Loose Tweets: An Analysis of Privacy Leaks on Twitter

Huina Mao, Xin Shuai, Apu Kapadia
School of Informatics and Computing
Indiana University Bloomington
Bloomington, IN, 47401, USA
{huinmao, xshuai, kapadia}@indiana.edu

ABSTRACT

Twitter has become one of the most popular microblogging sites for people to broadcast (or "tweet") their thoughts to the world in 140 characters or less. Since these messages are available for public consumption, one may expect these tweets not to contain private or incriminating information. Nevertheless we observe a large number of users who unwittingly post sensitive information about *themselves and other people* for whom there may be negative consequences. While some awareness exists of such privacy issues on social networks such as Twitter and Facebook, there has been no quantitative, scientific study addressing this problem.

In this paper we make three major contributions. First, we characterize the nature of privacy leaks on Twitter to gain an understanding of what types of private information people are revealing on it. We specifically analyze three types of leaks: divulging vacation plans, tweeting under the influence of alcohol, and revealing medical conditions. Second, using this characterization we build automatic classifiers to detect incriminating tweets for these three topics in real time in order to demonstrate the real threat posed to users by, e.g., burglars and law enforcement. Third, we characterize who leaks information and how. We study both self-incriminating primary leaks and secondary leaks that reveal sensitive information about others, as well as the prevalence of leaks in status updates and conversation tweets. We also conduct a cross-cultural study to investigate the prevalence of leaks in tweets originating from the United States, United Kingdom and Singapore. Finally, we discuss how our classification system can be used as a defense mechanism to alert users of potential privacy leaks.

Categories and Subject Descriptors

H.5.m [**Information Interfaces and Presentation (e.g., HCI**]: Miscellaneous

WPES'11, October 17, 2011, Chicago, Illinois, USA.

General Terms

Human Factors, Measurement, Security

Keywords

privacy, leaks, social networks, Twitter

1. INTRODUCTION

As a microblogging service, Twitter has become one of the most popular social networking tools today. As of June 2010 about 65 million tweets are posted each day, equaling about 750 tweets sent each second, according to Twitter [18]. With a wealth of information being broadcast publicly by individuals, one may wonder how much sensitive information is contained in these messages. Some may argue that by 'definition' information posted publicly on Twitter cannot be private and that Twitter users ought to realize that. A recent *qualitative* study of 'regrets' on Facebook shows that users "do not think about ... the consequences of their posts" and they regret posts made when "they are in a 'hot' state of high emotion when posting, or under the influence of drugs or alcohol" [19]. We perform the first *quantitative* study along these lines on Twitter and show there is a plethora of sensitive information revealed by Twitter users, not only about themselves but about other users. While users may themselves not think their posts are sensitive, *we focus on categories of leaks where there is a clear potential for negative consequences to the user.* In general it is hard to anticipate what other forms of leaks may occur from 'public' tweets, but recent news provides yet another note of caution. The New York Times reported on how various companies are now 'scoring' users based on Twitter (and other) feeds along various dimensions.[1] Some applications of such scoring may be helpful for the user, when, for example, targeted advertising may be welcome, but others may be especially harmful, when, for example, insurance companies score people based on their reported behaviors and increase premiums, or worse, deny them insurance.

The purpose of our work is threefold. Our first goal is to *characterize the nature* of these leaks and we particularly focus on topics such as tweeting about vacation and travel plans (vacation tweets), tweeting under the influence of alcohol (drunk tweets), and tweeting about medical conditions (disease tweets). We find that many people reveal their travel plans making them vulnerable to theft. We identify several sensitive topics that people are more likely to reveal

[1]Got Twitter? You've Been Scored: `http://www.nytimes.com/2011/06/26/sunday-review/26rosenbloom.html`

through drunk tweets as opposed to 'sober tweets,' as well as revelations about drunk driving.

Our second goal is to *demonstrate the real threat of automated attacks* based on the previous characterization, where burglars may automatically receive alerts about vacation messages, law enforcement may receive alerts about drunk driving, and insurance agencies may receive alerts about people with medical conditions. We build binary classifiers to detect sensitive vacation tweets (with 76% precision) and drunk driving tweets (with 84% precision), and also show that a simple classification rule can detect sensitive tweets about diseases such as cancer (with 76% precision).

Finally, our third goal is to *characterize who* leaks private information and *how*. We characterize both *primary leaks* (by the user itself) and *secondary leaks* (by some other user) of sensitive information on Twitter, and whether these leaks are through *status messages* or *conversation messages*. For example, users may inadvertently reveal information through Twitter-based, world-readable conversations with other Twitter users. We show that a large number of leaks are seen in all these categories. We also perform a cross-cultural exploration of users in different countries to see how people in different countries are prone to revealing more or less sensitive information in different categories. In particular, we compare the extent of privacy leaks in three countries: United States (US), United Kingdom (UK) and Singapore (SG), and explore how cultural differences might affect the extent of privacy leaks.

2. RELATED WORK

To the best of our knowledge, there has been only a limited amount of research into privacy leaks on Twitter. Humphreys et al. [10] found that personally identifiable information (such as email addresses, home addresses, and phone numbers) are rarely found in tweets, but a quarter of tweets do include information regarding when people are engaging in activities and where they are. Meeder et al. [13] demonstrated that the 'retweet' mechanism led to privacy leaks when followers retweeted sensitive protected tweets publicly.[2] For example, some retweets publicized family and contact information and sometimes even messages about bosses. Gomez-Hidalgoy et al. [9] proposed a mechanism to detect tweets about named entities (e.g., a company, brand, or person) using Named Entity Recognition (NER). They aim to prevent data leaks about other entities and don't necessarily focus on users leaking information about themselves. While these studies touch on some aspects of privacy leaks on Twitter, they do not present a systematic approach to *detect* various categories of privacy leaks on Twitter in real time. In addition to a larger characterization of privacy leaks on Twitter, our work aims to fill this gap by showing how privacy leaks can be detected automatically.

Other researchers have examined privacy issues on social networking sites in general. Bhagat et al. [4] studied the privacy loss from social graph prediction; Dwyer and Hiltz [7] discussed the trust and privacy concerns within Facebook and MySpace. Acquisti et al. [1] comprehensively investigate privacy issues surrounding Facebook. Nevertheless, such work does not look at what information is revealed through the messages themselves. Some privacy concerns have been raised about Twitter specifically. For example, recently Twitter added a mechanism to let other people automatically know where you are whenever you tweet and this has caused concern [16]. This constitutes another leak of important information through tweets, and would allow global-scale location tracking of Twitter users.

[2] "Protected" tweets are obtained from protected profiles, which are readable only by the tweeter's friends.

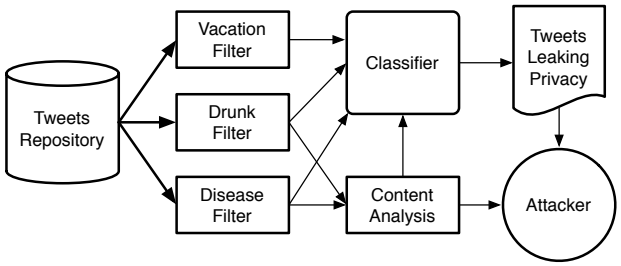


Figure 1: Architecture of our privacy detection approach. The three filter modules identify topical tweets based on keyword matching. These tweets are then fed through classifiers to detect sensitive tweets. The content analysis module characterizes the nature of privacy leaks for each of the three categories (vacation, drunk, and disease tweets).

3. OVERVIEW OF APPROACH

Figure 1 demonstrates the architecture of our privacy detection approach. Our Twitter data is acquired in real-time via the Twitter Streaming API (i.e., the Twitter "gardenhose") and stored in a *Tweets Repository*. This repository represents about 15% of the total tweets from public profiles sampled randomly. On average, over 1 million tweets are obtained everyday, and we use the tweets from January to September 2010 in our analysis. For our analysis, we choose tweets containing pronouns such as *I, me, my, we, us, she, her, he, him, his* and we filter out retweets.[3] After this simple filtering and sanitization step we obtained 162,597,208 tweets in our corpus stored in the *Tweets Repository*. Next, we apply the *Vacation Filter*, *Drunk Filter* and *Disease Filter*, three modules used to filter out all topically relevant tweets via keyword matching. Examples of topical keywords include "holiday", "fly to", and "travel" for vacation tweets, "I am drunk" for drunk tweets, and "cancer", "depression" and "diabetes" for disease tweets. In total we obtained 575,689, 21,297, and 149,636 tweets for vacation, drunk, and disease topics respectively. For drunk and disease tweets we also filter out tweets containing URLs because we found a large number of such tweets were spam or ads (Gao et al. analyze URL-based spam in social networks [8]). We keep vacation tweets with URLs because we observed URLs in vacation tweets were much less likely to be spam or ads than in drunk and disease tweets.

After picking out tweets through the filters, further processing is often needed, depending on what type of analysis will be done with the data. We will discuss these details below. After all related tweets are filtered out, the *Classifier* module is used to automatically detect sensitive tweets. The *Content Analysis* module provides information about what private topics are revealed from drunk and disease re-

[3] *Retweets* are tweets reposted by users.

lated tweets, which also can be utilized by *Classifier* to select classification labels. We did not apply *Content Analysis* towards vacation tweets because we only cared about whether people would go on vacation, and not what the details about their vacation. The system thus outputs a stream of sensitive tweets leaking private information through the *Classifier* as well as relevant topical information through the *Content Analysis* module, which can be analyzed by the *Attacker* to decide what other topics may be of interest in terms of implementing further attacks. We now describe these two modules in a little more detail.

Classification is the task of designing a classifier that can assign the correct *class label* for a given input. In basic classification tasks each input is considered in isolation from all other inputs, and the set of labels is defined in advance. If only two labels exist, then the classification is "binary". Our privacy information classifier is binary because each tweet can be classified as either *sensitive* or *non-sensitive*. A classifier is called *supervised* if it is built based on training corpora containing the correct label for each input. Feature extractors are used to extract features or characteristics from the input that can be used to classify the input. Common machine learning algorithms include Naive Bayes and SVM, which are used to learn the classification rules applied later in the *testing* phase. In most of the cases, as with machine learning algorithms, *training* is necessary because the classification rule is not clear. However, in some special cases we can directly deduct the classification rule without training. We will show such a special case in relation to our classification of *cancer* tweets as described in Section 6. For the classification of vacation and drunk tweets, we used both Naive Bayes and SVM but obtained better results with Naive Bayes, and present those results.

The most important step in creating a classifier is deciding what features of the input are relevant, and how to encode those features. While the basic process of designing the classifier is the same for the three topics we study, the feature selection is different because each domain has its own textual features that are more suitable for classification. These features include the words as well as the grammar. To capture the words features, we use the *bag-of-words* model [6]. Given a set of documents, all the words from the documents constitute the lexicon. Each document can be represented by a vector where each dimension represents a word in the lexicon. If that document contains some word, the corresponding dimension will be 1 otherwise it will be 0. This is also the baseline feature that is used the most in different types of document classification systems. In addition, we manually pick out some keywords as well as key phrases that are most relevant to classification in each domain. To capture the grammar features, we make use of natural language processing techniques such as name entity recognition and part-of-speech tagging.

For binary classification of sensitive tweets, the common metrics for evaluating classifier performance include:

- *Accuracy* – the fraction of correctly classified sensitive tweets among all samples in the testing set;
- *Precision* – the fraction of correctly classified sensitive tweets among all samples that are classified as sensitive in the testing set. A higher precision means fewer samples are misclassified as sensitive samples;
- *Recall* – the fraction of correctly classified sensitive tweets among all actual sensitive samples in the testing set. The higher the recall the more percentage of true sensitive samples are included in the final classified sensitive samples;
- *F-measure* – the harmonic mean of precision and recall, which gives a balanced measure of the two measures.

We use these measures to evaluate our classifiers, but emphasize the *precision* metric. In our attack scenarios we want to reduce misclassifications as a primary goal. Of course, while the recall must be reasonable, as long as a 'sufficient' number of sensitive tweets are identified the precision metric is more useful. For example, a burglar would be interested in *several*, but not *all* potential homes to burglarize.

An important consideration is whether the labeling of tweets as sensitive or not in the training and testing sets is correct. One approach would be to obtain 'ground truth' from the users themselves (who issue the tweets) and get their opinions on whether they consider the tweets sensitive in retrospect. While such ground truth would be desirable, contacting such a large number of users is very difficult practically. Instead, we focused on the manual labeling of data where we assess whether there is a potential for harm from those tweets. While such an endeavor may be difficult in general, it is easy for the cases we analyze. For example, answering the following questions does not require ground truth from the tweeters: Does the tweet reveal concrete vacation plans? If so, then there is real potential for burglary. Does the tweet mention drunk driving? If so, then there is the real potential for law enforcement action. Does the tweet mention disease related information? If so, then there is the real potential for exploitation by insurance companies.

Content analysis [11] is a research tool used to determine the presence of some concepts within text documents, mainly through manual annotation. Researchers quantify and analyze the presence, meanings, and relationships of textual words and concepts, then make inferences about the messages within the texts, the author(s), the audience, and even the culture and time. Content analysis is widely used in the training phase of classifier design in generating labeled documents. Content analysis can also extract richer semantic information that is difficult to detect automatically.

4. VACATION TWEETS

A few existing websites analyze location information in tweets. Pleaserobme.com uses Twitter's search functionality to show location-based messages and Foursquare's GPS-enabled mobile devices to target the location information. Icanstalku.com leverages the photo information shared in Twitter to infer the location information of users. Though our analysis of vacation is also related to location, the focus is a little different. Instead of trying to find where people are, we care more about when they will be away from home during vacation based on the textual content of tweets (thus our technique applies also to tweets without location information). As long as someone tweets about going on vacation it places the person's dwelling at risk of being burglarized, regardless of whether location information is revealed. Twitter users have indeed been burglarized in this way [14]. We call vacation tweets "sensitive" when the reveal concrete travel plans. Otherwise, if the tweets only mention "vacation" but do not actually reveal any travel plans, then we call them "non-sensitive". For example, we observed tweets

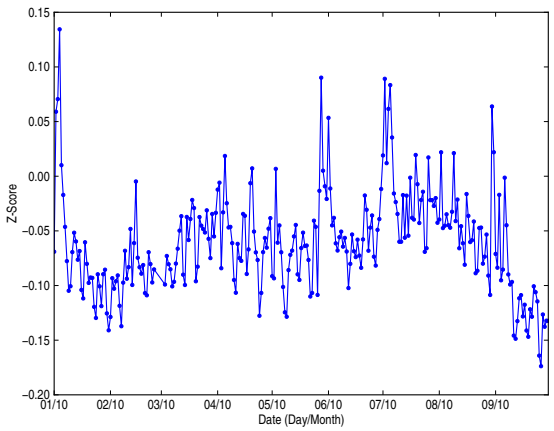


Figure 2: z-score of ratio of vacation tweets in the period of January to September 2010. Four vacation spikes are seen at the beginning of January, the end of May, beginning of July, and end of August.

with phrases such as *"i need a vacation already"*, which are clearly non-sensitive vacation tweets.

4.1 Data description

The keywords we used for searching vacation tweets are: *vacation, holiday, travel, trip, leave for, fly to.* The total number of vacation tweets in our dataset was 575,689, and their daily temporal pattern is shown in Figure 2. We found four vacation spikes during the beginning of January, end of May, beginning of July, and end of August. The *y-axis* shows the *z-score*[4] of the ratio of number of vacation tweets to the total number of tweets on that day.

We randomly sampled 1,000 tweets and annotated them by giving each tweet a class label as either "sensitive" (i.e., specified concrete vacation plans) or "non-sensitive" (i.e., did not specify concrete vacation plans). The percentage of sensitive tweets in the 1,000 vacation tweets is 10.8%. Notably, for these 108 vacation sensitive tweets, we found that the occurrence of location, time and person is 90.7%, 55.5%, and 44.4%, respectively. *Location* occurs much more frequently than *person* and *time* in sensitive tweets, implying that *location* is a more important feature than the other two for following classifier design.

4.2 Classifier implementation

4.2.1 Feature selection.

Most of the time, choosing the representative words instead of all words as features can improve the classification. Moreover, in vacation tweets, we found more important features than just words. After extracting vacation topic tweets from our repository, we manually checked a small sample of tweets and found three features that are most relevant to our sensitive vacation tweets detection: location name, person, and time (LPT). The common tool that we used to automatically detect LPT information from text is NER. For the person's name, besides common names the Twitter screen name starting with @ is also factored in.

After we pick out those sensitive vacation tweets via the NER-based Classifier system, we can have a list of users who post these tweets. Then the next step is to figure out where those users live. While we do not perform such analysis in this paper, we point out the attacker could check the user's Twitter account profile or use existing geo-inferencing algorithms to infer the user's residence address [5].

There are two well-known NER tools: one is Afner NER [2] and the other is Alchemy NER [3]. We found that Afner NER is bad at identifying person features, while Alchemy NER cannot identify time features well. Therefore, we used Afner to identify time, Alchemy to identify person (together with Twitter account screen names starting with @), and both of them to identify location.

In addition to LPT features, we also picked out some representative words that we found may or may not help to tell whether a vacation tweet is sensitive. Some places representative of vacation (e.g., beach, hotel), as well as air transportations (e.g., airport, flight), frequently occur in sensitive vacation tweets, even though they are not specialized location names. Additionally, some words are good indicators of sensitive tweets (e.g., leave, pack, booked, plan) implying the preparation for vacation travel. Some words are good indicator of non-sensitive tweets, including negative words (e.g., not, no, didn't), virtual words (e.g., should, wish, need, if) as well as past-tense verbs (e.g., went, got), which implies that the travel is already past or is not real, even though the name of the travel destination occurs. Finally, some tweets' special structural features, i.e., URL and hashtag, are also considered additional features, through detecting the occurrence of *http* and #. All the representative words are listed in Table 4.2.1.

4.2.2 Evaluation.

We manually annotated a total of 600 tweets consisting of 300 sensitive tweets and 300 non-sensitive tweets. Then we used two thirds of them as the training set and the rest as the testing set. We used both Naive Bayes and SVM Classifiers from the Natural Language Toolkit package [15]. We use the default parameters offered by the toolkit and don't explore parameter tuning in this paper. As mentioned above, the features we selected are: person, time, location, all words, and representative words.

To test how feature selection affects classifier performance, we selected different combinations of the above 5 features and compared their accuracy, precision, recall, and F-measure. The results, obtained by using Naive Bayes Classifiers, are listed in Table 2 and show that among the 5 single features, *location* gave the best precision, while the *representative words* gave the best recall value. It is not surprising that the combination of the two features produces the overall best performance in all evaluation indices. However, the *Person* feature did not perform well. When we combined the 5 features, the classification performance did not improve.

We also tried an SVM Classifier, but the results obtained were worse in all evaluation indices than with Naive Bayes Classifiers. The best performance we got from SVM is almost the same as the result we got from Naive Bayes classifier when using only the *location* feature.

We highlight the precision of 76% obtained by our classifier, which tells us that 76% of alerts supplied to burglars

[4]The *z-score* indicates how many standard deviations an observation is above or below the mean.

Table 1: List of representative words for classifying vacation tweets.

Category	Words
place and facility	*beach, coast, hotel, conference, island, airport, flight*
positive	*go, going, gonna, leave, leaving, pack, booked, before, will, until, wait, plan, ready, here I come, looking forward*
negative	*need, wish, not, no, want,wanna, back, went, may, might, maybe, had, recent, was, were, could, should, hope, got, suppose, if, didn't*
url	http
hashtag	#

Table 2: Naive Bayes classification results from different combinations of features for vacation tweets.

Features	Accuracy	Precision	Recall	F-measure
Person	0.415	0.383	0.28	0.324
Time	0.605	0.63	0.51	0.563
Location	0.715	0.717	0.71	0.713
All words	0.7	0.682	0.75	0.714
Representative words	0.61	0.567	0.93	0.704
Location +Representative	**0.785**	**0.761**	**0.83**	**0.794**

would be of true vacation events. While there is room to improve in terms of precision, we believe that this number is high enough to be filtered by manual inspection (e.g. by burglars) and thus demonstrates the real threat of automated classification of vacation tweets.

After sensitive tweets are detected by our classifier, one could further extract *who* will go on vacation, *where*, and during what *time* through NER and the user's profile. Again, we seek to analyze sensitive tweets with the potential for harm to the user, and assume burglars can use existing methods to obtain the address of the residence. In simple cases the name of the person may be available for lookup in the telephone directory. Absent such information, recent work shows how one can infer Twitter users' residence addresses through geotagged tweets and/or photos [5].

5. DRUNK TWEETS

We mainly focus on two interesting problems around drunk tweeting: one is *What private information is revealed during drunk tweeting?*; the other is *Are people more likely to reveal private information when they are drunk?* To answer these two questions, we analyzed the drunk-related topics revealed from 100 Twitter users and compared the percentage of private topics revealed between their drunk and sober tweets. Then we turn our attention to drunk driving for automatic detection. In particular we designed a binary classifier to filter out those *drunk driving* tweets.

5.1 Data description

First, we plotted the raw number of tweets containing *I'm/am/Im drunk* posted per hour from April to June 2010 as shown in Figure 3, where the *x-axis* is the z-score of the ratio of drunk tweets over all the tweets posted at that hour. From the autocorrelation, we can see the periodicity pattern that emerges: i.e. observations 24 hours apart are strongly positively correlated, while observations separated by 12 hours are negatively correlated. Further, we found that drunk tweet traffic rises in the evening and early morning and greatly diminishes in the day time. This pattern provides some useful information for the attacker about which time to monitor drunk tweets on each day.

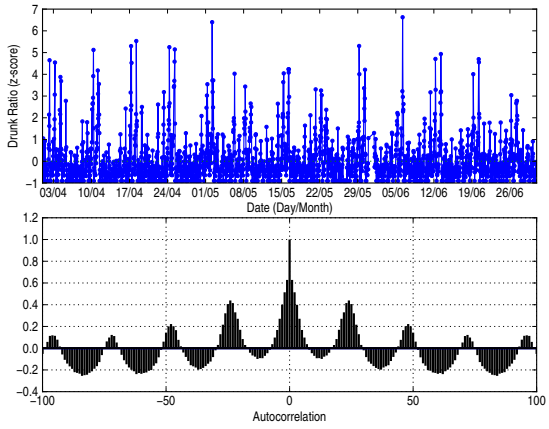


Figure 3: Drunk tweets/hour between April and June 2010. The autocorrelation graph shows there is indeed a daily (24-hour) pattern of drunk tweets.

5.2 Topic categorization and comparison

We selected the top 100 Twitter users who have the most drunk tweets from January to September, 2010. For each user's timeline we can detect many discrete time points when his/her tweets contain phrases such as *"I'm/am drunk"*. Then we aggregate all tweets within 3 hours after those discrete time points as drunk tweets. To identify sober tweets we pick those tweets between 6 and 3 hours preceding the time when the drunk tweet is posted (e.g., if a drunk tweet was recorded at 6pm, we consider tweets between noon and 3pm to be sober tweets). Thus we collect drunk and sober tweets coming from the 100 top drunk users. As a result we obtained 645 drunk tweets and 208 sober tweets. This

may confirm that people get more talkative when they are drunk. For drunk tweets, we performed content analysis to find out all private topics revealed. Then we manually annotated those drunk tweets and classified them into sensitive tweets or non-sensitive tweets. We found that a small fraction of drunk tweets were made jokingly, and so we filtered out these tweets. If a tweet is classified as sensitive, a further topic label will be attached to it. During our annotation, we found the following 6 topics that related to privacy issues (we give some excerpted and slightly reworded examples to protect privacy):

- Sexuality – revelation of sexual orientation or sexual activities and desires
- Expressed Emotions – expression of love/hate for somebody, or emotional outbursts about self:

  @[anonymized] I LOVE YOU!!!
- Confessions – revelation of personal affairs about self or others:

  me and my girl broke up. i'm single again :(.
- Disrespectful Behaviors – rants and embarrassing behaviors:

  ...just taking my pants off in front of all you
- Bodily Harm – some accident or adverse reaction:

  i just fell down the stairs, hitting my head really bad.
- Illegal Activities – drunk driving or other illegal activities:

  I drove drunk around the corner!

We plotted the private topics distribution in drunk tweets in Figure 4. We found that *Sexuality* and *Disrespectful Behaviors* are the most mentioned topics in drunk tweeting, constituting about 25% each of the drunk tweets. *Expressed Emotions* and *Confessions* occurred frequently too, at 22% and 16.7% respectively. *Bodily Harm* and *Illegal Activities* together constituted about 11% of drunk tweets.

Next we analyzed what topics occur more often in drunk tweets as opposed to sober tweets. Based on our coding scheme for drunk tweets, we also annotate those sober tweets and labeled them into the above 6 topics.

In Figure 5 we see that except for *Expressed Emotions* and *Confessions*, all other topics show increased rates in drunk tweets. Thus we observe that users are more susceptible to privacy leaks when they are 'tweeting under the influence.'

5.3 Drunk driving classification

Among all 6 topics, we consider *Illegal Activities* to be the most serious privacy leak. Specifically, most of the illegal activities are about drunk driving. To demonstrate the threat of automated detection of such tweets, we designed a classifier to automatically detect those drunk driving tweets.

5.3.1 Feature selection.

To make sure that *drunk* and *drive* occurred together, we used those tweets containing *drunk* co-occurring with the words *drove* or *drive*. Similar to vacation tweets, we first used *all words* as our basic features. Other than that, we considered some other textual features which we found during manual annotation. First, we analyzed the relative distance between keywords *drunk*, and *drive*, *drove*. The value of the distance is calculated by the position index of *drive*/*drove*, minus that of *drunk* in the whole tweet.

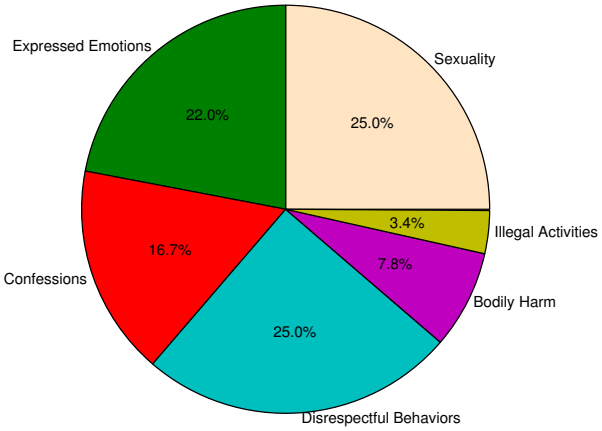


Figure 4: Distribution of sensitive topics found in drunk tweets

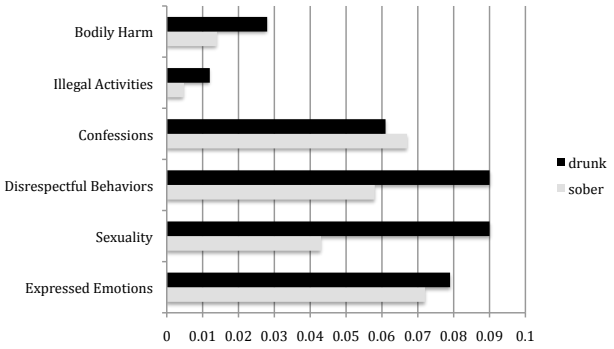


Figure 5: Comparison between drunk and sober tweets by percentage of sensitive topics.

We found that the smaller the distance, the higher probability that the tweet is a sensitive sample. Second, we looked at some representative negative words or phrases such as *don't, not, didn't, wasn't, too drunk to drive*, etc., frequently occur in non-sensitive samples. Once such words or phrases were found in a tweet, we marked the negative feature as *true* for that tweet. Third, we considered a pattern that shows the person who is driving and the person who is drunk is the same person. Specifically, the pattern *I/I'm/me...drive/drove...I/I'm/me...drunk* or *I/I'm/me...drunk...I/I'm/me...drive/drove* are found to frequently occur in sensitive drunk driving tweets, indicating that the poster was driving while he/she was drunk. If such a pattern is found in a tweet, then we mark the pattern feature as *true* for that tweet. Fourth, we employed part-of-speech tagging. We found that in addition to words themselves, the category of words also plays an important role. The word just after the key word *drunk*, *drive*, or *drive* was

Table 4: Naive Bayes classification results from different combinations of features for drunk tweets.

Features	Accuracy	Precision	Recall	F-measure
All words feature	0.75	0.797	0.67	0.728
Textual features	0.695	0.741	0.6	0.663
Both	**0.79**	**0.845**	**0.71**	**0.771**

worth attention. For example, if the word after *drive* or *drove* is a personal pronoun, like *me, him, her* etc., then in our sample, such a tweet probably means that some person drove the other drunk person home, implying that it is not a drunk driving incident. The total features we used in our classification are listed in Table 5.3.1.

5.3.2 *Evaluation.*

Then we implemented Naive Bayes and SVM Classifiers using words features and/or textual features and only the Naive Bayes results are listed in Table 4. We had the overall best performance when we used both words feature and textual features. Clearly, adding textual features to words feature can obviously improve the classifier's performance. As the same for vacation tweets, the performance of SVM is not as good as Naive Bayes, except for recall value, reaching as high as 0.9. In particular, the precision of 85% tells us that 85% of alerts supplied to law enforcement would be of true drunk driving incidents. We believe this precision is high enough to demonstrate the real threat of automated classification of drunk driving tweets.

6. DISEASE TWEETS

To find out at what level Twitter users reveal their health-related issues, especially disease problems, we conducted a broad and deep study to investigate how many people in general talk about diseases and how many types of disease are revealed. We further analyzed the tweets about several sensitive diseases to find out who was actually implicated in the tweets, i.e. the tweeter or others (e.g., friends and family). To ensure accuracy, we manually annotated those tweets and divided them into subcategories.

6.1 Data description

We extract all the disease names containing the words *"disease"*,*"syndrome"*, and *"disorder"* from *Mendelian Inheritance in Man* (OMIM). OMIM is a continually updated catalog of human genes and genetic disorders.[5] Moreover, we added the other eight diseases that are usually heard, including *tumors*, *depression*, *cancer*, *obesity*, *HIV*, *HPV*, *AIDS*, and *diabetes*. In total this led to 390 types of diseases we search for in the tweets. When reviewing the tweets, we ensured the presence of the words "has", "had", and "have" before the mentioned disease names. After performing the search, we found 24,346 tweets with 45 types of disease provided by 21,508 Twitter users during the period of January to September, 2010. We do not show a timeline of these tweets because we did not find any discernible patterns. While one may expect to see patterns related to some illnesses such as the flu,[6] we didn't see any overarching patterns when looking for disease tweets in general.

[5] `http://www.ncbi.nlm.nih.gov/Omim/omimfaq.html#db_descr`

[6] `http://www.google.org/flutrends/`

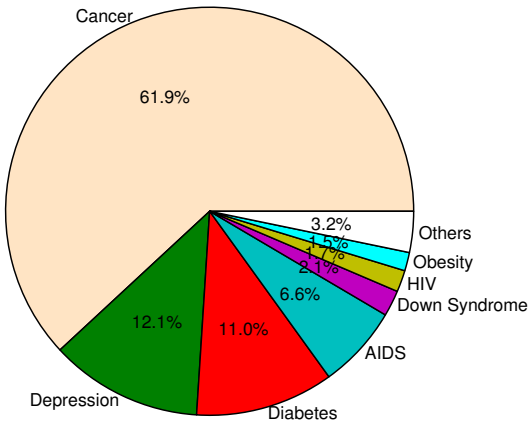


Figure 6: Distribution of names of different diseases mentioned in disease tweets.

6.2 Disease categorization

We plot the distribution of these tweets in Figure 6. We can see that *cancer* is the most mentioned disease, followed by *depression*, *diabetes*, and *AIDS*.

6.3 Cancer classification

Similar to our exploration for detecting vacation and drunk driving tweets, we tried to automatically detect sensitive disease tweets, which leak personal disease records. Specifically, among the four categories we defined, those tweets that reveal that others/oneself either have or have had some disease at one point are considered sensitive. During our manual annotation, we found that designing such a classifier for all diseases was challenging. For example, we also annotated *AIDS* tweets, but it turned out to be hard to judge whether someone really has *AIDS* solely based on words (unfortunately, in multiple senses of the word, many of these tweets appeared to be joking about *AIDS*). Such ambiguity was largely reduced in the *cancer* tweets based on our manual annotation analysis. Thus we only tried to classify sensitive *cancer* tweets here.

Instead of carefully designing and training a classifier like we did for vacation and drunk driving tweets, we found an easier way to achieve automatic classification based on regular expressions. In addition we filter out tweets containing pets (e.g., dog, cat, kitty, puppy) or negative expressions (e.g. *don't/doesn't have/has/had*). So for cancer, our simple classification rule is:

- If *has/have/had...cancer* exists **and** (*dog/cat/kitty/puppy* **and** *don't/doesn't have/has/had*) doesn't exist in a tweet, it is sensitive; otherwise, it is not.

We tested the above rule for the randomly sampled 200 annotated *Cancer* tweets, and found that the final accuracy,

Table 3: Textual features for classifying drunk driving tweets

Feature	Detail
key words distance	position index difference between *drunk* and *drove/drive*
negative words or phrases	don't, not, no, couldn't, can't, didn't, wasn't, won't, wouldn't, if, wish, too drunk to drive
regular expression pattern	*...I/I'm/me...drunk...I/I'm/me...drive/drove...* or *...I/I'm/me...drove/drive...I/I'm/me...drunk...*
words tagging	category of word after *drunk* category of word after *drove/drive*

precision, recall, and F-measure are 0.782, 0.759, 0.82 and 0.788, respectively. Compared with our best results obtained from vacation and drunk driving classifications, our classification result for cancer disease again demonstrates that alerts to an insurance company, for example, would have 76% precision, and demonstrates a real threat.

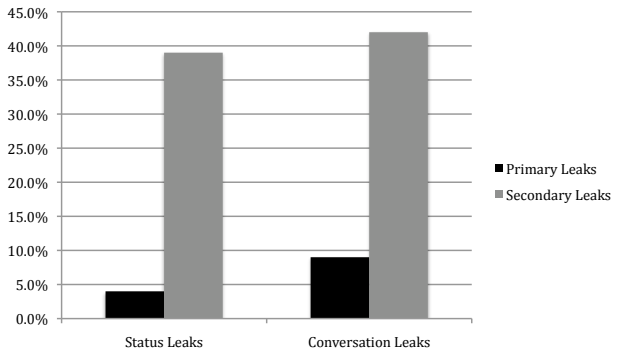
7. DIFFERENT TYPES OF LEAKS

In the previous sections we analyzed the sensitive content from three private topics: vacation, drinking, and disease, as well as designed binary classifiers to automatically detect sensitive tweets in those categories. In other words, we studied *what* leaks in sensitive tweets. In this section, we shift our focus to two other questions:

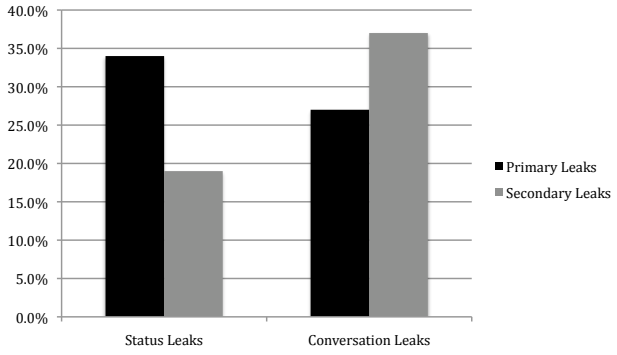
- *how* does privacy leak from sensitive tweets?
- *who* is revealed in sensitive tweets?

Specifically, we have two ways to categorize sensitive tweets. Based on the type of tweet, we have *status leaks* (privacy leaks from status-update tweets) and *conversation leaks* (privacy leaks from conversation-based tweets). Given a tweet, if it starts with @username, then it is a conversation tweet. Otherwise, it is a status tweet. Next, based on who is revealed, we have *primary leaks* (where the original tweet poster implicates himself/herself) and *secondary leaks* (where a tweet poster implicates some other person). We can distinguish such primary and secondary leaks though content analysis. An example in our dataset of a secondary leak through a status update is *"...My mom 'borrowed' my car and drove it around drunk"*, and an example of a secondary leak through a conversation tweet is *"@[anonymized] I will pray for your mom. I had 2 family members diagnosed with cancer in the past 2 days. My brother and aunt. :-("*. Note in the latter example, it may have been unlikely for the author to implicate his/her brother and aunt in a regular status update, but he/she provides this information as part of a conversation. Thus there are four types of leaks, which we analyze in the categories of: outgoing vacation, drunk driving, and three representative diseases, i.e., cancer, diabetes and HIV.

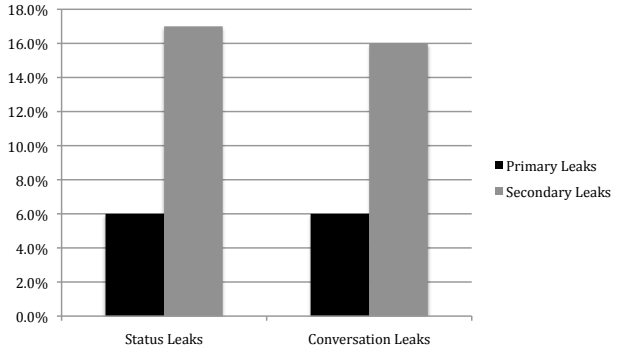
Vacation Tweets. From all vacation tweets in our corpus (including both sensitive and non-sensitive tweets), we randomly sampled 3,000 tweets and divided them into status and conversation tweets. We found that the ratio of status tweets to conversation tweets is close to 2:1. Then we randomly selected 20% for each category and obtained 398 status tweets and 201 conversation tweets. Contrary to the classifier design, we not only annotated those sensitive tweets, but further divided the sensitive tweets into *primary leaks* or *secondary leaks*. For instance, if a conversation tweet



(a) Cancer

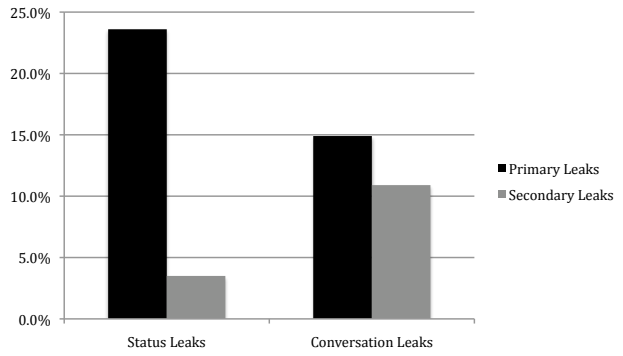


(b) Diabetes

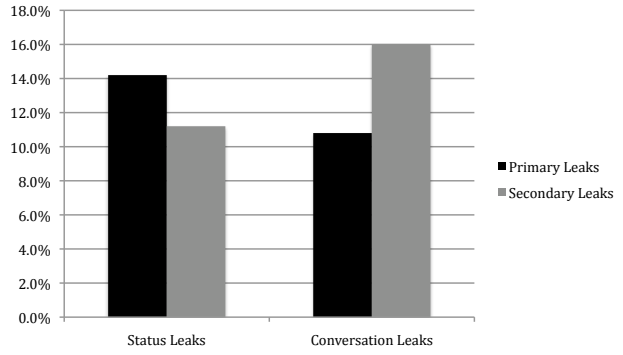


(c) HIV

Figure 7: Comparison between primary/secondary and status/conversation leaks in disease tweets.



(a) Vacation tweets.



(b) Drunk tweets.

Figure 8: Comparison between primary and secondary leaks, status and conversation leaks for vacation and drunk tweets

revealed that the poster itself would be on vacation and not home, then the tweet was annotated as a *primary leak* and a *conversation leak*.

Figure 8(a) shows that primary leaks occur more frequently than secondary leaks, and primary leaks occur more in status messages. However, secondary leaks occur more frequently in conversation tweets than in status tweets. Thus as part of a public conversation on Twitter, users are more likely to reveal sensitive vacation information about each other or other users than through status updates.

Drunk tweets. We found a total of 2832 drunk driving tweets in our corpus and also divided them into status and conversation tweets. Again we found the ratio of status tweets to conversation tweets to be close to 2:1. We then randomly selected 20% for each category and obtained 393 status tweets and 175 conversation tweets. The annotation is the same as with vacation tweets.

Figure 8(b) shows that the frequency of secondary and primary leaks in drunk tweets is roughly the same. Similar to vacation tweets, status tweets leak more primary privacy information while conversation tweets leak more secondary privacy information. Thus as part of a public conversation on Twitter, users are more likely to reveal sensitive drunk driving information about each other or other Twitter users than through status updates.

Disease tweets. For disease tweets, we found that the amount of tweets was quite different among the three types

Table 5: Percentage of vacation, drunk, disease tweets across these three countries, US, UK and SG (Singapore).

Privacy types	US	UK	SG
Vacation	0.34	0.40	0.34
Drunk	0.01	0.01	0.006
Disease	0.02	0.02	0.008

of diseases with cancer having the most and HIV the least. We also found that the number of status tweets is very close to that of conversation tweets for all three diseases. Therefore, we randomly sampled 1000 cancer tweets, 1000 diabetes tweets and took a total of 567 HIV tweets. After dividing them into status and conversation tweets, we further sampled 100 status tweets and 100 conversation tweets for all three diseases and then annotated them.

From Figure 7, we observe two interesting findings. First, tweets about *cancer* and *diabetes* are more likely to be sensitive as compared to *HIV*. Next, secondary leaks occur much more frequently than primary leaks in both *cancer* and *HIV*, which is alarming. For both these diseases we see a low number of primary leaks, suggesting that users consider this to be private information. Originally, we thought perhaps people primarily reveal this information about friends and family who have passed away, but these surprising results were achieved once we had removed such tweets. For *diabetes*, the frequency of primary and secondary leaks are almost the same, and similar to the drunk driving situation.

8. CROSS-CULTURAL ANALYSIS OF LOOSE TWEETS

In this section we explore the question of whether the prevalence of privacy leaks is different for different cultures. We chose three countries for our comparison: the United States, United Kingdom, and Singapore, which represents one country each from the North America, Europe, and Asia, respectively. We analyzed tweets originating in these countries based on location information in the tweets. We leave a more detailed analysis (with more countries) to future work.

From our tweets repository, we further parsed the location information and then chose only the US, UK and Singapore tweets, which resulted in 21,469,824 US tweets, 4,908,225 UK tweets, and 952,716 tweets in total. We counted the percentage of vacation, drunk, and disease tweets for each of these countries. The results are shown in Table 5. We see all three countries have the similar percentage of vacation tweets, with UK having a slightly higher percentage. For drunk and disease tweets, US and UK have similar ratios, while Singapore has about half as many (as a percentage) drunk and disease tweets. It could be that people in Singapore are more conservative about these topics.

Next we investigate the prevalence of sensitive tweets in *vacation*, *drunk*, and *disease* tweets. We randomly sampled 200 vacation tweets for each country. Since Singapore has few drunk and disease tweets, we kept all the tweets for those topics, i.e. 77 disease tweets and 59 drunk tweets. For US and UK, we selected 100 tweets each for these topics for manual annotation. The results are shown in Figure 10. We can see that tweets originating in the US for the various topics are more likely to contain sensitive information, and tweets from the UK are least likely to contain sensitive in-

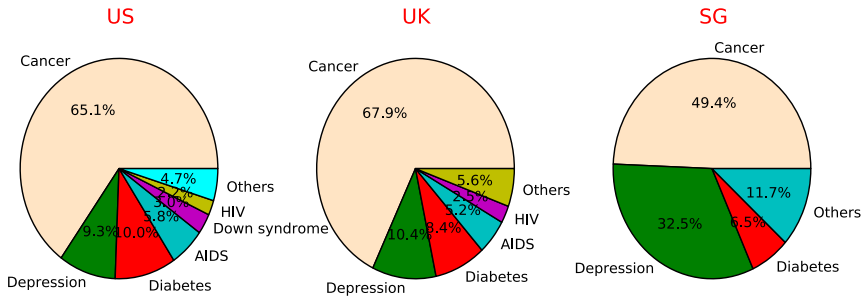


Figure 9: Distribution of diseases over US, UK and Singapore

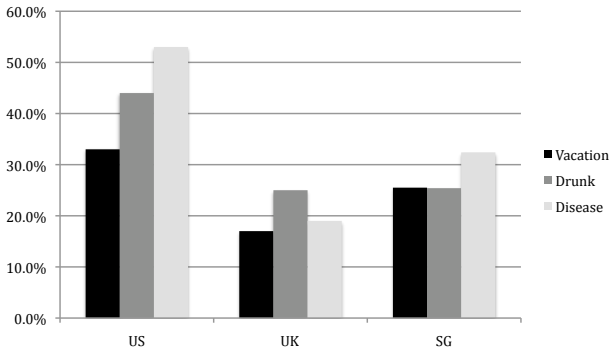


Figure 10: Fraction of sensitive tweets across countries.

formation. We can see that tweets from Singapore had fewer privacy leaks (as a percentage) than the US, but more than the UK. Thus our data suggests that people in the US are more prone to 'loose tweeting' about these topics, and people in the UK and Singapore exercise a little more restraint, but are loose with their tweets nevertheless.

We analyzed the distribution of disease topics covered by the tweets from these three countries (see Figure 9). We found the top 3 diseases mentioned in these countries are *cancer, depression,* and *diabetes*. Singapore had fewer tweets mentioning HIV, and had a relatively high fraction of depression tweets than the other two countries. US had 4973 disease tweets, mentioning 30 types of diseases; UK had 1025 tweets, in which 15 types of diseases were mentioned, Singapore had 77 disease tweets mentioning 8 types of diseases.

9. DEFENSIVE TECHNIQUES AND FUTURE WORK

The purpose of this paper is *not* about creating a defensive system; instead our focus is to raise awareness of the extent of privacy leaks on Twitter, where there is the clear potential for exploitation of unwitting users. Nevertheless, it is important to discuss defensive measures (both practical and research opportunities) for users who want to use services such as Twitter, but also want to rein in their privacy. We list some possible measures below beyond the obvious countermeasures of "thinking twice before you tweet" or using protected accounts (which would limit access to friends only):

Guardian angel services. The classifiers that we have built can be used by unwelcome parties, but we emphasize our work can also be used defensively. Services can offer Twitter users to monitoring of their tweets (and other tweets that mention them) and alert users of potential privacy violations. Such a service could send users warnings if their status messages reveal private information about themselves (or others). For example, these services could caution users against revealing too much information while under the influence of alcohol. We note that users are also vulnerable to damage resulting from information spread through the social network and not only limited to automated attacks. Thus, more research in automated methods for *detecting* privacy leaks is needed as such techniques will benefit a large range of users whose real threat may be their own social network.

Social network and relationship analysis. It has been widely observed that users' attributes and behaviors tend to correlate with their social connections [17]. If some private attribute is correlated with a social network, we expect actors sharing the same privacy property to be positively correlated with social relationship. The most well-known example is Gaydar [12], which leverages the social connections on Facebook to predict one's sexual orientation. On Twitter, latent social relationships can be revealed through conversation. In particular, we would like to study what information can be revealed as a *combination* of tweet content and social relationships through conversations. For example, more intimate language with LGBT individuals may inadvertently reveal your sexual orientation, which may be private for some people. Detection of such instances would serve as defenses through guardian angels. Another possibility is that social network analysis may identify which users are more likely to leak information, i.e., who are the likely 'gossips'. Guardian angel systems could monitor and score users through a combination of such content and network analyses to caution users against revealing too much through offline conversations with *gossiping* users.

Privacy leaks in other social networks. While this paper focuses on Twitter, it would be interesting to explore privacy leaks in other popular social networks such as Facebook and Google+. For example, Google+ (albeit in limited field trial during the time of writing), which also affords the social interactions provided by Facebook, allows people to post status updates publicly. People responding via comments to public updates may not realize their replies are

publicly viewable and it would be interesting to characterize such leaks. The detection of such leaks may differ for the various social networks. For example, not only are status posts in Facebook and Google+ longer than 140 characters in general, but leaks may occur through the action of "liking" various items, through picture tags, and so on. Finally, while we focus our analysis on settings where an individual may be harmed, it would be interesting to see if and how employees leak sensitive information about their companies.

10. CONCLUSIONS

We hope this paper highlights the privacy threats faced by users on social networking and microblogging sites such as Twitter. These users may not realize the implications of tweeting information publicly, about themselves and others, but as we show in at least three categories users can be implicated and impacted negatively. Our hope is that future *guardian angel* systems using similar classification techniques can be built to alert users of privacy leaks, giving them an option to remove their tweets, or think twice about posting a sensitive tweet.

11. ACKNOWLEDGMENTS

We would like to thank Minaxi Gupta, Nathaniel Husted, and our anonymous reviewers for their helpful comments. We would also like to thank the members of the Security Reading Group at Indiana University for their feedback.

12. REFERENCES

[1] A. Acquisti and R. Gross. Imagined communities: Awareness, information sharing, and privacy on the Facebook. *Privacy Enhancing Technologies (PET) Lecture Notes in Computer Science*, 4258:36–58, 2006.

[2] AFNER-Named Entity Recognition. `http://afner.sourceforge.net/`.

[3] AlchemyAPI. `http://www.alchemyapi.com/company/`.

[4] S. Bhagat, G. Cormode, B. Krishnamurthy, and D. Srivastava. Prediction promotes privacy in dynamic social networks. In *WOSN'10 Proceedings of the 3rd conference on Online social networks*, 2010.

[5] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating Twitter users. In *CIKM*, Toronto, Canada, 2010.

[6] L. David. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 4–15. Chemnitz, DE: Springer Verlag, Heidelberg, DE, 1998.

[7] C. Dwyer, S. R. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proceedings of the Thirteenth Americas Conference on Information Systems*, Colorado, August 2007.

[8] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *IMC '10 Proceedings of the 10th annual conference on Internet measurement*, New York, 2010.

[9] J. M. Gomez-Hidalgoy, J. M. Martin-Abreuy, J. Nievesx, I. Santosx, F. Brezox, and P. G. Bringas. Data leak prevention through named entity recognition. In *Proceedings of the 1st InternationalWorkshop on Privacy Aspects of Social Web and Cloud Computing*, 2010.

[10] L. Humphreys, P. Gill, and B. Krishnamurthy. How much is too much? Privacy issues on Twitter. In *Conference of International Communication Association*, Singapore, June 2010.

[11] Introduction to content analysis. `http://writing.colostate.edu/guides/research/content/pop2a.cfm`.

[12] C. Jernigan and B. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[13] B. Meeder, J. Tam, P. G. Kelley, and L. F. Cranor. RT@ IWantPrivacy: Widespread violation of privacy settings in the Twitter social network. In *Web 2.0 Privacy and Security Workshop, IEEE Symposium on Security and Privacy*, 2010.

[14] E. Mills. Twitter user says vacation tweets led to burglary. `http://news.cnet.com/8301-1009_3-10260183-83.html`, June 2008.

[15] Natural language toolkit. http://www.nltk.org/.

[16] Privacy, schmivacy! Twitter now lets you broadcast your location too... `http://www.csmonitor.com/From-the-news-wires/2010/0311/Privacy-Schmivacy!-Twitter-now-lets-you-broadcast-your-location-too`, March 2010.

[17] P. Singla and M. Richardson. Yes, there is a correlation: - from social network to personal behavior on the web. In *WWW '08: Proceedings of the 17th international conference on World Wide Web*, New York, 2008.

[18] Big goals, big game, big records. `http://blog.twitter.com/2010/06/big-goals-big-game-big-records.html`, June 2010.

[19] Y. Wang, S. Komanduri, P. Leon, G. Norcie, A. Acquisti, and L. Cranor. "I regretted the minute I pressed share": A qualitative study of regrets on Facebook. In *Symposium on Usable Privacy and Security*, July 2011.