# Is More Always Merrier?
# A Deep Dive Into Online Social Footprints

Terence Chen[*†], Mohamed Ali Kaafar[*⋆], Arik Friedman[*], and Roksana Boreli[*†]
[*]National ICT Australia  [†]University of New South Wales, Australia  [⋆]INRIA, France
firstname.lastname@nicta.com.au

## ABSTRACT

We present an empirical study of personal information revealed in public profiles of people who use multiple Online Social Networks (OSNs). This study aims to examine how users reveal their personal information across multiple OSNs. We first consider the number of publicly available attributes in public profiles, based on various demographics and show a correlation between the amount of information revealed in OSN profiles and specific occupations and the use of pseudonyms. Then, we measure the complementarity of information across OSNs and contrast it with our observations about users who share a larger amount of information. We also measure the consistency of information revelation patterns across OSNs, finding that users have preferred patterns when revealing information across OSNs. To evaluate the quality of aggregated profiles we introduce a consistency measure for attribute values, and show that aggregation also improves information granularity. Finally, taking Australian phone directory as a case study, we demonstrate how the availability of multiple OSN profiles can be exploited to improve the success of obtaining users' detailed contact information, by cross-linking with publicly available data sources such as online phone directories.

## Categories and Subject Descriptors

H.4.m [**Information Systems**]: Miscellaneous

## General Terms

Measurement

## Keywords

online social network, social footprint, privacy

## 1. INTRODUCTION

Sources of information about individuals, increasingly collected in numerous locations, include many aspects of people's lives. Databases with records of health, education, employment, finance and other personal information are being created and updated by various government departments and businesses that individuals deal with. Search engine providers and online retailers collect data about end-user interests, spending habits and other online activities. Finally, the individuals also contribute to the information available about them by making it accessible on their homepages and on various online personal profiles, and by participating and expressing their views in selected online fora.

Online Social Networks[1] (OSNs) are a rich source of information about individuals. It may be difficult to justify the claim that the existence of public profiles breaches the privacy of their owners, as they are the ones who entered the data and made them publicly available in the first place. However, aggregation of multiple OSN public profiles is debatably a source of privacy loss, as profile owners may have expected each profile's information to stay within the boundaries of the OSN service in which it was created.

In this paper we describe an empirical study based on 180,510 online profiles of 34,559 users across ten major OSNs. Based on this study, we make the following novel contributions:

**Cross-OSN profile analysis:** we evaluate the size of the online social footprint [5], represented by the number of publicly available attributes in linked OSN profiles, and characterize it based on various demographics, including occupation and the use of pseudonyms. In addition, we introduce a measure of **complementarity** between OSNs to evaluate how much information is gained through aggregation of profiles. We find that on average, more than half of the combined set of attributes from any two profiles of the same user are complementary to each other. However, there is also a **correlation** between the average number of attributes revealed in a single OSN and the number of OSN profiles owned by the user, indicating that a larger footprint size may also result from users' inclination to share more information.

**Consistency and data quality:** we measure the consistency of attribute revelation patterns across OSNs, and compare it with respect to a random model in which revelation patterns on different OSNs are independent. Our observations indicate that users have preferred patterns when revealing information across OSNs. We also evaluate the quality of aggregated profiles by measuring the consistency of attribute values. We find that over 85% of the users have more than 50% matching attribute values across different OSNs. Moreover, profile aggregation also enables an increase in the granularity of the collected information, further improving its quality. The high level of consistency of attribute revelation pattern and attribute values suggests that users are vulnerable to linking attacks which combine their online profiles and other datasets.

**Beyond online social footprints:** taking the Australian phone directory as a case study, we demonstrate that profile aggregation

*WOSN'12,* August 17, 2012, Helsinki, Finland.

[1] In this paper we use the term OSN broadly and apply it to any online service in which users share information with friends, work colleagues or others.

goes beyond larger online social footprints, making it easier to gather user information from other public data sources. Our results show that additional information obtained from aggregated profiles improves the success of record linkage, increasing up to five times the number of uniquely identified users (and their corresponding phone numbers and residential addresses), and up to three times the number of users who have five or fewer matching records in the phone directory.

We note that our analysis considers only publicly revealed information and as such we do not claim that our findings apply to users who tend to hide most of their personal data. Although the profiles we examined are biased toward Google users (we used more than 35K Google profiles as the starting point for the data collection), the techniques and metrics discussed in the paper are general, and provide insights into user behaviour over multiple online services.

The paper is organized as follows. In Section 2, we describe the data collection process. We study demographic correlations in Section 3 and analyze the complementarity and consistency of OSN profiles in Section 4. Cross-linking with the Australian phone directory is presented in Section 5. We discuss related work in Section 6 and conclude in Section 7.

## 2. DATA COLLECTION

We selected ten target OSNs based on their popularity and intended use, aiming to represent major OSNs for both social and professional. We crawled user profiles by visiting their profile pages and by APIs that are provided by the target OSNs, between May and August 2011.

As a starting point, we randomly selected around 35K profiles out of the 3 million Google Profiles[2] collected by Perito el. al. [12]. We limited the number of target profiles because of the complexity and time involved in the crawling process. Google Profiles allow users to link their profile to accounts on other OSNs and web services. As a first step, we leveraged this feature to collect linked accounts on other OSNs. In a second step, we crawled additional information sources, such as external web pages and social aggregation services, to associate users accounts with profiles on additional OSNs. Overall, we successfully crawled and parsed a total of 179,188 profiles across the ten considered OSNs associated to our initial set of  35K users. Table 1 lists the target OSNs and the respective number of collected profiles. Interested readers may refer to the technical report [2] for further information about crawler design.

| OSN | Sample size | OSN | Sample size |
|---|---|---|---|
| Google | 34,559 | Blogger | 16,357 |
| Facebook | 26,499 | Flickr | 13,713 |
| Last.fm | 7,661 | LinkedIn | 15,620 |
| LiveJournal | 2,055 | Myspace | 9,782 |
| Twitter | 27,702 | YouTube | 26,562 |

Table 1: Number of collected profiles in each OSN

## 3. ONLINE SOCIAL FOOTPRINTS

In this section, we first examine the average number of publicly available attributes in each of the OSNs individually and when aggregated across multiple OSNs. We then consider different demographics and evaluate the relation between the use of pseudonyms and the footprint size.

We unified attribute naming variations across the target OSNs and identified 40 types of available attributes representing different personal information in public profiles (see Table 2 for the full list). Figure 1(a) shows the total number of attributes available in each OSN and the average number of publicly available attributes in the collected profiles. While the number of available attributes varies significantly between different OSNs, ranging from 6 to 27 attributes, in practice the average number of publicly available attributes for each user spans a smaller range, between 5.4 to 10.2 attributes.

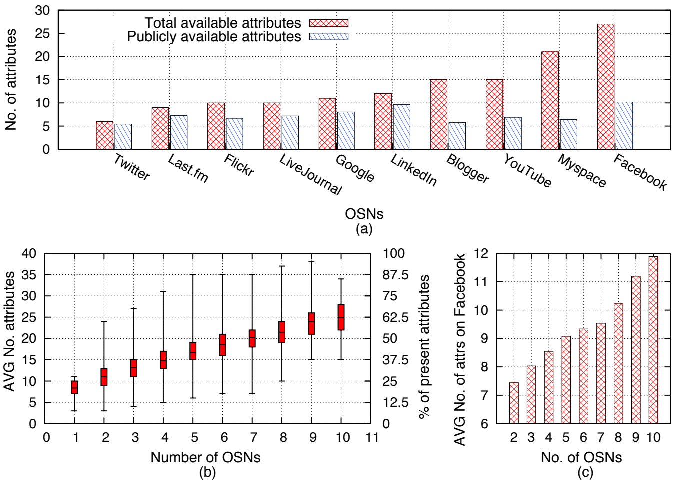

Figure 1: (a) Attribute availability in each OSN. (b) Average number of attributes v.s. number of OSNs. (c) Facebook footprint size v.s. number of OSN profiles owned by users.

about me/bio
age
anniversary
birth date
body type
books
children
companies
connections
education
email
ethnicity
family
gender
home town
income
industry
interests
interested in
language
links
lived
location
movies
music
name
occupation
organization
orientation
phone
political
quote
religion
schools
smoke/drink
sports
status
tv shows
user name
zodiac sign

Table 2: Attribute list

To quantify the amount of information revealed across multiple OSNs, we use the concept of online social footprint proposed by Irani et. al. [5], which represents the collection of all pieces of information that a user exposes on online social sites. For each set of users who have profiles on $n$ social networks, we measured the average number of (non-overlapping) publicly available attributes. The results are shown in Figure 1(b). We found that the average online social footprint size steadily increases with the number of OSN profiles, in line with the results reported by Irani et. al. [5]. To better understand this trend, we further explored two contributing factors. First, as different OSNs provide different types of attributes, users with more profiles have a larger selection of attributes from which to share. Moreover, even for overlapping attributes (e.g., the location attribute is available on many OSNs), a user may make an attribute public on one OSN, while it is private or not populated

[2] http://www.google.com/profiles

on another. We take a closer look at these aspects in Section 4. Second, possessing a large number of online profiles may indicate a tendency to share more information online, not only across several OSNs, but also in each individual OSN. For example, Figure 1(c) shows the average number of publicly available attributes on Facebook, for each set of users with a certain number of OSN profiles, indicating that the two are correlated. A similar pattern was observed in other OSNs.

## 3.1 Demographic Analysis

We further analyzed the data set to characterize users by various demographic groups: gender, age, country of residence and occupation. We note that out of the total 34,599 users in our data set, 75.94% revealed their gender in one or more OSN profiles; 68.24% disclosed their age; 90.96% disclosed their occupation and all users disclosed their country of residence.

Within the subset of users who revealed their gender, we did not detect statistically significant relation between gender and footprint size. Similarly, no such relation was evident for country.

Figure 2(a) shows the average online social footprint size by age group. In general, the average footprint size does not appear to be significantly different between age groups, with a slight increase for older age groups. However, we observed that users in different age groups tend to share different attributes, for example older users are more likely to provide religion and political view on Facebook while younger users are more interested in sharing their favourite music and books. A smaller footprint size was observed for users who did not share their age publicly.

Figure 2(b) presents ten of the most frequent occupations declared by users, and the average online social footprint size for each occupation. We found that users with customer-facing occupations, e.g., real estate or marketing, tend to disclose a significantly higher number of attributes than users with other occupations, e.g., teachers or students.

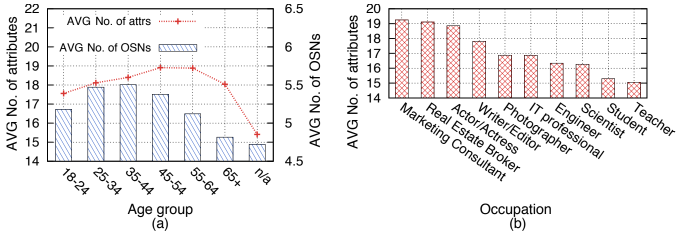

Figure 2: Average footprint size by (a) age group and (b) occupation.

## 3.2 Pseudonyms and Footprint Size

Although use of real names in OSNs has become a common practice (either by inclination or to comply with OSN policy), a significant number of users use pseudonyms. In this section, we characterize the amount of revealed information based on the use of pseudonyms. For this analysis we consider only English names, and exclude names in other languages.

We distinguished real names from pseudonyms by comparing names in profiles to those observed on LinkedIn, under the assumption that users provide genuine information in their professional profile.[3] We divided profile names into three categories: full real names, partial names (e.g., first or last name or some combination of them) and unrelated strings, considering the last as an indicator of a pseudonym. We used string matching and regular expression techniques to distinguish between these categories. To reduce the influence of spelling variations, e.g., name aliases, we used flexible fuzz matching with Jaro-Winkler distance metric [7]. Figure 3(a) shows the distribution of user name categories in each OSN. Users on Facebook, Google and Flickr mostly identify themselves by their real full or partial names, while pseudonyms are more prevalent on YouTube, Blogger and LiveJournal (24%, 29.4% and 41.2% respectively).

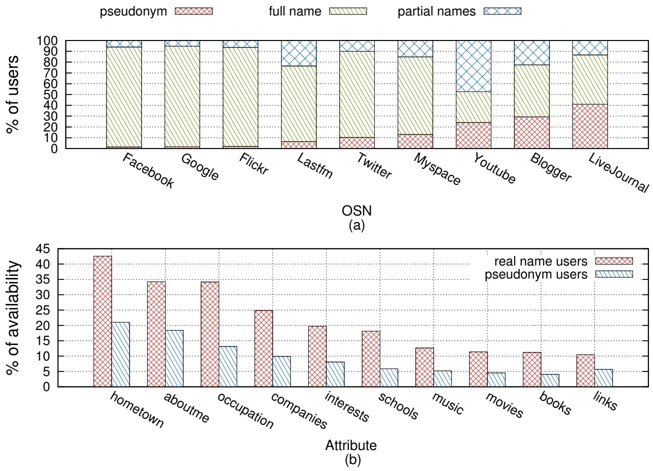

Figure 3: (a) Distribution of name categories in OSNs; (b) Online social footprint size on YouTube per name category.

We compared the average online social footprint size between users who use their real names (full or partial) and those who use pseudonyms, and found that on average the latter have a smaller footprint. As an example, Figure 3(b) demonsrates this for YouTube users. Users who identify themselves by their real names make more personal information publicly available: in addition to mandatory information like name, location and age, they share on average 2.2 more attributes, while pseudonym users only reveal 0.95 additional attributes.

# 4. CROSS-OSN ANALYSIS

The previous section identified the increased availability of information that can be obtained by combining multiple OSN profiles. We now investigate in detail the level of information complementarity across OSNs. Then, we question the existence of patterns in publicly revealed information across different OSNs and evaluate the consistency of the information users share through their multiple profiles.

## 4.1 Complementarity of Information

Consider a user $U^i$ who has two profiles $p^i_x$ and $p^i_y$ on OSNs $x$ and $y$ respectively. In each of these profiles, say $p^i_x$, we represent attributes availability as a binary vector $v^i_x = [a^i_{1,x}, a^i_{2,x}, ..., a^i_{t,x}]$, where $a^i_{k,x}$ denotes the availability of the $k^{th}$ attribute for user $i$ in OSN $x$. We set $a^i_{k,x}$ to 0 if the attribute is not available (either not populated by the user, private or altogether not available to be filled in that particular OSN) and $a^i_{k,x}$ is set to 1 if the attribute is populated and publicly available. We stress that in this initial analysis, we do not (yet) consider the consistency of the values across different OSNs profiles of the same user, so binary availability vectors are sufficient. We are interested in the complementarity of two profiles, i.e., the portion of attributes that are available in one profile

[3] This assumption is in line with the statistics reported by Irani et. al. [5], Fig. 3, for the use of real names in OSN accounts.

but not in another. To this end, we measure the Jaccard distance (or Jaccard dissimilarity) between profiles, given by:

$$D(v_x^i, v_y^i) = \frac{\sum(v_x^i \oplus v_y^i)}{\sum(v_x^i \vee v_y^i)} \ , \qquad (1)$$

where $\sum(v_x^i \oplus v_y^i)$ is the number of mismatches (XOR) and $\sum(v_x^i \vee v_y^i)$ is the number of non-zero instances in both $v_x^i$ and $v_y^i$ (OR). The average complementarity score of user $U^i$ possessing $n$ different OSNs profiles is:

$$\hat{D}^i = \frac{\sum_{x=1}^{n-1} \sum_{y=x+1}^{n} D(v_x^i, v_y^i)}{n(n-1)/2} \ . \qquad (2)$$

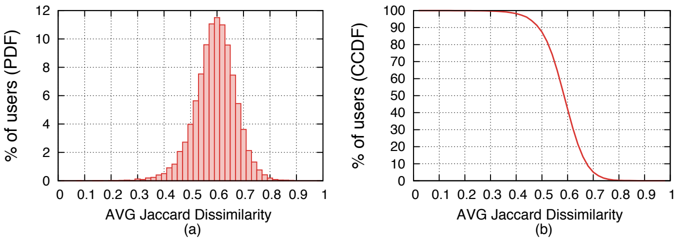

Figure 4: Information complementarity distribution: Jaccard dissimilarity PDF and CCDF.

Figure 4 shows the PDF and CCDF of the average complementarity scores. We observe in Figure 4(b) that almost 90% of the users have an average complementarity score over 0.5, which indicates that on average more than half of the set of combined available attributes from any two OSN profiles of the same user are complementary to each other.

## 4.2 Consistency of Information Revelation Patterns Across OSNs

To evaluate whether users actually have what could be considered a personal privacy policy, i.e., they reveal or hide attributes across different OSN profiles in a consistent manner, we compare the respective information revelation patterns. To this end, we again consider the attribute availability as a binary vector. For the purpose of this analysis, when comparing two profiles $p_x^i$ and $p_y^i$, we consider only the subset of attributes that are available in both OSN x and OSN y (regardless of whether these attributes are populated for any particular user or not). While our measure is sensitive to variations introduced by OSN design updates, or the introduction of new default settings, we still believe it is fit for purpose, as it provides a snapshot of users' online social footprint characteristics at the time the data was collected.

To capture the consistency of information revelation patterns we consider both positive and negative matches and we use the Sokal & Michener similarity metric [3]. The metric takes two binary vectors and returns a normalized similarity score between 0 (no match) and 1 (full match). The similarity score for vectors $v_x^i$ and $v_y^i$ is computed as:

$$S(v_x^i, v_y^i) = \frac{\sum(v_x^i \leftrightarrow v_y^i)}{\sum(v_x \wedge v_y)} \ , \qquad (3)$$

where $\sum(v_x^i \leftrightarrow v_y^i)$ is the number of binary instances that are both equal 1 or 0 (XNOR), and $\sum(v_x \wedge v_y)$ is the number of attributes available to be filled in both OSN $x$ and OSN $y$. The average consistency of information revelation patterns for a user with $n$ OSN profiles is then computed as:

$$\hat{S}^i = \frac{\sum_{x=1}^{n-1} \sum_{y=x+1}^{n} S(v_x^i, v_y^i)}{n(n-1)/2} \ . \qquad (4)$$

Some attributes may be publicly available by design or very common (e.g., name and location are highly represented in most OSNs), leading to higher similarity scores between OSNs that share those attributes. Similarly, very rarely revealed attributes can also lead to high similarity scores. Therefore, as a baseline for comparison we also plot the expected consistency distribution for a model where attributes are revealed randomly and independently in each OSN. We generate random patterns using a biased coin function, based on the public availability of each attribute as observed in each OSN. For instance 9.66% of of users in Facebook revealed the attribute "birthday", then we set the bit in the baseline pattern vector to "1" with a probability of 9.66%. We then compute the similarity score with other OSN for the obtained random binary vectors.

Figures 5(a) and 5(b) respectively show the PDF and the CCDF of consistency scores for all users. The observed distribution has a mean of 0.7, and more than 97% of users have a consistency score higher than 0.5. Compared to the random model, the observed distribution shows significantly more consistent patterns, which suggests that users do have preferred patterns when revealing information across different OSNs.

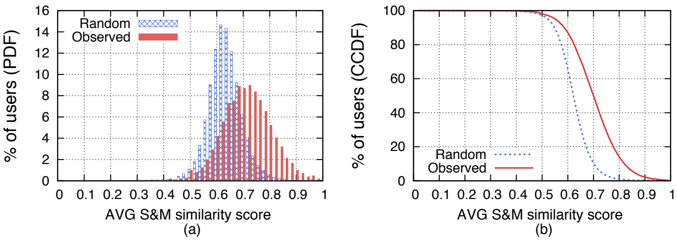

Figure 5: Information revealing pattern consistency across multiple OSNs: Sokal & Michener similarity PDF and CCDF

## 4.3 Consistency of Attribute Values

In the following, we measure the consistency of the attributes values as populated by users across multiple OSNs. The average consistency of attribute values for user $U^i$ (with $n$ OSNs) is computed as:

$$\hat{C}^i = \frac{\sum_{x=1}^{n-1} \sum_{y=x+1}^{n} \frac{\left|\{p_x^i\} \cap \{p_y^i\}\right|}{\sum(v_x^i \wedge v_y^i)}}{n(n-1)/2} \ , \qquad (5)$$

where $\{p_x^i\}$ and $\{p_y^i\}$ are the sets of publicly available attribute values in user $U^i$'s profiles $p_x^i$ and $p_y^i$ respectively; $\sum(v_x^i \wedge v_y^i)$ is the number of attributes that are publicly available in both profiles, and $\left|\{p_x^i\} \cap \{p_y^i\}\right|$ is the number of attributes with matching values. Because attribute values may have different formats on different OSNs, we apply different matching decision processes for different types of attributes rather than relying on plain string matching. For instance, we standardize the format of birth date, age, URLs and location before comparing the attribute values.

The average value consistency distribution is shown in Figure 6(a), with a mean of 65.9% matching attributes. The CCDF in Figure 6(b) shows that over 85% of the users have more than 50% matching attribute values across different OSNs.

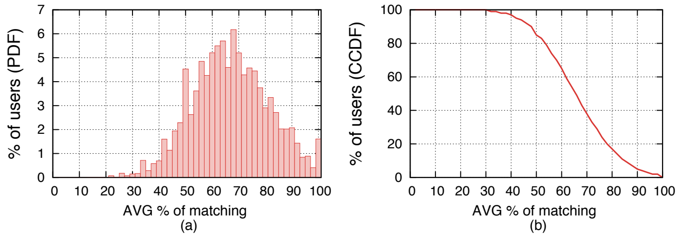

Figure 6: Attribute value consistency across multiple OSNs: PDF and CCDF

### 4.3.1 *Consistency of Location Information*

We used Google Geocoding API[4] to format the available location information to the precision levels of city, state and country. The availability of location information across 9 OSNs (location information is not shown in MySpace profiles) is shown in Figure 7(a). Note that by design, the location in Last.fm and YouTube profiles is restricted to country level. We observed that for over 88.2% of Facebook users, and over 60% of users on all other OSNs, city-level location was publicly available. Since most OSNs do not restrict the format of location in the profiles, a number of users filled in informal location names which may not be recognized by Geocoding API. Such portion is highest in Twitter, approximate 8% and follow by Blogger and Google, about 5% .

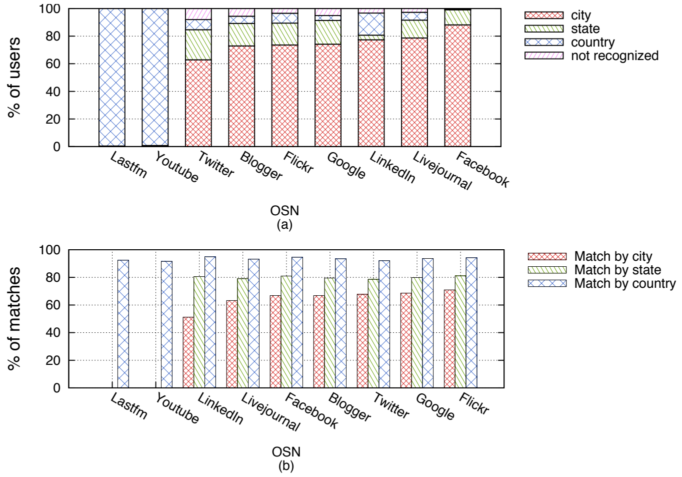

Figure 7: Location information: (a) Levels of granularity in different OSNs; (b) Consistency between OSNs.

Due to the mobility of users, location information may be inconsistent if their profiles across multiple OSNs are not up-to-date. To measure such inconsistency, we compare the location information in each OSN profile to other profiles with the granularities of city, state and country. Figure 7(b) shows the percentage of location information matches when the information retrieved from one particular OSN is used as a reference (considering the smallest available location granularity in both OSNs). As illustrated in Figure 7(b), a vast majority of users provide consistent country and state-level information (on average 93% and 80% respectively) across the OSNs. The proportion of city-level matches is surprisingly high, indicating that at least half of the users are willing to share the city where they live across multiple OSNs. For example, when considering LinkedIn as a reference, the city-level location provided by users on the other OSNs matches in 51% of the cases.

## 5. BEYOND ONLINE SOCIAL FOOTPRINTS – THE AUSTRALIAN PHONEBOOK CASE

In Section 4 we showed how aggregation of information across OSNs allows gathering additional information through complimentarity, and evaluating its quality through consistency. However, aggregation contributes also to improved information granularity. Figure 8 shows the improved availability of location information in different granularity levels on average as more OSN profiles are aggregated.

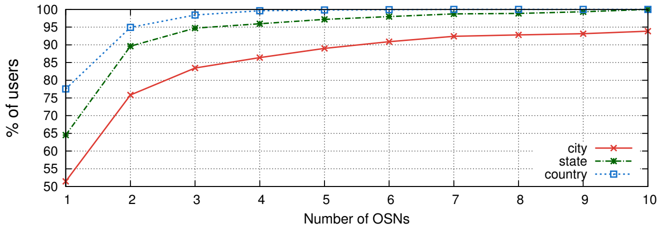

Figure 8: Increasing location granularity by aggregation

We stipulate that the aggregation of information over multiple OSNs may go beyond the increased online social footprint size and the higher granularity of information. It also makes it easier to gather user information from other public data sources, thereby contributing to privacy loss and increasing the risk of identity theft. To demonstrate this, we use an electronic copy of the Australian phone directory[5], which includes the registered users' surname and initial, full address (street number, street name, suburb and state) and phone number. In particular, we show how granular location information from aggregated OSN profiles helps uncover even finer address details as well as phone numbers of the owners of these profiles, although they may have preferred these details not to be associated with their profiles (e.g., phone numbers were publicly available for only 2% of Facebook users in our dataset. In contrast, Dey et. al. [4] observed that hometown and city details are increasingly shared publicly).

As a first step, we calibrated the search radius used to match locations reported in OSNs to phone directory listings. To this end, we used a control data set obtained from 42 volunteers around the greater regions of Sydney, Melbourne and Brisbane, of whom 25 (59.5%) were listed in the phone directory, and 17 (40.5%) were not. The volunteers provided their full names, OSN identifiers for two or more OSNs from the list in Table 1, and their residential addresses. To measure accuracy we considered the true positive rate (TPR), i.e., the rate of listed users whose real records were found within the given radius, and the false positive rate (FPR), i.e., the rate of unlisted users for whom false records were retrieved. Using the locations extracted from users' OSN profiles, we progressively increased the radius around the provided locations and measured the resulting TPR and FPR. The distance between two locations, knowing their latitude and longitude, was computed based on Haversine formula [10].

Table 3 shows sample values of the search radius and the corresponding search accuracy. We found 20km to provide the best balance between TPR and FPR. We observed that most Australian OSN users indicate a greater city area as their location (for exam-

[4] http://code.google.com/apis/maps/documentation/geocoding

[5] The Australian Residential Database, http://www.yell123.com

ple, in the phone directory, Sydney refers only to the central business district rather than the greater Sydney area), and the chosen radius is consistent with this observation.

| Range | TPR | FPR | Range | TPR | FPR |
|---|---|---|---|---|---|
| city | 33.3% | 27.8% | 30 km | 95.8% | 77.8% |
| 2 km | 41.7% | 38.9% | 50 km | 95.8% | 83.3% |
| 5 km | 45.8% | 50% | 100 km | 95.8% | 83.3% |
| 10 km | 50% | 50% | state | 100% | 88.9% |
| **20 km** | **91.7%** | **55.6%** | country | 100% | 94.4% |

Table 3: Search radius v.s. TPR and FPR values.

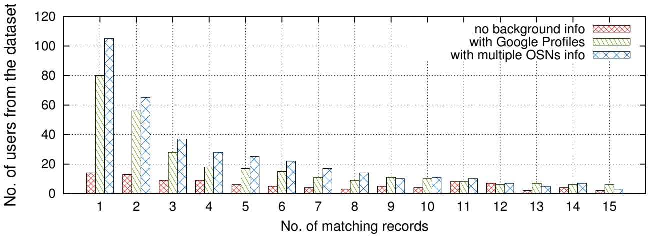

Figure 9: Combining OSN profiles and phone directory to gather contact details.

In the second step, we used a subset of the dataset described in Section 2, taking only Australian users whose surname was listed in the phone directory, resulting in a data set with 605 user profiles. For each profile, we measured the number of matching records in the directory for three cases: matching using only the surname; including also location from a single OSN (Google Profiles); and including granular location information aggregated over multiple OSNs. Figure 9 shows the distribution of the number of matching records for each of these cases. The results indicate that the added information makes record linkage much easier, increasing up to 5 times the number of users who are uniquely identified, and up to 3 times the number of users who have 5 or fewer matching records.

## 6. RELATED WORK

Prior work on evaluating the information revealed by aggregation of public profiles from multiple OSNs focused on measuring the quantity of attributes present in these profiles. Krishnamurthy et. al. presented results for both mobile [8] and fixed [9] OSNs. Irani et. al. [5] presented the online social footprint, consisting of the aggregate information of OSN account owners. In a later work [6] they expanded the concept further and analyzed the threat the online social footprint poses in identity theft and account password reset attacks. In continuation to this line of work, our work investigates the factors that contribute to increased online social footprint, based on a cross-OSN analysis of the information available in public profiles. Our work provides a novel evaluation of the complementarity, consistency, and quality of information, which are important factors when considering cross-linking between OSNs and with other data sources.

De-anonymization of publicly released databases by combining multiple information sources was studied in several works [1, 11]. Also, the similarity of user names has been explored in [12], showing how easy it may be for a third party to guess an account name and gain access to a user account. Our work demonstrates cross-linking with a different type of data source, which is publicly available in a large number of countries and is not anonymous.

As part of investigating privacy trends, Dey et. al. [4] characterized OSN information for New York City Facebook users based on various demographics, including age and gender. Our work considers a larger set of OSNs and provides an analysis based on several demographics, including occupation and the use of pseudonyms, which were not considered before.

## 7. CONCLUSION

This paper presents an analysis of information revealed by linking the public profiles of people who use multiple OSNs and the potential implications on the user's privacy from combining this information with another publicly available data source, the online phone directory. Our most interesting findings highlight the high level of consistency and complementarity of user provided information on OSNs and illustrate how aggregation of information over multiple OSNs may make the retrieval of additional personal information from other public sources easier.

## 8. ACKNOWLEDGEMENT

This research work has been supported by funding from National ICT Australia (NICTA).

## 9. REFERENCES

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore Art thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography. In *WWW*, pages 181–190. ACM, 2007.

[2] T. Chen, R. Boreli, M.-A. Kaafar, and A. Friedman. An analysis of social footprints across multiple online social networks. Technical report, NICTA, March 2012.

[3] S. S. Choi, S. H. Cha, and C. Tappert. A Survey of Binary Similarity and Distance Measures. *Journal on Systemics, Cybernetics and Informatics*, 2010.

[4] R. Dey, Z. Jelveh, and K. W. Ross. Facebook Users Have Become Much More Private: A Large-Scale Study. In *4th IEEE International Workshop on Security and Social Networking*, 2012.

[5] D. Irani, S. Webb, K. Li, and C. Pu. Large Online Social Footprints–An Emerging Threat. In *Proceedings of the 2009 International Conference on Computational Science and Engineering*, 2009.

[6] D. Irani, S. Webb, C. Pu, and K. Li. Modeling Unintended Personal-Information Leakage from Multiple Online Social Networks. *IEEE Internet Computing*, 15, May 2011.

[7] M. A. Jaro. Probabilistic Linkage of Large Public Health Data Files. *Stat Med*, 1995.

[8] B. Krishnamurthy and C. Wills. On the Leakage of Personally Identifiable Information via Online Social Networks. *ACM SIGCOMM Computer Communication Review*, 40:112–117, 2010.

[9] B. Krishnamurthy and C. E. Wills. Privacy leakage in mobile online social networks. In *Proceedings of the 3rd conference on Online social networks*, 2010.

[10] MobileReference. *Trigonometry Quick Study Guide for Smartphones and Mobile Devices*. MobileReference, 2007.

[11] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. *IEEE Symposium on Security and Privacy*, 2008.

[12] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How Unique and Traceable Are Usernames? In *PETS*, 2011.