# Identity vs. Attribute Disclosure Risks for Users with Multiple Social Profiles

Athanasios Andreou*
EURECOM, France
andreou@eurecom.fr

Oana Goga
MPI-SWS, Germany
ogoga@mpi-sws.org

Patrick Loiseau
EURECOM, France
MPI-SWS, Germany
patrick.loiseau@eurecom.fr

***Abstract*—Individuals sharing data on today's *social computing systems* face privacy losses due to *information disclosure* that go much beyond the data they directly share. Indeed, it was shown that it is possible to infer additional information about a user from data shared by other users—this type of information disclosure is called *attribute disclosure*. Such studies, however, were limited to a single social computing system. In reality, users have identities across several social computing systems and reveal different aspects of their lives in each. This enlarges considerably the scope of information disclosure, but also complicates its analysis. Indeed, when considering multiple social computing systems, information disclosure can be of two types: attribute disclosure or *identity disclosure*—which relates to the risk of pinpointing, for a given identity in a social computing system, the identity of the same individual in another social computing system. This raises the key question: *how do these two privacy risks relate to each other?***

**In this paper, we perform the first combined study of attribute and identity disclosure risks across multiple social computing systems. We first propose a framework to quantify these risks. Our empirical evaluation on a real-world dataset from Facebook and Twitter then shows that, in some regime, there is a tradeoff between the two information disclosure risks, that is, users with a lower identity disclosure risk suffer a higher attribute disclosure risk. We investigate in depth the different parameters that impact this tradeoff.**

## I. Introduction

Individuals publicly share large amounts of data about themselves on *social computing systems* such as Facebook, Twitter, LinkedIn, Reddit, IMDB, and Yelp. Although they receive great utility from those systems, users are also concerned that such data sharing negatively affects their *privacy*; but what exactly is the privacy loss is not always clear. To a large extent, privacy losses relate to *information disclosure*, that is to the information revealed about the user from the data shared. Privacy advocates often argue that making users aware of information disclosure risks could enable them to make better judgements on the data they share.

In the past decade, a large body of research has shown that (hidden) sensitive information about a user such as ethnicity or political affiliation can be inferred by mining publicly available data within a single social computing system [1]–[4]. This type of information disclosure is called *attribute disclosure*: it consists in inferring the value of an attribute (e.g., ethnicity) that was hidden (i.e., not directly shared by the user). All these studies use either *friendship* or *user behavior* data (or both) and exploit homophily to make the inference (i.e., the fact that it is possible based on friendship, or user behavior data to construct groups of users sharing similar attributes).

In parallel, many works appeared in recent years on *matching* identities across multiple social computing systems [5]–[18], that is on building algorithms to find, for a given identity in a social computing system, the identity in a second social computing system that belongs to the same individual (termed the matching identity). This type of information disclosure is called *identity disclosure*. The proposed matching algorithms typically use publicly available attributes (such as name and bio) and leverage the fact that individuals share attributes across social computing systems that might be unique enough to identify them. Indeed, the most recent studies (see [19]) concluded that, in real-world social computing systems, one can precisely pinpoint the matching identity for about 30% of the users.

Surprisingly, few studies seem to have noticed that, in addition to identity disclosure risks, considering multiple social computing systems also introduces significant new attribute disclosure risks due to the possibility of inferring a hidden attribute in a profile by looking at another social computing system (either through homophily or by finding the matching identity). Such attribute disclosure is powerful because individuals reveal different pieces of information on different social computing systems [20] (e.g., personal life on Facebook, profession on LinkedIn, interests on Twitter).

Even more importantly, to the best of our knowledge, no study has *jointly* analyzed identity and attribute disclosure risks. Doing this joint analysis is particularly important because the research community recently gained interest in building defenses against privacy attacks (such as privacy advisors). Defenses were proposed separately in the context of attribute disclosure [21] (warning users when their behavior put them at risk of attribute disclosure, e.g., "liking this will reveal that about you"), and in the context of identity disclosure (advising users to blend into the crowd, that is to share information at a granularity that makes them less uniquely identifiable [22]). However, it is not clear that, in the context of multiple social computing systems where both risks are present, one type of defense also helps against the other type of risk. Intuitively indeed, while blending into the crowd might help against identity disclosure, it might also offer more opportunities to learn attributes and hence increase the attribute disclosure risk. This raises the key questions: *what is the link between the two disclosure risks? does a lower identity disclosure risk always correspond to a lower attribute disclosure risk? do defenses against identity disclosure risk also reduce the attribute disclosure risk?*

In this paper, we perform the first combined empirical study of identity and attribute disclosure across social computing systems and of their relationship. We first propose a framework to quantify the two aforementioned risks, as well as methods for attribute inference from another social computing system in §III. We measure the two risks on a real-world dataset from Facebook and Twitter in §V. Finally, we then investigate in depth how attribute disclosure evolves with identity disclosure using our real-world dataset in combination with an original

*This work was partially done while this author was visiting MPI-SWS.



*ASONAM '17*, July 31 - August 03, 2017, Sydney, Australia

ACM ISBN 978-1-4503-4993-2/17/07?/$15.00
http://dx.doi.org/10.1145/3110025.3110046

semi-synthetic data generation process in §VI.

To the best of our knowledge, this is the first work studying systematically attribute disclosure from another social computing system. More importantly though, our key contribution lies in the analysis of the relationship between attribute and identity disclosure risks. Indeed, we show that, in some regimes, there is a tradeoff between attribute and identity disclosure risks; that is users facing a higher identity disclosure risk face a lower attribute disclosure risk and vice-versa (we also investigate the different parameters that affect this tradeoff). Our findings argue for providing two different risk assessments for users wishing to protect themselves against cross-site linking attacks, and that risk assessment tools and defenses recently proposed for identity disclosure risks do not work well for attribute disclosure.

## II. Background and Related Works

**Disclosure in social computing systems:** Many previous works have shown that one can exploit friendship graphs or the content users provide in order to infer various kinds of information about users such as their location, ethnicity, gender or political preferences [1]–[4], [23]–[34]. *Essentially all of the aforementioned studies show that we can group users together in various ways (based on their public attributes), and then infer additional information about an individual user by studying the properties of the user groups she belongs to.* However, none of the previous works analyzed attribute disclosure across different social computing systems and its relationship to identity disclosure.

A number of recent studies proposed techniques that leverage different attributes of public user data to match the identities users maintain across different social computing systems [5]–[18]. These studies focus on building algorithms that match identities accurately, but to our knowledge, no study has investigated the resulting attribute disclosure.

Finally, to estimate attribute and identity disclosure and study their relation, we need a framework with precise definition of both risks adapted to the context of multiple social computing systems, which no prior work provides. We propose a framework that extends traditional definitions of disclosure risks in databases.

**Privacy definitions in traditional databases:** The database community offers several measures to quantify identity and attribute disclosure risks involved by publishing anonymized databases. First, *k-anonymity* [35] estimates identity disclosure risks by measuring how identifiable a record is in a database. *A record is k-anonymous if it is indistinguishable from at least $k-1$ other records (its anonymity set)*, where two records are said indistinguishable if they have the same quasi-identifiers. k-anonymity guarantees that an adversary who knows the quasi-identifiers of a user will not be able to precisely pinpoint the corresponding record. It is well known that k-anonymity falls short of protecting against attribute disclosure because even if the adversary is not able to exactly pinpoint the targeted record, he can still learn statistical information by analyzing the properties of the anonymity set of the targeted record. To quantify such risks, previous work introduced two extra measures: l-diversity [36], which measures the number of distinct values of an attribute in an anonymity set; and t-closeness [37], which measures the distance between the distribution of attribute values in the anonymity set and in the whole database. Small l-diversity and high t-closeness mean that an adversary is able to learn more precise statistical information about the targeted record.

While the notions of k-anonymity, l-diversity and t-closeness offer attractive concepts to measure attribute and identity disclosure, their definitions in the closed setting of databases are not directly usable in social computing systems such as Facebook and Twitter: *in social computing systems there are no such notions as quasi-identifiers and sensitive attributes*, and any attribute could be considered both a quasi-identifier and sensitive attribute at the same time. As a consequence, no two identities are exactly the same (e.g., even if two identities have the same name and location, it is unlikely that they also have the same likes). The few works that used these concepts in social computing systems are limited to anonymity in the graph structure of a social network (and ignore other information that can be associated with a node such as name, location, interests) [38], [39]. As such, they are still defined in a closed setting and are not directly applicable in open settings such as social computing systems.

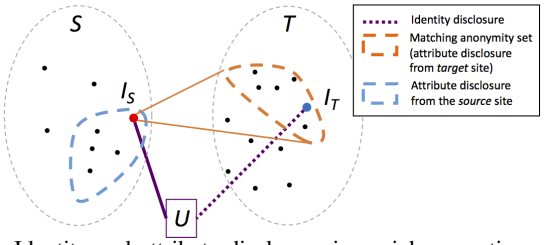


Fig. 1: Identity and attribute disclosure in social computing systems (note that $I_T$ need not be in the matching anonymity set).

Privacy definitions based on k-anonymity (and mechanisms to enforce them) are known to suffer from important limitations, in particular sensitivity to background attacks: guarantees do not hold if the adversary has extra information besides quasi-identifiers. To solve this issue, much recent work used instead the notion of differential privacy [40] which guarantee that, from outputs of queries computed on a database, an attacker cannot infer whether a given user record participated in the computation. The literature proposed mechanisms based on adding noise to the query answer in an interactive setting, or to the data itself before publishing it in a non-interactive setting to guarantee differential privacy given a set of queries [41]. Hence, differential privacy could be a useful tool for the data publisher (in our case the social computing systems) in order to guarantee small disclosure risks (provided that one could transport the definition from the closed database to the open social computing system setting); but it is not directly usable in our study to *measure* the disclosure risks from the given data which is out there. Here, we view the privacy problem from user's perspective and aim to provide her with a framework that can sufficiently capture the disclosure risks (and in particular the relation between attribute and identity disclosure risks) that she faces given the way social computing systems operate and how their data can be linked. To this end, we deem that using a framework based on $k$-anonymity is appropriate and provides useful intuition, despite the known weaknesses of $k$-anonymity.

## III. Framework to Study Disclosure Risks

In this section, we propose measures of the identity and attribute disclosure risks in the context of multiple social computing systems. Our intention is not to introduce a framework to *enforce* privacy in social computing systems, but to create a framework that allows us *quantify*, and subsequently investigate, the tradeoff between attribute and identity disclosure.

### A. Concepts and definitions

**Identity:** An account $I_S$ created in a social computing system $S$ and managed by a user to access services offered by the system. An identity is always associated with a social computing system and a user $U$. We call the identities created by a user across various social computing systems *matching identities*.

**Attributes:** We call all categories of information that can be associated with an identity *attributes*. These attributes can be either *public* or *hidden*. We denote the $j$-th attribute of a user by $a_j$.

### B. Identity disclosure risks

It is important to clarify both *what are* and *how to measure* identity disclosure risks for users with identities across multiple social computing systems. Many previous works have only focused on proposing methods to increase the accuracy of linking identities belonging to the same user across social computing systems without understanding how the accuracy of a linkage scheme actually translates to an *individual*'s identity disclosure risks. Thus, many questions remain unanswered. For example, if a scheme achieves a certain precision and recall to link identities between $S$ and $T$, what does this say about the identity disclosure risk of a given user $U$ that owns $I_S$ and $I_T$? (some users will likely be at higher risks than others). Also, is the identity disclosure risk specific to a user $U$, or to an identity $I_S$ or $I_T$?. Finally, is the identity disclosure risk of a user different when linking identities from $S$ to $T$ than the reverse?.

To answer and clarify such questions, we draw inspiration from previous works on database anonymization. Traditionally, identity disclosure has been defined as the risk of an adversary that has some information $\lambda$ about a user to pinpoint the record corresponding to the user in an anonymized database. Consequently, in the context of social computing systems, *identity disclosure should translate to the risk of an attacker with some information about a particular user (that can be potentially acquired from a social computing systems!) to find his identity $I_T$ in a target social computing system $T$* (see Fig. 1); more precisely, the probability of an adversary knowing the attributes $a_j$ associated with $I_S$ to find the matching $I_T$.

**Measures of identity disclosure risks:** Traditionally, identity disclosure risks have been quantified through k-anonymity. A recent study by Backes et al. [22] proposed to adapt k-anonymity by considering two identities as indistinguishable if they are *similar enough* (rather than having the same quasi-identitifiers) – thus define k-anonymity in social computing systems as the number of identities in $T$ that are similar enough with $I_T$. There are, however, two problems with this measure: (1) it is not clear after which threshold two identities are "similar enough"; and (2) it does not give any guarantees on the probability of an attacker to identify $I_T$ (given $I_S$). Indeed, even if there are multiple identities in $T$ "similar enough" to $I_T$, it is possible that these identities are not "similar enough" with $I_S$ (transitivity does not necessarily hold in social computing systems). Consequently, a large k-anonymity can give a false sense of safety. Contrary, k-anonymity gave guarantees on how easy an adversary can identify a record in a database (if the adversary only knows the quasi-identifiers of a user): if a record has a particular k-anonymity ($k-1$ other users have the same quasi-identifiers), an adversary has a probability of $1/k$ to pinpoint the targeted record. Based on this reasoning, we define *($\theta$, k)-matching anonymity*, an adapted version of k-anonymity for social computing systems that measures the difficulty of an adversary to identify the matching identity:

*Definition 1 ($(\theta, k)$-matching anonymity):* Given a user $U$ and a reference identity $I_S$, we define the *matching anonymity set* of $U$ with respect to a target social computing system $T$ as the set of identities in $T$ that have a probability $p_i$ of matching with $I_S$ higher than $\theta$. An identity has a $(\theta, k)$-matching anonymity if its matching anonymity set is of size $k$.

*Observations:*
$(i)$ The $(\theta, k)$-matching anonymity estimates the risks of an adversary to pinpoint $I_T$. Suppose that the adversary chooses a threshold $\theta$ to declare two identities as matching. Then, if an identity has a $(\theta, k)$-matching anonymity, the adversary will be able to pinpoint the matching identity with a probability of $1/k * recall$. Here, the recall accounts for the probability of the matching identity to be above the threshold $\theta$ while $k$ accounts for the number of identities in $T$ with a probability of matching higher than $\theta$.[1]
$(ii)$ The $(\theta, k)$-matching anonymity (see Fig. 1) can be seen as the projection of $I_S$ on $T$. Thus, we can see that the $(\theta, k)$-matching anonymity measures the identity disclosure risk of $I_T$ with respect to $I_S$ (which is different than the identity disclosure risk of $I_S$ with respect to $I_T$ which would be the projection of $I_T$ into $S$).
$(iii)$ $I_T$ need not be in the matching anonymity set. Whether or not it is depends on how consistent $I_S$ and $I_T$ are.
$(iv)$ The number of identities in the matching anonymity set depends on the uniqueness of $I_S$ with respect to identities in $T$ and not the uniqueness of $I_T$.
$(v)$ The $(\theta, k)$-matching anonymity makes few assumptions about the adversary strategy. We assume that, given two identities, the adversary computes the probability they belong to the same person, and links the identities if the probability is high enough. The threshold adversary is a very common strategy in the literature [19] and our model abstracts the way of computing the probability that two identities belong to the same person. However, our model does not give guarantees against all attackers, e.g., for unreasonable attackers, that link identities at random.

**Typical identity matching strategy:** The typical approach to match identities and compute $p_i$ is to build a binary classifiers that, given two identities $I_S$ and $I_T$, determines whether $I_S$ and $I_T$ are matching or not [5], [7], [8], [10]–[12], [14], [19], [42]. The binary classifier takes as input a feature vector $f(I_S, I_T)$ that captures the similarity between each attribute of a pair of identities $(I_S, I_T)$; and then outputs the probability $p_i$ of $I_S$ and $I_T$ to match. By selecting a cut-off threshold $\theta$ for $p$ the classifier returns 1 (i.e., matching identities) if $p_i$ is larger than the threshold; and 0 otherwise. For the adversary to learn accurate information the matching should be precise, thus the threshold must ensure a small number of false matches.

The choice of $\theta$ corresponds to the adversary's choice of matching accuracy, which entails the standard trade-off between precision and recall.[2] To account for different adversary strategies we will study information disclosure for different $\theta$.

### C. Attribute disclosure risks

Attribute disclosure happens when an adversary is able to learn the value of an attribute associated with a user. Given an identity $I_S$, attribute disclosure can happen either from $S$ (extensively studied by previous works) or $T$ (mostly omitted by previous works), see Fig. 1. We detail next the different ways attribute disclosure can happen from multiple social computing system.

**Attribute disclosure through identity matching:** The simplest kind of attribute disclosure happens when an adversary exploits the public attributes $a_j$ associated with the identity $I_S$ to find its matching identity $I_T$. Then, to learn more information about $I_S$, the adversary can search whether the attributes that are hidden in $I_S$ are public in $I_T$. We call this type of disclosure *attribute disclosure through precise matching*. It is limited by how well an adversary is able to precisely pinpoint the matching identity of a user. As noticed in [19], it is possible to precisely identify the matching identity for a sizable (but limited) number of users in practice (only 30% when matching Twitter to Facebook).

[1]Note that the recall takes into account the fact that a user might not have a matching identity in a social computing system (i.e., it accounts for the overlap between two social computing systems).

[2]Precision quantifies how often the adversary is right when he labels a pair as matching (true matches / predicted matches); and recall quantifies how many pairs of matching identities the adversary is able to detect out of *all* (predicted matches / all matches).

The large literature on vulnerabilities of k-anonymity taught us that attribute disclosure could however happen even if we cannot precisely pinpoint the matching identity. Recall that l-diversity and t-closeness were specifically defined for cases where it is not possible to find the target record. Thus, similarly, we can exploit the set of identities in $T$ with the highest probabilities of matching with $I_S$ (the matching anonymity set) for attribute inference. We call this *attribute disclosure through probabilistic matching*.

**Attribute disclosure through attribute correlation:** In this category we consider all previous works on attribute disclosure through homophily (e.g., people that like X are likely students). However, attribute disclosure through attribute correlation can happen from both the source as well as the target social computing system. In the target site, to exploit the correlation between different attributes (as was traditionally done) we can, for example, build a classifier that given a pair of identities outputs the probability of the two identities to have the same value for an attribute (e.g, have the same country of origin). As for attribute disclosure from probabilistic matching, we can infer the value of an attribute from the group of users in $T$ that have the highest probability to have the same value for the attribute as $I_S$. Note that, to infer attributes from $T$ we have fewer features than for inferring from $S$ (only features that are present in both $S$ and $T$). However, attribute disclosure form $T$ is powerful because individuals reveal different pieces of information about themselves on different social computing systems.

*Observation:* Attribute disclosure through precise matching is easy to control by adjusting the information a user provides in $I_T$. Attribute disclosure from probabilistic matching or correlation depends on what others share, thus it is harder if not impossible to control.

**Measures of attribute disclosure risks:** When inferring attributes the values inferred can be *precise* or *probabilistic* and can also be *correct* or *incorrect*. As noted by Lambert [43], both correct and incorrect inferences can harm the user, thus disclosure is only limited to the extent to which an adversary is *discouraged to make any inference at all*. Since we do not know the point where an adversary becomes discouraged, we use multiple measures of attribute disclosure in both cases where information is correct or not. Thus, we measure the *l-diversity* and *t-closeness* of attribute values in the $(\theta, k)$-matching anonymity set (i.e. users with high probability to have the same attribute value) and the frequency of the most frequent attribute value in the set.

## IV. DATASET

We study disclosure risks across two major social computing systems: Facebook (as source $S$) and Twitter (as target $T$). Besides their popularity, this case study is of particular interests for a number of reasons: (1) users tend to share different kinds of information in each (personal life events on Facebook, interests on Twitter); (2) Facebook is taking measures to limit personal information shared publicly while Twitter doesn't, so a scenario where an attacker leverages information in Twitter to learn more about a Facebook identity is realistic; and (3) due to the availability of the data in Twitter, an attacker can employ state of the art methods to infer very sensitive information about a user ranging from his political views [33], to his interests [44].

To measure information disclosure risks across social computing system we first need ground truth of matching identities in Twitter and Facebook. We obtained a sample of 2,000 pairs of matching Twitter-Facebook identities with limited bias from [19].[3] The authors of [19] collected this data for a study aiming to evaluate the accuracy of matching schemes in practice. Out of the 2,000 pairs of matching Twitter-Facebook identities, we only kept 1,333 for which we were able to infer the location and the interests of the Twitter identity. We call the resulting dataset the MATCHING-DATASET. For each Facebook identity in the MATCHING-DATASET, we collected a candidate set of identities in Twitter that have attributes similar with the Facebook identity: using the Twitter query API, we collected all Twitter identities that have the same or a similar real name, screen name and bio.[4] In total, we collected 7,421,390 Twitter identities and we call them the CANDIDATE-DATASET. 80% of Facebook identities have candidate sets higher than 1,000. A matching scheme should pinpoint the Twitter matching identity out of the candidate set of a Facebook account.

For all Facebook identities in MATCHING-DATASET, we crawled their *about* page and extracted their real names, screen names, location, profile photo, genders, year of birth and bio (the fields work, education, skills, basic-info, bio and quote). Note that not all users provide all these attributes as each user chooses what to make public.In our dataset there are 50% of Facebook identities that provide their location, 4% their age, 75% their gender, and 88% their bio.

For all Twitter identities in MATCHING-DATASET and CANDIDATE-DATASET we used the Twitter API to collect their real name, screen name, location, profile photo, bio, who they follow, and their tweets. As we already mentioned, on Twitter we can use state of the art techniques to make additional inferences about Twitter identities [44]–[46]. For example, to infer the age of a Twitter identity we can use a method that exploits the popularity of given names by decade [46]. Out of all the Twitter identities in MATCHING-DATASET we inferred the age for 45% (the method only works for users in US). To infer the gender of a Twitter identity we can use the *genderize.io* service. We were able to infer the gender for 84% of Twitter identities in MATCHING-DATASET. Finally to infer the interests of a Twitter identity we can use an algorithm that was recently proposed in [44]. At a high level, the method exploits who the user is following and returns a vector of values (a histogram) for how much a user is interested in the following 19 broad interests categories: *arts and crafts, automotive, business and finance, career, education and books, entertainment, environment, fashion and style, food and drink, health and fitness, hobbies and tourism, media, paranormal, politics and law, religion and spiritualism, science, society, sports* and *technology*. In our dataset, in median, users are interested in 8 topics.

In our evaluation we exploit the *real name, screen name, bio* and *profile photo* to link identities between Facebook and Twitter and we attempt to infer the *country, state (for US users), age, gender* and *interests* of users.

**Ethical concerns:** For our evaluation, we only collected publicly available data shared by users in Facebook and Twitter. The data used about matching identities is covered by an IRB. The IRB does not allow sharing the original data, but we can generate anonymized versions to share for the purpose of checking individual results. In addition we consulted a local privacy lawyer, who confirmed that our research is in accordance with our institute ethics guidelines.

## V. EVALUATION OF DISCLOSURE RISKS

In this section we examine identity and attribute disclosure risks for users with identities on Facebook and Twitter (where the adversary knows the identity on Facebook) and how they are affected by $\theta$.

### A. Measuring identity disclosure risks

**Building the matching scheme:** For matching identities we follow the method we described in §III-B. We represent a pair of identities

[3]The strategy is close to picking identities uniformly at random. Please check [19] for more details.

[4]We queried the Twitter API with each unigram and bigram extracted from the bio of the Facebook identity.

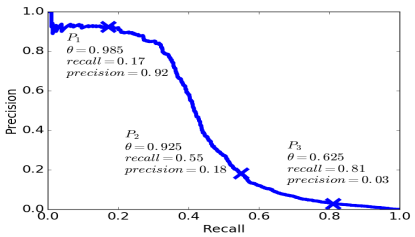
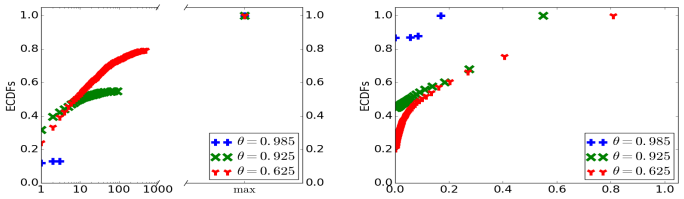



Fig. 2: Precision-recall curve for identifying the Twitter matching identities of Facebook identities in MATCHING-DATASET.




(a) $(\theta, k)$-matching anonymity (b) $\frac{1}{k} \times recall$

Fig. 3: $(\theta, k)$-matching anonymity and probability of pinpointing the matching identity ($\frac{1}{k} \times recall$) for different $\theta$.

$(I_S, I_T)$ with four features, each corresponding to the similarity score between $I_S$ and $I_T$ for a profile attribute. To compute the similarity between real names and screen names we use the Jaro distance [47], for bios we just count the number of common words between the bios of the two profiles and finally, for profile photo we use the Phash [48] and SIFT [49] algorithm to detect whether two photos are the same. We chose SVM as binary classifier and we follow the exact steps proposed in [19] to train the classifier.

**Evaluation of identity disclosure risk:**

For the evaluation, we aggregate the Twitter identities in MATCHING-DATASET (i.e., the true matches) and the identities in the CANDIDATE-DATASET (i.e., the false matches), and we investigate the accuracy of the matching scheme to correctly detect the matching Twitter identity out of the aggregated set. Fig. 2 shows the corresponding precision-recall curve (obtained by varying the threshold $\theta$ on the probability of two identities to match computed by our SVM classifier). We can match 22% of identities with a precision of 90%, but the precision drastically drops afterwards (which is consistent with [19]).

To study how a specific accuracy (i.e., precision-recall) of a matching schemes translates to an individual's identity disclosure risks, and how the risk varies with $\theta$ we pick three thresholds displayed on Fig. 2 corresponding to three main observable states: $P_1$ corresponds to a point with high precision; $P_2$ corresponds to a point where the precision of the matching drops significantly; and $P_3$ corresponds to a point with high recall, but the precision is very low. Fig. 3a shows the ECDFs of the size of the matching anonymity sets for $\theta$ corresponding to $P_1, P_2, P_3$ while Fig. 3b shows the corresponding probability to pinpoint the matching identity (i.e., $\frac{1}{k} \times recall$). If we assume that identities with a $k$ smaller than 10 are at high risk, the figure shows that there are more identities at high risk for larger $\theta$ (about 50% for a $\theta = 0.925$ and $\theta = 0.625$) than lower $\theta$ (less than 20% for $\theta = 0.685$). While counterintuitive, this is happening because when matching identities in practice attackers are not able to achieve high recalls when aiming for high precision [19].

We analyze next to which extent is the identity disclosure risk of a user differ when linking identities from $S$ to $T$ than the reverse. The scatter plot in Fig. 4 presents the relation between the $(\theta, k)$-matching anonymity of $I_T$ (when matching Facebook to Twitter) vs. the $(\theta', k)$-matching anonymity of $I_S$ (when matching Twitter to Facebook), where $\theta$ and $\theta'$ corresponds to a recall of 80%. While for 23% of the users the absolute

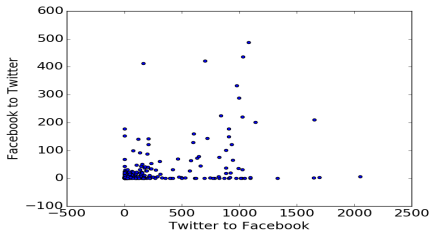
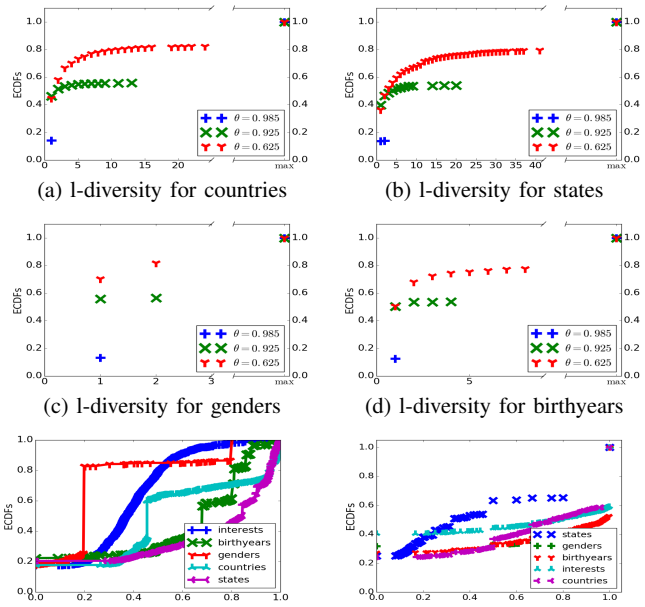



Fig. 4: $(\theta, k)$-matching anonymity for $I_T$ (Facebook to Twitter) vs. $(\theta', k)$-matching anonymity for $I_S$ (Twitter to Facebook).




(a) l-diversity for countries (b) l-diversity for states

(c) l-diversity for genders (d) l-diversity for birthyears

(e) t-closeness for $\theta = 0.625$. (f) confidence for $\theta = 0.625$.

Fig. 5: l-diversity, t-closeness and confidence for different attributes and different $\theta$ corresponding to $P_1$, $P_2$, and $P_3$.

difference between the two $(\theta, k)$-matching anonymity values is $\leq 1$, for most users the risk varies greatly when changing the linkage direction. This shows that one needs to take a twofold approach when attempting to mitigate identity disclosure risks.

### B. Measuring attribute disclosure risks

Since attribute disclosure from attribute correlation has been studied in depth by previous works, we focus in this section on attribute disclosure from probabilistic matching. While there is undeniably a risk of attribute disclosure from probabilistic matching (as there is a risk of attribute disclosure from k-anonymous databases), the interesting question is what is the amplitude of the risk in practice?

Fig. 5 shows the ECDFs of l-diversity, t-closeness and confidence for different attributes computed in the matching anonymity sets of all identity-paris in MATCHING-DATASET for different $\theta$. We have the highest attribute disclosure: (1) if l-diversity is 1 – all identities in the matching anonymity set have the same value for an attribute; (2) if t-closeness is 1 – happens when the distribution of attribute values in the matching anonymity set is very different than the distribution of the whole dataset; and (3) if confidence is 1 – the frequency of the most frequent value in the matching anonymity set is 100% (i.e., everyone has the same attribute value).[5] From Fig. 5 we can again see that, overall,

[5]For cases where the $k$ is zero, we count l-diversity as $max$ (and t-closeness and confidence as 0) as the value of the attribute cannot be inferred.

TABLE I: Precise vs. probabilistic matching ($\theta = 0.625$).

| | birthyears | countries | genders | states |
|---|---|---|---|---|
| **Precise matching** | 0.15 | 0.24 | 0.19 | 0.11 |
| **Probabilistic matching** | 0.45 | 0.45 | 0.65 | 0.25 |

attribute disclosure is higher for lower $\theta$ values. For $\theta = 0.625$ the l-diversity is less or equal to 2 for 60% of identities for country, for 45% of identities for states, and for 70% of identities for age; and the confidence is 1 for 40% to 60% of users depending on the attribute. These results confirm our intuition that we can indeed learn additional information about identities through probabilistic matching.

In terms of accuracy, Table II shows, for each attribute, how often the true attribute value appears in the $(\theta, k)$-matching anonymity set and how often it is actually the most frequent one. For interests, the first row represents the (mean) fraction of the true interests that appear in the $(\theta, k)$-matching anonymity set, and the second represents what (mean) fraction of the interests appearing in the matching anonymity set are actually true interests.[6] In general we can see that for most attributes all the fractions are really high (we make accurate inferences in more than 90% of cases) with the sole exception of states when $\theta$ is small. US states differ because their values are not so concentrated as for other attributes like countries, and when $\theta$ is low and the anonymity sets become much bigger, their distribution is closer to uniform. However, for the same reason it is really frequent for the anonymity sets to contain the true value for US states when they are big.

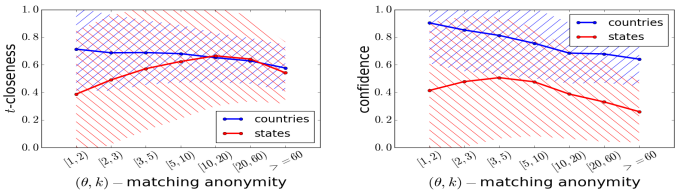
Finally, Table I presents a comparison between precise and probabilistic matching for different attributes. To quantify attribute disclosure through precise matching we analyze for how many Facebook identities we can find a matching Twitter identity with available country, gender, state, age group and interests. To quantify attribute disclosure through probabilistic matching we take a conservative approach: we only measure the fraction of Facebook identities for which we have an l-diversity of 1 (this is a lower bound on attribute disclosure). The figure shows that attribute disclosure through probabilistic matching is significantly higher than through precise matching. The reason behind this result is the fact in practice one can only match a small fraction of identities with high precision; if it would be possible to match all identities, there would not be any interest for probabilistic matching.

## VI. The identity-attribute disclosure risks interplay

In this final section, we investigate how identity and attribute disclosure risks relate to each other in order to answer the question: *do users facing a lower identity disclosure risk (i.e., blending into the crowd) also face a lower attribute disclosure risk?* Due to space constraints, we only show results for countries and states because they cover all interesting observations and other attributes do not bring more insights. For the same reasons, we also focus on attribute disclosure through probabilistic matching rather than correlation.

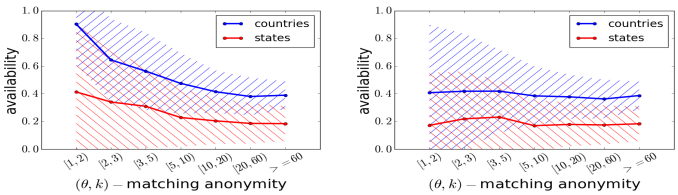
### A. Evaluating the attribute vs. identity disclosure risks

Fig. 6 shows the evolution of $t$-closeness and confidence as a function of $(\theta, k)$-matching anonymity. For countries for both, $t$-closeness and confidence, the attribute disclosure decreases as the anonymity increases, however, for states, $t$-closeness and confidence first increase with $(\theta, k)$-matching anonymity and after a point they start decreasing. That means that, in some cases, *users which are more anonymous suffer a higher attribute disclosure* which indicates that there is indeed a tradeoff between the two types of privacy risks.




(a) t-closeness (b) confidence

Fig. 6: Mean $t$-closeness and confidence area w.r.t. $(\theta, k)$-matching anonymity ($\theta = 0.625$).




(a) When the matching identity is in the sets. (b) When the matching identity is removed from the sets.

Fig. 7: Attribute availability in the matching anonymity set as a function of $(\theta, k)$-matching anonymity ($\theta = 0.625$).

In some sense, this tradeoff is natural: if a user blends into a larger crowd, that gives more opportunities for attribute inference; but then why is the tradeoff not observed for countries? Besides, after reaching a maximum (which we call the knee), the attribute disclosure for states then decreases; why does it start to decrease at this particular point? The rest of the paper investigates the reasons behind the existence or not of the tradeoff (and the knee) and parameters impacting it. Understanding these parameters will allow us to understand to which contexts this tradeoff generalizes.[7]

### B. Factors that determine the tradeoff

We identified three factors that might explain the attribute disclosure: *availability*, *uniformity*, and *correlation*.[8]

*Availability* (in the matching set): At extreme, no information can be learned if no identity in the matching anonymity set provides a particular attribute. However, the question is whether and how availability varies with the $(\theta, k)$-matching set, and what the resulting effect is.

Fig. 7a shows the probability that an identity of the matching anonymity set has the country or state available as a function of the $(\theta, k)$-matching anonymity. We observe that, for countries, the availability decreases sharply, while for states, the decrease is less sharp. The decrease of availability with $(\theta, k)$-matching anonymity is explained (at least partly) by the fact that in our dataset we have selected identity-pairs for which the matching identity has available location. To confirm this, Fig. 7b presents the attribute availability after removing the matching identity from the matching anonymity sets. Indeed, we observe a flat availability which confirms that the higher availability for smaller matching sets previously observed is due to a higher than normal attribute availability of the matching identity.[9] Second, we observe from the figures that states

[6]For interests the l-diversity, t-closeness and confidence do not portray well the disclosure risk as users might be interested in multiple topics, thus, we only present the accuracy of the inference.

[7]The same observations hold for higher $\theta$, but, due to smaller values of k, we do not always observe the decreasing part of the knee.

[8]There might be other factors that play a role, but our results show that they are enough to explain the most interesting properties of the tradeoff.

[9]We verified the availability of countries/states for all identity pairs in the initial dataset –before filtering (see §IV)– and the availability for matching identities is sightly higher than for non-matching identities. Therefore, in practice, the availability will decrease with the $(\theta, k)$-matching anonymity but with a very small slope.

TABLE II: Accuracy of attribute disclosure with probabilistic matching.

| | Countries | | | US states | | | Genders | | | Birthyears | | | Interests | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | **0.985** | **0.925** | **0.625** | **0.985** | **0.925** | **0.625** | **0.985** | **0.925** | **0.625** | **0.985** | **0.925** | **0.625** | **0.985** | **0.925** | **0.625** |
| **True value in anonymity set** | 0.99 | 0.95 | 0.95 | 0.98 | 0.87 | 0.89 | 1 | 1 | 0.99 | 1 | 1 | 1 | 0.99 | 0.84 | 0.71 |
| **True value majority** | 0.99 | 0.89 | 0.84 | 0.98 | 0.75 | 0.58 | 1 | 1 | 0.97 | 1 | 0.98 | 0.93 | 0.99 | 0.88 | 0.83 |

has a globally lower availability than countries.

*Uniformity* (of the global distribution – the distribution of attribute values of a random set of identities in a social computing system): For t-closeness, if the global distribution is very far from uniform, there is less to infer since a lot of information is disclosed *a priori*. The confidence on the other hand does not account for a-priori information, thus, the confidence will be always high for attributes with very non-uniform global distributions. In our dataset, for countries, most users come from the US – 43%, while the median of all countries is 0.02% – which translates to a very non-uniform global distribution, while for states the global distribution is much more uniform. Specifically, the three most frequent states in the global distribution appear 18%, 9%, 8% while the respective median value is 1.2%

*Correlation* (between the attribute to infer and the attributes used for matching): We expect a higher attribute disclosure from probabilistic matching when the attributes used for matching and the attributes to infer are correlated. Our intuition is that names or bios (i.e., attributes used for matching) correlate with countries but less with states.

In order to understand the effect of these parameters on the identity-attribute disclosure tradeoff, we proceed by creating controlled artificial datasets that vary the three parameters.

### C. Artificial datasets to study the tradeoff

*Fixing availability:* We create controlled artificial datasets using sampling with replacement from appropriately constructed distributions to maintain all parameters but availability identical to the original dataset. Specifically, for each Facebook identity with a given location and a given value $k$ of $(\theta, k)$-matching anonymity, we generate an artificial matching anonymity set as follows: $(i)$ For each of the $k$ identities in the matching anonymity set, we draw a Bernoulli random variable with probability equal to the expected availability we want to impose; and $(ii)$ We draw the available locations as follows. For the matching identity (if relevant), we put the location of the Facebook identity. For the others, we draw the locations with replacement from the union of all matching anonymity sets of Facebook identities with the same location in the original dataset.

We generate using this procedure 1000 Facebook identities per $(\theta, k)$-matching anonymity bin, with the same location distribution as in the original Facebook dataset. By construction, each artificial datasets maintains the probability that a Facebook identity contains its matching Twitter identity in its $(\theta, k)$-matching anonymity set and maintains the global attribute distribution; but they allow to control the availability (either keeping it as in the original dataset or setting it to a new, constant, value).

*Fixing uniformity:* To investigate the effect of uniformity on the tradeoff we do not need to generate a separate artificial dataset as we can opportunistically investigate the differences in attribute disclosure as measured by t-closeness and confidence. Indeed, t-closeness measures the EMD distance between the global distribution and the observed distribution of attribute values in a set, thus, eliminating the influence of the global distribution in the measure.

*Fixing correlation:* Fixing correlation without introducing unwanted biases, is a very challenging task. Instead, we compare attribute disclosure from countries (high correlation) and states (low correlation).

### D. Analysis of parameter's impact on the tradeoff

Fig. 8 presents the identity-attribute disclosure tradeoff curve for different imposed availabilities (0.3, 0.5 and 0.8) for countries and states and measured using t-closeness or confidence. To understand the tradeoff we focus on understanding the *position* of the knee on the XY-axes (i.e., what is the $(\theta, k)$ matching anonymity where the maximum attribute disclosure happens); and the *slopes* of the knee (i.e., how steep are the ascending and descending slopes). From the figures we make the following observations:

$(i)$ The first result we clearly see is that, with constant availability, except for high values, we always have an identity-attribute disclosure tradeoff (i.e., a knee) for both countries and states as well as t-closeness and confidence. This was not evident in Fig. 7 for countries, because attribute availability decreased with increasing $(\theta, k)$-matching sets (in part due to our biased way of sampling identity-pairs).

$(ii)$ The position (in terms of x-axis – $(\theta, k)$-matching anonymity) of the knee is strongly impacted by availability: the lower the availability, the further away the knee (see Fig. 8a, 8b, 8c, 8d). Secondly, the position of the knee when comparing t-closeness with confidence is slightly shifted away for t-closeness.

$(iii)$ For the position in terms of the y-axis, for both t-closeness and confidence, the maximum attribute disclosure decreases with availability (see Fig. 8a, 8b, 8c, 8d). For confidence, the maximum attribute disclosure is higher for countries than for states, while for t-closeness the reverse holds. This is just a consequence of the fact that t-closeness takes into account a-priori information and it removes the effect of uniformity of the global distribution.

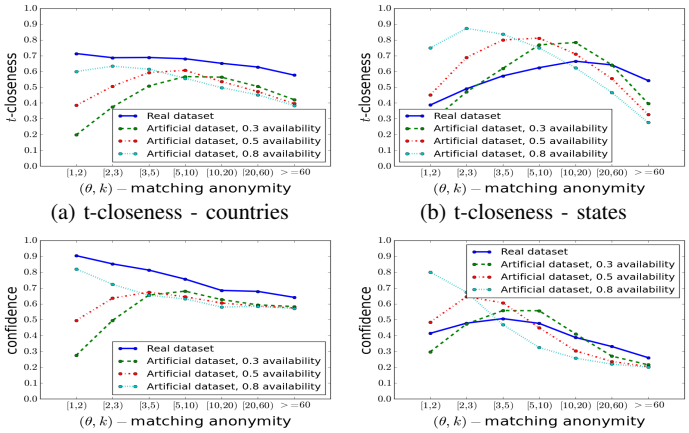
$(iv)$ Regarding the descending slope, for confidence (see Fig. 8c, 8d), we observe the indirect consequence of the uniformity, namely the fact that for countries most users come from US, thus the confidence remains relatively high for high values of $(\theta, k)$-matching anonymity whereas it drops faster for states. The correlation affects the curve in the same direction, for $t$-closeness (see Fig. 8a, 8b) (where the effect of uniformity is removed).

$(v)$ Finally, for the ascending slope, the confidence for a $(\theta, k)$-matching anonymity of one is equal to the availability which is natural since there is only one identity in the matching anonymity set which has location available with a probability equal to the availability. The attribute disclosure increases afterwards because of the increased inference opportunity with a larger matching anonymity set.

We conclude that the effect of uniformity and correlation are second order compared to the effect of availability. The availability imposes the existence of the knee, as well as its position. The uniformity and correlation mostly impact the slope of the descending part of the curve. Results not displayed in the paper due to space constraints show that these conclusions generalize to attribute inference through correlation. Finally, given that the trade-off is mostly imposed by availability, our results will generalize to any other dataset and attribute with availability less than 1. Thus, it is crucial for privacy advisors that aim at limiting privacy risks to take into account attribute availability and its impact on identity and attribute disclosure risks.

## VII. Concluding remarks

In this paper we highlight the existence of different information disclosure risks when reasoning across multiple social computing




(a) t-closeness - countries (b) t-closeness - states

(c) Confidence - countries (d) Confidence - states

Fig. 8: Mean $t$-closeness/confidence w.r.t. $(\theta, k)$-matching anonymity with artificial datasets of various availability ($\theta = 0.625$).

systems. Specifically, we expose the existence of a tradeoff between identity disclosure and attribute disclosure and bring attention to the fact that preventing identity linkage by blending into the crowd is not enough to limit information disclosure and can even make the attribute disclosure risk worse. Additionally, a small $(\theta, k)$-matching anonymity set might be preferable over a large one in some cases as users can have control over the information disclosed trough precise matching whereas they have no control over the information disclosed trough probabilistic matching.

We performed our study on attribute disclosure from the target network only, and using a small set of features available across social computing systems. In future work, we will tackle two interesting generalizations. First, we will investigate the case where the attacker complements attribute inference with inference in the source social computing system. We expect to find similar results because the set of identities to learn from in the source and target social computing system should be of similar size. Second, we will investigate disclosure based on the graph structure of a social computing system, that is when attribute disclosure is done from friends and identity disclosure from graph de-anonymization. Finally, we plan to build a privacy advisor based on our results that will inform users on the possible identity and attribute disclosure risks when creating an account or sharing information on a social computing system.

## ACKNOWLEDGEMENTS

The authors wish to thank Krishna Gummadi for his helpful advice. This work was supported by Institut Mines-Telecom through the "Futur & Ruptures" program and we acknowledge funding from the Alexander von Humboldt foundation.

## REFERENCES

[1] L. Backstrom, E. Sun, and C. Marlow, "Find me if you can: improving geographical prediction with social and spatial proximity," in *WWW*, 2010.
[2] 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.
[3] 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.
[4] J. Chang, I. Rosenn, L. Backstrom, and C. Marlow, "epluribus: Ethnicity on social networks." in *ICWSM*, 2010.
[5] M. Motoyama and G. Varghese, "I seek you: searching and matching individuals in social networks," in *WIDM*, 2009.
[6] D. Perito, C. Castelluccia, M. Ali Kâafar, and P. Manils, "How unique and traceable are usernames?" in *PETS*, 2011.
[7] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, and V. Almeida, "Studying user footprints in different online social networks," in *CSOSN*, 2012.
[8] P. K. Paridhi Jain and A. Joshi, "@i seek 'fb.me': Identifying users across multiple online social networks," in *WoLE*, 2013.
[9] A. Acquisti, R. Gross, and F. Stutzman, "Faces of facebook: Privacy in the age of augmented reality," in *BlackHat*, 2011.
[10] G.-w. You, S.-w. Hwang, Z. Nie, and J.-R. Wen, "Socialsearch: enhancing entity search with social network matching," in *EDBT/ICDT*, 2011.
[11] J. Vosecky, D. Hong, and V. Shen, "User identification across multiple social networks," in *NDT*, 2009.
[12] E. Raad, R. Chbeir, and A. Dipanda, "User profile matching in social networks," in *NBiS*, 2010.
[13] C. T. Northern and M. L. Nelson, "An unsupervised approach to discovering and disambiguating social media profiles," in *MDSW*, 2011.
[14] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Entity matching in online social networks." in *SocialCom*, 2013.
[15] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon, "What's in a name?: An unsupervised approach to link users across communities," in *WSDM*, 2013.
[16] R. Zafarani and H. Liu, "Connecting users across social media sites: A behavioral-modeling approach," in *KDD*, 2013.
[17] ——, "Connecting corresponding identities across communities," in *ICWSM*, 2009.
[18] S. Labitzke, I. Taranu, and H. Hartenstein, "What your friends tell others about you: Low cost linkability of social network profiles," in *SNA-KDD*, 2011.
[19] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi, "On the reliability of profile matching across large online social networks," in *KDD*, 2015.
[20] T. Chen, M. A. Kaafar, A. Friedman, and R. Boreli, "Is more always merrier?: A deep dive into online social footprints," in *WOSN*, 2012.
[21] J. A. Biega, K. P. Gummadi, I. Mele, D. Milchevski, C. Tryfonopoulos, and G. Weikum, "R-susceptibility: An ir-centric approach to assessing privacy risks for users in online communities," in *SIGIR*, 2016.
[22] M. Backes, P. Berrang, O. Goga, K. Gummadi, and P. Manoharan, "On profile linkability despite anonymity in social media systems," in *WPES*, 2016.
[23] D. Jurgens, "That's what friends are for: Inferring location in online social media platforms based on social relationships." in *ICWSM*, 2013.
[24] P. Wang, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Your cart tells you: Inferring demographic attributes from purchase data," in *WSDM*, 2016.
[25] J. Otterbacher, "Inferring gender of movie reviewers: Exploiting writing style, content and metadata," in *CIKM*, 2010.
[26] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen, "Demographic prediction based on user's browsing behavior," in *WWW*, 2007.
[27] N. Z. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," in *USENIX Security*, 2016.
[28] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla, "Inferring user demographics and social strategies in mobile social networks," in *KDD*, 2014.
[29] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie, "You are where you go: Inferring demographic attributes from location check-ins," in *WSDM*, 2015.
[30] A. Chaabane, G. Acs, M. A. Kaafar *et al.*, "You are what you like! information leakage through users' interests," in *NDSS*, 2012.
[31] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang, "Towards social user profiling: unified and discriminative influence model for inferring home locations," in *KDD*, 2012.
[32] K. Ryoo and S. Moon, "Inferring twitter user locations with 10 km accuracy," in *WWW*, 2014.
[33] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in *SMUC*, 2010.
[34] Y. DONG, N. V. CHAWLA, J. TANG, and Y. YANG, "User modeling on demographic attributes in big mobile social networks."
[35] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002.
[36] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, "l-diversity: Privacy beyond k-anonymity," *TKDD*, 2007.
[37] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *ICDE*, 2007.
[38] A. Campan and T. M. Truta, "Privacy, security, and trust in kdd," in *Data and Structural k-Anonymity in Social Networks*, 2009.
[39] B. Zhou and J. Pei, "The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks," *KAIS*, 2011.
[40] C. Dwork, "Differential privacy," in *ICALP*, 2006.
[41] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, 2014.
[42] M. A. Mishari and G. Tsudik, "Exploring linkability of user reviews," in *ESORICS*, 2012.
[43] D. Lambert, "Measures of disclosure risk and harm," *Journal of Official Statistics*, 1993.
[44] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi, "Inferring user interests in the twitter social network," in *RecSys*. ACM.
[45] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. N. Rosenquist, "Understanding the demographics of twitter users." in *ICWSM*, 2011.
[46] http://goo.gl/G4ZLgA, accessed: 2017-06-22.
[47] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string distance metrics for name-matching tasks," in *IIWeb*, 2003.
[48] "Phash," http://www.phash.org.
[49] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.