A Measurement Study on Amazon Wishlist and Its Privacy Exposure

Yue Li*, Nan Zheng†, Haining Wang‡, Kun Sun§, Hui Fang‡

*Department of Computer Science, College of William and Mary
†Shape Security
‡Electrical and Computer Engineering, University of Delaware
§ Center for Secure Information Systems, George Mason University

***Abstract*—User preference plays an important factor in E-commerce websites for advertising and marketing, and the disclosure of user preference could also raise privacy concerns. As one of the largest E-commerce platform, Amazon features a wishlist that allows users to keep track of their desired products. In this paper, we investigate Amazon wishlist, and its possible privacy exposure. To this end, we collect complete wishlists of over 30,000 users, by analyzing which we are able to make interesting observations based on user online shopping preference in multiple dimensions. Specifically, we show user preference variation from different demographical groups, including gender and geo-locations. Taking timing factors into consideration, we also observe that unlike traditional walk-in-shop type of shopping, there is no significant difference in the dynamics of Amazon wishlists between weekdays and weekend. In the investigation of user information exposure in Amazon wishlists, we parse and analyze list-descriptions, illustrating which and to what extent user personal information is exposed to the public. Finally, we demonstrate that the information in wishlists has potential to leak a user's private personal information. Based on the collected user data, we can predict user gender with over 80% accuracy by just exploiting items present in Amazon wishlists.**

I. Introduction

Entering the era of big data, online user data has been exploited to generate revenues in many fields. Electronic commerce is one of the most data-driven markets. There have been many works studying how users behave in E-commerce websites [8], [9], [16], bearing the goal of understanding the market and thus producing high financial outcomes, e.g., user data can be monetized by feeding targeted advertising or performing price discrimination [11]. Although there are various kinds of user data on e-commercial environments, the shopping history and preference is one of the most sensitive and valuable data. However, shopping history is considered private information that users are not willing to publicize. To meet this expectation, most E-commerce websites keep user shopping history private, which inevitably raises a major challenge in studying user shopping preferences.

Nonetheless, users do not always hide their purchase intentions. Wishlist, a list-type data in Amazon, is made publicly available by default. Users put items into their wishlists for tracking their desired products or for gift reference. As the wishlists is usually used to record desired products, it largely reflects the shopping preference of the user. In other words, if a user adds a item in his/her wishlist, the user is prone to purchase the item in the near future. Besides, wishlists are also the indicators of shopping history since the items will not be removed after these items are purchased unless manually done so. As the largest electronic retailer in the U.S., Amazon was reported to have over 270 million active customer accounts [14]. With such a large user base, user behaviors in Amazon significantly imply the market pattern and trend, which is critical for E-commerce and advertisement ecosystems. However, such an online user behavior has not been systematically studied before.

In this paper, we analyze the wishlist in Amazon to investigate user shopping preference as well as its privacy implication. Our objective is to shed light on general user preference on E-commerce and help enterprises to refine both their marketing strategies and privacy policies.

In this paper, we first collect the complete profile and wishlist information of more than 30,000 users and approximately 2 million items in Amazon by web scraping. Based on the collected data, we conduct analysis on user shopping preference in three dimensions: (1) product categories, (2) product prices, and (3) timing. Specifically, we compare the user preference between different gender, regions, and time periods. Our results suggest that there is a significant difference between male and female in their wishlists. By contrast, people from different regions in the U.S. demonstrate similar shopping preference. Moreover, the dynamics of wishlist, in terms of the number of "added items", are similar between weekend and weekdays, yet vary in different holidays throughout a year.

Besides user behavior analysis, we also explore the possibility of user privacy exposure in Amazon. It is believed that users are prone to expose their private information in public websites inadvertently [6], [7], making themselves vulnerable to information leakage. However, security and privacy protection is critical for an E-commerce website since it is closely tied with user purchasing intentions [4], [16]. To study to what extent the information in wishlists poses a threat on user privacy, we dissect both the items in wishlists and the user input list-descriptions to uncover user personal information. We first evidence user personal information exposure by analyzing their list-descriptions, showing that users inadvertently disclose sensitive information such as profession, education background, and relatives' information, in their list-descriptions.

978-1-4673-8999-0/17/$31.00 ©2017 IEEE

In addition to inadvertently written information, more could be inferred from public data in user's wishlists. Such an information leakage has been proved practical in many works [5], [13], [18]. In our study, we also aim to learn user personal information by analyzing publicly available data. In particular, we employ Support Vector Machine (SVM) to predict user gender based solely on these items in their wishlists. The results indicate that our prediction can achieve more than 80% accuracy.

We emphasize that though our study is mainly on Amazon, wishlist-type objects can be found on many major e-commerce websites such as Ebay and Bestbuy, etc. Thus, our methodology of deriving user shopping preference and personal information from publicly available wishlists can essentially be extended to other platforms.

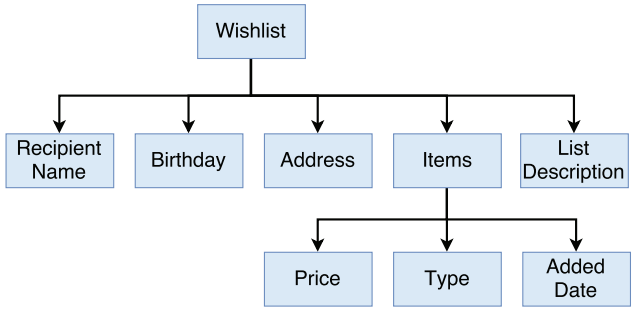
The remainder of this paper is organized as follows. Section II introduces the data structure of a wishlist and our data collection methodology. Section III measures the collected data and present our observations. Section IV shows to what extent personal data is exposed through list-descriptions. Section V presents an inference attack, in which we exploit wishlists to uncover user personal information. Section VI discusses the limitation of this work. Section VII surveys related work, and finally Section VIII concludes this paper.

II. Data Structure and Collection Methodology

Every registered user in Amazon has a public profile that contains various objects, such as a profile photo, rated items, reviews, and wishlists. Wishlist is a list-type data object that contains the owner's desired products. A user may have multiple wishlists to accommodate different types of products. Besides items, each wishlist maintains a separate recipient profile, including name, birthday, and shipping address. When a new wishlist 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, as the user fills the wishist, birthday only includes month and day in a wishlist, thus the age of the recipient is unknown. Similarly, address information only shows the state and city. Note that though every user has a public profile, not all items in the profile are public. Wishlists are configurable yet public in default. We focus on the analysis of user wishlists. Therefore, other items in a user profile, such as reviews, are beyond the scope of this work.

Figure 1 shows the data hierarchy of an Amazon wishlist. We can see that other than items, each list has its own recipient name, birthday, and address. Besides these formatted information, each wishlist also has a list-description. The list-description is plain text keyed in by a user, usually used to briefly describe the wishlist or the user. Two examples are as follows: *"I LOVE music! Buy me a CD!"* or *"Son of Alice and Bob, brother of Calvin, husband of Deanna. I moved from Milton Keynes, England to Smithfield, North Carolina, USA on 1/1/2000 and recently moved from there to the Raleigh/Garner border in the same state"*. The first example indicates the hobby of a user, and the second example (with sensitive information arbitrarily altered for privacy preserving purpose) exposes much more personal information including parents' names, marriage status, spouse's name, and moving history.

Fig. 1: Data Hierarchy.



As the wishlist stores a user's desired products, it directly reflects the user's online shopping preference. In particular, we are interested in the products a user prefers to buy, the price a user may need to pay, and the time when the wishlist is the most active or the least active. Therefore, we need to collect a sufficient amount of data from Amazon for wishlist analysis.

A. Data Collection Methodology

One way to collect data from Amazon is to use its Product Advertising API [1]. However, the API does not provide wishlist or product type access, which is essential to our study. In addition, Since there is a one request per second limit on non-profiting API users[1], using a few API accounts cannot boost the speed. Moreover, signing up for more accounts requires much more effort. As such, we opt for crawling Amazon by web scraping. We implement the crawler using a python library, BeautifulSoup [3], to extract specific data in certain HTML tags.

Generally, the data collection consists of three steps. In each step, we store certain data and collect input data for the next step. First, we collect substantial amount of user profiles to expand our crawling targets. From each 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 that are associated with the searched name. Note that wishlists under a user profile may have different personal information, such as name and address. The search engine returns the user name, birthday, address, and list-description in the latest updated wishlist.

Next, we collect the item information in a wishlist. We directly visit the wishlist URLs extracted in the first step. In a wishlist, items are listed with links to their own web pages. We cannot know the price and type of a item until we visit these pages. Thus, up to this point, we are only able to collect item name, item page URL, and the date when the item was added.

Finally, we retrieve detailed information for each item. We visit all the item URLs and scrape the product pages. In this item page, the type and price can be easily extracted. However, there may not always be only one price for an item.

[1]The actual rate is $1 + round(\frac{S}{\$4600})$ [1], in which S denotes the sales in the user's website in last 30 days

For example, for the same item, there could be prices from different retailers. There are also price differences between new products and used products. We only record the price chosen by the wishlist owner (wishlists remember which version of the items are selected in most cases). If a user does not choose a price, we record the lowest price for which the user needs to pay. We believe that it is reasonable to choose the lowest price, because people tend to pay less to purchase the same item. However, occasionally we cannot find information for an item. There are three main reasons. (1) The item is removed from Amazon, and thus there is no web page for it. However, it dangles in the user's wishlists. (2) The item information is not available in the item page. For example, an item is no longer in stock, the price shown will be "Currently unavailable.". Or the item is not under any type so the type information cannot be retrieved. (3) Web failures and anti-crawler mechanisms thwart the attempt to retrieve the price.

B. Data Overview

We search the top 300 common male and female names[2] twice to harvest user profiles. Eventually we collect 1,233,095 unique users.Their profile information and wishlist links are stored in our database. However, collecting all users' wishlists is very time-consuming, considering that one Amazon product page usually has over 10,000 LOC, which is around 300KB data. Therefore, we only collect part of the user profile pool that is large enough to fulfill our measurement purpose. As we are also interested in personal information of a user, we collect the wishlists whose owners have include both some products and list-descriptions. To this end, we collect 30,057 complete user data, including their wishlists and all the items in the wishlists. As we search male and female names to reach the users, we naturally infer the gender of a user based on his/her first name. Our data consists of 19,976 male users and 9,541 female users. The rest of the users are unclear in gender since their first names appear in both male and female name searches. In total, we collect 76,923 wishlists and 5,710,674 items, among which 2,248,142 are unique. Note that some items do not include complete information (price, type, etc). In certain analysis, we ignore the data with missing information. For example, when computing the average price of all products, those products with unknown price are omitted.

C. Personal Information

Wishlists enable us to study the problem of personal information exposure in Amazon. As wishlists contain user personal information such as birthday and location, we explore how many users have listed these information. Table I illustrates such personal information in user profiles. It can be observed that a considerable number of users list their birthday and location information in their user profiles. Furthermore, over half of the users who have a list-description publicize their birthday and address information. Our findings confirm with a previous study [6], which presents that users tend to expose their personal information in open websites.

III. Data Analysis

Now we try to dissect our dataset to study the dynamics of user wishlists. Specifically, we are interested in the products a user added, which indicate the user purchase intentions, and the time when the products are added. We reveal user preferences in general and perform comparative analysis among different demographical locations. Moreover, we illustrate the dynamics of wishlist, in terms of the number of added items, with respect to different time periods, and compare the dynamics between weekdays and weekends, as well as between holidays and normal days.

[2]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: Data Distributions.

Distribution	Mean	Max	SD	γ_1	κ
Wishlist in Profile	2.6	326	5.0	19.7	854.2
Items in Wishlist	74.2	7,350	187.3	8.3	111.1
Item Price	36.0	105,065	188.8	227.1	94,335.4

A. Basic Statistics

First we conduct a simple analysis on the dataset to gain a basic understanding on how wishlists are used by users.

For all 1,233,095 users searched out in the first step of our data collection, we collect a total of 2,121,173 wishlists (among which 76,923 are collected together with the items in them) and over 5,700,000 items (in which 2,248,142 items are unique). We describe the distributions in Table II. As we can see from the table, every user has 2.56 wishlists on average with a standard deviation of 4.97. The average number and standard deviation of items in a wishlist are 74.2 and 187.3, respectively. And the average and standard deviation of product price are $35.06 and $172.0, accordingly. Note that the maximum number of wishlists a user can reach (similarly, items in wishlists) is unknown to us, the maximum numbers listed in Table II just reflect our observations from the dataset. We also notice that all the three distributions have very high skewness (γ_1) and kurtosis (κ), indicating that the distributions are very skewed and heavily tailed.

B. User preference

An essential question on user preference is what users would like to purchase and how much they may need to pay. To categorize the products, we directly leverage item types collected from product pages. From all product pages that are visited, we observe 50 types of products in total. Note that Amazon was reported to perform price discrimination on E-books [11], which indicates that the same E-book may be priced differently at different locations. However, these locations are in scope of countries. In our study, we focus on the U.S. only. Moreover, since only E-books are price discriminated in Amazon, our major results are still meaningful even from a global perspective.

TABLE III: Overall User Preference.

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	2,018,907	43.59%	$23.03
2	Movies & TV	544,418	11.75%	$25.60
3	Buy a Kindle	392,249	8.47%	$9.41
4	CDs & Vinyl	351,873	7.60%	$18.50
5	Toys & Games	233,584	5.04%	$41.09
6	Video Games	104,926	2.27%	$47.09
7	Amazon Fashion	102,391	2.21%	$59.04
8	Kitchen & Dining	100,100	2.16%	$46.33
9	Sports & Outdoors	95,970	2.07%	$59.88
10	Home & Kitchen	85,439	1.84%	$59.09

The top 10 nation-wide most favourate product types are listed in Table III. The average price of all products is $34.13. With over 40% of items in a wishlist being books, we can see that books are in domination in user wishlists. Besides normal paperback books, E-books in Kindle are also very popular. We believe that the low price of an E-book is the main reason for its prosperity (Books are 144.7% more expensive than E-books on average). Other than books, entertainment products also play an important role in wishlists, as Movies & TV, CDs & Vinyl, Toys & Games, and Video games rank 2, 4, 5, and 6, respectively.

After knowing the general shopping preference of Amazon users over the entire U.S., it is meaningful to detail the preferences of people from different demographical groups. Specifically, we compare the shopping preference of males and females, as well as the preference of users from three different geo-locations—east coast[3], west coast[4], and inland of the U.S.

1) Different Gender: We show the top 10 popular products of male users in Table IV and those of female users in Table V. The average product price for male is $37.53 and that for female is $26.47. It is clear that male and female users have different shopping preference. First of all, on average the products in males' wishlists are 41.8% more expensive than those in females' wishlists. Different genders also prefer different types of products. While "Books" and "Buy a Kindle" account around 50% of the products for both genders, males are more likely to buy a paperback book, instead of E-books, than females. Not surprisingly, males prefer sports and electronics products while females prefer fashions and beauty-related products. Interestingly, females like "Arts, Crafts & Sewing" (with rank 10) much more than males as such products only count 0.28% of all products and ranks 24 in all categories for males. Although male and female behave similarly in the top 5 product categories, their differences, in terms of the percentage of a product type in wishlists, are significant in the rest of the categories. In particular, the differences are more than 100% in 13 categories[5] out of the 24 most popular ones.

2) Different Region: In our dataset, there is very small difference among three regions: east coast, west coast, and inland. For example, users in the inland expose the slightly higher price sensitivity, specifically, products in their lists are 6.6% less than the other 2 parts. Due to space limitation, we omit discussion about these small variations among regions, which may not be much useful.

[3]en.wikipedia.org/wiki/East_Coast_of_the_United_States

[4]en.wikipedia.org/wiki/West_Coast_of_the_United_States

[5]*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 IV: Male User Preference.

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	1,458,599	45.78%	$25.16
2	Movies & TV	395,867	12.42%	$26.37
3	CDs & Vinyl	277,802	8.72%	$19.45
4	Buy a Kindle	208,237	6.54%	$10.60
5	Toys & Games	143,937	4.52%	$45.40
6	Video Games	81,858	2.57%	$48.39
7	Sports & Outdoors	76,460	2.40%	$62.10
8	Home Improvement	59,929	1.88%	$66.17
9	All Electronics	50,460	1.58%	$135.09
10	Amazon Fashion	48,622	1.53%	$71.58

TABLE V: Female User Preference.

Rank	Item Type	Number of Items	Percentage of Items	Average Price($)
1	Books	529,011	38.72%	$17.49
2	Buy a Kindle	176,876	12.94%	$8.01
3	Movies & TV	139,406	10.20%	$23.58
4	Toys & Games	82,767	6.06%	$33.99
5	CDs & Vinyl	69,132	5.06%	$14.92
6	Amazon Fashion	51,794	3.79%	$48.08
7	Kitchen & Dining	49,170	3.60%	$38.47
8	Home & Kitchen	42,404	3.10%	$50.12
9	All Beauty	25,447	1.86%	$21.82
10	Arts, Crafts & Sewing	21,497	1.57%	$22.75

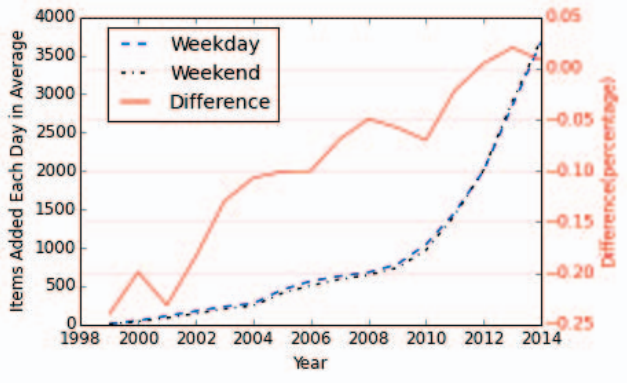
In summary, we observe that male and female bare quite different online shopping preferences, while that of people from the three different regions do not differ much.

C. Time Factor

We then study the impact of time factor on the dynamics of wishlists as a further characterization of user preferences. From over five million collected items, we can see that users have started to add items to wishlists since 1999. While only 2,056 items were added in 1999, there was an 85.6%. increase on average for each year thereafter. This rapid increase is due to the fact that electronic commercials have been gaining more popularity each year. We first explore and compare the dynamics of wishlists, in terms of the number of added items, during weekdays and weekend. Then we conduct the similar comparison between normal days and holidays, and we are able to learn the time that are most appealing to online shoppers.

1) Weekdays and Weekends: Intuitively users may browse and shop online more on weekends than weekdays since they have more free time. Surprisingly, our analysis indicates that this intuition does not hold for the dynamics of wishlists. Overall, 3,905,728 and 1,523,764 items are added in 4,174 weekdays and 1,670 weekend days, respectively, in our dataset. On average, 935.7 items are added on each weekday and 912.4 items are added on each weekend day, implying that a weekend day has even 2.55% less added items than a weekday. We show the general trend in Figure 2, in which the gap between the number of added items in weekdays and in weekend days is larger in earlier years but keeps shrinking. The number of items added in weekends start to exceed that in weekdays after 2012. However, there is no evidence showing that this trend would continue. It is fair to say that currently people browse and update their wishlists almost equally during weekdays and weekends. A reasonable explanation for this equality lies in the fact that it becomes drastically easy for users to browse and update their wishlists online, in oppose to the traditional walk-in-shop style which an extensive amount of time is needed.

Fig. 2: Items Added in Weekdays and Weekends.



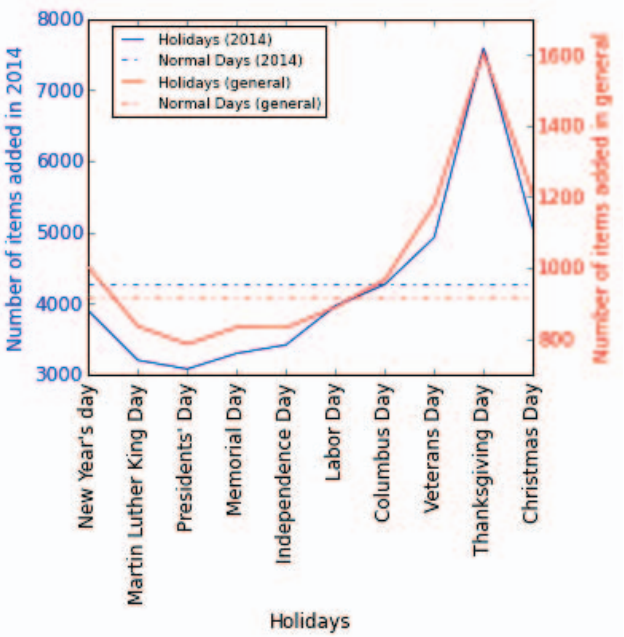
2) Normal days and Holidays: we also compare the change of wishlists between holidays and normal days. Since there are many unofficial and regional holidays, our study only focuses on the 10 nation-wide federal holidays, which can be seen in Figure 3.

When shopping for holidays, users may not always buy products on the exact date. For example, people usually prepare gifts before Thanksgiving, or they may forget to buy someone a gift but try to make up afterwards. Therefore, we define the five consecutive days centered on a holiday as a holiday season (i.e., the previous two days, the holiday, and the following two days). For example, we include December 23 to December 27 to represent the Christmas season. Note that as a side effect, the last two days of the previous year are also included when analyzing a New Year's holiday season. However, it does not affect our results since we are only interested in the average number of added items in normal days and holidays.

The average numbers of added item in normal days and holidays are 964.4 and 893.3 respectively. Surprisingly, we observe that the difference between holiday average and normal day average is minor. The difference between the two averages is only 8%.

Our observation above does not match the common impression that people shop significantly more in holidays. Such a counter-intuitive conclusion makes us realize that it is not appropriate to group all national holidays together since different holidays may have very different social and economical interpretation. As such, we handle each individual holiday separately. To show both the overall case and the current trend, we use the data both from 1999 to 2014 and the year 2014 solely. The results of this analysis are depicted in Figure 3. It is obvious that the dynamics of wishlists during a holiday season are very different from others. We notice that 3 holiday seasons (Veterans day, Thanksgiving, and Christmas) that have significant shopping inclination. By contrast, in 5 holiday seasons (Martin Luther King Jr. Day, Presidents' Day, Memorial Day, Independence Day, and Labor day) people add less items in their wishlists than normal days. One possible explanation is that people are more likely to be engaged in social activities rather than shopping online during these holidays, since four out of the five holidays are memorial-type days. The curve for year 2014 incur deeper pit in these holidays, which indicates that nowadays people's shopping behavior becomes more seasonal. I.e., People are more prone

Fig. 3: Items added in Different Holidays.



to bind their online purchasing to specific time. Thus, the general impression that people shop online more in holidays only partly stand. On the contrary, quite a few holidays do not seem enticing for online shoppers.

D. Result Implications

We analyze three aspects of user online shopping preference in Amazon, including product categories, price, and time factor, which may be used to affect user shopping behaviors in the future. Since Amazon is the largest general-purpose electronic commercial in the U.S., we believe our analysis can shed light on the entire e-commerce industry.

Our insights on user shopping preference can be used both in general education purpose and commercial related purpose. Understanding user shopping preference is an important factor in e-commerce ecosystem. Equipped with this knowledge, online retailers are able to better predict the popularity of products, refine marketing strategies, and perform timely promotions. Besides, since our results show various user shopping preference from different user groups, it can help online advertisers to develop and distribute more accurate targeted advertisements.

IV. Privacy Information Exposure

While personal data such as location and birth date are recorded in user wishlist profiles, we perform a study to explore more potential information exposure in Amazon wishlists. Amazon features a list-description for each wishlist in plaintext that is open to public access. However, besides simply describing the wishlist, users may also inadvertently put personal information that has nothing to do with the list. Naturally, we have interests to better understand how users may leak privacy information in their online wishlist profiles. We assume all users are telling the truth in their list-descriptions since users have no incentive to lie in their list-descriptions

and those who concerns about their privacy can simply leave the description empty.

A. List-Descriptions

We study the list-descriptions associated to the default lists of all 30,057 users. In average, a wishlist consists of 11.9 words. However, the median number of words in a description is only 6, which implies that while few users put long monologues in their list-descriptions, most users are more likely to put a few sketchy words. by using Stanford Part-of-Speech (POS) tagger tool [15] to parse the descriptions, we observe that approximately 40% of the words are nouns, where high ratio of nouns is able to provide more information regarding to the users. Some of the most frequent meaningful nouns are "university" (959 occurrences), "books" (878 occurrences), "music" (704 occurrences), "school" (604 occurrences), and "movies" (363 occurrences). However, these discreet words cannot be used to generalize the types of personal information users tend to put in their wishlists. Appropriate level of abstraction is necessary to make better depiction. For instance, to summarize to what extend users include their education information in list-descriptions, we may abstract all hyponyms of the word "education," which include "University," "College," "School," "graduate," etc. To abstract the hyponymy relations in words of list-descriptions, we use Wordnet [12], which is a directed graph database that connects English words with relations. Wordnet enables us to group related words into more general words. A previous work on password semantics [17] uses similar approach to extract semantic pattern from passwords and focuses on leveraging tree-cut model [10] to balance size of the cut and abstraction level. However, their approach is not particularly suitable in our case, where we are specially interested in optimal levels of word abstractions for identifying privacy exposure. For example, we may prefer the abstraction "relative" more than "person" or "sister" since "person" is too general and "sister" is too specific.

However, it is very hard to automatically generalize words in an optimal level. Thus, we manually select seven representative word abstractions and use them to show how much personal information users expose in list-descriptions. We first find all synonyms of the word. Then, for each of the synonyms, we find its hypernyms up to five levels. Note that we do not need to find all its hypernyms since higher level words are usually too general (such as people, entity, etc.) to be useful. Moreover, we keep the semantic sense close to the target word by restraining its hypernym levels. After obtaining a word pool with words related to the target word in a similar or more general level, we search the abstraction word (that we preselect) in the word pool. If a match is found, the word is then considered related to the abstraction. To achieve a higher accuracy, we apply such abstraction only on nouns since other words such as verbs carry little information and the generalization of such words may make the semantic meaning drifted.

The selected abstraction words as well as their frequencies are listed in Table VI, which clearly shows that users do publicize their personal information in list-descriptions. Particularly, we notice that a significant portion of users expose their activities (32.68%) and affiliations (24.09%). 8.47% of users mention their education background, and 4.56% of users put occupations-related information in list-descriptions. These essential information can help re-construct the user profiles. Furthermore, 9.58% of users talk about their relatives, and 2.6% of users mention their spouses in wishlists, which also indicates their marital status.

TABLE VI: Personal Information Exposure.

Abstraction	Occurrence in list-descriptions	Percentage
activity	9,822	32.68%
social_group	7,240	24.09%
relative	2,880	9.58 %
educational_institution	2,545	8.47%
professional	1,371	4.56%
sport	1,289	4.29%
spouse	787	2.60%

V. Personal Information Identification

User information exposure in Amazon wishlists is directly attributed to user behaviors. Since users choose to publicize their personal information, they are largely responsible for such privacy leakage. Now we study the potential leakage of privacy information that users choose not to publicize. Particularly, we conduct a pilot study on how to identify user gender from the products in their wishlists. We adopt Support Vector Machine (SVM) to learn the collected data and predict gender of a newly entered user.

Note that we choose gender to study because of the cleanness and easier access to ground truth, which is obtained from users that provide their real names. All users in our dataset are collected through a name search, and it is easy to distinguish male and female users. We ignore those users whose names are found in both male and female searches. We selected four user features to train our SVM, and they are: the fraction of number of products in one category to the total number of products in the 13 categories that show strong gender implications as mentioned in Section III-B1, the total number of items, the average item price, and the highest item price. To this end, a 16-dimensional vector is used to describe a user.

We conduct five experiments to better illustrate how accurately wishlists can be used to imply user gender. For each experiment, we trim the training and testing sets differently. We randomly select 2,500 qualified males and females and use 80% of the males and females as the training set and the rest 20% as the testing set. Thus, our experiments consist of 4,000 training users and 1,000 testing users, in both sets half of users are males and half of users are females. Furthermore, we use 5-fold cross-validation to ensure our results are more accurate. Table VII shows the result of prediction. The accuracy is 72% in Experiment 1 when no date selection is made. Though it is helpful in predicting the gender of a user, it may be biased since many users only have few items in their wishlists. We construct Experiment 2 that requires the users to have at least 1 of the 13 categories that have strong gender implications and Experiment 3 that only studies users with a relatively abundant products in their wishlists. These date refinements improve the prediction accuracy to 76%-78%. When we further refine users with abundant products in the 13 gender implication categories

TABLE VII: 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 the strongest gender indication in the training sets and the testing sets. Similarly, the column *Train All* and *Test All* specifies the number of products under all categories in the training sets and the testing sets.

in either the training set or the testing set (Experiment 4 and 5), we can achieve a better prediction with over 80% accuracy.

Our SVM gender model can be used to predict the user genders solely based on the product items in the user's wishlists, especially when users have a large number of products in their wishlists. Thus, even if a user wishes to protect his/her personal information (such as gender) from the public, the users' wishlist-type data may be used to successfully derive such personal information.

VI. Data Limitation

The user wishlists are not always publicly accessible, since users can change the accessibility of their wishlists (although the default setting is public). Therefore, privacy-aware users may choose to publicize only part of their wishlists and share no information on privacy-sensitive items such as pregnancy test, firearm-related products, and medicine & drugs. Since we can only retrieve the products in public wishlists, the collected data may be biased and not 100% representative of one user's shopping intention and behavior.

VII. Related Work

Researchers have been working on analyzing online e-commerce data and presenting critical observations [8], [9], [11]. For example, Ghose et al. [8] studied the review text of 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 influenced by product reviews.

Online privacy is becoming a major user concern; Brown et al. [4] showed that privacy invasion has significant negative impacts on online purchasing behaviors and Tsai et al. [16] showed that users are prone to use privacy protective websites when privacy information is salient. It has been proven that when inferring user personal information based on public data is possible, user privacy is under serious threat [5], [13], [18]. For instance, Wondracek et al. [18] leveraged user group membership to uniquely identify an individual user or at least reduce the uncertainty. Chaabane et al. [5] studied the privacy leakage through user interest in music and were able to infer user personal information.

VIII. Conclusion

In this paper, we investigate Amazon wishlists that record users' desired products. After collecting over 30,000 wishlists, we first measure the user preference from different demographical groups. We observe that there is a large discrepancy between male and female on both preferred product categories and prices, but people from different regions inside the U.S. do not differ much in their shopping preferences. Furthermore, we reveal that users may not have more online shopping items added in wishlists during holidays or weekends, over half of the national holidays with even lower enthusiasm for shopping than normal days. We then study user information exposure in Amazon wishlists by parsing and analyzing list-descriptions. It is also shown that the information in wishlists has potential to leak a user's personal information.

Acknowledgement

We would like to thank the anonymous reviewers for their insightful comments. This work is partially supported by NSF grant CNS-1618117 and ONR grants N00014-16-1-3214 and N00014-16-1-3216.

References

[1] Amazon product advertising api. https://affiliate-program.amazon.com/gp/advertising/api/detail/main.html.

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

[3] Beautifulsoup. http://www.crummy.com/software/BeautifulSoup/.

[4] 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.

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

[6] 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.

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

[8] A. Ghose and P. G. Ipeirotis. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE TKDE*, 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] H. Li and N. Abe. Generalizing case frames using a thesaurus and the mdl principle. *Computational linguistics*, 1998.

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

[12] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995.

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

[14] Statista. Number of worldwide active amazon customer accounts from 1997 to 2014. http://www.statista.com/statistics/237810/number-of-active-amazon-customer-accounts-worldwide/.

[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] 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.

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

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