Provable De-anonymization of Large Datasets with Sparse Dimensions

Anupam Datta, Divya Sharma, and Arunesh Sinha

Carnegie Mellon University
{danupam,divyasharma,aruneshs}@cmu.edu

Abstract. There is a significant body of empirical work on statistical de-anonymization attacks against databases containing micro-data about individuals, e.g., their preferences, movie ratings, or transaction data. Our goal is to analytically explain why such attacks work. Specifically, we analyze a variant of the Narayanan-Shmatikov algorithm that was used to effectively de-anonymize the Netflix database of movie ratings. We prove theorems characterizing mathematical properties of the database and the auxiliary information available to the adversary that enable two classes of privacy attacks. In the first attack, the adversary successfully identifies the individual about whom she possesses auxiliary information (an *isolation attack*). In the second attack, the adversary learns additional information about the individual, although she may not be able to uniquely identify him (an *information amplification attack*). We demonstrate the applicability of the analytical results by empirically verifying that the mathematical properties assumed of the database are actually true for a significant fraction of the records in the Netflix movie ratings database, which contains ratings from about 500,000 users.

Keywords: Privacy, database, de-anonymization

1 Introduction

In recent years, there has been a steady increase in the number of publicly released databases containing micro-data about individuals, e.g., their preferences, movie ratings, or transaction data. There are a number of reasons for this phenomena, for example, enabling useful tasks such as improving recommendation systems [8] and providing transparency about the activities of government agencies, such as the justice system [1].

At the same time, these databases raise privacy concerns because they contain personal information about individuals that they may not want to share with the whole world. In order to alleviate these concerns, various techniques have been proposed to "anonymize" databases before releasing them. These anonymization techniques have been developed in response to specific classes of attacks observed in practice. It is now well known that just removing obvious identifiers, such as names, social security numbers and IP addresses, is not sufficient for anonymization—an adversary can use auxiliary information acquired from

other sources to de-anonymize individual data records by computing database joins. Examples of attacks of this form include de-anonymizing records in a hospital discharge database and AOL search logs [2, 16]. More recently, a class of statistical de-anonymization attacks have been presented that work on high-dimensional micro-data and the applicability of the attack has been *empirically* demonstrated on the publicly available Netflix movie ratings database; the attacks work even when the released data has been perturbed and the auxiliary information available to the adversary is noisy [11].

Our goal is to *analytically* explain why such attacks work. Specifically, we analyze a variant of the Narayanan-Shmatikov weighted algorithm that was used to effectively de-anonymize the Netflix database of movie ratings. Roughly, this algorithm takes as input noisy auxiliary information about an individual (e.g., movie ratings) and a database, and outputs the record in the database that has the highest score on the common attributes. The score is a weighted sum of the similarity of individual attributes where rare attributes are assigned higher weights. We prove theorems characterizing mathematical properties of the database and the noisy auxiliary information available to the adversary that enable two classes of privacy attacks. In the first attack, the adversary successfully identifies the individual about whom she possesses auxiliary information (an *isolation attack*), i.e., the algorithm outputs the correct record. In the second attack, the adversary learns additional information about the individual, although she may not be able to uniquely identify him (an *information amplification attack*), i.e., the algorithm outputs a record of a 'similar' individual. We empirically verify that the mathematical properties assumed of the database are actually true for a significant fraction of the records in the Netflix movie ratings database, which contains ratings from about 500,000 users, even when the auxiliary information is noisy. Thus, our theorems formally explain why these attacks work on the Netflix database.

The analytical and empirical study led to several insights about the nature of de-anonymization attacks. First, it provides a technical characterization of an observation due to Narayanan and Shmatikov [12] that *"any information that distinguishes one person from another can be used to re-identify anonymous data"*. This intuition is reflected in the weighted scoring algorithm: rare attributes directly correspond to distinguishing attributes because, by definition, a record's non-null value for a rare attribute means that that record is different from the many records that have null value for the rare attribute. In addition, the weighted linear combination is combining different distinguishing attributes into a single metric that distinguishes the records better than the individual attributes. While the effectiveness of this idea has been empirically demonstrated [11], to the best of our knowledge our theorem about the isolation attack is the first analytical characterization of this idea. (Note that while Narayanan and Shmatikov present analytical results about a simpler unweighted algorithm, they do not analyze the weighted algorithm that was actually used in the empirical study.)

Second, we formulate and prove the theorem about the information amplification attack under the following assumption: *records which agree on distinguishing (rare) attributes must be similar overall.* This assumption is justified by the observation that people with similar tastes, e.g., in rare movies are likely to also share similar opinions on other movies. It is important to note that this assumption may not hold for all databases, but our empirical results demonstrate that it holds for the Netflix database.

Third, in formulating our theorems, a guiding consideration was that the *assumptions should be empirically verifiable* on a released database even if we do not know what distribution the database was drawn from. We conduct experiments to verify the assumptions on the Netflix database. We find that the assumptions for both the theorems for the weighted algorithm are true for a significant fraction of the records. In particular, the assumptions required for the isolation attack hold for 90% of the records when the perturbation in the auxiliary information is less than 10%. As expected, the percentage of records for which the assumption holds decreases as the perturbation is increased, and increases as the number of attributes in the auxiliary information is increased. For the information amplification attack, we verify that if auxiliary information aux_y about a target record y is not too perturbed ($< 10\%$) and a significant fraction of the attributes in aux_y (> 0.75) are rare, then for a significant fraction of target records (> 0.90), if any record r is similar (similarity value > 0.75) to aux_y, then r is also similar (similarity value > 0.65) to y. Also, as the fraction of rare attributes in auxiliary information increases and the threshold for similarity between auxiliary information and the output record increases, the similarity between the target record and the output record increases.

Finally, we comment on the relation of our results to prior work on *quasi-identifers.* Observing that de-anonymization attacks are possible even when obviously identifying information (such as names and social security numbers) is removed from micro-data databases, Samarati and Sweeney [13, 15] introduced the concept of quasi-identifers—attributes that could be used to re-identify individuals by linking with another source. An important challenge that this line of work does not address (see also [9,10,17]) is *how to identify quasi-identifiers.* As mentioned earlier, our formalization captures the intuition that any attribute can be a quasi-identifier—the rarer the attribute, the greater is its contribution towards distinguishing an individual. Thus, one might view our results as providing a semantic characterization of database properties and auxiliary information that *provably* enable de-anonymization by linking in this more general setting. Note that in our characterization, the analog of a quasi-identifier (the property that enables linking attacks) is not just a property of the database; it also depends on the adversary's auxiliary information.

The rest of the paper is organized as follows. Section 2 describes related work. Section 3 presents preliminary definitions. Section 4 presents an analysis of the simpler generic (unweighted) algorithm for de-anonymization [11]. Section 5 presents the main technical results of the paper—the analysis of isolation and information amplification attacks using the weighted algorithm. Section 6

presents empirical results demonstrating that the analytical results apply to the Netflix database. Finally, Section 7 presents conclusions and directions for future work.

2 Related Work

Dwork and Naor [5] prove a fundamental tradeoff between utility and a strong form of privacy property (due to Dalenius [4]) capturing the intuition that nothing about an individual should be learnable from the database that cannot be learned without access to the database. They prove that it is not possible to satisfy this definition if the database is to have any utility assuming that the adversary has *arbitrary* auxiliary information. In contrast, in our work we seek to characterize a restricted class of auxiliary information (which adversaries may realistically possess) and database perturbation techniques (employed in practice to release micro-data) for which de-anonymization attacks provably work. The starting point of our analysis is the formal model proposed by Narayanan and Shmatikov, which they used to analyze a simpler algorithm [11] (see also [3]). In the original paper by Narayanan and Shmatikov [11], two algorithms have been proposed- generic and weighted scoring algorithms. The first algorithm (generic scoring algorithm) is analyzed, however, it is not used in the actual attack. In our work, we analyze a minor variant of the weighted scoring algorithm, used in the attack on Netflix database by Narayanan and Shmatikov. We present an alternative definition of the similarity metric (as mentioned in Section 3), a different notion of "eccentricity" (as described in Section 5) and prove theorems characterizing both the isolation and information amplification attacks using the weighted algorithm that was only empirically evaluated in their paper. In addition, we empirically validate that the assumptions hold on the Netflix database.

Boreale et al [3] present an alternative approach to analyzing de-anonymization attacks. Specifically, the authors model the process of de-anonymization of a database using an Information Hiding System (IHS) whose input includes identified records, output includes observable information (e.g., perturbed attributes), and the conditional probability matrix models the process of acquiring auxiliary information. They prove theorems characterizing information leakage using a sparsity assumption about rows in the database (which roughly captures the idea that no two records in the database are similar except with low probability) and assuming that the auxiliary information includes attributes sampled uniformly at random. In contrast, we use an assumption about sparsity of columns (rare attributes) and leverage knowledge of rare attributes in the auxiliary information to provide an analysis of the weighted algorithm of Narayanan and Shmatikov, which as also remarked by Boreale et al., allows more effective de-anonymization.

In a separate attack, ratings from an "anonymized" database released by another recommender service, Movielens, were linked to individuals by using movies publicly mentioned by users online [7]. We use a similar scoring method-

ology as was proposed by Narayanan and Shmatikov [11] and Frankowski et al [7], however the algorithms we analyze allow for information to be perturbed, unlike the attack on the MovieLens database.

Differential privacy is useful for releasing privacy-preserving statistics [5, 6]. However, the focus of this work is on databases containing microdata.

3 Definitions

In this section, we describe notation used throughout the paper and present definitions of the asymmetric similarity metric and perturbed auxiliary information.

Let D_0 denote the original database containing individuals' records (which have not been anonymized). An 'anonymized' version of this database is released as D with n rows and c columns. D is obtained by running an anonymization algorithm on the original database D_0. Each row in the database corresponds to a different individual's record and each column corresponds to an attribute. Let $r(i)$ denote the value of i^{th} column for the row corresponding to record $r \in D$. The target record (i.e., the record the adversary is looking for), denoted by y, is assumed to be always present in the released database D. The set of non-null values of any record r is denoted by $supp(r)$; similarly, the set of non-null values in any column i is denoted by $supp(i)$.

In order to compare the values of any two records in the database, we define a similarity metric S.

Definition 1 (Asymmetric Similarity Metric). *Similarity of record r when compared against record y is defined as:*

$$S(y, r) \triangleq \sum_{i \in supp(y)} \frac{T(y(i), r(i))}{|supp(y)|} \tag{1}$$

where

$$T(y(i), r(i)) \triangleq 1 - \frac{|y(i) - r(i)|}{p} \tag{2}$$

and p is the maximum possible difference between values of the column i.

Here, $T(y(i), r(i))$ is defined as a scaled measure of difference between two records y and r when compared on the i^{th} attribute. The value of each column is scaled by p (the range of values for the column), so that the value for $T(.,.)$ lies in the interval $[0, 1]$.

In contrast, Narayanan and Shmatikov use a symmetric similarity measure Sim that compares two records on the union of the non-null attributes in the two records [11]. Observe that when $S(y, r)$ is high, r reveals information about the attributes of y. However, even if $S(y, r)$ is high, $Sim(y, r)$ could be low if r has a large number of a non-null attributes that do not overlap with non-null attributes in y. Thus, we believe that S is a better measure to use in the design and analysis of de-anonymization attacks than Sim.

Definition 2 ((m, γ)-perturbed auxiliary information). *Auxiliary information about record $y \in D$, denoted by aux_y, contains perturbed values of m non-null attributes sampled from attributes in record y. aux_y is defined to be (m, γ)-perturbed if $\forall i \in supp(aux_y).T(y(i), aux_y(i)) \geq 1 - \gamma$ where $0 \leq \gamma \leq 1$.*

Note that the definition above abstracts away from whether the perturbation is in the released database or in the auxiliary information by noting that the relevant property is a lower bound on the attribute-wise similarity between the auxiliary information and the target record. The structure of aux_y is similar to that of a record in the database, with m columns ($|supp(aux_y)| = m$).

4 Analysis of Generic Scoring Algorithm

In this section, we analyze and obtain provable bounds for the generic scoring algorithm proposed by Narayanan and Shmatikov. The generic algorithm considers all the attributes in the auxiliary information as equally important for re-identification. Our theorem uses the asymmetric similarity metric defined in Section 3 and gives a lower bound on the similarity of the record output by the algorithm with the target record y. However, this algorithm might not de-anonymize records effectively as the effect of perturbation in even a single attribute of the auxiliary information would lower the overall score [11]. We include this analysis for completeness.

The scoring function used by the generic scoring algorithm is defined below.

Definition 3 ($Score_g$). *$Score_g(aux_y, r)$ of a record $r \in D$ w.r.t. auxiliary information aux_y about target record y is defined as:*

$$Score_g(aux_y, r) = min_{i \in supp(aux_y)} T(aux_y(i), r(i)) \tag{3}$$

The Narayanan-Shmatikov generic scoring algorithm is described in Algorithm 1.

Algorithm 1 Generic Scoring Algorithm

- Fix a target record y
- aux_y is (m, γ)-perturbed auxiliary information about target record y
- Compute $Score_g(aux_y, r)$ for every record r in the dataset
- Form a matching set of records that satisfy:

$$M = \{r \in D : Score_g(aux_y, r) \geq 1 - \gamma\} \tag{4}$$

- Output a randomly chosen record from the matching set.

We prove a lower bound on the similarity of the record output by the algorithm with the target record, assuming that the auxiliary information is (m, γ) perturbed. The full proof is included in the appendix.

Theorem 1. *Let y denote the target record from given database D. Let aux_y denote (m, γ)-perturbed auxiliary information, uniformly sampled from the attributes in record y. Let $\epsilon > 0$. Then with probability $\geq 1 - g$, a record o can be found in the dataset such that the value of $S(y, o)$ is greater than $1 - 2\gamma - \epsilon$, where $g = e^{-2*\epsilon^2*m}$*

5 Analysis of Weighted Scoring Algorithm

In this section, we analyze a variant of the weighted scoring algorithm proposed by Narayanan and Shmatikov. The weighted scoring algorithm gives higher weight to 'rare' attributes in the auxiliary information. We present the algorithm and two theorems characterizing the effectiveness of the algorithm for isolation and information amplification attacks. Specifically, we prove that if the score of a record is significantly higher than the scores of other records, then the record can be isolated using the weighted scoring algorithm. We also prove a theorem that quantifies the probability and degree of an information amplification attack assuming that (a) a fraction of the attributes in perturbed auxiliary information is rare; and (b) if people agree on several rare attributes, then with high probability they are also similar on other attributes.

We begin by presenting definitions that are used in the description of the algorithm and its analysis.

Definition 4 (Weight of an attribute). *The weight of an attribute i is denoted by w_i and is defined as $w_i = \frac{1}{\log_2 |supp(i)|}$*[1]*.*

We denote the scaled sum of weights of attributes in aux_y by $M = \frac{\sum_{i \in supp(aux_y)} w_i}{|supp(aux_y)|}$ where aux_y refers to the perturbed auxiliary information corresponding to the target record y. Next, we formalize the notion of rarity of an attribute.

Definition 5 (t-rare attribute). *An attribute is said to be t-rare if $w_i = \frac{1}{\log_2 |supp(i)|} \geq t$ where t is a threshold value and $0 < t \leq 1$.*

Definition 6 ((δ, t)-rare auxiliary information). *Auxiliary information about record $y \in D$, denoted by aux_y, is said to be (δ, t)-rare if the fraction of t-rare attributes in auxiliary information aux_y, denoted by δ_{aux_y} equals δ where $0 < \delta, t \leq 1$*

Definition 7 ($Score_w$). *$Score_w(aux_y, r)$ of a record $r \in D$ w.r.t. auxiliary information aux_y about target record y is defined as:*

$$Score_w(aux_y, r) = \sum_{i \in supp(aux_y)} \frac{w_i * T(aux_y(i), r(i))}{m} \tag{5}$$

[1] We assume that $|supp(i)| > 2$; for the Netflix dataset we have $\min_i |supp(i)| = 3$

Definition 8 (Eccentricity). *We define eccentricity e as*

$$e(aux_y, D) = \max_{r \in D}(Score_w(aux_y, r)) - \max_{2, r \in D}(Score_w(aux_y, r)) \quad (6)$$

where $r \in D$, y is the target record and aux_y refers to the perturbed auxiliary information obtained from target record y. $\max_{r \in D}(Score_w(aux_y, r))$ and $\max_{2, r \in D}(Score_w(aux_y, r))$ refer to the highest and second highest value, respectively, of $Score_w(aux_y, r)$ taken over the scores of all the records r in D.

Eccentricity is a measure of how far apart the highest scoring record is from the second highest score when a scoring algorithm is employed. The eccentricity measure would be useful in eliminating false positives in the result output by the algorithm, as described in Algorithm 2.

Algorithm 2 Weighted Scoring Algorithm

- Fix a target record y
- aux_y is (m, γ)-perturbed auxiliary information about target record y
- Compute $Score_w(aux, r)$ for every record r in the dataset
- Output the record with the highest score if $e(aux_y, D) > T$, where T is a preset threshold[2], else output NULL. Let o denote the record output by the algorithm.

Isolation Attack In the first attack, an adversary with some auxiliary information successfully isolates an individual from a database. We prove that for a given target record y, if auxiliary information aux_y is (m, γ)-perturbed and if the score of the record o output by the algorithm differs from the second-highest score by a certain threshold, then $o = y$. The intuition behind the assumption in this theorem is that if a record is significantly different from other records on attributes present in auxiliary information, then the record can be isolated using the weighted scoring algorithm.

The proof proceeds as follows: by using the assumption that aux_y is (m, γ)-perturbed, we derive a lower bound for the score of target record y when compared with aux_y. We prove the main result in the theorem by contradiction. We assume that the maximum possible value of the scoring function when computed over all records in the database is not equal to the score of the target record y. We show that this assumption leads to the conclusion that the maximum possible score is greater than M, which is not possible since M equals the value of the scoring function assuming that every attribute in aux_y matches completely with the attributes in the record being compared against.

Theorem 2. *Let y denote the target record from given database D. Let aux_y denote (m, γ)-perturbed auxiliary information about record y. If the eccentricity measure $e(aux_y, D) > \gamma M$ where $M = \frac{\sum_{i \in supp(aux_y)} w_i}{|supp(aux_y)|}$ is the scaled sum of weights of attributes in aux_y, then*

1. $\max_{r \in D}(Score_w(aux_y, r)) = Score_w(aux_y, y)$.
2. *Additionally, if only one record has maximum score value* $= Score_w(aux_y, y)$, *then the record returned by the algorithm* o *is the same as target record* y.

Proof. By definition of Similarity metric $S(.,.)$, for any record $r \in D$ and given target record y, $S(y, r) = \sum_{i \in supp(y)} \frac{T(y(i), r(i))}{k}$, where $k = |supp(y)|$.

Also, by definition of $Score_w(.,.)$,

$$Score_w(aux_y, r) = \frac{\sum_{i \in supp(aux_y)} w_i * T(aux_y(i), r(i))}{m} \tag{7}$$

where $w_i = \frac{1}{\log_2 |supp(i)|}$. Given the assumption $\forall i \in supp(aux_y).T(y(i), aux_y(i)) \geq 1 - \gamma$, we can use equation 7, to conclude that

$$\begin{aligned} Score_w(aux_y, y) &= \frac{\sum_{i \in supp(aux_y)} w_i * T(aux_y(i), y(i))}{m} \\ &= \frac{\sum_{i \in supp(aux_y)} w_i * T(y(i), aux_y(i))}{m} \\ &\geq \frac{\sum_{i \in supp(aux_y)} (1-\gamma) w_i}{m} \geq (1-\gamma) * \frac{\sum_i w_i}{m} = (1-\gamma) * M \end{aligned}$$

since $T(y(i), aux_y(i)) = T(aux_y(i), y(i))$ by definition.

We prove the result in our theorem by contradiction. We assume that

$$\max_{r \in D}(Score_w(aux_y, r)) \neq Score_w(aux_y, y) \tag{8}$$

Observe that

$$\max_{r \in D}(Score_w(aux_y, r)) > Score_w(aux_y, y) \text{ (from equation 8)} \tag{9}$$

If $\max_{r \in D}(Score_w(aux_y, r))$, is greater than $Score_w(aux_y, y)$ then

$$\max_{2, r \in D}(Score_w(aux_y, r)) \geq Score_w(aux_y, y) \tag{10}$$

since $\max_{2, r \in D}(Score_w(aux_y, r))$ is the second highest value of all scores.

Further, it is assumed that $e(aux_y, D) > \gamma M$, therefore,

$$\begin{aligned} \max_{r \in D}(Score_w(aux_y, r)) &> \gamma M + \max_{2, r \in D}(Score_w(aux_y, r)) \\ &> \gamma M + Score_w(aux_y, y) > \gamma M + (1-\gamma) * M = M \end{aligned}$$

which is not possible since M is the maximum possible score for any record r in the database as shown below

$$\begin{aligned} Score_w(aux_y, r) &= \frac{\sum_{i \in supp(aux_y)} w_i}{m} * T(aux_y(i), r(i)) \\ \max_{r \in D}(Score_w(aux_y, r)) &\leq \frac{\sum_{i \in supp(aux_y)} w_i}{m} \leq M \end{aligned}$$

since $\max(T(.,.)) = 1$

Therefore, our assumption is wrong and we conclude that

$$\max_{r \in D}(Score_w(aux_y, r)) = Score_w(aux_y, y)$$

Also, since we assumed that target record y is always part of the released database D, therefore, if there is only one record with the maximum score and $Score_w(aux_y, y)$ is same as the maximum score then the the record with maximum score has to be y, which is returned by the algorithm. Hence proved.

Information Amplification Attack In the second attack, although an adversary may not be able to uniquely isolate a record, she can still obtain additional information about the randomly chosen target record y under certain assumptions about the database. The intuition that if people agree on several rare attributes, then with high probability they are similar on other attributes, guided us to define a function f_D for database D. We use f_D to measure the overall similarity of the target record y and r by an indirect comparison of the rare attributes of y and the record r. The comparison is indirect because we use aux_y as a proxy for y and compare the rare attributes of aux_y with r. To capture the intuition that the agreement must happen on rare attributes the function f_D depends on the fraction of rare attributes in aux_y (η_1). To capture the intuition that there should be agreement on the rare attributes, f_D also depends on a lower bound (η_2) for $S(aux_y, r)$. In addition, to capture the fraction of target records (η_3) for which the overall similarity of the target record y and r is given by f_D we also include η_3 as a parameter for f_D. We define two parameterized sets before formalizing this intuition in Property 1.

Definition 9 (D_{m,η_1}). *D_{m,η_1} is the subset of the records of a database D that have no less than m non-null attributes and at least η_1 of those attributes are t-rare.*

We denote the above set as D_{m,η_1}, ignoring the parameter t for notational ease.

Definition 10 (Aux_{y,m,η_1}). *Aux_{y,m,η_1} is the set of all (m,γ)-perturbed and (η_1, t)-rare sets of auxiliary information about record y.*

Again, we ignore some parameters in Aux_{y,m,η_1} for the sake of notational ease. We assume that for the given database D there exists a function f_D with $Range(f_D) \subseteq [0,1]$ and the following property:

Property 1 *Choose any m and η_1. Let y be chosen uniformly at random from D_{m,η_1}. Let aux_y be chosen uniformly at random from Aux_{y,m,η_1}. Then*

$$\forall \eta_2, \eta_3, r.\ (S(aux_y, r) \geq \eta_2) \rightarrow Pr[S(y, r) \geq f_D(\eta_1, \eta_2, \eta_3)] \geq \eta_3$$

where $r \in D$. The probability is over the random choices made in choosing y.

We state the theorem below.

Theorem 3. *Let t and γ be in $(0,1)$. Fix any $l_1 \in (0,1)$. Let y denote the target record chosen uniformly at random from D_{m,l_1}. Let aux_y denote a (m,γ)-perturbed and (l_1,t)-rare auxiliary information about record y chosen uniformly at random from Aux_{y,m,l_1}. Additionally, we assume the existence of function $f_D(.,.,.)$ that satisfies Property 1. Then $Pr[S(y,o) \geq f_D(l_1,l_2,\eta_3)] > \eta_3$, where $l_2 = \frac{(\sum_{i\in supp(aux_y)} w_i)^2}{\sum_{i\in supp(aux_y)} (w_i)^2} \frac{(1-\gamma)^2}{m}$, o is the record output by the Weighted Algorithm and the probability is taken over the random choices made in choosing y.*

The proof proceeds as follows:

1. We derive a relationship between $S(aux_y,r)$ and $Score_w(aux_y,r)$ by using the Cauchy-Schwarz inequality [14] for any record r.
2. By using the assumption that aux_y is (m,γ)-perturbed, we derive a lower bound for $Score_w(aux_y,o)$. Using this and the last step we obtain a lower bound for $S(aux_y,o)$.
3. By using this bound in conjunction with the function f_D stated above, we give a probabilistic guarantee about $S(y,o)$.

Proof. Let $x_i(y,r) = T(y(i),r(i))$ for any record $r \in D$. Therefore, $S(y,r) = \sum_i \frac{x_i(y,r)}{k}$, where $k = |supp(y)|$. Also,

$$\begin{aligned} Score_w(aux_y,r) &= \frac{\sum_{i\in supp(aux_y)} w_i * T(aux_y(i),r(i))}{m} \\ &= \frac{\sum_{i\in supp(aux_y)} w_i * x_i(aux_y,r)}{m} \end{aligned}$$

We prove the first part of the proof by Cauchy Schwarz inequality,

$$\left(\sum_i A_i B_i\right)^2 < \sum_i A_i^2 \sum_i B_i^2$$

Therefore

$$\left(\sum_{i\in supp(aux_y)} w_i * x_i(aux_y,r)\right)^2 < \left(\sum_{i\in supp(aux_y)} w_i^2\right)\left(\sum_{i\in supp(aux_y)} x_i(aux_y,r)^2\right)$$

Since $0 \leq T(aux_y(i),r(i)) \leq 1$

$$\left(\sum_{i\in supp(aux_y)} x_i(aux_y,r)^2\right) \leq \left(\sum_{i\in supp(aux_y)} x_i(aux_y,r)\right)$$

Therefore,

$$\left(\sum_{i\in supp(aux_y)} w_i * x_i(aux_y,r)\right)^2 < \left(\sum_{i\in supp(aux_y)} (w_i)^2\right)\left(\sum_{i\in supp(aux_y)} x_i(aux_y,r)\right)$$

By definition of $Score_w(aux_y, r)$ and $S(aux_y, r)$ we get

$$(m * Score_w(aux_y, r))^2 < (\sum_{i \in supp(aux_y)} (w_i)^2) S(aux_y, r) m$$

$$S(aux_y, r) > \frac{(m * Score_w(aux_y, r))^2}{m \left(\sum_{i \in supp(aux_y)} (w_i)^2\right)} = \frac{m * (Score_w(aux_y, r))^2}{\sum_{i \in supp(aux_y)} (w_i)^2}$$

For the second step of the proof we use the assumption $\forall i \in supp(aux_y).T(y(i), aux_y(i)) \geq 1 - \gamma$. We can use the definition of $Score_w(.,.)$ (equation 7), to conclude that

$$\begin{aligned} Score_w(aux_y, y) &= \frac{\sum_{i \in supp(aux_y)} w_i * T(aux_y(i), y(i))}{m} \\ &= \frac{\sum_{i \in supp(aux_y)} w_i * T(y(i), aux_y(i))}{m} \\ &\geq \frac{\sum_{i \in supp(aux_y)} (1-\gamma) w_i}{m} \\ &\geq (1-\gamma) * \frac{\sum_i w_i}{m} \geq (1-\gamma) * M \end{aligned}$$

since $T(y(i), aux_y(i)) = T(aux_y(i), y(i))$ by definition.

Also since o has the max score $Score_w(aux_y, o) \geq Score_w(aux_y, y)$ and hence

$$\begin{aligned} Score_w(aux_y, o) &\geq \frac{\sum_{i \in supp(aux_y)} w_i * T(aux_y(i), y(i))}{m} \\ &\geq (1-\gamma) \frac{\sum_{i \in supp(aux_y)} w_i}{m} \geq (1-\gamma) M \end{aligned}$$

Substituting in equation derived for $S(aux_y, o)$ above,

$$S(aux_y, o) > \frac{m(Score_w(aux_y, o))^2}{(\sum_{i \in supp(aux_y)} w_i^2)} > \frac{m((1-\gamma)M)^2}{(\sum_{i \in supp(aux_y)} w_i^2)} > \frac{(\sum_{i \in supp(aux_y)} w_i)^2}{\sum_{i \in supp(aux_y)} (w_i)^2} \frac{(1-\gamma)^2}{m}$$

Thus, $S(aux_y, o) > l_2$.

Finally for the last part of the proof, we use the assumption that y was chosen uniformly at random from D_{m,l_1}, aux_y was chosen uniformly at random from Aux_{y,m,l_i} and the result above that $S(aux_y, o) > l_2$ to invoke Property 1 and claim the following:

$$Pr[S(y, o) \geq f_D(l_1, l_2, \eta_3)] \geq \eta_3$$

To summarize, we use a function f_D parametrized by a database D in formulating and proving the theorem about the information amplification attack. The idea here is that the theorem provides bounds on the information amplification attack for 'any' database D for which there exists an f_D such that the assumptions in the above stated theorem holds. Note that the bounds will be

good (i.e., the information amplification attack is effective) if η_3 is close to 1 (i.e., the attack succeeds with high probability) and the value of f_D is also close to 1 (i.e., the target record is very similar to the record output by the algorithm) given (a) the fraction of rare attributes in the auxiliary information (l_1) and (b) the similarity between the auxiliary information and the record output by the algorithm (l_2). We demonstrate in the next section that the function f_D computed for the Netflix database enables us to claim that with high probability, the output of the Weighted Algorithm run on the Netflix database will be similar to the target record.

6 Empirical Results

For empirically testing the assumptions in our theorems, we use the 'anonymized' Netflix database with 480,189 users and 17,770 movies, also used by Narayanan and Shmatikov. We run the modified Narayanan-Shmatikov weighted scoring algorithm as described in Section 5. Note that when we use m attributes in auxiliary information, we filter out records that have less than m attributes. Additionally, when we have the condition that δ_{aux_y} is a fixed fraction, this leads to more records being filtered out as the criteria is not met for these records. The percentage values claimed in all our results are percentage of records that are not filtered out. The following table shows the fraction of records that get filtered out for different values of m and t.

m	t	Percentage of records	m	t	Percentage of records
8	0.07	28.4	8	0.075	38.4
10	0.07	31.3	10	0.075	41.4
20	0.07	46.6	20	0.075	56.9

Table 1. Percentage of records that get filtered out, when t= 0.07, 0.075

We list some of our key findings and explain these in detail.

1. Isolation Attack
   - We verify the percentage of records in the database for which both the assumptions in Theorem 2 presented in Section 5 hold true, over the Netflix database. Our empirical analysis verifies that the assumptions hold true for majority of records.
   - We also test the assumptions for varying levels of perturbation in aux_y.
   - Additionally we compute the percentage of records for which the eccentricity assumption holds when we vary threshold for rarity of an attribute, t and number of attributes in aux_y, m.
   - As compared to the attack demonstrated by Narayanan and Shmatikov [11], we do not use dates for analysis. However, we consider perturbation in ratings in aux_y, as opposed to exact ratings being present in aux_y.

2. Information Amplification Attack
   - We develop an algorithm that computes the value of f_D for different values of the parameters γ, η_1, η_2 and η_3 for any database D and auxiliary information aux_y.
   - Our results show that for the Netflix database the function f_D is monotonically increasing in η_1, η_2 and tends to 1 as η_1, η_2 increases. Then the weighted scoring algorithm output will be quite similar to the target record for Netflix database, hence the Narayanan-Shmatikov weighted scoring algorithm was successful in finding attacks.

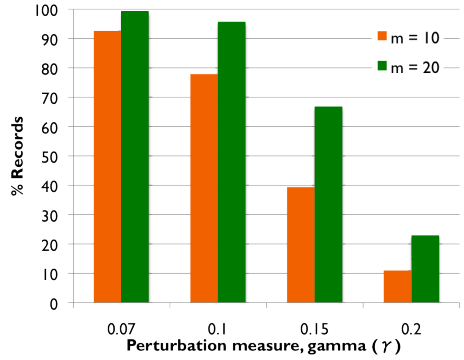
6.1 Isolation attack

Verifying assumptions for varying levels of perturbation in aux_y For the first result based on the isolation attack, we plot the fraction of records for which the eccentricity assumption holds, against the value of perturbation in auxiliary information aux_y, where fraction of rare attributes in aux_y (δ_{aux_y}) = 0.75. These results have been obtained by averaging the results from a sample of $10,000$ records randomly chosen with replacement. We obtain results for varying levels of perturbation in auxiliary information, $\gamma = 0.07, 0.1, 0.15, 0.2$. The results are shown and plotted in Figure 1. In this figure, we consider an attribute as rare if the column corresponding to the movie has weight ≥ 0.07 ($t = 0.07$) which implies that any column with less than $\sim 19,972$ entries will be defined as rare.

We conclude that when perturbation in aux_y is less than 10%, then the score of the best-match record exceeds the second-best score by a value greater than our theoretic threshold ($= \gamma * M$) for a significant fraction of the records (> 0.90), which implies that $> 90\%$ of the records can be successfully isolated. Also, we observe that as perturbation in auxiliary information (γ) increases, the number of records for which the assumption holds decreases. One factor causing this decrease could be that an increase in γM implies that the best match record would need to be different from the second highest score by a much higher value than when γ is lower, which may not always be true. However, we note that even with 20% perturbation, the assumption holds for $> 10\%$ of the records when the auxiliary information set contains 10 attributes. There are approximately 500000 users in the database; without considering the records that get filtered out, the attack still affects more than 34,000 users, which is quite significant.

Additionally in Figure 1, we also vary the number of attributes m in the auxiliary information aux_y; specifically we run the algorithm for $m = 10, 20$. We observe that as the number of attributes in auxiliary information set increases, the fraction of target records for which the eccentricity assumption holds and thus fraction of target records which can be isolated from a database, increases.

Verifying assumptions for unperturbed aux_y We compute the fraction of records that are isolated for $m = 8, \gamma = 0, \delta_{aux_y} = 0.75$. Since the perturbation γ in aux_y is 0, the score of the best-match record exceeds the second-best score by γM trivially. So for $> 99\%$ of the records that have greater than 8 attributes and



m	γ	% of records	m	γ	% of records
10	0.07	92.45	20	0.07	99.18
10	0.1	77.73	20	0.1	95.54
10	0.15	39.29	20	0.15	66.73
10	0.2	10.97	20	0.2	22.88

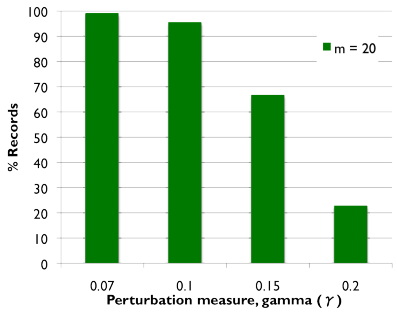
Fig. 1. Percentage of records for which eccentricity assumption holds when $t = 0.07$

more than 6 rare attributes and 2 non-rare attributes, there is only one record with the highest score and all these target records can be isolated. However, if we do not filter out records that have less than 8 attributes we get the result that 72% of all the records can be isolated when threshold for rarity of an attribute, $t = 0.07$ and 61% of all the records can be isolated when $t = 0.075$. This conclusion is not as good as the results obtained by Narayanan and Shmatikov, as they de-identified 84% of the records in the database with exact ratings and no dates at all. However, our results are computed using the generalized variant of the weighted scoring algorithm and not the heuristically tuned algorithm that Narayanan and Shmatikov actually use in the experiments. Our guarantees are supported by the theorems in Section 5, however as the authors themselves point out, the specifically tuned parameters in their algorithm might not work for another database.

Additionally in Figure 2, we plot the fraction of records for which eccentricity assumption holds when we consider an attribute as rare if the weight of the column i corresponding to the attribute has $w_i \geq 0.075$ ($t = 0.075$) which implies that any column with less than $\sim 10,000$ entries will be defined as rare. We plot the results for $m = 20$. In Figure 2, overall less attributes are considered as rare as compared to Figure 1.

6.2 Information Amplification Attack

Computing $f_D(\eta_1, \eta_2, \eta_3)$ We compute $f_D(.,.,.)$ using the routine shown in Algorithm 3. In the given code we would ideally want to take n as large as possible, but, that is not feasible. Hence we take n as 50 and then run the code 60 times and take the average value of f_D over the 60 runs as the final computed value. This is not the exact value of f_D, but is a good estimate.



m	γ	% of records
20	0.07	99.22
20	0.1	96.49
20	0.15	74
20	0.2	29.84

Fig. 2. Percentage of records for which eccentricity assumption holds when t= 0.075

Algorithm 3 Calculation of f_D

Require: m, t, γ
  for $\eta_1 : \{0.7, 0.8, .., 1.0\}$ **do**
    for $\eta_2 : \{0.5, 0.6, .., 1.0\}$ **do**
      for $i : \{1, 2, .., n\}$ **do**
        Choose y uniformly at random from D_{m,η_1}
        Choose aux_y uniformly at random from Aux_{y,m,η_1}
        $k_i = \min_{r \mid S(aux_y, r) \geq \eta_2} S(y, r)$
      end for
      $f_D(\eta_1, \eta_2, \eta_3) = \eta_3$ percentile of $k_1, .., k_n$
    end for
  end for

Value of $f_D(\eta_1, \eta_2, \eta_3)$ for varying levels of perturbation in aux_y and η_3 We plot the value of $f_D(\eta_1, \eta_2, \eta_3)$ by varying values of η_3, i.e. the probability of a record y having greater than $f_D(\eta_1, \eta_2, \eta_3)$ similarity with r given $\delta_{aux_y} = \eta_1$ and $S(aux_y, r) \geq \eta_2$. We obtain results for varying levels of perturbation in auxiliary information, $\gamma = 0.07, 0.1$, keeping the number of attributes in aux_y, $m = 10$. The results are plotted in Figures 3, 4. In each of these figures, we plot the value of $f_D(\eta_1, \eta_2, \eta_3)$ when $\eta_1 = \{0.7, 0.8, 0.9, 1.0\}$ and $\eta_2 = \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Figures 3, 4 show the value of $f_D(\eta_1, \eta_2, \eta_3)$ when $(\eta_3 = 0.9, \gamma = 0.07)$ and $(\eta_3 = 0.9, \gamma = 0.1)$ respectively. We conclude that, keeping γ, η_1, η_2 constant, the value of $f_D(\eta_1, \eta_2, \eta_3)$ decreases as η_3 increases, which reinforces the intuition that a higher probability of a record y having greater than $f_D(\eta_1, \eta_2, \eta_3)$ similarity with r, given $\delta_{aux_y} = \eta_1$ and $S(aux_y, r) \geq \eta_2$, is accompanied by a lower guarantee $f_D(\eta_1, \eta_2, \eta_3)$ of similarity.

Additionally, we observe that, for a constant value of η_3, the value of $f_D(\eta_1, \eta_2, \eta_3)$ increases as γ increases, but the value of γ is still small, and the function also becomes smooth with increasing γ, which implies that small perturbation of rare attributes does not decrease the knowledge of similarity between y and r that is gained from the knowledge of aux_y.

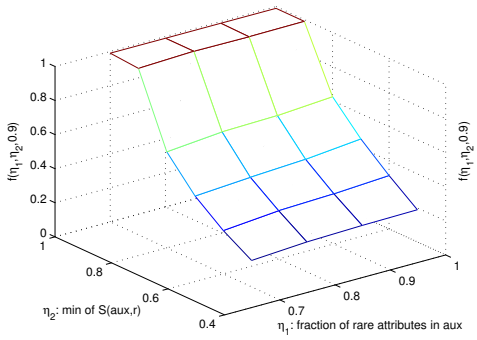


Fig. 3. Value of $f(\eta_1, \eta_2, \eta_3)$ when $\eta_3 = 0.9$ and $\gamma = 0.07$

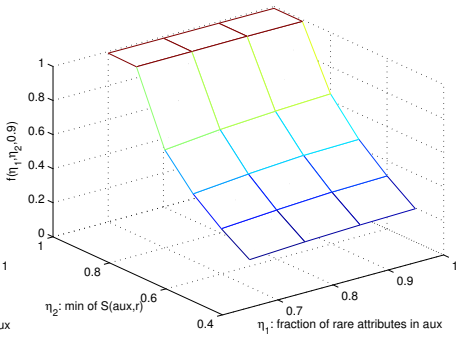


Fig. 4. Value of $f(\eta_1, \eta_2, \eta_3)$ when $\eta_3 = 0.9$ and $\gamma = 0.1$

$f_D(\eta_1, \eta_2, \eta_3)$ for unperturbed aux_y We also compute the value of $f_D(\eta_1, \eta_2, \eta_3)$ when $\gamma = 0$, which implies that the auxiliary information has no noise. The results are plotted in Figures 5, 6 for $m = 20$. Figures 5, 6 show the value of $f_D(\eta_1, \eta_2, \eta_3)$ when $(\eta_3 = 0.75, \gamma = 0.0)$ and $(\eta_3 = 0.9, \gamma = 0.0)$ respectively. All these graphs show that $f_D(\eta_1, \eta_2, \eta_3)$ is monotonically increasing in η_1 and η_2, and also tends to 1 as η_1, η_2 increase.

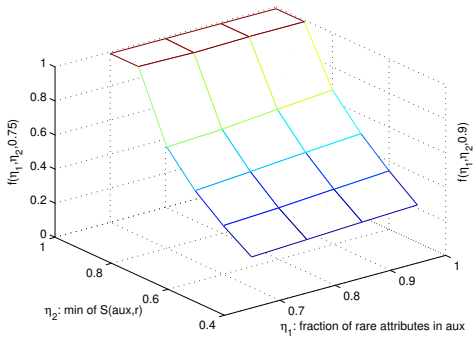


Fig. 5. Value of $f(\eta_1, \eta_2, \eta_3)$ when $\eta_3 = 0.75$ and $\gamma = 0$

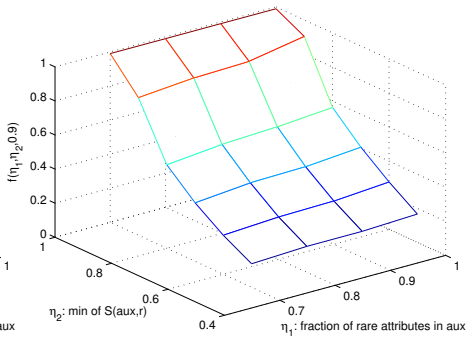


Fig. 6. Value of $f(\eta_1, \eta_2, \eta_3)$ when $\eta_3 = 0.9$ and $\gamma = 0$

7 Conclusion

We have presented a mathematical analysis of the effectiveness of the Narayanan-Shmatikov weighted algorithm in isolating individuals and carrying out information amplification attacks. Our empirical study of the Netflix database of movie ratings demonstrates that the assumptions about the database used in proving the theorems hold for a substantial fraction of records in the database. Thus, our theorems formally explain why these attacks work on the Netflix database. Indeed enabling this form of empirical validation without requiring knowledge of the distribution from which the database was drawn was a desideratum for our approach.

Our empirical results for the isolation attack are not as strong as those reported by Narayanan and Shmatikov (72% vs. 84% for parameter settings where a head-to-head comparison was possible). The difference could be caused by the generality of our assumptions. At a technical level, it would be interesting to understand if it is possible to prove an isolation theorem with stronger bounds using different assumptions about the dataset.

The technical result about the information amplification attack is formulated in terms of an abstract function f_D that depends on the database D. Our empirical results demonstrate that for the Netflix database $f_D(\eta_1, \eta_2, \eta_3)$ is monotonically increasing in η_1 and η_2, and also tends to 1 as η_1, η_2 increase. Our theorem predicts that this behavior of f_D implies that the Netflix database is de-anonymizable by the weighted scoring algorithm. It would be interesting to identify a class of distributions from which if databases are drawn they would satisfy this property.

Acknowledgments. We thank Anupam Gupta for suggesting the asymmetric similarity metric. We also thank Arvind Narayanan for useful discussions during the course of this work.

References

1. PACER- Public Access to Court Electronic Records. `http://www.pacer.gov`, [Last accessed 2011.12.16]
2. Barbaro, M., Zeller, T.: A Face Is Exposed for AOL Searcher No. 4417749. New York Times (August 09, 2006), available at `http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all`
3. Boreale, M., Pampaloni, F., Paolini, M.: Quantitative information flow, with a view. In: ESORICS. pp. 588–606 (2011)
4. Dalenius, T.: Towards a methodology for statistical disclosure control. Statistics Tidskrift 15, 429–444 (1977)
5. Dwork, C.: Differential privacy. In: ICALP. pp. 1–12. Springer (2006)
6. Dwork, C.: Differential privacy: a survey of results. In: Proceedings of the 5th international conference on Theory and applications of models of computation. pp. 1–19. TAMC'08, Springer-Verlag, Berlin, Heidelberg (2008), `http://dl.acm.org/citation.cfm?id=1791834.1791836`

7. Frankowski, D., Cosley, D., Sen, S., Terveen, L., Riedl, J.: You are What You Say: Privacy Risks of Public Mentions. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 565–572. SIGIR '06, ACM, New York, NY, USA (2006), http://doi.acm.org/10.1145/1148170.1148267
8. K. Hafner: And if You Liked the Movie, a Netflix Contest May Reward You Handsomely. New York Times (October 02, 2006), available at http://www.nytimes.com/2006/10/02/technology/02netflix.html
9. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Data Engineering, 2007. ICDE 2007. IEEE 23rd International Conference on. pp. 106 –115 (april 2007)
10. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: L-diversity: Privacy beyond k-anonymity. ACM Trans. Knowl. Discov. Data 1 (March 2007), http://doi.acm.org/10.1145/1217299.1217302
11. Narayanan, A., Shmatikov, V.: Robust De-anonymization of Large Sparse Datasets. In: Proceedings of the 2008 IEEE Symposium on Security and Privacy. pp. 111–125. IEEE Computer Society, Washington, DC, USA (2008), http://dl.acm.org/citation.cfm?id=1397759.1398064
12. Narayanan, A., Shmatikov, V.: Myths and fallacies of "personally identifiable information". Communications of the ACM 53, 24–26 (June 2010)
13. Samarati, P.: Protecting respondents' identities in microdata release. IEEE Trans. on Knowl. and Data Eng. 13, 1010–1027 (November 2001), http://dl.acm.org/citation.cfm?id=627337.628183
14. Schwarz, H.A.: ber ein Flchen kleinsten Flcheninhalts betreffendes Problem der Variationsrechnung. Acta Societatis scientiarum Fennicae XV: 318 (1888)
15. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. Int. J. Uncertainty, Fuzziness and Knowledge-Based System 10, 571–588 (October 2002), http://dl.acm.org/citation.cfm?id=774544.774553
16. Sweeney, L.: k-anonymity: a Model for Protecting Privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10, 557–570 (October 2002), http://dl.acm.org/citation.cfm?id=774544.774552
17. Xiao, X., Tao, Y.: M-invariance: towards privacy preserving re-publication of dynamic datasets. In: Proceedings of the 2007 ACM SIGMOD international conference on Management of data. pp. 689–700. SIGMOD '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1247480.1247556

Appendix

Theorem 1. *Let y denote the target record from given database D. Let aux_y denote (m, γ)-perturbed auxiliary information, uniformly sampled from the attributes in record y. Let $\epsilon > 0$. Then with probability $\geq 1 - g$, a record o can be found in the dataset such that the value of $S(y, o)$ is greater than $1 - 2\gamma - \epsilon$, where $g = e^{-2*\epsilon^2*m}$*

Proof. Let $x_i(y, r) = T(y(i), r(i))$ for any record r . Therefore, $S(y, r) = \sum_i \frac{x_i(y,r)}{k}$, where $k = |supp(y)|$.

Let $\mathbf{Y_1}$, $\mathbf{Y_2}$,..$\mathbf{Y_m}$ be m random variables which take a value equal to any of the x_j's and are chosen independently. $\mathbf{Z}$ is another random variable defined as $\mathbf{Z} = \frac{\sum_i \mathbf{Y_i}}{m}$ where $i \in \{1, ..., m\}$

We form the matching set M such that,

$$\begin{aligned} M &= \{r \in D : Score_g(aux_y, r) \geq 1-\gamma\} \\ &\quad \text{Using definition of } Score_g(aux_y, r) \\ M &= \{r \in D : min_{i \in supp(aux_y)} T(aux_y(i), r(i)) \geq 1-\gamma\} \\ M &= \{r \in D : \forall i \in supp(aux_y).T(aux_y(i), r(i)) \geq 1-\gamma\} \end{aligned}$$

Also, given $\forall i \in supp(aux_y).T(y(i), aux_y(i)) \geq 1-\gamma$, we calculate $T(y(i), r(i))$ for any record r in the matching set, and $\forall i \in supp(aux_y)$.

$$\begin{aligned} &T(y(i), r(i)) \triangleq 1 - \frac{|y(i) - r(i)|}{p} \\ &|y(i) - r(i)| \leq |y(i) - aux_y(i)| + |aux_y(i) - r(i)| \\ &|y(i) - r(i)| \leq 1 - (1 - p*\gamma) + 1 - (1 - p*\gamma) = 2*p*\gamma \end{aligned}$$

Thus, for any record r in the matching set

$$\forall i \in supp(aux_y).T(y(i), r(i)) \geq 1 - 2\gamma$$

Also, since $\mathbf{Y_i}$ has an uniform distribution

$$E[\mathbf{Y_i}] = x_1(y,r) * \frac{1}{k} + x_2(y,r) * \frac{1}{k} + \cdots + x_k(y,r) * \frac{1}{k} = S(y,r)$$

We show that expectation of $\mathbf{Z}$ is also $S(y,r)$

$$E[\mathbf{Z}] = \frac{\sum_i E[\mathbf{Y_i}]}{m} = \frac{m * E[\mathbf{Y_1}]}{m} = \frac{mS(y,r)}{m} = S(y,r)$$

One-sided Hoeffding bound states that given n independent random variables $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_n}$ where $Pr(\mathbf{X_i} \in [a_i, b_i]) = 1$ and $\bar{\mathbf{X}} = \frac{\mathbf{X_1} + \mathbf{X_2} + \cdots + \mathbf{X_n}}{n}$, the following inequality holds: $Pr[\bar{\mathbf{X}} - E[\bar{\mathbf{X}}] \geq \epsilon] \leq \exp\left(\frac{-2*\epsilon^2*n^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$. Using Hoeffding bound for $\mathbf{Z}$ with the observation that $\mathbf{Z}$ takes values in $[0,1]$ we get

$$Pr[\mathbf{Z} - E[\mathbf{Z}] \geq \epsilon] \leq \exp\left(\frac{-2*\epsilon^2*m^2}{\sum_{i=1}^{m}(1-0)^2}\right) = \exp(-2*\epsilon^2*m)$$

We can consider the complementary event and get

$$Pr[\mathbf{Z} - E[\mathbf{Z}] \leq \epsilon] \geq 1 - \exp(-2*\epsilon^2*m)$$

Let $g = e^{-2*\epsilon^2*m}$. Therefore, with probability $\geq 1-g$, z_i (realized value of $\mathbf{Z}$) $\leq E[\mathbf{Z}] + \epsilon$. Thus, with probability $\geq 1-g$, $E[\mathbf{Z}] \geq z_i - \epsilon$, and by substituting the value of $E[\mathbf{Z}]$ we get that with probability $\geq 1-g$, $S(y,r) \geq z_i - \epsilon$. Additionally, we have shown that for $r \in M$, $z_i \geq 1 - 2\gamma$ and hence for $r \in M$, $S(y,r) \geq (1 - 2\gamma - \epsilon)$. This implies that the record output by the generic algorithm described above, is guaranteed to have similarity greater than $1 - 2\gamma - \epsilon$ with the target record y, with probability $\geq 1-g$.