Privacy in Online Review Sites

Matthew Burkholder and Rachel Greenstadt

Department of Computer Science
Drexel University
Email: mpb37@drexel.edu and greenie@cs.drexel.edu

Abstract

The increasing use of online review sites is creating new challenges for user privacy. Although reviews are public, many users inadvertently disclose private information about relationship, location, and temporal attributes to the world. This research protects users of online review sites from the inadvertent disclosure of private information in three ways. First, the types of unstructured and structured information made public by online review sites are characterized and used to grade those sites on their attention to privacy. Second, a privacy-check tool that uses keyword matching and named-entity recognition to annotate potentially sensitive review text is presented. Third, we raise awareness of the privacy threat in online review sites through examples and statistics derived from the privacy-check tool.

I. Introduction

Online review sites are sites which publish user reviews to help other users make decisions. The use of these sites is increasing. According to its about page, Amazon is the global leader in e-commerce, and each Amazon customer is encouraged to review their purchased products. The Netflix Q3'11 financial press release stated that they had over 25 million subscribers. Each Netflix subscriber has the ability to post and read online reviews. According to its fact sheet, Yelp had more than 61 million monthly visitors in Q3'11. Yelp specializes in user-generated reviews of restaurants and local businesses. The OpenTable about page says that it is a leading provider of free, real-time online restaurant reservations and has seated more than 250 million diners since its inception. Users are encouraged to review each restaurant they visit. The TripAdvisor about page boasts that it is the largest travel site and a pioneer of user-generated content. It has over 20 million members.

Many users of online review sites inadvertently disclose private information in their reviews. First, private information can include the relationships of users, such as the fact that a user has a brother. Second, it can include the locations that users visit or the location of user homes. Third, private information can include temporal data such as time-of-day, a specific date, or a special occasion. For the purposes of this research, an attack is defined as the ability to acquire private information about a person from their online reviews. There are two attack scenarios. In the first scenario, the attacker does not know the identity of the user and must determine it directly from the review. In the second scenario, the attacker does know the identity of the user and can combine private information disclosed in reviews with information from other sources. This research focuses on the second attack scenario.

This research is not concerned with the significance or impact of the disclosure of private information. For instance, users may or may not care about the disclosure of more detailed location information when reviewing restaurants or hotels, which already reveal a location implicitly. This work only tries to determine whether or not such detailed private information is disclosed. This research is not concerned with business relationships connected to the items or places reviewed. For instance, revealing that a user dined at a restaurant with his or her manager is considered the disclosure of private information, but revealing the name of the restaurant manager is not considered the disclosure of private information with respect to the user.

The contributions of this work are the characterization of types of unstructured and structured information made public by online review sites, the development of a privacy-check tool that uses keyword matching and named-entity recognition to annotate potentially sensitive review text, and the increased awareness of the privacy threat in online review sites through examples and statistics. Section 2 of this paper gives an overview of related work. Section 3 discusses the approach taken in this research. Section

DOI 10.1109/SPW.2012.23

4 evaluates the selected online review sites. Section 5 describes the technical implementation of the privacy-check tool. Section 6 provides the results of checking a sample of online reviews and evaluates the effectiveness of the privacy-check tool. Section 7 discusses the implications of the results and how the privacy-check tool might be made available to actual users. Section 8 suggests future work, and section 9 concludes this paper.

II. Related work

Mao et al. [4] characterized the nature of privacy leaks on Twitter and focused on users divulging vacation plans, tweeting under the influence of alcohol, and revealing personal medical conditions. They used keyword matching and built automatic classifiers to detect incriminating tweets for these three topics in real time. They also characterized who leaks information and how by studying self-incriminating primary leaks versus secondary leaks that reveal sensitive information about others. The researchers hoped that future guardian angel systems using similar classification techniques could alert users to privacy leaks and give them the option to remove tweets or think twice about posting a sensitive tweet. The differences between privacy leaks in Twitter and privacy leaks in online reviews are the shorter length of tweets, the real-time nature of Twitter, and the general-purpose use of Twitter. Additionally, the research in this paper focuses on broad keyword categories rather than specific topics and also compares results across multiple review sites.

Most of the research in online reviews is focused on opinion mining, which is the extraction of opinion features from sets of reviews. Dave et al. [1] developed an opinion mining tool that would generate a list of product attributes (quality, features, etc.) and aggregate opinions about each of them (poor, mixed, good). They developed a classifier for distinguishing between positive and negative reviews. Minqing Hu and Bing Liu also mined product features from customer reviews [3]. More recent work on opinion mining and sentiment analysis by B. Pang and L. Lee [5] briefly acknowledges the broader issues of privacy, manipulation, and economic impact of online reviews. A difference between the opinion mining research and this work is the focus on privacy as well as the inclusion of restaurant, hotel, and movie reviews rather than just product reviews.

One other related research area worth mentioning is data loss prevention. Hart et al. [2] presented an automatic text classification algorithm for classifying enterprise documents as sensitive or not sensitive, which had a false-negative rate of less than 3.0% and a false discovery rate of less than 1.0%.

III. Approach

The approach to this research started with the selection of online review sites to analyze. Then, the selected sites were graded according to how much structured information about users they reveal. A privacy-check tool was developed and a sample of items and item reviews were checked. Finally, the effectiveness of the privacy-check tool was measured by false-positive rates.

A. Online Review Site Selection

Online review sites were selected based on popularity and on the types of items reviewed. The five sites chosen for this research were Amazon, Netflix, Yelp, OpenTable, and TripAdvisor. Amazon was chosen for being the most popular online retailer and for its user reviews of consumer products. Netflix was chosen for its large DVD rental user base and for its user reviews of movies. Yelp was chosen for its large user base and for its user reviews of restaurants and local businesses. OpenTable was chosen because of its extensive use in making online restaurant reservations and for its user reviews of restaurants. TripAdvisor was chosen for its popularity with travelers and for its user reviews of hotels and other travel related content.

B. Online Review Site Grading

The selected online review sites were graded based on whether or not they made certain types of personal information public. A comprehensive list of types of personal information was determined by observing all selected review sites. For instance, one review site revealed the review history of its users, so this was added to the list. Another site, which does not reveal the review history of its users would then score positively for not revealing that personal information. Sites were graded based on structured information, which is information that is always present for a user or a user review and able to be parsed. Unstructured information such as relationships, location and temporal attributes, and the names of people, places, and organizations are only present in user profiles or review text. Sites were not graded based on unstructured information.

C. Development of Privacy-Check Tool

The privacy-check tool was developed to scrape review text from the online review sites, match and annotate keywords, recognize and annotate named-entities, and gather statistical counts. Three keyword categories were used for keyword matching: relationship, location, and temporal. Relationship keywords are intended to catch potentially

inadvertent disclosures of relationships or information tied to relationships. Examples include a product purchased for a sibling or dining at a particular restaurant with a friend or co-worker. Location keywords are intended to catch potentially inadvertent disclosures of user location or the location of a user's home. Examples include dining at a restaurant a block away from the user's apartment or watching a movie that was filmed nearby where the user lives. Temporal keywords are intended to catch potentially inadvertent disclosures of personally meaningful dates, where the user was at a given time, or how often the user goes somewhere. Examples include dining at a restaurant for a friend's birthday or revealing that the user is on a business trip once every month. Technical details for the privacy-check tool are discussed in section 5.

D. Item Selection

For each online review site, 10 items were selected for review scrape and analysis. Items were selected based on popularity because of the expectation that more popular items would have a greater number of lengthy reviews. For Amazon, the focus was narrowed to include men's and women's accessories and kids toys. The idea was that reviews for these types of items might have a higher incidence of relationship and temporal disclosures since they are often given as gifts. For Netflix, the focus was narrowed to romantic comedies and family films. These types of items were expected to have a higher incidence of relationship disclosures. For Yelp and OpenTable, the focus was narrowed to popular restaurants in the Philadelphia area. All restaurant reviews were expected to have a higher incidence of relationship and location disclosures regardless of cuisine or style. For TripAdvisor, the focus was narrowed to popular hotels in the Philadelphia area. All hotel reviews were expected to have a high incidence of all three types of disclosures.

E. Results and Evaluation of Privacy-Check Tool

The privacy-check tool contains functions for deriving aggregate match counts from the reviews. The analysis of the privacy-check results focused on the number of matched words as a percentage of total word count by site for each keyword category. The privacy-check tool also dumps annotated reviews to an HTML file. This file was used to perform a manual inspection of the reviews for the determination of false-positives and false-negatives. A false-positive occurs when a word is annotated but does not disclose any private information. An example of a false-positive is if a user were to discuss the brother of a movie character in their review, which reveals nothing about the user's brother. A false-negative occurs when no word is annotated but private information is disclosed. An example of a false-negative would be if a user visited a particular island to use a product they purchased, but the keyword "island" was not included in the location keyword list.

	Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No
Item location		X		X	X		X		X	
Item rating	X		X		X		X		X	
Date of review	X			X	X		X		X	
User identity	X			X	X			X	X	
User join date		X		X	X		X		X	
User rating	X			X	X			X	X	
User location	X			X	X			X	X	
User age		X		X		X		X	X	
User gender		X		X		X		X	X	
User friends		X		X	X			X		X
Review History	X			X	X			X	X	

Fig. 1: Structured information revealed by online review sites

IV. Site Evaluation

The online review sites were graded based on how much structured personal information they made public. The publication of unstructured personal information is under the control of the user and is the target of the privacy-check tool. The lists of structured and unstructured personal information are:

- Structured—item location, item rating, date of review, user identity, user join date, user rating, user home location, user age, user gender, user friends, review history
- Unstructured—relationships, proximity, location, travel history, travel plans, visitation or purchase time, visitation frequency, special date, personal schedule, names of people, names of places, names of organizations

Figure 1 shows what structured information online review sites revealed. Netflix revealed the least structured information revealing only the user ratings alongside their reviews. OpenTable also revealed minimal structured information revealing only item location and rating, date of review, and user join date. Neither site associated user identity with reviews, which is a huge privacy advantage for these sites. Amazon revealed an average amount of structured information including item rating, date of review, user identity, user rating, user location, and user review history. Amazon did not reveal user join date and there is no notion of Amazon friends. Yelp and TripAdvisor revealed the most structured information. Yelp revealed everything except user age and gender. TripAdvisor revealed everything except user friends.

There are some things worth noting about structured information. First, Amazon reveals a user ranking based on the number of and helpfulness of reviews, and this was counted as a user rating. Second, Amazon optionally reveals user birthdays, which is likely to encourage more frequent exchange of gifts. Third, OpenTable reveals user join date but only by the year. Fourth, gender can sometimes be inferred from user profile pictures or pseudonyms in Amazon, Yelp, and TripAdvisor reviews. To summarize the results of the site evaluation:

- Netflix revealed the least amount of structured information
- OpenTable was a close second with respect to not revealing structured information
- Amazon revealed an average amount of structured information
- Yelp and TripAdvisor revealed the most amount of structured information

V. Technical Implementation

Review scraping, keyword matching, name recognition, review annotating, and match counting were done using the privacy-check tool developed specifically for this research. The tool is implemented in Python making it well-suited for interactive use. The tool takes local HTML files containing review text as input.

All but one of the review sites retrieve review text from the site server after the page as finished loading using asynchronous JavaScript requests. For this reason, the input HTML files must contain dynamic content and cannot merely be retrieved using a tool such as wget. We use Firebug, a web development add-on for the Firefox web browser that supports real-time inspection of page contents, to dump the dynamic content of HTML pages containing review text to local HTML files.

Reviews were scraped from these HTML files using Beautiful Soup, which is a Python HTML/XML parser designed for quick turnaround projects like web scraping. Review scraping is specific to each review site and how its developers decide to structure their HTML.

A. Keyword Matching and Name Recognition

Each of the three keyword categories was further divided into sub-categories for the process of brainstorming keywords. The sub-categories for relationship keywords are significant other, family, extended family, friend, teammate, classmate, co-worker, and formal relationship such as doctor, landlord, or priest. The sub-categories for location keywords are transportation mode, distance, vacation or travel, administrative division, building, and landmark. The sub-categories for temporal keywords are time relativity, time unit, named days and months, time periods, time frequency, special occasions, holidays, and scheduled events. Keywords were brainstormed for each sub-category and stored as static Python lists in the privacy-check tool source code. A listing of the keywords can be found on our website[1].

Keyword matching is done using regular expressions to split each review into a list of words and non- words (i.e., white space, punctuation, and numbers). Words are then matched against keywords in the given keyword list. For each match, the word is annotated before re-joining the list of words and non- words, and a counter corresponding to the keyword is incremented for that review. Name recognition is done using NLTK to scan each review. Names are annotated in the same way as words matching keywords.

B. Annotating and Match Counting

Keyword matches and recognized names are highlighted a certain color by enclosing them in a HTML font tag with a specific value for the color attribute. Relationship, location, temporal keywords are highlighted red, green, and blue respectively. Named-entities are highlighted orange. The tool dumps annotated reviews separated by a line using the HR tag to a HTML file. This HTML file is what was used to manually count false-positives and false-negatives. Keyword matches and recognized names are counted per review. The tool contains functions for deriving aggregate counts. Aggregate counts include the total number of matches for a given keyword category and the total number of matches for individual keywords. These counts can be determined by item, site, or all sites.

VI. Results and Tool Evaluation

Results include the number of keyword matches by category as a percentage of total word count and the number of named-entity matches and actual named-entity counts. The privacy-check tool evaluation includes an analysis of the false-positive and true-positive rates and a list of false-negatives found.

A. Keyword Matches

Figure 2 shows keyword matches by category as a percentage of the total word count for each online review site. Total word count is the number of words in all reviews checked for a given site. The percentage of total

[1] http://psal.cs.drexel.edu/files/reviewkeywords.pdf

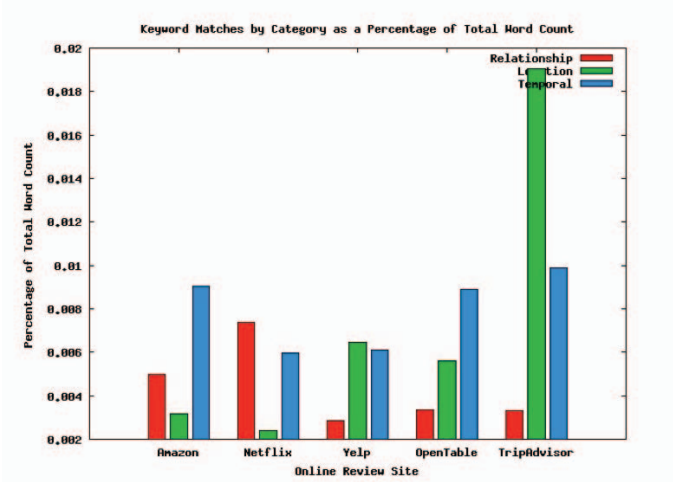

Fig. 2: Keyword matches by category as a percentage of the total word counts

Online Review Site	Matches	Actual Count	True Positives
Amazon	233	6	4
Netflix	697	0	0
Yelp	787	12	0
OpenTable	247	15	1
TripAdvisor	476	41	0

TABLE I: Number of named-entity matches, number of actual named-entities, and number of true-positives (names of users or people known by the users)

False-positive Type	Example
Homonyms	"she said that many handbags **state** they are all leather"
Wrong Context	"I've **moved** onto a Codura nylon wallet"
Ambiguous	"think short, matronly woman in **airport** with rolling laptop case"
Non-specific	"thankfully I was **nearby** and quickly got to him"

Fig. 3: Types of false-positives with corresponding examples of location false-positives

words matched ranged from 0.2427% for Netflix location keywords to 1.9071% for TripAdvisor location keywords.

The percentages of relationship keywords were similar for TripAdvisor, Yelp, and OpenTable. The percentage of relationship keywords for Amazon was noticeably higher than that of TripAdvisor, Yelp, or OpenTable, and the percentage of relationship keywords for Netflix was noticeably higher than that of Amazon. However, as will be shown below, Netflix had a high false-positive rate for relationship keywords while Amazon had the lowest false-positive rate for relationship keywords. Overall, Amazon had the highest percentage of true-positive relationship keywords, and this is likely due to a higher incidence of product recommendations by people the users know as well as giving products as gifts.

The high percentage of location keywords for TripAdvisor makes intuitive sense since users primarily review hotels when they are traveling. The low percentages of location keywords for Amazon and Netflix are likely due to users trying products or watching movies at home. The percentages of location keywords for Yelp and OpenTable were several times greater than those for Amazon and Netflix but still nowhere near that of TripAdvisor. This is probably due to TripAdvisor users describing multiple locations to which they traveled in a single review, while Yelp and OpenTable users describe only the location of the single restaurant they visited.

The percentages of temporal keywords were more consistent across sites. Yelp and Netflix were close to a tie for the lowest percentages and TripAdvisor had a slightly higher percentage than either Amazon or OpenTable. As will be shown below, other than for Netflix, the false-positive rates for temporal keywords are also fairly consistent. This seems to imply that users are almost equally as likely to reveal private information about age, time, date, or special occasion on any review site.

B. Named-entity Matches

Named-entity recognition using NLTK was mostly ineffective. Table I shows the number of named-entity matches found using NLTK. The number of actual named-entities was determined by manual verification of the matches. Most of these names were names of waiters, managers, directors of guest services, or other business relationships tied to the item being reviewed. A true-positive is when a user reveals their name or the name of someone they know. Only 5 true-positives were found in all 1500 reviews checked. Furthermore, one name which was discovered using relationship keyword analysis was not recognized using named-entity recognition. The only true-positives found were the names of users that signed their reviews or began their review by stating their name.

C. False-positive and True-positive Rates

False-positives were counted per keyword for each online review site. In order to identify false-positives, four false-positive types were defined. Figure 3 shows each of these types along with an example. A homonym false-positive is when the keyword is used with a different meaning. A wrong- context false-positive is when the keyword is used with the intended meaning but in a context that does not reveal relationship, location, or temporal information about the user. An ambiguous false-positive is when the keyword is used with the intended meaning and context, but the user may be speaking hypothetically. Finally, a non-specific false-positive is when the keyword is used with the intended meaning and context, but no

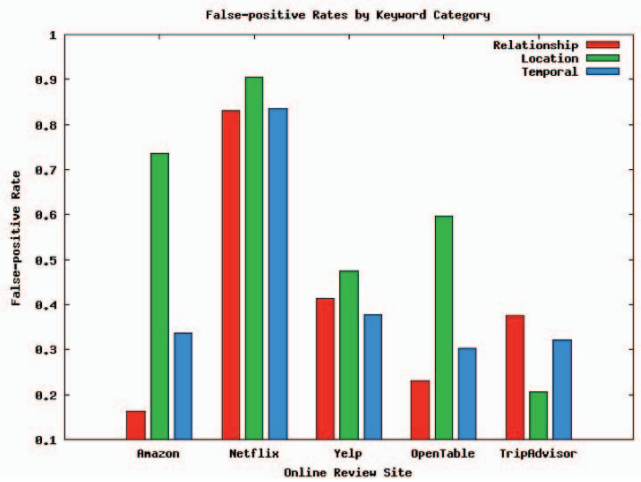

Fig. 4: The false-positive rates by keyword category for each of the sites

specific location or temporal information is revealed. For the example in table 3, the user could have been anywhere when they were nearby him. On the other hand, if a user were to reveal that they were at the beach without revealing which beach, this is considered specific enough to be a true-positive since anyone reading the review now knows that this person visits a beach. This means that a true-positive does not have to reveal an exact location or time. Taking the four types of false-positives into account, a true-positive is when the keyword is used with the intended meaning and context and is not ambiguous or non-specific.

Figure 4 shows the false-positive rates as a percentage of the total words matched in each category for each site. It is clear from the histogram that Netflix reviews had the highest percentage of false- positives for all three categories. This can be explained by the fact that Netflix users discuss the characters, plots, and settings of movies, so any keywords matched are more likely to be about those movies. That is not to say that there were no private information disclosures in Netflix reviews. The following Netflix true-positives are a sample of those found through keyword matching:

- "This is one of my favorite movies. My **friend** Cristen is also a huge fan of it."
- "I had the opportunity to see Howl's Moving Castle at a **preview** screening about a **month** ago"

Amazon had the lowest false-positive rate for relationship keywords. This seems to be in part due to users mentioning their relationships when they review products intended as gifts. Amazon users tended to mention the ages of the recipients of reviewed products and also special occasions or events related to the purchase. Yelp, OpenTable, and TripAdvisor also had relatively low false-positive rates for relationship keywords, and this is likely due to users dining or traveling with people they know. The following Amazon true-positives are a sample of those found through keyword matching:

- "My **son** decided to use it for a **trip overseas** instead of the "regular" passport holders."
- "We are going to a **birthday party** this weekend and our **daughter** will be giving the **birthday boy** the..."
- "My **son** will turn 3 **years old** next **week**."

TripAdvisor had the lowest false-positive rate for location keywords. This is due to users describing in detail their location when on vacation or traveling. It is likely that users are less concerned with revealing this information since their general location is already revealed by the review itself. The following TripAdvisor true-positives are a sample of those found through keyword matching:

- "I only **live** a few **blocks** from the hotel."
- "We **stayed** there on 10/21/11 and 10/22/11 in room 1722."
- "We just checked in earlier **today**."
- "I'm here right now (09/18/11). Arrived **yesterday** and leave **tomorrow** for a total of 3 **nights**."

Yelp and OpenTable both had average false-positive rates for location keywords. OpenTable had somewhat lower false-positive rates for relationship and temporal keywords. Users tended to reveal information about the location of their homes as well as information about special occasions. The following Yelp and OpenTable true-positives are a sample of those found through keyword matching:

- "Honey's is a short 5-**blocks** from my **house**, a pretty nice **walk**."
- "I came to this restaurant because I **live** less than one **block** away at the Ben."
- "We recently **moved** to Philadelphia from the **bay** area."
- "I go probably every other month since I **live** in the **neighborhood**."
- "While living in Philly for the **past** two **years** Village Whiskey was a staple."
- "Went here for my **birthday** on 10-6."
- "Paul suggested that we become his **Sunday** night **regulars**, and I think that I will have to make that happen."

Figure 5 summarizes the most frequent false-positives by keyword category and by online review site. Some of the most frequent false-positives for relationship keywords were: children, kids, family, group, manager, and pa. The word "manager" was used to refer to the manager of a restaurant or hotel rather than the manager of the user. The word "pa" was used to refer to the state of Pennsylvania rather than a father. Frequent false-positives for location keywords were: far, minutes, home, and capital. Many

Most Frequent Relationship False-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
baby	4	children	15	group	8	group	4	manager	13
child	4	kids	14	boy	5	family	4	group	5
children	4	family	13	family	4	manager	3	pa	5
Most Frequent Location False-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
far	19	minutes	9	far	17	capital	13	minutes	23
minutes	14	far	8	minutes	14	walked	9	walk	14
floor	8	home	6	home	12	house	8	town	12
Most Frequent Temporal False-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
fall	15	old	27	party	15	party	12	old	23
old	14	year	15	old	12	old	8	independence	20
return	12	years	12	late	6	later	6	morning	18

Fig. 5: Most frequent false-positives by keyword category and by online review site

Most Frequent Relationship True-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
son	30	kids	5	friend	19	husband	21	husband	20
daughter	24	wife	5	husband	12	wife	15	wife	12
baby	20	daughter	4	boyfriend	9	boyfriend	10	family	11
Most Frequent Location True-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
house	6	school	3	street	22	visit	12	stayed	110
home	5	college	2	visit	12	visited	7	walk	55
trip	4	home	1	trip	11	visiting	5	walking	53
Most Frequent Temporal True-positives									
Amazon		Netflix		Yelp		OpenTable		TripAdvisor	
old	52	old	5	years	11	week	36	weekend	44
year	29	year	4	afternoon	11	evening	21	nights	26
months	28	months	3	sunday	11	birthday	16	morning	14

Fig. 6: Most frequent true-positives by keyword category and by online review site

users said that something had "far" to go rather than use the word to describe location distance. Frequent false-positives for temporal keywords were: fall, old, party, and independence. The word "old" was used to describe the age of things other than people. The word "party" was used to describe a dinner party. The word "independence" was used to refer to Independence Hall.

Figure 6 summarizes the most frequent true-positives by keyword category and by online review site. The most frequent true-positives for relationship keywords were: son, daughter, kids, husband, wife, boyfriend, family, and friend. Frequent true-positives for location keywords were: house, home, school, college, street, visit, trip, and stayed. Amazon users often revealed that they used a product at home. TripAdvisor users revealed more detailed information about their "stays." Frequent true-positives for temporal keywords were: old, year or years, month or months, afternoon, evening, weekend, and Sunday. Common special occasions revealed were birthdays and anniversaries. The word "old" was used to describe the age of gift recipients or fellow movie watchers.

Relationship	Location	Temporal
Grandparents	Boardwalk	Several
Sisters	Living	Celebrate
Brothers	District	Month abbreviations (Sep, Oct, etc.)
In-laws	Cab	Festival

TABLE II: Missed keywords found by manual inspection of reviews

D. False-negatives

A false-negative occurs for each missed keyword not included in the keyword lists that is found through manual inspection of the reviews. Private information associated with the keyword must be disclosed in order for the keyword to be considered a false-negative. Table II lists the false-negatives that were found. Note that this list should not be considered comprehensive or complete since finding false-negatives is difficult and prone to error.

VII. Discussion

By combining the results of the site evaluation with the results from the privacy-checks of unstructured review content, an overall impression of the state of privacy can be formed for each site. Netflix revealed the least amount of structured information and had average keyword match percentages but high false-positive rates. This means that minimal structured and unstructured information was revealed about Netflix users by their reviews. Of the sites evaluated and checked, Netflix poses the least privacy threat. Netflix once had community features that included the association of reviews with user identities and the ability to network but phased out its community features starting in March 2010 and completing in September 2010. It is clear from this research that the phasing out of community features increased the privacy of Netflix users. This supports the notion that there is a trade-off between privacy and social networking.

OpenTable is not far behind Netflix when it comes to privacy. OpenTable does not reveal the identity of users posting reviews and reveals very little structured information overall. OpenTable users reveal an average amount of unstructured information about themselves in their reviews but have a tendency to reveal more temporal information. Amazon is similar to OpenTable when it comes to privacy but reveals slightly more structured information. Amazon users also reveal less location information and significantly more relationship information in their reviews. This is likely due to reviewing products from home and also discussing the people intended to receive products as gifts.

Yelp and TripAdvisor pose the greatest privacy threats of the sites analyzed. Both sites reveal a significant amount of structured information about their users. Yelp users reveal about as much unstructured information as OpenTable or Amazon reviewers. TripAdvisor users reveal the most unstructured information overall and TripAdvisor keyword matches had the lowest false-positive rates overall. This means that although both Yelp and TripAdvisor reveal a high amount of structured information, TripAdvisor users are more likely to reveal private information in their reviews.

A. Suggestions for Improving Privacy

Online review sites can learn something about privacy from the steps Netflix took when it phased out its community features. The biggest step towards increasing privacy is to not reveal the user identities associated with reviews. Review sites can still do what OpenTable does and reveal the year that review authors joined as a way to indicate the reviewer's level of maturity or expertise. Revealing the month a review was written can also help readers know if the review is still relevant without revealing the exact day of the review. In general, there does not appear to be much usefulness in knowing the location, age, or gender of reviewers, so review sites should not reveal these structured data.

B. Protecting Users from Inadvertent Disclosures

A privacy-check tool could be deployed on the servers of the online review sites, as a third-party "guardian angel" service, or as a client-side browser extension. Since users cannot rely on the review sites to do this and a third-party service is likely to cost money, a client-side browser extension is the best solution. A Firefox extension could easily be implemented to use keyword matching and NLTK to protect users from the inadvertent disclosure of private information. Such a tool could include other useful features. For example, the tool could make local copies of all reviews posted online to provide the user with a search-able database of all of their past reviews. The tool could provide a spell-check- like interface with ignore, ignore all, always ignore, and accept options for keyword matches. The tool could include descriptions by category or by keyword explaining the potential privacy implications of using a word. Finally, the tool could provide useful statistics, i.e., the total number of marked words accepted or ignored or the total number of "always ignore" words.

VIII. Future work

Future work could expand on the efforts of this research by studying the level of user control over their online reviews, including a larger number of online review sites, or studying the implications of cross- site trace-ability based on user pseudonyms or other personally identifying information. The level of user control over their online reviews is defined by their ability to delete their profile, edit or retract reviews, access all review data, and control third-party sharing. Another way of expanding this work would be to include additional keyword categories such as personal attributes or behaviors. Attributes might include diseases, beliefs, and ethnicity. Behaviors might include addiction, abuse, and dating. Furthermore, keyword list generation could be made user- driven by surveying the users of online review sites for keywords. Future work could focus on analyzing multiple reviews of single users in order to learn and customize privacy protection to the user. Keyword matching could be enhanced with classifiers or with the filters based on context. For instance, do not try to match the word "flight" if the review is about a toy helicopter sold by Amazon.

IX. Conclusion

The goal of this paper was to bring attention to privacy threats in online review sites. Users of online review sites are shown to be prone to inadvertent disclosures of private information such as their relationship, location, and temporal attributes. There is a need for a client-side privacy-check tool to protect users from these inadvertent disclosures.

Acknowledgements: We would like to thank the Intel Science and Technology Center (ISTC) for Secure Computing for supporting this work.

References

[1] Kushal Dave, Steve Lawrence, and David M. Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *WWW*, pages 519—528, 2003.

[2] Michael Hart, Pratyusa Manadhata, and Rob Johnson. Text classification for data loss prevention. In *Privacy Enhancing Technologies: 11th International Symposium, PETS 2011*, July 2011.

[3] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168—177, 2004.

[4] Huina Mao, Xin Shuai, and Apu Kapadia. Loose tweets: An analysis of privacy leaks on twitter. In *Workshop on Privacy in the Electronic Society (WPES)*, 2011.

[5] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January 2008.