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

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, the data 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 her 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. In each case, the attacker has access to a dataset that he is looking to de-anonymize, and *auxiliary information*. This additional information can be used is several ways, such as [2]:

- **Graph matching** is the most common approach in the case of social network de-anonymisation studies.
- ...

2.2 The Netflix case

The attack that will be reproduced is the one presented in [3]. 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 [3] 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 second threshold, referred to as the eccentricity ϕ. Again, 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 !

3 Approach

4 Results

5 Conclusions

[1] See its IMDb page: https://www.imdb.com/title/tt0342707/

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. "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.