FamilyID: A Hybrid Approach to Identify Family Information from Microblogs

Jamuna Gopal, Shu Huang, Bo Luo

FamilyID: A Hybrid Approach to Identify Family Information from Microblogs

Jamuna Gopal[1], Shu Huang[2]*, and Bo Luo[3]*

[1] IBM, San Jose, CA, USA
[2] Microsoft, Seattle, WA, USA
[3] Department of EECS, University of Kansas, Lawrence, KS, USA

Abstract. With the growing popularity of social networks, extremely large amount of users routinely post messages about their daily life to online social networking services. In particular, we have observed that family related information, including some very sensitive information, are freely available and easily extracted from Twitter. In this paper, we present a hybrid information retrieval mechanism, namely FamilyID, to identify and extract family related information of a user from his/her microblogs (tweets). The proposed model takes into account part-of-speech tagging, pattern matching, lexical similarity, and semantic similarity of the tweets. Experiment results show that FamilyID provides both high precision and recall. We expect the project to serve as a warning to users that they may have accidentally revealed too much personal/family information to the public. It could also help microblog users to evaluate the amount of information that they have already revealed.

1 Introduction

With the growing popularity of online social networks, the data that is publicly available has increased by numerous folds. This data includes personal, employment, education, relationship, and family-related information. Figure 1 shows a microblog example – a tweet message that was broadcasted to the public, and effectively reveals his mother's Twitter ID, birthdate and last name.

Numerous commercial products or research projects have been developed to discover user information from online social networking data. Such information is used to improve the accuracy of advertisement delivery, to make sensible suggestions to users, and to predict events or trends. Moreover, the media industry (radio, movie, television) now highly depends on feedback from public OSN data for market study, user preference analysis, hot topic identification, etc. Although such products/projects may benefit both OSN providers and end users, they pose significant privacy threats to all users, while many of them are

* Corresponding authors: Shu Huang [shuang@microsoft.com]; Bo Luo [bluo@ku.edu].
This work was partially supported by NSF CNS-1422206, NSF IIS-1513324, NSF OIA-1308762, and University of Kansas GRF-2301876.

Fig. 1. A Tweet message that reveals sensitive family-related information.

unaware of such threats. An online stalker with limited hacking capability but ample time can effectively figure out lots of details about a targeted user with this publicly available data. For instance, message with birthday or anniversary wishes exposes users' age, date of birth and family information.

Extracting family-related information from Twitter is challenging: (1) it is cumbersome to manually identify such posts, as we have discovered that less than 1% of the tweets are family-related; and (2) although it is possible to develop an automated mechanism to identify family-related tweets, the task is nontrivial, due to the size of data, the use of short text and informal language, and large amount of synonyms. In this paper, we present FamilyID, a multi-phase approach that automatically identifies family-related information from publicly available Twitter data. Our algorithm considers multiple features of tweets, including part-of-speech tagging, term distribution similarity, and semantic similarity. Experimental results show that FamilyID produces good accuracy.

The key contributions of this paper are: (1) We make the first attempt to automatically identify family-related microblogs – they usually disclose sensitive personal information, and they are the primary targets for both adversaries and defenders. (2) The proposed mechanisms exploit multiple lexical and semantic features, with a good balance of efficiency and precision. Our approach could handle large amount of data and provide relatively high accuracy.

2 Related Work

Private information disclosure. People may publicize private information for social advantages [7]. Users' privacy settings violate their sharing intentions, and they are unable or unwilling to fix the errors [11]. [13] explores three types of private information disclosed in the textual content of tweets. Impersonation attacks are proposed in [2] to steal private (friends-only) attributes.

Information aggregation attacks. Information aggregation attacks were introduced in [10, 8, 17]: significant amount of privacy is recovered when small pieces of information submitted by users are associated. [1] confirms that a significant amount of user profiles from multiple SNSs could be linked by email addresses.

Inference attacks. Hidden attributes are inferred from friends' attributes with a Bayesian network [5]. [4] developed a model to predict user's birth year (i.e., age). Unknown user attributes could be accurately inferred when as few as 20% of the users are known [14]. Friendship links and group membership information can be used to identify users [16] or infer sensitive hidden attributes [18].

Microblog Mining. Knowledge discovery in social networks is a hot research area. For instance, methods have been proposed to identify user attributes, such as

gender, age, location [12], location type [9], activities [6], personalities [15], etc. There are also proposals to make predictions based on information and activities in social networks, e.g., to predict stock rates based on user tweets [3].

3 The FamilyID Approach

3.1 Problem Definition and Solution Overview

The goal of this research is to identify family-related posts from microblogs. Due to the volume of the data, manually reading each tweet and classifying it is an almost impossible task. Formally, the objective of the research is: *For each tweet, efficiently and accurately identify whether it is related to one or more family members of the message owner (the user who posted the message).*

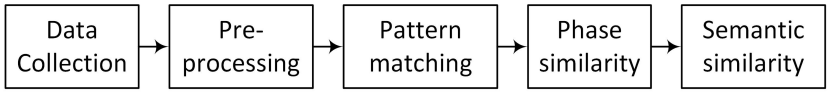

Fig. 2. Overview of the FamilyID Approach

As illustrated in Figure 2, we first use a customized crawler to collect user information and messages from Twitter. Each message is pre-processed to remove all the special characters and other unwanted contents, such as multimedia data (images, audio and video files). Each message from a user (denoted as the *owner* of the account/tweet) is processed through three steps: pattern matching, lexical (phrase) similarity measurement, and semantic similarity measurement. These steps are used to predict the likelihood of each tweet being family-related.

3.2 Data collection

Using the *twitter4j* API, we have collected 150 twitter users' information, including username, screen name, friends (follower and following) list, tweets and tweets time-stamp. Twitter does not have the concept of *friends*. Hence, we considered the intersection of the followers list and the following list as the friends list. We have randomly selected users with the following criteria: (1) Users with more than 1500 followers are omitted as they have higher chances of being celebrities. Tweets of celebrities are not used in this research, since they demonstrate significantly different styles and contents from tweets of regular users. (2) Users with fewer than 2000 tweets are not crawled. (3) Users with majority of tweets in foreign languages (anything other than English) are discarded.

3.3 Pre-Processing

Messages from Twitter are extremely noisy. We develop several heuristics to pre-process raw tweets: *(1) Term Expansion.* Twitter users like to use abbreviations

and very informal terms that do not exist in the dictionary. Certain steps in Figure 2 cannot process irregular words. Hence, we construct a table for Twitter term expansion for family-related terms (some examples are shown in Table 1). *(2) URL Truncation.* Tweets sometimes have URLs embedded in them. Since these URLs are not utilized in pattern matching, lexical similarity or semantic similarity assessments, we truncate all URLs. *(3). Stop Words.* FamilyID does not remove stop words, since words like "my", "our" are important in predicting family relationships. *(4) Special Characters.* All special characters other than the English words and numbers are truncated. Although we do not process numbers, we keep them for future use, e.g., to identify patterns related to year.

Table 1. Term Expansion Examples

Base word	Expanded word	Base word	Expanded word
mum	mother	sissy	sister
gf	girlfriend	bro	brother

3.4 Pattern Extraction and Matching

In Sections 3.4 to 3.6, we present a series of operations to identify family-related tweets. The design philosophy is to first employ computationally inexpensive methods to eliminate the majority of irrelevant tweets, and then refine the results with methods that are more effective but expensive.

Iterative Pattern Discovery. The first step in family-related tweet identification is to discover natural language patterns that are highly likely to mention family member(s). We first employ the Stanford NLP tagger for part-of-speech tagging on all crawled tweets. Next, we extract N-Gram histograms ($N = 2, 3, 4$) across the dataset to collect the common patterns containing family terms. Pattern discovery is performed in an iterative manner: for each discovered pattern, we attempt to relax it, and validate the relaxed pattern on the dataset.

Example 1: In our dataset, POS-tagged text snippet

```
my_PRP$ little_JJ sister_NN
```

has repeated 48 times, while text snippet

```
my_PRP$ little_JJ sister_NN @UserName_NN
```

has appeared 32 times. Therefore, we have extracted the following pattern:

```
_PRP$ _JJ _NN
```

Pattern Matching. Every POS-tagged tweet is matched against the seed patterns. With a matched pattern, the tweet has the potential to contain family-related information. Note that pattern matching is the first filter in the whole process, it leads to lot of noise outputs since many phrases could match one of our seed patterns. For instance, phrases such as "my dear dog", "my sweet neighbor" are matched to the `_PRP$ _JJ _NN` pattern, although they have nothing to do with family members.

3.5 Lexical Similarity Assessment

This phase finds if a pattern-matched tweet contains family-related words. We first create a seed tweet set covering all possible relationships and frequent non-relationship components from the patterns. We then employ the *UMBC ebiquity* text similarity system to calculate the lexical similarities for pairs of tweets. *Stanford WebBase Corpus* is used to find possible synonyms of the given words. Table 2 shows some examples of similarities computed in FamilyID.

Lexical similarity assessment effectively eliminates most of the noise from pattern matching. In particular, messages such as "my dog", "my neighbors", are effectively eliminated. However,, tweets such as "`my dear dog is my best companion`" pass the pattern matching phase ("my dear dog" matches `_PRP$ _JJ _NN`), and the lexical similarity assessment phase, due to the existence of terms "dear", "best", "companion". Since such tweets are clearly not family-related, we need another layer of semantic analysis to handle them.

Table 2. Lexical Similarity Examples

Text Compared	Score
Happy Birthday mother *vs.* happy birthday father	0.902
Happy anniversary sister *vs.* birthday wishes sister	0.749
grandma is the best *vs.* my life is boring	0.033
I love you the most father *vs.* Jesus is great	0.122

3.6 Semantic Similarity Assessment

Semantic similarity assessment, which is relatively slower, is the last step to remove irrelevant tweets that have passed through the first two filters.

To generate a seed set for this model, we first take a seed such as "my little sister", and ran the sliding window algorithm on it. This is a recurring model that matches patterns in windows' length of up to 5. It replaces each word in the seed, and finds substitutions for the word, as shown below:

```
my little sister
*** little sister      my *** sister      my little ***
my *** little sister      my little *** sister
```

To calculate semantic similarity, we employ the *UMBC GetStsSim API*. This API takes 2 text snippets and returns a value between 0 and 1 as a similarity measure. Every candidate tweet is compared with the seed tweets, to measure the pairwise semantic similarity. As shown in Table 3, similarity score of 0.75 or above indicates an almost perfect match, while similarity score of 0.6 or above indicates relatively similar texts. Tweets with the highest similarity scores higher than the threshold are finally labeled as family-related. As shown in the previous

example, tweet "`my dear dog is my best companion`" passes first two phases. When we evaluate its semantic similarity with the seed tweets in this phase, the highest similarity score is 0.33, which indicates that it is not similar with any of the seeds. In this way, this message is labeled as non-family-related.

Table 3. Semantic Similarity Examples

Compared Tweets	Scores
happy birthday mother *vs.* birthday wishes mother. you are the best	0.611
my sweet little sister *vs.* my handsome young brother	0.624
my sweet sister *vs.* my awesome dog	0.21
long day. i miss you my dear mother. Come back soon *vs.* feeling extremely tired. its a long day	0.38

4 Experimental results

Tweet Identification. First, we have performed *tweet identification* on the collected dataset (150 Twitter users, more than 450,000 tweets). On average, FamilyID has identified approximately 30 tweets from each user as family-related, as shown in Figure 3 (users are sorted by total number of tweets crawled). Less than 1% of the tweets are identified to be related to family members. These include a small amount of false positives (to be discussed later). With the numbers and by looking into the identified tweets, we have found that the results reflect our previous observations: (1) for most of the Twitter users, family-related tweets are very sparse. It is extremely time-consuming, if not impossible, to manually identify such tweets. (2) The identified family-related tweets almost always bring additional information about the family members, including the relationship, Twitter username, date of birth, age, interests, etc.

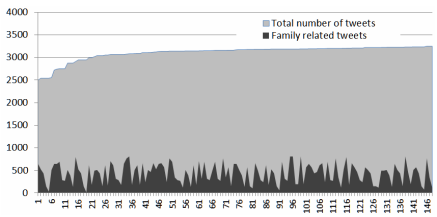

Fig. 3. Total number of tweets and family-related tweets for each user.

Comparing with Keyword-based Retrieval. To evaluate the effectiveness of FamilyID in reducing false-positives, we compare it with a keyword-based approach – identifying family-related tweets with keyword spotting. That is, when a pre-selected relationship keyword (e.g., "sister", "mother", the same as we used in Section 3.4) is found in the tweet, it is labeled as "family-related".

In order to manually examine the results, we perform keyword-based retrieval on 75 randomly selected users. We have evaluated 225,886 tweets. Keyword-based retrieval has found 6,121 tweets to be family-related, while FamilyID has identified 2301 of them as family-related. Note that due to the selection of the

Table 4. Examples of true positives and false positives.

True Positives (Family-related tweets)
I'm gon be an uncle *smiles* "@Bintah_Adam: I can't imagine my mum having another baby now"
Oh my god my sister is annoying
False positives
When one of my boys tells me he's in love
If your not my girl don't be jealous of my other girls

keywords, each tweet identified by keyword spotting is a candidate tweet in FamilyID. Therefore, more than 62% of the tweets containing family-related keywords are identified as *irrelevant* to family relationships through content-based analysis in FamilyID. We further manually look into such irrelevant tweets, and find that more than 90% of them are true negatives (not relevant to family members). This also indicates that the precision of the keyword spotting approach is low, since it has included large amount of non-family tweets.

Precision. We invite human evaluators to examine the tweets identified as family-related from 50 random users, to determine whether each tweet is truly related to family members. As the most important performance metric of FamilyID, the *precision* is defined as: $Precision = \frac{TP}{P}$, where TP indicates the number of true positives (tweets labeled as family-related that are determined to be family-related by human evaluators), and P indicates the number of positives (tweets labeled as family-related by FamilyID).

The evaluators have examined 1346 tweets that are identified as family-related by FamilyID. They have found 1110 tweets to be true positives. Therefore, the *precision of FamilyID is 83%*. Table 4 shows examples of true/false positives. The precision is high, especially consider the difficulty of the task. For some tweets, the human evaluator could hardly determine if they are family-related. For instance, for the message "`When one of my boys tells me he's in love`", the evaluator has referred to many other posts from the user, to find that she is a teacher and she is very likely talking about a student, instead of a child. However, the evaluator is less confident about the verdict.

Finally, we would like to point out that we have not evaluated the overall recall of FamilyID, for two reasons: (1) the size of the data set (450K tweets in total) makes it infeasible to manually examine all tweets; and (2) due to the heavy use of urban slang, abbreviations and short texts, it is even difficult for human evaluators to determine whether some of the tweets are family-related.

5 Conclusion

With the growing popularity of online social networks, large amounts of private information have been voluntarily posted to the Internet. From attackers' perspective, they could stalk a targeted user and attempt to extract such private information. However, manually identifying family-related tweets that are scattered in millions of microblog posts is very labor intensive. The FamilyID project

demonstrates the capabilities of an automated mechanism to identify family-related microblogs and extract family member information from the microblogs. By utilizing lexical and semantic features in a multi-phase approach, we are able to achieve high accuracy. Moreover, most of the identified tweets carry additional (very sensitive) information about the family, such as birthdates, hobbies, family events, etc. FamilyID could be used by social network users to self-assess the amount of family-related information that they have posted to the public. We also expect the project to serve as a warning to Twitter users who carelessly disclose too much information in online socialization.

References

1. M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing social networks for automated user profiling. In *Recent Advances in Intrusion Detection*, volume 6307, pages 422–441. 2010.
2. L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: automated identity theft attacks on social networks. In *WWW*, 2009.
3. J. Bollen, H. Mao, and X.-J. Zeng. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003, 2010.
4. R. Dey, C. Tang, K. Ross, and N. Saxena. Estimating Age Privacy Leakage in Online Social Networks. 2012.
5. J. He, W. W. Chu, and Z. Liu. Inferring privacy information from social networks. In *IEEE Intl. Conf. on Intelligence and Security Informatics*, pages 154–165, 2006.
6. S. Huang, M. Chen, B. Luo, and D. Lee. Predicting aggregate social activities using continuous-time stochastic process. In *Proceedings ACM Intl. Conf. on Information and knowledge management*, 2012.
7. B. A. Huberman, E. Adar, and L. R. Fine. Valuating privacy. *IEEE Security and Privacy*, 3(5):22–25, 2005.
8. F. Li, J. Y. Chen, X. Zou, and P. Liu. New privacy threats in healthcare informatics: When medical records join the web. In *BIOKDD*, 2010.
9. H. Liu, B. Luo, and D. Lee. Location type classification using tweet content. In *ICMLA*, volume 1, pages 232–237. IEEE, 2012.
10. B. Luo and D. Lee. On protecting private information in social networks: A proposal. In *M3SN Workshop*, 2009.
11. M. Madejski, M. Johnson, and S. M. Bellovin. The failure of online social network privacy settings. Technical Report CUCS-010-11, Columbia University, 2011.
12. J. Mahmud, J. Nichols, and C. Drews. Home location identification of twitter users. *ACM Transactions on Intelligent Systems and Technology*, 5(3), 2014.
13. H. Mao, X. Shuai, and A. Kapadia. Loose tweets: an analysis of privacy leaks on twitter. In *WPES*, 2011.
14. A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, 2010.
15. D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *IEEE PASSAT*, 2011.
16. G. Wondracek, T. Holz, E. Kirda, and C. Kruegel. A practical attack to de-anonymize social network users. In *IEEE Security & Privacy*, 2010.
17. Y. Yang, J. Lutes, F. Li, B. Luo, and P. Liu. Stalking online: on user privacy in social networks. In *ACM CODASPY*, 2012.
18. E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *WWW*, 2009.