Social Footprints - A Comprehensive Analysis

Hamza Karachiwala, Kedar Amrolkar, Mickey Vellukunnel

September 22, 2015

1 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[12] 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 every 3 out of 4 persons can be found online if you look!

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[13] 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. It is of prime importance to reach out to users and explain this hard-to-visualize problem in a simplistic and intuitive way.

2 Scope

1. We intend to build a users social footprint by aggregating public and personal data disclosed on -

   - Facebook - Users share a lot of information on Facebook. It being the most popular social network having the most users, the average time a person spends on Facebook is quite large. We could have access to a persons name, address, date of birth, profile picture, place of work, place of study, close friends and relatives, favourite places to visit, emotional situation, financial situation, etc.
   - Twitter - Basic profile information as with Facebook along with list of followers, list of people followed, political ideologies, sensitive opinions, etc.
   - Google+ - While Google+ is not as widely adopted, its integration with Google which is the biggest search engine may link profiles with different types of data. This needs to be explored as well.
   - Pinterest - This site is different from the rest as it focuses on a variety of personal interests and choices. Coupled with image sharing, it is a rich source for profile building which can be used for social engineering.
   - LinkedIn - Being the most popular professional network, most data here is public. There is more likelihood of disclosing contact information on this site.
   - Instagram - This popular photo sharing app has become a single destination for user images. Users share photos on Instagram disregarding privacy settings as they do not realize the risks involved.
   - YouTube - This could be explored although there may not be as much useful information as could be retrieved from the other sites above. However personal videos have their own value when it comes to harvesting data as it provides a source to observe personality, mannerisms, etc. first hand.

2. We then propose to calculate the vulnerability of a footprint. By analyzing the weight of a footprint, we want to quantify how much risk is associated with the aggregated public profile.

3. We then would like to identify ways to reduce the weight of a social footprint.By identifying data that is unnecessary or outright dangerous, users need to be suggested to take down some sections of their public profiles.

3 Approach

Social networking sites provide integration endpoints for third party applications. Larger social networks often play the role as an upstream system of data for smaller applications. Or in many cases, they simply provide a simple means of authentication. Once integrated with a particular OSN,these applications can access a users data based on the permissions granted by that user.

- We will build a simple application that will integrate with some popular sites and request users to provide us access to their public profile. We will initially work with data visible on

the public profile only and build aggregations on this data. There may be a few unknown obstacles in accessing data via APIs in this fashion but we are optimistic about the integration capabilities provided by the websites. By the end of *week 3* starting from the date of proposal submission, we aim to integrate APIs from at least three OSNs.

- Once we obtain the data, we intend to build a build a weighted math model as depicted in [2] from the data based on empirical weights associated with all bits of data. For example - disclosing an address is riskier than disclosing the middle name OR disclosing the nickname-birthdate combination. This risk associated with data can be understood by observing patterns of crime and misdemeanours that have taken place over the last few years in connection to SSNs, as well as robberies, kidnappings, credit card fraud and online identity theft, all which had their roots from a user's unwittingly extensive online social media footprint. Within the timeframe, we will analyze as many instances of cyber crimes reported to derive weights for individual fields of profile information. We are aiming to propose weights for personal data by the end of *week 6*.

- Based on the model derived above,we will then propose a threshold value for social footprints that will quantify how much that user is at risk with the amount of data disclosed in public. As a simplistic approach, we would like to simply identify a limit in weight of the social footprint. However, it might not be as simple due to the range and variety in data available. But the aim is to be able to provide a reliable metric to determine the risk element in publicly visible online content. By the end of *week 9*, we will be looking to propose a mathematical model associates with ones social foorprint and the calculate the risk involved with it. We will build upon work done in [6] and [2].

- We can then provide suggestions or ways to the users to reduce these weights via a readable graphical interface along the lines of previous tools [11] and [8]. From the users perspective, all he would be interested in is privacy and safety. This interface should be intuitive and make it relatively simple for a user to understand the problem at hand. And at the same time, it should suggest a solution, which in this case could possibly be a list of data that is largely contributing towards footprint weight. We are aiming to achieve a large magnitude of users but the timeline constraints may not let us achieve the numbers as in [1] or [3]. We will begin with a comparatively small user base and data set to start with and continue to scale on this base as much as possible. We attempt to have a user friendly interface for the same by the end of *week 12*.

4 Literature Survey

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 rely on intuition to weigh leaked profile attributes. We will attempt to provide weights by studying certan cyber 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 represenation by Irani et al [2] of a social footprint as τ_s^u is very intuitive. They then create an aggregate represenation $P^u = \bigcup \tau_i^u$ for the footprint spanning various sites. Another notable mention is the *PIDX* proposed by Nepali et al. [6]

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

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.

5 Conclusion

We attempt to highlight the leakage of profile attributes on online social networks. We will build a social footprint of an individual by accessing his public profile data across various OSNs. 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 cyber crime patterns related to OSNs over the past few years. By comparing this footprint to a range of threshold values, we will then determine the level of risk a person is subject to with his current OSN public profile data. Following this we can provide suggestions to reduce the public 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] 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] 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] Statistics Portal at statista.com
`http://www.statista.com/topics/1164/social-networks/`

[13] Socialnomics
`http://www.socialnomics.net/2014/03/04/the-shocking-truth-about-social-networking-crime/`