Amazon Wishlists

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 Wishlists and items of over 30,000 users, in the mean time showing interesting results on user online shopping behavior. Specifically, we show behavior difference in gender and geo-location. Besides, we also study shopping patterns in normal days and holidays. Surprisingly we found that user may not always like to shop during holidays. Finally, we demonstrate that information in Wishlists has potential to leak user personal information. We use state-of-art machine learning techniques (SVM) to train and test the data. Our result indicates that based solely on the items in users' Wishlists, their gender, education, and hobbies can be inferred with a considerable success rate.**

I. Introduction

Entering the era of big-data, Internet has been largely extended to be data-rich. Network users are prone to expose their private information in public websites inadvertently [9], making themselves face the threat of information leakage. To study how the information bundle can be used to threat user privacy, we investigate user profile and their wishlists in Amazon, unveiling its user behaviour exposure threat and potential privacy leakage.

We collected complete profile and wishlists information of over 30,000 users in Amazon by web scraping. Based on the data, we conduct measurement study on user behavior. Our measurement study concentrates on user shopping preference, which are organized in 3 dimensions 1)What to buy 2)When to buy 3)Product prices users are willing to pay. Specifically, we compare the user preference in different regions and gender. We also investigate shopping peaks and pits through a year. Surprisingly we found that there is little shopping increase in holidays and not all holidays are appealing to shoppers. 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 conduct a preliminary study on user privacy information identification using wishlists. We prove that wishlists have potential to leak user personal information. We used the well-know machine learning algorithm – Support vector machine (SVM) to train and test data. Ground truth are retrieved from user wishlist descriptions in their Amazon profiles. based on the types of products in user wish lists, we are able to identify the user gender with over 70% success rate, which is substantially more than guessing.

II. Data and Methodology

As the largest electronic retailer in the U.S, Amazon was reported to have 270 active customer accounts in 2014 (http://www.statista.com/statistics/237810/number-of-active-amazon-customer-accounts-worldwide/). With 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.
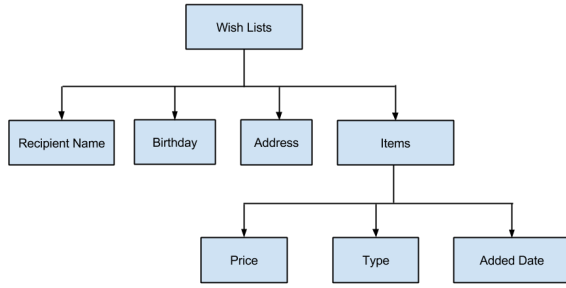
Fig. 1: data hierarchy

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 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 [3] to search common names. The wishlist search engine will return at most 2016 users and that are associated with a name. [1] 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 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 600 common names (300 males [2] and 300 females [1] 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. "MENTION MALE AND FEMALE NUMBER".

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 [9], which states that users tend to expose their personal information in open websites.

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.
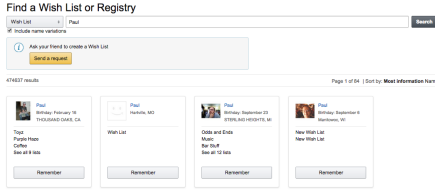

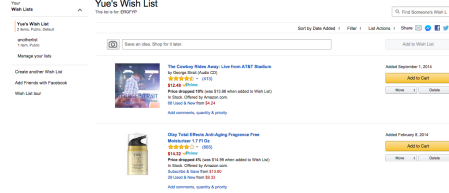[1] The search engine usually states that hundreds of thousand users are found, but only the first 2016 users are viewable.

Fig. 2: Wishlist Search Engine

Fig. 3: Wishlist

Fig. 4: Item

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%

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.

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

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 [**?**], 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.

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

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, which are – east coast [4], west coast[5], 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
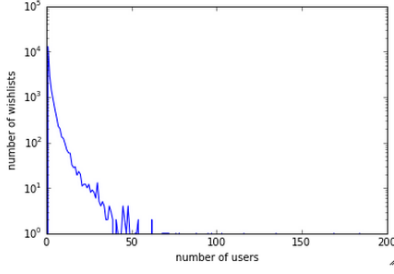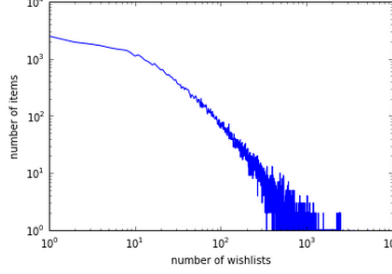
Fig. 5: Number of lists the users have

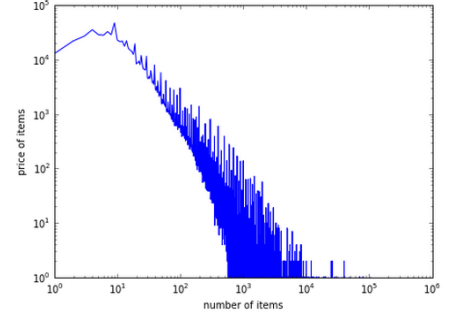Fig. 6: Number of items the lists have

Fig. 7: Item Price

find that male and female customers have different 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.

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

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 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. We explore shopping trend during weekdays and weekends, as well as normal days and holidays. By investigating the added date of items we are able to learn the days that are most appealing to shoppers. Learned from 5,710,674 items, users start to add items to wishlists since 1999. Each year the items added increase 85.6% in average. This rapid increase is reasonable since electronic commercials are gaining popularity.

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. 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 trend in Figure 8.

Fig. 8: Items Added in Weekdays and Weekends
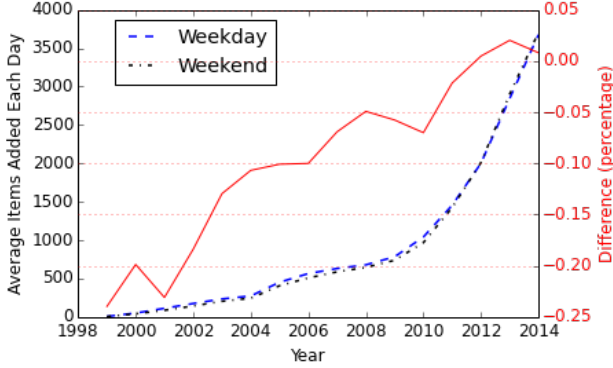
2) Normal days and Holidays: In this section we measure how the items are added in the wish lists during holidays and normal days. As there are many unofficial or regional holidays, our study considers only nation-wide federal holidays. There are totally 10 qualified holidays. These holidays (http://www.usa.gov/citizens/holidays.shtml) and the date of the holidays are listed in Table VI.

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

We analyze the data in 2 dimensions. First, We measure the items added in these holidays and compare the average items added in holidays and normal days. Then we compare the added items among different holidays. When calculating items added in a certain holiday, we consider the nearest consecutive 5 days. That is to say, we include the previous 2 days and the next 2 days in our holiday shopping session. For example, the Christmas day is December 25. We will include December 23, December 24, December 25, December 26, December 27 in Christmas day calculation. As a result, calculate items added during a year in a little different manner – we include the last 2 days in previous year because the 2 days are considered New Year's day. We do not include the first 2 days in the next year because there is no holidays in the end of Decembers.

3) Holidays and Normal Days: According to our data, people start to add items in their wish lists since 1999. We analyze the items added in normal days and holidays in each year by calculating the average items added in the 2 groups. Figure **??** shows the items added in normal days and holidays

1) The items added in wish lists increase year by year, no matter in normal days or holidays. We believe that the reason behind is that electric commercial gains its popularity.
2) There is increase in number of items added during holidays. The average increase rate is 10.89%. However, as the data for year 1999 is not sufficient to be representative, we exclude year 1999 in our calculation and eventually we conclude that there is around 5.9% increase in holidays.
3) Generally people do buy more stuff during holidays. However, the increase is very limited. In 2001, the items added during holidays are even less than normal days.

Fig. 9: Items Added in Normal Days and Holidays
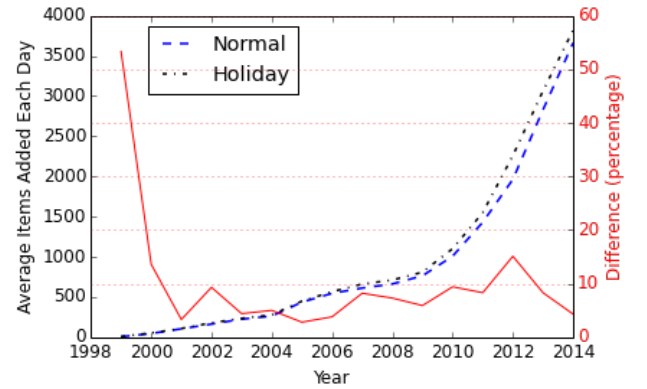
4) Different Holidays: Previously we compared the shopping difference between holidays and normal days. We found that although people shop more in holidays, the increase is not that much. Our findings are against common impression that people do shop a lot in holidays, especially Thanksgiving and Christmas. We realized that it is rough to group all national holidays together to compare to normal days. We believe that holidays are also different from each other. For example people tend to shop more in Thanksgiving and Christmas. Therefore we compare different holidays in our study. Similarly, we include the last 2 days in the previous year in our analysis. There are 10 national holidays as introduced before. We again calculate the average number of items added in the nearest consecutive 5 days and compare to the average number of normal days.

We use the data from year 1999 - 2014 and the year 2014 solely. We wish to show both the average case(from 1999-2013) and the current trend (only year 2013). We do not use data from Year 2014 because during the time of the study, the years 2014 has not been finished. Note that when computing the average item number added in normal days, we leave out all the holidays instead of simply leaving out the one that is being analyzed. The result is shown in Figure 10 for 1999-2013 and Figure 11 for the year 2013. In Figure 11, The x-axis indicate the holidays accordingly (0-New Year's day, 1-Birthday of Martin Luther King, Jr, 2-Washington's Birthday, 3-Memorial Day, 4-Independence Day, 5-Labor Day, 6-Columbus day, 7-Veterans Day, 8-Thanksgiving day, 9-Christmas day).

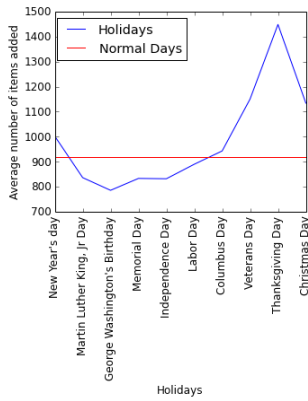According to Figure 10 and Figure 11, we find that holidays
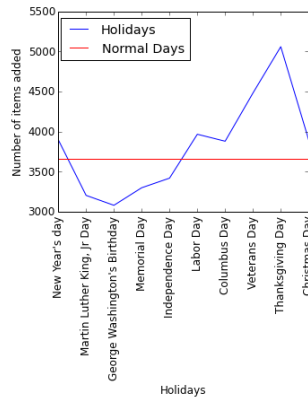
Fig. 10: Different Holidays in 1999-2014

Fig. 11: Different Holidays in 2014

have huge difference between each other. In addition, we make the following observations:

1) The 2 figures match very well, which means that people do not tend to change their shopping behaviors at least during a short period of time.

2) Some of the holidays are not much different from normal days (For example, Columbus Day). What is more, some of the holidays make the people less willing to add items in their wish lists.

3) 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.

4) Among the 4 holidays, Thanksgiving day is the most shopping appealing holiday, which is 72.8% higher than normal days. It is reasonable because people always get considerable discount during thanksgiving days. Christmas is the second most shopping appealing holiday because people always need to buy presents during Christmas.

5) At the first glance it is surprising that 4 holidays – Birthday of Martin Luther King. Jr, Washington's Birthday, Memorial Day, Independence Day, and Labor day – clearly have even less items added to user Wish Lists. The drop percentage is 10.7%, 11.8%, 9.0%, 11.3%, and 4.6%. One possible reason is that People are more likely to be involved in other activities other than shopping during these holidays. 3 out of the 4 holidays are memorial days. Certainly these days are considered not good for shopping.

We believe our quantitative results are helpful in guiding merchant marketing strategies and targeted advertising.

IV. PRIVACY INFORMATION IDENTIFICATION

From our measurement study, we have studies 3 dimensions of user shopping preference – 1)What to buy 2)When to buy 3)Price.

From our study of shopping preference difference, it has been clear that different people have different shopping pref-

erence. One question is: Can we identify user personal information from the their Wish Lists. That is to say, Items added in Wish Lists have potential to leak personal information. We explore the possibility using machine learning to help identify user personal information from their Wish Lists.

A. Information in list-description

To successfully identify the user personal information. One challenge we face is how to obtain the ground truth of the user. There are several possible ways to find out the result. First, Services such as yahoo people search can be used as ground truth at a reasonable monetary cost. We can input the name and location information in the search engine to identify a user. Second, similar to the first approach, we can search people in online social networks such as Facebook or Twitter. Third, we can use the user list-description as the ground truth. Users sometimes expose personal information in their list-description and we can take advantage of it.

The first approach does not really provide an accurate result since the information given is not rigorous enough to identify the exact person. If wrongly identified, it has much negative impact on the identification process, which is by itself not very accurate already. The second approach is similar to the first one. It is even more inaccurate. Due to the unreliable nature of the first 2 approaches, we took the third approach. Besides the accuracy, collecting ground truth from list-descriptions have the benefit of not depending on external data source. Recall that we collect items for only users with list-descriptions. All our 20,099 users collected have list-descriptions. Note that there are also limitations for the third approach. For example, the list-description is plain-text, making it hard to process and analyze. Secondly, there may not always be useful information in list-descriptions.

We assume that everybody is telling the truth in the list-descriptions. If somebody is lying, we do not have a way to identify it. However, it is meaningless for a user to lie in list-descriptions since lying there will not bring them any benefit. If a user concerns his/her privacy, he/she can simply leave the list-description blank instead of lying.

Among the users who have list-descriptions, 233,595 words are found (44,387 punctuations are excluded). The average number of words in list-descriptions is 11.6. We can see that the list-descriptions are not long – basically the length of a normal sentence. While there are scattered long list-description (the longest list-description has 682 words), more than half of the list-descriptions have length less than or equal to 6 (10,592 out of 20,099). And only 29.9% of the list-descriptions have over 10 words. Considering that there are even less words that carry information, One possible consequence is that there might not be enough information in one list-description to identify personal information.

Next we retrieve information from list-descriptions.

One way to retrieve information is to search the keyword. If a keyword appears in one user's self-description, there is high probability that the user is related to the keyword. For example, if a keyword "college" appears in the self-description, the user is likely to have a college degree.

We used Stanford Part-of-Speech (POS) tagger [13] to tag the user list-descriptions to retrieve useful information. There are a number of tags a word can belong to. However, we only focus on noun-family words (with tag NN, NNS, NNP, and NNPS in Stanford POS tagger). The reason behind is that most of the nouns carry certain amount of information while most of other words do not carry useful information.

There are totally 95,359 out of all 233,595 words (around 40.8%) belonging to the noun family. The ratio of nouns is quite large. The most frequent ones are listed in Table VII

TABLE VII: Frequent Nouns IN list-descriptions

Rank	Item Type	frequency
1	s	1498
2	love	1327
3	m	1314
4	university	1098
5	music	1095
6	list	1064
7	books	1001
8	school	659
9	movies	652
10	t	628
11	games	593
12	things	553
13	college	495
14	stuff	443
15	work	419
16	fan	411
17	wish	389
18	video	380
19	history	371
20	photography	353

From Table VII We can see that except for meaningless single characters ("s" and "m" in this case), the noun family words carry some information of a user. The information involves hobbies (such as music, books, movies, etc) and education level (such as university, college). These words can be used as keywords in the information retrieval step. On the other hand, if we look into the verbs, we will find the most popular 20 verbs are "is", "am", "have", "are", "live", "be", "love", "buy", "want", "enjoy", "know", "used", "do", "don", "get", "married", "go", "went", "read", "loves", and "was". Except "married", the other words can not assist us to find any useful information about a user.

B. Information Identification

We do personal information identification on a subset of users who have more than 30 items in their wish lists in our dataset. We filter part of the users out because they do not have sufficient information in their Wish Lists. After the filtering, we have 13,464 users as our target users. We choose users with certain keywords in their list-descriptions in the user pool. We use the portion of different types of products, in user wish lists as input to the classifier. Therefore we have an input vector of length of 50. We use SVM classifier to train and test the data. 80% of our data is used as training data and the rest of 20% of the data as testing data. The personal information identification process is binary, which means that we use only 2 classes in each identification cycle. The reasons for having only 2 classes. It is reasonable to have 2 classes in a classifier. We intend to answer questions like "Is he a professor?", "Does he have a university diploma" or "Does he like hiking?". A prediction of the answer is very helpful in identifying a user. We try 2 different types of personal information – 1) gender 2)hobbies

C. gender

There are only 2 genders in general, it is natural to have only 2 classes. To save the training resource while maintaining high representation, we chose 389 males and 168 females as our dataset. We used the first 80% person as the training set and the rest as the test set. The ground truth comes from the name of the users. We identify a user "male" if this user is using a male name in the profile. females are identified using a similar approach. As a result, the success rate to classify whether a given user is a male or female is around 84%. We can see that the classifier works pretty well in the case of gender, which means that given only the Wish List information of a user, we can identify whether this user is male or female with pretty high certainty.

D. Hobbies

Afterwards, We use hobbies as examples, we choose all pairs of hobbies in the top 20 nouns. 7 keywords are selected. They are -"music", "books", "movies", "games", "video", "history", and "photography" and try to differentiate users between any pair of them. The success rate between any pair of them is shown in Table VIII

TABLE VIII: Success Rate of Hobbies

	music	books	movies	games	video	history	photography
music	NA	0.506	0.653	0.713	0.691	0.772	0.743
books	0.506	NA	0.631	0.667	0.647	0.526	0.807
movies	0.653	0.631	NA	0.617	0.617	0.762	0.662
games	0.713	0.667	0.617	NA	0.521	0.776	0.703
video	0.691	0.647	0.617	0.521	NA	0.777	0.700
history	0.772	0.526	0.762	0.776	0.777	NA	0.680
photography	0.743	0.807	0.662	0.703	0.700	0.680	NA

We can see that the success rate can be high (80.7%) or even no better than guessing (50.6%). Such big variance may be due to the relativity of different hobbies. When 2 hobbies are not much related, we can find the success rate be usually higher than 70%. For example, "history" has very limited relativity to other hobbies except for "books". We can see from the table that history has higher success rate except for the "books" entry because history relates to books pretty much. It is difficult to have further increase in success rate because people often have more than 1 hobbies. Therefore one person may belong to both group, which will have quite negative impact on the result.

We conclude that wish lists are helpful to identify a user's hobby. However, depending solely on wish lists are not a very reliable option to identify user hobbies.

V. Limitation and Future Work

A. Limitation

There are limitations in our work. We discuss the limitations and possible consequences from these limitations.

1) The data may be biased. The user wish lists 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 accurately representative of one users' shopping behavior.

2) When data of price is involved, some products are ignored. We have mentioned that the price cannot always be retrieved. There are multiple reasons and some of them cannot be easily solved such as price not being displayed or javascript being used for displaying the price. Besides, some prices do not accurately reflect the user's preference because there might be multiple prices (from various retailers, used ones or new ones), and we do not know which one does the user prefer. We take only the new one and the first retailer's price.

3) When doing personal information identification. We actually do not have complete ground truth. Using the keyword approach will give us a group of people with certain features. However, we cannot identify a large enough group without the features. For example, searching "engineer" gives us a group of engineers, however, we do not know whether the users not mentioning "enginner" in their list-descriptions are engineers or not. This lack of complete ground truth may have big negative impact on the success rate of our information identification step.

B. Future Work

We discuss future work that can be done.

1) Use possibly better information retrieval method. In our project, the information retrieval step is straightforward – we search the keywords in users' list-descriptions to label the users. It works with relatively high accuracy. However, this method is a little too strict so that not many users are selected for a certain class. One way to improve our result is to use a better information retrieval method. We can possibly obtain more users for a specific group while maintaining relatively high accuracy. It will provide more stable data and probably more well-trained SVM.

2) There are 2 ways to improve the user personal information identification process. The first way is to improve the accuracy of the classification step. The second way is to use our method as a filter to filter out irrelevant users and use some other strategies to do further identification.

VI. Related Work

There are a large amount of works that have been done to learn plain text. [8] mines e-mail content for author identification. They study an extended set of email document features, which include structure, linguistic patterns, etc. They use SVM learning algorithm to differentiate the authors of emails. [12]presented algorithm for mining users' opinions of an item from their reviews of the items, classifying documents in overall sentiment. They use keywords appearing in the user comments of an item to infer whether the user like this item or not and to what extend do they like/dislike the item.

[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. [7] 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. [10]attempts to do on a set of Amazon wish lists spiders from web. They started by doing a wish lists search of people names(first names and nicknames from the baby names databases) and then retrieve the results linking to the wish lists of the first 25 matches.

VII. Conclusion

In this project, we investigate Amazon public wish lists, where users store their desired products. We collect over 20,000 users' complete profile and wish list items and take a 2-step approach to analyze our data. The first step is that we try to measure the user behavior from their wish lists. We answer the following questions: 1) Are users from different regions follow the same shopping patter? Is there any difference among people in different regions? 2) Is the common belief that people shop more during special date, especially holidays reflected in Amazon wish list? Our result shows that although people from different regions are following a general shopping behavior, they still have differences. Interesting observations such as east coast people are less willing to pay high price on clothing are presented in our measurement result. Furthermore, we investigate user shopping preference in holidays and normal days. We found that people do shop more in holidays, but the increase is not significant. We then explore different holidays through a year, we found that although there are holidays that people shop much more than normal days, there are over half of the holidays that have no increase or even lower shopping records. The second step we take is trying to identify personal information from the items in a user's wish list. We retrieve our ground truth information from the data using a keyword approach – searching a keyword in users' list-descriptions to divide users into different classes. This approach works well because it is accurate while it labels sufficient people for machine learning purpose. We use SVM to train and classify the data obtained from the information retrieval step. We examined 2 types of personal information – gender and hobbies. We use the first 80% data for training and the rest 20% of data for testing. The gender classification gives a good success rate of over 80%. However, identifying hobbies does not always achieve a satisfying success rate. This accuracy is not good enough to identify users' personal information. However, it can provide help when accuracy requirement is

not high. It is also able to serve as a filter for a more complex identification tool.

VIII. ADDITIONAL DATA

TABLE IX: East User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	248357	38.73%	$21.87
2	Movies & TV	76642	11.95%	$26.82
3	Buy a Kindle	46996	7.33%	$10.15
4	CDs & Vinyl	46811	7.30%	$18.33
5	Toys & Games	42784	6.67%	$40.24
6	Video Games	19585	3.05%	$48.05
7	Amazon Fashion	18805	2.93%	$54.12
8	Sports & Outdoors	17064	2.66%	$54.92
9	Kitchen & Dining	16583	2.59%	$48.08
10	Home & Kitchen	13568	2.12%	$56.90
11	Home Improvement	11019	1.72%	$57.58
12	All Electronics	9457	1.47%	$123.46
13	Health & Personal Care	7527	1.17%	$33.49
14	All Beauty	6609	1.03%	$25.14
15	Computers	6439	1.00%	$125.72
16	Camera & Photo	6076	0.95%	$201.96
17	Patio, Lawn & Garden	5263	0.82%	$68.42
18	Grocery & Gourmet Food	4437	0.69%	$21.45
19	Digital Music	4410	0.69%	$8.77
20	Automotive	4181	0.65%	$57.23

TABLE X: West User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	170646	40.55%	$22.02
2	Movies & TV	53326	12.67%	$25.96
3	Buy a Kindle	32443	7.71%	$10.30
4	CDs & Vinyl	30563	7.26%	$17.73
5	Toys & Games	25275	6.01%	$41.67
6	Video Games	10809	2.57%	$45.89
7	Amazon Fashion	10598	2.52%	$56.69
8	Sports & Outdoors	9801	2.33%	$59.55
9	Kitchen & Dining	9791	2.33%	$48.12
10	Home Improvement	8094	1.92%	$70.88
11	Home & Kitchen	7745	1.84%	$63.56
12	All Electronics	5786	1.37%	$124.45
13	Health & Personal Care	4365	1.04%	$36.16
14	Camera & Photo	4249	1.01%	$203.55
15	Digital Music	3981	0.95%	$8.42
16	Computers	3925	0.93%	$128.50
17	All Beauty	3389	0.81%	$25.50
18	Arts, Crafts & Sewing	2921	0.69%	$27.85
19	Grocery & Gourmet Food	2778	0.66%	$21.29
20	Automotive	2651	0.63%	$64.46

TABLE XI: Mid User Preference

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	449377	40.49%	$22.79
2	Movies & TV	146774	13.23%	$25.12
3	Buy a Kindle	96586	8.70%	$9.13
4	CDs & Vinyl	78067	7.03%	$17.02
5	Toys & Games	67700	6.10%	$38.92
6	Video Games	31220	2.81%	$47.69
7	Amazon Fashion	27008	2.43%	$55.90
8	Kitchen & Dining	25975	2.34%	$45.25
9	Home & Kitchen	22353	2.01%	$56.94
10	Sports & Outdoors	22192	2.00%	$59.24
11	Home Improvement	16485	1.49%	$62.64
12	All Electronics	13294	1.20%	$118.96
13	Health & Personal Care	11506	1.04%	$34.83
14	All Beauty	9856	0.89%	$22.30
15	Digital Music	9309	0.84%	$8.86
16	Computers	8884	0.80%	$130.28
17	Camera & Photo	8243	0.74%	$194.83
18	Arts, Crafts & Sewing	7803	0.70%	$24.75
19	Grocery & Gourmet Food	7692	0.69%	$21.61
20	Patio, Lawn & Garden	7661	0.69%	$65.30

REFERENCES

[1] 1000 most common female names in the u.s. http://names.mongabay.com/female_names.htm.

[2] 300 most common male names in the u.s. http://names.mongabay.com/male_names.htm.

[3] Amazon wish list search engine. http://www.amazon.com/gp/registry/search.

[4] East coast of the united states. http://en.wikipedia.org/wiki/East_Coast_of_the_Unite

[5] West coast of the united states. http://en.wikipedia.org/wiki/West_Coast_of_the_United_States.

[6] CATALYST. Women's earnings and income, 2015. http://www.catalyst.org/knowledge/womens-earnings-and-income.

[7] K. Dave, S. Lawrence, and D. M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[8] O. De Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Record*, 30(4):55–64, 2001.

[9] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You are what you say: privacy risks of public mentions. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 565–572. ACM, 2006.

[10] M. A. Kraft. Mashing up the internet. *Medical Reference Services Quarterly*, 26(S1):191–207, 2007.

[11] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: Social honeypots + machine learning. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 435–442, New York, NY, USA, 2010. ACM.

[12] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[13] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.