Stop the Tweet Show : Preventing Harm and Embarrassment to Twitter Users

Xiao Xiao and Chris Varenhorst

April 3, 2009

Contents

1 Introduction **4**

2 About Twitter **5**

2.1 Posting and Following Tweets 5

2.2 Privacy Controls 5

2.3 Searching 6

3 Privacy and Twitter **7**

3.1 Twitter's Privacy Issues 7

3.2 Data Collection 8

3.2.1 Technical Details 8

3.3 What Data-miners Can Do 9

4 Related Work **10**

5 Proposed Systems **12**

5.1 Evaluation Criteria 12

5.2 Basic Changes to Twitter 14

5.2.1 Description 14

5.2.2 Implementation 15

5.2.3 Evaluation 15

5.3 System that alerts users 16

5.3.1 Description 16

5.3.2 Implementation 17

5.3.3 Evaluation 23

5.4 Privacy Controls for Users 25

5.4.1 Description 25

5.4.2 Implementation 26

5.4.3 Evaluation 27

5.5 Data-mining checks 27

5.5.1 Description 27

5.5.2 Implementation 28

5.5.3 Evaluation 28

6 Conclusion **29**

A About Twitter **35**

B Search Queries **37**

C Raw data **38**

1 Introduction

Twitter is a microblogging service that allows users to make frequent postings of short messages, or "tweets" so that friends, family, and co-workers can communicate and stay connected [4]. Since its inception in 2006, it has gained much popularity and now has at least 4 million users [6]. Even though Twitter's stated purposes is to enable users to communicate with only a small circle of friends, most often the contents of their posAts are viewable by a much wider audience. Tweets are public by default, which means that they are accessible by anyone, are indexed in search engines, and are shown on Twitter's "public timeline", a real time feed of all tweets. Furthermore, all tweets on Twitter are searchable through Twitter's search page and search API.

Because most users only intend to communicate with their friends on Twitter, they often post tweets that contain sensitive or embarrassing information that can cause long-term harm to them. For example, potential employers can find a user's Twitter account by email and search for tweets evident of unprofessional behavior. Insurance companies can deny users coverage based on evidence of pre-existing conditions from tweets. Law enforcement can also use Twitter as a tool. As an example of how casual online postings can have serious consequences, in 2006 the University of Colorado posted photos from Facebook of students smoking marijuana at a protest for the legalization of marijuana and offered a monetary reward for anyone who can identify people in the photos [1]. In a more recent example a Republican Congressman posted on his public twitter account detailed updates of his whereabouts during a confidential diplomatic trip to Baghdad. [5]

We propose and evaluate five systems that could prevent long-term harm to users that post personal, sensitive, and embarrassing information about themselves on Twitter, including basic changes to Twitter's tweet indexing, a pre-tweet and a post-tweet notification system, more privacy controls for users, and changes to Twitter's search function.

2 About Twitter

2.1 Posting and Following Tweets

Twitter allows its users to post short messages of 140 characters or less. Users can post tweets through the Twitter website itself (Figure 1), text messages on mobile phones, and various third-party Twitter clients such as Twhirl and Twitterific.

Users can subscribe to other users' updates by "following" them. The Twitter updates that a user follows shows up in the user's personal feed. Users can also set Twitter to send tweets from followed accounts to their mobile phones via text messages. In addition to the feed of updates from followed accounts, users can also monitor Twitter's global public timeline on Twitter.com to see what everyone else is updating in real time. In a 2008 study, social networking researcher Balachandran Krishnamurthy describes three broad categories of Twitter users based on the ratio of their followers and those they follow. Those with a long followers list but short following list are generally "broadcasters" such as news services and popular, public blogs while those with a long following list but short followers list are often spammers. The vast majority of personal Twitter users are what Krisnamurthy categorizes as "acquaintance users", those whose followers and following lists are roughly the same size [7]. These users use Twitter to communicate with their friends and are the users whose tweets we discuss in this paper.

2.2 Privacy Controls

There are very limited privacy settings on Twitter. Users can set their accounts to be either public or private. Tweets from public accounts are displayed on Twitter's global public timeline, are searchable through Twitter's search page and API, and are indexed by

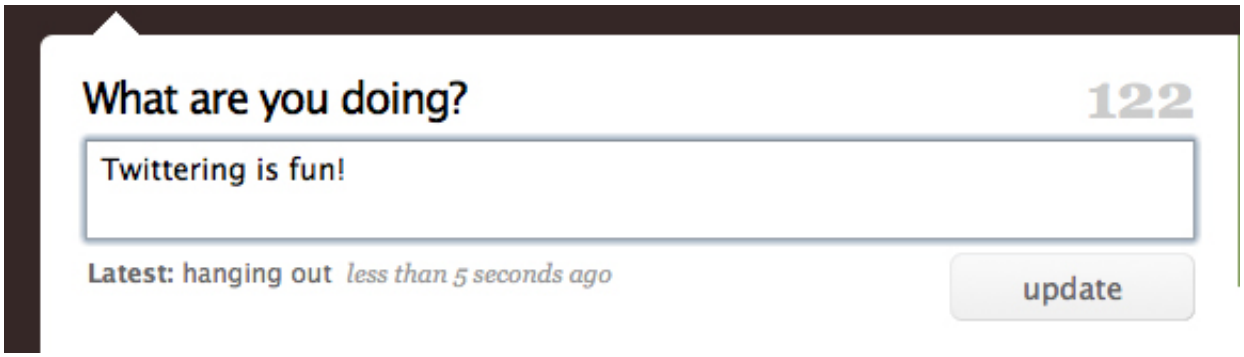

Figure 1: Screenshot of someone submitting a tweet through Twitter's web interface.

search engines. Tweets from private accounts are only visible to approved followers of the accounts and are not searchable. Accounts are public by default; 99% of Twitter users keep the default public setting [7].

2.3 Searching

Twitter has a search page, where users can search for other users' tweets by keywords or phrases. Results are displayed with the most recent tweets first. Twitter also has a search API that allows third party applications to query tweets. Twitter attempts to limit data-miners by placing checks on the search API. An application can only make 200 requests in 1 hour, and the results only include tweets from the past 11 days. However, Twitter does grant requests for a higher limit to some applications. Third party applications can also scrape Twitter data directly from Twitter's pages such as the public timeline, users' individual pages, and Twitter's online search page. Unlike the Twitter search API, there is no limit to scraping.

In addition to searching for tweets, Twitter also allows users to search for other users by email, name, or location. Twitter has a "Find Your Friends" feature that allows a user to check if people in their email address book have Twitter accounts. While users' email

addresses are not present on their Twitter pages, the users can be easily found by their email through the "Find Your Friends" feature. Twitter encourages users to place their real name and location on their Twitter page so that their friends and acquaintances can more easily find them on Twitter. However, because real names and locations are completely public on every user's Twitter page, this information is also available to everyone.

3 Privacy and Twitter

3.1 Twitter's Privacy Issues

Although Twitter is a powerful communication tool for both friends and strangers, it has several privacy issues that must be addressed. First, Twitter's current privacy setting is very misleading to users. A tweet posted from an account when it is set to public will always remain public, even if the user changes the account to private in the future. This means that the tweet will always be searchable even when it is no longer viewable from the user's Twitter page. Similarly, a post that is deleted can still show up in searches even when it is no longer visible on the user's Twitter page.

Second, the intimacy of Twitter conversations among groups of friends can give users a false sense of privacy. Most Twitter users use Twitter to update their small group of friends and expect only their friends to read their tweets, but these tweets are not only shown on Twitter's public timeline but are also searchable trough Twitter and Google and are visible by anyone. In addition, Twitter users themselves can be found by people who are not their close friends through email, real name, or location. This make it easy for electronic eavesdroppers to tune in to the updates and conversations of their targets.

3.2 Data Collection

We programmatically collected embarrassing tweets on Twitter for analysis. Our data collection exercise demonstrates how individuals could systematically collect sensitive information from Twitter. Overall we collected 988 embarrassing tweets over 11 days (See Table 1).

3.2.1 Technical Details

To collect data from Twitter, we wrote Perl scripts that accessed Twitter's JSON API. We discovered the rate limiting mechanism in Twitter's API which prohibits a particular IP address from making more than 200 requests in an hour. We often needed to collect data faster than this, so to bypass this restriction, the data collection program systematically rotated IPs on the network it was connected to. Since MIT's network has an overabundance of address space, this method got around Twitter's restrictions.

To demonstrate how embarrassing tweets can be programmatically harvested we first identified 4 categories of sensitive or embarrassing information: drug use, irresponsible drinking behavior, intimate sexual details, and damaging work related tweets. We then assembled 16 different queries by hand that returned tweets the majority of which easily fall in one of our categories. We queried Twitter using its search API to collect the results. (See Appendix C for the raw data) From the data we collected, we drew some conclusions about the practices of Twitter users posting embarrassing tweets.

We found that only 2% of the offending tweets were posted by users that had their account set to private. Two percent is slightly above the Twitter-wide average of 1% [7] but is still extraordinarily low. Our finding suggests that users are either not aware that their embarrassing tweets are accessible by everyone or do not consider exposure of these tweets

Table 1: **Summary of data mining. Tweets were collected used 16 hand made search queries which by manual testing proved to return embarrassing tweets.**

	Total	Protected Tweets
Drug Use	340	11
Irresponsible Drinking Behavior	150	2
Intimate Sexual Details	98	3
Irresponsible Work Behavior	400	4
Total	988	20

to be harmful.

3.3 What Data-miners Can Do

The most worrisome scenario is when an adversary uses Twitter to find information about a particular target. Twitter makes this relatively easy by indexing users names, location, and email addresses. Twitter's "find your friends" feature lets anyone match an email address to a Twitter account. Because search engines also index Twitter, even if a user takes down damaging tweets, an adversary can still find cached copies of the tweets.

Table 2: Source of embarrassing tweets, their percentage of that total, and how that deviates from the Twitter wide average. From this we can conclude that people generally post the most embarrassing tweets from mobile devices.

Source	# of tweets	percentage of total	difference from Twitter average
web	480	49.5 %	-7.5
3rd party	302	31.1%	-4.4
mobile	186	19.2%	+ 11.8

4 Related Work

We looked at three existing systems that address the leakage of personal, sensitive, or embarrassing information on the Internet, each of which approach the problem from a different perspective. First, we studied Latanya Sweeney's Privacy Angel, a project that tries to prevent identity theft by informing people that their birthdays and Social Security numbers are on the Internet. Many people put their birthday and Social Security number on their resume and have their resume avaliable online. They don't realize that identity thieves can easily find that information through their resumes and impersonate them. Privacy Angel crawls the web to find instances of people's personally identifying information avaiable on the Internet and sends a message to inform them of that fact, telling them about the possible negative consequences and recommending them to take the personally identifying information off the Internet. Similarly, as can be seen from the result of our data analysis that very few people keep their accounts private, Twitter users also seem to post personal information in their tweets without realizing the implications. Following the Privacy Angel model, we can prevent longterm harm to Twitter users posting sensitive information about themselves on Twitter by informing them and recommending that they remove the personal tweets and think twice about tweeting personal information in the future.

Another system that we examined is Google Labs's Mail Goggles feature. Mail Goggles tries to prevent people from sending potentially embarrassing emails late at night when they are likely drunk or highly sleep deprived. Users can choose to turn on the feature; Those with Mail Goggles enabled are forced to correctly complete several simple arithmetic problems when they try to send emails on times they indicate they are likely to be drunk. This system is an example of a preemptive guard system. Unlike Sweeney's Privacy Angel that informs users after the personal information is already on the Internet, Mail Goggles tries to prevent embarrassing informations from being sent out in the first place. Following

the Mail Goggles model, we can add a feature to Twitter that sets up an extra check before users post potentially personal, sensitive, or embarrassing information.

The third system that we analyzed is Livejournal, a popular blogging community. Unlike Twitter, Livejournal allows users to dictate the privacy setting of each post. Users can specify whether entries are available for everyone, only for approved "friends", or for a subset of approved "friends". Friends on Livejournal are similar to followers on Twitter. If Alice adds Bob as a friend on Livejournal, Alice gets Bob's public entries on her friends page, a feed of entries from her friends. If Bob adds Alice back, Alice can then see all of Bob's "friends only" entries as well. It is much more difficult for users to search for individual people on Livejournal. Users choose whether someone can find they by their primary email address on Livejournal. Users can specify "yes", "yes but not show the Livejournal username", or "no". Even though users can still potentially be found on Livejournal by their email, the users have more control over their degree of privacy than on Twitter. Finding specific information in individual entries is also difficult. Livejournal entries are not indexed by Google, are not present in a global public feed, and are not searchable through Livejournal. People can search Livejournal users by the user's interests, but interests is something that users can easily control. Although Livejournal's privacy settings are not perfect, they give users a higher degree of control of what to make public and what to make private. Livejournal also limits searching so that users cannot be easily found and targeted.

Each of the existing systems that we examined takes a different angle in preventing the long term harm of users posting personal information on the Internet. Our proposed systems take inspiration from these systems and approach the Twitter problem from several angles.

5 Proposed Systems

5.1 Evaluation Criteria

We evaluate our proposed systems according to four criteria. The first and most important criteria for any solution is how effectively it solves the problem. Since our proposed solutions targets different areas and vary in scope, we evaluate what part of the problem each system focuses on and how effectively it addresses the specific part. Target groups include people who are not aware that they are posting embarrassing information on Twitter, people who are aware but post the information anyway because of Twitter's lack of controls, and malicious third parties.

Second, we consider how much our changes take away from the current Twitter experience. Despite its unresolved privacy issues Twitter is nevertheless a useful and effective tool for communication. Twitter's biggest innovation is convenience and its openness. Twitter's convenience is no doubt a big reason for its popularity. Users frequently tweet because submitting a tweet only takes a few seconds and can be done from a computer or a mobile phone. Although Twitter's openness makes it a powerful forum of speech online. Users can watch the public timeline or search to see what other from all around the world are saying. When a major event happens, people can watch real-time updates from Twitter users on the scene about the event as it unfolds. Political candidates and news agencies can speak more directly to people through Twitter updates. While Twitter's openness is part of the reason why users may experience long-term harm from tweeting personal or sensitive information, Twitter's overall philosophy of sharing speech is positive and should be preserved as much as possible in any proposed solution.

Third, we consider how likely the system will be used and accepted. We first examine how easily ignored the system is. A system that is easily ignored may not have as much effect

on the user as one with a louder message, but the latter may annoy users and fail to achieve its goals. How likely the system will be used and accepted also ties in with how much the system affects the user experience of Twitter. A system that modifies very little of the Twitter user experience should generally be more accepted.

Finally, we analyze the ease and cost of implementation. A simple, elegant system that is easy to implement is obviously preferable to a complex system that takes up more money and development time. An easy to implement system is also more likely to be realized and accepted. Some systems may call for drastic changes to Twitter's underlying architecture while others can be developed completely by a third party. While a change to Twitter itself could potentially have more impact, third-party systems are more convenient to implement. Whether the solution involves Twitter or only a third party is also related to how likely the system will be accepted. In general, changes coming from Twitter would be used more than third party systems because third party systems are necessarily opt-in systems for Twitter users.

In general, the criteria of how much changes take away from the Twitter experience and how likely the system will be accepted have more weight than the ease and cost of implementation unless the system is near impossible to implement or require an extravagant cost.

Table 3: Criteria that the proposed systems are evaluated on

Evaluation Criteria
Target problem that the system tries to solve
How much of the target problem does it solve
Amount of loss of current Twitter user experience
Likeliness of system to be used and accepted
Ease and cost of implementation

5.2 Basic Changes to Twitter

5.2.1 Description

The first system that we propose include basic changes to remove loopholes that allow people to view private and deleted tweets. We noticed while doing our datamining exercise that tweets deleted from private accounts, even accounts are visible as search results. We did some tests with our own Twitter accounts and discovered that once a tweet is indexed by Twitter's search feature, it is not removed even when the user deletes the tweet.

New tweets from a public account are pushed to the public timeline and are indexed by Twitter's search as they are posted. New tweets from a private account are not pushed to the public timeline and do not get indexed by Twitter's search. If an account changes its privacy setting, it only effects new tweets that are posted after the setting change. Old tweets that were already indexed by Twitter's search are not removed, and old tweets that were never indexed by search do not get indexed retroactively. This creates a discrepancy between the setting of an account and the actual privacy of individual tweets in the account.

That tweets can never be removed from Twitter's search index and the discrepancy between the privacy of an account and the privacy of individual tweets are big problems in Twitter's design because even when users attempt to hide or delete tweets that contain sensitive or embarrassing information from their personal Twitter pages, people can still view the tweets through Twitter's search function. In effect once the tweet has been published a user can never take it back, whether by setting the account private or by deleting the tweet. To see a user's hidden tweets in addition to the visible ones, all someone has to do is use Twitter's advanced search and query tweets from that user. Furthermore, the fact that private and deleted tweets are visible through Twitter's search

deceives users and gives them a false impression of privacy.

5.2.2 Implementation

We propose a simple solution to fix Twitter's problem with private and deleted tweets. Since private and deleted tweets are visible only in search results, Twitter can implement simple checks on tweets that match search queries before displaying them. To prevent tweets from private accounts from showing up in search, Twitter can check the current privacy status of a tweet's parent account before displaying it in the search results. If either the tweet or the parent account is set to private, do not show the tweet. To prevent deleted tweets from appearing in searches, Twitter can do an extra check to see whether a tweet has been deleted before displaying it in search results.

5.2.3 Evaluation

This system targets the problem of people unable to hide a tweet that could contain personal information once it has been posted. Since 99% of Twitter users keep their accounts public and only 2% of tweets found in our datamining demonstration were private or deleted, the system only directly addresses a small number of Twitter users. However, it fixes a huge loophole in Twitter's search and result display. Thus, even though it targets a small part of the big problem, it does so very effectively. This system does not modify user experience at all. A user still tweets the same way and can still search for tweets. Nothing is affected except for the loopholes in Twitter. As a result, users of Twitter should not have any complaints about this system. Because the system only involves small changes to Twitter in the display of tweets and does not involve large changes to data or database structure, it should not be difficult or costly to implement.

5.3 System that alerts users

5.3.1 Description

The second system we propose alerts Twitter users when they post a tweet that may contain sensitive or embarrassing information by sending a message to the user notifying them that they have posted some personal information and explaining the possible consequences. This system targets users who many not be aware of the extent that tweets are public on Twitter and may not realize what can happen when they post personal information in tweets. From the results of our data-mining exercise, this seems to be the majority of users who post embarrassing tweets.

Alerting users can either occur before or after a tweet is sent to Twitter. In the pre-tweet alert system, users get an alert when they try to send in a tweet to Twitter if the tweet has a high likelihood of containing sensitive or embarrassing information. The alert asks whether users are sure if they want to submit the tweet anyway. If they select yes, the tweet is then published on Twitter. If they select no, the tweet is never published.

In the post-tweet alert system, users submit a tweet as usual but receive a private message on Twitter soon after the tweet is published. The message sends the user a link to the tweet, explains that the tweet may contay embarrassing information, and suggests to the user that they delete the tweet. The pre-tweet and post-tweet system have several other differences. The post-tweet system may be easiser to implement than the pre-tweet system, but requires that the loophole for deleted tweets to be fixed in order to be completely effective. Both systems not only alert users about the tweet in question but also make users more aware when they tweet in the future.

5.3.2 Implementation

Filter

Both the pre-tweet and the post-tweet system need an automated mechanism for figuring out whether a tweet contains sensitive or embarrassing information. One way to find tweets that contains embarrassing information is by creating a filter that searches for key words and phrases, much like what we did for our datamining exercise. First, implementers of the system need to determine out what are personal and embarrassing areas about which people can tweet. In our datamining exercise, we came up with four specific categories of embarrassing or character damaging information - illegal drug use, irresponsible drinking behavior, explicit intimate details, and complaints about work- for which to come up with search terms. The user alert systems can use a similar process to gather search words and phrases for sensitive tweets. The categories for sensitive information do not have to be limited to the ones we used for our user study and can include other areas such as personal health. After deciding on search terms, implementers can test them out by searching for the terms in Twitter's existing tweets. There is a balance between high hit rate of a search query and the number of relevant tweets actually found. A narrow search may have a very high hit rate but may miss many more relevant tweets than a wider search. After settling on a set of search terms, the filter can be used to see whether a tweet has a high likelihood of containing personal, sensitive, or embarrassing information.

Pre-Tweet

The pre-tweet system must have an implementation for web, for text, and for third party Twitter clients in order to be most effective because tweets can come from any of these sources. It only checks tweets that are posted from accounts set as public. For the web version, after a user finishes a tweet and clicks the "update" button on Twitter, the new tweet goes through the previously described filter. If the filter determines that the tweet has a high likelihood of containing sensitive information, Twitter displays a message

explaining that the user has posted a tweet that may contain sensitive or embarrassing information in the specific category and what harm that could cause the user. The message could even give a link to a separate page on Twitter with a more thorough explanation of the problems of posting personal information on Twitter and what this system does. At the bottom of the message are two buttons- "do not update" or "update anyway", which lets the user decide whether they still want to publish the tweet. An option to ignore future warnings in that category should also exist (See Figure 2).

For the text message version, after a user texts a tweet to Twitter's number, Twitter passes the tweet through the filter. If the filter determines that the tweet has a very high likelihood of containing sensitive information, Twitter sends back a text message explaining that the user has posted a possibly personal tweet, and the tweet is put in "limbo". The message instructs the user to text "1" back in order to publish the tweet anyway. If the user never texts "1" back, the tweet is stored in limbo for a certain number of days on Twitter and then deleted. User can see a list of limbo tweets and can decide to either publish or delete the tweets from the Twitter web interface. Due to the 140 character limit for SMS messages, the text version must send a shorter alert message to the user than the web version. A sample message sent back from Twitter may look like Figure 3. A page with all the limbo tweets may look like Figure 4.

The pre-tweet alert system for third-party clients will require a bit more work. A new set of API calls dealing with messages in "Limbo" will have to be developed. It would be important to maintain backward compatability for a while, so at first third-party Twitter clients would likely not be required to deal with limbo messages.

When the pre-tweet system is first implemented, it should automatically notify all users with public accounts when they try to submit tweets that may contain personal information. After a notification has been given to the user at least once, the user can choose to disable future notifications from the Twitter setting page, indicating that they

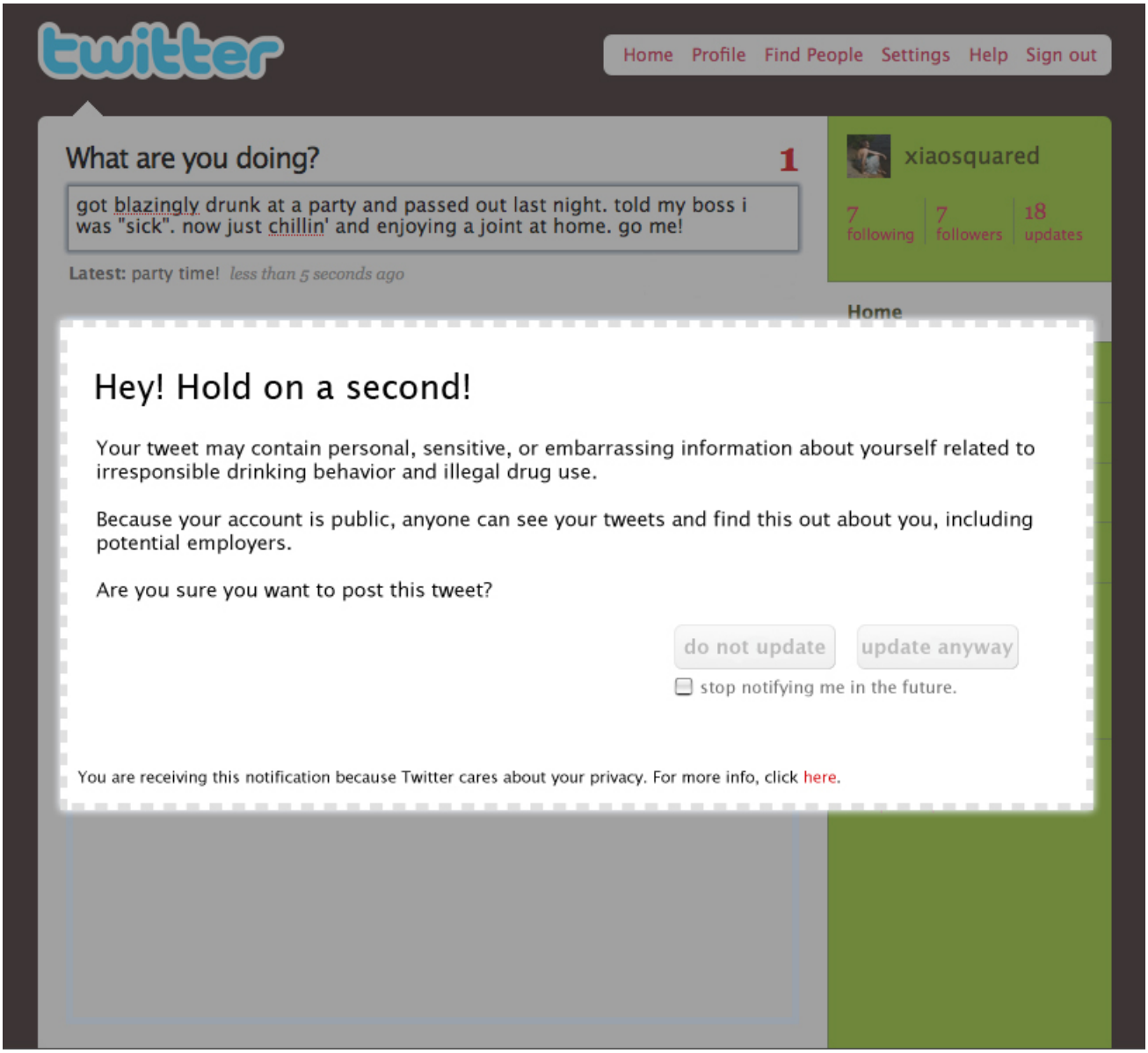

Figure 2: Twitter mockup of proposed embarrassing tweet warning system. The system would identity potentially damaging tweets based on keywords.

Figure 3: Twitter mockup of proposed embarrassing tweet warning system as a text message. At the users request, Twitter holds tweet it thinks the user may find embarrassing until the user approves them.

understand that everything they post is public and that they understand the possible ramifications.

Post-Tweet

The post-tweet system can be implemented either as part of Twitter or by a third party. The system constantly monitors Twitter's public timeline and filters all public tweets with the previously described filter. If the filter determines a tweet has a high probability of containing sensitive information, the system sends a message to the user who posted the tweet. If the system is implemented by Twitter, the message can be an email sent to the address provided to Twitter. The email can display the tweet in question and explain that the tweet may contain sensitive information, that the tweet is completely public and can be seen by anyone, and that this could harm the user in the long run. The email can give a link to the user where they can delete the tweet. If the system is implemented by a third party, email cannot be used to contact the user because a user's email is not directly

Figure 4: Twitter mockup of proposed embarrassing tweet warning system. At the users request, Twitter holds tweet it thinks the user may find embarrassing until the user approves them. All tweets shown are real tweets we recorded during data collection.

exposed by Twitter. Thus, a direct message is sent to the user on Twitter. Because direct messages are limited to 140 characters, the notification must be very brief and could look like the following: "your tweet http://tinyurl.com/a3d5b3 may contain sensitive information. since it is public, anyone can see it. please consider deleting it." More information about the system can be displayed on the Twitter page of the account sending the direct message as demonstrated in figure 5. The post-tweet system can not only alert users who are posting new tweets but can also go through all previous tweets on Twitter, search for ones that may contain sensitive information, and notify the users.

Figure 5: Twitter mockup third party accoun that sends direct message warnings to users who have posted personal, sensitive, or embarrassing tweets. Because direct messages are limited to 140 characters, the Twitter page of this account gives more information to users.

5.3.3 Evaluation

Pre-Tweet System

Because the pre-tweet and post-tweet alert systems differ significantly in function and implementation, they are evaluated separately. The pre-tweet system acts as a watchdog to prevent users from posting personal information publicly on Twitter without consideration. It also promotes awareness in users of how public tweets really are and familiarizes users with the harms of making personal information completely public on Twitter. The system targets both users who are not aware and users who post material on Twitter in a spur of the moment without considering the consequences and is fairly effective in addressing its goals. However, the effectiveness of the system is related to how good the filter is, and sometimes the filter may miss tweets.

In terms of user experience, the pre-tweet system creates a minor to moderate inconvenience to the user depending on the version. The web version and the third party client version have only minor inconveniences because they only force the user through one single prompt. The mobile phone text message version has a bigger inconvenience because it sends a text message to the user and forces the user to text in an additional message in order to post something the system considers embarrassing information. If users do not get the alert text immediately, their original tweet may be very delayed if they do decide to publish it anyway. This detracts from Twitter's user experience because a main appeal of Twitter is that users can send in rapid, up to the minute updates.

How likely the pre-tweet system will be used will largely depend on how accurate the filter system is. If the filter produces too many false positives, users may get annoyed and disable the filter. Despite that, because a large portion of the system is simply to make users more aware when tweeting, as long as users receive and read the notifications, the response of the system is passed along to them.

The pre-tweet alert system has some complications when it comes to implementation. First, because people can tweet from web, phone, and third party clients, both Twitter and all third party clients of Twitter must be modified in order for the system to reach everyone. However, it may be sufficient just to add the functionality to the few most popular third party clients. Second, making a good filter for the system may take much experimentation. To find the most effective search terms for the filter, people need to come up with terms, search for them on Twitter, and read through many search results to decide on the quality of the terms. Making a thorough filter would take time.

Post-Tweet system

The main goal of the post-tweet notification system is to promote awareness in users so that they think twice before posting personal information on Twitter in the future. If implemented on top of Twitter as it currently is, asking users to delete questionable tweets does not actually help protect users' privacy because of the loophole on Twitter that deleted tweets can still be seen through Twitter's search. This system would only be effective if it were implemented alongside a fix for the deleted tweet loophole.

The post-tweet notification system does not take away anything from the current user experience because all aspects of tweeting remain the same. Reading the notification message or email can be seen as an inconvenience for the user, but given that direct messages on Twitter are limited to 140 characters and that Twitter users often receive direct message and emails anyway, the inconvenience is negligible.

As is the case for the pre-tweet system, how likely the post-tweet system will be used and not ignored depends on the accuracy of the filter system. One thing about the post-tweet system is that it is easier to ignore the notification because it is not immediate. The pre-tweet system forces the user to make a decision about whether to post the tweet or not as soon as the user tries to update. The post-tweet system sends a message that can be more easily ignored so its effectiveness is less than the pre-tweet system's.

In terms of implementation, the post-tweet notification system is significantly easier to implement than the pre-tweet alert system because only one version needs to be implemented. Both the Twitter dependent and completely third party versions are simple to realize. The difficulty of implementation lies in constructing a good filter as discussed earlier.

5.4 Privacy Controls for Users

5.4.1 Description

We propose a systems that gives users more granular privacy options so that they can have precise control of who has access to particular tweets. As discussed above, Twitter only has an account-wide binary privacy setting, either making every new tweet public, or every new tweet private. By allowing users to control privacy at a per-tweet level, users will be able to continue to tweet and share information as they do, but when they wish to make an occasional more personal tweet, they can do so without making it public and without changing account wide settings.

The biggest potential problem with giving users such control is finding a way this can fit into the current Twitter paradigm without altering the experience. To address this, we propose a special character, $ at a beginning of a tweet that will let Twitters system know that this tweet is private. This is consistent with other practices on Twitter of using special characters to communicate auxiliary data. The @ character before a username indicates the tweet is addressed to that user. The # character before a word indicates a category for that tweet.

Users will place a "$" before any potentially embarrassing tweet, making it private. A private tweet will act the same way a tweet made from a private account will work. The

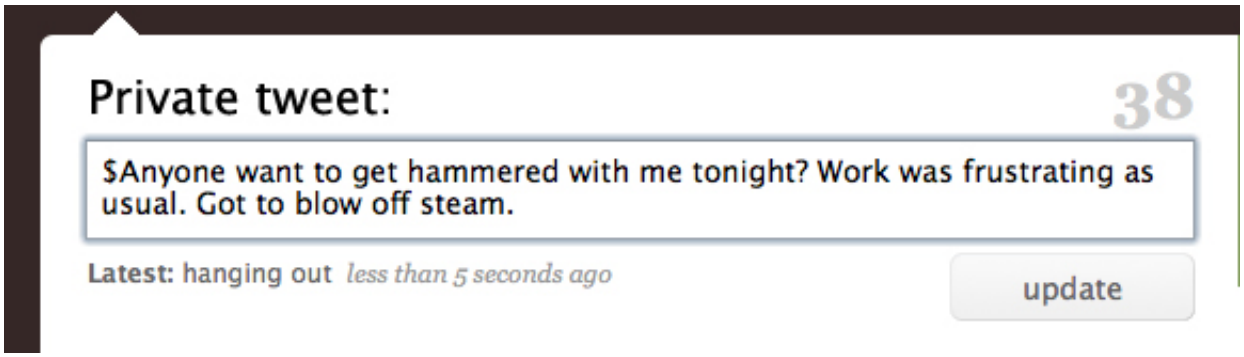

Figure 6: Twitter mockup of proposed private tweet system. Tweets that begin with the $ character would only be sent to approved followers.

tweet will not go on the public timeline, will not be returned in searches, and will only be visible to approved followers on the user's Twitter page. This system also proposes a change to Twitter's follower system. In Twitter's current follower system, users public accounts do not approve their followers while users in private accounts must approve all followers. In this new system, every account has "approved followers" in addition to the normal followers currently present on Twitter. Anyone can become a follower of a user and get the user's public updates sent to their feed. A user can request to be an approved follower them to see all of a someone's tweets on the Twitter page and to have both public and private updates sent to their feed. As indicated by the name, approved followers must be approved by the user they are trying to follow.

5.4.2 Implementation

Neither the Twitter search API nor tweeting through 3rd party client need to change at all, and tweeting through txt and third party clients can continue as it always has. The key difference for Twitters technology is that each tweet will have a property specifying whether it is public or private. A tweet is public by default, unless a $ character is the first character of the tweet, which indicates it is private. Twitter will also need to develop a simple interface supplement to allow users to retroactively mark a tweet as private or

public. When displaying tweets, Twitter will not only check whether the account is public or private but also check whether a tweet is public or private. If a tweet is private, it is not displayed on the timeline, on the Twitter page, and in search results.

5.4.3 Evaluation

Private tweets will allow users that want to communicate an occasional personal note only to their close friends a much more convenient way to do so. Thus, its primary target is people who are aware that they are posting personal information on Twitter but do so anyway because they would rather keep most of their tweets public. For the primary target audience, this feature is very effective. It also does not negatively impact the current user experience at all, especially because the mechanism to do so is only a single character. Those wishing to ignore this feature are welcome to do so. However, the effectiveness of this feature is unclear for other users of Twitter is unclear. This feature will only help users keep damaging information private when they recognize the dangers of that information being public in the first place. As a result, this feature will probably be used and accepted but only by people who have already some sort of awareness of the privacy issues of Twitter. The cost of implementation is minimal beceause the systen can easily integrate into the existing Twitter service.

5.5 Data-mining checks

5.5.1 Description

Embarrassing tweets are only a problem when they can be found by people searching for them with malevolent intent. Something can be public on the internet, without it being searchable or easily located. We proposes a series of changes to Twitter that obstructs the

finding of specific information. There are three main changes: reducing the amount of history Twitter stores, preventing search engines from indexing tweets, and disabling the lookup of Twitter accounts using email addresses without prior approval from users.

5.5.2 Implementation

Twitter can programmatically remove tweets older than 7 days. In most cases this would not affect users very much because Twitter's purpose is for friends to share streams of up to the minute tweets. Twitter can keep tweets from being indexed by search engines by applying a simple set of tags to its pages. The document A Method for Web Robots Control outlines an HTML meta tag that lets search engine crawlers know that a page is not to be index. This standard is follow by every major search engines.

Twitter can follow these procedures to prevent tweets from being indexed in search engines. Twitter currently has a system in place for finding people's usernames given their email address. Twitter can alter this function so that the username is only revealed to the searcher if that user has approved the searcher. This lets users control who looks them up with their email and prevents strangers from identifying a target user's account without the user's knowledge and consent.

5.5.3 Evaluation

This system targets data-miners and tries to hinder their ability to find information about specific Twitter users. The system addresses the problem with a reasonably high effectiveness as it would severely limit a user's ability to find compromising information about particular Twitter users. However, the system is not 100% effective and does come at a marginal loss of utility. Potentially damaging information is still "public" and freely viewable by anyone; it just takes more work to find it with this system in place. It is

conceivable that some particularly determined individual could deploy their own web crawler that does not abide by Web Robot Controls.

This system also takes away some aspects of the current Twitter experience. With a reduced Twitter history, users will not be able to look up old tweets from users, and infrequent Twitter users may miss some tweets altogether. These changes will also make it more difficult for people to find other Twitter users they know, because they will no longer be able to look them up using their email address as easily. This protects users because they can more easily control who has their Twitter name then who has their email address. However, most changes are barely noticible to users. The average user will likely not even notice that tweets are no longer indexed by search engines such as Google. Implementation costs for this system are relatively low, requiring only a few minor changes.

6 Conclusion

Even though personal users of Twitter use it as a way to communicate with their group of friends, almost all the users keep all their tweets completely public and accessible to anyone. Users sometimes post tweets for their friends that contains sensitive, or embarrassing information about themselves, which can cause long term harm to the users especially since individual people can be found on Twitter through their email. Users' personal tweets may be found by potential employers, insurance companies, and even law enforcement, which can have negative consequences for the users.

In this paper, we proposed five systems that help prevent long term harm to Twitter users who post personal information, each of which targets a different group of users. First, we proposed basic changes to Twitter to fix the loophole that tweets from private accounts and deleted tweets can be seen through search. Second, we discussed pre-tweet and

Table 4: Summary of Evalution of proposed systems

Systems	*Users that are targeted*	*Effectiveness*	*Effect to Twitter experience*	*Likelihood to be used and accepted*	*Ease and cost of implementation*
Basic changes to Twitter	Users who tried to hide already posted information	Very effective at target users	Almost none	Very likely	Twitter dependent. Not difficult or costly to implement.
Pre-tweet notification system	Users who post personal tweets and don't realize how public their tweets are	Somewhat effective. Users can choose not to heed the message.	Very little for tweets from web. Some annoyances for tweets from text	Likely but some users may choose to turn it off after getting the message.	Twitter dependent. Some complications of implementation because of various ways of tweeting
Post-tweet notification system	Users who have posted personal tweets who don't realize who can see them	Somewhat effective. Users can choose not to heed the message. Also, unless basic changes are implemented, deleted posts are still visible	Very little	Easy to be ignored.	Can be Twitter dependent or not Twitter. Not difficult of costly to implement
Per-tweet privacy controls	Users who want some tweets private but can't easily do so currently on Twitter	Very effective at target users Difficult to measure how much message actually influences people	Almost none. Difficult to measure how much message actually influences people	Very likely to be used but may not be used as much by people who aren't the target	Twitter dependent. Not difficult or costly to implement.
Limits on Data-mining	Data-miners who targets users	Somewhat effective. determined data-miners can still find ways to get around limits.	Various- limits on users finding their friends by email, users may miss tweets because of reduced history	Likely to be accepted by users but there are some annoyances.	Twitter dependent. Not difficult or costly to implement.

Systems	Effectivness	Negative impact to Twitter experience	Likelihood to be used and accepted	Ease and cost of implementation
Basic Changes to Twitter				
Pre-tweet notification system				
Post-tweet notification system				
Per-tweet privacy controls				
Limits on Datamining				
legends	very somewhat not very	almost none very little some	very likely likely less likely	easy and cheap some difficulties significant difficulties

Figure 7: Comparative chart by color

post-tweet notification systems that inform the user that their tweet may contain personal information. Third, we suggested adding a per-tweet privacy feature on Twitter. Finally, we presented changes to Twitter that make it more difficult for data-miners to target individual users. We analyzed each of the systems according to how effective the systems are for the target group, how much the system detracts from the overall Twitter experience, how likely users are to use the system, and the difficulty and cost of implementation.

Based on our proposals and our evaluations, the best way to address the problem should incorporate all of our proposed systems. The most effective system would also involve Twitter itself to make some changes. Twitter definitely needs to fix deleted tweets and tweets from private accounts from persisting in searches. It also should implement a per-tweet privacy setting. These two changes would give people more control over their privacy on Twitter. Twitter should also make it more difficult for data-miners to target individual people on Twitter. Finally, Twitter could implement some sort of notification system that alerts users when they are posting something personal, sensitive, or embarrassing, which would inform users who are not aware of how public Twitter is. This would promote awareness among users to use Twitter more responsibly. A combination of parts of our four proposals would greatly prevent users from long term harm of posting public tweets intended only for friends.

References

[1] University of colorado puts photos online of students smoking marijuana. http://www.associatedcontent.com/article/31074university_of_colorado_puts_photos.html, 2006.

[2] Livejournal. http://www.livejournal.com, 2008.

[3] Mail goggles. http://gmailblog.blogspot.com/2008/10/new-in-labs-stop-sending-mail-you-later.html, 2008.

[4] Twitter's api documentation. http://apiwiki.twitter.com/REST+API+Documentation, 2008.

[5] Congressman becomes latest twitter blunderer. http://riverscrap.typepad.com/home/2009/02/congressman-becomes-latest-twitter-blunderer.html,
2009.

[6] Social networks site usage: Visitors, members, page views, and engagement by the numbers. http://www.web-strategist.com/blog/2008/11/19/social-networks-site-usage-visitors-members-page-views-and-engagement-by-the-numbers-in-2008,
2009.

[7] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 19–24, New York, NY, USA, 2008. ACM.

[8] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, pages 47–42, New York, NY, USA, 2008. ACM.

[9] L. Sweeney. Protecting job seekers from identity theft. *Internet Computing, IEEE*, 10(2):74–78, March-April 2006.

A About Twitter

The following information comes from Twitter's own About page. It can be found on the web at www.twitter.com/about

About us Twitter is a privately funded startup with offices in the SoMA neighborhood of San Francisco, CA. Started as a side project in March of 2006, Twitter has grown into a real-time short messaging service that works over multiple networks and devices.

In countries all around the world, people follow the sources most relevant to them and access information via Twitter as it happensfrom breaking world news to updates from friends.

Where did the idea for Twitter come from? Jack Dorsey had grown interested in the simple idea of being able to know what his friends were doing. Specifically, Jack wondered if there might be an opportunity to build something compelling around this simple status concept. When he brought the idea up to his colleagues, it was decided that a prototype should be built.

Twitter was funded initially by Obvious, a creative environment in San Francisco, CA. The first prototype was built in two weeks in March 2006 and launched publicly in August of 2006. The service grew popular very quickly and it soon made sense for Twitter to move outside of Obvious. In May 2007, Twitter Incorporated was founded.

Why do so many people seem to like Twitter?

Simplicity has played an important role in Twitter's success. People are eager to connect with other people and Twitter makes that simple. Twitter asks one question, "What are you doing?" Answers must be under 140 characters in length and can be sent via mobile texting, instant message, or the web.

Twitter's core technology is a device agnostic message routing system with rudimentary social networking features. By accepting messages from sms, web, mobile web, instant message, or from third party API projects, Twitter makes it easy for folks to stay connected.

Isn't Twitter just too much information? No, in fact, Twitter solves information overload by changing expectations traditionally associated with online communication. At Twitter, we ask one question, "What are you doing?" The answers to this question are for the most part rhetorical. In other words, users do not expect a response when they send a message to Twitter. On the receiving end, Twitter is ambient–updates from your friends and relatives float to your phone, IM, or web site and you are only expected to pay as much or as little attention to them as you see fit.

The result of using Twitter to stay connected with friends, relatives, and coworkers is that you have a sense of what folks are up to but you are not expected to respond to any updates unless you want to. This means you can step in and out of the flow of information as it suits you and it never queues up with increasing demand of your attention. Additionally, users are very much in control of whose updates they receive, when they receive them, and on what device. For example, we provide settings for scheduling Twitter to automatically turn off at dinnertime and users can switch off Twitter updates at any point.

Simply put, Twitter is what you make of it–receive a lot of information about your friends, or just a tiny bit. It's up to them.

How is Twitter built? Our engineering team works with a web application framework called Ruby on Rails. We all work on macintosh computers except for testing purposes. Our web site and user interface were designed using Omnigraffle and Photoshop.

We built Twitter using Ruby on Rails because it allows us to work quickly and easily–our

team likes to deploy features and changes multiple times per day. Rails provides skeleton code frameworks so we don't have to re-invent the wheel every time we want to add something simple like a sign in form or a picture upload feature.

B Search Queries

Irresponsible Drinking Behavior

- "my fake ID"
- "got hammered" OR "get hammered"
- shitfaced

Illegal Drug Use

- I smoked high -"high school" -"junior high" -"high horse" -"High Fidelity"
- rolled OR roll OR rolling blunt OR joint -"blog joint" -"roll up" -sushi -mccain -obama -"this joint"
- i smoked pot -"don't" -"dont" -"pot-luck" -"hickory" -"never" -"democrats" -"republicans" -"obama" -"not"
- i smoke pot -"don't" -"dont" -"pot-luck" -"hickory" -"never" -"democrats" -"republicans" -"obama" -"not"
- vicodin -"pain" -"hurts" -"ibuprofen" -"cramps" -"headache" -"tooth" -"dentist" -"jaw" -"lip" -"arthritis"

Intimite Sexual Details

- i "had sex" -"never had sex"
- "fuck her" -"fuck her up" -"the fuck" -"palin"
- "i hooked up with" -"dream"
- condom broke -"the condom broke"

Work Relate

- "hate my job"
- "hate my boss"
- twittering OR twitter "at work" OR "from work"
- "do nothing" OR "done nothing" "at work"

C Raw data

```
"my fake ID" : 5
# of restricted tweets: 0
web:3
3rd party:1
txt:1
"got hammered" OR "get hammered" : 100
# of restricted tweets: 1
web:52
mobile web:4
3rd party:40
txt:3
shitfaced : 45
# of restricted tweets: 1
web:24
3rd party:10
txt:6
```

```
mobile web:4
I smoked high -"high school" -"junior high" -"high horse" -"High Fidelity" : 5
# of restricted tweets:
3rd party:2
web:2
txt:1
rolled OR roll OR rolling blunt OR joint -"blog joint" -"roll up" -sushi -mccain
-obama -"this joint" : 23
# of restricted tweets: 0
web:10
3rd party:9
txt:4
i smoke pot -"don't" -"dont" -"democrats" -"republicans" -"obama" -"not" : 24
# of restricted tweets: 0
web:10
mobile web:1
3rd party:8
txt:5
i smoked pot -"don't" -"dont" -"pot-luck" -"hickory" -"never" -"democrats"
-"republicans" -"obama" -"not" : 6
# of restricted tweets:
3rd party:4
txt:2
vicodin -"pain" -"hurts" -"ibuprofen" -"cramps" -"headache" -"tooth" -"dentist"
-"jaw" -"lip" -"arthritis" : 282
# of restricted tweets: 11
web:129
mobile web:10
3rd party:92
txt:40
i "had sex" -"never had sex" : 57
# of restricted tweets: 1
web:22
mobile web:4
3rd party:17
txt:13
"fuck her" -"fuck her up" -"the fuck" -"palin" : 33
# of restricted tweets: 2
web:15
mobile web:1
3rd party:12
txt:3
"i hooked up with" -"dream" : 6
# of restricted tweets: 0
3rd party:3
```

```
web:2
txt:1
condom broke -"the condom broke" : 2
# of restricted tweets:
3rd party:1
web:1
"hate my job" : 185
# of restricted tweets: 0
web:71
3rd party:44
txt:58
mobile web:12
"hate my boss" : 8
# of restricted tweets: 0
web:4
3rd party:2
txt:2
twittering OR twitter "at work" OR "from work" : 200
# of restricted tweets: 4
web:129
mobile web:5
3rd party:56
txt:6
"do nothing" OR "done nothing" "at work" : 7
# of restricted tweets: 0
web:6
3rd party:1

Totals:
web:480
3rd party:302
txt:186
```

[8] [2] [3] [4] [9]