How Much is too Much? Leveraging Ads Audience Estimation to Evaluate Public Profile Uniqueness

Terence Chen[1,2], Abdelberi Chaabane[3], Pierre Ugo Tournoux[4], Mohamed-Ali Kaafar[1,3], and Roksana Boreli[1,2]

[1] National ICT Australia
{firstname.lastname}@nicta.com.au
[2] School of Electrical Engineering & Telecommunications, UNSW, Australia
[3] INRIA, Grenoble, France
{firstname.lastname}@inria.fr
[4] IREMIA, LIM, Université de la Réunion, France
tournoux@gmail.com

Abstract. This paper addresses the important goal of quantifying the threat of linking external records to public Online Social Networks (OSN) user profiles, by providing a method to estimate the uniqueness of such profiles and by studying the amount of information carried by public profile attributes. Our first contribution is to leverage the Ads audience estimation platform of a major OSN to compute the information surprisal (IS) based uniqueness of public profiles, independently from the used profiles dataset. Then, we measure the quantity of information carried by the revealed attributes and evaluate the impact of the public release of selected combinations of these attributes on the potential to identify user profiles. Our measurement results, based on an unbiased sample of more than 400 thousand Facebook public profiles, show that, when disclosed in such profiles, *current city* has the highest individual attribute potential for unique identification and the combination of *gender*, *current city* and *age* can identify close to 55% of users to within a group of 20 and uniquely identify around 18% of users. We envisage the use of our methodology to assist both OSNs in designing better anonymization strategies when releasing user records and users to evaluate the potential for external parties to uniquely identify their public profiles and hence make it easier to link them with other data sources.

1 Introduction

The potential to uniquely identify individuals by linking records from publicly available databases has been demonstrated in a number of research works, e.g. [12, 20, 23]. In [23] Sweeney reported on the uniqueness of US demographic data based on the 1990 census and showed that, 87% of the US population can be uniquely identified by *gender*, *ZIP code* and *date of birth*. The resulting loss of privacy, i.e. the potential for re-identification of a person's private data which

may exist in any other publicly released dataset, was also demonstrated by the author. A more recent study [12] also produced similar conclusions. Therefore, anonymization of databases (*e.g.* medical records or voting registers) with the aim of protecting the privacy of individual's records when such are publicly released, in reality cannot be successful if the released database contains potentially unique combinations of attributes relating to specific individuals.

Today, with the proliferation of public online data, Online Social Networks (OSNs) are a rich source of information about individuals. For either social or professional purposes, users upload various, in most cases highly personal and up to date information, to their OSN accounts. User's personal data exposure is managed by public profiles, which contain a selected (in some case mandatory) subset of the total information available in their private OSN profiles. In fact, public profiles represent an easily accessible public dataset containing user's personal details which, depending on the OSN, can include their age, gender, contact details [5] for home and workplace, interests, etc (for a full list, see [16]).

The existence of public profiles creates a valuable new source of information that has to be considered when releasing anonymized personal records. Also, the anonymized OSN private (profile) data is being released by OSN's to profiling and advertising companies, including in some cases additional information (e.g. political orientation such as in [1]), thus increasing the number of already available anonymized datasets used e.g. for medical or other research. These can be henceforth linked to public profiles, allowing the re-identification (and the de-anonymization) of the personal records i.e. the exposure of individual's identities[6]Previous research has addressed the release of online data in public OSN profiles [15, 16] and re-identification mechanisms aimed at e.g. anonymized OSN graphs [21].

In this paper, we aim to revisit the study of the uniqueness of demographics, however we consider *online public data* available for individuals. As a first step towards such analysis, we consider the evaluation of the uniqueness of public OSN profiles, consisting of the publicly available attributes e.g. *gender*, *age*, *location*, etc. associated with individual OSN accounts. We use information surprisal and entropy, established information theory metrics for measuring the level of information contained in random variables, to quantify the level of uniqueness. Having a higher information surprisal of the attribute values released in the public OSN profile can be directly related to being more unique in a set of OSN users, and therefore more easily re-identifiable when combining with other publicly available datasets containing the same attribute values. Then, this work also answers the question of the appropriate selection of attributes to be included when releasing anonymized personal records.

We note that quantifying the user's revealed information is a challenging task, as data that needs to be acquired in order to obtain a reliable estimation of profile uniqueness, is either only partially accessible (private attributes are

[5] Recently released Graph Search (`https://www.facebook.com/about/graphsearch`) service from Facebook illustrates well the ease of access to individual profiles

[6] The policy of major OSNs is to use real names (`http://goo.gl/2DkG6`)

by definition hidden), protected by OSNs providers, or of too large volume for the data collection to be practical. Our study provides a novel probabilistic framework that leverages the global private attribute statistics retrieved from a major OSN ad platform (Facebook), to obtain an *unbiased* quantification of uniqueness. We present an approach that takes user specific privacy policy into account and allows us to calculate the uniqueness of public profiles, computed over the entire Facebook dataset.

The first contribution of our paper is our proposed **methodology for computing the uniqueness of public OSN profiles**, independently from the dataset on which the analysis is performed. This methodology can, more generally, be applied to any set of attributes that comprise a user's profile. To calculate the probability of publicly revealing a combination of attributes and evaluate the measure of uniqueness, we combine statistics derived from the captured dataset of publicly revealed attributes and the ads audience estimation platform. We consider both independence and dependence of the probabilities to reveal different attributes.

Our second contribution is that we **evaluate the quantity of information** carried in individual attributes and attribute combinations present in user's profiles of a major OSN (Facebook). We show that there is a wide range of values for the amount of identifying information carried by different attributes, with *gender* being the lowest with $1.3bits$ of entropy and *current city* the highest with the entropy of $13.6bits$.

In our third contribution, we **identify the key attribute combinations that contribute to profile uniqueness in Facebook**. Consistent with reported results for linking anonymous US datasets [12, 23] but also applicable globally, we show that the combination of *gender*, *place of residence* and *age* (directly related to date of birth used in [12, 23]) has the highest impact on the potential for re-identification of user's anonymized data. The higher information granularity available in [12,23] and the difference in the type of community studied (online and global versus US population) results in a lower, although still significant, potential for identification. We show that 55% of Facebook users that reveal this attribute combination can be identified as a group of 20 and around 18% of such users can be considered unique with an information surprisal of $29bits$.

Finally, we **show the impact of user's privacy policy on the amount of information carried in Facebook public profiles** and highlight how policy uniqueness contributes to potential re-identification of users in anonymized datasets. We show that some attributes may allow users to hide in the crowd if revealed, as opposed to hiding them from public access.

The remainder of this paper is organized as follows. In Section 2 we provide a summary of the datasets used for this study. In Section 3 we describe the methodology for computing the uniqueness of public profiles. We present results and identify the key attributes that contribute to uniqueness in Section 4, followed by the discussion in Section 5. Related work is presented in Section 6 and we conclude in Section 7.

2 Our Data Source

For the purpose of our study, we have collected two datasets from Facebook: a set of public user profiles and a set of statistics collected from the **F**acebook **A**ds audience estimation **P**latform (hereafter referred to as FAP). In the following, we start by providing a brief description of user's profile as implemented by Facebook, then we describe the methodology used to collect the data in use in this paper. Finally, we describe the characteristics of our datasets.

2.1 Facebook Profiles

Facebook, similarly to a number of other OSNs, utilizes user profiles that are a collection of attributes which describe the user's personal data. An attribute may take one of a pre-determined set of values, e.g. *gender* can be male or female, while *current country* may take any of the global country names. Also, some attributes may be in free form text and may also have a number of values, e.g. *interests* may include books, movies, shopping, etc. The availability of these attributes conforms to a set of privacy rules (i.e., ACL) defined by Facebook and selected (with the exception of a small number of mandatory attributes) by the user. According to the privacy settings, an attribute can be visible to anyone, shared with (a set of) user's social links (e.g., friends) or only visible to the owner of this profile. Hereafter, we consider an attribute (resp. a set of attributes, i.e. profile) to be *public* if it is visible to anyone and *private* otherwise.

2.2 Public Facebook Profiles Dataset

Collecting data from a large OSN is a challenging task, as the huge volume of data necessitates use of a sampling approach, which should produce a uniform representation of the overall dataset. In this study we use the dataset of Facebook public profiles from [5]. This dataset was obtained by first scraping all unique user names in the latin character part of the the Facebook public directory[7], resulting in 100 Million user identifiers (IDs). Then, we sampled, randomly, a subset of 494,392 IDs for which we retrieved the corresponding public profiles (i.e. attributes). We finally processed the collected data to unify the values of country of origin and current country using the Geocoding API[8]. This resulted in a set of 445,024 profiles[9] used for this study, that we refer to as PubCrawl.

It is worthwhile noting that, as per [11], an unbiased sample of a population can be obtained by True Uniform sampling of the total population, i.e. for

[7] `http://www.facebook.com/directory/`

[8] `https://www.developers.google.com/maps/documentation/geocoding/`

[9] The size's mismatch is due to profiles where the geolocation was unsuccessful or simply could not be used (e.g. some of the 49 K profiles that have been removed correspond to locations in China for which at the time of the data collection the Ads platform did not provide demographic information. Note that currently the FB ads platform provides no information for Iran, while China has been enabled).

Facebook, by sampling the 32-bit space of user IDs. However, in practice only 25% of this space is currently allocated to existing users. A close approximation of True Uniform sampling for Facebook IDs can be achieved by randomly sampling the Facebook public directory which lists all IDs of searchable profiles. We have verified the number of searchable profiles in Facebook public directory, as of April 2013, was 1.14 Billion. This is very close to the reported number of Facebook monthly users for December 2012, 1.06 Billion[10] (whether the corresponding profiles are searchable or not[11]). We believe then that our dataset extracted from the Facebook public directory is a good representative of the Facebook population of public profiles.

2.3 Facebook Ads Platform Dataset

Facebook offers a platform to estimate the audience of targeted ad campaigns[12]. Advertisers can select different criteria such as user's *locations* (country or city), *gender*, *age* (or range of ages), etc.[13] These criteria can also be combined in a conjunctive manner. According to the selected combination, FAP outputs the ***audience*** which represents the number of Facebook users that match the criteria.

Although there is no full report on how Facebook generates the audience values, Facebook document [14] states that it uses *all* provided information to calculate the audience size for targeted ads which implies that both public and private attributes are utilized. The only exception is the use of IP address to determine the current location of users (i.e. *current city* and *current country*)[15] To build the FAP dataset we proceed as follows. We use a subset of six attributes: *gender*, *age*, *relationship status*, *Interested in*, *current city* and *current country*. First, for every Facebook profile in PubCrawl, we extract the set of revealed attribute's values (e.g., male, New York). Then, for each extracted attribute set, we retrieve the corresponding audience size from FAP. In addition, we collect statistics for each attribute and for all possible attribute values (e.g., all possible locations).

To collect the statistics from FAP, we have developed a customized automated browser based on the Selenium WebDriver[16] which sends requests to FAP with an acceptable rate. We share our collected dataset on: `http://planete.inrialpes.fr/projects/Adsstatistics`. Finally, it is interesting to note that

[10] `http://goo.gl/GEJyH`

[11] Note that with the current privacy settings in Facebook, users can no longer opt-out of the Facebook public directory (`http://goo.gl/AufHN`)

[12] `http://www.facebook.com/advertising/`

[13] The advertiser can also target user's interests (e.g., beer and wine), interested-in (men or/and women), relationship, language, education and workplace.

[14] `http://goo.gl/wxcgX`

[15] We acknowledge that users connecting to the OSN service through e.g. proxies may introduce errors into the location distributions extracted from Facebook statistics compared to actual values.

[16] `http://seleniumhq.org/`

Facebook deliberately reduces the granularity of estimated audience size by only returning "fewer than 20" for audience numbers lower than 20 users. In our methodology presented in the following section, we conservatively consider "fewer than 20" as being exactly 20 users.

3 Methodology for Computation of Public Profile Uniqueness

This section presents our proposed method to leverage the ad platform audience estimation provided by OSNs operators (focusing on Facebook) to estimate the uniqueness of users profiles. The uniqueness of a random variable is related to the amount of information that it carries and is commonly measured by Information Surprisal (IS) and entropy. These are probability based metrics, therefore to compute the IS or entropy associated with a user's profile, we need a way to estimate the probability to observe the set of attribute values comprising the profile, independently from the population of profiles we consider.

We first introduce the required theoretical background and notations used in this paper, followed by the description of our mechanism to estimate the profile uniqueness.

$\mathcal{A}$	A set of attributes $(a_1, a_2, ...)$.
$V(a_i)$	The values of attribute a_i.
$u^{\mathcal{A}}$	A profile defined over the attributes in $\mathcal{A}$.
pub, priv	Denote the public and private OSN profiles.
$\varnothing^{a_i}$	The set of profiles in which an attribute a_i is not available.
$P^{\varnothing}(a_i)$	Probability that the attribute a_i is not present in a profile.
$P^{rev}(\mathcal{A})$	Probability to publicly reveal every attribute in $\mathcal{A}$ knowing that they are present in the private profile.

Table 1: Notations used in this paper

3.1 IS and Entropy Computation for OSN Profiles

Table 1 introduces the notations used in this paper. We denote Tot as the set of *all* user profiles of a given OSN. Every user profile $u^{\mathcal{A}}$ in Tot comprises a set of k attributes $\mathcal{A} = (a_1, .., a_i, .., a_k)$. The profile $u^{\mathcal{A}}$ and all the associated variables may refer to a private, priv or a public, pub profile. An attribute a_i can be seen as a random variable, X^{a_i}, with values in $V(a_i) = \{x_1^{a_i}, x_2^{a_i}, .., x_n^{a_i}\}$ which follow a discrete probability function $P(a_i = x_j^{a_i})$. Similarly, a user's profile $u^{\mathcal{A}}$ defined on a set of k attributes $\mathcal{A}$ can be seen as the outcome of the k-dimensional random vector $(X^{a_1}, X^{a_2}, .., X^{a_k})$.

Information surprisal and entropy IS or self-information measures the amount of information contained in a specific outcome of a random variable. IS of a user profile u which includes a set of attributes $\mathcal{A}$ is given by: $IS(u^{\mathcal{A}}) = -log_2(P(u^{\mathcal{A}}))$, with $P(u^{\mathcal{A}}) = \frac{|u^{\mathcal{A}}|}{|Tot|}$ *i.e.* the proportion of users having the values of $u^{\mathcal{A}}$ for the set of attributes $\mathcal{A}$. IS is measured in bits and every bit of surprisal adds one bit of identifying information to a user's profile and thus halves the size of the population to which $u^{\mathcal{A}}$ may belong.

Entropy, denoted $H(\mathcal{A})$, on the other hand, quantifies the amount of information contained in a random variable (here a multi-dimensional random vector). Entropy and IS are closely related, as entropy is the expected value of the information surprisal, i.e. $H(\mathcal{A}) = E(IS(u^{\mathcal{A}}))$. The entropy of a set of attributes $\mathcal{A}$ is given by: $H(\mathcal{A}) = -\sum_{u^{\mathcal{A}} \in V(\mathcal{A})} P(u^{\mathcal{A}}) IS(u^{\mathcal{A}})$, and can be seen as the amount of information carried by the attributes in $\mathcal{A}$. E.g. a user in our public dataset of $4.45 \cdot 10^5$ profiles is unique if IS reaches $19bits$. For the Facebook population estimate, we use the value provided by FAP of 722 Million users, therefore a user profile is unique with an IS of $29bits$.

In the following, we focus on the use of the IS and entropy as a convenient way to measure the uniqueness of $u^{\mathcal{A}}$ amongst the OSN user profiles, which can be further utilised to derive the related level of anonymity of user profiles e.g. by using k-anonymity [24].

The freq method – Is PubCrawl enough A naive approach to compute the uniqueness of profiles is to rely on an unbiased sample of the entire OSN's profiles, such as PubCrawl, and adopt a frequency-based approach (denoted freq) to provide a rough approximation of the probability $P(u^{\mathcal{A}})$, used to compute IS and entropy. Assuming we have a dataset of $|Tot|_{crawl}$ profiles, we can then estimate the probability of each profile simply as $\frac{|u^{\mathcal{A}}|}{|Tot|_{crawl}}$ if u^A belongs to PubCrawl, and 0 otherwise, where $|u^{\mathcal{A}}|$ represents the number of occurrences of $u^{\mathcal{A}}$ in PubCrawl. In the remainder of the paper, we will refer to the frequency-based computation of IS as IS_{freq}, computed by:

$$IS_{\mathsf{freq}} = -log_2(\frac{|u^{\mathcal{A}}|}{|Tot|_{crawl}}) \tag{1}$$

This approach has at least two drawbacks. Unless all possible combinations of attribute values (as observed in the entire set of profiles Tot) are collected in the PubCrawl dataset, the frequency-based approach would provide a very coarse estimation and the IS value is lower bounded by the sample size of the dataset. Therefore if freq method is used, a maximum value of $19bits$ can be reached, as opposed to the maximum IS value of $29bits$, based on a full dataset. For the same reason, we would not be able to estimate the uniqueness of profiles corresponding to a set of attribute values that are not in PubCrawl. Whereas collecting such a large dataset is technically challenging, we propose a new methodology based on the audience estimation provided by the advertising systems of OSNs, which, as per Section 2, have access to the full set of private user's profiles.

3.2 Computing Profile Uniqueness from Advertising Audience Estimation

Ideally, to compute IS and entropy of a set of attributes $\mathcal{A}$ that are free from sampling bias and granularity constraints, we need to know the frequency of each profile, i.e. $|u^{\mathcal{A}}|$, in the full dataset Tot. Leveraging the audience size estimation from the OSN ad platform, we are now able to obtain such statistics that are based on the entire set of profiles. As discussed in Section 2.3, the audience size is estimated from both public and private profiles, resulting in overestimation of frequency for public profiles. This is because user's privacy policy limits the amount of information released on public profiles, which is often significantly lower than that available in private profiles and as such $|\varnothing^{\mathcal{A}}|_{pub} \gg |\varnothing^{\mathcal{A}}|_{priv}$.

However, the bias induced by users' privacy policy can be corrected by noting that: $|u^{\mathcal{A}}|_{pub} = |u^{\mathcal{A}}|_{priv} \cdot P^{rev}(\mathcal{A})$, where $P^{rev}(\mathcal{A})$ is the probability to publicly reveal attributes in $\mathcal{A}$ knowing that they are disclosed in the private profile.

In the following, we propose two methods to compute P^{rev}, trading off accuracy of the IS estimation and measurement costs (reflected by the number of requests to the ad platform) as discussed in Section 3.2. These methods are respectively denoted indep and dep, as they differ in the assumption regarding the mutual independence of the probabilities to reveal specific attributes.

The indep method – assuming independence between the likelihood of revealing specific attributes Here, we assume the probabilities to reveal selected attributes in user's public profile are mutually independent. The probability to reveal an attribute a_i, $P^{rev}(a_i)$, can then be obtained as follows.

First, we highlight the fact that the total number of public and private profiles is equal, $|Tot|_{pub} = |Tot|_{priv}$, i.e. there will always exist a corresponding public and private user's profile. We also observe that the number of public profiles in which an attribute is not present, i.e. $|\varnothing^{a_i}|_{pub}$, strictly depends on the probability that this attribute isn't publicly present, i.e. $P^{\varnothing}_{pub}(a_i)$, and as such: $|\varnothing^{a_i}|_{pub} = P^{\varnothing}_{pub}(a_i) \cdot |Tot|_{pub}$. Similarly, we can calculate the probability that an attribute is not disclosed in private profiles as: $|\varnothing^{a_i}|_{priv} = P^{\varnothing}_{priv}(a_i) \cdot |Tot|_{priv}$.

The number of profiles which define a_i as a private attribute but in turn hide this attribute from public access can then be obtained from equation (2):

$$\begin{aligned} &|\varnothing^{a_i}|_{pub} - |\varnothing^{a_i}|_{priv} \\ &= P^{\varnothing}_{pub}(a_i) \cdot |Tot|_{pub} - P^{\varnothing}_{priv}(a_i) \cdot |Tot|_{priv} \end{aligned} \tag{2}$$

On the other hand, we note that $(|Tot|_{priv} - |\varnothing^{a_i}|_{priv})$ accounts for the number of private profiles where a_i is revealed, and that $P^{rev}(a_i) \cdot (|Tot|_{priv} - |\varnothing^{a_i}|_{priv})$ is the total number of public profiles where a_i is revealed. Hence, the difference $(|Tot|_{priv} - |\varnothing^{a_i}|_{priv}) - P^{rev}(a_i) \cdot (|Tot|_{priv} - |\varnothing^{a_i}|_{priv})$ accounts for the number of users who have profiles where a_i is revealed on private but not on public profiles. We can then compute:

$$\begin{aligned} &|\varnothing^{a_i}|_{pub} - |\varnothing^{a_i}|_{priv} \\ &= (1 - P^{rev}(a_i)) \cdot (|Tot|_{priv} - |\varnothing^{a_i}|_{priv}) \end{aligned} \tag{3}$$

Hence from equations (2) and (3) we have:

$$P^{\varnothing}_{pub}(a_i) \cdot |Tot|_{pub} - P^{\varnothing}_{priv}(a_i) \cdot |Tot|_{priv} \\ = (1 - P^{rev}(a_i)) \cdot (|Tot|_{priv} - |\varnothing^{a_i}|_{priv})$$

i.e.

$$P^{\varnothing}_{pub}(a_i) - P^{\varnothing}_{priv}(a_i) = (1 - P^{rev}(a_i)) \cdot (1 - P^{\varnothing}_{priv}(a_i))$$

$$P^{rev}(a_i) = 1 - \frac{P^{\varnothing}_{pub}(a_i) - P^{\varnothing}_{priv}(a_i)}{1 - P^{\varnothing}_{priv}(a_i)} \tag{4}$$

with $P^{\varnothing}_{pub}(a_i)$ is the probability that attribute a_i is not available in public profiles.

Note that $P^{\varnothing}_{pub}(a_i)$ is computed from PubCrawl: $P^{\varnothing}_{pub}(a_i) = \frac{|\varnothing^{a_i}|_{pub}}{|Tot|_{crawl}}$. On the other hand, $P^{\varnothing}_{priv}(a_i)$, the probability that attribute a_i is not available in private profiles, is computed from FAP dataset: $P^{\varnothing}_{priv}(a_i) = \frac{|\varnothing^{a_i}|_{priv}}{|Tot|_{priv}}$, where $|\varnothing^{a_i}|$ is not directly available but can be computed by using the aggregate number of profiles queried from the ad platform for all possible values of the attribute a_i:

$$|\varnothing^{a_i}|_{priv} = |Tot|_{priv} - \sum_{u^{a_i} \in V(a_i)} |u^{a_i}|_{priv}$$

For example, for the attribute *age*, the number of private profiles in which this attribute is not included can be obtained by: $|\varnothing^{age}|_{priv} = |Tot|_{priv} - \sum_{j=13}^{j=65^+} |u^{age=j}|_{priv}$ (*age* can be queried from FAP for a range of values between $13 - 65^+$, where 65^+ refers to the "Nomax" *age* attribute in FAP).

According to the assumed independence between attributes a_i, the probability to reveal every attribute in $\mathcal{A}$ is obtained by: $P^{rev}_{\mathsf{indep}}(\mathcal{A}) = \prod_{a_i \in \mathcal{A}} P^{rev}(a_i)$

Finally, the IS estimation of public profile $u^{\mathcal{A}}$ using indepmethod can be computed as:

$$IS_{\mathsf{indep}} = -log_2(\frac{|u^{\mathcal{A}}|_{priv} \cdot P^{rev}_{\mathsf{indep}}(\mathcal{A})}{|Tot|_{priv}}) \tag{5}$$

The dep method – considering dependence between the likelihood of revealing specific attributes Although the indep method offers a simple way to compute $P^{rev}(\mathcal{A})$, the estimation of probabilities can be inaccurate if the independence assumption does not hold. To verify this, we evaluate the dependence between the likelihood of revealing specific attributes, based on our PubCrawl dataset. Table 2 shows the calculated probabilities to reveal each of the six example attributes: *gender*, *interested in*, *relationship*, *age*, *current city*, and *country* along the rows knowing that another attribute along the columns has been already revealed. Table 2 also includes the overall probability to reveal specific attributes ($1 - P^{\varnothing}_{pub}(a_i)$).

We can observe that there is indeed a correlation between probabilities to reveal specific attributes on public profiles. To properly assess the correlation

$1 - P^{\varnothing}_{pub}(a_i)$	0.76	0.15	0.22	0.024	0.21	0.23
	Gend.	Int. In	Rel.	Age	City	Country
Gender	1.00	0.88	0.86	0.86	0.8	0.8
Interested In	0.17	1.00	0.46	0.35	0.24	0.24
Relationship	0.25	0.68	1.00	0.48	0.33	0.33
Age	0.01	0.04	0.03	1.00	0.03	0.03
City	0.23	0.34	0.32	0.41	1.00	0.97
Country	0.23	0.35	0.33	0.43	0.99	1.00

Table 2: Probabilities to reveal attribute a_1 (rows) knowing that attribute a_2 is shown on public profile, e.g. $P(gender = revealed|age = revealed) = 0.86$

between two attributes, the probability $P(a_i = revealed|a_j = revealed)$ must be considered jointly with $1 - P^{\varnothing}_{pub}(a_i)$, the overall probability to publicly reveal a_i. The highest dependance can be observed for users' interest (*Interested In*), where users who reveal this attribute have a much higher probability to reveal any other attributes, *e.g.* the probability to reveal the *relationship status* when *Interested In* is revealed is over three times higher than the overall probability to reveal the *relationship status*.

We note that the values of the probabilities from Table 2 may be driven either by information sensitivity and user's privacy awareness, or simply by natural dependency between attributes from a semantic perspective, however the dependency analysis is out of the scope of this paper.

In the following, we present a methodology to compute $P^{rev}(\mathcal{A})$ taking into account the dependency between probabilities to reveal attributes. Addressing the dependency between $P^{rev}(a_i)$ with $a_i \in \mathcal{A}$, requires us to compute the frequency of a disclosed combination of these attributes.

$P^{rev}_{\mathsf{dep}}(\mathcal{A})$ can be computed similarly to equation (4), as:

$$P^{rev}_{\mathsf{dep}}(\mathcal{A}) = 1 - \frac{P^{\varnothing}_{pub}(\mathcal{A}) - P^{\varnothing}_{priv}(\mathcal{A})}{1 - P^{\varnothing}_{priv}(\mathcal{A})} \tag{6}$$

with $P^{\varnothing}_{pub}(\mathcal{A})$ and $P^{\varnothing}_{priv}(\mathcal{A})$, the probability that a set of attributes $\mathcal{A}$ is not available in a public (resp. private) profile being defined as : $P^{\varnothing}_{pub}(\mathcal{A}) = P(\bigvee_{a_i \in \mathcal{A}} a_i = \varnothing)$ and

$$P^{\varnothing}_{priv}(\mathcal{A}) = \frac{|Tot|_{priv} - \sum_{u^{\mathcal{A}} \in V(\mathcal{A})} |u^{\mathcal{A}}|_{priv}}{|Tot|_{priv}} \tag{7}$$

We note that the computation of $P^{\varnothing}_{priv}(\mathcal{A})$, and $P^{rev}_{\mathsf{dep}}(\mathcal{A})$, requires the audience estimation of every value $u^{\mathcal{A}}$ in $V(\mathcal{A})$. This is implemented by requesting every possible set of attributes from the ad platform. For example, to obtain $P^{\varnothing}_{priv}(\mathcal{A})$ where $\mathcal{A} = \{$*Interested In*, *gender*$\}$, we query the ad platform for the number of profiles corresponding to all combinations of *gender* = {man, woman} and *Interested In* = {man, woman, both}.

This represents an overhead in terms of measurement costs for the dep method, as compared to the indep method which requires a fewer number of queries. However, we note that this overhead may not be prohibitive, as the audience size estimation requests are sent to the ad platform only once for any set of attribute values.

The IS of the public profile $u^{\mathcal{A}}$, assuming the dependency of publicly revealing attributes, denoted by IS_{dep}, can be estimated as:

$$IS_{\mathsf{dep}} = -log_2(\frac{|u^{\mathcal{A}}|_{priv} \cdot P^{rev}_{\mathsf{dep}}(\mathcal{A})}{|Tot|_{priv}}) \tag{8}$$

4 Findings on Public Profile Attributes

In this section we study the uniqueness of users within the PubCrawl dataset, using the methodology presented in Section 3. We stress that our main focus is on the uniqueness resulting from the presence of specific attributes and attribute combinations in user's public profile, in line with our goal to have a generic mechanism for evaluating uniqueness. The impact of specific values is only presented for selected examples and used for illustration purposes and a comprehensive analysis based on attribute values is subject for further study.

4.1 Information Surprisal for a Single Attribute

We first consider the IS and entropy (average IS) for individual attributes, calculated using the freq and indep/dep methods and based on the PubCrawland FAP datasets. Note that in this section, as we are calculating IS (and entropy) for a single attribute, the IS_{dep} and IS_{indep} (and corresponding entropy) values are identical, and denoted as $IS_{\mathsf{dep/indep}}$.

Figure 1 (a)–(l) shows the PDF and CDF of the calculated IS values (y-axis, left and right hand side, respectively). For the sake of clarity, entropy H is included as a numerical value on top of each sub-figure (a)–(l). In addition, an absence of an attribute value may also be related to the profile uniqueness. To illustrate this, suppose that all users but one show their gender, as such the user who is hiding this information is *uniquely* identifiable since he has a unique "disclosing" pattern. Hence, the entropy is not only derived from the attribute value, but also from its presence (or absence). Therefore, we also show above each of sub-figures 1 (a)–(l) the number of users who hide a specific attribute and the associated IS as numerical values.

Overall, we can observe that there is a considerable difference in the range of IS and entropy values for selected attributes, with *gender* shown in Figure 1 (c)–(d) having the lowest and *current city* shown in Figure 1 (k)–(l) the highest entropy (and IS) values, respectively 1.3$bits$ and 13.6$bits$. This follows the definitions of IS and entropy, which are related to the number of values an attribute may take and the number of users with specific attribute values, so higher information granularity and lower number of users for a specific value both result in a higher uniqueness.

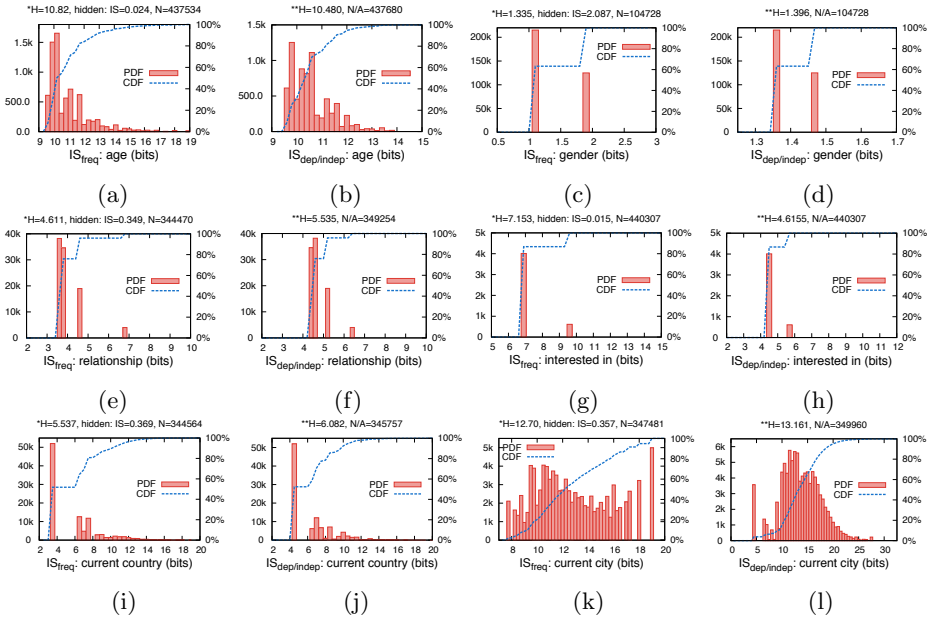

Fig. 1: PDF (left) and CDF (right) of *IS* values and Entropy H (shown on top of the sub-figures) for single attribute computed by IS_{freq} and $IS_{\mathsf{dep/indep}}$ methods (note $IS_{\mathsf{dep}} = IS_{\mathsf{indep}}$ for single attribute). Values are shown for: *age, gender, relationship, interested in, current city* and *country*. * *hidden IS=information surprisal for users hiding this attribute; N: number of users who hide this attribute; ** N/A: Total of N and the number of users for whom the disclosed attribute value is not available on FAP, e.g. age > 65*

Age: Considering the $IS_{\mathsf{dep/indep}}$ values in Figure 1 (b), we can observe that over 70% ($\sim$ 5.5k) of users have an IS value higher than 10.5*bits*, corresponding to an identifying user group size of about 500k users. This supports the conclusion that *age* is an identifying attribute which users should be careful about disclosing. In line with this, the users who hide this attribute (representing 98.4% of the total population) are highly anonymous with an IS value of 0.024*bits*. We remind the reader that each bit of information increase in IS halves population size to which the user represented by their public profile with corresponding IS may belong.

Gender: We can observe that Facebook users who reveal the *gender* attribute disclose less information (with average $IS_{\mathsf{freq}} = 1.34bits$ shown in Figure 1 (c) and average $IS_{\mathsf{dep/indep}} = 1.4bits$) shown in Figure 1 (d) than the users who consider this information private (with IS of 2.08*bits*). In Section 4.4 we will show the impact of hiding a common combination of attributes, including *gender*. Note that this is a highly popular attribute, with around 75% of Facebook users disclosing it in their profiles. Consequently, the population that hides it displays a high IS value for this attribute.

Relationship status: The calculated IS_{freq} shows, in Figure 1 (e)–(f), that for more than 60% of the users, the *relationship status* reveals a low value of IS, with $IS_{\mathsf{freq}} = 4bits$ and $IS_{\mathsf{dep/indep}} = 4.4bits$. Hiding this information has a very low associated IS of $0.35bits$. We note that only a subset of IS results are presented here, due to the supported values in FAP [17].

Interested In: We observe in Figure 1 (g)-(h) that the vast majority of profiles in our dataset do not disclose this attribute, resulting in a low IS value for such users (0.24bits). The average IS_{freq} values for users who display this attribute are moderate ($7.53bits$). Similarly, the $IS_{\mathsf{dep/indep}}$ values also do not indicate high user uniqueness, with users being identifiable to within a group of 3.9 Million, only by revealing this single attribute.

Current country: There is a wide range of IS values for users who have disclosed this attribute, as can be seen in Figure 1 (i)–(j). The average IS values are moderately high, with $IS_{\mathsf{freq}} = 5.54bits$ and $IS_{\mathsf{dep/indep}} = 6.08bits$, while hiding this information reveals very little ($0.4bits$). We note that 210 different countries appear as values for this attribute in our PubCrawl dataset. By examining the data values, we have observed that close to half of the total population (of those who have revealed their *current country*) have US as this attribute value. Therefore, the corresponding IS, for both IS_{freq} and $IS_{\mathsf{dep/indep}}$ methods, is low with a value of around $4bits$. For all other users with the *current country* attribute set, the calculated IS values for both methods range between a moderate value of $7bits$ to $15bits$, a significant amount of information which increases the uniqueness of the user resulting in an identifiable group of around 22k users.

When comparing the IS_{freq} and $IS_{\mathsf{dep/indep}}$ values, we can observe a lower IS_{freq} for the US, indicating that the IS_{freq} method overestimates the representation of US in the IS calculation.

Current city: The large range of potential values for this attribute and correspondingly high potential to distinguish users intuitively flags it as sensitive personal information. We can observe from Figure 1 (k)–(l) that the average IS values are quite high, with $IS_{\mathsf{freq}} = 12.7bits$ and $IS_{\mathsf{dep/indep}} = 13.16bits$, while hiding this information reveals very little ($0.4bits$). Also, more than 75% of the users who display this attribute value lose more than $11bits$ (based on both IS_{freq} and $IS_{\mathsf{dep/indep}}$ values). Note than more than 20% of the users in PubCrawl reveal this information, which makes it a valuable attribute for unique identification.

4.2 Expected Information Surprisal as a Function of the Number of Attributes

We now consider multiple attributes in IS calculations. Figure 2 shows the expected IS and the average entropy values calculated for a varying number of

[17] The Facebook Ads Platform (FAP) allows display of *relationship* statistic based only on a subset of values supported in Facebook profiles: single, married, engaged and in a relationship; queries based on divorced and widowed status are not supported.

attributes. We show the minimum, 25^{th} percentile, median, 75^{th} percentile and maximum of the IS values for all users. Both IS and entropy are averaged over all combinations of the selected number of attributes. As can be expected, increasing the number of disclosed attributes results in higher IS and entropy values and the corresponding amount of revealed information about the users. In Section 4.3, we will explore the specific attribute combinations which will result in higher IS values and therefore present a higher privacy risk for users.

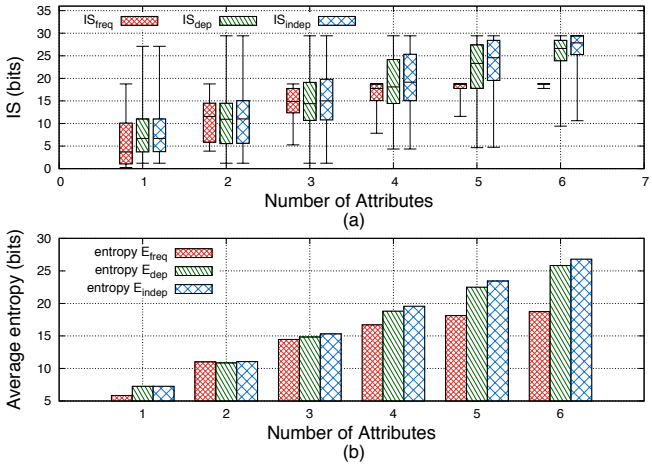

Fig. 2: (a) Information surprisal and;(b) Expected entropy for a varying number of attributes.

Comparing the results obtained using the three calculation methods, in Figure 2 we can observe that the values of IS_{freq} are consistently lowest for all attribute combinations, followed by IS_{dep} and IS_{indep}. As previously discussed, IS_{freq} presents a rough calculation of values, which can be used as an indication of the relevance (to privacy) of both attributes and attribute combinations. Increasing the complexity of obtaining data (i.e. the number of required queries from FAP) increases the accuracy of the result. Consequently, the IS_{indep} values can be calculated for the combinations not present in the collected dataset. However, this method results in higher IS and entropy values than what is obtained by the more precise IS_{dep} method, which in turn requires the highest amount of information from FAP.

We can observe the most significant difference in the IS and entropy values obtained by different methods when considering the users who have revealed six attributes in Figure 2 (b). The IS_{indep} and IS_{dep} values reach an average entropy higher than $25bits$, representing a corresponding uniqueness within a set of 22 users, while the IS_{freq} value underestimates IS and only reaches $19bits$ of entropy with a significantly lower corresponding unique user set of around 1300 users. Although there may be a number of factors contributing to the low

IS_{freq} values, the most relevant one is the dependency of the frequency-based entropy estimation on the used dataset. Regardless of the large size and the unbiased sample of the full dataset that we have used for IS_{freq} calculations, a number of combinations of attributes among the profiles may still be missing and will influence the result.

4.3 On the Relevance of Disclosed Attribute Combinations

We now consider the IS values for different attribute combinations, enabling us to draw conclusions about the dominant (and less relevant) parameters contributing to privacy loss. Figure 3 shows the cumulative distributions function of the IS, for: (a) IS_{freq}, IS_{indep} and IS_{dep} with all six attributes considered; (b)–(d) IS_{freq}, IS_{indep} and IS_{dep} for selected attribute combinations that were shown to have extreme (both low and high) IS values. Similarly to the results shown in Figure 2, we can observe that revealing 6 attributes (regardless of their values) results in a high IS value for the majority of users, e.g. observing the IS_{indep} and IS_{dep} CDF values in Figure 3 (a), more than 80% of the Facebook population with six disclosed attributes has IS of more than *22 bits*. This represents users uniquely identifiable within a set size of 170.

Considering specific attribute combinations shows the importance of having a carefully considered personal privacy policy with selectively disclosed attributes. Figures 3 (b)–(d) indicate that users should be wary of concurrently disclosing the combination of *age*, *gender* and *current city*, as this reveals almost as much information as as the total of six disclosed attributes. Although the granularity of our data is significantly lower (only age is available in the PubCrawl dataset, as compared to full birth date in the dataset used in [12,23]) and we study a different community (global and online, as compared to US only and based on Census data in [12, 23]), our results are in line with the previously published studies on the uniqueness of demographics [12, 23], which show that the combination of *gender*, *ZIP code* and *date of birth* has a very high uniqueness. We can observe in Figure 3 (c) that around 55% of users have IS of around *25 bits* and can therefore be identified in a set of 20 users. Further to this, around 18% of users can be uniquely identified, having the IS value of their public profile at *29 bits*. This represents a significant potential threat, as the corresponding number of Facebook users is around 7.7 Million for being identifiable to within a set of 20 and 2.7 Million for unique identification.

On the other hand, revealing the *gender* and *interested in* may not be harmful from a privacy perspective for most of the Facebook users (IS is less than *5 bits* for 90% of the users). Disclosing the *relationship status* along with the *gender* and *country* of residence reveals a higher amount of information, with more than 70% of users loosing at least 11*bits* of IS.

We note that, although our results have been derived for a sample of the Facebook population, the unbiased nature of the sample (as argued in Section 2) makes them applicable to the whole Facebook population.

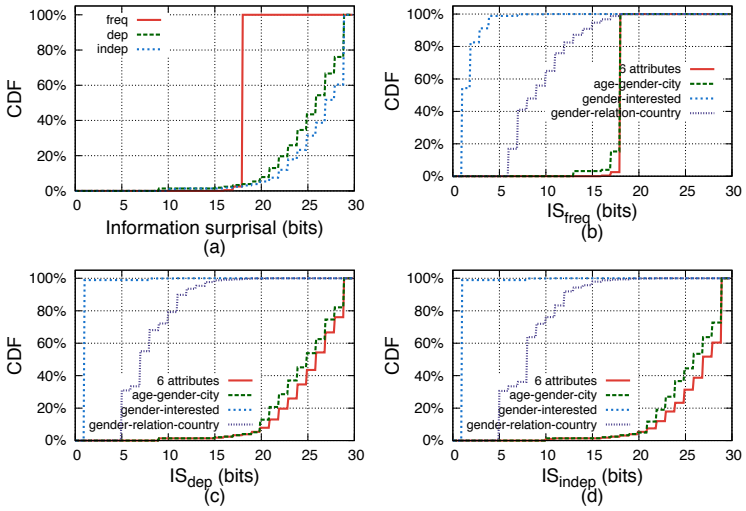

Fig. 3: Multiple attribute information surprisal distribution

4.4 Impact of Privacy Policy

This section complements the understanding of the key parameters that influence the resulting IS and entropy values of users' profiles, by studying the potential impact of users' privacy policy.

The likelihood to reveal specific attributes varies significantly amongst user profiles and there are some combinations that users may potentially prefer to hide (the dependency between probabilities to reveal pairs of attributes, shown in Table 2 partially illustrates this). One consequence of users' restrictive privacy policies is that other users revealing a rare set of attributes may be more easily identifiable, independently from the values of the attributes (i.e. regardless whether the attribute's values are rare). To verify this claim, we show in Figure 4 the information surprisal IS_{dep} as a function of the P^{rev}_{dep} (note the log scale on the x-axis).

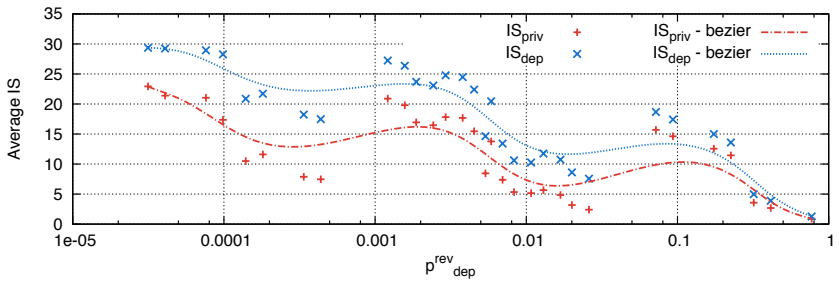

Fig. 4: Average IS as a function of P^{rev}_{dep}

As expected, the lower the probability with which users reveal attributes, the higher the value of IS_{dep}. However, to understand whether the increase in profile uniqueness is contributed by the attributes' values themselves or by the hiding of these attributes (setting restrictive privacy policies), we also include in the figure the values of IS_{priv} denoting the information surprisal of the user's private profiles (i.e. reflecting the profile uniqueness contributed to by all the attribute values). IS_{priv} is obtained from the Facebook Ads platform statistics, for each set of attributes corresponding to a specific value of IS_{dep}. For improved clarity of the figure, we also plot the Bezier approximation of the data.

We again observe the general trend for both IS_{dep} and IS_{priv}: A decrease in P^{rev}_{dep} values corresponds to an increase in average IS_{dep} and IS_{priv} values, which illustrates that independently from the attributes values, the more a set of attributes is hidden the more unique the corresponding profiles will be. In other words, a restrictive privacy policy is also a good identifier of profiles. On the other hand, we highlight an interesting observation that can be made from Figure 4, which shows that the gap between the IS_{dep} and IS_{priv} also increases as P^{rev}_{dep} decreases. This result suggests that the more users tend to hide a combination of attributes, the more identifying this set of attributes will be for other users that do reveal it. The paradox here is that when a combination of attributes becomes rare, due to the majority of user's choice to hide it, it also becomes very identifying when revealed on other profiles.

5 Discussion

Potential extensions As previously mentioned, a frequency-based approach to compute profile uniqueness is dependent of the collected public profiles i.e. the captured set of combinations of attributes. It is therefore impractical, as e.g. it cannot estimate the uniqueness of a profile with a combination of attributes absent from the dataset. An alternative approach could be to adopt a smoothing-based (e.g. Good-Turing [10]) frequency estimation technique to predict the occurrences of the non-observed combinations of attributes, by relying on observed distributions of individual attributes. However, a smoothing-based method still cannot take into account the dependence between the likelihood of revealing specific attributes. In this work, we have deliberately chosen to focus on a fixed set of attributes. Our paper illustrates the extent to which the chosen attributes may identify the users revealing them. Although the results presented in this paper are only applicable to Facebook population, it is important to note that our methodology is general enough to be extended not only to other attributes but also to other online services providing similar statistics platforms, e.g. LinkedIn[18], Yelp[19]. We also did not include users' names in this study, as the application scenarios we consider focus on anonymized profiles where the identity of users is unknown.

[18] `https://www.linkedin.com/ads`
[19] `http://www.yelp.com/advertise`

Addressing possible limitations We stress that in this work we did not investigate in depth the impact of attribute values on the uniqueness of attributes. In Section 4, we have observed the impact of some attribute values (e.g. US in *current country*) on the IS values computed for a single attribute. Such analysis can and should be extended to other attributes and attribute combinations. In this paper we took a first step towards the analysis of profile uniqueness by examining the combinations of attributes that globally, across a selected OSN (i.e. considering a large sample of users), enable re-identification of profiles, independently from whether such re-identification would result from the rare combinations of attributes or the rarity of the attribute values.

Finally, as discussed in Section 2, our use of FAP as a statistical database for private attributes heavily relies on how these statistics are extracted by the OSN operator. Our methodology assumes that the audience statistics are only extracted from *private* profiles (which by definition include public profiles). While Facebook claims that indeed this is the case, a number of other indicators could be used to populate users' (missing) attributes. While this is possible in theory, we believe that this unlikely to happen in practice as the OSNs business model heavily depends on their credibility, and as such attribute inference, which is prone to errors, is unlikely to be adopted.

We acknowledge however that the accuracy of our profile uniqueness estimation is tightly linked to the accuracy of the statistics collected from the FAP.

6 Related work

The study [23] was the first to show that seemingly coarse-grained information such as birth date or ZIP code, if combined, can uniquely identify their owners. Following studies such as [12] emphasize the same finding that: "few characteristics are needed to uniquely identify an individual." Our work complements these by proposing a way to measure the uniqueness of every public profile in a large Online Social Network (e.g. Facebook). Moreover, these studies differ from ours in at least two aspects. First, studied datasets are released by a third-party (e.g., government) who decides which attributes to disclose. This implies "a one rule fits all" approach, where all users are subject to the same privacy policy. Our work considers a dataset where each user has significant control over the revealed data. Second, the targeted populations are significantly different, i.e. US census data versus our world-wide and online population. Hence, our work can be viewed as a new technique to quantify user anonymity in a dynamic environment (e.g. OSNs), where both self-selected and crowd-driven privacy policies impact user anonymity. Moreover, while our approach does not explicitly address user re-identification and record linkage, it can be leveraged to assess the feasibility of such attacks. Specifically, the magnitude of the IS value of a user's profile (in the attacked dataset) can be directly related to the level of user's vulnerability to linkage. In the following, we discuss the works most closely related to ours.

Entropy has been used at various levels: to measure the fingerprint size of a web browser [9], of a host [25] or to track users across multiple services based on their usernames [22].

Privacy leakage in OSN: Numerous works studied the information leakage resulting from OSN use. Krishnamurthy et. al. presented results for both mobile [15] and fixed [16] OSNs. Irani et. al. [13] introduced the online social footprint, consisting of the aggregate information of OSN account owners. [6] investigated the factors that contribute to increased online social footprint, based on cross-OSN public profiles analysis. [7] investigated the evolution of user's privacy policies. A second line of research studied how to exploit publicly revealed information to infer hidden/private data using communities and friendship relations [19, 26] or profile information [5, 17]. Finally, [14] leverages the micro-targeting platform provided by Facebook to target specific users and infer their hidden information. After the publication of [14], several targeting attributes were removed from the Facebook Ads platform and Facebook claims that an ad requires a minimum audience size of 20 to receive any prints. We note that, to the best of our knowledge, none of the work related to privacy issues in micro-targeting platforms considers them as a statistical database.

De-anonymization: Multiple studies explored the feasibility of de-anonymizing both statistical databases [2, 3] and micro-data [4, 20, 21]. Different types of data structures have been targeted, ranging from movie rating data [20] to social network graphs [21]. All these studies share the same (attacker) goal, that is to produce an efficient algorithm that leverages background knowledge to de-anonymize users. Other works utilize obfuscation techniques like differential privacy to ensure privacy of publicly accessible (or released) data [8], [18]. Our work can be seen as a possible aid to techniques which rely on obfuscation, as it provides a measure of privacy and can potentially be utilized e.g. directly by users to control their released data prior to obfuscation.

7 Conclusion

This paper proposes a novel method to compute the uniqueness of public profiles independently from the dataset used for the evaluation. We exploit the Ads platform of a major OSN, Facebook, to extract statistical knowledge about the demographics of Facebook users and compute the corresponding IS and entropy of user profiles. This is used as a metric to evaluate Facebook profile uniqueness and hence the magnitude of the potential risk to cross-link a user with other public data sources. Our findings highlight the relevance of choosing the right combination of attributes to be released in user's public profile and the impact of user's privacy policy on the resulting anonymity.

References

1. How microsoft and yahoo are selling politicians access to you. `http://goo.gl/90d6j`.

2. A. Acquisti and R. Gross. Predicting Social Security Numbers from Public Data. *Proceedings of the National Academy of Sciences*, 106, July 2009.
3. A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In *PODS*, 2005.
4. J. Calandrino, A. Kilzer, A. Narayanan, E. Felten, and V. Shmatikov. You might also like: Privacy risks of collaborative filtering. In *S&P*, 2011.
5. A. Chaabane, G. Acs, and M.-A. Kaafar. You Are What You Like! Information Leakage Through Users' Interests. In *NDSS*, 2012.
6. T. Chen, M. A. Kaafar, and R. Boreli. Is More Always Merrier? A Deep Dive Into Online Social Footprints. In *SIGCOMM WOSN*, 2012.
7. R. Dey, Z. Jelveh, and K. W. Ross. Facebook Users Have Become Much More Private: A Large-Scale Study. In *WSSN*, 2012.
8. C. Dwork. Differential privacy: a survey of results. In *TAMC*, 2008.
9. P. Eckersley. How Unique is Your Web Browser? In *PETS*, 2010.
10. W. A. Gale and G. Sampson. Good-turning Frequency Estimation without Tears. *Journal of Quantitative Linguistics*, 1995.
11. M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou. Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM*, 2010.
12. P. Golle. Revisiting the Uniqueness of Simple Demographics in the US Population. In *WPES*, 2006.
13. D. Irani, S. Webb, K. Li, and C. Pu. Large Online Social Footprints–An Emerging Threat. In *SIAM CS&E*, 2009.
14. A. Korolova. Privacy Violations Using Microtargeted Ads: A Case Study. In *CDMW*, 2010.
15. B. Krishnamurthy and C. Wills. On the Leakage of Personally Identifiable Information via Online Social Networks. *ACM SIGCOMM CCR*, 2010.
16. B. Krishnamurthy and C. E. Wills. Privacy Leakage in Mobile Online Social Networks. In *SIGCOMM WOSN*, 2010.
17. J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham. Inferring private information using social network data. In *WWW*, 2009.
18. F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *KDD*, 2009.
19. A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *WSDM*, 2010.
20. A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. In *S&P*, 2008.
21. A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *S&P*, 2009.
22. D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How Unique and Traceable Are Usernames? In *PETS*, 2011.
23. L. Sweeney. Uniqueness of Simple Demographics in the U.S. Population. *LIDAP-WP4 Carnegie Mellon University*, 2000.
24. L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, Oct. 2002.
25. T.-F. Yen, Y. Xie, F. Yu, R. P. Yu, and M. Abadi. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications. In *NDSS*, 2012.
26. E. Zheleva and L. Getoor. To Join or not to Join: the Illusion of Privacy in Social Networks with Mixed Public and Private user Profiles. In *WWW*, 2009.