Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Privacy Preference Inference via Collaborative Filtering

Taraneh Khazaei,* **Lu Xiao,*** † **Robert E. Mercer***
University of Western Ontario
London, ON, Canada

Atif Khan
InfoTrellis Inc.
Toronto, ON, Canada

Abstract

Studies of online social behaviour indicate that users often fail to specify privacy settings that match their privacy behaviour. This issue has caused a dilemma whether to use publicly available data for targeted advertisements and personalization. As a possible approach to manage this dilemma, we propose a collaborative filtering method that exploits homophily to build a probabilistic model. Such a model can indicate the likelihood that a given public profile is meant to be private. Here, we provide the results of an analysis of a set of observable variables to be used in a neighbourhood-based manner. In addition, we establish a social graph augmented with privacy information. Users in the graph are then transformed into a set of latent features, uncovering informative factors to infer privacy preferences.

Introduction

Providing personalized content can mutually benefit businesses and customers. Effective user profiling, however, hinges on collecting large amounts of data from users. Hence, privacy is the cornerstone of any personalization activity that, if disregarded, will influence gratifications derived from the personalization and can lead to irrecoverable trust issues. This issue may be especially problematic in the context of social networks, wherein massive numbers of daily social interactions are taking place nowadays.

Even though users are enabled to protect their data by using privacy controls, such privacy decisions are often complex, requiring careful examination of the trade-offs between the potential social gain and possible privacy risks. Therefore, many users avoid the hassle of privacy configuration and follow the default settings (Strater and Lipford 2008). However, they are normally unaware that the default setting is often open and permissive. Even among the ones that make the effort to manage their privacy, many are still unaware of the implications of their decisions (Liu et al. 2011).

Although limited in quantity, earlier research on privacy behaviour has had modest success in predicting user's privacy preferences by relying on their social footprints (Khazaei et al. 2016a; Dong, Jin, and Knijnenburg 2015). Here, we propose a Collaborative Filtering (CF) approach that combines neighbourhood-based techniques with a latent factor model inferred from the social graph of users. Our privacy prediction approach is novel and aims to detect privacy-concerned users with publicly available profiles. The contributions of this paper are as follows: 1) adapting a hybrid CF method to infer social privacy preferences, 2) exploring the usefulness of profile attributes in privacy preference detection, 3) establishing and analyzing the properties of a privacy-enhanced social graph, and 4) discovering a set of latent factors related to privacy attributes.

Our study is conducted on Twitter, wherein privacy control follows a binary specification. In Twitter, users can either follow the default *public* setting, which indicates that their tweets and contacts are publicly available, or they can change the setting to *protected*, which makes their tweets and contacts accessible only by their approved followers.

Related Work

CF methods have shown great promise in the development of recommendation systems, where unknown preferences of users are identified using known preferences of other users. Such approaches can be particularly valuable in the context of social media due to the existence of additional social relations. However, limited attempts have been made to adapt CF techniques for the prediction of privacy preferences.

For instance, Squicciarini et al. (Squicciarini et al. 2014) first form social circles based on users' characteristics (e.g., gender and hobbies). When a new object is uploaded by the focal user, the system then seeks the social circles that are most likely to deal with the object in a similar way as the user. Then the privacy policies used by the selected circle are the basis for predicting the policy for the added object.

In (Shehab and Touati 2012), active learning and the properties of the social graph are first used to detect a set of informative contacts to be labeled as training samples. In the labeling process, the user specifies whether he/she is willing to share a specific item with the selected contact. Then labels are propagated from labeled instances to unlabeled ones in the graph. This propagation is guided by the user similarity metric that is computed based on contacts' profile information, along with their network and community metrics.

CF is also followed in (Ghazinour, Matwin, and Sokolova 2013), where a set of profile features, users' interests, and

*Department of Computer Science
†Faculty of Information & Media Studies

privacy configurations are used to find a set of users similar to the focal user. In their work, users are first characterized according to their privacy preference as privacy fundamentalist, privacy pragmatist, or privacy unconcerned. Users are assigned to these categories based on the number of their public, customized, and private photo albums. Then K-nearest neighbour algorithm is used to determine which privacy categorization the focal user belongs to.

CF-based techniques to privacy prediction mainly used neighbourhood-based techniques, where privacy observations from neighbouring or similar users are the basis of the prediction. However, latent factor models, which are the current state-of-the-art for CF (Agarwal and Chen 2009), are yet to be explored. Earlier work also lacks the study of hybrid techniques that have shown significantly better results compared to pure neighbourhood-based and pure latent factor models in other domains (Koren 2008).

Neighbourhood-based Latent Factor Model

Two popular techniques to CF are neighbourhood-based methods and latent factor models. In neighbourhood-based methods, observations from neighbouring and/or similar users is used to detect attributes and preferences of the focal user. Detecting privacy preferences using such user-oriented techniques introduces unique challenges due to the lack of observable information associated with private accounts. In Twitter, a limited number of profile attributes are visible for both *public* and *protected* accounts. Hence, such profile attributes can be used to measure user similarity. The privacy preference of the focal *public* user i can then be determined using the following:

$$y_i = \frac{\Sigma_{i' \in \Omega_i} \omega_{ii'} x_{i'}}{\Sigma_{i' \in \Omega_i} \omega_{ii'}}$$

where $\omega_{ii'}$ measures the similarity between user i and its neighbours $i' \in \Omega_i$. In addition, $x_{i'} \in \{\varepsilon, 1\}$ indicates the actual privacy setting of the neighbour:

$$x_{i'} = \begin{cases} \varepsilon & \text{if } i' \text{ is } \textit{public} \\ 1 & \text{if } i' \text{ is } \textit{protected} \end{cases}$$

Based on this formula, a larger value of y_i indicates a higher level of privacy concern for user i. To gain insight into the potentials and limitations of Twitter profile attributes for our task, we carried out a set of experiments. These experiments, which are briefly presented in the next section, suggest the value of profile attributes in the inference of privacy preferences and indicate their potential for similarity measurement in a neighbourhood-based approach.

Meanwhile, latent factor models have gained tremendous success in the context of recommender systems. Such methods aim to discover a set of informative latent factors regarding users and later use such attributes to infer preferences. We can calculate the likelihood that the *public* user i is privacy-concerned using the following:

$$p(y_i) = p(y_i|\theta_1, ..., \theta_k) = \Pi_{j=1,..,k} p(y_i|\theta_j)$$

where $u_i = \theta_1, ..., \theta_k$ and $p(y_i|\theta_j)$ indicates the probability of user i being privacy-concerned if user i is associated with the latent factor θ_j. Later in this document, we propose a technique to discover such latent variables from a social graph of users. In this graph, users are first transformed into a set of latent variables. The probability of a latent attribute being associated with private people can then be calculated based on the number and/or the ratio of its *protected* neighbours. Finally, these two approaches can be merged in a single model to effectively capture privacy preferences (Koren 2008).

Profile Attributes for Privacy Prediction

To analyze the relations of profile attributes and privacy behaviour, we first built a directory of Twitter users by collecting the followers of several famous Twitter accounts (e.g., "Facebook", "Katy Perry", "Obama"). For each account, we then calculated the percentage of the *protected* followers to the total number of followers. The results indicate that the percentage is considerably higher for "CNN Breaking News"(11%) compared to the other follower sets (between 5% to 7%) and the average percentage in Twitter (4.8%). We thus selected this set for our study since the privacy attitude-behaviour dichotomy seems to be minimized.

We then analyzed this user set, which includes a balanced set of almost 1M users, to gain insight into the potential differences in how profile attributes of *public* and *protected* accounts are configured. We focused our analysis on a set of profile features that are readily available from Twitter[1] accounts, along with additional features developed based on the existing profile attributes. For instance, we used a directory of English names to analyze if the declared name includes an actual person name. In addition, we examined linguistic attributes of profile *descriptions* using LIWC[2] and keyword frequency analysis.

As a result, a feature set of size 27 is developed, a summary of which is provided in Table . The first column in the table shows the Twitter API features. The second column presents the linguistic attributes extracted from profile *descriptions*. In addition to a set of LIWC categories[2], this list includes four keyword-based features that indicate the presence of the corresponding keyword in the *description*. The differences between *protected* and *public* accounts are statistically significant across all these features (as determined by chi-square or t-test results depending on the feature type). For four of these features, the differences are practically significant as well (as determined by Cramer's V or Cohen's d depending on the feature type). These four features are marked by asterisk in the table. Hence, profile attributes are distinctive across *protected* and *public* users and proved to be of value for our task of privacy preference inference in a neighbourhood-based approach. Utilizing this feature set in a regression-based supervised algorithm resulted in an F-score of 0.72, outperforming a random baseline by over 20%. The details of the feature set and machine learning experiments can be found in (Khazaei et al. 2016b)

[1] https://dev.twitter.com/overview/api/users

[2] http://liwc.wpengine.com/

Twitter Attributes	Linguistic Attributes
Tweet Count*	Six Letter Words
Friend Count	Function Words
Favirote Count	Clout
List Count	Emotional Tone
Actual Name Count	Authentic
Is Geo-Enabled*	Analytical Thinking
Has Actual Name	Affect Words
Username Has Name	Social Processes
Has URL	Cognitive Processes
Has Location*	Relativity
Has Default Image	Drivers and Needs
Has Default Profile*	"follow"
	"business"
	"smile"
	"@username"

Table 1: Profile features to detect *protected* accounts.

Privacy Graph and Latent Attributes

Graph Construction and Properties

Our approach is intended to exploit homophily (McPherson, Smith-Lovin, and Cook 2001) and is based on the fact that people of similar interests tend to connect with each other. Therefore, instead of using the asymmetric follow or friend relation in Twitter, we built an undirected mutual graph of users that only includes the edges that are reciprocated. Reciprocated relations are expected to indicate a stronger relationship between the two users, and they distinguish the social network section of the Twitter sphere from its information network (Myers et al. 2014).

Starting from a random *public* user, we iteratively built a mutual graph of users in a Breadth First Search (BFS) manner. For each *public* user, we first counted the number of *protected* mutual neighbours as well as the ratio of *protected* to all mutual neighbours. This user is then annotated with these metrics and is added to the graph. We then check if the new node has a reciprocated relationship with any existing node in the graph and add the corresponding edges. This process is repeated with a new *public* user pulled from the BFS queue. Users with less than 10 tweets or less than 30 followers/friends are considered inactive and thus are not added to the graph. In addition, *verified* users and users with more that 1K followers/friends are not included since they often represent brands and celebrities and are not from the general public. We collected the total of 3K *public* nodes that are annotated based on their privacy ratio metric. Figure shows a snapshot of a small portion of the graph visualized using a force-directed layout, wherein the privacy ratio metric is mapped to the node size. In this dataset, each Twitter account is mutually connected to an average of 77 contacts. Among these neighbours, an average of 69 are *public* and 8 are *protected*. Figure shows the distribution of *protected* neighbours, where mean and median are marked by a solid and a dotted line, respectively. As can be seen, the distribution is skewed to the right, indicating that despite the smaller number of users with a large number of *protected* neighbours, these numbers are considerably large so that the mean is dragged to the right.

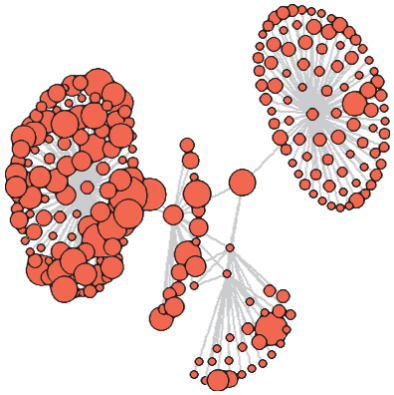

Figure 1: A snapshot of the privacy graph.

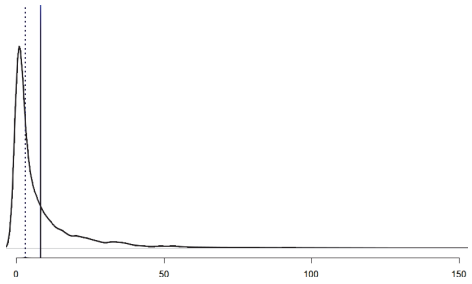

Figure 2: The distribution of *protected* contacts.

To ensure that homophily applies in the context of privacy, we calculated the correlation between the privacy ratio of each node and the average privacy ratio of the neighbours. That analysis of about 700 users with at least 10 mutual contacts in the graph showed a strong positive correlation between the two variables (Pearson correlation coefficient $r = 0.88$). This result indicates that users' privacy behaviour is either influenced by their close social contacts or individuals with similar privacy behaviour tend to cluster together in social networks. In either case, this finding implies the great potential of CF for privacy preference prediction.

Latent Attribute Detection

To discover a set of latent attributes that are of interest to privacy-concerned users, we transformed each node into a set of attributes. For each attribute node, we then calculated the number and the ratio of *protected* neighbours. The resulting graph is expected to enable the computation of $p(y_i|\theta_j)$, which indicates the likelihood of user i being private in case he/she is labeled with attribute θ_j. For instance, the privacy ratio of each latent factor may serve as such a probability value. We conducted experiments using unigrams and hashtags extracted from user tweets as latent variables. These attributes are selected from the 500 most recent tweets published by each user. Table 2 displays the top 10 hashtags and unigrams that are associated with at least 10% (n = 30) of the users. These features are extracted based on the privacy ratio metric, which is also provided in the table.

The top hashtag in table is "#neverforget", which is a

Hashtags		Unigrams	
#neverforget	19.89	seperate	32.07
#USA	19.57	reward	31.54
#respect	19.25	dull	31.1
#cantwait	18.49	deepest	29.55
#FML	17.62	unforgettable	28.66
#YOLO	16.61	activities	27.98
#instantfollow	16.52	lyric	27.55
#aquarius	16.46	circumstances	25.96
#winning	15.83	somethings	25.83
#soundcloud	15.52	forgiving	25.62

Table 2: Latent factors extracted from the privacy graph.

commemorative political slogan that encourages remembrance for national and international tragedies. The high privacy ratio of this hashtag, along with the top-ranked hashtags of "#USA" and "#respect" can indicate that communities interested in political topics tend to be more private. This finding is interesting and can be considered inline with a high percentage of *protected* CNN followers. However, given that the graph was collected in a particular timeframe, these results can be time-specific. Another time-oriented hashtag apparent in the list is "#aquarius". Collecting data over time and clustering hashtags into high-level topics can shed light on the potential relations between users' interests in politics and zodiac signs and their privacy preference.

We also explored the collected tweets to understand the context in which the top unigram "separate" is used. The majority of these tweets are related to user's love lives and relationships. Interestingly, "#relationshipgoals" and "#lastrelationshiptaughtme" have a high privacy ratio in our list. Therefore, one may conclude that users who share about their relationships in social media are more likely to be privacy-concerned. Again, grouping hashtags into general concept and topics can provide more accurate results.

Surprisingly, in the ranked list of hashtags, "#personal" was placed last with the lowest privacy ratio of 0.25. Although this finding may seem counterintuitive, our examination of the tweets showed that this hashtag is mainly used in conversations, wherein users are asked to provide some information, but they refuse to do so by replying a tweet that contains "#personal". Hence, it is likely that they are privacy-aware people that deliberately use the *public* setting, which is inline with what the latent factor model probability indicates. Overall, despite the limited size of the graph, the results are sensible and can reveal interesting information about private neighbourhoods in Twitter.

Conclusion

To predict one's privacy preference, we proposed a CF method that utilizes both neighbourhood-based and latent factor models. We analyzed the benefits of using profile attributes to measure user similarity and the use of hashtags and unigrams as latent features. The data collection process is currently ongoing, and we will run similar studies on multiple graphs built using different seed users. We also will examine a variety of other user attributes for the latent factor model. Robust evaluation methods will be developed to verify the usefulness of the approach. While we are focused on a simplied form of privacy here (i.e., binary specification), attempts will be made to analyze complex forms and strategies of privacy protection.

Acknowledgments

This project is supported through the MITACS Accelerate program of Canada. We also thank our industry partner, InfoTrellis, for their support.

References

Agarwal, D., and Chen, B.-C. 2009. Regression-based latent factor models. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 19–28.

Dong, C.; Jin, H.; and Knijnenburg, B. 2015. Predicting privacy behavior on online social networks. In *Proceedings of the AAAI Conference on Web and Social Media*.

Ghazinour, K.; Matwin, S.; and Sokolova, M. 2013. Monitoring and recommending privacy settings in social networks. In *Proceedings of the Joint EDBT/ICDT Workshops*, 164–168.

Khazaei, T.; Xiao, L.; Mercer, R.; and Khan, A. 2016a. Detecting privacy preferences from online social footprint: A literature Review. In *Proceedings of the iConference*.

Khazaei, T.; Xiao, L.; Mercer, R.; and Khan, A. 2016b. Privacy Behaviour and Profile Configuration in Twitter. In *Proceedings of the Conference on World Wide Web - Companion Volume*.

Koren, Y. 2008. Factorization meets the neighborhood: A multifaceted collaborative filtering model. In *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 426–434.

Liu, Y.; Gummadi, K. P.; Krishnamurthy, B.; and Mislove, A. 2011. Analyzing Facebook privacy settings: User expectations vs. reality. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement Conference*, 61–70.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.

Myers, S. A.; Sharma, A.; Gupta, P.; and Lin, J. 2014. Information network or social network? The structure of the twitter follow graph. In *Proceedings of the International Conference on World Wide Web*, 493–498.

Shehab, M., and Touati, H. 2012. Semi-supervised policy recommendation for online social networks. In *Proceedings of the Conference on Advances in Social Networks Analysis and Mining*, 360–367.

Squicciarini, A.; Karumanchi, S.; Lin, D.; and DeSisto, N. 2014. Identifying hidden social circles for advanced privacy configuration. *Computers & Security* 41:40 – 51.

Strater, K., and Lipford, H. R. 2008. Strategies and struggles with privacy in an online social networking community. In *Proceedings of the British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction*.