measurement study on user behaviors. Our study concentrates on user shopping preference, which are organized in 3 dimensions 1)Product Categories 2)Product prices 3) Timing. Specifically, we compare the user preference in different gender and regions. We also investigate shopping pattern in different time through a year. Surprisingly we found that there is little shopping increase in holidays or weekends and even some holidays are shopping repelling. We believe our findings provide insight that is able to help sellers to revise their marketing strategies, as well as to help advertisers to make more accurate targeted advertising.

Beside user behavior analysis, we study user privacy information exposure in Amazon. Network users are prone to expose their private information in public websites inadvertently [4], [5], making themselves face the threat of information leakage. Privacy protection is very important for websites not only because it involves legal issues but also that it correlates to user purchasing intentions [1], [17]. To study to what extent does the information in wishlists threat user privacy, we investigate both the items in wishlists and the user input list-descriptions to identify user personal information. We first illustrate user personal information exposure by analyzing their list-descriptions, showing that users are mentioning sensitive personal information such as profession, education background, relatives' information, etc in their list-descriptions.

Furthermore, public information may be exploited to infer user personal information. Such information leakage has been proved in many works [2], [7], [8], [14], [19]. In our study, we also try to identify user personal information leakage through analyzing publicly available data. Specifically, we use Support Vector Machine (SVM) to predict user gender based solely on items in their wishlists. The result indicates that user gender information can be identified with fairly high accuracy.

II. Data and Methodology

As the largest electronic retailer in the U.S, Amazon was reported to have 270 active customer accounts in 2014 [1]. With

[1] http://www.statista.com/statistics/237810/number-of-active-amazon-customer-accounts-worldwide

such large user base, user behavior in Amazon significantly imply the market pattern and preference, which is critical for electronic commercials and advertisement ecosystem. However, such online user behavior has not been systematically studied before. In this paper we unveil customer shopping pattern in Amazon through analyzing Wishlists of a large amount of users.

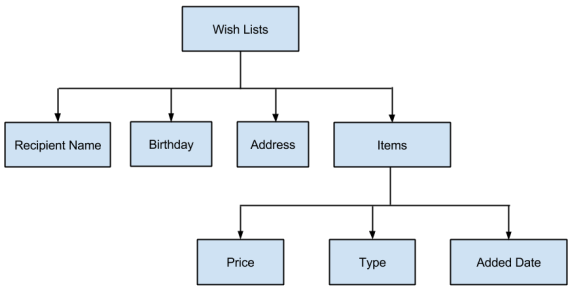
Every registered user in Amazon has a public profile that contains various objects such as profile photo, rated items, reviews, and Wishlists, etc. Wishlist is a list-type object that contains the desired products added by the user. A user may have multiple Wishlists to accommodate different types of products. Wishlists can be useful in many ways. For examples, a user may use Wishlists to record the products he/she wishes to buy in the future. Or the user may use the Wishlist to show guidelines for gift givers. Beside items, each wishlist maintains a separate recipient profile, including name, birthday, and shipping address. When a new Wishlist for the user is created, birthday and address are defaulted to be empty, and recipient name is pre-set as the user's profile name. To preserve user privacy, birthday has only month and day in a Wishlist, thus the age of the recipient is unknown. Similarly, address information shows only the state and city. Note that although every user has a public profile, not all objects in this profile are public. Wishlists are configurable yet public in default. We focus on the analysis of user Wishlists. Therefore, other objects that are beyond the scope of this paper such as reviews, activities are deliberately omitted.

Figure 1 shows the data hierarchy in our measurement. We illustrate only the layers under the Wishlist. As we can see from the figure, other than items, each list has its own recipient name, birthday, and address. Beside these formated information, each Wishlist can have a list-description. List-description is plain text keyed in by the users, which is which is usually used to briefly introduce the wishlists and the users. Examples are *"I LOVE music! Buy me a CD!"* or *"Son of Sue and Kevin, brother of Roger, husband of Kathy. I moved from Milton Keynes, England to Smithfield, North Carolina, USA on 8/8/2000 and recently moved from there to the Raleigh/Garner border in the same state"*. The first example list-description shows the hobby of a user and the second example exposes much more personal information such as parents' names, marriage status, wife name, moving history, etc.

As the Wishlist stores the user desired products, it directly reflect the user online shopping pattern in Amazon. Particularly we are interested in the products users prefer to buy, the time when shopping peak or pit happens, and the price users are willing to pay. Therefore we need to collect adequate number of data from Amazon for analysis.

A. Data Collection Methodology

One way to collect data from Amazon is to use its Product Advertising API[]. However, the API does not provide wishlist or product type accesses, which are essential in our study. Besides, there is a 1 request/second limit on non-profiting API users (The actual rate is $1 + round(\frac{S}{\$4600})$[], in which S denotes the sales in the user's website in last 30 days). Using few API accounts does not boost the speed while signing up



.

Fig. 1: data hierarchy

.

for more accounts require much more effort. Therefore we opt for crawling Amazon by web scraping. Although web scraping may not be reliable, it provides all information that are needed. We implement the crawler using python package BeautifulSoup [] to extract specific data in certain HTML tags.

Generally the data collection process consists of 3 steps. In each step we store certain data and collect input data for the next step.

First we aim to collect substantial amount of user profiles to expand our crawling targets. From the user profile, we are able to extract the user name, birthday, address, Wishlist names, list URLs, and Wishlist list-descriptions. To this end, we leverage Amazon wishlist search engine [2] to search common names. The wishlist search engine will return at most 2016 users and that are associated with a name. [3] We have mentioned that each Wishlist under a user profile may have different personal information such as name, address, etc. The search engine returns the user name, birthday, address, and list-description as the ones in the latest updated Wishlist. Figure 2 shows the result of a typical search, in which we searched user named "Paul".

Then we further investigate the items in the Wishlist. We directly visit the Wishlist URLs extracted in last step. In the page of Wishlist as shown in Figure 3, the items are listed with links to their own pages. We cannot know the price and type of the item until we visit these pages. Therefore to this point we are only able to collect item name, item page URL and the date the items were added.

Finally We retrieve detailed information for each item. We visit all the item URLs and download the product pages. One example of the product web pages is shown in Figure **??**. In a page like this, the item type and the item price (In red box) can be easily extracted. However, there may not always be only one price for an item. For example, for the same item, there can be prices from different retailers. There are also differences

[2]http://www.amazon.com/gp/registry/search

[3]The search engine usually states that hundreds of thousand users are found, but only the first 2016 users are viewable.

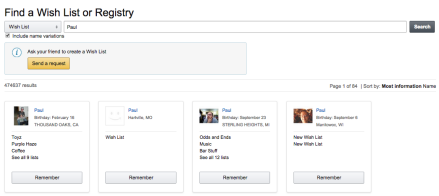


Fig. 2: Wishlist Search Engine

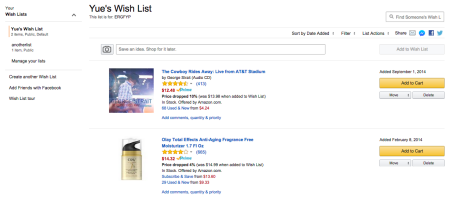


Fig. 3: Wishlist

Fig. 4: Item

between new ones and used ones. We record only the price that the wishlist owner chose (Wishlists remember which version of the items are selected in most cases). In cases which the user does not choose a price, we record the lowest prices that a user needs to pay to get an item. In these rare case we believe it is reasonable to choose the lowest price because it is natural to assume that people prefer to pay less to purchase an item. Note that we cannot always find the price for an item. There are multiple reasons – (1) The item is removed from Amazon so there is no web page for it. However, it still stays in user Wishlists. (2) The item is no longer in stock, the price shown will be "Currently unavailable" (3) Web failures and anti-crawler mechanisms.

B. Data Overview

We searched the top 300 common male and female names [4] to harvest user profiles. Eventually we collected 1,233,095 unique users.Their profile information and wishlist links are stored in our database. However, collecting Wishlists of all the users is very time-consuming (Consider that one Amazon product Page usually has over 10 thousand lines of code, which is around 300KB data). Therefore we collect only part of the user profile pool that is large enough to conduct sound measurement. As we are also interested in personal information of a user, we collect the wishlists and items of users who potentially have input personal information – users with list-description. To this end, we collected 30,057 complete user, together with all the items and wishlists. In total we collected 76,923 wishlists and 5,710,674 items, among which 2,248,142 are unique. The size of the data is approximately 1.6GB.

C. Personal Information

Although we did not collect wishlists for all the recorded users. We can still use their wishlist profile to study personal information exposure in Amazon especially when we are particularly interested in potential user information leakage. Table I illustrates information exposure in user list profiles. We found that a considerable number of users put their birthday and location information in their list profiles. Furthermore, most of the people who have a list-description have exposed their birthday and address information. Our findings agree with [4], which states that users tend to expose their personal information in open websites.

[4]http://names.mongabay.com/

TABLE I: Personal Information in User Profile

Personal Info	User Number	Percentage
Birthday	280,328	29.0%
Location	221,298	22.9%
Birthday & Location	150,004	15.5%
List-description	104,846	10.8%
List-description & Birthday	94,284	9.7%
List-description & Location	59,731	6.2%

TABLE II: Distributions

Distribution	Mean	Max	SD	γ_1	κ
Wishlist in Profile	2.56	326	4.97	19.7	854.15
Items in Wishlist	74.2	7,350	187.3	8.25	111.11
Item Price	35.06	105,065	172.0	299.4	144,743.2

III. Data Analysis

Now we aim to dissect our dataset to reveal how users use their wishlists. Particularly we are interested in the products user prefer to buy, as well as the time that is appealing to shoppers. We compare user preferences in 3 different regions to show market trend. Besides, we also quantify the increase in different time of a year. Specifically we show different shopping behaviors in various holidays.

A. Basic Statistics

First we conduct fundamental measurements on the dataset to help build basic understanding on how wishlists are used and to what extend the users expose their personal information in Amazon.

For all 967,603 users collected in the first step of data collection, we record totally 2,121,173 wishlists and 5,700,000 items (in which 2,248,142 items are unique). We show distribution metrics in Table II. As we can see from the table, every user has 2.56 wishlists in average with standard deviation of 4.97. The average number and standard deviation of items in a wishlist is 74.2 and 187.3 accordingly. And the average and standard deviation of product price is $35.06 and $172.0. Note that although we list the maximum number of wishlists a user has (Similarly, items in wishlists and item price) in our dataset, we did not probe the limit set by Amazon. All three distributions have very high skewness (γ_1) and kurtosis (κ), which means the distributions are very skewed and heavily tailed. For clearer presentation, we show the distribution of wishlist in Figure 5, distribution of items in Figure 6, and distribution of price in Figure 7 using log-scaled y axis.

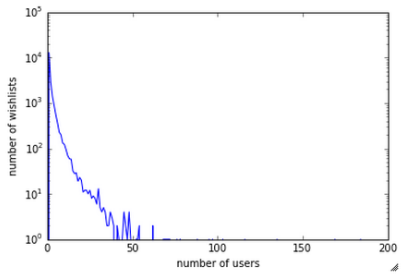


Fig. 5: Number of lists the users have

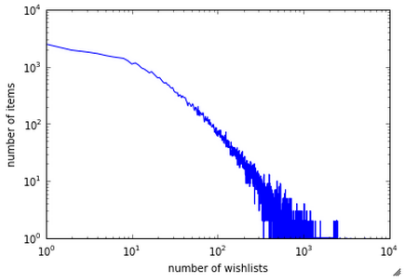


Fig. 6: Number of items the lists have

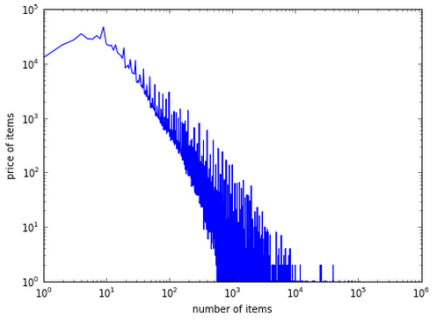


Fig. 7: Item Price

B. User preference

A core question on user preference is what the users would like to buy and how much they are willing to pay for certain products. To categorize the products, we directly use item type that are in the product pages (See Fig 4. From all product pages that are visited we found 50 types in total. Note that Amazon was reported to do price discrimination on E-books [13], which indicates a same E-book may be priced differently in various locations. However, these locations are in scope of countries. In our study, we focus on the U.S. Besides, we believe our results are also meaningful in a global perspective since only E-books are price discriminated. Most of our conclusions still stand.

We show national user preference in Table III. With over 40% of items in wishlist being books, we can see books are in domination. Beside normal paperback books, E-books in Kindle are also being very popular. We believe the cheap price of E-book is a main reason for its prosperity (Books are 136.7% more expensive than E-books in average). Other than books, entertainment products play second important roles. Movies & TV, CDs & Vinyl, Toys & Games, and Video games rank 2, 4, 5, 6 accordingly, which indicate that users are normally pursuing leisure products on Amazon. The following popular products are fashions, home related, and sport items. We also found that Computers, All Electronics, and Camera are only 3 types of products users are paying more than $100 in average, which shows that users are paying substantially more in electronic devices than other products.

After analyzing the general shopping preference of Amazon users, it is natural to study the preferences of people from different demographical and geographical groups. For demographical factor, we compare the shopping preference of males and females. For geographical factor, we compare customers from 3 different geo-location [5], which are – east coast, west coast, and middle of the US.

1) Different Gender: To start with, we show the top 10 popular products of male users in Table IV and female users in Table V. In addition, the average product price for male and female is $36.95 and $26.53 correspondingly. Clearly we can find that male and female customers have different

TABLE III: Overall User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	1829657	41.55%	$22.25
2	Movies & TV	533504	12.12%	$25.26
3	Buy a Kindle	391967	8.90%	$9.40
4	CDs & Vinyl	337188	7.66%	$17.36
5	Toys & Games	231522	5.26%	$40.42
6	Video Games	103951	2.36%	$46.87
7	Amazon Fashion	102012	2.32%	$57.77
8	Kitchen & Dining	99647	2.26%	$46.00
9	Sports & Outdoors	95522	2.17%	$59.16
10	Home & Kitchen	85088	1.93%	$58.50
11	Home Improvement	72194	1.64%	$62.49
12	All Electronics	58146	1.32%	$121.55
13	Health & Personal Care	47912	1.09%	$34.54
14	Digital Music	38375	0.87%	$8.94
15	Computers	38066	0.86%	$123.31
16	All Beauty	37946	0.86%	$23.45
17	Camera & Photo	37725	0.86%	$192.99
18	Arts, Crafts & Sewing	31029	0.70%	$24.91
19	Patio, Lawn & Garden	30747	0.70%	$66.83
20	Grocery & Gourmet Food	27616	0.63%	$21.67

shopping preference. First of all, products in males' wishlists have 39.3% higher price than those in females' wishlists averagely. Furthermore, males are willing to pay higher price in all categories in the two tables. The gap between the product price is even larger than the income gap between males and females, which are reported that males earn 21.1% more than females in 2014 [6]. Therefore we concluded that males are prone to spend more online. Beside the price difference, different genders also prefer different categories of products. While "Books" and "Buy a Kindle" account around 50% of the products for both genders, male customers are more likely to buy a paperback book instead of E-books than female customers. Males prefer sports and electronics while females prefer fashions and beauty-related items. Interestingly, females like "Arts, Crafts & Sewing" much more than males as such items count only 0.3% of all products and ranks 26 in all categories for males. In general, 13 out of the most popular 24 categories (both male and female have over 5000 products in these categories) have over 100% difference, which means that the percentage of certain product of a gender is at least double as the other gender.

2) Different Region: After presenting strong variances in gender preference with quantitative analysis. We now show geographical difference among 3 regions as mentioned – east coast, west coast, and the US. Note that only %57 users have location information in their list descriptions. Following analysis include only the users that input their locations. Due to

[5]East coast includes Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina and Georgia. West coast includes California, Oregon and Washington. Middle of the US includes the rest states

[6]http://www.catalyst.org/knowledge/womens-earnings-and-income

TABLE IV: Male User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	1306784	43.56%	$24.16
2	Movies & TV	386769	12.89%	$26.01
3	CDs & Vinyl	264310	8.81%	$18.10
4	Buy a Kindle	208003	6.93%	$10.60
5	Toys & Games	142277	4.74%	$44.52
6	Video Games	81003	2.70%	$48.13
7	Sports & Outdoors	76039	2.53%	$61.28
8	Home Improvement	59606	1.99%	$65.00
9	All Electronics	48746	1.62%	$126.75
10	Amazon Fashion	48251	1.61%	$69.00

TABLE V: Female User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	493964	37.22%	$17.50
2	Buy a Kindle	176831	13.32%	$8.00
3	Movies & TV	137732	10.38%	$23.30
4	Toys & Games	82399	6.21%	$33.71
5	CDs & Vinyl	68012	5.12%	$14.65
6	Amazon Fashion	51788	3.90%	$48.08
7	Kitchen & Dining	49058	3.70%	$38.30
8	Home & Kitchen	42372	3.19%	$50.04
9	All Beauty	25440	1.92%	$21.74
10	Arts, Crafts & Sewing	21475	1.62%	$22.51

space limitation we do not show all popular products. Instead we present some of the interesting observations.

The total amount of items from east coast, west coast, and mid of US are 641,252, 420,828, and 1,109,846. The average product price in east coast, west coast, and mid of US are $35.80, $35.83, and $33.45. While users from east coast and west coast have almost the same product price, users in the middle of US expose highest price sensitivity – they prefer to pay 6.6% less than other 2 groups in products they desire. It also agrees with the common knowledge that coastal areas have higher income than mid area. However, generally the distributions of items are very similar in the 3 regions. Some noteworthy differences are (1) users from east coast has 2.66% "sports & Outdoors" products. The percentage is 14.2% and 33.0% more than that of west coast and mid of the U.S. However, they are prone to pay 8.4% and 7.9% less in these products. (2) West coastal users are willing to pay $70.88 in "Home Improvement" products, which are 12.9% and 22.8% more than east and mid area. The percentage of "Home Improvement" is also higher (1.92% compared to 1.49% in mid area and 1.72% in east coast).

To conclude, we show that different genders expose vastly different online shopping preference. We also show that people from 3 different regions have similar shopping pattern. Although in most cases it is true, we still point out noticeable diversity among users in these 3 regions.

C. Time Factor

After analyzing product categories and price, we study another important factor that describes user behaviors – time factor. Learned from 5,710,674 items, we found users start to add items to wishlists since 1999. There are only 2,056 items added in 1999. However, in each year the items added increase 85.6% in average. This rapid increase is reasonable since electronic commercials are gaining popularity. Next we explore shopping trend during weekdays and weekends, as well as normal days and holidays. By investigating the added date

Fig. 8: Items Added in Weekdays and Weekends

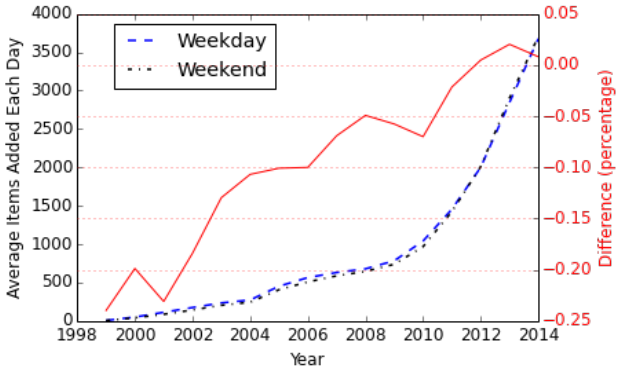


TABLE VI: U.S Federal Holidays

New Year's Day	January 1
Birthday of Martin Luther King, Jr.	third Monday in January
Washington's Birthday	third Monday of February
Memorial Day	last Monday of May
Independence Day	July 4
Labor Day	first Monday of September
Columbus Day	second Monday in October
Veterans Day	November 11
Thanksgiving Day	fourth Thursday in November
Christmas Day	December 25

of items we are able to learn the days that are most appealing to shoppers.

1) Weekdays and Weekends: Intuitively users may browse and shop online more on weekends since they have more free time. Surprisingly our analysis indicates that this intuition is in fact premature. Generally 3,905,728 and 1,523,764 items are added in 4,174 weekdays and 1,670 weekends from 1999 to 2014. Therefore averagely 935.7 items are added on a weekday and 912.4 items are added on a weekend, which implies weekend has 2.55% less added items. We further present our result in terms of each year to show the general trend in Figure 8. As the figure shows, the gap between items in weekdays and weekends is large in earlier years but keeps shrinking. Items added in weekends start to exceed weekdays after 2012. However, there is no evidence showing that weekends would keep growing faster than weekdays. Therefore we conclude that currently people browse and shop online almost equally during weekdays and weekends. A reasonable explanation for this equality is that online shopping consume little effort from users – all the they need to do is to sit in front of a computer at any time – so they are not motivated to devote large chunks of free time in weekends to online shopping.

2) Normal days and Holidays: After studying the shopping favor in weekdays and weekends, similarly we measure holidays and normal days. As there are too many unofficial and regional holidays to study, our study focuses on nationwide federal holidays. There are totally 10 qualified holidays [7], which are listed in Table VI.

When shopping in holidays, users may not always buy products on the exact date. For example, People usually

[7] http://www.archives.gov/news/federal-holidays.html

Fig. 9: Items Added in Normal Days and Holidays

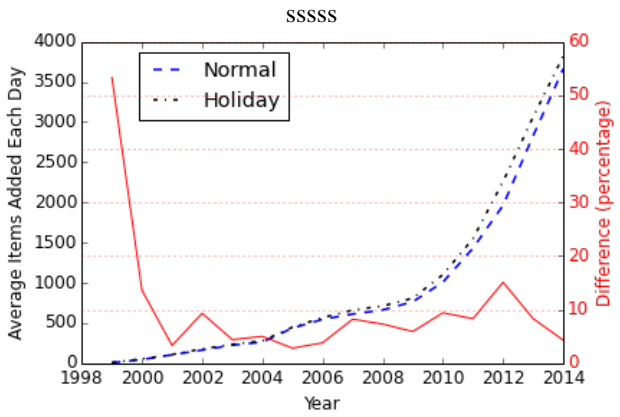
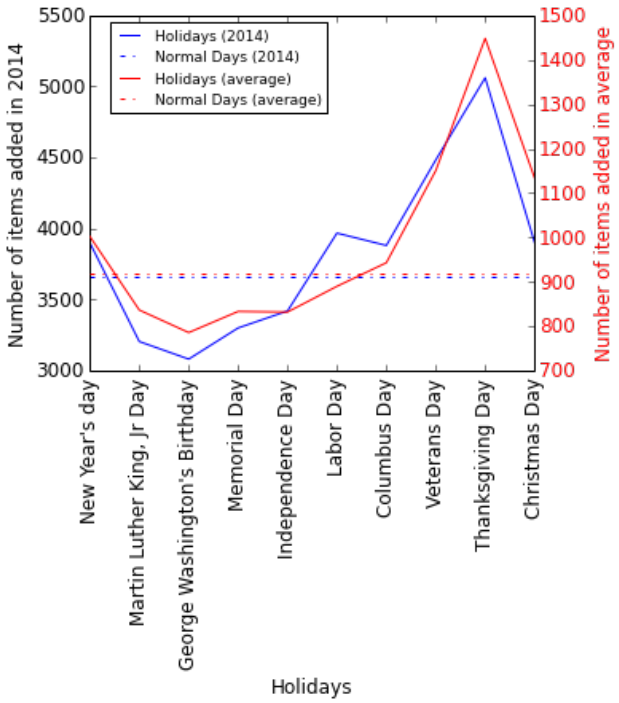


Fig. 10: Different Holidays



prepare gifts before Thanksgiving. Therefore, we consider the nearest consecutive 5 days in our holiday shopping period (previous 2 days and next 2 days). Any day in this period will be counted as shopping in the holiday. For example, We include December 23 to December 27 to calculate the shopping behavior in Christmas day. Note that as a side effect of our methodology, the last 2 days of the previous year are also included when analyzing New Year's day. However, it does not affect our results since we are only interested in the average items in normal days and holidays.

First, we show the average number of items in normal days and holidays in each year as well as the increase rate in Figure 9. Interestingly we found that there is very limited increase from normal days to holidays in average. The average increase is 10.14%. However, as the data of year 1999 is not sufficient to be representative (Only 2,056 items in 1999), we exclude this year in calculation. Therefore the average increase becomes 5.8%. We then conclude that generally people do shop online more during holidays. However, the increase in holidays is merely 5.8%.

Our findings are against common impression that people do shop a lot in holidays, especially Thanksgiving and Christmas. We realized that it is rough to group all national holidays together to compare to normal days. Instead we could treat holiday individually since they are also different from each other. Therefore we compute the average number of items added in each of the holiday. Similar to previous general study, we count the nearest consecutive 5 days of the holiday date as the holiday period. To show both the general case and the current trend, we use the data from 1999 to 2014 and the year 2014 solely. We do not use data from year 2015 because during the time of this study, year 2015 has not been finished. Note that when computing the average item number added in normal days, we leave out all the holiday periods instead of the one that is being analyzed. We show the result in Figure 10. Clearly we can see that holidays are very different from each other. Based on the figure we also made following interesting observations.

1) The sub-figures for year 2014 and average case match well, which indicates that people do not tend to change their shopping behaviors in the long run.
2) New Year's day, Veterans day, Thanksgiving, and Christmas day are 4 holidays that have quite obvious shopping increase. The result indicates that people tend to buy more stuff during the 4 holidays. The increase rate is 4.1%, 25.2%,72.8%, and 30.2% accordingly for all 15 years we analyzed.
3) Among the 4 holidays, Thanksgiving day is the most shopping appealing holiday, which is 72.8% higher than normal days. Christmas ranks second with 30.2% increase. We believe the result is reasonable because people always get considerable discount during thanksgiving. Besides, Thanksgiving and Christmas usually involve lots of presents.
4) Some of the holidays are close to normal days in terms of number of added items in user wishlists (For example, Columbus Day). Furthermore, some of the holidays make the people less willing to add items in their wish lists. 4 holidays – Birthday of Martin Luther King. Jr, Washington's Birthday, Memorial Day, Independence Day, and Labor day – have clearly less items added to user Wish Lists. The drop percentage is 10.7%, 11.8%, 9.0%, 11.3%, and 4.6% accordingly. One possible reason is that People are more likely to be engaged in other activities other than shopping during these holidays since 3 out of the 4 holidays are memorial-type days.

D. Result Implications

Up to this point, we have analyzed 3 aspects regarding to user online shopping behaviors in Amazon. They are product categories and price users prefer in general and on a gender basis. And time factor that affects user shopping behaviors Since Amazon is the largest electronic commercial in the world that offers almost all kinds of products, we believe our analysis

TABLE VII: Frequent meaningful nouns in list-descriptions

Word	Occurrence in list-descriptions	Percentage
university	959	3.19%
books	878	2.92%
music	704	2.34 %
school	604	2.00%
movies	363	1.21%

is also able to shed light on the general case in e-commerce industry.

Our insights on user shopping patterns can be used both in general education purpose and commercial related purpose. Understanding user behavior can be an important factor in e-commerce ecosystem. Equipped with the knowledge, companies are able to better investigate the potential of products, refine marketing strategies, and do timely promotions. Besides, our result show user preference from different user groups, it can also help develop more accurate targeted advertising.

IV. Privacy Information Exposure

Other than the general user behavior regarding to online shopping, we would also like to explore privacy information exposure in Amazon wishlists. it has been shown that many users put their location, birth month and day in their wishlist profiles. We now investigate more such privacy threats. As introduced before, a user is able to write a list-description for each of his/her wishlist in plaintext. Beside the designated purpose of describing the list, users also put much privacy information in these list-descriptions inadvertently or purposefully. For example, in list-descriptions users may mention their relatives' names, education, age, hobbies, etc even though these information has no relation to their wishlists. We believe list-descriptions are a relative source of user information. It is meaningless for a user to lie in list-descriptions since it would not bring any obvious benefit. For privacy concern users, they can simply leave the description empty instead of putting wrong information. Therefore we assume all users are telling the truth in their list-descriptions. By studying to what extent users expose their privacy information, we gain understanding on how users can lose information in their online shopping profiles.

A. List-Descriptions

As list-descriptions are plaintexts keyed-in by the owners. Averagely a list-description consists of 11.9 words (after removing punctuations). However, the median number of words in a description is only 6, which means that while few users put a long speech in their list-descriptions, users are more likely to write down merely a short sentence. Using Stanford Part-of-Speech (POS) tagger [15], we found that approximately 40% of the words are nouns. High ratio of nouns is able to provide more information regarding to the users. Some of the most frequent meaningful nouns and their percentage in all words are listed in Table VII.

We can see that people did put personal information in their wishlists. A representative example is the education background (3.19% users mention "universities" and 2.00% users mentioned "school" in list-descriptions). However, these discreet words cannot generalize the types of personal information users tend to put in their wishlists. If we would like to summarize to what extend do users put their education information in list-descriptions, we need to abstract all hyponyms of the word "education", which may include "University", "College", "School", "graduate", etc. Toward this end, we use Wordnet[] to abstract the ISA relations in words of list-descriptions. Wordnet[] is a directed graph database that connects English words with relations. Synsets relation connect words with similar semantic senses. Super-subordinate relations connect words with their hypernyms and hyponyms. Using Wordnet we are able to group related words into more general words. There is another work that uses similar approach to extract semantic pattern from passwords [18]. However, unlike [18] which leverages tree cut model [12] to balance size of the cut and abstraction level, we are particularly interested in certain levels of abstractions which is very hard to automate. For example, we may desire "relative" more than "person" or "sister "in our word abstraction since "person" is too general while "sister" is too specific. In order to generalize words in list-descriptions in a proper level, we manually selected 8 representative word abstractions to show how much personal information users exposed in list-descriptions. A word is generalized as following. First, we find all synonyms of the word. Then, for each of the synonyms, we find its hypernyms to up to 5 levels. We do not find all its hypernyms because very high level word is usually too general (such as people, entity, etc.). Besides, we also keep the semantic sense close to the target word by restraining levels. Now we have a word pool where each word inside is related to the target word in a similar or more general level. Finally we search the abstraction word we selected in the word pool. If a match is found, we know the word is related to the abstraction. Note that to ensure higher accuracy, we apply such abstraction on nouns only since other words such as verbs carry little information and the generalization of such words may drift the semantic meaning too far away.

TABLE VIII: Personal Information Exposure.

Abstraction	Occurrence in list-descriptions	Percentage
educational_institution.n.01	2545	8.47%
professional.n.01	1371	4.56%
relative.n.01	2880	9.58 %
spouse.n.01	787	2.6%
activity.n.01	9822	32.68%
sport.n.01	1289	4.29%
social_group.n.01	7240	24.09%

".n.01" appended to the words specifies attributes of the words. "n" indicates it is used as a noun. "01" means the first meaning of the word. A word is likely to have different meanings. "01" usually represents the most commonly used meaning.

We show the selected abstraction words as well as their frequencies in Table VIII. Clearly users have publicized much personal information in list-descriptions. Specifically we made the following observations.

1) A considerable portion of users expose their activities (32.68%) and affiliations (24.09%). In a finer-granular sense of activities, 4.29% users mention certain sports. We found that users are extensively mentioning what they did and what group they belong to.

2) 8.47% users talked about their education background. 4.56% users put occupations-related information in list-descriptions. With these essential information, to re-construct the user profiles can be much easier.
3) Generally 9.58% users talked about their relatives. Besides, 2.6% users put their spouses in wishlists, indicating their marital status.

V. Personal Information Identification

We have discussed user information exposure in Amazon wishlists. Such information exposure is directly attributed to user behaviors. That is to say, no matter inadvertently or not, users choose to publicize these information. One may argue that users are responsible for such privacy information loss. Now we study the potential leakage of privacy information that users did not publicize. Specifically, we conduct a pilot study on how to identify user gender from the products in their wishlists. In Amazon, it is possible that a user's gender is unknown. For example, users may only put a nickname in their wishlists to hide their real names and thus their genders.

As our measurement study shows, male and female has significant difference in their shopping behaviors, which is reflected on product type and price in their wishlists. Equipped with this knowledge, we can see that inferring gender information from user wishlists is possible. To show this point, we adopt Support Vector Machine (SVM) to learn the pattern and predict gender of a newly entered user.

We obtain the ground truth from user names as the name is a direct indicator of gender. All users in our dataset are collected through a name search. Therefore we can easily distinguish male and female users. However, there are users that are return to both male and female searches. We ignore these users in our experiment since their gender are unclear. We selected 4 user features to train our SVM. They are (1) The fraction of number of products in one category to the total number of products in the 13 categories[8] that show strong gender implications as mentioned in Section III-B1. (2) Total number of items (3) Average item price (4) Largest item price. To conclude, a 16-dimensional vector is used to describe a user.

To better illustrate how accurately wishlists imply user gender, we setup 5 experiments, in which we trim training and testing sets differently. In each experiment, we set restrictions on training set and testing set. We randomly select 2,500 qualified males and females and we use 80% of the males and females as training set and the rest 20% as testing set. Therefore, our experiments consist of 4000 training users and 1000 testing users, in both sets half of users are males and half of users are females. The results are shown in Table IX. Experiment 1 shows the most general case where no selection is made. The prediction accuracy is 72% under such case, which means it is helpful in predicting the gender of a user. However, it may not be very accurate since there are many users have very few items in their wishlists who are very hard to identify. We tuned our experiments such as requiring the users to have at least 1 of the 13 categories that have strong gender implications (Experiment 2) or to have a relatively abundant products in their wishlists (Experiment 3). After the tuning, the SVM prediction accuracy is increased slightly to 76%-78%. We further select users with abundant products in the 13 gender implying categories in training or testing sets (Experiment 4 and 5). In the 2 experiments, we achieved a fairly good performance with over 80% accuracy.

[8]They are *Arts, Crafts & Sewing, Home Improvement, All Beauty, Grocery & Gourmet Food, All Electronics, Baby, Pet Supplies, Computers, Office Products, Kitchen & Dining, Amazon Fashion, Camera & Photo, Home & Kitchen*.

TABLE IX: SVM Results.

Experiment	Train 13	Train all	Test 13	Test all	Accuracy
1	≥ 0	≥ 0	≥ 0	≥ 0	72%
2	≥ 0	≥ 1	≥ 0	≥ 1	76%
3	≥ 20	≥ 0	≥ 20	≥ 0	78%
4	≥ 0	≥ 0	≥ 0	≥ 20	80%
5	≥ 0	≥ 20	≥ 0	≥ 20	83%

The column Train 13 and Test 13 specifies the number of products under the 13 categories that have strongest gender indication in training sets and testing sets. Similarly, the column Train all and Test All specifies the number of products under all categories in training sets and testing sets.

A user may hide his/her name for privacy reason and expect the related personal information (such as gender) is unknown to the public. However, we proved that from solely the products in users' wishlists, we are able to predict such personal information. Especially when users have a relatively large number of products in their wishlists, the gender prediction can be fairly accurate.

VI. Discussion

A. Data Limitation

There is certain limitation in the data of this study. One problem is that the data may be biased. User wishlists are not always public – users have the ability to change the accessibility of their wish lists (although the default is public). Therefore privacy-aware users may choose to publicize some of their wish lists while keep other wish lists from strangers. Besides, users may choose not to share certain items in their wish lists. For example, privacy-sensitive items such as pregnancy test, firearm-related products, and medicine & drugs. In our work we can only retrieve the product in public wish lists. The data may not be 100% representative of one users' shopping behavior.

B. Personal Information Identification

We have done a pilot study on predicting user gender from their wishlists. We now discuss the potential to use wishlists to identify other types of personal information. To conduct such experiments, first we may need to identify the ground truth. Other than the several available information, we can extract useful information in the list-descriptions. However, it involves accurate human language processing and may require much more effort. Some other way to obtain ground truth is to search online social network, which is shown to be fairly easy [10]. In order to maintain high accuracy, finer-granularity wishlist information may be retrieved. For example, subcategories, ranking of items, timing factor could be effective features to create better machine learning models. We believe identifying more personal information using wishlists is very promising, which will be left as a future work.

VII. Related Work

There have been papers working on analyzing data and present interesting observations based on the data. For example, Traud et al. [16] studied social graph of 100 colleges and universities on Facebook and showed interesting observations such as different institutions have different characteristics. [11] does statistical analysis of the properties of spam profiles collected from social network communities for creating spam classifiers to actively filter out existing and new spammers. [3] describes the system infrastructure, identify the unique properties of list of product attributes and develops a method for automatically distinguishing between positive and negative reviews in Movie Lens to link forum posts to mentioned items.

When coming to e-commercial data measurement, Mikians et al. [13] explored price discrimination problem on several electronic commericals such as Amazon and staples, finding that the same item may have different prices in different regions. Especially some works focus on analyzing user input text such as reviews. Ghose et al. [6] studied the review text of several hundred most popular products and explored its impact on economic outcomes such as sales on Amazon. Similarly, Ivanova et al. [9] studied the review system in Amazon, revealing that user purchasing intention is greatly impacted by product reviews.

Online privacy is a major user concern. However, users may still leak or expose their data on websites inadvertently. Friedlan et al. [5] illustrated that users are often unaware the privacy implication of publishing locations. Even worse, users do not even know they published such sensitive information. However, when users realized the privacy implication, their shopping behaviors change significantly. Brown et al. [1] shows privacy invasion puts significant negative impact on online purchasing behaviors and Tsai et al. [17]shows that users are more willing to purchase in privacy protective websites if privacy information is salient.

Besides, it has been proven that inferring user personal information based on other data is practical. Using easily accessible public data, user privacy is under huge threat. Narayanan and Shmatikov [14] created a new framework to De-anonymize users based on social network topology. Wondracek et al. [19] leveraged user group membership to uniquely identify an individual user or at least largely reduce candidates. Chaabane et al. [2] studied the privacy leakage through user interest in music. They can infer user personal information such as gender, age, location, etc based on user self-declared interest. Hecht et al. [8] derives user geo-locations from their tweets through machine learning. Goga et al. [7] correlates features such as timestamp and writing style of user posts on different websites to identify same user.

VIII. Conclusion

In this paper, we investigate Amazon wishlists, where users store their desired products. We collect over 30,000 users' complete wishlsits and took a 2-step approach to analyze our data. First we try to measure the user behavior from their wish lists by analyzing wishlists in multiple dimensions. Our result shows that there is huge discrepancy between male and female online shopping pattern. The 2 groups differ in both preferred product categories and prices. Furthermore, we investigate user shopping preference taking time as a factor. We reveal that users are in fact not shopping significantly more during holidays or weekends in terms of online shopping. We also investigate different holidays that are shopping appealing or repelling. We highlight that although there are certain holidays are very shopping involving, over half of the national holidays have no increment or even lower shopping favor than normal days. After studying user shopping pattern, we try to identify personal information from list-descriptions or items in the lists. We apply simple human language processing on the plaintext in list-descriptions to generalize the information users tend to expose. Beside the information users mentioned themselves, we also explore the possibility to infer user personal information based on their wishlists. To this end, we selected representative features from wishlists and use SVM to infer user gender. Our results show that given a user that has abundant wishlists, we can predict the user's gender with fairly high accuracy.

References

[1] M. Brown and R. Muchira. Investigating the relationship between internet privacy concerns and online purchase behavior. *American Academy of Advertising. Journal of Electronic Commerce Research*, 2004.

[2] A. Chaabane, G. Acs, M. A. Kaafar, et al. You are what you like! information leakage through users interests. In *NDSS*, 2012.

[3] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *ACM. WWW*, 2003.

[4] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You are what you say: privacy risks of public mentions. In *ACM. SIGIR*, 2006.

[5] G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geo-tagging. In *USENIX. HotSec*, 2010.

[6] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE TKDE*, 2011.

[7] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *ACM. WWW*, 2013.

[8] B. Hecht, L. Hong, B. Suh, and E. H. Chi. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *ACM. SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[9] O. Ivanova, M. Scholz, and V. Dorner. Does amazon scare off customers? the effect of negative spotlight reviews on purchase intention. In *Wirtschaftsinformatik*, 2013.

[10] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. In *ACM COSN*, 2009.

[11] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *ACM. SIGIR*, 2010.

[12] H. Li and N. Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 1998.

[13] J. Mikians, L. Gyarmati, V. Erramilli, and N. Laoutaris. Detecting price and search discrimination on the internet. In *ACM HotNets*, 2012.

[14] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *IEEE Security & Privacy*, 2009.

[15] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *ACL. NAALC*, 2003.

[16] A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Elsevier. Physica A: Statistical Mechanics and its Applications*, 2012.

[17] J. Y. Tsai, S. Egelman, L. Cranor, and A. Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. *INFORMS Information Systems Research*, 2011.

[18] R. Veras, C. Collins, and J. Thorpe. On the semantic patterns of passwords and their security impact. In *NDSS*, 2014.

[19] G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *IEEE Security & Privacy*, 2010.