Your Love is Public Now: Questioning the use of Personal Information in Authentication

Payas Gupta, Swapna Gottipati, Jing Jiang and Debin Gao
{payas.gupta.2008, swapnag.2010, jingjiang, dbgao}@smu.edu.sg
School of Information Systems, Singapore Management University, Singapore

ABSTRACT

Most social networking platforms protect user's private information by limiting access to it to a small group of members, typically friends of the user, while allowing (virtually) everyone's access to the user's public data. In this paper, we exploit public data available on Facebook to infer users' undisclosed interests on their profile pages. In particular, we infer their undisclosed interests from the public data fetched using Graph APIs provided by Facebook. We demonstrate that simply liking a Facebook page does not corroborate that the user is interested in the page. Instead, we perform sentiment-oriented mining on various attributes of a Facebook page to determine the user's real interests. Our experiments conducted on over 34,000 public pages collected from Facebook and data from volunteers show that our inference technique can infer interests that are often hidden by users on their personal profile with moderate accuracy. We are able to disclose 22 interests of a user and find more than 80,097 users with at least 2 interests. We also show how this inferred information can be used to break a preference based backup authentication system.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Authentication; H.2.4 [**Systems**]: Textual databases

General Terms

Security

Keywords

Facebook; preference based authentication; Graph API; semantic analysis

1. INTRODUCTION

With the rise of online social networks (OSNs), more and more personal information of users is available on the web. It also forms a fertile ground for a variety of research efforts. The information shared on OSNs can be classified into two broad categories, *private* (shared with a limited set of users) and *public* (shared with the whole world). Prior research shows that the information shared with a limited set of users can leak undisclosed privacy attributes, e.g., users' interests and even sexual orientation [28, 21]. Authors of [4] crawled Facebook users' personal profiles to infer users' undisclosed interests. However, getting access to such information (without explicit permission) that is shared with a limited set of users is non-trivial as it is not available to public via any APIs. Moreover, OSNs are adopting ways to restrict crawling unless explicit permission is granted [25, 31].

In comparison to the information available only to a limited set of users, public information is readily available. In many cases APIs are provided by OSNs for anyone to efficiently download such public data. Facebook, for example, has made all the fan pages public by default. Access to the data of these pages can be conveniently obtained through Graph APIs [8]. It is generally believed that these public pages hardly contain any secret information, and mining some useful information from them is not easy mainly due to the large amount of noise contained in the heterogeneous pages, and the huge amount of unstructured data involved. For example, Facebook has little control on the titles and descriptions of fan pages; posts from users may contain text and multimedia content; users use a lot of short sentences and slang (e.g., "LOL", "LMAO", etc); off-topic discussions go on frequently (e.g., on a "Jazz" page, we found users discussing the latest soccer game). All this adds to the noise in public data on OSNs. Moreover, almost 15 percent of user-submitted content on large Facebook fan pages is spam [10]. Such noise and the huge amount of unstructured data to be processed usually makes mining interesting information not practical.

In this paper, we show that this belief might not be true in certain aspect. In particular, we show how we use *publicly available* data from Facebook to infer users' interests that are usually only on their personal profile pages. We make use of the graph APIs provided by Facebook to obtain public fan pages [7]. As these pages are public irrespective of the users' privacy settings, an attacker can grab the unique profile IDs of those who have interacted with the page. We show that by aggregating different interests of the users found across different pages, one could build users' interests profiles from the public data without gaining access to the personal profile pages of any of the users. This collective information can be used in many ways including targeted spamming, show-

ASIA CCS'13, May 8–10, 2013, Hangzhou, China.

ing ads without the consent of users, or even breaking into specific authentication systems, in our case preference based authentication [17]. A system based on user preferences was proposed in order to reduce the vulnerability to data-mining and maximize the success rate of legitimate reset attempts. The viability of such an approach is supported by findings in psychology, showing that personal preferences remain stable for a long period of time.

To demonstrate the security and privacy implication of this, we base our experiments on mining personal interests to break into Blue Moon^{TM} [15] introduced by RavenWhite as a backup authentication system to provide better security and usability. From the dataset in our experiments involving 1.1 million different user IDs from 34,000 Facebook public pages, we detected 80,097 (6.89%) users with two or more interests. Out of these 80,097 users, there are 66 who have been found with more than 8 interests, which is enough to break their corresponding Blue Moon accounts (if they have) with reasonable accuracy under certain assumptions. In one case, we were able to build a user profile with as many as 22 interests by mining the data we collected. We also present valuable lessons we learned in our experiments, among which the most notable one being that users' sentiment orientation might not be inclined towards the sentiment orientation of the page i.e. simply liking a page does not corroborate enough that the user is really interested in the page. Therefore, we performed sentiment mining to find out the actual sentiment orientation of the user.

In summary, this paper makes the following contributions.

- We use publicly available data on Facebook to infer users' interests and aggregate this information across different pages. This differs from prior research as we do not use user's personal data posted on their profile page (e.g., gender, current location, activities, interests, etc.).

- We find that liking a page does not corroborate a user's inclination towards a page or interest category. We performed an in-depth analysis (sentiment) using text mining to find the real sentiment orientation or polarity (like or dislike) of the user towards a page and an interest.

- We use Facebook's public Graph API [8] to obtain the public pages. Unlike crawling which is usually restricted in its usage by OSNs to a small number of partners, our method could be easily used by anyone with little restriction.

- We demonstrate the severe implication of interests mining by showing that interests inferred from the public data can be used to exploit a previously proposed preference based authentication system.

The rest of the paper is structured as follows. We provide background and related work in Section 2 where we show some of the important prior work to abuse OSNs. We explain our technique to mine user interests from Facebook public pages in Section 3, and report experimental results in Section 4. We then discuss the limitations and errors that could have occurred in our technique in Section 5 and conclude the paper in Section 6.

2. RELATED WORK AND BACKGROUND

In this section, we first discuss related work in obtaining a user's private information by abusing OSNs in general. After that, we discuss the more specific interests inference techniques in social networks.

2.1 Abusing OSN data

With the increasing popularity of OSNs, people start to find ways of abusing it, e.g., illegitimate use by spammers with ad deals. In this paper, we focus on the abuse in which a user's privacy attributes are inferred from information hosted on OSNs. In general, attackers could base their attacks on two types of data obtained in different ways.

One is to use restricted pages by crawling. Prior research shows that information on restricted pages (shared with a limited set of users) can leak undisclosed privacy attributes about the users [28, 21]. Existing techniques have demonstrated that private information can be crawled to obtain attributes like mother's maiden name, date of birth, hometown, first school attended to break into backup authentication mechanisms that are based on such privacy attributes [16]. Attackers can also correlate information from different OSNs to retrieve undisclosed attributes of the users [21]. Authors of [4] crawled users' personal profiles of Facebook to infer their undisclosed interests. [3] describes how an attacker could query popular social networks for registered e-mail addresses on a large scale and information from different social networks can be aggregated to launch sophisticated and targeted attacks.

An important limitation to using restricted pages by crawling is that most OSNs restrict crawling to a small number of partners only. That is, crawling restricted pages is not a technique available to general attackers.

The other type of data to use is public pages. As compared to crawling restricted pages on OSNs, anyone can use a legitimate channel (usually by using public APIs provided by OSNs) to gather public information. Although these public pages are more readily available for anyone to analyze, as pointed out in Section 1, it is generally believed that mining interesting private information from these public pages is difficult due to the noise in it and the huge amount of unstructured data to be analyzed. In this paper, we show that mining users' otherwise undisclosed interests from public pages on OSNs is, in fact, practical.

There are strong security and privacy implications to such abuse of OSN data because the private information mined could potentially be used to break existing authentication systems, typically those that use challenge questions as a backup to the main authentication mechanism. In Table 1, we highlight noticeable differences between this work and prior research. Previous work has shown that OSN data and public databases can be used to infer or guess sensitive information about users [28, 21]. A number of incidents, e.g., in 2008 the Republican vice presidential nominee Sarah Palin's email account was compromised by an attacker who guessed her personal authentication question (where did you meet your spouse?) [26], in 2009 a vandal successfully guessed a Twitter executive's password and leaked the company's internal documents [6], have shown the severe damage such attacks could have. Personal authentication questions are usually a weaker link in authentication systems [24]. Authors show that answers to predefined questions can be easily guessed or obtained from OSNs [24]. Instead of specific

#	Paper	Datasets source	Dataset type	Count	Inferred information	Collected information	Dataset gathering technique
1	Mislove et al. [21]	Rice University Facebook network	Personal profiles	6,156	Missing attribute on personal profile page	Attributes of others (friends etc.)	Crawling
		New Orleans Facebook network	Personal profiles	90,269			Crawling
2	Chaabane et al. [4]	Facebook	Personal profiles	104,000	Gender, Relationship status, Country-level location and Age	Interests	Crawling
		Volunteers from Facebook	Personal profiles	200			Voluntarily provided
3	Lindamood et al. [20]	LiveJournal	Personal profiles	66,766	Nodes in the social graph	Friendship relations	Crawling
4	Goga et al. [13]	Twitter	Personal profiles	93,839	Correlating missing attributes	Profile and privacy information	Friend-finder
		Flicker		59,476			
		Yelp		24,176			
5	Avello et al. [12]	Twitter	Personal Profiles	4.98M	Sex, age, political orientation, religious affiliation, race	Personal details and tweets	Crawling
			Tweets	27.9M			
6	Zheleva et al. [33]	Flicker	Personal Profiles	9,179	Location	Information on profile	Crawling
		Facebook	Personal Profiles	1,598	Gender, political orientation	Information on profile	Crawling
		Dogster	Dog Profiles	2,632	Breed category	Information on profile	Crawling
		BibSonomy	Personal Profiles	31,715	Spammer	Information on profile	N/A
7	*This paper*	Facebook	*Public pages*	34,000	Interests	Public page's data	*API*
			Personal profiles	1.1M			

Table 1: Comparison with the related work

attack incidents where one or two particular accounts are compromised, our work presented in this paper shows an attack to the authentication system and evaluates the extent to which thousands of users of such a system could be attacked.

2.2 Interests mining from OSN data

This paper focuses on personal interests mining because personal interest is one of the most popular choices used in challenge questions. Authors of [22] leverage on the friendship network to mine users' interests. [11] employs feature engineering to generate hand-crafted meta-descriptors as fingerprints for a user. However, such models alone may not derive the complete interest list of any user [32]. [19, 5] resort to collaborative filtering techniques to profile user interests by collaboratively uncovering user behaviors.

The "Like" function on OSN provides a more intuitive way of estimating user interests as compared to non-direct indicators such as user-service interactions. Clicking on the "Like"/"Dislike" button associated with an object usually indicates that (s)he is highly interested/disinterested in the object [32]. Recent approaches like LikeMiner [18] assumes that clicking the "like" button demonstrates the user's liking towards the object. In this paper, we show complications in using such an assumption on large datasets and propose solutions to it.

3. INTERESTS INFERENCE

Although users understand that public pages are for everyone to view and should not contain sensitive or private information, these pages nevertheless reveal what users do and what users think. Therefore, it is probably not difficult to be convinced that such pages still contain private information, probably indirectly and to a limited degree, e.g., by reflecting what users like and dislike. This paper is not to argue this, but to rather investigate how practical it is to mine interesting personal interests from the large amount of unstructured data on public pages that contain a lot of noise.

To do this, we first introduce the data source on which our analysis is performed, i.e., the public pages on Facebook (see Section 3.1). Section 3.2 presents our methodology to fetch information from these public pages. Finally, we present our methodology to infer users' interests (i.e. likes and dislikes) for different categories (e.g., music, cars, sports) in Section 3.3.

3.1 Facebook Page Layout

A Facebook Page is a public profile where users can talk (and comment/like) about a particular topic. As shown in Figure 1, it usually contains many attributes including title, page description, profile picture, wall posts, likes, etc. Any registered user can create a page and by default all Face-

book pages are public. Please note that anyone can view the page, however, to interact with a page (i.e. comment or post something), a registered Facebook user must "like" it first by pressing the like button on the page. After liking the page, the user can post a message/link/photo/video which will appear on the wall of that page. Other users who have already liked that page can post comments on a post, like the post, etc.

As an illustration, Figure 1 shows a Facebook page "LSU Football" with one post "LSU Tigers in the NFL – Week 10" from the user "LSU Football". 244 users have liked that post; a user has commented on it and 4 users have liked it.

As shown in Figure 2, in general, a public page P is a collection of many attributes. First, there is a title and description t of the page. Each page P may contain a number of posts p^1, p^2, etc. Each post p^i may have a few likes L_1^i, L_2^i, etc. and a few comments c_1^i, c_2^i, etc. Each comment c_j^i might have a few likes $l_{j,1}^i$, $l_{j,2}^i$, etc.

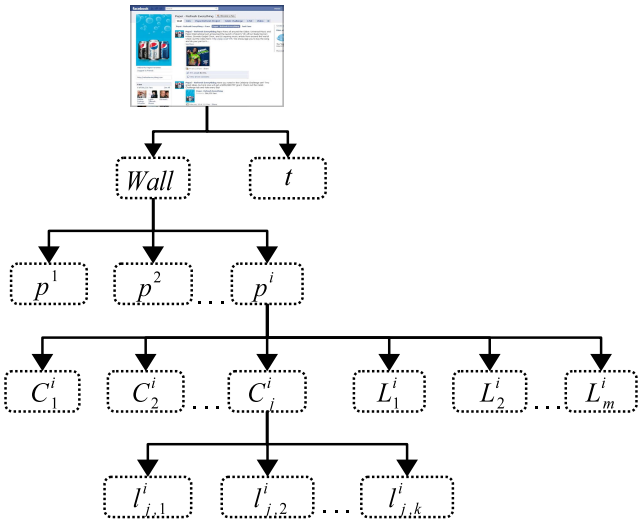

Figure 2: Structure of a Facebook page

3.2 Data collection

As discussed in Section 2, since we make use of public pages only, we can use Facebook's public Graph API to fetch data of any Facebook public page. Information that we manage to fetch for each page includes its title and description t, all posts p^i, comments c_j^i and likes L_m^i of each post, as well as likes of each comment $l_{j,k}^i$.

A small difficulty we faced was the authentication needed to use the Graph API. To fetch pages from Facebook using Graph API, one requires an authentication code. This authentication code is generally provided to Facebook applications for a limited amount of time. However, we did not build any Facebook application to obtain this authentication code. Instead, we created a php script to automatically login to Facebook and then parse the webpage returned at the following URL and look for the authentication code `https://developers.facebook.com/tools/access_token/`. Access token is only granted for a limited time, therefore this process can be repeated whenever the access token is expired. This work around was possible at the time of writing this paper.

There are limitations in using the public APIs only. For example, we were not able to obtain the IDs of those people who only liked the page and did not comment or posted anything on the page. We were also limited by the number of API calls we can make in a certain duration. At the time when the experiments were conducted, Facebook used to provide the list of users who liked a page, this feature is not supported anymore.

3.3 Automated profiling with attributes

In this subsection, we first present how we analyze an individual attribute on a Facebook page to figure out if a user has personal interests in the topic covered in that page. This might sound simple, as the user's interaction with the attributes of a page reveals some inclination towards that page. For example, if a user posts a positive message on the wall of a soccer page, it can be inferred that the user may be interested in soccer. For this we propose a technique SPM (**S**im**P**le **M**ining) to mine the information of users' interest (see Section 3.3.1). However, we also observed during our analysis that many users may "like" a page even though they are not interested in the corresponding topic, or if they have strong negative opinions on the topic. To solve this problem, in Section 3.3.2 we propose a more advanced technique called SOM (**S**entiment **O**riented **M**ining) to use sentiment analysis of the attributes to find the actual sentiment orientation of the users.

3.3.1 SPM: SimPle Mining

In SPM, we simply assume that a user's involvement in any of the attributes of P in whichever way indicates that the user has a same interest on the topic of the page, which can be inferred from t. For example, if a user likes a comment or adds a comment on a post on P, then we believe that the user is interested in P. If multiple users have interacted with P, we add all these users into the set $u_q^=$ which denotes the set of users who are interested in P about the interest q.

SPM is simple, but can easily introduce errors to $u_q^=$ because there could be a group of users of P (denoted $u_q^{\neq}$) who hold an opinion opposite to the focus of P. For example, a user who liked the cats page posted the following posts "I hate cats", "Lewis is a mad cat", "go doggies cats are crap" etc. They "liked" the page not because they really like it, but simply because Facebook does not allow them to add a post until they "like" it. To minimize this noise, we perform sentiment mining on textual data to find out the inclination of all users towards that page. See the following section for details.

3.3.2 SOM: Sentiment Oriented Mining

Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations [30]. Sentiment analysis has been used in many fields where users have subjective agenda such as movie reviews [23]. Intuitively, the content of the posts/comments should be accounted in deriving the users' interest. Hence, the polarity of the sentiment information of the text aids in conforming the users' interest.

We first define two sets of attributes on a Facebook page P. $\mathcal{A}_\mathcal{C} = \{t, p, c\}$ is the set of text-based attributes which consists of text, while $\mathcal{A}_\mathcal{D} = \{L, l\}$ is the set of dependent attributes which does not contain text. We separate these

Figure 1: A public Facebook Page

attributes into two groups because those that consist of text can go through a more thorough sentiment analysis on the text, while attributes of the other set are more dependent on the post or comment upon which the "like" was applied.

Sentiment analysis on $\mathcal{A}_C$.

We propose to use lexicon approach [14], which is one of the most popular methods used in sentiment analysis to detect the opinion bearing words. Lexicon approach concerns the use of lexical resources such as a dictionary of opinionated terms or opinion words. Collectively, they are called the opinion lexicon and are instrumental for sentiment analysis. Opinion words are the words that are commonly used to express positive (s^+) or negative (s^-) sentiments. For example: 'beautiful', 'wonderful', 'good', and 'amazing' are positive opinion words, and 'bad', 'poor', and 'terrible' are negative opinion words. Many opinion words are adjectives and adverbs. Sometimes, nouns such as 'rubbish', 'junk', and 'crap' and verbs such as 'hate' and 'like' also indicate opinions. Words which are neither positive nor negative are marked as neutral ($s^{\bigcirc}$).

Several opinion lexicons are available and SentiWordNet [2] is one such resource containing opinion information on terms extracted from the WordNet database and made publicly available for research purposes. SentiWordNet is a lexical resource built on top of WordNet. WordNet [9] is a thesaurus containing descriptions of terms, and relationships between terms and part-of-speech (POS) types. For example "car" is a subtype of vehicle and car has same concept as automobile. Hence, a synset (a synonym set) in WordNet comprises of all the terms with the same concept, e.g., the synset is car, automobile.

SentiWordNet assigns three sentiment scores to each synset of WordNet: positivity, negativity, objectivity/neutral. The sentiscores are in the range of [0, 1] and sum up to 1 for each triplet. For example, in SentiWordNet, the sentiscore of the term "good" is (pos, neg, obj) = (0.875, 0.0, 0.125). For our experiments, the scores are approximated with labels/part-of-speech of term in the text or sentence. First, the text is tagged using a standard POS tagger. A standard POS Tagger [27] is a piece of software that reads text in some language and assigns parts of speech to each word, such as noun, verb, adjective, etc. Then the SentiWordNet is used to get the scores for each term in the text. As our sentiment analysis is domain independent, we choose the general lexicon method as compared to the corpus-based method which is a domain dependent approach.

We now explain in detail how the sentiment is derived for the attributes in $\mathcal{A}_C$. We call them text-based attributes because their sentiment orientation is based on the text of the attribute. For a particular attribute $a \in \mathcal{A}_C$, we count the total number of words/phrases with positive sentiment (p_s) and that of words/phrases with negative sentiment (n_s) for all posts and comments he/she has, and use the term-counting method proposed by [29] to determine $\Psi(a)$ i.e. the sentiment orientation of a to be

$$\Psi(a) = \begin{cases} s^+, & \text{if } p_s > n_s \\ s^-, & \text{if } p_s < n_s \\ s^{\bigcirc}, & \text{otherwise} \end{cases}$$

Sentiment analysis of $\mathcal{A}_D$.

Attributes in $\mathcal{A}_{\mathcal{D}}$ does not contain text, but they also contribute to a user's sentiment orientation. We call them dependent attributes because their sentiment orientation is dependent on other attributes. For example, if a user u_1 has a post p^i with negative opinion, and user u_2 likes that post L^i_m, then both u_1 and u_2 share negative sentiment orientation on the topic. Similarly, if a user u_1 has a comment c^i_j with positive opinion, and user u_2 likes that comment $l^i_{j,k}$, then both u_1 and u_2 share positive sentiment orientation on the topic. That is,

$$\Psi\left(L^i_m\right) = \Psi\left(p^i\right)$$

and

$$\Psi\left(l^i_{j,k}\right) = \Psi\left(c^i_j\right)$$

Aggregating interests profiling from multiple attributes on multiple pages.

A user might have multiple posts, comments, and likes on a single Facebook page, and multiple Facebook pages might be about the same interest. Therefore, we have to aggregate the sentiment analysis results on multiple attributes from multiple pages in order to figure out the sentiment orientation of the user on that interest.

Let $A = \{a_1, a_2, ..., a_k\}$ be the set of posts, comments, and likes of a user u on a page P about a particular interest q. For each $a \in A$, we compute the sentiment orientation. Then, the sentiment orientation of u towards P, S^u_P is s^+/s^- if the number of attributes with positive sentiment is greater/lesser than the number of attributes with negative sentiment respectively, otherwise $s^{\bigcirc}$. Aggregating all Facebook pages about q, sentiment orientation of u towards q (S^u_q) is s^+/s^- if the number pages with positive sentiment orientation is greater/lesser than the number of pages with the negative sentiment orientation respectively; otherwise $s^{\bigcirc}$. If the sentiment orientation of q and S^u_q is same, then, u is added to the set $u^{=}_q$ otherwise to $u^{\neq}_q$. That is,

$$u^{=}_q = \left\{ u \in U \;\middle|\; \left(\left(\Psi(q) = s^+ \;\&\&\; S^u_q \in \{s^+, s^{\bigcirc}\} \right) \right.\right.$$
$$\left.\left. \| \left(\Psi(q) = s^- \;\&\&\; S^u_q = s^- \right) \right) \right\}$$

$$u^{\neq}_q = \left\{ u \in U \;\middle|\; \left(\left(\Psi(q) = s^+ \;\&\&\; S^u_q = s^- \right) \right.\right.$$
$$\left.\left. \| \left(\Psi(q) = s^- \;\&\&\; S^u_q \in \{s^+, s^{\bigcirc}\} \right) \right) \right\}$$

4. EXPERIMENTAL RESULTS

To base our analysis on a concrete example, we focus on breaking Blue Moon™ [15], a backup authentication system which can be used by a user to reset his lost or forgotten credentials. For example, if a user forgets his password of an email account, he or she can use Blue Moon to reset the password. The idea is to use personal preferences as challenge questions for authentication. Figure 3 shows a screenshot of Blue Moon[1].

During enrollment, the user is asked to select 8 items which he likes and 8 items which he dislikes from a list of 76 common interests. During authentication, the user is presented with a set containing the chosen items in a randomized fashion. The user categorizes the items to like and dislike. A user is not required to pick all the interests correctly. Instead, the user just need to correctly categorize 8 items[2] to reset his password [17]. To make our analysis consistent and the evaluation comparable, in the rest of the paper we assume that a user has to correctly categorize 8 items from the entire list of 76 interests which are shown in Table 2.

4.1 Dataset Description

In order to attack the Blue Moon system, we assume a strategy taken by an attacker as follows. He first constructs a set of interests Q from Table 2, and another set $Q^{'}$ containing the corresponding negated items like "I hate golf" and "I hate jazz". He then leverages the Facebook's public Graph API to 1) find all public Facebook pages related to $q \in \{Q \cup Q^{'}\}$; 2) fetch all attribute data of these pages; and 3) use the technique described in Section 3 to find $u^{=}_q$ and $u^{\neq}_q$ for all $q \in \{Q \cup Q^{'}\}$.

Note that this methodology does not cover all those pages which are semantically related with the query term. For example, query term "cars" may not fetch pages of Mercedes, Hyundai or Porche which are indeed the pages of cars. [4] provides a solution to fetch these pages using semantic search with the help of an ontology build upon Wikipedia. We are sure that this could significantly increase accuracy of the attack, although that comes with a price of longer processing of a large set of pages. We leave this as a future work to increase the size of the corpus.

Table 3 summarizes the availability of the attributes with their counts. From 34K pages fetched for 152 categories we found 2.5 million posts, 7.5 million likes and 4.3 million comments on these posts, and 1.3 million likes on those comments.

Attribute	Count
Categories	152
Pages	34,738
Posts	2,538,987
Post likes	7,574,965
Comments	4,381,967
Comment Likes	1,361,361

Table 3: Number of attributes found

We apply both SPM and SOM mining techniques as discussed in Section 3.3.1 and 3.3.2 respectively to the dataset `PubProf`, and discuss the results in the next subsections.

4.2 Inferred interests using SPM

We found a total of $1,162,575$ unique users whose interests can be inferred from the pages analyzed. These users

[1]This image is taken from `http://www.ravenwhite.com/iforgotmypassword.html`.

[2]The threshold where false acceptance rate and false rejection rate meets.

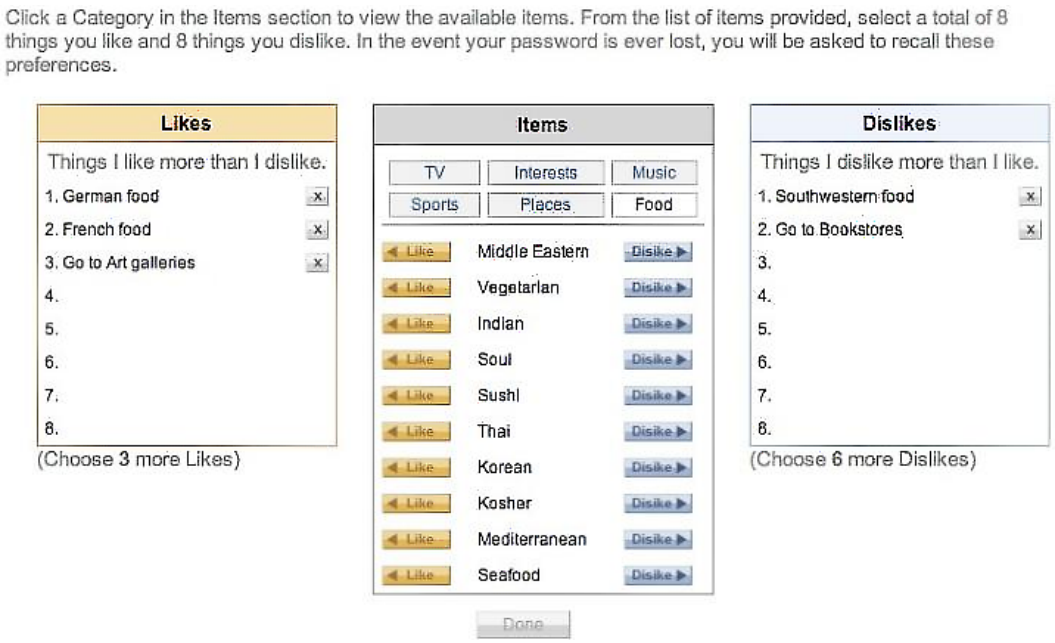

Figure 3: Blue Moon™

are the users who have either posted something on the pages, commented on the posts, or liked the posts/comments. Applying the SPM approach to our dataset, we detected 80,097 users with 2 or more interests. This amounts to 6.89% of all the user IDs collected in our dataset. Figure 4 shows breakdown of users with different number of interests found. We were able to build a user profile with as many as 22 interests.

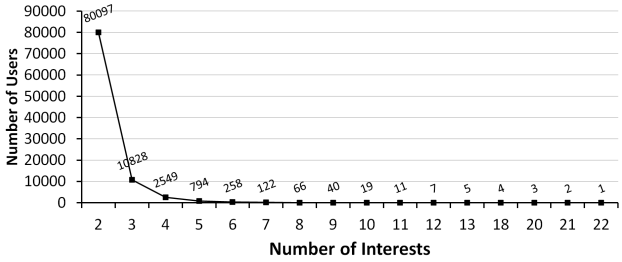

Figure 4: Inferred interests of the users using SPM

Note that although SPM might not be accurate in finding the true sentiment orientation of the user over an interest, the numbers presented here is not affected by this inaccuracy. That is, results presented in Figure 4 actually applies to SOM as well.

Results show that the number of users whose interests could be inferred from the public pages is significant, and this would have an important impact on the possibility of breaking into such users' Blue Moon account. For example, for those users found to have two interests, the search space for breaking into their Blue Moon account is reduced by about a factor of 3,000 ($^{76}C_2$). Please also note the dataset we use represents a tiny subset of the Facebook pages.

4.3 Inferred interests using SOM

In this section, we first investigate the inaccuracies when applying SPM on our data. As discussed in Section 3.3.2, these inaccuracies happen when users like a page but oppose to the topic in it. For example, if the page's title is "I hate Cats", we want to find those users who clicked on the like button on this page, however, actually like cats. Figure 5 shows the result of our sentiment analysis on all the comments, in particular, the number of negative comments different user posts. We can observe that about 10% of the comments posted are negative, meaning that the comment itself does not have the same sentiment orientation as that of the page. This suggests that more careful analysis and handling of the sentiments of posts and comments are important in order to find out the users' interests.

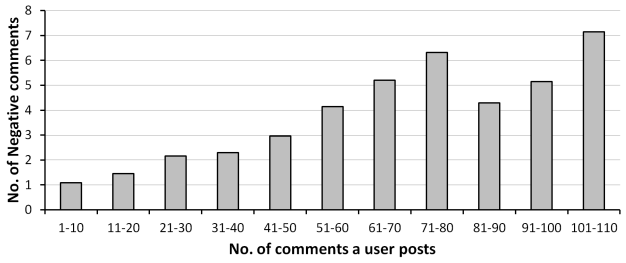

Figure 5: No. of negative comments posted by users

One interesting finding we also observe from Figure 5 is that users who comment a lot (more than 70 comments) tend to have a smaller percentage of negative comments. It is our future work to investigate whether the same is observed on a larger dataset.

To further investigate the result of our sentiment analysis, Table 4 shows 15 interests with the largest percentage of users who disagree with the corresponding topic of the page. $|u_q^{=}|$ is the total number of users whose sentiment orientation is same as that of the category, and $|u_q^{\neq}|$ is the total number

Sports	Music	Places	Interests	TV	Food
Aerobics	Instrumental	Garage sales	Cars	Watching Extreme sports	Soul food
Billiards	Symphony	Bookstores	Crafts	Documentaries	Indian food
Racing	Folk	Political events	Creative writing	Watching Auto racing	Korean food
Martial arts	Easy listening	Art galleries	Casino	Watching News	Kosher food
Baseball	Gospel	Raves	Painting	Watching Figure skating	Middle Eastern food
Soccer	Electronics	Antique stores	Gardening	Watching Diving	Southwestern food
Golf	Classical	Museums	Religion	Watching Baseball	French food
Running	Jazz	Flea markets	Politics	Watching Hockey	Seafood food
Yoga	Big Band	Libraries	Poetry	Watching Golf	Thai food
Skating	Reggae	The opera	Gaming	Game shows	Vegetarian food
Cycling	Show tunes	Politics	Reading comics	Watching Soccer	German food
Hockey	Heavy Metal		Cats	Watching Bowling	Mediterranean food
Pool			Gambling		
Motocross					
Basketball					
Football					

Table 2: User interests domain

of users whose sentiment orientation is opposite to that of the category. We see that although the inaccuracies from the SPM technique exist, most pages, especially those with a large number of users discussing, tend to have less than 10% of the users with negative sentiment orientation.

Category (q)	$\lvert u_q^{\neq}\rvert$ (%)	$\lvert u_q^{=}\rvert$ (%)	Total
Hate motocross	25.00	75.00	8
Hate skating	25.00	75.00	4
Hate heavy Metal	14.28	85.72	14
Hate poetry	12.50	87.50	8
Hate hockey	11.53	88.47	26
Watching baseball	11.12	88.88	9
Hate baseball	8.33	91.67	48
Hate basketball	5.52	94.48	181
Hate cats	5.39	94.61	260
Hate religion	5.36	94.64	56
Hate football	3.35	96.65	359
Raves	3.22	96.78	280
Gamble	3.08	96.92	195
Hate cars	3.08	96.92	130
Game shows	2.95	97.05	34

Table 4: The percentage of users whose sentiment orientation is not inclined / inclined towards the sentiment orientation of the page across all interests categories.

Another interesting observation is that most of the entries in Table 4 are of pages with a negative sentiment, i.e., q is "Hate xxx". We believe that it is because there are more people who want to voice out their disagreement with such pages than people who disagree with pages with a positive sentiment.

4.4 Comparing SPM and SOM

Table 5 shows the number of users found liking/disliking selected categories for both the SPM and SOM techniques. We only show a few categories with the largest discrepancies due to space constraint. Note that these are accumulated categories, e.g., we combined all 12 musical categories like jazz, classical, etc.

Results show that although there are users who like the page while having a different sentiment orientation as shown in the previous subsection, these users are minorities, and that is why we do not see a large discrepancy between results from SPM and SOM. In this respect results here seem to be consistent with those presented in Table 4.

Also, we observe that sports is the most popular category where 12.62% of users are inclined to the sports. These numbers could potentially be used to obtain the a priori probability for an unknown user having different interests, and subsequently used in attacking the Blue Moon system. We leave more detailed analysis on this for our future work.

Main Category	#Users using SPM	#Users using SOM
Like Sports	146756	145141
Like Music	29597	29255
Like General	65354	64571
Like Entertainment	163	158
Like Food	4031	3990
Do not like Sports	1120	1079
Do not like Music	30	28
Do not like General	825	800

Table 5: Likes and Dislikes

4.5 Errors in sentiment analysis

As text mining is prone to errors, in this section we evaluate the correctness of our SOM approach in detecting the correct sentiment orientation. We manually label 300 sentences randomly chosen from various categories including sports, music, religion, politics, cats, and food. An independent human annotator then labels each sentence to either s^+, $s^{\bigcirc}$ or s^- depending on his/her understanding of the sentence. We then evaluate SentiWordNet's accuracy using precision metrics (automated labeling against manually annotated labels). This measure have been commonly used to evaluate the accuracy of various retrieval, classification, and mining algorithms. Precision refers to the proportion of true positives over the sum of the true positives and false positives. The sentiment mining technique provided an overall

accuracy of 69.33%, see Table 6 where the diagonal figures represent the accurate labeling, while off diagonal figures represent false positives.

		Estimated/Sentiwordnet		
		s^+	$s^\bigcirc$	s^-
Actual/Human	s^+	**66.33**	17.35	16.33
	$s^\bigcirc$	23.68	**63.16**	13.16
	s^-	18.25	06.35	**73.02**

Table 6: Confusion matrix for sentiment (%)

Sentiment analysis failed for conjugate and multi sentences. For example, "*As far as intelligence goes cats have a different kind of intelligence than that of dogs. They can MANIPULATE their environment to SURVIVE can hunt on their own and...*" is labeled negative and "*really no 1 can say our governance in 9ja is understood... we are just driven here and there no prosperous direction... we really don't know... Im shot of words 4 my dear country*" is labeled as positive. Sentiment approach also failed for sarcastic statements like "*GREAT.. now I can not get ONLINE...*" which has been labeled positive.

The inaccurate estimation of positive to negative or negative to positive labels has more impact on building the user profile. The neutral messages are overwritten by the page polarity and hence no impact on user profile.

4.6 Concentrated group with ground truth

We tried to get some ground truth to be compared with the results obtained. We chose 450 users from $u^=$ obtained using the SPM approach with the largest number of interests inferred (more than 4 in particular). Out of the 450 user profiles, 47 have either been deactivated or deleted. We manually sent Facebook friend requests and messages to the remaining 403 Facebook users to know more about their interests in certain categories. In particular, we provided the user with a list of all the inferred interests (using SPM) of the user and asked him/her to classify into three categories i.e. s^+, s^-, $s^\bigcirc$. Due to privacy settings imposed by many users, we were only able to send 334 friend request (70 accepted) and 299 messages (15 replied back).

We expected those who accepted our friend request would reply to our messages sent; however, there were only 12 (out of 70) who replied to our message. 56 did not reply and we were not able to send message to 2 users because of the privacy settings imposed by them.

Majority users (212) neither accepted our friend request nor responded to our message sent. There were 2 participants who did not accepted our friend request but still replied to the message. Please refer to Table 7 for a summary of the responses we got.

		Disallow Message	Allow Message	
			No reply	Replied
Friend req. not allowed		52	16	1
Friend req. allowed	Not added	50	212	2
	Added	2	56	12

Table 7: Users category settings in `VolProf`

We take the responses of those 15 users who replied to our messages and compared them with the corresponding

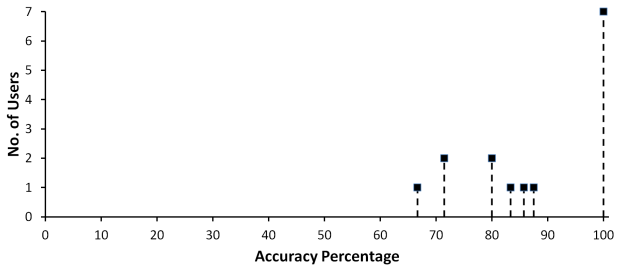

Figure 6: Users Interest from `VolProf`

interests inferred using SPM. Figure 6 shows the percentage of the correctly inferred interests of these users. It shows that we were able to infer approximately half (7 out of 15) of the users' interests with 100% accuracy. Also, we can see from the figure that at least two-third of the interests are inferred correctly for all the 15 users. Although we did not manage to get a larger pool of people with ground truth due to the manual work involved, available data seems to suggest that our technique provides reasonable accuracy in inferring users' interests from public pages on Facebook.

5. DISCUSSION AND LIMITATIONS

In this section we first discuss another threat model and results under this new threat model with our manual analysis. We then discuss other factors that could have contributed to inaccuracies and point out limitations in our results.

5.1 A different threat model

Apart from the threat model discussed so far, we discuss the strategy that can be adopted by an attacker who has access to *all* the pages the victim has liked on Facebook. The objective of the attack is still to guess the Blue Moon^{TM} categories. This is a reasonable threat model when the attacker is a malicious app or some close friend/colleague of the victim. For example, spouses check their partners accounts without their permission, which can lead to divorces later on [1]. Partners are usually friends on Facebook who have access to most of the information including "likes" etc.

We conducted a small experiment by inviting some users to the lab and collecting their page "likes" (only after seeking their permission). We hired two research assistants to independently categorize these pages according to the Blue Moon^{TM} categories. This manual analysis on one user having the highest (926) number of page likes shows that 71 out of 926 pages can be categorized to Blue Moon^{TM} categories. These 71 pages fall into 15 Blue Moon^{TM} categories. This shows that an attacker who has access to the user's like pages could possibly use them to break into the victim's Blue Moon^{TM} system.

5.2 Limitations

One of the most important contribution to inaccuracies in our analysis is noise in the public pages on Facebook. This noise could come from advertisements, spamming in general, conversations in comments, and others.

Facebook is popular marketing media and some users are actually advertisers. For example, a person selling Nike shoes may post his ad in all the categories corresponding

to sports. Our system might therefore believe that this user has an interest in sports. To have a sense on the noise level, we search for advertisements in selected categories by randomly choosing some posts and manually labeling them as advertisements. Table 8 shows the number of advertisements found in a number of posts for selected categories.

Category	No of posts scanned	% of Ads found
Sports	1,348	0.445
Food	2,312	0.346
Cats	2,312	0.216
Music	7,000	0.171
Politics	9,928	0.060

Table 8: Percentage of Advertisement posts

It appears that the noise level in our dataset is very low, however, our manual process in finding advertisements might be error prone, too. Another step we took to minimize this error was to manually check whether the users inferred with more than eight interests are real users. Our simple manual checking revealed that except a few users whose accounts had been deactivated or deleted, all of them seem to be legitimate. We also filtered off users who posted same message in more than four pages.

We have seen many people get into their personal conversations on public pages that are not related to the topic of the page. Unfortunately we do not find a scalable way of filtering out such noise, and therefore it might have contributed to errors made.

There are also limitations in the techniques that we use. First, we use a context independent text mining algorithm. This limits our capability in analyzing the sentiment of certain pages, e.g., "Lets help the dogs in the streets and kick the cats." The sentiment scoring without context will fail to identify the user interest in this example. To solve this problem, we need additional scoring models that can handle the sentiment with the context.

Second, we only managed to obtain ground truth for a small set of users. We wish we could find a better approach to obtain the ground truth, but sending out message to a large number of users had one of our Facebook account suspended, and that was why we did not go further to target a larger group.

Last but not the least, manual work was involved in a number of steps in our experiments, including evaluation of the accuracy of sentiment analysis, spam detection, etc. This manual work could potentially introduce errors into the evaluation.

6. CONCLUSION AND FUTURE WORK

In this paper we raised two important questions, as why public information is made public? And is there any information that can be mined to break preference based authentication. We present two mining based approaches to predict undisclosed users' interests. Using only simple mining approach we extracted unobservable Interest topics by analyzing the corpus of Interests obtained via legitimate use of Graph API provided by Facebook. From our experiments, we were able to disclose 22 interests of a user and found more than 80,097 users with at least more than 2 interests. We also show how this inferred information can be used to break preference based backup authentication system. We also demonstrated that simply liking a Facebook page does not imply the users' inclination towards that page. We showed that there exists many users who liked a Facebook page, however they post negative comments on the page. We also compared our work with prior research work which involves crawling personal profiles (either public or private). In future, we would like to improve our mining approach to infer more privacy attributes apart from interests.

7. REFERENCES

[1] F. L. Attorney. Maintaining privacy and starting a separate life during divorce. `http://goo.gl/Nm5Wc`.

[2] A. E. S. Baccianella and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[3] M. Balduzzi, C. Platzer, T. Holz, E. Kirda, D. Balzarotti, and C. Kruegel. Abusing social networks for automated user profiling. In *Proceedings of the 13th international conference on Recent advances in intrusion detection*, RAID'10, pages 422–441, Berlin, Heidelberg, 2010. Springer-Verlag.

[4] A. Chaabane, G. Acs, and M. A. Kaafar. You are what you like! Information leakage through users' Interests. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium*, Feb. 2012.

[5] Y. Chen, D. Pavlov, and J. F. Canny. Large-scale behavioral targeting. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 209–218, New York, NY, USA, 2009. ACM.

[6] N. Cubrilovic. The anatomy of the twitter attack. `http://techcrunch.com/2009/07/19/the-anatomy-of-the-twitter-attack/`.

[7] Facebook. Facebook pages. `http://www.facebook.com/directory/pages/`.

[8] Facebook. Graph api. `https://www.developers.facebook.com/docs/reference/api/`.

[9] C. Fellbaum. Wordnet: An electronic lexical database. `http://wordnet.princeton.edu/`.

[10] L. Gannes. Mole-whacking: Vendor says spam is growing on facebook fan pages. `http://goo.gl/a89Aa`.

[11] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli. The adaptive web. In P. Brusilovsky, A. Kobsa, and W. Nejdl, editors, *The Adaptive Web*, chapter User profiles for personalized information access, pages 54–89. Springer-Verlag, Berlin, Heidelberg, 2007.

[12] D. Gayo Avello. All liaisons are dangerous when all your friends are known to us. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 171–180, New York, NY, USA, 2011. ACM.

[13] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. On exploiting innocuous user activity for correlating accounts acrossăsocial network sites & twitter & personal profiles. Technical report, International Computer Science Institute, 2012.

[14] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[15] R. Inc. The blue moon authentication system. `http://www.ravenwhite.com/iforgotmypassword.html`.

[16] D. Irani, S. Webb, K. Li, and C. Pu. Modeling unintended personal-information leakage from multiple online social networks. *IEEE Internet Computing*, 15(3):13–19, May 2011.

[17] M. Jakobsson, E. Stolterman, S. Wetzel, and L. Yang. Love and authentication. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 197–200, New York, NY, USA, 2008. ACM.

[18] X. Jin, C. Wang, J. Luo, X. Yu, and J. Han. Likeminer: a system for mining the power of 'like' in social media networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 753–756, New York, NY, USA, 2011. ACM.

[19] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems, 2009.

[20] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 1145–1146, New York, NY, USA, 2009. ACM.

[21] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 251–260, New York, NY, USA, 2010. ACM.

[22] J. Owyang. The many challenges of social network sites, 2008.

[23] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[24] A. Rabkin. Personal knowledge questions for fallback authentication: security questions in the era of facebook. In *Proceedings of the 4th symposium on Usable privacy and security*, SOUPS '08, pages 13–23, New York, NY, USA, 2008. ACM.

[25] Robots.txt. A standard for robot exclusion. `http://www.robotstxt.org/orig.html`.

[26] U. T. Ted Bridis, Associated Press Writer. Hacker impersonated palin, stole e-mail password. `http://www.usatoday.com/news/politics/2008-09-17-152224562_x.htm`.

[27] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[28] A. Tsotsis. Hacker proves facebook's public data is public. `http://techcrunch.com/2010/07/28/hacker-proves-facebooks-public-data-is-public/`.

[29] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[30] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics*, 35(3):399–433, 2009.

[31] Yahoo. Rate limiting for yahoo! search web services. `http://developer.yahoo.com/search/rate.html`.

[32] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. Like like alike: joint friendship and interest propagation in social networks. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 537–546, New York, NY, USA, 2011. ACM.

[33] 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 *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 531–540, New York, NY, USA, 2009. ACM.