Large Online Social Footprints - An Emerging Threat

Danesh Irani [#1], Steve Webb [*2], Kang Li [&3], Calton Pu [#4]

[#] *College of Computing, Georgia Institute of Technology*
225 North Ave NW, Atlanta, GA 30332, USA
[1] danesh@cc.gatech.edu, [4] calton@cc.gatech.edu

[*] *Purewire, Inc.*
14 Piedmont Center NE - Suite 850, Atlanta, GA 30305, USA
[2] swebb@purewire.com

[&] *Department of Computer Science, University of Georgia*
415 Graduate Studies Research Center, Athens, GA 30602, USA
[3] kangli@cs.uga.edu

***Abstract*—We study large online social footprints by collecting data on 13,990 active users. After parsing data from 10 of the 15 most popular social networking sites, we find that a user with one social network reveals an average of 4.3 personal information fields. For users with over 8 social networks, this average increases to 8.25 fields. We also investigate the ease by which an attacker can reconstruct a person's social network profile. Over 40% of an individual's social footprint can be reconstructed by using a single pseudonym (assuming the attacker guesses the most popular pseudonym), and an attacker can reconstruct 10% to 35% of an individual's social footprint by using the person's name. We also perform an initial investigation of matching profiles using public information in a person's profile.**

I. Introduction

Social networking sites are big. MySpace and Facebook both have over 250,000,000 accounts [1], [8] and are still growing at a rapid pace [3]. As a social network gets larger, they attract more users and share even more information.

Unfortunately threats of private information leakage increase along with the growth of social networks. As a prerequisite to participate in most social networking sites, a user has to create a profile, enter some basic information, and is encouraged to enter detailed information relating to the purpose of the site. For example, Last.fm is a music listening service and it encourages users to enter details about their favorite artists. Due to the large number of social networking sites currently out there and the increasingly social nature of the web, most users belong to more than one social networking site. They assume that the information provided will be kept within the boundaries of the social networking site, and that the privacy policies across sites are standard.

The danger of this implicit assumption is that many users don't realize how much information could be revealed by blurring the boundaries of these social networking sites. One danger in blurring these boundaries is that any information disclosed at one site, could be combined with information at other social networking sites. We call the resulting combination of the information revealed by multiple social networking sites, a user's *online social footprint*.

We study the threats associated with an online social footprint by leveraging data collected from an online-identity management site. Online-identity management sites allow a user to provide links to all their social network profiles. We crawled one such site retrieving over 13,990 profiles and 80,357 potential links to social network profiles. We use this to preform quantitative measures in respects to the size of a person's online social footprint and the ability of an attacker to reconstruct such a footprint, given that such sites did not exist.

From this study, we have two main contributions:

- *Measure the size of a user's online social footprint* - We find that an active member has on average 5.7 profiles of which 1.6 profiles are in the top 15 social networking sites, which we retrieve for further analysis. Based on a sample of 9 fields, we find that a person's online social footprint size increases from an average of 4.3 fields to 8.25 fields, when the number of profiles in a person's online social footprint increases from 1 to 8 profiles.
- *Investigate how easy it is to reconstruct a user's online social footprint* - We show that an attack on targeted individuals is possible without having to use network based de-anonymization techniques [9], [10]. Our first method involves assuming an attacker has prior knowledge of a single pseudonym and measuring the amount of other social networking sites he can find. Our results show that an attacker would be able to find over 40% of a person's other social networking sites for most of the data-set in the best-case (assuming the attacker knows the most-common pseudonym). The second method involves assuming an attacker knows the person's name. In this case, the attacker will try and guess a person's pseudonym. In this case, we are able to find approximately 10-35% of a person's other social networking sites. Although this method might yield a few false-positives, we evaluate a technique which could be used to calculate the likelihood

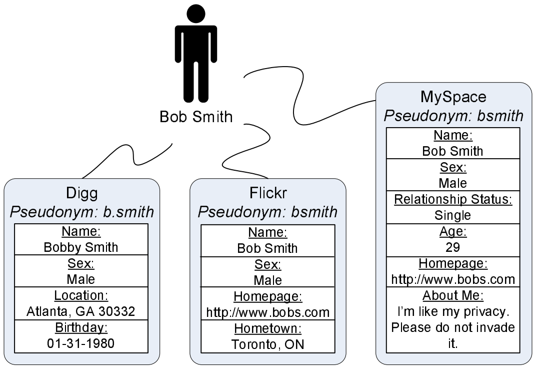

Fig. 1. Online Social Footprint

of two profiles belonging to the same user.

We expect that our findings will increase the awareness of the threat caused by large online social footprints and promote protection mechanisms against this threat.

II. Background

A. Online Social Footprints

A user's online social footprint is the online information that is available about him by aggregating his social networking profiles. Essentially, this footprint characterizes a user's social networking activities. To illustrate, Figure 1 shows a user named Bob Smith and his online social footprint, which was constructed with information from three social networking sites (identified by two unique pseudonyms). Individually, each of these sites reveals between 4 and 6 pieces of information (e.g., age, sex, etc.); however, if the sites are linked together, 8 pieces of information are displayed about Bob. This combined view of Bob's information represents his online social footprint.

B. Threats of Information Leakage

Threats associated with Information Leakage have been discussed in previous research [4], [5], but we briefly describe them here for completeness.

- *Online stalking* - As users spend more time online and reveal more information about their activities, they become more susceptible to digital stalking. Using Twitter, stalkers can view "tweets" about their victims' current activities. On Last.Fm, stalkers can determine what music their victims are listening to. Utilizing Del.icio.us, stalkers can identify the surfing habits of their victims.
- *Compromising personal accounts* - Most login-based Web sites allow users to "recover" their passwords by answering some personal questions about themselves. Answers to common questions such as date of birth, address, or hometown are sometimes inadvertently made public on social networking sites. A recent example of this type of compromise occurred when Sarah Palin's e-mail account was hijacked because the recovery question for her Yahoo account was discovered. For further details on the types of password recover questions, please refer to our report [7].
- *Customized spam/phishing* - Occurrences of spear spam/phishing have already been observed, and with an abundance of personal information, such techniques can be made more effective to the point where the user might be fooled into believing that the email is legitimate due to the amount of personalized information it contains.

III. Online Identities and Data Collection

Identity management sites allow users to manage their online identities by enabling them to provide links to their social networking sites. One can use an online identity to determine a user's social footprint by visiting each profile that is associated with the online identity and identifying the pieces of information that are revealed by each profile.

In August 2008, we crawled the publicly available online identities stored by one such identity management site. During this crawl, we collected 54,600 users' online identities, and of those identities, 13,990 were labeled active (i.e., the identity contained one or more links to a profile on a social networking site). From the 13,990 active identities, we found 80,357 links to social networking profiles. Then, we identified links pointing to the top 15 most common social networking sites, which accounted for 21,764 profile links. Next, we proceeded to crawl the profiles associated with each of these links, obtaining a total of 21,764 profiles. The distribution of the number of profiles crawled for each site is shown in Table I.

TABLE I
Number of profile links.

Social Site:	*# of Profiles*
Blogspot	2625
Del.icio.us	1959
Digg	759
Facebook	1840
Flickr	3646
Last.Fm	1540
LinkedIn	1879
LiveJournal	645
MySpace	1677
Technorati	497
Tumblr	607
Twitter	1901
Wikipedia	280
Wordpress	926
YouTube	747

The profiles and links were entered manually by the users of the identity management site. However, some of the links were not associated with the user entering the data. For example, a number of the links pointed to profiles belonging to the users' friends or celebrities. In a few extreme cases, the links formed a link farm that was used to promote other sites/profiles. To minimize the effect of these links and maintain an accurate representation of a person's online identity, we removed these unrelated links by implementing a few heuristics. For example, we removed any links that were duplicated across profiles because we had no way of knowing which profile was the legitimate owner of those links.

To investigate the amount of information leaked by this

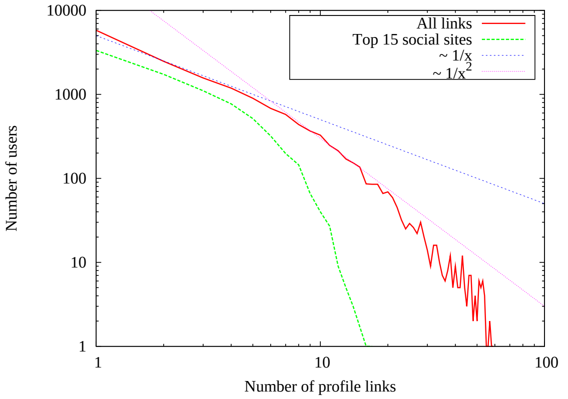

Fig. 2. Number of users against size of social network footprint

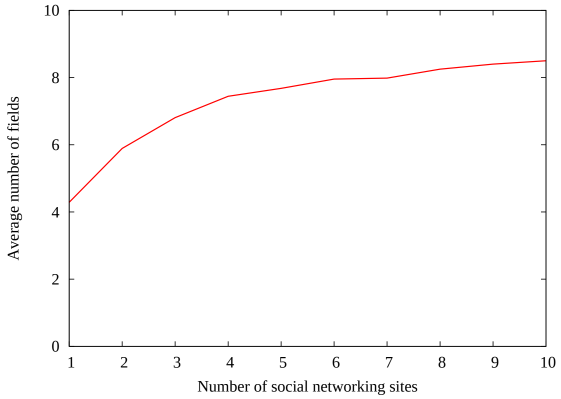

Fig. 3. Average size of an online social footprint increasing with number of social network sites

dataset, we wrote parsers for 10[1] of the top 15 social networking sites. The parsers performed a great deal of post-processing to merge fields with semantically similar meanings but syntactic differences. An example of this is "Age" and "Date of Birth", as Age can be represented as a coarse granularity date of birth.

Table II shows the amount of information parsed from each profile for a small subset of available fields. For each field, we represent the amount of information as a percentage of the number of profiles that displayed this value on a particular social networking site. If a '-' is present, it means the field was not revealed by the site in question.

IV. Size of a user's online social footprint

From the data we collected, the number of users with profiles on multiple social networking sites follows a Zipfian distribution as shown in Figure 2. The number of users exponentially decreases with an increase in the number of profiles (e.g., we have 5,797 users with one profile link and 327 users with 10 profile links). We observe a similar trend with the top 15 social network profile links (e.g., we have 3,335 users with one top 15 profile link and 40 users with 10 top 15 profile links). On average, a user has 5.7 links to other sites, and an average of 1.6 of those links point to a social network in the top 15. Users with a link to the top 15 social profiles have an average of 2.6 profiles.

Looking at 9 basic fields, a subset of which are shown in Table II, we measure the size of an online social footprint. For a user with one social networking site, the online social footprint size is 4.3 fields on-average, whereas the online social footprint size is over 8.25 fields, on-average, for a person with 8 or more social networking sites. Figure 3 shows the size of a user's online social footprint growing, with an increase in the number of social networking sites.

Depending on the social network being used, the information can vary widely. For example: Delicious tracks a user's favorite URLs; Flickr stores user's pictures, and Last.Fm tracks the music a user is listening to. We see that sites like Facebook, Flickr, and YouTube collect and display more basic personal information than sites such as Delicious and Twitter.

V. Reconstructing a person's online social footprint

Each social networking site reveals a limited, sometimes unique, amount of information about a given user, but if an attacker coalesced the information from multiple social networking profiles for that user, the amount of disclosed information would increase significantly. In this section, we explore the feasibility of profile aggregation techniques that can be performed by attackers.

Previous work on aggregation techniques (or de-anonymization techniques) [9], [10] have focused primarily on network based matching techniques. Unfortunately, these approaches typically require knowledge of a large portion of a social network graph with overlapping sections, which is very computationally expensive and often impossible to obtain. Thus, to avoid these issues with previous techniques, we propose using pseudonyms to match profiles across multiple social networking sites.

A. *Prior knowledge of pseudonyms*

Our first aggregation technique assumes that an attacker knows one of a user's pseudonyms from an e-mail address or another source. Using a known pseudonym, we investigate what percentage of a user's social networking sites can be found by an attacker. In the best-case (from an attacker's standpoint), the known pseudonym is the most frequently used by a user, and as a result, it allows the attacker to access a maximum number of the user's social networking sites. On the other hand, the known pseudonym could be the least frequently used by a user, revealing a minimum number of the user's sites (this represents the worst-case).

Figure 4 illustrates the success rate of this approach by an attacker. The x-axis shows the number of unique id's for a user, and the y-axis shows the percentage of the user's total number of sites that an attacker can find. For users with two unique pseudonyms, the figure shows that an attacker can obtain 62%

[1] A few sites were not parsed due to technical and legal challenges associated with parsing the site or the low value of information found on the site.

TABLE II
SUBSET OF FIELDS COLLECTED FOR EACH SOCIAL NETWORKING.

Social Site:	***Name***	***Location***	***Sex***	***Relationship***	***Hometown***	***Homepage***	***Birthday***
Del.icio.us	-	-	-	-	53	-	-
Digg	100	67	55	-	-	-	30
Flickr	73	58	82	59	51	74	-
Last.Fm	82	-	87	-	76	77	-
LinkedIn	100	88	-	-	-	-	-
LiveJournal	93	69	-	-	-	68	64
MySpace	94	98	100	72	40	-	100
Technorati	94	-	-	-	-	-	-
Twitter	100	93	-	-	-	89	-
YouTube	68	-	-	-	29	57	73

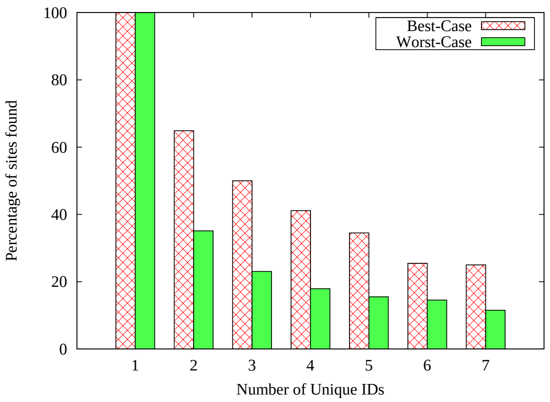

Fig. 4. Number of sites identified with a single pseudonym in the best-case and worst-case.

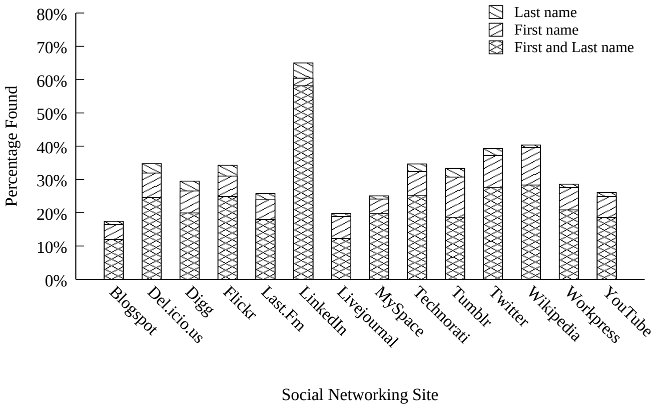

Fig. 5. Percentage of pseudonyms found by inference rule category for each social network site.

of their social networking sites in the best-case (i.e., with knowledge of a user's most frequently used pseudonym) and 38% of their sites in the worst-case (i.e., with knowledge of a user's least frequently used pseudonym). Over 98% of the users in our dataset have 4 or less unique pseudonyms. Thus, an important observation from the figure is that an attacker can find more than 40% of those users' social networking profiles in the best-case. Even in the worst-case, an attacker can still find over 17% of those profiles.

Although the figure only illustrates the situation in which an attacker knows a single pseudonym, we also investigated scenarios in which an attacker knows two or more of a user's pseudonyms. As expected, the success rate for these attacks increases dramatically. With knowledge of only two of a user's pseudonyms, an attacker will find over 60% of the user's social networking profiles in the best-case (and more than 35% of the profiles in the worst-case).

B. Inferring a pseudonym from a name

Another method of aggregating a user's profiles involves guessing that user's pseudonyms based on real name information. This method has not been sufficiently explored in previous literature, and in this section, we investigate the technique's ability to match a user's profiles.

To perform the aggregation process, we use three categories of inference rules. To achieve a baseline result, the inference rules are intentionally simplistic, and each category is limited to approximately 30 rules. The three categories are as follows:

1) *First name and last name* - includes guesses of a pseudonym that include the first name and last name. This category also includes tests that incorporate a subset of the first name and/or last name. Examples of guesses include "{FirstName}{Separation-Character}{LastName}", "{FirstName}{LastInitial}", and "{FirstNameSubstr(3)}{LastName}" (where Separation-Character is one of [-,;_.+,]).
2) *First name* - includes guesses of a pseudonym that include the first name or a substring of the first name. Examples of guesses include "{FirstName}", "{FirstNameSubstr(3)}", "{FirstName}007/69", etc.
3) *Last name* - includes guesses of a pseudonym that include the last name or a substring of the last name. Examples of guesses include "{LastName}", "{LastNameSubstr(3)}", "{LastName}007/69", etc.

To illustrate the power of pseudonym guessing, we begin by guessing a user's pseudonym in each of the top 15 social networking sites. Figure 5 shows a stacked bar graph that illustrates the number of pseudonym matches found for each site using each of the inference rule categories. Each bar consists of three sub-components (one sub-component for each category). The figure shows that LinkedIn has the highest number of matches for our guess-based approach, presumably due to the professional nature of the site. Blog web sites such as Blogspot, LiveJournal, and Wordpress have a lower than average matching percentage. One hypothesis for this result is

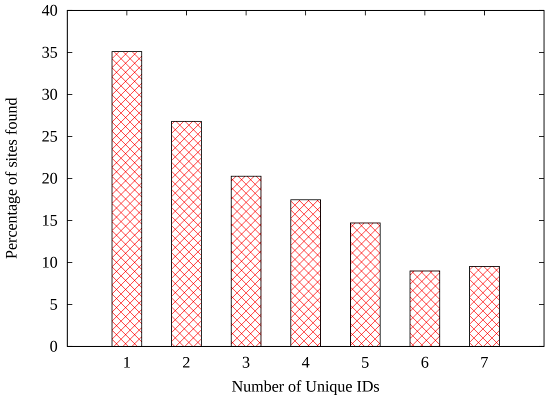

Fig. 6. Number of sites identified per pseudonym using names as guesses.

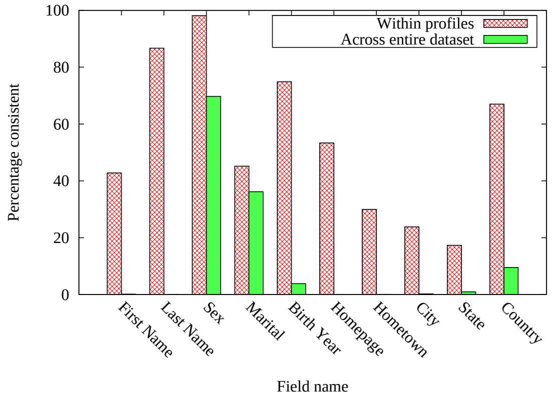

Fig. 7. Consistency of fields within a profile and across the entire dataset.

that most people name their blogs about the topic they plan to write about (i.e., they do not use their real name to title their blogs).

Another important observation from Figure 5 is that a large percentage of the matches appear in the "First and last name" category. In fact, over 50% of the overall matches resulted from one of this category's inference rules: {FirstName} {LastName}. Additionally, 14% of these matches were found on LinkedIn.

Using only the simple guesses described above, we also investigated the percentage of a user's profiles that can be found using guess-based aggregation. Figure 6 shows the percentage of a user's profiles that were found using this guessing approach. From the figure, we see that even with limited guesses, an attacker can find up to 35% of a user's social networking sites (if a user only has one unique pseudonym). Additionally, for users with six or seven unique pseudonyms, attackers are still about to find around 10% of their sites.

Although we used relatively simple matching criteria, two profiles with identical pseudonyms might not belong to the same user (i.e., we might have a false positive). This problem is exacerbated if one were to guess pseudonyms more aggressively. As the aggressiveness of the guesses increases, the number of false positives increases as well. To truly estimate the false positive rate for our guesses, we would need to check if the pseudonym existed on the respective social networking sites for every one of the generated guesses. This would be expensive both computationally and in terms of network bandwidth. To overcome these limitations, we introduce a method in the next section that allows us to gain confidence about whether or not two profiles are the same by comparing information within the profiles.

C. Matching profile information

In this section, we offer an approach for determining if two profiles belong to the same user. This technique can be used in multiple ways including matching profiles together in the case when there are no pseudonyms (or pseudonyms are hidden, as is the case with Facebook) as well as eliminating false-positives.

To check if two profiles belong to the same user we need to check if fields common to both profiles are equal or similar in value. In this section, we focus our attention on categorical or single text fields (e.g., "First name", "Hometown", etc.) and leave investigating similarity across free text entry fields (e.g., "About me", "Interests", etc.) as future work.

In our dataset, for profiles that have duplicate fields, we investigate the consistency of information entered. The consistency of a field is calculated by taking all possible pairwise combinations ($\binom{n}{2}$, where n is the number of values a field takes on within a profile and across the dataset respectively) and checking if the values are equal.

Figure 7 shows the amount of consistency between fields within a profile and across the dataset. We performed this evaluation to determine how well two matching fields can predict two matching profiles. From the figure, we observe that "Sex" is the field with the highest consistency across profiles; however, this field is not a strong signal because the consistency of "Sex" against the dataset is very high. As a result, "Sex" is not a very discriminatory field. Last name, birth year and country all seem to be strong signals with higher than average matching ability within profiles and low matching ability across the dataset. Therefore, any of these fields should be a good candidate for checking if two profiles belong to the same user.

The consistency of certain fields like Martial status and Country could be improved by cleaning up the data and assigning synonyms to certain values. For example, one social network site might not distinguish between "Marriage" and "Common-law" whereas another site might allow the user the option to do so. The value in this case is not inconsistent; it is simply semantically different. Similarly, more complex location-based fields such as Street, State or City could be improved by taking into account different ways of representing the same value, (e.g., a state Georgia could be represented in one social network as Georgia and as GA in another). Ideally, one would simply be able to feed two locations into a mapping service and find the distance between the two locations, applying a radius threshold to the value to determine if they are in the same locality.

Instead of assigning a simple binary value (i.e., consistent or not) to particular field matches, a more flexible approach involves assigning a consistency probability based on a measure of "how far" one value is from another. For numeric fields, a simple measure of how close the two numbers are divided by the standard deviation can be used as a measure of how similar two fields are. For textual fields, string similarity metrics (e.g., Cosine similarity, QGramsDistance, etc.) can be applied.

To illustrate the value of this approach, consider the "Birth Year" field. It is consistent 75% of the time within profiles and only consistent 4% of the times across the dataset. As a result, this field is a good choice for matching between profiles. Given two profiles that possess the "Birth Year" field, if those two values match, it means that (with a certain probability) the two profiles belong to the same user. After investigating the distance measure (in the numeric case, the absolute value of the subtraction), we found that within profiles the sample standard deviation (including Bessel's correction) is an average of 1.8 years. However, across the dataset the sample standard deviation is 12.7 years. In the event that two "Birth Year" values do not match, this allows us to say with some probability what the likelihood is that the profiles belong to the same user (if the difference between values is under 1.8 years, the likelihood will be much higher than if it is closer to or over 12.7 years).

Even after performing the aforementioned correlation analysis, it might not be possible to easily match profiles. Different social networking sites might not reveal the same pieces of information, or they might not reveal enough of the same pieces of information to be able to make a strong match. To help alleviate this situation, Table II shows the amount of information that is available for the most popular sites. This information can be used to determine which sites are most susceptible to this type of matching, and it also provides an optimal match order.

VI. Related Work

A brief outline of areas in social networking sites which can compromise a user's privacy are discussed in Chew [2]. Some of the areas she discusses include 'Activity Streams', 'Unwelcome Linkage', and 'Merging Social Graphs'. Further Gross and Acquisti [4], highlight potential attacks on privacy and also show that a minimal percentage change default lax privacy settings.

Krishnamurthy [6] present statistics on privacy settings of two popular Social Networks. They show that 79% of MySpace allowed their profile, friends, comments and user content to be viewable. They also showed that a majority of users don't change default privacy settings of Facebook allowing anyone in the same "Regional Network" to view their profile.

Narayanan and Shmatikov [10] discuss techniques to de-anonymize large social networks using only network structure. The work presents another method by which relationships between profiles of the same user on multiple social networks can be identified.

VII. Conclusion and future work

As social networks continue to grow in size and importance, they begin to pose a number of interesting privacy challenges for their users. In this paper, we investigated one of those challenges: large online social footprints. Specifically, we have shown that targeted attacks on individuals are possible using techniques based on social networking pseudonyms. Based on our experiments, we have shown that an attacker can reconstruct over 40%, in the best-case from his perspective, of an individual's social footprint by using a single pseudonym. Additionally, an attacker can reconstruct 10% to 35% of an individual's social footprint by using the person's name.

The outlook we present in this paper makes hope seem bleak, especially with the aim of social networks being to promote creating online-identities and sharing information between friends. Normal anonymization methods of k-anonymity and anonymized network leakage aren't used when routinely displaying information to a potential friend, a curious wanderer, or a malicious attacker. As future work, we plan on investigating multidimensional k-anonymity techniques when applied to large online-social footprints.

Acknowledgment

This research has been partially funded by National Science Foundation grants ENG/EEC-0335622, CISE/CNS-0646430, CISE/CNS-0716484, AFOSR grant FA9550-06-1-0201, NIH grant U54 RR 024380-01, IBM SUR grants and a Faculty Partnership Award, Hewlett-Packard, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

[1] Facebook - press room: Statistics. http://www.facebook.com/press/info.php?statistics, 2009.

[2] Monica Chew, Dirk Balfanz, and Ben Laurie. (Under)mining privacy in social networks. In *W2SP: Web 2.0 Security and Privacy*, 2008.

[3] comScore. Social networking goes global. http://www.comscore.com/press/release.asp?press=1555, 2007.

[4] Ralph Gross, Alessandro Acquisti, and H. John Heinz, III. Information revelation and privacy in online social networks. In *WPES '05: Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, pages 71–80, New York, NY, USA, 2005. ACM.

[5] H. Jones and J.H. Soltren. Facebook: Threats to privacy. *Project MAC: MIT Project on Mathematics and Computing*, 2005.

[6] Balachander Krishnamurthy and Craig E. Wills. Characterizing privacy in online social networks. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 37–42, New York, NY, USA, 2008. ACM.

[7] Kang Li and Danesh Irani. Sample of password recovery questions. http://www.cc.gatech.edu/~danesh/download/KLi_Dirani_PasswordRecovery_2009.pdf, 2009.

[8] Sean Michaels. Facebook challenges myspace's music monopoly. http://www.guardian.co.uk/music/2008/feb/29/news1, 2008.

[9] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *IEEE Symposium on Security and Privacy, 2008. SP 2008*, pages 111–125, 2008.

[10] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. *Imprint*, 2009.