Brussels Faculty of Engineering

Université Libre de Bruxelles

Vrije Universiteit van Brussel

Academic Year 2017-2018

Project Report

Data De-Anonymisation of the Netflix Prize dataset

Cédric Hannotier, Mathieu Petitjean, Hasan Can Yildirim

ELEC-Y591: Machine Learning and Big Data Processing

Contents

List of Figures

1 Introduction

Operators of social networks as well as companies are increasingly sharing information about their users. Would it be to support research or for commercial purposes, related data is typically protected by anonymization. Often, this "anonymization" is carried out by removing sensitive fields such as the name, address or Social Security Number of the user. Still, the scientific community has expressed doubt as to whether those methods guaranteed effective user privacy. Several successful attacks have been demonstrated, and this report aims to reproduce (with limitations such as reduced computing power capabilities) one of the most famous of those privacy breaches: the Netflix Prize dataset de-anonymization.

This report is structured as follows. First, in section 2, a state of the art of de-anonymization techniques is presented. It summarizes the major existing attacks, then focuses on the Netflix case. Secondly, in section 3, our approach is detailed and our choices are described. Eventually, in section 4, our results are shown, analyzed, and compared to the state of the art.

2 State of the art

2.1 De-anonymization attacks

One of the most mentioned de-anonymization deeds dates back to 2006, when New York Times journalists identified Thelma Arnold in the "anonymized" search queries released by AOL for research purposes [1]. By manually searching in the 20 millions search queries coming from 657,000 users, the reporters could tie Arnold's identity to some quite embarrassing queries.

The computer-aided attacks, being able to push such results to a much larger scale, can be separated in several categories depending on their approach. The two most represented methods are[2]:

- **Graph matching** is the most common approach in the case of social network de-anonymisation studies and is based on social graphs. One meaningful example is in [3], where Flicker and Twitter accounts were linked together with a 12% error rate. Several complex strategies can be used to improve graph matching, such as Seed & Grow [4] or Threading [5].
- **Similarity matching** is based on similar features between the target and auxiliary information. In [6], users were de-anonymised using the similarity between tweets and the content of their resume. In [7], victims of homicides were re-identified using "anonymous" homicides public records of Chicago and records in the Social Security Death Index.

2.2 The Netflix case

The attack that will be reproduced is the one presented in [8], an example of similarity matching. In this paper, researchers attacked a dataset released by Netflix in the context of a contest to improve their recommendation system. The 100 millions movie ratings by over 480,000 users were correlated to another movie rating database: the Internet Movie Database (IMDb). In a very small sample of the IMDb (50 users only), 2 users of the Netflix dataset were identified with statistical quasi-certainty. As the authors summarized, given a few of an user's reviews that he chose to make *public*, their algorithm is able to access all of his *private* Netflix ratings.

The algorithm is based on the similarity measure denoted Sim. It is defined, for two records r_1 and r_2, with $supp$ denoting the non-null attributes of a record:

$$Sim(r_1, r_2) = \frac{\sum Sim(r_{1i}, r_{2i})}{|supp(r_1) \cup supp(r_2)|} \tag{1}$$

The function Sim maps the records r_1 and r_2 to an interval $[0, 1]$, representing the notion of them being similar. This concept now needs to be adapted to the specific content of a movie review dataset. In particular, the scoring function needs to give higher weight to statistically rare attributes. Indeed, a review on "The Longest Most Meaningless Movie in the World[1]" helps identify a user much more than the knowledge of the fact that he liked the last episode of "Game of Thrones". The final scoring function that was used in [8] is:

$$Score(r, aux) = \sum_{i \in supp(aux)} \frac{1}{\log|supp(i)|} \left(e^{\frac{\rho_i - \rho'_i}{\rho_0}} + e^{\frac{d_i - d'_i}{d_0}} \right) \tag{2}$$

In the $Score$ function that compares a record r and auxiliary information aux, ρ and ρ' denote the score given to the same movie, while d and d' refer to the date of the rating. ρ_0 and d_0 are constants empirically determined to respectively 1.5 and 30.

Ultimately, two records are considered a match only if the difference between the best and second-best scores is higher than a threshold, referred to as the eccentricity ϕ. Its value was found experimentally to be 1.5 times the standard deviation. It is worth noting that the two matches with the 50 samples from IMDb had an eccentricity of respectively 28 and 15 ! These especially high numbers lead to the belief that two matches were found.

In [9], theorems that formally demonstrate why the scoring expressed by Equation 4 works on the Netflix dataset were introduced. In addition to providing a mathematical framework, they propose another scoring algorithm, slightly less performing but based on more general assumptions. It is based on another similarity measure that is now

[1]See its Wikipedia page: https://en.wikipedia.org/wiki/The_Longest_Most_Meaningless_Movie_in_the_World

asymmetrical, and that will be denoted $Sim2$:

$$Sim2(r_1, r_2) = \frac{1}{|supp(r_1)|} \sum_{i \in supp(r_1)} T(r_1, r_2) = \frac{1}{|supp(r_1)|} \sum_{i \in supp(r_1)} \left(1 - \frac{|r_1(i) - r_2(i)}{p_i}\right) \tag{3}$$

In Equation 3, p is the maximum difference value between the values of the i^{th} column. It is used to scale the values of $Sim2$ to $[0, 1]$. Then, the scoring is defined as:

$$Score(r, aux) = \sum_{i \in supp(aux)} \frac{1}{\log|supp(i)|} \frac{T(r, aux(i))}{m} \tag{4}$$

In both [8] and [9], the matching algorithm is the same (only the metrics that are used differ) and it is summarized in Algorithm 1.

Algorithm 1 Matching algorithm based on weighted scale scoring.

```
Starting from datasets R and aux

for each record r_i in R do
    for each entry aux_i in aux do
        Compute Score(r_i, aux_i)
    end for
    if eccentricity > φ then
        Match found !
    end if
end for
```

[Mention the difference between Sim and $Sim2$, cfr Netflix ++ paper, definition 1]

3 Approach

3.1 Scope of the work

In this work, it is intended to reproduce the method proposed by the original Netflix attack. Hence, Algorithm 1 has been implemented in Python. To evaluate the performance of the two different scoring metrics that have been introduced, both have been implemented and they will be compared later.

However, several differences with the original paper are to be highlighted. First, it used 50 records from the IMDb. However, the only datasets that are currently publicly available are the ratings for each movie, but not the ratings from a given user. A solution would be to get this data directly from the IMDb website because the ratings of a user are public if he also wrote a review. A data miner that parses the content of random IMDb users could do the job, but there are two obstacles:

- it would be of significant complexity, and is out of scope of the project.
- the terms and conditions of IMDb prohibit the usage of "data mining, robots, screen scraping, or similar data gathering and extraction tools"[2]. It can be suspected that it is the reason only 50 entries we used in the original attack.

As a workaround, we propose to use the MovieLens dataset as auxiliary information. MovieLens is a web-based movie recommender system that makes its database available for research. This database has already been used in privacy-related research, such as [10]. As opposed to the IMDb case, the "anonymous" user IDs are consistent across all the movie ratings that are registered, which makes it suitable for the user re-identification. Also, it is the occasion to test another dataset against the Netflix one.

Finally, due to the significant sizes of both the Netflix and MovieLens datasets (100 and 20 millions entries), it was necessary to subsample them. Results are expected to suffer from the reduced number of samples.

3.2 Data pre-processing

Here: data presentation (numbers + structure), necessary processing before actual similarity computation

4 Results

5 Conclusions

[2] see https://www.imdb.com/conditions

References

[1] The New York Times. *A Face Is Exposed for AOL Searcher No. 4417749*. [Online, last accessed 9 April 2018]. URL: https://www.nytimes.com/2006/08/09/technology/09aol.html.

[2] Dalal Al-Azizy et al. "A Literature Survey and Classifications on Data Deanonymisation". In: *Risks and Security of Internet and Systems*. Ed. by Costas Lambrinoudakis and Alban Gabillon. Cham: Springer International Publishing, 2016, pp. 36–51. ISBN: 978-3-319-31811-0.

[3] A. Narayanan and V. Shmatikov. "De-anonymizing Social Networks". In: *2009 30th IEEE Symposium on Security and Privacy*. 2009, pp. 173–187. DOI: 10.1109/SP.2009.22.

[4] A. Narayanan, E. Shi, and B. I. P. Rubinstein. "Link prediction by de-anonymization: How We Won the Kaggle Social Network Challenge". In: *The 2011 International Joint Conference on Neural Networks*. 2011, pp. 1825–1834. DOI: 10.1109/IJCNN.2011.6033446.

[5] Xuan Ding et al. *De-Anonymizing Dynamic Social Networks*. DOI: 10.1109/GLOCOM.2011.6133607.

[6] T. Okuno et al. "Content-Based De-anonymisation of Tweets". In: *2011 Seventh International Conference on Intelligent Information Hiding and Multimedia Signal Processing*. 2011, pp. 53–56. DOI: 10.1109/IIHMSP.2011.57.

[7] Salvador Ochoa et al. *Reidentification of Individuals in Chicago's Homicide Database: A Technical and Legal Study*. 2001.

[8] 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. DOI: 10.1109/SP.2008.33.

[9] Anupam Datta, Divya Sharma, and Arunesh Sinha. "Provable De-anonymization of Large Datasets with Sparse Dimensions". In: *Principles of Security and Trust*. Ed. by Pierpaolo Degano and Joshua D. Guttman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 229–248. ISBN: 978-3-642-28641-4.

[10] Dan Frankowski et al. "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*. SIGIR '06. Seattle, Washington, USA: ACM, 2006, pp. 565–572. ISBN: 1-59593-369-7. DOI: 10.1145/1148170.1148267. URL: http://doi.acm.org/10.1145/1148170.1148267.