Robust De-anonymization of Large Datasets

Rui Fernandes (up202103071)
Department of Computer Science
Faculty of Sciences of the University of Porto

***Abstract*—Privacy preservation is a major concern when it comes to data analysis. When a dataset is released to third parties, privacy-preserving techniques are often necessary to reduce the possibility of identifying sensitive information about individuals. Usually, the data owner modifies the data in a way that the modified data can guarantee privacy while retaining sufficient utility. This process is usually referred to as privacy-preserving data publishing.**

In this report, I provide an overview of a statistical de-anonymization attack against high-dimensional micro-data and its application in de-anonymizing the Netflix Prize dataset.

***Keywords*—Data Anonymization, Privacy Preservation, Privacy-Preserving Data Publishing**

I. Introduction

A common reason for publishing anonymized micro-data[1] is *collaborative filtering*, *i.e.*, predicting a consumer's future choices based on his past behavior using the knowledge of what similar consumers did.

However, the privacy risks associated with publishing micro-data are well known. Even if personal identifiers have been removed, an adversary can use background knowledge and cross-correlation with other databases to re-identify individual data records. Famous attacks include the de-anonymization of a Massachusetts hospital discharge database by joining it with a public voter database, in which the combination of both was used to determine the values of medical attributes for each person who appears in both databases [1].

II. Problem

In October 2006, Netflix, the world's largest online movie rental service, announced the $1-million Netflix Prize, a machine learning and data mining competition for movie rating prediction [2]. To win the prize, a contestant would have to design a system that is more accurate than the company's recommendation system by at least 10%. To support the research, a dataset was made available consisting of about 100 million movie ratings for 17,770 movies given by 480,189 users. Each rating consists of four entries: user, movie, date of grade, grade. Users and movies are represented with integer IDs, while ratings range from 1 to 5.

This seemingly innocuous dataset actually has real privacy implications which will be addressed in the following sections.

[1]Micro-data are sets of records containing information on individuals.

III. De-anonymization of the Netflix Prize Dataset

Narayanan and Shmatikov [3] have shown that it is possible to de-anonimyze the Netflix Prize Dataset using the algorithm described in Section III-A.

In this dataset, there are no attributes that can be used directly for de-anonymization. Indeed, there are hundreds of records with the same value for a certain attribute. However, knowledge that a particular individual has a certain attribute value reveals some information since attribute values and even the mere fact that a given attribute is non-null vary from record to record.

Formally, we sample a record r randomly from a database D and give auxiliary information related to r to the adversary. Given this auxiliary information and an anonymized sample $\hat{D}$ of D, his goal is to reconstruct the attribute values of the entire record r.

A possible source of background knowledge is the Internet Movie Database – IMDb. Netflix subscribers that also use IMDb are expected to have a high correlation between their private Netflix ratings and their public IMDb ratings. In many cases, even a few movies that are rated by a subscriber in both services would be sufficient to identify his record in the Netflix Prize dataset with enough statistical confidence to rule out the possibility of a false match.

A. Algorithm

The algorithm use to de-anonymize the Netflix Prize dataset has three main components:

- A **scoring function** $Score$ which assigns a numerical score to each record in $\hat{D}$ based on how well it matches the adversary's auxiliary information aux. This function gives higher weights to statistically rare attributes, which captures the intuitive

notion that statistically rare attributes help de-anonymization much more than the knowledge of a common attribute;

- A **matching criterion** which is the algorithm applied by the adversary to the set of scores to determine if there's a match. The score of a candidate record is determined by the least similar attribute between itself and the adversary's auxiliary information. To improve the robustness of the algorithm, the matching criterion requires the top score to be significantly higher than the second-best score. This measures how much the first record "stands out" from other candidate records;
- A **record selection** which selects a "best-guess" record, or a probability distribution over the candidate records.

The algorithm, whose inputs are a sample $\hat{D}$ of a database D and auxiliary information $aux = Aux(r),\ r \leftarrow D$ and outputs either a record $r' \in \hat{D}$ or a set of records and a probability distribution over those records, works as follows:

1) Compute $Score(aux, r'),\ \forall\ r' \in \hat{D}$;
2) Apply the matching criterion to the resulting set of scores;
3) If a "best-guess" is required, output $r' \in \hat{D}$ with the highest score. If a probability distribution over the candidate records is required, compute and output a probability distribution based on the scores.

A more mathematical approach is provided by the authors in the original paper.

It is important to note that is algorithm may fail in two different scenarios: when an incorrect record is assigned the highest score and, on the other hand, when the correct record does not have a significantly higher score when compared to the second-highest score.

B. Results

Using this algorithm, it was possible to infer that very little auxiliary information is needed to de-anonymize an average subscriber record from the Netflix Prize dataset.

In fact, 99% of the Netflix subscribers were shown to be uniquely identifiable by a limited knowledge of no more than 8 movie ratings (2 of which may be wrong) with their corresponding rating dates. This emphasizes the relevance of background information for de-anonymization and re-identification.

Furthermore, a considerable privacy breach occurs even without any date, especially when the auxiliary information consists of movies that are not blockbusters, considering the fact that, as previously stated, statistically rare attributes help de-anonymization much more than the knowing a common attribute.

It is also important to note that *partial de-anonymization* may still pose a serious threat, considering the fact that there are many things the adversary might know about his target, such as the approximate number of movies rated, that can be used together with human inspection to complete the de-anonymization. In some cases, knowing the number of movies the target has rated, even if with a 50% error, can more than double the probability of complete de-anonymization.

Finally, even if it is hard to collect such information for a large number of subscribers, *targeted de-anonymization* still presents a serious threat to privacy.

IV. Critical Analysis

The utility of a data source lies in its ability to disclose data, and privacy aspects have the potential to hurt utility. Indeed, utility and privacy may be competing goals. The central question concerning privacy and utility of data is: "Can a higher level of privacy be achieved while maintaining utility?". Another important question is how to design micro-data sanitization algorithms that provide both privacy and utility. This is particularly relevant in the field of Privacy Preserving Data Mining, whose goal is to extract relevant knowledge from large amounts of data and provide accurate results while preventing sensitive information from disclosure.

V. Conclusions

It has been demonstrated that with very limited background knowledge, even if imprecise, it is possible to de-anonymize move viewing records released in the Netflix Prize dataset. It is also worth noting that anonymization operations such as generalization and suppression do not ensure privacy, and in any case fail on high-dimensional data[2]. For most records, simply knowing which columns are non-null reveal as much information as knowing the specific values of these columns.

References

[1] L. Sweeney, "Weaving Technology and Policy Together to Maintain Confidentiality," *Journal of Law, Medicine & Ethics*, vol. 25, no. 2-3, p. 98–110, 2021.

[2] "And if You Liked the Movie, a Netflix Contest May Reward You Handsomely - The New York Times," https://www.nytimes.com/2006/10/02/technology/02netflix.html, (Accessed on Oct 24 2021).

[3] A. Narayanan and V. Shmatikov, "Robust De-anonymization of Large Sparse Datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008, pp. 111–125.

[2]High dimensional data refers to a dataset in which the number of features p is larger than the number of observations N, often written as $p \gg N$.