

Social Footprints : Using Profile Attributes to Determine Vulnerability in Online Social Networks

Hamza Karachiwala*, Kedar Amrolkar[†] and Mickey Vellukunnel[‡]

College of Engineering

University of Florida

Email: *hskarachiwala@ufl.edu, [†]kamrolkar@ufl.edu, [‡]m.vellukunnel@ufl.edu

Abstract—There are different types of Online Social Networks today, each with very large user bases. These users tend to disclose much of their identifying information on these sites via public profiles which can be accessed by malicious users. While there may not be enough public data on a single OSN, the public data across all the OSNs aggregated together results in a comprehensive collection of user profile data. This aggregated data is called a social footprint. We intend to measure how heavy a users aggregated footprint may be, given the user has accounts on different Online Social Networks. We calculate this footprint weight by giving weights to all the attributes that build this footprint. Attributes are information like name, address, place of work, etc. These weights are assigned based on the sensitivity of the data. We then identify if the weight crosses a threshold value making it vulnerable to malicious outsiders. If the footprint is vulnerable, we suggest possible ways of reducing the weight of this social footprint.

I. INTRODUCTION

We have witnessed the advent of various social networking sites over the last decade. And along with this, we have also witnessed the phenomenal adoption of these social networks. Users on these sites continues to multiply daily mirroring the willingness of the audiences to join the OSN bandwagon. Statistics show[13] that the number of users worldwide has increased from 0.97 billion to 1.96 billion in just the last five years. This number is expected to shoot up to 2.5 billion by 2018. 75% of the US population has a social network profile. Which means that 3 out of every 4 people can be found online.

Moreover, these sites cater to different aspects of our lives. One would often have separate social networks for his personal and professional lives, a separate platform for photo sharing and at the same time, a public channel for videos. These social networks are ever evolving continue to be more and more pervasive in our lives. Although this remarkable progress is utterly commendable, it does come with its own perils. After all, it is a “social” network with humans communicating with each other. This leads us to sometimes unwittingly disclose information about ourselves on these websites which would actually be quite sensitive or private. While we see it as a little bit of information about ourselves, these bits and pieces spread over varying social networks aggregated together as illustrated in [3], makes it possible for people with malicious intent to gain information about us which may be used in a harmful way.

Consider this report[14] from 2014 which dishes out a few astonishing numbers related to cyber crimes. In the surveys conducted, it was found that 78% burglars used the geotagging features across OSNs to locate their victims and plan their strikes. Almost 50% of sex crimes committed against a minor have been involve the perpetrator obtaining profile pictures. But here is the most worrying statistic - 66% of Facebook users had no notion of privacy settings were and what they had disclosed online. 15% also admitted that despite the knowledge, they never bothered to check because there are so many sites.

Our aim is to warn users about the risk they may be exposed to because of this level of information disclosure in cyberspace. By accessing various degrees of their data on these social network websites, we want to build an aggregated profile of a user - his social media footprint, along similar lines as explained in [1]. This social media footprint will be an intuitive metric of an individuals presence across social networks. We want to further provide this footprint to a user with the intention of encouraging him to trim down the personal information he might have disclosed without knowing the consequences. While various surveys have been conducted over the same concerns, there have been no known attempts to formalize the vulnerability of a social footprint. Also, there has not been a formal evaluation of the number of crimes arising because of a particular privacy vulnerability. It is of prime importance to reach out to users and explain this hard-to-visualize problem in a simplistic and intuitive way.

II. RELATED WORKS

Researchers have since long identified the need for privacy in social networks. They have studied the implications of leakage and the consequences as also suggested methods to avoid this [1],[4],[3],[5]. Moreover, they have observed that while it is risky to disclose information on an OSN, that is also the base of its success. In order to gain the benefits of an OSN, it is necessary to share certain information and be identifiable to a certain set of people at least. This is the reason that makes it important to study footprints and aggregated data as proposed in [1],[2] and [7]. Various proposals have been made to calculate an aggregated footprint on a single website or spanning multiple sites and then measure the risk associated with these footprints [9],[10]. These methods

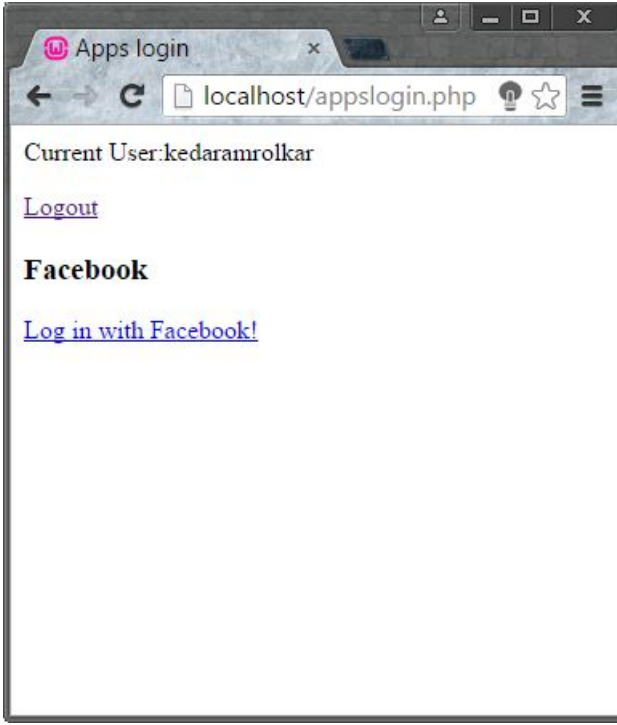


Fig. 1. Facebook Login Page

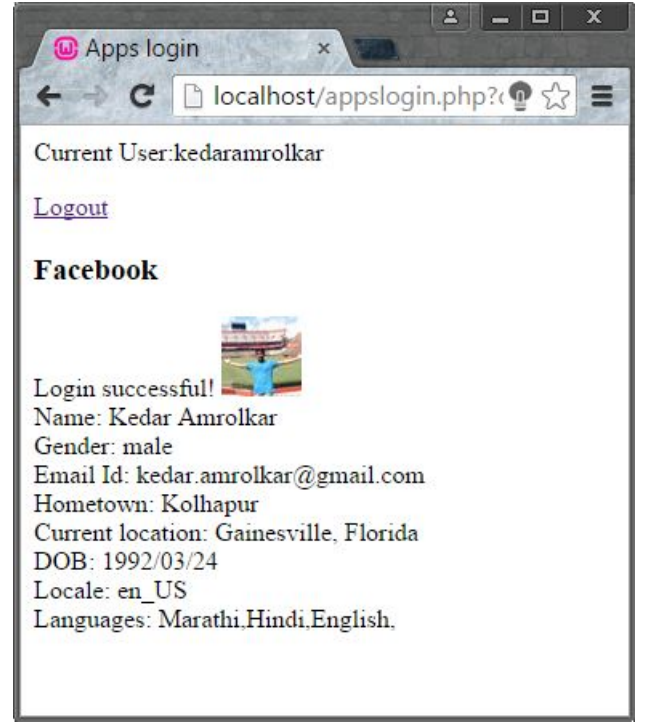


Fig. 2. User Data retrieved from Facebook

rely on intuition to weigh leaked profile attributes. We will attempt to provide weights by studying past crime instances that occurred as a result of attribute leakage.

There have also been some interesting and in depth mathematical models designed which we would be interested to incorporate in our analysis. The representation by Irani et al [2] of a social footprint as τ_s^u is very intuitive. They then create an aggregate representation $P^u = \bigcup \tau_i^u$ for the footprint spanning various sites. Another notable mention is the *PIDX* proposed by Nepali et al. [6]

$$PIDX(i, j) = \frac{w(i, j)}{w(j)} \times 100 \quad (1)$$

While there are tools proposed to analyze data on Facebook [8] and [11], we are attempting to provide a common application for thorough footprint measurement. While Facebook is undoubtedly the major contributing factor to a footprint, the aggregation from multiple sites is what largely includes the risk element. We have access to more of these OSNs via APIs than it was possible before, with the advent of REST architectures in the web industry.

III. PROBLEM AND SOLUTION

Our main aim is to generate awareness among OSN users about the concept of social footprinting. While a user may be aware of the risk of disclosing information on a single OSN, he would most probably not be aware that profile information across multiple OSNs can be aggregated to generate a more comprehensive list of user information. As mentioned in [1],

while a user may disclose 4 unique fields in a single OSN, this number doubles when he is a user of five OSNs. This information can then be exploited by miscreants and criminals to cause harm to the user. This may range from offences such as harassment or stalking to crimes such as identity theft.

To solve this problem, we first need to collect aggregated data from a user spanning across his multiple social network accounts to build his social footprint. We will achieve this using the APIs exposed by these OSNs. We have so far integrated our application with the Facebook API to pull a users profile data. We aim to use this approach for more popular OSNs which have a large user base. These include Google+ which is similar to Facebook. We then have Twitter which might provide a list of persons that the user might be following signifying an ideological similarity. Instagram is a photo sharing website that may have its own set of interesting data. LinkedIn is extremely important as a professional network and exposes a different kind of data. SoundCloud would list music preferences and so on.

Figures 1 and 2 show a rudimentary functioning of the Facebook API integration. Following this integration, the retrieved data is stored in our database where it will await more data from other OSNs. Repeated data can be ignored. The new data can be aggregated with what has been found so far.

Once we have built a footprint, we need to measure its weight. In order to do this, we must have weights associated

TABLE I
ATTRIBUTE WEIGHTS BASED ON CRIME RECORDS

First Name	1
Second Name	1
Middle Name	1
Age	1
Sex	1

with all the attributes of a users profile data. We propose to assign these weights based on empirical estimates derived from criminal records. We will list out which attributes contribute the most towards a cybercrime caused by data leakage via public profiles. These attributes will have higher weights. This is a significant challenge as this data is not only crucial to build the mathematical model, it is also very difficult to collect. We have so far scanned over various online resources and collected some numbers. However, it would be important to find and utilize a consolidated data source of crime records and causes. We will continue looking for such data and at the same time, build our weight function using external online sources.

Table 1 illustrates the weight table we have built so far. We will keep extending on it as and when we encounter new attributes or identify weights for known attributes from crime records.

Based on the weights assigned to attributes, we will compute the aggregated weight of the social footprint. This weight can be calculated trivially as the sum of the weights of all the attributes that have aggregated into the footprint. Then comes the challenge of identifying the risk associated with that footprint. We need to identify a threshold value beyond which the weight of a social footprint would classify as at-risk. We intend to assign a trivial random value for this due to the time constraints. However, this value should be calculated by observing crime patterns as well.

Our next task will involve identifying ways to reduce the weight of a social footprint. In this step, the user will be suggested various attributes that could be left out or hidden from his public profile. These suggestions will be made based on the weight of these attributes, and the objective will be to minimize the footprint weight by leaving out these heavy attributes that may cause harm to a user by disclosing it as a public profile attribute.

IV. ROADMAP FOR REMAINING WORK

A. Collecting footprint data from more OSNs

The increase in OSNs coupled with the boom in API integrations gives us a rich source of information. We are hoping to integrate the tool with as many Social Networks via their APIs as possible. These will also span various domains - while Facebook may have most of the information that contributes to a footprint, there are many other interesting OSNs today that may contain interesting bits of information that could increase the weight of a footprint. LinkedIn will have professional information about a person which may help draw inferences about his travel routes, work hours and most importantly, income. We also have interesting OSNs such as Pinterest which is an easy listing of personal interests. Photo sharing sites such as Instagram and music preference sites such as SoundCloud can further help somebody maliciously gain personal information. From our perspective, it is helpful that these OSNs usually expose APIs for developers to integrate with them. This provides us an easy way to pull data for our experiments.

B. API Integration

The technology stack that we use for the web application is Apache as the server, PHP for the server side scripting and MySQL as the database. PHP being the choice to integrate with as many APIs as possible. No prior experience with PHP comes as a significant challenge to us and is a very important hurdle to overcome as this is the selected source way of collecting footprints for us. This slows down our process of collecting data from multiple sites and building the social footprints. However, we have integrated with Facebook which should be very helpful in building forward and integrating with more OSNs. Once we integrate with multiple OSNs, a user will access the web application and then provide his authentication for all the OSNs that he uses. This is needed to call the API to fetch that users data. On retrieving the data from a single OSN, we can store it in a database. The user can then continue to authenticate himself for one OSN after the other, and we can access his data and aggregate it in the database. Once, all the data is collected we can retrieve the entire footprint to calculate the weight.

C. Assigning attribute weights

We then have to solve the problem of associating weights with these attributes. Prior works assign these weights based on intuition with no real well defined meaning to these weights. We propose that these weights should be based on how the attribute leakage has already harmed users historically. Consider a popular attack in the past where a user might have disclosed his mothers maiden name as a public profile field. This is a popular security question on most websites. This makes the attribute heavy, that is, it would contribute more towards making a social footprint vulnerable. At the same time, disclosing your first name is not as risky as it cannot be used much with a malicious intent. However, aggregated attributes may have more weightage. For example, the first

name, last name and date of birth together can create a vulnerable footprint. While these attributes individually are safe to disclose, when they are disclosed together they can be used effectively in identity theft. This in some ways explains how aggregated data is more dangerous than data on a single social network. For an attacker, an aggregated footprint provides much more information and means to cause harm than the data on an individual social networking site. Another aspect in attributes is identifying a person more precisely. Some attributes such as age and sex may seem generic but these are useful in guaranteeing that the person is the same across multiple OSNs. This helps in creating a social footprint again, which makes these attributes risky to disclose in a group.

D. Calculate and compare with threshold

After having a list of assigned values for different attributes, we then calculate the weight for that users footprint. We must check what attributes are present in the users footprint. According to that we will fetch those attribute weights and do a simple sum on those weights. This will amount to the total weight of that footprint. Mathematically, this can be represented as,

$$w(\text{footprint}) = \sum_{i=1}^n \text{attribute}_i \quad (2)$$

E. Check for vulnerability

After we calculate the weight of the footprint, we compare it to a threshold value to check if it is vulnerable. We plan to do this simply by considering the average footprint weights of five users who can be considered to have safe footprints. That is, these users will be having very minimal public profile attributes which guarantee bare minimum visibility on these websites. This average footprint weight will be considered as the threshold for our remaining checks. While this is crude as compared to some previous works as in [6], it will temporarily accomplish the requirements and we can extend on this approach when time permits.

F. Suggestion reduction in weights

Once we have the threshold value and the result of the comparison with that value, we can determine the vulnerability of the footprint. In simple words if the footprint weight crosses the threshold value, it is vulnerable and we must bring the footprint weight down below the threshold. In order to achieve this, some attributes need to be removed from the users footprint or effectively from their OSN profile. This must be done keeping in mind that visibility must not be reduced drastically. We need to minimize the weight of the footprint and at the same time keep the visibility constant. While this is a classic linear programming problem, we will simply provide suggestions to users in decreasing order of weight of attributes. Considering that the weights will be assigned keeping in mind visibility, we can assume that heavy attributes will also be ones that will be affecting visibility less. The user can then decide

if he wants to modify his profile or not. We will only provide suggestions.

V. CONCLUSION

We attempt to highlight the dangers of disclosing a range of personal data via public profile attributes on online social networks. We will build a social footprint of an individual by accessing his public and other profile data across various OSNs and aggregating them. We will then calculate this footprint weight by assigning weights to these attributes. The weights will be based on the risk measure of the attributes by observing past crime patterns related to OSNs over the past few years. By comparing this footprint to a threshold value, we will then determine the level of risk a person is subject to with his current OSN's public and other profile data. Following this we can provide suggestions to reduce the public and other data and trim the weight of the footprint.

REFERENCES

- [1] Danesh Irani, Steve Webb, Kang Li, Calton Pu: *Large Online Social Footprints -An Emerging Threat.*
- [2] Danesh Irani, Steve Webb, Calton Pu, Kang Li: *Modeling Unintended Personal-Information Leakage from Multiple Online Social Networks*
- [3] Balachander Krishnamurthy, Craig E. Wills: *Characterizing Privacy in Online Social Networks*
- [4] Ralph Gross, Alessandro Acquisti: *Information Revelation and Privacy in Online Social Networks*
- [5] Monica Chew, Dirk Balfanz, Ben Laurie: *(Under)mining Privacy in Social Networks*
- [6] Yong Wang, Raj Kumar Nepali, Jason Nikolai: *Social Network Privacy Measurement and Simulation*
- [7] E. Michael Maximilien, Tyrone Grandison, Tony Sun, Dwayne Richardson, Sherry Guo, Kun Liu: *Privacy-as-a-Service: Models, Algorithms, and Results on the Facebook Platform*
- [8] Nilothalpal Talukder, Mourad Ouzzani, Ahmed K. Elmagarmid, Hazem Elmeleegy, and Mohamed Yakout: *Privometer: Privacy Protection in Social Networks*
- [9] Cuneyt Gurcan Akcora, Barbara Carminati, Elena Ferrari: *Privacy in Social Networks: How Risky is Your Social Graph?*
- [10] Kun Liu, Evimaria Terzi: *A Framework for Computing the Privacy Scores of Users in Online Social Networks*
- [11] Justin Becker, Hao Chen: *Measuring Privacy Risk in Online Social Networks*
- [12] L.Sweeney: *Uniqueness of Simple Demographics in The US Population*
- [13] Statistics Portal at statista.com
<http://www.statista.com/topics/1164/social-networks/>
- [14] Socialnomics
<http://www.socialnomics.net/2014/03/04/the-shocking-truth-about-social-networking-crime/>