Case Study #1 - Robust De-anonymization of Large Dataset

José Donato nº 2016225043

I. Introduction

This paper talks about statistical deanonymization attacks against datasets published by a certain company. In this paper's case the company is Netflix. The publishment of datasets by companies or governments is becoming a common strategy in order to support data mining research or open government laws.

II. Problems and why companies publish datasets

In 2006, Netflix announced a $1 million prize for improving their movie recommendation service releasing a dataset containing 100 million movie ratings from 500 thousand Netflix subscribers. In this situation, the sensitive data is the movies that a certain subscriber rates but the information can be far more critic. For example, it can contain health information or previous felonies from a certain individual. In 1997, crossing the medical dataset with another database containing the voter list from USA, the governor of Massachusetts was uniquely identified and they discovered his diagonis using their Zip, Birth date and Sex [1].
If there are problems with releasing datasets, why do companies still want to publish them? As I said before, data mining is one of the biggest reasons and it is a hot area nowadays. The companies want to understand the patterns of a certain user in order to make suggestions and make recommendations to them in the future. In the case of Netflix, they released the dataset in order to improve the algorithm of movie recommendations.

III. Important concepts

In order to understand the attacks that can be done against this datasets, there are some important concepts we need to understand:

- Quasi-identifier (QID): do not identify a person directly (like key attributes such as name, for example) but combined with other datasets can identify a person (in this case, the movie ratings combined with other datasets can identify the person)
- Sparsity: Average record has no similar records. If it is high, it means that a record has fewer similar records increasing the probability of de-anonymization.
- Background Knowledge/Auxiliary Information: extra information that the attacker has (in this case, the attacker can know the movies a subject likes or dislikes)
- Privacy Breach: finding the anonymized subject in the public sample or at least getting knowledge about some of his attributes
- De-anonymize: the process of identifying a subject in the public sample that was supposedly anonymous
- Re-identification: the outcome of a successful de-anonymization, i.e., when we identify a subject in a dataset

IV. Attack

The adversaries, i.e., the people that want to do a privacy breach and identify an individual subscriber in the dataset can cross their background knowledge with this public dataset and contradict what the Netflix promised in the prize FAQ: "Even if someone knows all movies ratings and dates, they could not identify them reliably in the dataset" [2]. Because this dataset is sparse, adversaries with very little background information about some Netflix subscribers can easily do a privacy breach on this dataset and de-anonymize and re-identicate them in the dataset. When saying little background information, the paper say that for 68% of the records only knowing two subscriber ratings and their dates (with a 3-day error) are sufficient to identify the target record. For 99% of the records, with 8 movie ratings (of which 2 can be completely wrong) and their respectively dates (with a 14-day error), the attacker will be successful in re-identifying the user.

This is a big problem because, knowing the subscriber patterns, even if he changes his virtual identity, the attacker can predict his future decisions when watching movies. Therefore, this situation does not have forward secrecy, i.e., once compromised, future situations are also compromised. As said in the paper it should be user deciding and not Netflix to choose whether to reveal their cinematographic interests publicly.

V. Question 1: Where did Netflix failed?

With the goal of making impossible to identify a target on the dataset, Netflix published only a small subset of the entire database (around 1/10). However, the selection algorithm to choose the 1/10 was not random. In addition to this, the level of noise (perturbation to the sample) was far too small to prevent the de-anonymization. In order to add perturbation, the paper referred two cases: one user had 1 out of 306 ratings altered and the other had 5 out of 229 altered. This level of perturbation is not enough to avoid a successful de-anonymization. However, there is a problem because if Netflix increases the noise to a level that would resist to such algorithm, the dataset utility would be destroyed. Other techniques must be researched in order to increase the anonymity of the dataset while keeping its utility. On the same paper, it is referred that in 2008 Netflix published another dataset that resulted from the application of an algorithm that brought together 108 different techniques to anonymize the dataset.

VI. Question 2: What can increase the probability of an attacker being successful in the Netflix dataset?

It is not always necessary to have a lot of background information. In fact, as referred before, having only two subscriber ratings and their dates, 68% of the records could be identified. However, the higher the background information the attacker has, the higher the probability he will succeed. Obtaining auxiliary information can be as easy as having a conversation with the subject about his cinematographic likes and dislikes, or search for it on his facebook page or the public ratings on IMDb. If that conversation is about films outside the top 100 of popularity, or even better outside the top 500, the results will be even better. The rarer the films are, the bigger is the probability of finding the target user in the sample dataset.

References

[1] LATANYA SWEENEy. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. https://epic.org/privacy/reidentification/Sweeney$_{A}rticle.pdf$.
[2] Arvind Narayanan and Vitaly Shmatikov. Robust De-anonymization of Large Sparse Datasets. https://www.cs.utexas.edu/ shmat/shmat$_{o}ak08netflix.pdf$.