Poster: Towards Understanding Health Information Leakage through Social Networks

Qinchen Gu
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: qgu7@gatech.edu

Shouling Ji
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: sji@gatech.edu

Raheem Beyah
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332–0250
Email: raheem.beyah@ece.gatech.edu

***Abstract*—In this poster, we measure the health-related information leakage through social networks. Leveraging 103,862 Twitter users' profiles and 233,080 WebMD users' online posts as auxiliary information, we demonstrate that a significant amount of Twitter users' health-related information can be correctly inferred. In the experiment, we use six different machine learning algorithms and compare their performance in inferring Twitter users' health information.**

I. Introduction

Social networks have become an indispensable part of our life nowadays. It is estimated that there are 1.55 billion active users on Facebook as of January 2016 [1]. As people share more enriched content on social networks, they may also have unintentionally leaked private information pertinent to their own lives. Health information, as demonstrated in this poster, is one of the sensitive information that most people try not to disclose to the unwanted public. In this poster, we study health information leakage through social media, leveraging Twitter as a representative among various social networks. We collect data from WebMD, which is one of the largest US online publisher of health-related information and also provides a public forum for registered users to share and discuss on different health topics. We also collect $103,862$ Twitter users' data and infer their health information on the individual account level using six different machine learning techniques. We compare our results with the ground truth obtained from the WebMD dataset and show that it is possible to infer health information from daily Twitter usage, and thus causing information leakage.

A. Data Collection

We collect all data used in this study from publicly available online resources. Two major datasets are collected from WebMD and Twitter, respectively.

*1) **WebMD**:* We collect WebMD users' data from its 395 distinctive public forums, each with a specific health topic, e.g., *acne*. The dataset contains a total of $233,080$ users along with their $1,795,429$ posts. Each post belongs to one of the 395 public forums and is therefore tagged with the specific health topic of each forum.

*2) **Twitter**:* The Twitter dataset is collected through name matching between our collected WebMD users and Twitter. Each user in WebMD has at most one correspondent in Twitter which shares the same case insensitive username (or screen name, as Twitter refers to it). Each Twitter user contains its user profile, timeline (collection of most recent Tweets posted by the user), follower IDs, following IDs, favorite IDs (IDs of Tweets liked by the user) and list IDs (IDs of lists the user subscribes to). For the $233,080$ WebMD usernames a total of $103,862$ matching Twitter users are collected, among which $63,467$ users have not turned on "Protect my Tweets" option, i.e., their Tweets are visible by any users including those who do not have following/follower relationship with the user.

B. Ground Truth

We treat our collected name-matching WebMD and Twitter users as ground truth. Note that neither WebMD nor Twitter allows duplicate usernames among their own users, therefore no pair of correspondents in the ground truth overlap with each other. The $103,862$ corresponding pairs are not treated with the same level of confidence as will be discussed in Section II-B.

II. Inferring Health Information

The main idea behind the health information inferring technique is the assumption that each Twitter user can be classified into the health topic class it associates with (including disease-related such as *Alzheimer's*, and non-disease-related such as *healthy life style*), based on features that can be extracted from publicly available information of the user. We first transform each user's data fields into a sparse feature vector, then use K-fold cross validation machine learning among Twitter users to test the accuracy of our prediction.

A. Feature Vector

Table. I shows features extracted from each Twitter user's data fields. Due to the enormous amount of features, the feature matrix cannot be represented regular array in the memory. However, the feature vector is highly sparse, e.g., the $5,041,590$ *followings* consists of all IDs of users followed by the $63,467$ Twitter users in our dataset, while the average

TABLE I
FEATURE VECTORS

Feature	Length	Comment
favorites	5,603,009	Bernoulli vector spanning IDs of all Tweets marked as "favorite" by users in our dataset
followings	5,041,590	Bernoulli vector spanning IDs of all Twitter users followed by users in our dataset
hashtag	1,234,281	Real value vector spanning all hashtags mentioned by users in our dataset
lists	19,313	Bernoulli vector spanning IDs of all lists joined by users in our dataset
post time	9	Mean, standard deviation, skew and kurtosis of Tweet posting time (time of day and day of year) and frequency of posting Tweet of an individual user
TF-IDF	varies	TF-IDF feature vector extracted from Tweets of users in our dataset

number of *followings* of each user is 155.7. Thus sparse matrix representation may be used to significantly lower memory complexity.

B. Ranking Users

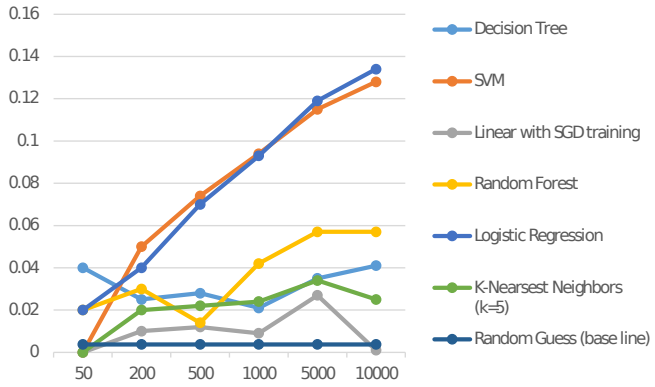
The 63,467 Twiitter users are ranked by the uniqueness of their screen name [2]. As suggested in [2], higher uniqueness of a username implies higher probability that the username is shared by the same real person. We therefore create a name uniqueness ranking for the 63,467 users.

C. Classification

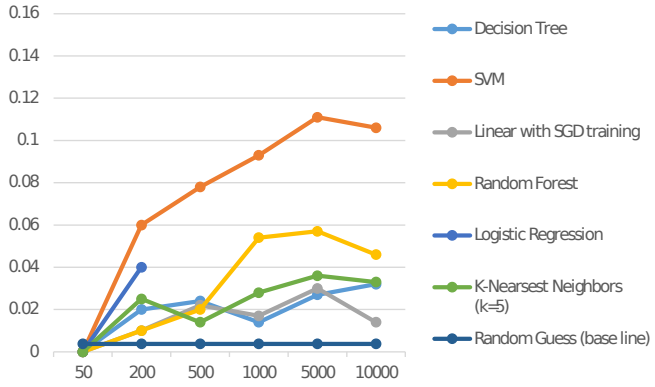
After all features have been extracted, the rest of the problem becomes a classification task. The goal is to correctly classify 63,467 "unprotected" Twitter users into 269 health topic categories. Note that the number of categories have decreased from 395 because not all WebMD users can find a matching Twitter user, and we also exclude "protected" Twitter users. The base line is randomly guessing topic category for each user, which leads to $1/269 \approx 0.0037$ correct rate.

For TF-IDF feature listed in Table. I, we vary the feature length by setting vocabulary parameter to none (all words in the Tweets are vectorized) or a pre-set medical dictionary (only words appeared in medical dictionary are vectorized). We vary the sample size by taking the top 50, 200, 500, 1000, 5000 and 10000 users in the uniqueness rank. We then use 10-fold validation to split the sample users into training and test set with each sample size. And for each setting, we employ six machine learning models to do the classification task and test their performance respectively. The result of correct rate is shown in Fig. 1.

As can be seen from the result, when sample size is between 50 and 5,000, both vocabulary setting has similar results, while experiment with medical vocabulary setting performs better than the one without vocabulary. However in general, all machine learning classification performs significantly better than the base line, and shows an increase in correct rate as



(a) Medical vocabulary



(b) No vocabulary

Fig. 1. Experiment result.

the sample size increases, since larger training set leads to better fit. Notice that the correct rate for 10,000 users in the second experiment setting drops below the extrapolated point, probably due to the lower ranked users have less probability being the same real user, which lead to lower confidence in ground truth.

III. Conclusion and Future Work

In summary, we demonstrate the plausibility of health-related information leakage through social networks, and show the effectiveness with six machine learning models under various experiment settings.

Future work includes improving the prediction correct rate by filtering out users with certain attributes. More auxiliary information may be useful in providing more confident ground truth.

References

[1] Statista. (2016) Leading global social networks 2016. [Online]. Available: http://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[2] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, *Privacy Enhancing Technologies: 11th International Symposium, PETS 2011, Waterloo, ON, Canada, July 27-29, 2011. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, ch. How Unique and Traceable Are Usernames?, pp. 1–17.