

Social Footprints : Public Profile Attributes to Determine User Vulnerability in Online Social Networks

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

University of Florida

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

Abstract—Social networks have been a very important advance over the last decade and a half. Today, there are various types of Online Social Networks or OSNs. Most of these OSNs have a very large user base which confirms their popularity and also their need in the present time. These users tend to disclose much of their identifying information on these sites via public profiles in order to be discovered by other users. However, these users may also be malicious and pose a threat to others in the network by collecting public profile data of any user. 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 measure this by assigning weights to all the individual public profile attributes that build this footprint. These weights are assigned based on the sensitivity and the need of having a particular attribute disclosed in a user's public profile. We then identify if the aggregated weight of this footprint crosses a threshold value making it vulnerable to a malicious attacker. 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 a half. And along with this, we have also witnessed the phenomenal adoption of these social networks. Users on these sites continue to multiply daily, which highlights the willingness of the masses to join this OSN bandwagon. Statistics show[17] 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. In fact, another striking statistic is that 75% of the US population has a social network profile. Which means that 3 out of every 4 people can be found online. These statistics are not only restricted to a single age group or income group, but are in fact applicable to a wide range of audience as can be observed from [18]. It is safe to say that social networks have definitely been a very influential advance in computing and the numbers and statistics listed above corroborate that.

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 that are ever evolving, continue to be more and more pervasive in our lives. Although this remarkable progress is highly commendable, it does come with its own perils. After all, it is a “social” network with humans communicating with each other. While providing a rich source of both information and interaction, it also requires us to disclose some basic facts about ourselves such as our name, gender and in some cases our photograph. The purpose behind disclosing these attributes is so that a user can be “visible” on the network. That is, a user can be discovered or recognized from the others on a network. However, while this visibility information can be bare minimum, we 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. While the independent bits and pieces of information may be harmless by themselves, a systematic collection of all that data paves the way for various crimes such as identity thefts, dictionary attacks, harrasment, etc

Consider this report [19] 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 involves the perpetrator obtaining profile pictures. But here is the most worrying statistic - 66% of Facebook users had no notion of what 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. This illustrates the carelessness of the masses in general when it comes to protecting profile data. And while this may not be as harmful on a single OSN, or the dangers might not be highlighted by a limited number of attributes, the problem magnifies when we pick data from multiple OSNs and build a social footprint by aggregating all this data. The amount of disclosed data increases drastically. Subsequently, it increases the vulnerability or susceptibility of a user to attacks that can be carried out by knowing that

information.

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 individual's 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. Users are often not aware of the consequences and through this work, we can generate that awareness. While various surveys have been conducted over the same concerns, there have been no known attempt to formalize the vulnerability of a social footprint. 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],[11]. These methods 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. For example, a burglary that is caused by a checkin would mean a higher weight for a checkin. Or accounts are hacked due to dictionary attacks. These dictionaries are created using profile data and something like a school name would possess a higher attribute weight if a crime has been carried out by using it.

However, there has been a marked rise in the number of OSNs since the past few years. And most of these OSNs have not been included in the various analyses and works done above. While the core contributors to a social footprint would remain very much the same, there are some interesting new attributes that are now added to a social footprint. These are worth investigating as well. Namely, interest lists that may be garnered from Pinterest or favourite public figures that may be enlisted from Twitter. These lists can be used for other activities such as targeted advertisements which would in a way be illegal as that profile data is not meant to be used.

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]

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

While there are tools proposed to analyze data on Facebook [8] and [12], we are attempting to provide a common application for thorough footprint measurement. Privometer was especially promising as a Facebook app which uses a similar procedure to calculate the weight of public profile information disclosed on Facebook. It also had some self explanatory graphic interfaces for the user to help him understand the dangers of disclosing this data and how these dangers could be averted by minimizing some of the data disclosed on the public profile. However Privometer is not available as of today for use and some of the Facebook regulations have changed over time. 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

In this section, we identify the scope of our problem and then illustrate our approach to solving it. 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/she 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. Most OSNs provide these APIs to integrate with other applications. In the case of Facebook, LinkedIn and Google+, they also act as an authentication medium. We will use this functionality to fetch user data from these three websites and avoid the cumbersome crawling or scraping data from these websites for a user. While Facebook and Google+ are complete social networks in a sense, LinkedIn is restricted to professional data. Together, we believe that these three sites would cover a majority of common public attributes and would make quite an accurate social footprint, if not the most accurate. We will

discuss some other sources of data in Section 5 and how they could improve on the comprehensiveness of the footprint. The remaining of this work, however, uses Facebook, Google+ and LinkedIn.

Once we have built a footprint, we need to measure its weight. In order to do this, we must have weights associated with all the attributes of a user's profile data. We will be assigning these weights based on statistics related to two kinds of attacks - identity theft and password recovery. For password recovery, we build on the work of Bonneau et.al [15] which highlights the risk of using password recovery questions. In both cases, the idea is to assign a higher weight to an attribute that is associated with more risk, or could be more harmful to divulge as part of the footprint. While the statistics we have to back our mathematical model are limited, we believe that it is trivial to strengthen these claims and would simply need some more work on data collection. We explore this in detail in Section 5. In the case of identity theft, we consider the contribution of an attribute towards identification of a victim and subsequent impersonation.

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 simple average value for this due to the time constraints. We will assume for now that a name and profile picture is a bare minimum requirement for visibility in these sites. Our threshold will be

$$w(first - name) + w(last - name) + w(profile - pic)$$

where $w(.)$ is the weight assigned to any attribute. This is the simple weight required by any profile to be visible or discoverable in the least. We will consider any other information above this as associated with some risk. We propose improvements to this approach in Section 5. Following this, 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. The objective will be to minimize the footprint weight by leaving out heavy attributes that would be associated with most risk. While we would like to find an optimal subset of attribute weights that maintain a balance of visibility and risk, in this work we restrict our suggestions to a greedy approach that suggests pruning attributes in decreasing order of weights.

In the following subsections, we give details on each of these steps required to solve this problem by outlining the approach and the obstacles that have to be overcome to improve this approach.

Attribute	Source
First Name	Facebook, Google+, LinkedIn
Last Name	Facebook, Google+, LinkedIn
Profile Picture	Facebook, Google+, LinkedIn
Gender	Facebook, Google+, LinkedIn
Date of Birth	Facebook, Google+, LinkedIn
Email	Facebook, Google+, LinkedIn
Phone number	Facebook, Google+, LinkedIn
Marital status	Facebook, Google+, LinkedIn
Lives In	Facebook, Google+, LinkedIn
Previous cities	Facebook, Google+, LinkedIn
Works at	Facebook, Google+, LinkedIn
Occupation	Facebook, Google+, LinkedIn
Studies at	Facebook, Google+, LinkedIn
Previous work	Facebook, Google+, LinkedIn
Previous education	Facebook, Google+, LinkedIn
Family	Facebook, Google+
Checkins	Facebook
Nickname	Facebook
Life events	Facebook
Projects	LinkedIn
Courses	LinkedIn
Skills	LinkedIn
Circles	Google+

TABLE I
LIST OF ATTRIBUTES IN A SOCIAL FOOTPRINT

A. 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 is the choice language to integrate with these OSNs via APIs as all of them provide comprehensive documentation to integrate with their services. Also, these technologies being Open Source is a contributing factor in using them. The authors do not have prior experience with PHP which is a significant challenge. The MySQL database can be used for historic data and statistic collection. However, our web interface does not require interactions with the database. Figures 1 and 2 illustrate the use of the interface. A user will access the web application and then provide his authentication for the three OSNs that he uses. This is needed to call the API to fetch that users data. We do not store his login credentials in any way and simply use the third party authentication service provided by these three OSNs. The user is then prompted for permission to access his public profile. On providing this permission, the user allows us to view his public profile attributes. We then build a union of attribute lists of all the three OSNs. This would adhere to the definition of social footprint that we mention in the previous section. Table 1 lists attributes and the OSN they can be retrieved from. We have built this list manually and may not be exhaustive.

One observation that we can make here is that while some of the attributes at the bottom of the table are unique to some of the sites, a majority of the attributes are common to the three

Fig. 1. Facebook Login Page



sites. This reinforces our motivation of a user accidentally leaking a comprehensive profile across various sites. While a user might be careful of protecting attributes on one site, it is extremely likely that the user may have leaked that same attribute information unintentionally on another site. This is how a social footprint might be exploited. We will revisit some of these possibilities in Section 4.

B. Weight assignments

We would like to assign weights to attributes based on the risk associated with divulging that attribute. Prior works assign these weights based on intuition which rely on the same idea. We propose that these weights should be based on how the attribute disclosure has already harmed users historically or statistically indicate that users could be harmed by disclosing that attribute. That is, we infer that attributes that are risky are the ones that can be used to harm the user in some way. The information held by those attributes would be of some private value and divulging them would not be safe for a user. Now, depending on the extent of the harm that may be caused by divulging an attribute, we give a weight to that attribute. In short, an attribute which, by being divulged in public causes more harm to a user, has a higher weight than the others.

We now talk about what can come under the definition of harm. As mentioned in the previous section, various crimes are tied to online social networks. Attackers use different tools, one of which is the use of information available on social networking sites. We explore two possible attacks - identity theft and password recovery. We will assume every attribute to contribute a maximum of $unitweight = 1$ to every footprint. This will give us a maximum weight of n for a footprint when we have n total attributes spanning across all the OSNs included in our study. In this work, we have considered a total of 23 attributes, making the heaviest possible footprint weight 23. However, realistically we will never assign a weight of 1 to any attribute as we would like to provide some allowance for visibility. The weights of all attributes will range from $0 \leq weight < 1$. Different attack possibilities will come with different weights and their sum total will be the total associated risk of displaying that profile attribute. That is,

$$w(attribute) = \sum_{i=1}^n attack_i$$

where the the two attacks we consider are, $attack_1$ as identity theft and $attack_2$ as password recovery. We will scale the weights of these two attacks for our work. Ideally, the weights will continue to be more accurate as we include more and more attacks in our analysis. So we can describe it even more simply as

$$w(attribute) = weight\ due\ to\ password\ recovery\ risk \\ + weight\ due\ to\ identity\ theft\ risk$$


1) *Password Recovery*: Most websites require users to create accounts. This provides a certain level of personalization with a user. On creating an account, the user chooses a username for the website and a password which would serve as his login credentials. However, passwords can be lost or forgotten. And moreover, in this age when there are so many different websites each requiring their own login, users tend to forget passwords often. According to statistics in [20], 80% people have forgotten passwords at some point. Also, close to 60% survey participants admitted to having the same password. This shows the need to have password recovery mechanisms. For this reason, many sites provide a mechanism to retrieve a password. In this process, users are asked to choose a security question, an answer of which would only be known to them. These questions are a general popular set appearing across various websites. Table 2 shows a list of some popular security questions.

The arrow between attributes signifies a second step needed to reach an answer. As shown in the table, which is a list of popular questions, almost all questions can be answered by an OSN profile attribute. So we can now outline the attack method. An attacker who would want to break into another user's account would need to guess the password. Instead of guessing the password, an attacker can simply invoke the the recovery mechanism and attempt to answer the security question. A famous example is the Sarah Palin case where the attacker simply answered security questions from social footprint attributes. Examples of fields used in password recovery questions are mothers middle name, high school name, etc. Existence of any one of these fields in a social footprint is enough to compromise the online account.

In order to assign weights to attributes keeping in mind the password recovery attack, we will use a recent survey paper by Bonneau et.al [15]. This paper closely surveyed the current state of password recovery questions and concluded that this

Fig. 2. User Data retrieved from Facebook

FACEBOOK




[Profile URL](#)
Name: Hamza Karachiwala
Gender: male
Email Id: hskarachiwala@gmail.com
Hometown: Pune, India
Current location: Gainesville, Florida
DOB: 1991/09/11
Languages: HIDDEN
Relationship Status: Single
CheckIns: HIDDEN

LINKEDIN



[in Profile URL](#)
First Name: Hamza
Last Name: Karachiwala
headline: Graduate Computer Science Student at the University of Florida
User ID: HIClvbmYmG
Location: Gainesville, Florida Area
Industry: COMPUTER SOFTWARE
Summary: Graduate student in the Department of Computer Science at the University of Florida. Expecting to graduate in December 2016. Currently have two years' industry experience in web application development.
Email Address: hskarachiwala@gmail.com

Google+



[Profile URL](#)
Name: Hamza Karachiwala
NickName: HIDDEN
birthday: HIDDEN
Gender: male
Email: hskarachiwala@gmail.com
Current Location: HIDDEN
Language: HIDDEN
Places Lived: Gainesville
Relationship Status: HIDDEN

vulnerable

Security Questions	Possible Answer Attribute	OSN
Mother's maiden name?	Family	Facebook, Google+
City of birth?	Previous City	Facebook, Google+, LinkedIn
Name of wedding reception venue?	CheckIns	Facebook
Person you first kissed?	Relationship Status	Facebook
Manager at first job?	Projects	LinkedIn
Name of primary school?	Education	Facebook, Google+, LinkedIn
Where does sibling live?	Family → Current Location	Facebook, Google+, LinkedIn
Vehicle Registration Number?	Life Event	Facebook
Pets Name?	Photos / Updates	Facebook, Google+
Year Father was Born?	Family → DOB	Facebook, Google+
Favourite restaurant?	Check In	Facebook
Make of first car?	Life Event	Facebook
High school mascot?	Education → Online search	Facebook, Google+, LinkedIn
Favourite Actor?	Likes / Interests	Facebook, Google+
Favourite Musician?	Likes / Interests	Facebook, Google+
Library Card Number?	-	-
First Telephone Number?	Phone	Facebook, Google+, LinkedIn
Childhood best friend?	Friends → Check among all	Facebook, Google+

TABLE II
SECURITY QUESTIONS AND POSSIBLE ANSWER ATTRIBUTES

mechanism is not reliable. They also briefly explore the attack avenue outlined in our work but not in detail. In the process of their work, they carried out extensive surveys related to password recovery questions. Along with this, another major source of our statistics is the work of Rabkin [14]. Rabkin has provided some information on 200 sample password recovery questions. While the motivation of his work was different, we can nevertheless make use of the numbers aggregated in his work. Some of the important statistics that we will use are -

$$\begin{aligned}
 \text{Name based questions} &= \frac{70}{200} = 0.35 \\
 \text{Favourites based questions} &= \frac{36}{200} = 0.18 \\
 \text{Specific personal questions} &= \frac{14}{200} = 0.07
 \end{aligned}$$

What this means is that from a set of 200 questions, 70 were answerable with a name (first, middle or last). What we

will interpret from this is that this is the probability of a user selecting a name based security question on any website. These statistics are not comprehensive, and hence, this heuristic may not be very accurate, but it helps us determine in a way which questions are likely to be chosen. And subsequently, if a user adopts those questions and at the same time has attributes that may answer those questions, then those attributes contribute in some way to the risk. We have deliberately kept these contributions small, as password recovery is a less serious threat in comparison to identity theft. So we can rewrite the above statistics as

$$\begin{aligned}
 P(\text{Name}) &= 0.35 \\
 P(\text{Favourites}) &= 0.18 \\
 P(\text{Personal}) &= 0.07
 \end{aligned}$$

This gives us the probability of a certain type of question being used as a security question by any user. Apart from this another statistic that we will use is that for all security

questions, 65% of the answers are true. That is, the probability of a user saving a correct answer for a question is $P(\text{correct}) = 0.65$. The consequence of this is that if a user does not save a correct response for his security question, then an attacker cannot in any way extract the answer from a social footprint. So we must assume that the weight of an attribute should depend on both, the probability of a question being selected and the probability of a correct answer being used. The correct answer here means that it would match with an attribute in the footprint. We reassign the probabilities above as

$$\begin{aligned} P(\text{NameCorrect}) &= P(\text{Name}) \cdot P(\text{Correct Answer}) \\ &= 0.35 \times 0.65 \\ &= 0.23 \end{aligned}$$

$$\begin{aligned} P(\text{FavouritesCorrect}) &= P(\text{Name}) \cdot P(\text{Correct Answer}) \\ &= 0.18 \times 0.65 \\ &= 0.12 \end{aligned}$$

$$\begin{aligned} P(\text{PersonalCorrect}) &= P(\text{Name}) \cdot P(\text{Correct Answer}) \\ &= 0.07 \times 0.65 \\ &= 0.04 \end{aligned}$$

We will manually map attributes to one of these three categories and assign them a corresponding weight. This weight will be scaled in our work as we consider only two attacks, but this scaling is proportional to the probabilities mentioned above. Attributes that do not fall in these categories will not have a contribution to footprint weight in terms of password recovery attacks.

2) *Identity theft*: One attack that immediately comes to mind on seeing a collection of personal information is identity theft. A section of the attributes in Table 1 can be used to not only clearly identify a user but also identify him. In this attack avenue, we base attribute weights on their identification character. Consider the concept of cross-site identifiers as outlined in [1]. There are some attributes which would not be heavy independently, but could be used to match profiles of a single user across multiple OSNs. Or in other ways, these attributes would have more value as identifiers rather than profilers. An example is when, on their own, gender, age and location might be benign, but when combined with name, it can be used to get info about the user from multiple OSNs to create a more complete social footprint. In the case of the password recovery attack, we assigned weights based on the likelihood of an attribute answering a security question. In this attack, we will be giving weights to attributes based on the likelihood of them identifying a person and subsequently being used to impersonate that victim.

Being able to guess a Social Security Number would be one of the ways of committing an identity theft. Also, for identity theft, we solely consider the risk of obtaining users

private SSN numbers based only on their publicly available data. This is an accurate assumption for identity theft because SSNs are the most widely used sensitive authentication devices, and its the most commonly sought after piece of information by identity thieves [10]. Its interesting to know that 34% of identity fraud cases in the US happen in Florida itself [21].

This mode of attack has been explained very concisely in [16]. The risk lies in making birthdate, hometown and current residence publicly available at the same time. In a social footprint these attributes are likely to appear. The first three digits of a social security number reveal where that number was created. Specifically, the digits are determined by the ZIP code of the mailing address shown on the application for a social security number. This is put at risk by disclosing the current location in the social footprint. The next two digits are group identifiers, which are assigned according to a peculiar but predictable temporal order. When that persons birthday is also known, and an attacker has access to SSNs of other people with the same birthdate in the same state as the target, it is possible to pin down a window of values in which the two middle digits are likely to fall. The last four digits are progressive serial numbers. The last four digits can be retrieved through social engineering. Here in lies another potential threat. Social engineering requires a lot of profiling, the effort of which is saved by simply analyzing a social footprint. In order to build trust with a victim attributes related to occupation, work place, education, etc. can be very useful. We can term these attributes as *profilers* - attributes which can be used to describe a user as opposed to identify him.

In order to build weighted relations of attributes for this attack, we will follow a similar approach as in the previous attack. We will base our proposals on the statistics found in [16]. We propose the following relations

$$\begin{aligned} w(\text{date of birth}) &\sim w(\text{home town}) \\ &\sim \text{Primary Risk Attributes,} \\ w(\text{cell phone number}) &\sim w(\text{location}) \sim w(\text{email}) \\ &\sim \text{Secondary Risk Attributes} \end{aligned}$$

What this means is that the Date of Birth and Hometown attributes are more important when one tries to guess an SSN. On the other hand, the contact number, location and email serve as auxilliary attributes. they would expedite the guess but are not essential. Due to this classification, we reasonably assume that the mean *Primary Risk Attribute Weight* $\sim > 4 \cdot (\text{Secondary Attribute Weight mean})$. We are giving an 80% to 20% severity ratio here to highlight the fact as we would not want it to be lost in the total footprint sum.

Apart from this, we now consider the concept of cross site linking. What cross site linking can achieve for an attacker is to fish a profile from an OSN even if it has low visibility. That is, in case the name does not match on one OSN with the others,

Attribute	Weight
Gender	0.02
Date of Birth	0.32
Email	0.15
Phone number	0.08
Lives At	0.28
Previous cities	0.20

TABLE III
IDENTITY THEFT

an attacker can use certain attributes to guarantee a unique identification of a victim. To incorporate a weight model for this aspect of the identity theft attack, we consider statistics and observations in [1] and [16]. These works have ordered the importance of some attributes as cross-site identifiers. These can be represented as $w(date\ of\ birth) > w(home\ town) > w(location)$. We propose to assign weights to them in this decreasing order as listed. Another statistic we use to back our weight assignments come from [2], we know that the 53% of the US population can be identified using the attributes *date of birth*, *gender* and *location* attributes alone. This being strong identifiers, need to be considered strongly for an identity theft attack.

$$w(date\ of\ birth + sex + location) > \\ w(Primary\ Attribute\ Weight\ Mean)$$

We consolidate the above assumptions into a relative ordering of weights as follows,

$$w(dateofbirth) > primaryAttributeMeanWeight > \\ w(hometown) > w(age) > w(location) > \\ w(cellphonenumber) > w(email) > \\ w(gender)$$

Using this relations between the attributes, we use the balance beam approach to obtain their respective normalized weights as depicted in table 3. These are a limited set of attributes as the remaining from our set do not contribute to an identity theft attack. The same goes with the previous attack. As mentioned before, considering more attacks spanning more statistics would make this study more inclusive of all attributes in a social footprint.

Based on the methods of assignment described above, we have assigned weights to a subset of the attributes in Table 1. Because we have two attacks, we scale their base weights by 50%. As and when we consider more attacks, the effect of each attack will be included in the final weight using such a scaling. These numbers will be used to calculate the total footprint weight. Once again, we keep in mind that name and profile picture would be required to provide basic visibility on these sites.

Attribute	Weight
First Name	0.25
Last Name	0.25
Profile Picture	0.20
Gender	0.04
Date of Birth	0.34
Email	0.20
Phone number	0.12
Marital status	0.02
Lives In	0.36
Previous cities	0.18
Works at	0.13
Occupation	0.13
Studies at	0.13
Previous work	0.13
Previous education	0.13
Family	0.13
Checkins	0.06
Nickname	0.06
Life events	0.02
Projects	0.02
Courses	0.02
Skills	0.06
Circles	0.06

TABLE IV
WEIGHTS FOR 2 ATTACKS - PASSWORD RECOVERY AND IDENTITY THEFT

C. Footprint Weight and Threshold value

After having a list of assigned values for different attributes as depicted in Table 3, we then calculate the total weight for that user's footprint. We must check what attributes are present in the user's 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,

$$w(footprint) = \sum_{i=1}^n attribute_i$$

After we calculate the weight of the footprint, we compare it to a threshold value to check if it is vulnerable. For our threshold, we consider a very strict minimum value that accounts for visibility. We believe that a name and profile picture would be sufficient to uniquely identify a user on the network (we do not consider the case of fake accounts). From our table, we have found these weights to be $weight_{name} = 0.3$ and $weight_{profilepic} = 0.2$ giving us a total of $weight_{name} + weight_{profilepic} = 0.70$ while the maximum total weight could be 2.80. That is, we allow approximately an 30% weight of total. This will be a simplistic threshold for the experiments we carry out on our initial users. While this is crude as compared to some previous works as in [6], it still takes into account the need for visibility on these sites. Apart from this, it is also safe to assume that any excess attribute as public information

continues to be a risk. While our proposal may be amplifying the risk by keeping a strict threshold, it still holds with this ground truth that needs public profile attributes to be minimum.

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. With our current model, we will be listing all the attributes apart from the name and profile picture in decreasing order of weights. However, once we consider visibility with more accuracy, we can include more than two attributes to be allowed on a public profile.

IV. RESULTS AND ANALYSIS

We have conducted a few experiments on a limited set of 23 users. This is largely due to a time constraint and the difficulty in usability of our interface. As our initial intention was not to build a sophisticated UI, it will need some more work to entice users to adopt it. In this section we list out some of our observations and our opinions about these results. For all the users, Facebook was the main contributor. In one way this is surprising, as although Facebook is the most popular OSN, we were expecting users to be more careful about their attributes disclosed on Facebook. However, in case of the same attribute appearing on all three websites, we pick the source OSN as the one into which the user logs in first. In this case, it is highly likely that the user will pick Facebook given its placement on our interface, hence giving us the impression that Facebook is a major contributor. Nevertheless, we feel that Facebook is still a rich contributor of footprint data.

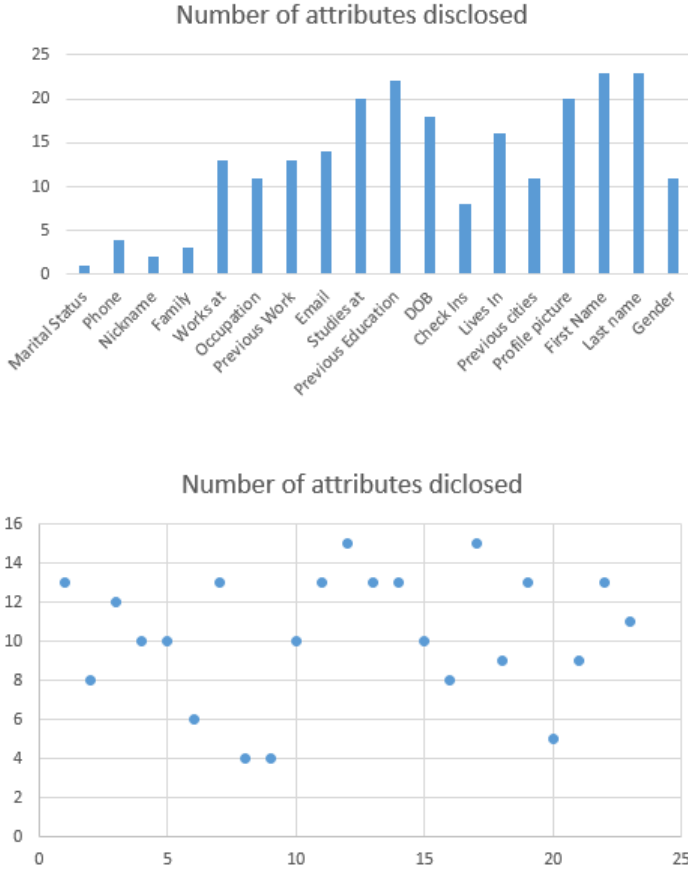
Of the 23 users we found that 22 users had risk higher than threshold associated with their profile. The one user not at risk had only a Facebook profile and no profile picture on it. The reason this number is high among remaining users is partly due to our strict threshold which allows only two attributes. While this approach will highlight most users as vulnerable irrespective of the extent of risk, the variations in the sum total of the footprints shows how some users allow more attributes to leak than the others. By enlisting all the attributes aggregated together with weights we are also able to clearly depict the need of privacy to a user. We observe that *current location city* attribute is the most common attribute among all users having the highest weight. This could mean

one of two things - either that many users are not aware of the risk associated with disclosing that attribute, or it is one that many users feel are needed to contribute to visibility. Such decisions can be made only by collecting more data and having a smarter way of calculating the threshold.

We list out some more observations here, based on results for our set of users that is illustrated in the chart

- The average weight of all user footprints was 1.59 depicting approximately 57% of risk weight for our users. We have our threshold currently at around 18%.
- There was more than one group of users having the same footprint weight. This indicates a possibility that a number of users disclose the same set of data publicly. This is likely due to the fact that they do not alter the initial privacy settings of their accounts. Numerous surveys have been carried out regarding this and the outcome remains that OSNs need to make their users more aware about their privacy. There is quite some progress done regarding that on major OSNs.
- Many attributes are dependent on each other, that is, users that have disclosed one attribute have more likely disclosed another one. For example, users having prior experience disclosed Occupation, Works At and Previous Work. It would be interesting to identify such groupings in order to make the task of reducing footprint weights easier.
- As expected, name and profile picture appear in almost every profile. This justifies our decision to not consider it with risk, as most users believe that it is required. After all, visibility is extremely important on social networking sites.
- Some of the attributes appeared on only one site. In such cases it would be prudent to inform the user to simply remove that attribute irrespective of the weight as it was not required on other sites and hence, may just be there accidentally.
- The average number of attributes disclosed by users is between 10 to 13. It is highly unlikely that most of the users would be unaware of all these attributes being disclosed. In fact, it is more plausible that the users intend to increase visibility by disclosing those attributes. This statistic would also be helpful to gain some more insights on users' perception of visibility and the need of so many attributes to be visible.
- Email was an attribute that more than often was picked from LinkedIn. This is fair as users on LinkedIn would want to be reachable via Email, it being a professional site. However, this does not mitigate the risks associated with disclosing the Email ID. Various websites use the Email ID as a username. That being said, this is another instance of the tradeoff between visibility and risk, and one that is hard to resolve in favour of any direction.
- Some statistics are biased in favour of the user group that participated on our experiment, which was primarily

classmates or students. An example of this bias is the number of users that have disclosed education details. However, we can use this bias to our advantage by identifying different groups of users and deciding their threshold accordingly. Attributes like place of study would be common among students, and dangerous for them to divulge as well, as they can be expected to visit educational campuses often.



V. FUTURE WORK

There is much scope for improvement in our approaches which would demand more time and statistical resources. While we have integrated with Facebook, Google+ and LinkedIn, there are many more OSNs which would be worth investigating in terms of attributes. Twitter, for example, is another example of a very popular social networking site and works on a different dynamic. One would, however, argue that the public profile attributes remain the same across these OSNs. While this is true, certain sites may contain specific content that could be more helpful towards profiling. Politicians followed on Twitter, music preferences on SoundCloud, artistic interests on Pinterest, image content on Instagram, video content on YouTube, etc. are some examples of OSNs that cater to very specific kinds of content. In some cases the entire profile content may be public,

which makes the user very vulnerable to profiling and social engineering. These avenues are interesting to explore for this reason. Consequently, all user content is visible to connections on a site. By connection, we mean a friend on Facebook or a circle member on Google+, etc. We assume these connections to be trustworthy, but in the event that a connection is malicious, it gives the attack an entirely different magnitude of potency.

For our weight assignments, there are possible improvements as also different approaches that we may take. While we currently do consider password recovery attacks, in the near future it is likely that such question based systems will be obsolete [15]. Even within the weight assignments for password recovery attacks, we could collect more data related to security questions such as a larger question set and question popularity in order to make our assignments more accurate. In terms of safety threshold, we have a trivially defined value in our system. If we focus on the tradeoff between risk and visibility, we currently have not worked on formalizing visibility in the same way as risk. Hence, it is essential to be more accurate to decide if an attribute is required for visibility or if it should be considered risky. In order to understand that, we would need to perform more experiments and conduct more surveys to understand the visibility requirement in a public profile. On gaining more knowledge about visibility, we could come up with a better proposal in calculating a true threshold beyond which a profile could be considered unsafe.

The suggestions that we make to reduce footprint weight are currently ordered in decreasing attribute weights. This is a greedy heuristic which should be quite accurate, under the assumption that every unsafe attribute should be removed. However, it is possible that some attributes may be important to visibility and should be accommodated in a public profile. Or it could be possible that a set of attributes together would be more risky rather than them appearing individually. For example the attributes *gender*, *age* together is riskier than simply having *gender* alone. Such combinations can be explored more as a follow up to the work on visibility in public profiles.

Finally, from the point of view of usability and as a product, we can improve the web interface and deploy this solution as an online web application. The interface is simply designed to pull data currently, and does not seem very trustworthy. However, we can have a more sophisticated User Interface with instructions and disclaimers to explain the motive of the interface more clearly to users. If the user is assured that we will not collect their data, he or she would not have any issues with using this interface. From a statistical point of view, what we could store in the database would be the number of users found at risk, most common risky attributes, etc. Such statistics would be helpful for these OSNs to enforce their own privacy measures. At the same time, these would also be

important numbers that could help make our measurements in experiments more accurate. Eventually, the more outreach we achieve, it will help in developing a better understanding on what users consider safe or risky.

VI. CONCLUSION

We attempt to show the risk in disclosing excess information on our public profiles on online social networks. From the limited statistics that we collected, it is quite evident that a user would not be aware of the comprehensive extent of information disclosed by him. We can show how the information safeguarded on one website, is readily available on another. This makes the collective social footprint of a user always susceptible to certain attacks by malicious entities. We cover two such attacks - identity theft and password recovery. Based on the vulnerability of public attributes to these attacks we verify if the collective social footprint of a user is at risk. We build a social footprint of an individual by accessing his public profile attributes available on three major OSNs - Facebook, Google+ and LinkedIn, and we aggregate these attributes. We assign weights to a majority of the attributes available on these sites, with respect to the amount of risk associated with them. We will then calculate the user's footprint weight by summing up the weights of all attributes disclosed by him. 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 profile information. Following this we can provide suggestions to reduce the public data disclosed by a user and encourage the user to trim the weight of the footprint. This will help achieve a certain degree of privacy on these ubiquitous social networks.

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] Nilothpal 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] Alessandro Acquisti and Ralph Gross: *Predicting Social Security numbers from public data*
- [11] Kun Liu, Evimaria Terzi: *A Framework for Computing the Privacy Scores of Users in Online Social Networks*
- [12] Justin Becker, Hao Chen: *Measuring Privacy Risk in Online Social Networks*
- [13] L.Sweeney: *Uniqueness of Simple Demographics in The US Population*
- [14] Ariel Rabkin: *Personal knowledge questions for fallback authentication: Security questions in the era of Facebook*
- [15] Joseph Bonneau, Elie Bursztein, Ilan Caron, Rob Jackson, Mike Williamson: *Secrets, Lies, and Account Recovery: Lessons from the Use of Personal Knowledge Questions at Google*
- [16] Ralph Gross, Alessandro Acquisti: *Information Revelation and Privacy in Online Social Networks (The Facebook case)*
- [17] Statistics Portal at statista.com
<http://www.statista.com/topics/1164/social-networks/>
- [18] Demographic Distribution os OSN users
<http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/>
- [19] Socialnomics
<http://www.socialnomics.net/2014/03/04/the-shocking-truth-about-social-networking-crime/>
- [20] Password Statistics
<http://passwordresearch.com/stats/statindex.html/>
- [21] Risk for identity theft
<http://www.bbb.org/blog/2013/06/identity-theft-on-social-media-are-you-at-risk/>