Robust De-anonymization of Large Sparse Datasets

Arvind Narayanan

Vitaly Shmatikov

Presented by Ben Hesford,
Oct. 30, 2008

Presentation Overview

- Introduction
- Model for Privacy Breach
- De-anonymization Algorithm
- Case study: Netflix Prize dataset
- Conclusion
- References

Introduction

- Micro-data, information about specific individuals, are characterized by:

ID #	$Attr_1$	$Attr_2$	$Attr_3$	…	$Attr_M$
1					
2					
3					
…					
N					

1) High dimensionality: many attributes per record

2) Sparsity: for the average record, there exist no “similar” records

- In sparse datasets, little background knowledge is needed to de-anonymize a record

Introduction (cont'd)

- Approaches like k-anonymity fail on high-dimensional datasets
- Data perturbation techniques or other imprecise knowledge are ineffective in most cases

Model for Privacy Breach

Define a database ***D*** to be an ***N*** x ***M*** matrix

- Each row is a record associated with some individual, and the columns are attributes
- Each attribute (column) may be boolean, integer, or tuple

The set of *non-null* attributes of a record ***r*** is denoted **supp(r)**

Null attributes are denoted ⊥

A similarity measure **Sim** maps a pair of records to the interval [0, 1]

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

Model for Privacy Breach (cont'd)

- **Sparsity** defined, where **ε** = sparsity threshold and δ = sparsity probability

Definition 1 (Sparsity) *A database D is (ϵ, δ)-sparse w.r.t. the similarity measure Sim if*

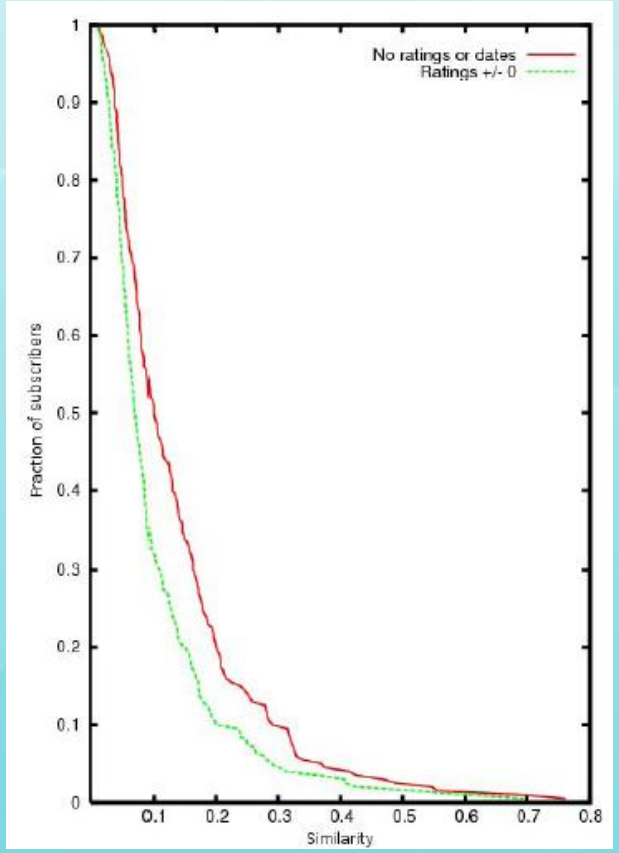
$$\Pr_r[Sim(r, r') > \epsilon \; \forall r' \neq r] \leq \delta$$

- Define ***Ď*** : arbitrary subset of dataset ***D***
  r: random record from ***D***
  Aux(r) : background knowledge about r

- Adversary model: Given ***Ď*** and ***Aux(r)***, the adversary should

  If $r \in Ď$, output r' s.t. $Pr(Sim(r,r') \geq \Theta) \geq \omega$
  If $r \notin Ď$, output $\perp$ with prob. at least ω

  where Θ = closeness of de-anonymized record,
  ω = probability that de-anonymization succeeds



Model for Privacy Breach (cont'd)

- In simpler terms, the adversary must:
  1) Output something with high probability if r is in the released sample that is similar to r
  2) Recognize when r is not in the released sample

- In the latter case, there is not enough auxiliary information to find a record similar to the target. The adversary produces a set ***D'*** of candidate records and a probability distribution Π such that

$$E[\ \min_{r' \in D',\ Sim(r,r') \geq \Theta} H_s(\Pi, r')\ ] \leq H$$

  - ***H_s*** **:** Shannon entropy measures how much information is needed to complete de-anonymization
  - ***H*** : De-anonymization entropy

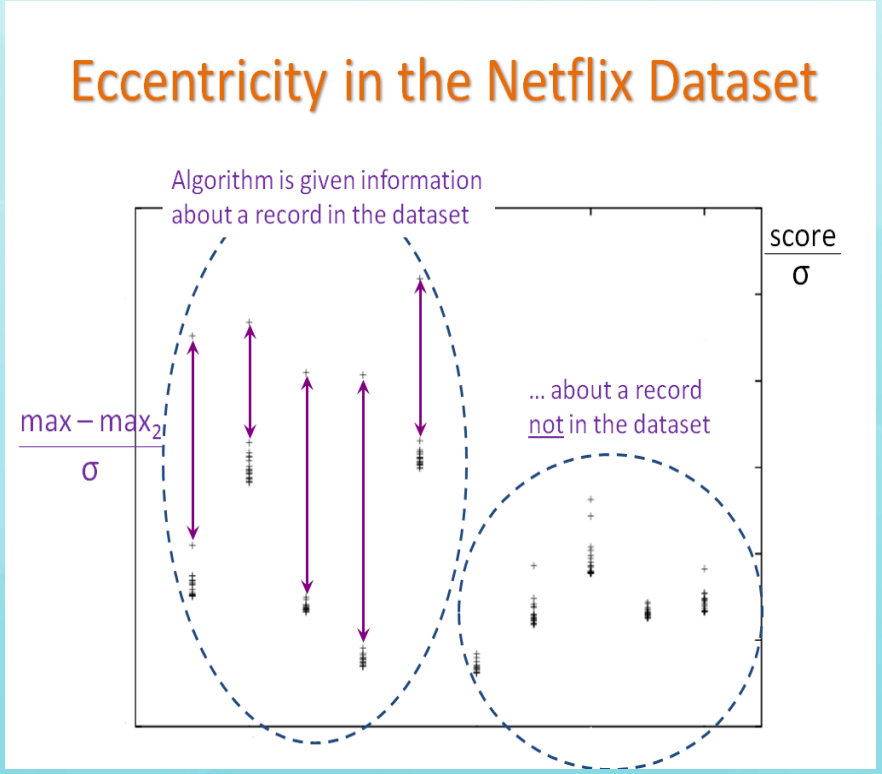
De-anonymization Algorithm

- Scoring function $\text{Score}(\text{aux}, r') = \min_{i \in \text{supp}(\text{aux})} \text{Sim}(\text{aux}_i, r'_i)$,
  - Determined by the least similar attribute among those known to the adversary as part of Aux

- Selection : Pick randomly from records with scores above threshold
  - Heuristic: $c \cdot e^{\frac{\text{Score}(\text{aux}, r')}{\sigma}}$ where c is a constant s.t. the distribution sums to 1
    - Gives higher weight to rare attributes & selects statistically unlikely high scores

“Eccentricity explained”

- Author Arvind Narayanan explains eccentricity on his blog (33bits.org). After finding the best match, he uses the notion of eccentricity to measure how much the matching record stands out from the second best matching record.
- Arvind explains “The trick is to look at the **difference between the best match and the second best match** as a multiple of the standard deviation … So, if the best match is 10 standard deviations away from the second best, it argues very strongly against the “null hypothesis,” which is that the match occured by fluke. Visually, the results of applying eccentricity are immediately apparent”



Netflix Prize Background

- Netflix released announced the $1 million Netflix prize for improving their movie recommendation service by 10%
- Data were collected from 1999 to 2005
  - Dataset is 10% of Netflix's 2005 database
  - Dataset contained 100+ million movie ratings created by ~500,000 users and when they were rated
  - Average user rated ~200 movies
- It is assumed the Netflix dataset is as sparse as the entire database

Netflix Case Study

- In the case of Netflix, the **Similarity** measure returns a 1 if two attribute values are within a threshold of each other.
- For movie ratings, on a 1-5 scale, thresholds of 0 and 1 are acceptable
- Rating dates were considered for 3 days, 14 days, and ∞ (no information about date of rating)
- A sample of ~ 50 IMDb users were compared with the Netflix dataset

De-anonymization Results

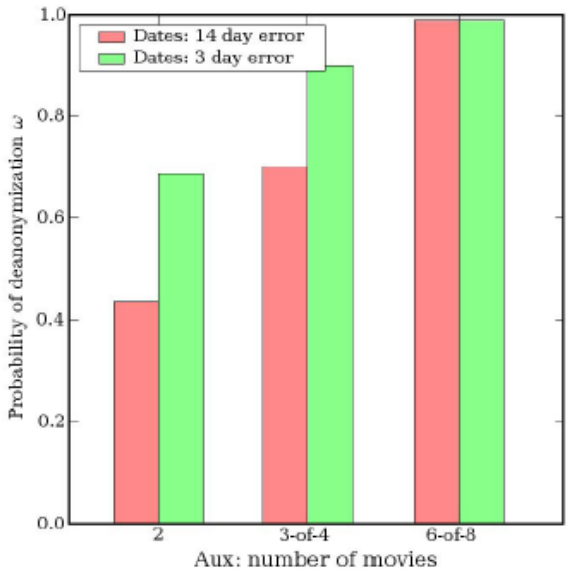


Figure 4. Adversary knows exact ratings and approximate dates.

- When 6 of 8 were known, the probability of re-identification is 99%

De-anonymization Results

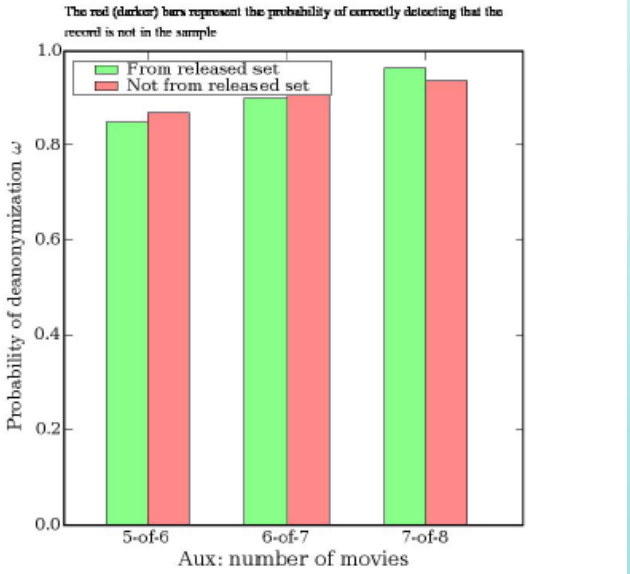
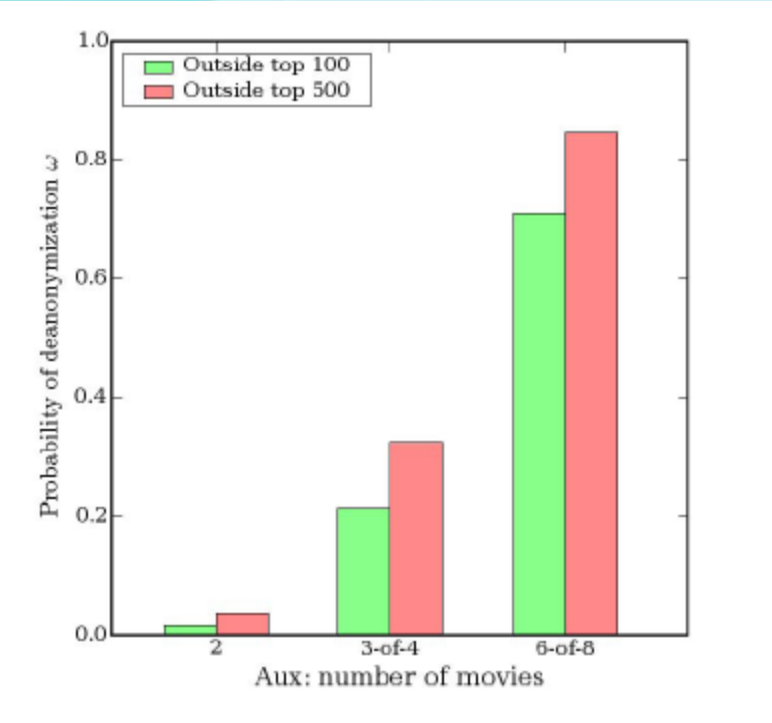


Figure 5. Same parameters as Fig. 4, but the adversary must also detect when the target record is not in the sample.

- The algorithm can recognize if a record has been released from set (deleted)

De-anonymization Results



- Excluding the top 100 & 500 films rated, with exact ratings known and no dates known, the probability of de-anonymization is still 84% if 6 out of 8 movies are known.

De-anonymization Results

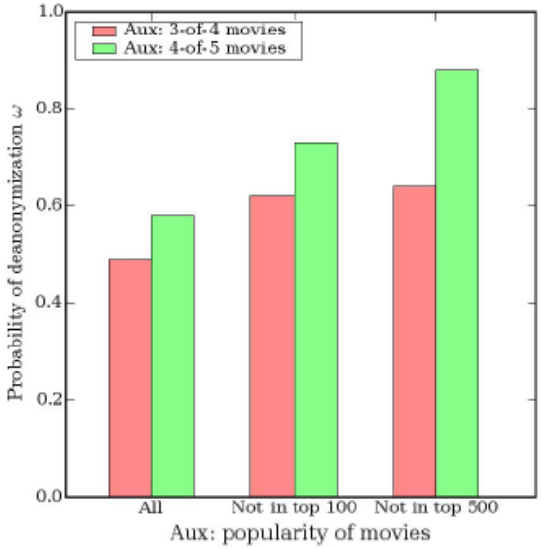


Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).

- Most users rate, and can be identified by, rare movies they rate

Netflix FAQ

- “Even if, for example, you knew all your own ratings and their dates you probably couldn’t identify them reliably in the data because only a small sample was included (less than one-tenth of our complete dataset) and that data was subject to perturbation. Of course, since you know all your own ratings that really isn’t a privacy problem is it?”

- Did no one at Netflix suspect a linkage attack?

Preventative Measures

- What could have Netflix have done to avoid this re-identification attack?
- Perturbing data significantly of individual attributes would have affected cross-attribute correlations & decrease data utility
- Perturbation may have caused a decrease in data utility while still allowing for a linkage attack

Questions

- How sensitive is your viewing data to you?
- In the paper, releasing the dataset without column identifiers (i.e., names of movies) is mentioned to protect privacy. Is this a good solution?
- Should new lines be drawn to guard against re-identification at the potential expense of data utility for researchers?

Conclusion

- Very little background data is needed to de-anonymize records.
- *k*-anonymity or data perturbation approaches do not succeed due to high dimensionality
- The average user had over 200 ratings, and only 4 ratings were needed to uniquely identify users (on average)

References

- http://33bits.org/2008/10/03/eccentricity-explained/ (Narayanan's blog)
- http://randomwalker.info/ppt/oakland.pptx (Talk given by authors)