Theoretical Results on De-Anonymization via Linkage Attacks

Martin M. Merener[*]

[*]York University, N520 Ross, 4700 Keele Street, Toronto, ON, M3J 1P3, Canada.

E-mail: merener@mathstat.yorku.ca

Abstract. Consider a database *D* with records containing history of individuals' transactions, that has been de-identified, i.e., the variables that uniquely associate records with individuals have been removed from the data. An adversary de-anonymizes *D* via a linkage attack if using some auxiliary information about a certain individual in the database, it can determine which record of *D* corresponds to such individual.

One example of this is given in the article Robust De-anonymization of Large Sparse Datasets, by Narayanan and Shmatikov [19], which shows that an anonymized database containing records with ratings of different movies rented by customers of Netflix, could in fact be de-anonymized using very little auxiliary information, even with errors. Besides the heuristic de-anonymization of the Netflix database, Narayanan and Shmatikov provide interesting theoretical results about database de-anonymization that an adversary can produce under general conditions.

In this article we revisit these theoretical results, and work them further. Our first contribution is to exhibit different simple cases in which the algorithm *Scoreboard*, meant to produce the theoretical de-anonymization in [19], fails to do so. By requiring $1 - sim$ to be a pseudo-metric, and that the algorithm producing the de-anonymization outputs a record with minimum support among the candidates, we obtain and prove de-anonymization results similar to those described in [19].

We then consider a new hypothesis, motivated by the fact (observed in heuristic de-anonymizations) that when the auxiliary information contains values corresponding to rare attributes, the de-anonymization achieved is stronger. We formalize this using the notion on *long tail* [4], and give new theorems expressing the level of de-anonymization in terms of the parameters of the tail of the database *D*. The improvement in the de-anonymization is reflected in the fact that when at least one value in the auxiliary information corresponds to a rare attribute of *D*, the size of auxiliary information could be reduced in about 50%, provided that *D* has a long tail.

We then explore a microdata file from the Joint Canada/United States Survey of Health 2004 [22], where the records reflect the answers of the survey respondents. While many of the variables are related to health issues, some other variables a related to characteristics that individuals may disclose easily, such as physical activities (sports) or demographic characteristics. We perform an experiment with this microdata file and show that using only some non-sensitive attribute values it is possible, with a significant probability, to link those values to the corresponding full record.

Keywords. Linkage attack, de-anonymization, database privacy, Netflix dataset

1 Introduction

Anonymization, also referred to as de-identification, is often used to provide privacy before publishing a set of records corresponding to individuals. It consist in not releasing the variables that uniquely associate records to the corresponding individuals, such as name, email address, social security number, and so on. This mechanism often fails in providing privacy, because the rest of the variables present in the data could work jointly as a fingerprint, despite the fact that each one isolated does not carry enough information to do so.

This is eloquently expressed in [18]: "While some attributes may be uniquely identifying on their own, any attribute can be identifying in combination with others. For example, the books a person has read: while no single element is a (quasi)-identifier, any sufficiently large subset uniquely identifies the individual." As a result, supposedly anonymized datasets can sometimes be de-anonymized by an adversary with access to auxiliary information [1, 17, 19, 21], producing privacy breaches.

One type of de-anonymization is called *linkage attack*. Consider a database D with records containing history of individuals' transactions, that has been de-identified, i.e., the variables that uniquely associate records with individuals have been removed from the data. An adversary de-anonymizes D via a linkage attack if using some auxiliary information about a certain individual in D, it can determine which record in D corresponds to such individual. Considering the increasingly available data on the Internet and in digital media [14], this scenario turns out to be realistic.

Such attacks are remarkable for the small amount of auxiliary information that the adversary needs in order to succeed. One example of this is given in the article Robust De-anonymization of Large Sparse Datasets, by Narayanan and Shmatikov [19], which shows that an anonymized database published by Netflix in www.netflixprize.com—meant to be used to improve the performance of the Netflix recommender system—could in fact be de-anonymized using very little auxiliary information, even with errors.

The Netflix database consists of records with ratings (and date of the ratings) of different movies rented by customers of Netflix (a DVD rental company). The performance of the de-anonymization in [19] was measured using auxiliary information coming from the same database, with introduced errors. They also show that the auxiliary information can be taken from the Internet Movie Database (IMDb), where users provide reviews and ratings of movies. While the identity of a user in IMDb may be public by choice (the user's profile could include the real name), this user may want to be anonymous in the Netflix database, for instance because it contains a larger history of movie rentals. So, by linking a record of IMDb to a record of the Netflix database (both corresponding to the same person), it is possible to reveal information that such a person has chosen to keep private.

Besides the heuristic de-anonymization of the Netflix database, Narayanan and Shmatikov also provide interesting theoretical results about the de-anonymization of databases that can be done under general conditions. These results apply to a large class of databases, and express the level of de-anonymization in terms of global parameters of the database (size and sparsity) and of the auxiliary information (accuracy).

In this article we start by focusing on these theoretical results, and work them out further. Our main contribution is to establish and prove mathematical results (Theorems 4, 6, 8, 9, 10 and 11) describing the de-anonymization that can be achieved by an adversary under general and realistic assumptions.

Our starting point is Theorem 1 in [19]. We describe some technical issues in its proof, which uses a specific adversary algorithm (different from the one used in the heuristic de-anonymization of the Netflix database). We believe that this algorithm does not produce the de-anonymization meant by the theorem. We consider an extra hypothesis on the similarity function *sim*, and change the adversary algorithm slightly, after which we prove a result analogous to Theorem 1 in [19].

Then we consider a hypothesis about how sparse the database D is, and show that in general, the de-anonymization improves due to sparsity (which is already observed in [19]). Also, we evaluate our theoretical results on the Netflix database, and compare—through the size of the auxiliary information denoted by m—the heuristic de-anonymization in [19], against the de-anonymization that our results guarantee (the former being more efficient, although both give values for m with the same order of magnitude).

Next we consider a new theoretical hypothesis, motivated by the fact—observed in the heuristic de-anonymization—that when the auxiliary information contains values corresponding to rare attributes, the de-anonymization achieved is stronger. We formalize the needed concepts using the notion on *long tail* from [4], and give new theorems expressing the level of de-anonymization in terms of the parameters of the tail of the database *D*. The improvement in the de-anonymization theorems is reflected on the fact that when the database *D* has a long tail, the size of the auxiliary information could be reduced to the half of what it is when it is not known if values correspond to rare attributes.

Then we work with a concrete microdata file produced by the Joint Canada/United States Survey of Health 2004 [22]. This database is less sparse than the databases with large number of attributes and records, so to guarantee high level of adversary success via the established theorems, the adversary is required to use large auxiliary information. Alternatively, we evaluate a linkage attack throughout an experiment and show that using significantly less (than the theorems) values of non-sensitive attributes, the adversary can successfully perform a linkage attack.

1.1 Related Work

The general problem in the field of Private Data Analysis (also known as Statistical Disclosure Control) is how to release useful information about a set of individuals, while preserving their privacy. To answer this question, one approach is to create mechanisms that sanitize the data before its publication [2, 3, 6, 8, 9, 10, 12, 16], and give empirical or theoretical evidence of the fact that the sanitized data has the desired properties regarding privacy. These are positive results in the sense that say how it is possible to provide privacy.

Other valuable contributions to Privacy Data Analysis come in the form of impossibility results, showing the limitations that commonly used mechanisms have. These results tell how *not* to release information if privacy wants to be preserved, which helps to re-design and improve the known sanitizers. These results usually consist in showing an explicit attack to a sanitizer. The adversary algorithm representing the attack would take the sanitized data as input, and produce an output that represents a privacy threat. Some of these attacks use only sanitized data as input [9, 11], while others use auxiliary information (separate from the sanitized data) [1, 17, 19].

The article [13] is a predecessor of [19], which uses the same scoring paradigm to re-identify users from a public web movie forum in a private movie ratings database. The work done in [15] shows that people's gender, ZIP code and full date of birth allow unique identification of 63% of the U.S. population, which means that any sensitive information joining these three variables could be linked to individuals' identities. In [1], the authors use files made public by the Social Security Administration, to determine the correlation between Social Security numbers and place and date of birth, which is data usually available in social networking sites. Many recent articles deal with de-anonymizations of social graphs [5, 17, 23, 24, 25], based on the topology of the graphs, *friendship* relationship, and group membership.

1.2 Organization of the Paper

Section 2 starts with the motivation of the problem, followed by some of the definitions needed throughout the paper, and simple remarks. In section 3 we review the main theoretical result in [19] about de-anonymization of databases, and expose three specific problems with its proof.

These problems induce changes that allow us to prove new de-anonymization results in section 4.1 (Theorems 4 and 6) and section 4.2 (Theorem 8). In section 4.3 we evaluate these results on real databases to determine the levels of de-anonymization that the results guarantee, and compare them with the heuristic de-anonymization in [19].

Section 4.4 contains new results on de-anonymization (Theorems 9 and 10) that take into account the distribution of the support size of the attributes of the database, and increase the level of de-anonymization of databases having a long tail.

In section 4.5 we work with a microdata produced by the Joint Canada/United States Survey of Health 2004. We perform a linkage attack using simulated auxiliary information coming from the same database, and compare the rate of success in this experiment to the adversary success ensured by one of the proved theorems. Section 4.6 explores an alternative way (Theorem 11) of estimating the probability of de-anonymization.

2 Preliminaries

2.1 Motivation

In our context, a database is a table with rows corresponding to individuals and columns corresponding to attributes. For example, a database representing the books that each person bought online from Amazon. Such a database would have many empty entries since the set of books bought by each person is typically small compared to all the books available in a catalogue. Empty entries are represented by the symbol $\perp$ (*null*), and occur whenever an individual does not manifest a certain attribute, which is common among databases reflecting histories of transactions, as well as surveys.

The Netflix database contains $\approx$ 100 million ratings from 1 to 5 (and the corresponding dates of the ratings) about $17,770$ movies, provided by $\approx$ 480,000 Netflix users. Note that 100 million is less than 1.2% of $17,770 \times 480,000$, so most of the entries of the Netflix database are null.

Ratings 1-5 (and Dates) on DVDs Rented From Netflix

Customer ID	Toy Story	Titanic	...	Brazil
ID$_1$	2 (10/03/01)	$\perp$	...	$\perp$
ID$_2$	$\perp$	1 (12/04/04)	...	4 (15/01/02)
	...	...	...	...
ID$_n$	5 (25/12/08)	3 (22/08/10)	...	$\perp$

This is an example of the kind of database on which our results are meant to be applicable. Another example could be a database detailing when web pages are visited by users of an Internet service provider:

Web Pages Visited by AOL Users

User ID	gmail.com	cnn.com	...	ebay.com
ID$_1$	...	...	...	...
...	...	...	...	...
ID$_n$	...	...	...	...

While databases like these are published and used for research and commercial purposes, it is likely that some of their attributes are sensitive for some of the individuals contributing their data. There are two seemingly opposing interests—utility and privacy—and our goal is to understand better how this trade-off works.

To model this general situation consider two parties interacting, represented by algorithms: the sanitizer and the adversary. The sanitizer algorithm takes a database as input, and outputs a sanitized version (to be published) meant to protect the privacy of the individuals who contributed their data to the original database. The adversary, in turn, takes the sanitized database and uses it to attempt a privacy breach.

To preserve privacy, the least thing that the sanitizer should do is to eliminate all the attributes that directly match individuals to records, such as full name, email address and social security number. This is called *de-identification* or *anonymization*.

The next question is if this is enough to preserve privacy. Under realistic assumptions it is not enough, and for an illustrative example, consider the two following databases [21].

De-Identified Database: Medical Conditions

Gender	*Postal code*	*Date of birth*	**Description**
...	...	...	...

This anonymous database describing (sensitive) medical conditions could be the input of the adversary algorithm. It is realistic to assume that the adversary has also access to auxiliary databases, such as:

Auxiliary Database: Voter List

Gender	*Postal code*	*Date of birth*	**Name**
...	...	...	...

If an individual who has records in both databases happens to have a unique value for $v =$(*gender, zip code, date of birth*), then v becomes an identifier that links both records. In that case, the name of such individual can be linked to the sensitive attribute *description* in the anonymized database. Individuals with unique values for v are not rare. In [15], the authors show that 63% of the U.S. population is expected to have unique values for $v =$(*gender, zip code, date of birth*).

The auxiliary information that the adversary has access to can be seen also as a collection of values of certain attributes, among which there could be an identifier. As in the example with the voter list, attributes of this kind are not used to link the auxiliary information to the corresponding record in the anonymized database (only *gender*, *postal code* and *date of birth* are used to link the databases).

Hence, to study the linkage attacks we can ignore identifiers, as well as any other attributes in the auxiliary information that are not attributes in the anonymized database.

So we simply assume that the attributes in both the auxiliary information and the anonymized database are the same and come in the same order, and we model them as arrays. The Netflix database would look like this:

$$D = \begin{pmatrix} 2\ (10/03/01) & \dots & \perp \\ \perp & \dots & 4\ (15/01/02) \\ \dots & \dots & \dots \\ 5\ (25/12/08) & \dots & \perp \end{pmatrix}$$

The auxiliary information is modeled as a *subrecord* (see Definition 3) of a perturbed record of D. This is, for a record r in D, r' represents some perturbation of r, and the auxiliary information $aux(r)$ is given by some of the values (not all of them) of r'. In this way, the adversary algorithm receiving the input $(aux(r), D)$, attempts to find (or approximate) the record r in the de-identified database D.

2.2 Definitions

Notation. If $m \in \mathbb{N}$, then $[m] = \{1,...,m\}$. We denote the set from where entries of matrices and vectors are taken by $\mathbb{V}$, which contains the element $\perp$, called *null*.

Definition 1. *Given an array $M \in \mathbb{V}^{n\times c}$, its* support *is $supp(M) = \{(j,i) \in [n]\times[c] : M_{j,i} \neq \perp\}$. If x is a row of M, then $supp(x) = \{i \in [c] : x_i \neq \perp\}$, and if $i \in [c]$ denotes a column, then $supp(i) = \{j \in [n] : M_{j,i} \neq \perp\}$. Two arrays M and M' of the same size* have the same support *if $supp(M) = supp(M')$, which means that the null values (if any) appear in the same places in both arrays.*

Definition 2. *If $n,c \in \mathbb{N}$, a* database *of size $n\times c$ is $D \in \mathbb{V}^{n\times c}$. We reserve D to denote databases and n and c to denote their sizes. Each row of D represents an* individual *and each column an* attribute*. Rows of D, also called* records*, are denoted by x,y,z, or by r to denote a particular record. Attributes are labeled by $i \in [c]$. That x is a record of D (i.e., a row of D), is denoted by $x \in D$ (by abuse of notation we sometimes treat D as the set of its rows).*

Definition 3. *Let D be of size $n\times c$, and let $m \in \mathbb{N}$, with $m \leq c$. D is m-*supported *if $\forall x \in D$, $|supp(x)| \geq m$. D is* fully supported *if it is c-supported. We say that x is a* subrecord *of $y \in \mathbb{V}^c$ if $\forall i \in supp(x)$, $x_i = y_i$; we denote this by $x \subseteq y$.*

Definition 4. *A* similarity function (on values) *is a function $sim : \mathbb{V}\times\mathbb{V} \to [0,1]$ satisfying:*
(i) $\forall a \in \mathbb{V}$, $sim(\perp,a) = 0$
(ii) $\forall a \in \mathbb{V}\setminus\{\perp\}$, $sim(a,a) = 1$
(iii) Symmetry: $\forall a,b \in \mathbb{V}$, $sim(a,b) = sim(b,a)$.

For some results we require that $1 - sim$ has to satisfy the triangle inequality over $\mathbb{V}\setminus\{\perp\}$. This means that if $d(x,y) = 1 - sim(x,y)$, then $\forall x,y,z \in \mathbb{V}\setminus\{\perp\}$, $d(x,y) \leq d(x,z) + d(z,y)$. This makes of $1 - sim$ a *pseudometric* over $\mathbb{V}\setminus\{\perp\}$, i.e., a metric, except for the fact that $d(x,y) = 0 \Rightarrow x = y$.

Remark 1. *A simple way of producing a sim function such that $1 - sim$ is a (pseudo)metric, would be to take a (pseudo)metric d and define $sim(x,y) := 1 - d(x,y)$. However, since we need $sim \geq 0$, we take instead $sim(x,y) := \max(1 - d(x,y), 0)$. In that case, $1 - sim(x,y) = \min(d(x,y), 1)$, which is a bounded (pseudo)metric.*

Unless we specify *sim*, the results we present apply to an abstract *sim*, i.e., any function satisfying Definition 4.

Definition 5. *A* similarity function (on records) *is a function $Sim : \mathbb{V}^c\times\mathbb{V}^c \to [0,1]$ such that $\forall x \in \mathbb{V}^c$, $Sim(x,x) = 1$.*

Although the definition of *Sim* is general, most of the results we give refer to a specific *Sim* given by a formula depending on a certain *sim* function (abstract or not). Note the use of “s” and “S” to distinguish between similarity functions on values and on records. The following are three *Sim* functions that we use in different results (assume $\mu \in (0,1]$).

$$Sim_\mu(x,y) = \frac{|\{i : sim(x_i,y_i) \geq \mu\}|}{|supp(x)\cup supp(y)|} \tag{2.1}$$

$$Sim(x,y) = \frac{\sum_i sim(x_i,y_i)}{|supp(x)\cup supp(y)|} \tag{2.2}$$

$$Sim(x,y) = \frac{|supp(x)\cap supp(y)|}{|supp(x)\cup supp(y)|} \tag{2.3}$$

It is straightforward to verify that the above equations define similarity functions on records, provided that *sim* is a similarity function on values.

Note that (2.1) becomes (2.3) when *sim* is defined by:

$$sim(a,b) = \begin{cases} 1 & \text{if } a \neq \perp \text{ and } b \neq \perp \\ 0 & \text{otherwise} \end{cases}$$

Definition 6. *Given $h > 0$, two databases D and D' of the same size are* h-similar *(with respect to sim) if $\forall j, i$ such that $d_{j,i} \neq \perp$ $\underline{\text{or}}$ $d'_{j,i} \neq \perp$, it holds that $sim(d_{j,i}, d'_{j,i}) \geq h$.*

Remark 2. *If D and D' are h-similar ($h > 0$), then D and D' have the same support. This follows from (i) in Definition 4: $\forall a \in \mathbb{V}$, $sim(\perp, a) = 0$.*

Definition 7. *Given a database D, an* auxiliary database *is a database D' with the same size and the same support of D. If $x \in D$, then x' denotes the corresponding record (same row) of D'. We reserve D' to denote auxiliary databases.*

Note that D' is not the auxiliary information of the adversary, but it is used to define it as we explain next.

The database D' represents a perturbation of D, from which the adversary observes only *some* of the values of a record r', based on which it tries to construct an approximation of r. To accomplish this, the adversary has access to the whole database D, so its goal consists in finding the record r in D, given the values of **some** coordinates of r' (a perturbation of r).

Notation. If $x \in \mathbb{V}^c$ and $s \subseteq [c]$, then $x|_s \in \mathbb{V}^c$ is defined by: $\forall i \in s$, $(x|_s)_i = x_i$, and $\forall i \notin s$, $(x|_s)_i = \perp$.

Definition 8. *An* adversary *is an algorithm that has input $(r'|_s, D)$, where $r \in D$ of size $n \times c$, D' is an auxiliary database, and $s \subseteq supp(r)$. The vector $r'|_s$ is the* auxiliary information *of the adversary. The output of the adversary algorithm is a vector in $\mathbb{V}^c$ meant to be highly similar to r.*

Defining auxiliary information as some of the coordinates of r' (those corresponding to s), reflects the fact that the adversary observes some particular values (possibly with errors) of the target record r.

Although we allow $D' \neq D$, we assume that D and D' have the same support. As a consequence, if an entry is non-null in $r'|_s$, then the corresponding entry in r must be non-null as well. For example, if the auxiliary information has a rating for a certain movie, then the corresponding record in D must have also a rating (maybe different) for that movie. The converse is true: if an individual rated a movie in the anonymized database D, then the auxiliary information $r'|_s$ may either give the adversary a different rating, or omit whether that individual have rated such a movie (in case the movie is not an attribute in s).

Remark 3. *For any $x \in D$, and any $s \subseteq [c]$: $supp(x|_s) \subseteq s$. Also: $supp(x|_s) = s \Leftrightarrow s \subseteq supp(x)$. In particular, that D and D' have the same support implies $s \subseteq supp(r) = supp(r')$, which combined with $(\Leftarrow)$ gives $supp(r'|_s) = s$.*

The auxiliary information could be associated to an identifier, so a breach of privacy occurs if the adversary links the auxiliary information to the record to which it corresponds in D (or a record equal to it).

The definition of *de-anonymization* that follows describes an adversary algorithm A, which on input given by D and the auxiliary information $r'|_s$, outputs a record that is similar to r. The interest is not upon a specific record r, but on the overall fraction of records for which a high similarity (parameterized by σ) is achieved. Thus, we consider a random variable (r.v.) R selecting random records from D, and a parameter p that measures the fraction of records that are at least σ-similar to the adversary's guess.

Now, for each record r, we have to consider the auxiliary information $r'|_s$ on which the adversary's guess is based. The set s is assumed to have size m and to be chosen uniformly at random among all the subsets of $supp(r)$ with m elements. Let S_r be the r.v. that outputs such a set s. Notice that once $r \in D$ is fixed, the random variables S_r and R are (probabilistically) independent. For simplicity we denote S_r just by S (the corresponding r should be clear from the context).

Definition 9. *Given $m \in \mathbb{N}$, $\sigma \in (0,1)$ and $p \in (0,1)$, an m-supported database D can be (σ, p, m)-*de-anonymized with respect to *Sim* and D'*, if there exists an adversary A such that:*

$$P[Sim(R, A(R'|_S, D)) \geq \sigma] \geq p \tag{2.4}$$

where R is a uniform r.v. selecting a record $r \in D$, and S is a uniform r.v. selecting a subset of $supp(r)$ with m elements. The probability is taken with respect to R, S and the random choices that A makes (if any). When (2.4) occurs, we say that D is (σ, p, m)-de-anonymized by A, or that A (σ, p, m)-de-anonymizes D (with respect to Sim and D').

That S is a uniform r.v. selecting a subset of $supp(r)$ with m elements means that each one of the $\binom{|supp(r)|}{m}$ sets has probability $1/\binom{|supp(r)|}{m}$ of being chosen.

Notation. We do not specify the size of D in every statement, but it is always assumed to be $n \times c$. We may omit saying that D is m-supported, but this is always assumed in statements in which a random subset s of size m is taken from $supp(r)$, for an arbitrary r. We always assume $\sigma, p \in (0,1)$ and $m \leq c$. We omit saying with respect to which *Sim* and D' is the de-anonymization in question, but this should be clear from the context.

3 A Theorem on De-Anonymization

Theorem 1 in [19] gives hypotheses under which a certain degree of de-anonymization is possible. The proof of the theorem shows a particular adversary that is meant to produce such a de-anonymization. One of the appealing aspects of the theorem is that it deals with two realistic situations: (i) the database D is not fully supported (Definition 3), i.e., entries of D are allowed to be null; and (ii) the auxiliary information is not a subrecord of an existing record (Definition 3), so $D' \neq D$.

However, it is precisely in these two situations that the proof of Theorem 1 in [19] fails. Our goal is to continue the work initiated in [19] by producing formal results on de-anonymization such as Theorem 1. The first thing we do is to point out the technical problems in its proof, and then we fix them by introducing an extra hypothesis on *sim*, and modifying the adversary algorithm.

Theorem 1 [19]. Let $\varepsilon, \delta \in (0,1)$ and let D be a database. Let *Aux* be such that $aux = Aux(r)$ consists of $m \geq \frac{\log(n/\varepsilon)}{-\log(1-\delta)}$ randomly selected non-null value attributes of the target record r, and $\forall i \in supp(aux)$, $sim(aux_i, r_i) \geq 1 - \varepsilon$. Then D can be $(1-\varepsilon-\delta, 1-\varepsilon)$-de-anonymized with respect to *Aux* and $Sim(x,y) = \frac{\sum_i sim(x_i, y_i)}{|supp(x) \cup supp(y)|}$.

Notation. In Theorem 1, both $Aux(r)$ and aux mean $r'|_s$ (as in Definition 8). With our notation, Theorem 1 says that when $m \geq \frac{\log(n/\varepsilon)}{-\log(1-\delta)}$ and D and D' are $(1-\varepsilon)$-similar, then there exists an adversary algorithm A such that $P[Sim(R, A(R'|_S, D)) \geq 1-\varepsilon-\delta] \geq 1-\varepsilon$, where the probability is taken as in Definition 9, and with *Sim* given by equation (2.2).

The conditions in Definition 4 given for *sim* are also assumed in [19], but there $1 - sim$ is not required to be a pseudometric. In fact, the *sim* function that the authors use for the heuristic de-anonymization of the Netflix database is:

$$sim(a,b) = \begin{cases} 1 & \text{if } a \neq \perp, b \neq \perp \text{ and } |a-b| < \rho \\ 0 & \text{otherwise} \end{cases}$$

with which $1 - sim$ is not a pseudometric (it does not satisfy the triangle inequality when $\rho > 0$). We will show that with this *sim* function there are some technical issues with the proof of Theorem 1.

The adversary algorithm *Scoreboard*, used in [19] to prove Theorem 1 consist of two steps. Given the input (aux, D):

1. Compute the set $D_{aux} = \{y \in D : \forall i \in supp(aux), sim(aux_i, y_i) \geq 1 - \varepsilon\}$

2. Choose an element of D_{aux} uniformly at random and use it as output.

Note that $D_{aux} = \{y \in D : score(aux,y) \geq 1 - \varepsilon\}$, where $score(aux,y) = \min_{i \in supp(aux)} sim(aux_i, y_i)$.

3.1 Cases Where *Scoreboard* Fails De-Anonymizing as Meant by Theorem 1

The following cases are meant to point out technical problems with the proof of Theorem 1 in [19]. They should be seen as a way of testing the proof, rather than real-world examples. They are quite simple, so the formal results on de-anonymization should contemplate them. The issues with the proof do not imply that Theorem 1 is not true, and in fact, these cases suggest how to fix the problems so that results similar to Theorem 1 can be established.

In cases 1 and 2, D is fully supported, so none of its entries are null. However $D' \neq D$, so the auxiliary information has errors with respect to the original records. In case 3 the auxiliary information has no error ($D' = D$), but D has null values. For the sake of clarity, these cases are given for specific values of n, c, ε, δ, D and D', but similar examples work for many other values (for instance, for any n even).

Case 1. Let $n = 10^6$, $c = 30$, $\varepsilon = 0.4$ and $\delta = 0.5$. Let D and D' be:

$$D = \begin{pmatrix} 1 & 1 & ... & 1 \\ 2 & 2 & ... & 2 \\ 3 & 3 & ... & 3 \\ 4 & 4 & ... & 4 \\ ... & ... & ... & ... \\ n-1 & n-1 & ... & n-1 \\ n & n & ... & n \end{pmatrix} \qquad D' = \begin{pmatrix} 1.5 & 1.5 & ... & 1.5 \\ 1.5 & 1.5 & ... & 1.5 \\ 3.5 & 3.5 & ... & 3.5 \\ 3.5 & 3.5 & ... & 3.5 \\ ... & ... & ... & ... \\ n-0.5 & n-0.5 & ... & n-0.5 \\ n-0.5 & n-0.5 & ... & n-0.5 \end{pmatrix}$$

We use:

$$sim(a,b) = \begin{cases} 1 & \text{if } |a-b| < 0.6 \\ 0 & \text{otherwise} \end{cases}$$

Note that $\forall r \in D, \forall i \in [c], sim(r'_i, r_i) = 1 \geq 1 - \varepsilon$, so if the algorithm $A = Scoreboard$ proves Theorem 1, it should happen that if $m \geq \frac{\log(10^6/0.4)}{-\log(0.5)} \approx 21.2$, then $P[Sim(R, A(R'|_S, D)) \geq 0.1] \geq 0.6$. However, whatever value m has (somewhere between 1 and 30), for any target r, it always happens that $D_{aux(r)}$ has two different records, one of which is the target (for example, if $r = (1, ..., 1)$, then $r' = (1.5, ..., 1.5)$, and so $D_{aux} = \{(1, ..., 1), (2, ..., 2)\}$). Since the output of A is taken from D_{aux}

with uniform probability, then with probability 0.5 the adversary A outputs the wrong record, in which case $Sim(r, A(aux(r), D) = 0$. Hence, for any $1 \leq m \leq 30$, $P[Sim(R, A(R'|_S, D) = 0] \geq 0.5$, which contradicts that $\forall m \geq 22$, $P[Sim(R, A(R'|_S, D)) \geq 0.1] \geq 0.6$, showing that *Scoreboard* cannot always prove Theorem 1 when $D \neq D'$.

What makes it possible to have this counterexample is that $sim(a,b) = 1$ and $sim(b,c) = 1$ *does not* imply $sim(a,c) = 1$. This does not happen if $1 - sim$ satisfies the triangle inequality, which suggests that this is a condition that would allow us to fix the problem that we are exposing.

Case 2. In this case we also have that D is fully supported, and now we work with a *sim* function for which $1 - sim$ is a metric, but still *Scoreboard* algorithm is not good enough for the proof of Theorem 1. Let $n = 10^5$, $c = 200$, $\varepsilon = 0.25$ and $\delta = 0.125$. Let D and D' be:

$$D = \begin{pmatrix} 0.5 & ... & 0.5 \\ 1 & ... & 1 \\ 1.5 & ... & 1.5 \\ 2 & ... & 2 \\ ... & ... & ... \\ n/2 - 0.5 & ... & n/2 - 0.5 \\ n/2 & ... & n/2 \end{pmatrix} \qquad D' = \begin{pmatrix} 0.75 & ... & 0.75 \\ 0.75 & ... & 0.75 \\ 1.75 & ... & 1.75 \\ 1.75 & ... & 1.75 \\ ... & ... & ... \\ n/2 - 0.25 & ... & n/2 - 0.25 \\ n/2 - 0.25 & ... & n/2 - 0.25 \end{pmatrix}$$

We use $sim(a,b) = \max(1 - |a-b|, 0)$. By Remark 1, $1 - sim$ is a bounded metric. Now, $\forall r \in D, \forall i \in [c]$, $sim(r'_i, r_i) = 0.75 \geq 1 - \varepsilon$, so if *Scoreboard* works to prove Theorem 1, we would conclude that if $m \geq \frac{\log(10^5/0.25)}{-\log(0.875)} \approx 96.6$, then $P[Sim(R, A(R'|_S, D)) \geq 0.625] \geq 0.75$.

As in case 1, for every target record r, when the adversary observes $r'|_s$, it constructs D_{aux} having two candidates. For example, when $r = (2, ..., 2)$, then $r' = (1.75, ..., 1.75)$, and then $D_{aux} = \{(1.5, ..., 1.5), (2, ..., 2)\}$. The similarity between the two candidates in D_{aux} (one of which is the target) is always 0.5, and since the probability of not outputting the target is 0.5, then for any $1 \leq m \leq 200$, $P[Sim(R, A(R'|_S, D)) \leq 0.5] \geq 0.5$, which is equivalent to $P[Sim(R, A(R'|_S, D)) > 0.5] \leq 0.5$, and contradicts that $\forall m \geq 97$, $P[Sim(R, A(R'|_S, D)) \geq 0.625] \geq 0.75$. Again, this shows that *Scoreboard* does not always work to prove Theorem 1 when $D \neq D'$. In this case, however, $1 - sim$ is a metric. What makes this counterexample possible is how the parameters of de-anonymization are given in Theorem 1. We are going to fix this by weakening a little the aimed level of de-anonymization.

Case 3. Let $n = 10^5$, $c = 100$, $\varepsilon = 0.2$ and $\delta = 0.6$. About *sim*, we only assume that $sim(a,b) = 0$ if $|a-b| \geq 1$; and $sim(a,a) = 1$ if $a \neq \perp$. An analogous case can be given with $sim(a,b) = 0$ if $|a-b| \geq \rho$, for any $\rho > 0$, but for simplicity we work with $\rho = 1$. Notice how general this *sim* is. Let D and D' be:

$$D = D' = \begin{pmatrix} 1 & ... & 1 & \perp & ... & \perp \\ 1 & ... & 1 & 1 & ... & 1 \\ 2 & ... & 2 & \perp & ... & \perp \\ 2 & ... & 2 & 2 & ... & 2 \\ ... & ... & ... & ... & ... & ... \\ n/2 & ... & n/2 & \perp & ... & \perp \\ n/2 & ... & n/2 & n/2 & ... & n/2 \end{pmatrix}$$

Assume that the number of columns on the left side of D that contain no $\perp$ are $c' = 19$. Thus, half of the rows have support of size 19 (*small*), and the rest have support of size 100 (*large*). Moreover, each row with small support is a subrecord (see Definition 3) of the record immediately below it, which has large support and is referred to as *its extension*. The similarity between a record of short support and its extension, using equation (2.2) and $sim(a,b)=0$ if $|a-b|\geq 1$, equals $\frac{c'}{c}=0.19$.

Since $D=D'$, then D and D' are $(1-\varepsilon)$-similar. If $A=Scoreboard$ proves Theorem 1, we should have that if $m \geq \frac{\log(10^5/0.2)}{-\log(0.4)} \approx 14.3$, then $P[Sim(R,A(R'|_S,D)) \geq 0.2] \geq 0.8$.

Suppose now that r has short support. Since $D'=D$, then the values that the adversary observes in $r'|_s$ come from the first 19 coordinates of r (since the rest are null). But since r has an extension, the adversary cannot distinguish if the values observed in $r'|_s$ come from r or its extension. Hence, D_{aux} consists of two records, r and its extension, from which the algorithm *Scoreboard* chooses one uniformly at random as its output.

Now, half of the records in D have short support, and when the target has short support, *Scoreboard* wrongly outputs its extension with probability 0.5. The similarity between a record of short support and its extension is 0.19, so we get that for any $1 \leq m \leq 19$, $P[Sim(R,A(R'|_S,D)) \leq 0.19] \geq 0.25$, which is equivalent to $P[Sim(R,A(R'|_S,D)) > 0.19] \leq 0.75$, and contradicts the fact that for $m \geq 15$, $P[Sim(R,A(R'|_S,D)) \geq 0.2] \geq 0.8$. This shows that *Scoreboard* cannot always prove Theorem 1 when D is allowed to have null values.

The problem with *Scoreboard* in this case is that after seeing that all the values in $r'|_s$ come from the first 19 attributes, it still considers equally likely that the target has short support and that it is fully supported. However, given that the values observed in the auxiliary information correspond to some of the first 19 attributes, it is lot more likely that the target has short support than full support. We fix this simply by requiring that the output of the adversary algorithm has minimum support size among the candidates in D_{aux}.

4 Results

We start with two technical lemmas that are used later.

Lemma 1. *Let Z be a finite set and $m \in \mathbb{N}$. Let S be a uniform r.v. taking one of the $\binom{|Z|}{m}$ subset of Z with m elements. Suppose $\Gamma \subseteq Z$ and $|\Gamma| < \gamma \cdot |Z|$, for some $\gamma \in (0,1)$. Then $P[S \subseteq \Gamma] \leq \gamma^m$.*

Proof. If $m > |\Gamma|$, then $P[S \subseteq \Gamma] = 0$. Assume $m \leq |\Gamma|$. Since $|\Gamma| < \gamma \cdot |Z|$, then $m \leq |\Gamma| \leq \lfloor \gamma \cdot |Z| \rfloor$, so:

$$P[S \subseteq \Gamma] = \binom{|\Gamma|}{m} \cdot \binom{|Z|}{m}^{-1} \leq \binom{\lfloor \gamma \cdot |Z| \rfloor}{m} \cdot \binom{|Z|}{m}^{-1} = \prod_{j=0}^{m-1} \frac{\lfloor \gamma |Z| \rfloor - j}{|Z| - j} \leq \gamma^m$$

The last inequality holds because $\forall j \in \{0,\ldots,m-1\}$, $\frac{\lfloor \gamma|Z|\rfloor - j}{|Z|-j} \leq \frac{\gamma|Z|-j}{|Z|-j} \leq \gamma$. □

Lemma 2. *Let U and V be two finite sets such that $|V| \leq |U|$, and let $\psi, \sigma > 0$ such that $\psi < \sigma \cdot |U \cup V|$ and $\psi \leq |U \cap V|$. Then, $\psi < \frac{2\sigma}{1+\sigma} \cdot |U|$.*

Proof. From $|U \cup V| = |U| + |V| - |U \cap V|$ follows $\psi < \sigma \cdot |U \cup V| = \sigma \cdot (|U| + |V| - |U \cap V|)$. Since $|V| \leq |U|$ and $\psi \leq |U \cap V|$, then $\psi < \sigma \cdot (2|U| - \psi)$, which implies $\psi \cdot (1+\sigma) < 2\sigma \cdot |U|$. □

The adversary algorithms that we use in this paper have all the following general form: on input (x,D)—where x represents the auxiliary information—the algorithm constructs a non-empty subset D_x of the set of records in D, and then outputs an element from D_x. What specifies the algorithm is how D_x is defined, and how the output is chosen from D_x.

The input of interest for our results is $x = r'|_s$, where D and D' have the same size and the same support, $r \in D$, and $s \subseteq supp(r)$ with $|s| = m \in \mathbb{N}$.

4.1 New Adversary Output: a Candidate With Minimum Support

Consider the algorithm $\hat{A}$, given by $D_x = \{y \in D : \forall i \in supp(x), sim(x_i, y_i) \geq \alpha\}$, where $\alpha > 0$ is a parameter of the algorithm, and the output is any element of D_x with minimum support. This algorithm is essentially *Scoreboard* from [19], with the only difference that the output of $\hat{A}$ is not a random element from D_x as in *Scoreboard*, but one with minimum support (any record from D_x with minimum support, in case there is more than one, will make the adversary succeed with the desired probability; we do not need to take probability with respect to $\hat{A}$).

Since the output is a record from D_x, it is required that $D_x \neq \phi$. The input of interest is $x = r'|_s$, so it is reasonable to expect $r \in D_{r'|_s}$, which is the same as $\forall i \in s, sim(r_i, r'_i) \geq \alpha$ (for this use that $supp(r'|_s) = s$ by Remark 3, and that *sim* is symmetric by Definition 4). Now, if D and D' are $(1-\varepsilon)$-similar (we assume this is known by the adversary), then $\forall i \in s, sim(r_i, r'_i) \geq 1-\varepsilon$. So to have $r \in D_{r'|_s}$, it suffices that $1-\varepsilon \geq \alpha$.

Remark 4. *To make the adversary output more accurate, α should be taken as high as possible. Thus, when D and D' are $(1-\varepsilon)$-similar, put $\alpha = 1-\varepsilon$, which ensures $r \in D_{r'|_s}$.*

Remark 5. *Recall that we always have $supp(r'|_s) = s$ (Remark 3), so if the input of an algorithm includes $r'|_s$, then the algorithm is allowed to use s in its computations.*

Algorithm $\hat{A}$. Given D, D', and m, the algorithm $\hat{A}$ can be described concisely as follows:
Input: $(r'|_s, D)$, with $r \in D$ and $s \subseteq supp(r)$ such that $|s| = m$.
Output: $\hat{A}(r'|_s, D)$ is any element of $D_{r'|_s} = \{y \in D : \forall i \in s, sim(r'_i, y_i) \geq \alpha\}$ with minimum support.

Since both r and $\hat{A}(r'|_s, D)$ belong to $D_{r'|_s}$, and $\hat{A}(r'|_s, D)$ has minimal support among the records in $D_{r'|_s}$, then $|supp(\hat{A}(r'|_s, D))| \leq |supp(r)|$.

Proposition 3. *For $\hat{A}$ to (σ, p, m)-de-anonymize D, it suffices that $\forall r \in D$, $P[Sim(r, \hat{A}(r'|_S, D)) < \sigma] \leq 1-p$, with probability taken with respect to S.*

Proof. First, that D is (σ, p, m)-de-anonymized by $\hat{A}$ (with respect to *Sim* and D'), is equivalent to:

$$P[Sim(R, \hat{A}(R'|_S, D)) < \sigma] \leq 1-p$$

Now,

$$\begin{aligned} P[Sim(R, \hat{A}(R'|_S, D)) < \sigma] &= \sum_{r \in D} P[Sim(R, \hat{A}(R'|_S, D)) < \sigma \big| R = r] \cdot P[R = r] = \\ &= \sum_{r \in D} P[Sim(r, \hat{A}(r'|_S, D)) < \sigma \big| R = r] \cdot P[R = r] = \\ &= \sum_{r \in D} P[Sim(r, \hat{A}(r'|_S, D)) < \sigma] \cdot P[R = r] \end{aligned}$$

where the last equality holds because for each fixed $r \in D$, the random variables R and S are independent. Hence, if $\forall r \in D$, $P[Sim(r, A(r'|_S, D)) < \sigma] \leq 1-p$, then $P[Sim(R, A(R'|_S, D)) < \sigma] \leq 1-p$. □

Theorem 4. *Let D and D' be $(1-\frac{\varepsilon}{2})$-similar (with $\varepsilon \in [0,1)$), and let $m \in \mathbb{N}$. Assume that $1 - sim$ satisfies the triangle inequality.*
(1) If $Sim_{1-\varepsilon}$ is given by (2.1) and $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, then $\hat{A}$ (σ, p, m)-de-anonymizes D.
(2) If Sim is given by (2.2) and $m \geq \frac{\log(n/(1-p))}{\log((1-\varepsilon+\sigma)/(2\sigma))}$, (assuming$\sigma < 1-\varepsilon$), then $\hat{A}$ (σ, p, m)-de-anonymizes D.

Proof. (1) Given $\sigma \in (0,1)$ and $r \in D$, let $F_{r,\sigma} = \{y \in D : Sim(r,y) < \sigma \ \& \ |supp(y)| \le |supp(r)|\}$. First we show that:

$$P[Sim(r, \hat{A}(r'|_S, D)) < \sigma] \le n \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \tag{4.1}$$

$Sim(r, \hat{A}(r'|_s, D)) < \sigma$ implies that $\exists y \in F_{r,\sigma}$ (the output of $\hat{A}$) such that $y \in D_{r'|_s}$. Hence, (4.1) follows:

$$\begin{aligned} P[Sim(r, \hat{A}(r'|_S, D)) < \sigma] &\le P[\exists y \in F_{r,\sigma} \text{ such that } y \in D_{r'|_S}] = \\ &= P\Big[\bigcup_{y \in F_{r,\sigma}} \{y \in D_{r'|_S}\}\Big] \le \\ &\le \sum_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \le \\ &\le |F_{r,\sigma}| \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \le \\ &\le n \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \end{aligned}$$

By Proposition 3 and (4.1), it suffices to show that $\forall r \in D$:

$$n \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \le 1 - p \tag{4.2}$$

with probability taken with respect to S. Up to here the proof follows the approach in [19].

Now fix r and $y \in D$. We bound $P[y \in D_{r'|_S}]$ from above. If $y \in D_{r'|_s}$, then $\forall i \in s$, $sim(r'_i, y_i) \ge \alpha = 1 - \frac{\varepsilon}{2}$ (see Remark 4). By hypothesis on the similarity between D and D', we have: $\forall i \in s$, $sim(r_i, r'_i) \ge 1 - \frac{\varepsilon}{2}$. Since $1 - sim$ satisfies the triangle inequality, we get: $\forall i \in s$, $sim(r_i, y_i) \ge 1 - \varepsilon$, which means: $s \subseteq \{i : sim(r_i, y_i) \ge 1 - \varepsilon\}$.

Denoting $\Gamma = \{i : sim(r_i, y_i) \ge 1 - \varepsilon\}$, we just proved $y \in D_{r'|_s} \Rightarrow s \subseteq \Gamma$, which implies:

$$P[y \in D_{r'|_S}] \le P[S \subseteq \Gamma] \tag{4.3}$$

Next, we use Lemma 2 with $U = supp(r)$, $V = supp(y)$ and $\psi = |\Gamma|$, to prove:

$$y \in F_{r,\sigma} \Rightarrow |\Gamma| < \frac{2\sigma}{1+\sigma} \cdot |supp(r)| \tag{4.4}$$

We check the hypotheses of Lemma 2. Note that $\sigma > 0$ by hypothesis, and $\psi > 0$ because $\psi = |\Gamma| \ge |s| = m \ge 1$. Assume $y \in F_{r,\sigma}$. Since $y \in F_{r,\sigma}$, then $|supp(y)| \le |supp(r)|$, which means $|V| \le |U|$. Also, $y \in F_{r,\sigma}$ gives $Sim(r,y) < \sigma$, which is $\psi < \sigma \cdot |U \cup V|$. Finally, $\{i : sim(r_i, y_i) \ge 1 - \varepsilon\} \subseteq supp(r) \cap supp(y)$ implies $\psi \le |U \cap V|$ (use $\varepsilon < 1$ and $\forall a \in \mathbb{V}$, $sim(a, \perp) = 0$). So by Lemma 2, $\psi < \frac{2\sigma}{1+\sigma} \cdot |U|$, which is exactly (4.4).

So, for any $y \in F_{r,\sigma}$, we have: $P[y \in D_{r'|_S}] \le P[S \subseteq \Gamma]$ by (4.3), and $|\Gamma| < \frac{2\sigma}{1+\sigma} \cdot |supp(r)|$ by (4.4). Now we apply Lemma 1 with $Z = supp(r)$ and $\gamma = \frac{2\sigma}{1+\sigma}$, to get $P[y \in D_{r'|_S}] < (2\sigma/(1+\sigma))^m$. So, to obtain (4.2), it suffices that $n \cdot (2\sigma/(1+\sigma))^m \le 1 - p$, which is equivalent to $m \ge \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$.

(2) In (1) we just proved that if $m \ge \frac{\log(n/(1-p))}{\log((1+\sigma')/(2\sigma'))}$, then $P[Sim_{1-\varepsilon}(R, \hat{A}(R'|_S, D)) \ge \sigma'] \ge p$.

Sim given by (2.2) satisfies that $\forall x, y, Sim(x,y) \ge (1-\varepsilon) Sim_{1-\varepsilon}(x,y)$. So, $m \ge \frac{\log(n/(1-p))}{\log((1+\sigma')/(2\sigma'))}$ implies $P[Sim(R, \hat{A}(R'|_S, D)) \ge (1-\varepsilon)\sigma'] \ge p$. The result follows by setting $\sigma' = \frac{\sigma}{1-\varepsilon}$. □

Remark 6. *In Definition 9 of de-anonymization, the probability is taken with respect R, S and A. However, in the proof given above, we only took probability with respect to S. Probability with respect to R is avoided by Remark 3, and probability with respect to A does not apply because A is not probabilistic, and the proof works for any output chosen from $D_{r'|_s}$ with minimum support.*

To illustrate Theorem 4, we evaluate it in the simple case $D = D'$ and D fully supported (no null entries). Since $D = D'$, we put $\varepsilon = 0$, and set:

$$sim(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

Note that $1 - sim$ is the discrete metric. Sim given by (2.1), with sim as (4.5), becomes $Sim(x,y) = \frac{1}{c}|\{i : x_i = y_i\}|$. By (1) in Theorem 4, if $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, then D can be (σ, p, m)-de-anonymized by $\hat{A}$. Using that $P[Sim(R, \hat{A}(R'|_S, D)) \geq \sigma] \geq p$ is equivalent to $P[|\{i : R_i = \hat{A}(R'|_S, D)_i\}| \geq \sigma \cdot c] \geq p$, we can formulate the following:

Corollary 5. *When the adversary $\hat{A}$ knows D (fully supported) and has access to $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$ random attribute values of a random record of D of size c, it can learn at least $\sigma \cdot c$ attributes of that same record, with probability at least p.*

Example 1. For Corollary 5 to be useful, we need $m << \sigma \cdot c$. If the adversary wants to double the number of values known about a record, then $2m \leq \sigma \cdot c$. Let $n = 10^6$ and $p = 0.99$. From $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$ and $2m \leq \sigma c$ we get $\frac{16}{\sigma \log((1+\sigma)/(2\sigma))} \leq c$, which determines c for each $\sigma \in (0,1)$. For example, $\sigma = 0.35$ gives $c \approx 160$. In fact, for $n = 10^6$, $c = 160$, $p = 0.99$ and $\sigma = 0.35$, Corollary 5 says that with 0.99 probability, when the adversary knows the whole database and 29 random attribute values of a random record, it can learn 56 attribute values of it (including the known values).

Remark 7. *(2) in Theorem 4 gives a result close to Theorem 1 in [19]. With $\sigma = 1 - \varepsilon - \delta$ and $p = 1 - \varepsilon$, we get that if $m \geq \frac{\log(n/\varepsilon)}{\log(1+\delta/(2-2\varepsilon-2\delta))}$, then D can be $(1 - \varepsilon - \delta, 1 - \varepsilon, m)$-de-anonymized. This is as Theorem 1 in terms of the Sim function (2.2) and the parameters (σ, p) of de-anonymization. The differences are: (i) D and D' are $(1 - \frac{\varepsilon}{2})$-similar, instead of $(1 - \varepsilon)$; (ii) $1 - sim$ satisfies the triangle inequality; (iii) and the lower bound on m, which in Theorem 1 is $m \geq \frac{\log(n/\varepsilon)}{\log(1/(1-\delta))}$.*

Now we give a result on de-anonymization regardless of the level of error in the auxiliary information. This is reflected on the fact that the result works for every D and $D\,'$ having the same support, regardless of the level of similarity. The Sim function (2.3) does not depend on the actual values of the records but only on their supports.

Algorithm $\hat{B}$. Given D, D', and m, the algorithm $\hat{B}$ works as follows:
Input: $(r'|_s, D)$, with $r \in D$ and $s \subseteq supp(r)$ such that $|s| = m$.
Output: $\hat{B}(r'|_s, D)$ is any element of $D_{r'|_s} = \{y \in D : s \subseteq supp(y)\}$ with minimum support size.

Recall that since $r'|_s$ is part of the input and $supp(r'|_s) = s$, then $\hat{B}$ is allowed to use s to construct $D_{r'|_s}$ (Remark 5).

Theorem 6. *Let D and D' have the same support, and Sim be given by (2.3). If $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, then $\hat{B}$ (σ, p, m)-de-anonymizes D.*

Proof. Use (1) in Theorem 4 with $\varepsilon = 0$ and a specific similarity function sim on values:

$$sim(a,b) = \begin{cases} 1 & \text{if } \perp \notin \{a,b\} \\ 0 & \text{otherwise} \end{cases} \tag{4.6}$$

which satisfies Definition 4, and with which D and D' are 1-similar. Since $1 - sim$ is zero over $\mathbb{V} \setminus \{\perp\}$, it trivially satisfies the triangle inequality there. By (1) in Theorem 4, if $m \geq \frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, then $\hat{A}$ (σ, p, m)-de-anonymizes D, with respect to Sim_1 given by (2.1). To conclude, it suffices to show that with the sim given by (4.6), Sim_1 coincides with (2.3), and $\hat{A} = \hat{B}$.

For $Sim_1(x,y) = \frac{|supp(x) \cap supp(y)|}{|supp(x) \cup supp(y)|}$ we need to check $|\{i : sim(x_i,y_i) \geq 1\}| = |supp(x) \cap supp(y)|$, which follows from $sim(x_i,y_i) \geq 1 \Leftrightarrow \perp \notin \{x_i,y_i\} \Leftrightarrow i \in supp(x) \cap supp(y)$.

For $\hat{A} = \hat{B}$, it suffices that $\{y \in D : \forall i \in s, sim(r'_i,y_i) \geq 1\} = \{y \in D : s \subseteq supp(y)\}$, which follows from: $(\forall i \in s, sim(r'_i,y_i) \geq 1) \Leftrightarrow s \subseteq supp(y)$. For $(\Rightarrow)$, take $i \in s$. By hypothesis $sim(r'_i,y_i) \geq 1$, so $y_i \neq \perp$, thus $i \in supp(y)$. For $(\Leftarrow)$, take $i \in s$. Since $s \subseteq supp(r) = supp(r')$, then $r'_i \neq \perp$. Also, by hypothesis $i \in supp(y)$, so $y_i \neq \perp$, hence $sim(r'_i,y_i) \geq 1$. □

4.2 Improvements by Sparsity

Large databases with individuals' transactions typically have high level of sparsity, which means that it is unlikely for a random record of the database to be highly similar to another record of the same database. It is shown in [19] that sparsity improves the de-anonymization of Theorem 1. We show in general that every result on de-anonymization turns into a result on *perfect* de-anonymization (see Definition 11 below), under a certain hypothesis on sparsity.

Definition 10. *Let K be a r.v. selecting an element from $[n]$ uniformly at random. Given $j \in [n]$, let D_j denote the j-th row of D. A database D is (σ, q)*-sparse with respect to *Sim if:*

$$P[\exists j \in [n] : j \neq K \ \&\ Sim(D_K, D_j) \geq \sigma] \leq q$$

with probability taken with respect to K.

Definition 11. *An m-supported database D can be* perfectly (p,m)-de-anonymized with respect to *D' if there exists an adversary A such that:*

$$P[R = A(R'|_S, D)] \geq p \tag{4.7}$$

The probability is taken with respect to R, S (as in Definition 9) and the random choices of A (if any).

We express (4.7) saying that A perfectly (p,m)-de-anonymizes D, or that D is perfectly (p,m)-de-anonymized by A.

Note that if A perfectly (p,m)-de-anonymizes D with respect to D', then A $(1, p, m)$-de-anonymizes D with respect to Sim and D' (for any Sim), since $x = y \Rightarrow Sim(x,y) = 1$.

The next theorem shows that in combination with the appropriate hypothesis on sparsity, every result on de-anonymization by an adversary outputting records of D, becomes a result on perfect de-anonymization, with a cost on the probability of success of the adversary.

Theorem 7. *Let $p > q$. Assume that under certain hypotheses $\mathbb{H}$ (e.g., about m, D and D'), the database D can be (σ, p, m)-de-anonymized by an algorithm A (outputting records of D) with respect to Sim and D'. Assume also that D is (σ, q)-sparse with respect to Sim. Then, $\mathbb{H}$ implies that D can be perfectly $(p - q, m)$-de-anonymized by A with respect to D'.*

Proof. We want $P[R = A(R'|_S, D)] \geq p - q$. The outputs of A and R are records of D, so it suffices that $P[row(R) = row(A(R'|_S, D))] \geq p - q$, where $row(R) \in [n]$ denotes the row given by R (similarly for $row(A(R'|_S, D))$. It is equivalent to show $P[row(R) \neq row(A(R'|_S, D))] \leq 1 - p + q$. Now:

$$\begin{aligned} P[row(R) \neq row(A(R'|_S, D))] &= P[(row(R) \neq row(A(R'|_S, D))) \cap (Sim(R, A(R'|_S, D)) \geq \sigma)] \\ &+ P[(row(R) \neq row(A(R'|_S, D))) \cap (Sim(R, A(R'|_S, D)) < \sigma)] \end{aligned}$$

We bound each term on the right-hand side. First:

$$P[(row(R) \neq row(A(R'|_S,D))) \cap (Sim(R,A(R'|_S,D)) \geq \sigma)]$$

$$\leq P[\exists j \in [n] : j \neq row(R) \;\&\; Sim(R,D_j) \geq \sigma]$$

$$= P[\exists j \in [n] : j \neq K \;\&\; Sim(D_K,D_j) \geq \sigma] \leq q$$

where the first inequality follows setting $j = row(A(R'|_S,D))$ and the last inequality holds by the hypothesis on sparsity. Second:

$$P[(row(R) \neq row(A(R'|_S,D))) \cap (Sim(R,A(R'|_S,D)) < \sigma)] \leq P[Sim(R,A(R'|_S,D)) < \sigma] \leq 1-p$$

since $\mathbb{H}$ implies that A (σ,p,m)-de-anonymizes D with respect to Sim and D'. □

The next result combines the de-anonymization theorems 4 and 6, with Theorem 7.

Theorem 8. *Assume D is (σ,δ)-sparse with respect to Sim used in each of the following cases.*

(a) Let D and D' be $(1-\frac{\varepsilon}{2})$-similar, for an $\varepsilon \in [0,1)$, and assume that $1-sim$ satisfies the triangle inequality. If Sim_μ is given by (2.1), with $\mu = 1-\varepsilon$, and $m \geq \frac{\log(n/\delta)}{\log((1+\sigma)/(2\sigma))}$, then $\hat{A}$ perfectly $(1-2\delta,m)$-de-anonymizes D.

(b) Let D and D' be $(1-\frac{\varepsilon}{2})$-similar, for an $\varepsilon \in [0,1)$, and assume that $1-sim$ satisfies the triangle inequality. If Sim is given by (2.2), and $m \geq \frac{\log(n/\delta)}{\log((1-\varepsilon+\sigma)/(2\sigma))}$ (assume $\sigma < 1-\varepsilon$), then $\hat{A}$ perfectly $(1-2\delta,m)$-de-anonymizes D.

(c) Let D and D' have the same support, and Sim be given by (2.3). If $m \geq \frac{\log(n/\delta)}{\log((1+\sigma)/(2\sigma))}$, then $\hat{B}$ perfectly $(1-2\delta,m)$-de-anonymizes D.

Proof. Call $\mathbb{H}$ the hypotheses on D, D', and on m. For (a) and (b), Theorem 4 gives that under $\mathbb{H}$, D is $(\sigma,1-\delta,m)$-de-anonymized by $\hat{A}$. For (c), Theorem 6 gives that under $\mathbb{H}$, D is $(\sigma,1-\delta,m)$-de-anonymized by $\hat{B}$. Now, D is (σ,δ)-sparse with respect to Sim, so from Theorem 7 it follows that D is perfectly $(1-2\delta)$-de-anonymized (using $\hat{A}$ for (a) and (b), and using $\hat{B}$ for (c)). □

4.3 Applying Theorem 8 on the Netflix Database

In this section we evaluate Theorem 8 in the parameter values of the Netflix database, here denoted by D_N. This database has approximately $n = 480,000$ records, each of which contains ratings—and the dates of these ratings—for $c = 17,770$ movies, entered by Netflix users. The movies are the attributes, and each element in $\mathbb{V}$ is a pair rating-date.

To evaluate Theorem 8 on D_N, we use one of the sparsity functions of D_N given in [19]. The definition of sparsity function is the following.

Definition 12. *The* sparsity function $f:[0,1] \to [0,1]$ *of D with respect to Sim is given by $f(\sigma) = P[\exists j \in [n] : j \neq K \;\&\; Sim(D_K,D_j) \geq \sigma]$, with probability taken with respect to K (as in Definition 10).*

Theoretical vs. Heuristic De-Anonymization. We compare the level of de-anonymization guaranteed by Theorem 8, with the heuristic de-anonymization given in [19] for the Netflix database, where it is concluded that when the adversary knows 8 exact ratings of movies (but ignores the dates), of which 6 are outside the top 500 most rated movies, then the adversary can uniquely identify the complete history of ratings and dates, with probability 0.84.

We want to obtain an analogous result using (b) in Theorem 8. Since the adversary knows the ratings exactly, but ignores the dates, we assume that D_N only contains the ratings, and that $D_N = D'_N$ (no error in the auxiliary information), so we set $\varepsilon = 0$. We use the same *Sim* as in [19], which is given by (2.2), with *sim* defined by:

$$sim(a,b) = \begin{cases} 1 & \text{if } a = b \text{ and } \bot \notin \{a,b\} \\ 0 & \text{otherwise} \end{cases}$$

which is referred to as "ratings $+/-0$" in [19], and makes of $1 - sim$ a discrete metric over $\mathbb{V} \setminus \{\bot\}$.

We have to apply (b) in Theorem 8 with the probability level at 0.84, so $1 - 2\delta = 0.84$, which gives $\delta = 0.08$. The sparsity function of D_N gives that D_N is $(0.25, 0.08)$-sparse [19], so $\sigma = 0.25$, and by (b) in Theorem 8, if $m \geq \frac{\log(480{,}000/0.08)}{\log(1.25/0.5)} \approx 17.03$, then $\hat{A}$ perfectly 0.84-de-anonymizes the Netflix database D_N.

The heuristic de-anonymization requires a smaller number of movies (8 instead of 18), and it recovers a record with the exact ratings and dates, while the theorem assures that the adversary recovers only the exact ratings (not the dates). This gap is to be expected, since in the heuristic de-anonymization the adversary algorithm is specifically calibrated for the Netflix database, whereas the theorems are proved using a fixed algorithm that works for all the databases satisfying general hypotheses. Moreover, in the heuristic de-anonymization the auxiliary information is stronger, in the sense that 6 of the 8 movies are outside the top 500 most rated, which is not assumed in Theorem 8. This motivates the next results.

4.4 Decreasing *m* When *D* Has a Long Tail

One of the parameters to evaluate the performance of the adversary is *m*, the number of non-null attribute values used by the adversary to de-anonymize a database (i.e., the size of the auxiliary information). In each of the previous results, a lower bound for *m* is given. These bounds are not necessary, but sufficient to prove the statements.

In this section we obtain better (from the adversary's point of view) lower bounds, assuming that the adversary knows at least one value corresponding to a non-popular attribute *in the long tail* of *D*. For this we need to introduce Definition 13.

Without loss of generality, we assume that the attributes of *D* are given in decreasing order with respect to their support sizes, i.e., $i \mapsto |supp(i)|$ is a decreasing function on $[c]$, where $i \in [c]$ is an attribute and $supp(i)$ is its support. We always assume that $\forall i \in [c]$, $supp(i) \neq \phi$.

Definition 13. *Let $\tau, \kappa \in [0,1]$. A database D of size $n \times c$ has a (τ, κ)-*tail *if:*

$$|supp(\lceil \tau \cdot c \rceil)| = \lfloor \kappa \cdot n \rfloor$$

Remark 8. *Let D of size $n \times c$ have a (τ, κ)-*tail. *Since $i \mapsto |supp(i)|$ is decreasing, then:*

$$\forall i \in [c] (i \geq \tau \cdot c \ \Rightarrow \ |supp(i)| \leq \kappa \cdot n)$$

Theorem 9 (below) is meant to formalize what is observed in the heuristic de-anonymization of the Netflix database: when the auxiliary information contains values corresponding to rare attributes, less amount of auxiliary information suffices to achieve perfect de-anonymization (at a certain probability level).

In [4], the author describes the *long tail* phenomenon, which occurs in many different databases where individuals (records) relate to attributes (columns). Examples include the Netflix, Amazon, and Rhapsody (an online music service) databases, where attributes correspond to movies, books, and music tracks, respectively. Something observed in these cases is that the majority of the items are related to a small fraction of the individuals, but because there are so many non-popular items, their sales represent a significant share of the total.

If the support sizes $|supp(.)|$ are plotted decreasingly, the *long tail* is the part of the graph corresponding to non-popular items, and their sales (in items) equals the area under the curve.

We use this simple description of the graph of $|supp(.)|$ to improve the de-anonymization results given above.

Table 1 shows the number of track downloads via Rhapsody, from a total of 1.5 million tracks, when there were 1.4 million subscribers (December 2005, http://investor.realnetworks.com/results.cfm). The most popular tracks were downloaded approximately 180,000 times, but as popularity decreases, the number of downloads decreases very rapidly.

If D is the database corresponding to Table 1, with size $1,400,000 \times 1,500,000$, then the third line of Table 1 says that D has a $(\frac{1}{300}, \frac{1}{140})$-tail.

Another characteristic of a database is the proportion of records with support containing non-popular attributes, which is formalized in the following definition.

Definition 14. *Let D be of size $n \times c$ and let $\tau \in [0,1]$. Then:*
(1) A record $x \in D$ is supported on τ-ranks *if $\exists i \in supp(x)$ such that $i \geq \tau \cdot c$.*
(2) $D_{\geq\tau}$ is the database formed with records of D that supported on τ-ranks.

Table 2 gives this information for D_N, the Netflix database. Since D_N has $c = 17,770$ attributes, the third line in Table 2 implies that 97% of the records in D_N are supported on 0.06-ranks (since $1000/17770 \approx 0.06$).

Table 1: Downloads From Rhapsody [4]

Popularity rank	Number of downloads ≈
1	180,000
1,000	30,000
5,000	10,000
25,000	700
85,000	150

Table 2: Netflix's % of Subscribers Supported at Different Ranks [19]

% of subscribers who rated a movie	not among the x most rated
100%	$x = 100$
99%	$x = 500$
97%	$x = 1,000$

The improvement we give consist in decreasing the lower bound on m, so that with less auxiliary information, we can still assure that the de-anonymization is successful. This will be possible when at least one attribute in s is not popular, so we need to change the way s is chosen. So far s was taken uniformly at random among those subsets of $supp(r)$ with $|s| = m$. For the results in this section, given $r \in D$ supported on τ-ranks (i.e., $r \in D_{\geq\tau}$), we take $i \in \{j \in supp(r) : j \geq \tau \cdot c\}$ uniformly at random, and $\tilde{s} \subseteq supp(r)$ with $|\tilde{s}| = m - 1$ uniformly at random. Then $s = \tilde{s} \cup \{i\}$ is used to define the input of the adversary. If $\tilde{S}$ and I are the random variables giving $\tilde{s}$ and i, we now write $S = \tilde{S} \cup \{I\}$.

Remark 9. *Definitions 9 and 11 of de-anonymization change accordingly: the probability is now taken with respect to the r.v. $S = \tilde{S} \cup \{I\}$, and the r.v. R, selecting a random record from $D_{\geq\tau}$.*

Analogously to Proposition 3, it holds that to get de-anonymization, it suffices that:

$$\forall r \in D_{\geq\tau}, P\left[Sim(r, A(r'|_S, D)) < \sigma\right] \leq 1 - p$$

with probability taken with respect to $\tilde{S}$ and I.

Now, because for each fixed $r \in D_{\geq\tau}$, the random variables $\tilde{S}$ and I are independent, to have $P\left[Sim(r, A(r'|_S, D)) < \sigma\right] \leq 1 - p$ with probability over both $\tilde{S}$ and I, it suffices that

$$P\left[Sim(r, A(r'|_S, D)) < \sigma\right] \leq 1 - p$$

for each possible output of I fixed, with probability taken only with respect to$\tilde{S}$. So we have this:

Remark 10. *For (σ, p, m)-de-anonymization (note Remark 9), it suffices that $\forall r \in D_{\geq\tau}$, $\forall i \in supp(r)$ such that $i \geq \tau \cdot c$, it happens that $P\left[Sim(r, A(r'|_S, D)) < \sigma\right] \leq 1 - p$ with probability over $\tilde{S}$, where $S = \tilde{S} \cup \{i\}$.*

The following theorem is very similar to part (1) of Theorem 4, except that it uses the extra hypothesis of D having a (τ, κ)-tail, which leads to a smaller lower bound on m. Also, the de-anonymization is with respect to the new random variables: $S = \tilde{S} \cup \{I\}$ and R taking a random record from $D_{\geq\tau}$.

Theorem 9. *Let D have a (τ, κ)-tail, D and D' be $(1 - \frac{\varepsilon}{2})$-similar (with $\varepsilon \in [0, 1)$), $m \in \mathbb{N}$, and assume that $1 - sim$ satisfies the triangle inequality. If $Sim_{1-\varepsilon}$ is given by (2.1) and $m > \frac{\log(\kappa \cdot n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, then $\hat{A}$ (σ, p, m)-de-anonymizes D with respect to $Sim_{1-\varepsilon}$ and D'.*

Proof. By Remark 10, if we fix $r \in D_{\geq\tau}$, and $i \in supp(r)$ such that $i \geq \tau \cdot c$, then it suffices to prove: $P\left[Sim(r, \hat{A}(r'|_S, D)) < \sigma\right] \leq 1 - p$, with probability taken with respect to $\tilde{S}$ (where $S = \tilde{S} \cup \{i\}$). This is similar to what is done in the proof of Theorem 4. Instead of (4.1), here we want:

$$P[Sim(r, \hat{A}(r'|_S, D)) < \sigma] \leq \kappa \cdot n \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \tag{4.8}$$

where $F_{r,\sigma} = \{y \in D : Sim(r, y) < \sigma \ \& \ |supp(y)| \leq |supp(r)|\}$. Once we prove (4.8), we then need:

$$\kappa \cdot n \cdot \max_{y \in F_{r,\sigma}} P[y \in D_{r'|_S}] \leq 1 - p \tag{4.9}$$

with probability taken with respect to $\tilde{S}$. As in the proof of Theorem 4, we get that $P[y \in D_{r'|_S}] \leq (2\sigma/(1+\sigma))^{m-1}$ (recall that $S = \tilde{S} \cup \{i\}$), and $\tilde{S}$ picks a subset $\tilde{s} \subseteq supp(r)$ such that $|\tilde{s}| = m - 1$). So to get (4.9), it suffices that $\kappa \cdot n \cdot (2\sigma/(1+\sigma))^{m-1} \leq 1 - p$, which is equivalent to our current hypothesis: $m > \frac{\log(\kappa \cdot n/(1-p))}{\log((1+\sigma)/(2\sigma))}$. So to conclude the proof we show (4.8).

We stress that in (4.8), the probability is taken with respect to $\tilde{S}$ (outputting $\tilde{s}$ such that $|\tilde{s}| = m - 1$), and $S = \tilde{S} \cup \{i\}$, where i remains fixed.

Denote $\hat{D} = \{y \in D : y_i \neq \perp\}$ and $\hat{n} = |supp(i)|$. Note $\hat{D}$ is in 1-1 correspondence with $supp(i)$, and since D has a (τ, κ)-tail and $i \geq \tau \cdot c$, then we have $|\hat{D}| = \hat{n} \leq \kappa \cdot n$.

Note that $\forall \tilde{s}$, $D_{r'|_s} \subseteq \hat{D}$. This is because if $y \in D_{r'|_s}$, then by definition of$\hat{A}$, $y \in D$ and $\forall j \in s = \tilde{s} \cup \{i\}$, $sim(r'_j, y_j) \geq 1 - \varepsilon > 0$. In particular, $j = i$ gives $sim(r'_i, y_i) > 0$, so $y_i \neq \perp$, and hence $y \in \hat{D}$.

Now, for every $\tilde{s}$ with $|\tilde{s}| = m - 1$, $Sim(r, \hat{A}(r'|_s, D)) < \sigma$ implies that $\exists y \in F_{r,\sigma}$ (the output of $\hat{A}$) such that $y \in D_{r'|_s} \subseteq \hat{D}$, and $\hat{D}$ remains fixed for the different subsets $\tilde{s}$. We can express this by saying that for every $\tilde{s}$ with $|\tilde{s}| = m - 1$, $Sim(r, \hat{A}(r'|_s, D)) < \sigma$ implies that $\exists y \in F_{r,\sigma} \cap \hat{D}$ such that $y \in D_{r'|_s}$.

Hence:

$$\begin{aligned}
P[Sim(r,\hat{A}(r'|_S,D)) < \sigma] &\leq P[\exists y \in F_{r,\sigma} \cap \hat{D} \text{ such that } y \in D_{r'|_S}] = \\
&= P\Big[\bigcup_{y\in F_{r,\sigma}\cap\hat{D}} \{y \in D_{r'|_S}\}\Big] \leq \\
&\leq \sum_{y\in F_{r,\sigma}\cap\hat{D}} P[y \in D_{r'|_S}] \leq \\
&\leq |F_{r,\sigma} \cap \hat{D}| \cdot \max_{y\in F_{r,\sigma}\cap\hat{D}} P[y \in D_{r'|_S}] \leq \\
&\leq \hat{n} \cdot \max_{y\in F_{r,\sigma}} P[y \in D_{r'|_S}] \leq \\
&\leq \kappa \cdot n \cdot \max_{y\in F_{r,\sigma}} P[y \in D_{r'|_S}]
\end{aligned}$$

□

Remark 11. *If $\tau_1 \leq \tau_2$ and $\kappa_1 \leq \kappa_2$ then the de-anonymization is more efficient when D has a (τ_1,κ_1)-tail than when it has a (τ_2,κ_2)-tail. In fact, the lower bound on m is smaller with κ_1, so less amount of auxiliary information is required, and $\tau_1 \leq \tau_2$ implies $D_{\geq\tau_1} \supseteq D_{\geq\tau_2}$, so the de-anonymization is successful over a larger set of records.*

By looking at the proofs given so far, we note that (2) in Theorem 4, Theorem 6 and Theorem 8, all follow from (1) in Theorem 4. Likewise, Theorem 9 (which is an updated version of (1) in Theorem 4), yields results analogous to (2) in Theorem 4, Theorem 6, and Theorem 8.

The updated statements are similar to the previous ones, with the following differences: (i) D has a (τ,κ)-tail; (ii) the lower bound on m, which previously looked like $m \geq \log(n/...)/\log(...)$, now becomes $m > \log(\kappa \cdot n/...)/\log(...)$; and (iii) the de-anonymization is with respect to the uniform random variables R, taking a record of $D_{\geq\tau}$, and $S = \tilde{S} \cup \{I\}$ (where $\tilde{S}$ takes a subset of $supp(r)$ with $m-1$ elements, and I takes an element from $\{i \in [c] : i \geq \tau \cdot c\}$).

For instance, the new version of (c) in Theorem 8 is:

Theorem 10. *Assume D has a (τ,κ)-tail. Let Sim be given by (2.3), and assume D is (σ,δ)-sparse with respect to Sim. If $m > \frac{\log(\kappa\cdot n/\delta)}{\log((1+\sigma)/(2\sigma))}$, then $\hat{B}$ perfectly $(1-2\delta,m)$-de-anonymizes D with respect to D' (having the same support as D).*

We want to know how is the new lower bound on m compared to the previous one. Suppose we want the new lower bound $\frac{\log(\kappa\cdot n/\delta)}{\log((1+\sigma)/(2\sigma))}$ to be half of the previous lower bound $\frac{\log(n/\delta)}{\log((1+\sigma)/(2\sigma))}$, so that the auxiliary information of the adversary can be reduced a 50%, still making the adversary able to de-anonymize.

From $\frac{\log(\kappa\cdot n/(1-p))}{\log((1+\sigma)/(2\sigma))} = \frac{1}{2}\cdot\frac{\log(n/(1-p))}{\log((1+\sigma)/(2\sigma))}$, we solve for κ to get $\kappa = \sqrt{\frac{1-p}{n}}$. So given p and n, when $\kappa = \sqrt{\frac{1-p}{n}}$ is satisfied, the new lower bound on m is half of the previous bound. But this new bound on m only works to de-anonymize the records of $D_\tau \subseteq D$, where τ is determined by κ (hence by p and n). Notice that for each κ there is a corresponding (minimum) τ for which the database has a (τ,κ)-tail.

To illustrate, consider a database D with $n = 480,000$, and let the probability of de-anonymization be $p = 0.99$. Then, the new lower bound on m would be half of the previous one if $\kappa \approx 0.0001$.

Now, for $\kappa = 0.0001$, we want τ for which D has a (τ,κ)-tail. The Rhapsody database—to consider a real example—has a $(0.056, 0.0001)$-tail. This follows from the last line in Table 1 and that the Rhapsody database has size $1,400,000 \times 1,500,000$. So the de-anonymization with the new lower bound applies to the records of $D_{\geq\tau}$, with $\tau = 0.056$. Note that by Table 2, 97% of the Netflix records are supported on some attribute (movie) beyond the 1000 most rated. Since the Netflix database has $c = 17,770$, this means that 97% of its records are in $D_{\geq\tau}$, for $\tau = 0.056$.

4.5 Linkage Attacks on Survey Data

In this section we evaluate one of the theorems on a database with relatively small number of records and attributes, which does not have a high level of sparsity. We then perform an heuristic linkage attack on the database, and compare the rates of adversary success given by both the theorem and the experiment.

We work with the microdata file of the Joint Canada/United States Survey of Health, 2004 [22] (JCUSH), produced by Statistics Canada and distributed by the Data Liberation Initiative[1]. The complete survey, supporting documentation and JCUSH file can be downloaded from the Ontario Data Documentation, Extraction Service and Infrastructure (ODESI) at *odesi.ca*.

The JCUSH file contains 366 variables (attributes) and 8688 records, each corresponding to an individual who participated in the survey. The variables are grouped into 24 different categories, mostly related to health, but also including socio-demographic characteristics. Among the different categories, some may be considered more sensitive than others. The categories *chronic conditions* or *depression*, could be considered more sensitive than the category *physical activity*, which consists of attributes such as whether the participant practices tennis or basketball, among other activities.

Of the 366 variables, we use 353; we ignore the following 13 variables: SAMPLEID, SPJ1_CP (sampled collection period), DPJ1DPP (depression scale), IWJ1DXUC (exchange rate U.S. to Can), IWJ1DXCU (Can to U.S.), WTS_STR (variance domain), WTS_SAM (sampling weight), and 6 other variables (survey administration).

The used 353 variables correspond to questions made to the participants, and the answers are encoded with non-negative integers. The majority of the variables contemplate answers such as *not applicable, don't know, refusal, not stated*, all of which we treat as null ($\perp$).

Of the 353 variables, 213 take values 1 (*yes*), 2 (*no*), and $\perp$. This is, 213 variables have only 2 non-null values in their ranges. Among the remaining variables, 97 have less than 14 non-null values in their ranges, and the other 43 have ranges with different amounts (up to 103) of non-null values.

Of the 8688 records, we ignore 69 records that have size support smaller than 80, which yields the database that we denote by D_J, having $n = 8619$ and $c = 353$. Of the $8619 \times 353 = 3,042,507$ entries of D_J, $1,515,364$ are non-null ($\approx 50\%$).

We first evaluate Theorem 8 (a) with the parameters of D_J. For simplicity we assume no error in the auxiliary information, which is $D' = D$, and so $\varepsilon = 0$. Set:

$$sim(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

so that $1 - sim$ is the discrete metric. *Sim* given by (2.1) becomes

$$Sim(x,y) = \frac{|\{i \in supp(x) \cap supp(y) : x_i = y_i\}|}{|supp(x) \cup supp(y)|}$$

Theorem 8 (a) then says that if a database D is (σ,δ)-sparse with respect to *Sim* and $m \geq \frac{\log(n/\delta)}{\log((1+\sigma)/(2\sigma))}$, then $\hat{A}$ perfectly $(1-2\delta, m)$-de-anonymizes D (in the sense of Definition 11).

Recall that the sparsity function of D_J is $f : [0,1] \to [0,1]$,

$$f(\sigma) = P[\exists j \in [n] : j \neq K \;\&\; Sim(D_K, D_j) \geq \sigma]$$

[1] The results or views expressed are those of the author and are not those of Statistics Canada.

so by definition of f, D_J is $(\sigma, f(\sigma))$-sparse. Hence, if $m \geq \frac{\log(n/f(\sigma))}{\log((1+\sigma)/(2\sigma))}$, then with probability $1-2f(\sigma)$, the adversary $\hat{A}$ successfully links the auxiliary information (of size m) of a random record in D with the right record.

We need to find a value for σ such that $f(\sigma)$ is small enough so that the probability of success is high, but at the same time the lower bound on m is as small as possible.

As σ approaches 1, $f(\sigma)$ approaches 0, however the lower bound on m given by $\frac{\log(n/f(\sigma))}{\log((1+\sigma)/(2\sigma))}$ increases (mostly dominated by the denominator).

Table 3 gives these parameters for different values of σ.

Table 3: Parameters in Theorem 8 (a) for D_J With $n = 8619$

σ	$f(\sigma)$	$m \geq$	P[adversary success] $\geq$
0.45	1.000	19	-1 (useless bound)
0.50	0.988	23	-0.98 (useless bound)
0.55	0.955	27	-0.91 (useless bound)
0.60	0.673	33	-0.35 (useless bound)
0.65	0.455	42	0.09
0.70	0.145	57	0.71
0.75	0.063	77	0.87
0.80	0.000	N/A (divide by 0)	1

To calculate each $f(\sigma)$ in Table 3 we take a random sample of 400 records of D_J, and for each record r in the sample we look at all the other records in D_J to determine whether there exists a record $x \neq r$ such that $Sim(r,x) \geq \sigma$. Then $f(\sigma)$ is calculated as the fraction of the sample of records for which such x exists. This is of course an approximation of $f(\sigma)$—to get the exact value this computation should be done over all the records of D_J instead of using a sample.

By Table 3 the value of σ to be used in Theorem 8 should satisfy $\sigma > 0.70$ if we want to ensure a probability of success greater than 0.71 (rather low), and also $\sigma < 0.75$ so that we can take $m < 77$, which seems large.

We calculate $f(0.73) = 0.079$ (again using a random sample of 400 records), which then by Theorem 8 (a), implies that if the adversary knows $m = 69$ randomly attribute values from $supp(r)$, for a randomly chosen record $r \in D$, then the adversary can find the 353 values of r with probability 0.842 (at least).

To test this empirically we perform a linkage attack on D_J using the adversary algorithm $\hat{A}$, which with the current sim works as follows:

Input: $(r|_s, D)$, with $r \in D$ and $s \subseteq supp(r)$ such that $|s| = 69$.
Output: $\hat{A}(r|_s, D)$ is any element of $D_{r|_s} = \{y \in D : r|_s = y|_s\}$ with minimum support.

In the implementation of this algorithm, for which we use R [20], the output is chosen uniformly at random from the elements of $D_{r|_s}$ with minimal support.

We take a random sample of 150 records of D, and for each r, a subset $s \subset supp(r)$ with $|s| = 69$ is then taken at random (this is possible since $\forall r \in D_J$, $|supp(r)| \geq 80$). The proportion of the 150 records (in the sample) for which $r = \hat{A}(r|_s, D)$ is an approximation of the probability of the adversary success, which by Theorem 8 should be ≥ 0.842. In our experiment all the 150 outputs $A(r|_s, D)$ were equal to r. The difference between the theorem and the concrete results of the experiment with D_J in terms of the rate of success (0.842 vs. 1) could be explained by the level of generality of Theorem 8 and the fact that the theorem gives a lower bound on the probability of success.

A Second Experiment. What we obtained is an adversary that is successful when it uses 69 non-null attribute values, which is large ($\approx$ 20% of c). Moreover, these 69 non-null values are taken among any of the attributes, including sensitive attributes such as those in the category *chronic conditions*. One could argue that using sensitive data to reveal more sensitive data is a flawed approach for a successful attack, since it starts by assuming the availability of the sensitive attributes.

This leads us to try a linkage attack on D_J based only on non-sensitive attributes. Among the 24 categories of variables in JCUSH, perhaps the less sensitive is *physical activity*, followed by *dental visits* and some socio-demographic variables. If we put these variables together, the restriction of D_J to these attributes is not sparse enough to obtain a meaningful application of our theorems: to make $1-2f(\sigma) > 0$ and have a useful lower bound on the probability of success, σ has to be taken so close to 1, that makes the lower bound on m even larger than the total number of non-sensitive attributes.

This means that our theorems are not strong enough to show a successful linkage attack on D_J based on the (seemingly) non-sensitive attributes. However, we can still experiment with $\hat{A}$ and D_J and evaluate the linkage attack heuristically.

We perform this linkage attack based on the following 35 variables: 25 belong to the *physical activity* category, and represent the answers to the question "have you done any of the following in the past 3 months?". The 25 variables correspond to 25 activities such as walking, swimming, and so on. These 25 variables take values in $\{1,2,\perp\}$. The other 10 variables used for this linkage attack are *age*, *sex*, *type of participant* (US or Canada), and the 7 variables of the *socio-demographic characteristics* category (4 them taking 2 non-null values, and the others no more than 6).

We run the same experiment using $\hat{A}$ as we did above (with the sample of 150 records), except that now we impose that for each record r, the random subset s that defines the auxiliary information $r|_s$ is such that $|s| = 20$ and $s \subseteq supp(r) \cap V$, where V consist of the described 35 non-sensitive attributes. To make sure every record in the database has at least 20 non-null values among those corresponding to the attributes in V, we kept in D_J only the records with support size ≥ 30 (so now $n = 8419$). Over a random sample of 150 records we get that on 110 of them the adversary $\hat{A}$ outputs r when the input is $r|_s$. In other words, an adversary that knows 20 random non-null values from the 35 non-sensitive variables described, can find all the 353 values (including the sensitive one) of the record from which those 20 are coming, with probability ≈ 0.73.

A comment on the term *de-anonymization*. Our work is motivated by [19], where the authors show that anonymous records of the Netflix database can be linked to profiles in the IMDb, which could be associated to the names of the users (a *real* de-anonymization). The concept is formalized in a definition and used in the cited Theorem 1.

We use the same definition for our formal results. However, what we do with the JCUSH database is to show a successful linkage attack through an experiment, in which we simulate the auxiliary information using random values of the non-sensitive variables from the same database. We do not link identifiers (such as name) to records of JCUSH, nor we explore the degree in which an adversary can find on the web, or any other data source, such non-sensitive attribute values associated to the identity of a participant of a survey.

Perhaps a valid implication of our experiments is that once sensitive and non-sensitive attributes are published together in microdata, the non-sensitive attributes become sensitive.

4.6 Another Approach

We explore another approach to perfect de-anonymization, without using the hypothesis on sparsity.

Let A^* denote an algorithm with the following general form: on input (x,D), where $x = r'|_s$ is the auxiliary information, the algorithm constructs a non-empty subset of candidates $D_x \subseteq D$, and then outputs an element from D_x chosen uniformly at random. Assume that the target record r is always among the candidates, i.e., $\forall r \in D$ and $\forall s \subseteq supp(r)$, $r \in D_{r'|_s}$. The *Scoreboard* algorithm in [19] has this form, for instance.

Theorem 11. *For any such A^*, it holds that:*

$$P[R = A^*(R'|_S, D)] = E[|D_{R'|_S}|^{-1}] \geq \frac{1}{E[|D_{R'|_S}|]}$$

with probability over R, S and A^*, where R and S are random variables giving a record from D and a subset of $supp(R)$, respectively.

Proof. For each $r \in D$,

$$\begin{aligned} P[r = A(r'|_S, D)] &= \sum_{k=1}^{n} P[r = A(r'|_S, D)\big||D_{r'|_S}| = k] \cdot P[|D_{r'|_S}| = k] =^{(*)} \\ &= \sum_{k=1}^{n} \frac{1}{k} \cdot P[|D_{r'|_S}| = k] = E\left[|D_{r'|_S}|^{-1}\right] \end{aligned}$$

(*) holds because $r \in D_{r'|_s}$, and the choice from $D_{r'|_s}$ made by A^* is uniformly at random and independent of S. Also, since for each r, the r.v.'s S_r and R are independent, we have (**) next:

$$\begin{aligned} P[R = A(R'|_S, D)] &= \sum_r P[R = A(R'|_S, D)\big|R = r] \cdot P[R = r] =^{(**)} \\ &= \sum_r P[r = A(r'|_S, D)] \cdot P[R = r] = \\ &= \sum_r E[|D_{r'|_S}|^{-1}] \cdot P[R = r] = \\ &= E[|D_{R'|_S}|^{-1}] \geq \frac{1}{E[|D_{R'|_S}|]} \end{aligned}$$

where the last inequality holds by Jensen's inequality [7]—which gives $E[\varphi(X)] \geq \varphi(E[X])$ for any convex function φ—and the fact that $\varphi(t) = \frac{1}{t}$ is convex on the positive real numbers. □

Remark 12. *Since $\forall r \in D$ and $\forall s \subseteq supp(r)$, it holds that $r \in D_{r'|_s}$, then $1 \leq E[|D_{R'|_S}|]$. We would like hypotheses on D, D′, and m implying $E[|D_{R'|_S}|] \leq \beta$, for $\beta \geq 1$ as small as possible. For the obtained β, D is perfectly $\frac{1}{\beta}$-de-anonymized by A^* with respect to D′.*

Toy Example. Assume $D' = D$ and suppose that the non-null values in D are taken uniformly and independently at random out of b different values. For simplicity put $b = 2$, so $\mathbb{V} = \{v_1, v_2, \perp\}$. Since $D' = D$, consider the adversary $\tilde{A}$ that uses as output an element from $D_{r|_s} = \{y \in D : y|_s = r|_s\}$. Fix $r \in D$ and let s be given by m non-null attribute values taken independently from $supp(r)$. Let, for each $j \in [m]$, $s_j := \{i_1, ..., i_j\}$ (note that $s = s_m$). For each $i \in [c]$, roughly 50% of $supp(i)$ corresponds to v_1's, and the rest to v_2's. Hence, $|D_{r|_{s_1}}| \leq 1 + \frac{n}{2}$, $|D_{r|_{s_2}}| \leq 1 + \frac{n}{2^2}$,..., $|D_{r|_{s_m}}| \leq 1 + \frac{n}{2^m}$, thus: $E[|D_{r|_S}| \leq 1 + \frac{n}{2^m}$. If $m = k + \log_2 n$, then $E[|D_{r|_S}| \leq 1 + \frac{1}{2^k}$, so by Theorem 11, $P[R = \tilde{A}(R'|_S, D)] \geq \frac{2^k}{2^k+1}$. With $k = 4$ we get that if $m \geq 4 + \log_2 n$, then $\tilde{A}$ perfectly 0.94-de-anonymizes D (roughly). With $n = 480,000$ as the Netflix database, we get $m \geq 4 + \log_2(480,000) \approx 22.9$.

The same argument holds for any $b \geq 2$, giving: $m \geq k + \log_b n \Rightarrow P[R = \tilde{A}(R'|_S, D)] \geq \frac{b^k}{b^k+1}$. If $b = 10$ (e.g., ratings), then $k = 2$ gives: $m \geq 2 + \log n \Rightarrow P[R = \tilde{A}(R'|_S, D)] \geq 0.99$, which for $n = 480,000$ gives $m \geq 7.7$. Thus, $m = 8$ non-null attribute values are sufficient to identify a whole record with probability 0.99.

This is an example of how Theorem 11 can be used. We would like to explore more realistic scenarios. For instance, $D' = D$ should be replaced by a different relationship between D and D'. Also, assuming that the non-null values are independent is unrealistic, since typically some pairs of attributes are highly correlated (e.g., a movie and its sequel).

5 Conclusions

In this paper we revisited Narayanan and Shmatikov work [19] on the de-anonymization of large sparse databases, such as the Netflix database. We discover and resolve technical issues with the proof of their main theoretical result (Theorem 1) for both the case where the auxiliary information has error and the case where the database contains null values.

The new results that we establish are very similar to Theorem 1 in [19], and can in turn be improved under assumptions on the sparsity of the database. These upgraded theorems, when evaluated on the Netflix database using its size and sparsity, guarantee levels of de-anonymization weaker but consistent with the heuristic de-anonymization of the Netflix database in [19].

We explore one aspect of the de-anonymization of the Netflix database, regarding the fact that the performance of the adversary improves when the auxiliary information includes rare attributes. This can be formalized via the notion of tail of the database, and the parameters describing the tail can be incorporated to the de-anonymization results, yielding more efficient attacks when the database has a long tail, since less auxiliary information can ensure the same level of success.

One of our theorems is then tested on the JCUSH database, which has a level of sparsity much lower than that of the Netflix database. This shows that high sparsity and a long tail are required in order to guarantee efficient de-anonymizations via our theoretical results. This becomes more evident when the same algorithm used to prove the abstract results is used to perform an empirical linkage attack on the JCUSH database, giving a satisfactory adversary success.

A path for future work is to establish formal results of de-anonymization dealing with a different type of auxiliary information. In our theorems we are assuming that the values in the auxiliary information could have error, but these are uniformly bounded by a parameter, which in addition is known by the adversary (D and D' are assumed to be $(1-\varepsilon)$-similar). Moreover, in our model, we assume that if $r_i = \perp$, then the corresponding value in the auxiliary information $(r'|_s)_i$ is also $\perp$ (since $s \subseteq supp(r)$). New results should contemplate the possibility of the auxiliary information taking non-null values on those attributes in which the original record is null.

Acknowledgements

The author is very grateful to Professor Stephen Watson for helpful discussions and comments on this article.

References

[1] Acquisti, A., Gross, R. (2009) Predicting Social Security Numbers From Public Data, PNAS, Vol. 106, no. 27, 10975-10980

[2] Adam, N. R., Wortmann, J. C. (1989) Security-Control Methods for Statistical Databases: A Comparative Study, ACM Computing Surveys, Vol. 21, Issue 4, 515-556

[3] Agrawal, R., Srikant, R. (2000) Privacy-preserving Data Mining, ACM SIGMOD Record, Vol. 29, Issue 2, 439-450

[4] Anderson, C. (2006) The Long Tail: why the Future of Business is Selling Less of More, Hyperion, New York

[5] Backstrom, L., Dwork, C., Kleinberg, J. (2007) Wherefore art thou r3579x? Anonymized Social Networks, Hidden Patterns, and Structural Steganography, Proc. of the 16th International Conference on World Wide Web, 181-190

[6] Chawla, S., Dwork, C., McSherry F., Smith A., Wee H. (2005) Toward Privacy in Public Databases, TCC, 363-385

[7] Chung, K. L. (2001) A Course in Probability Theory, Academic Press, San Diego

[8] Dalenius, T. (1977) Towards a Methodology for Statistical Disclosure Control, Statistik Tidskrift, 15, 429-444

[9] Dinur, I., Nissim, K. (2003) Revealing Information While Preserving Privacy, Proc. of the 22nd Symposium on PODS, 202-210

[10] Dwork, C., McSherry, F., Nissim, K., Smith, A. (2006) Calibrating Noise to Sensitivity in Private Data Analysis, Proc. of the 3rd TCC, 265-284

[11] Dwork, C., Yekhanin, S. (2008) New Efficient Attacks on Statistical Disclosure Control Mechanisms, Advances in Cryptology, Vol. 5157, 469-480

[12] Evfimievski, A. (2002) Randomization in Privacy Preserving Data Mining, ACM SIGKDD Explorations Newsletter, Vol. 4, Issue 2, 43-48

[13] Frankowski, D., Cosley, D., Sen, S., Terveen, L., Riedl, J. (2006) You are What You Say: Privacy Risks of Public Mentions, Proc. of 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 565-572

[14] Garfinkel, S. (2000) Database Nation: The Death of Privacy in the 21st Century, O'Reilly Media, Cambridge

[15] Golle, P. (2006) Revisiting the Uniqueness of Simple Demographics in the US Population, Proc. of the 5th ACM WPES, 77-80

[16] Martin, D. J., Kifer, D., Machanavajjhala, A., Gehrke, J., Halpern, J. (2007) Worst-case Background Knowledge for Privacy-preserving Data Publishing, Proc. of the ICDE, 126-135

[17] Narayanan, A., Shmatikov, V. (2009) De-anonymizing Social Networks, Proc. of the 2009 30th IEEE Symposium on Security and Privacy, 173-187

[18] Narayanan, A., Shmatikov, V. (2010) Myths and Fallacies of Personally Identifiable Information, Communications of the ACM, Vol. 53, Issue 6

[19] Narayanan, A., Shmatikov, V. (2008) Robust De-anonymization of Large Sparse Datasets, Proc. of the 2008 IEEE Symposium on Security and Privacy, 111-125

[20] R Development Core Team (2011) R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, www.R-project.org

[21] Sweeney, L. (1997) Weaving Technology and Policy Together to Maintain Confidentiality, Journal of Law, Medicine and Ethics, 25 (2-3), 98-110

[22] Joint Canada/United States Survey of Health (2004), Statistics Canada, Identification number jcush_82M0022_E_2004.

[23] Thompson, B., Yao, D. (2009) The Union-split Algorithm and Cluster-based Anonymization of Social Networks, Proc. of the 4th International Symposium on Information, Computer, and Communications Security, 218-227

[24] Wondracek, G., Holz, T., Kirda, E., Kruegel, C. (2010) A Practical Attack to De-Anonymize Social Network Users, Proc. of the 2010 IEEE Symposium on Security and Privacy, 223-238

[25] Zheleva, E., Getoor, L. (2009) To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles, Proc. of the 18th International Conference on World Wide Web, 531-540