Robust De-anonymization of Large Sparse Datasets

researched by Arvind Narayanan and Vitaly Shmatikov
2008 IEEE Symposium on Security and Privacy (S&P 2008)

presented by Haein Lee

What is de-anonymization?

- Public data are usually **anonymized** to protect privacy
- The anonymity could be **removed** by cross-referencing with other data sources

AOL search log release (‘06)

- AOL released a list of 20M web search queries for academic research

The New York Times

A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

ONLINE MEDIA DAILY

AOL Settles Data Valdez Lawsuit For $5 Million

by Wendy Davis @wendyndavis, February 19, 2013

The Netflix Prize (‘06 ~ ‘09)

- Netflix released movie rating dataset to improve movie recommendation system

NYC taxicab tip (‘14)

- New York City released taxi trip record dataset to public for data analysis
- https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Introduction

- A formal model for privacy breaches
  - A probability of successful de-anonymization
  - The amount of information that the attacker needs
- De-anonymization algorithm: Scoreboard-RH
  - Very little auxiliary information is needed for sparse datasets
  - Robust to the imprecision of data
- Practical analysis: The Netflix Prize dataset

Most data points are dissimilar to each other

Previous Work

- Frankowski, et al. “You Are What You Say: Privacy Risks of Public Mentions” (ACM SIGIR 2006)
  - Proposed algorithm is **not robust** to simple perturbations in data
  - Algorithm uses the **entire** public record of MovieLens dataset
  - Use **movie mentions in forums** via text-mining

De-anonymization Workflow

- Original Dataset (Name, Movie, Rating)

Publisher Perspective

Name	Movie 1	Movie 2	Movie 3	Movie 4	…
…					
Alice	-	5	1	?	
Bob	3	2	2	-	
Charlie	-	-	1	-	
…					

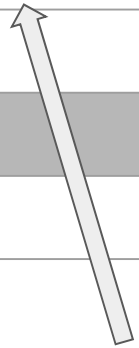
De-anonymization Workflow

- Released Dataset (Id, Movie, Rating)

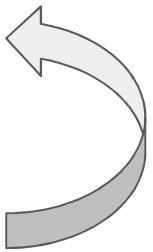
Publisher Perspective

Id	Movie 1	Movie 2	Movie 3	Movie 4	…
…					
10001	-	~~5~~ → 3	1	?	
10002	3	2	2	-	
~~10003~~	-	-	~~1~~	-	
…					

Use number instead of name



Change the value



Remove some record
(sampling)

De-anonymization Workflow

- Attacker's auxiliary information

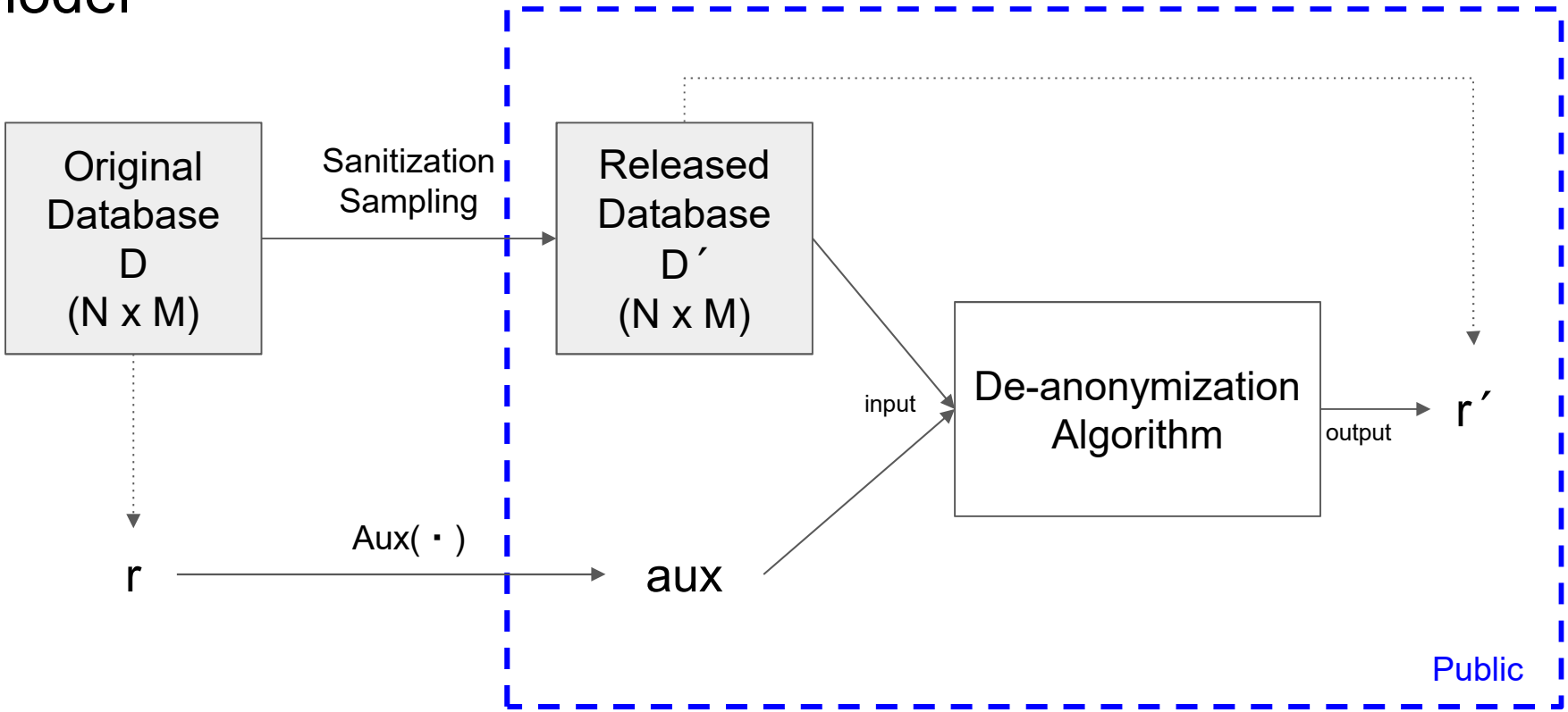
Attacker Perspective

Id	Movie 1	Movie 2	Movie 3	Movie 4	…
…					
10001	-	3	1	?	
10002	3	2	2	-	
…					

Attacker →	3	1	??	-	

Auxiliary Information

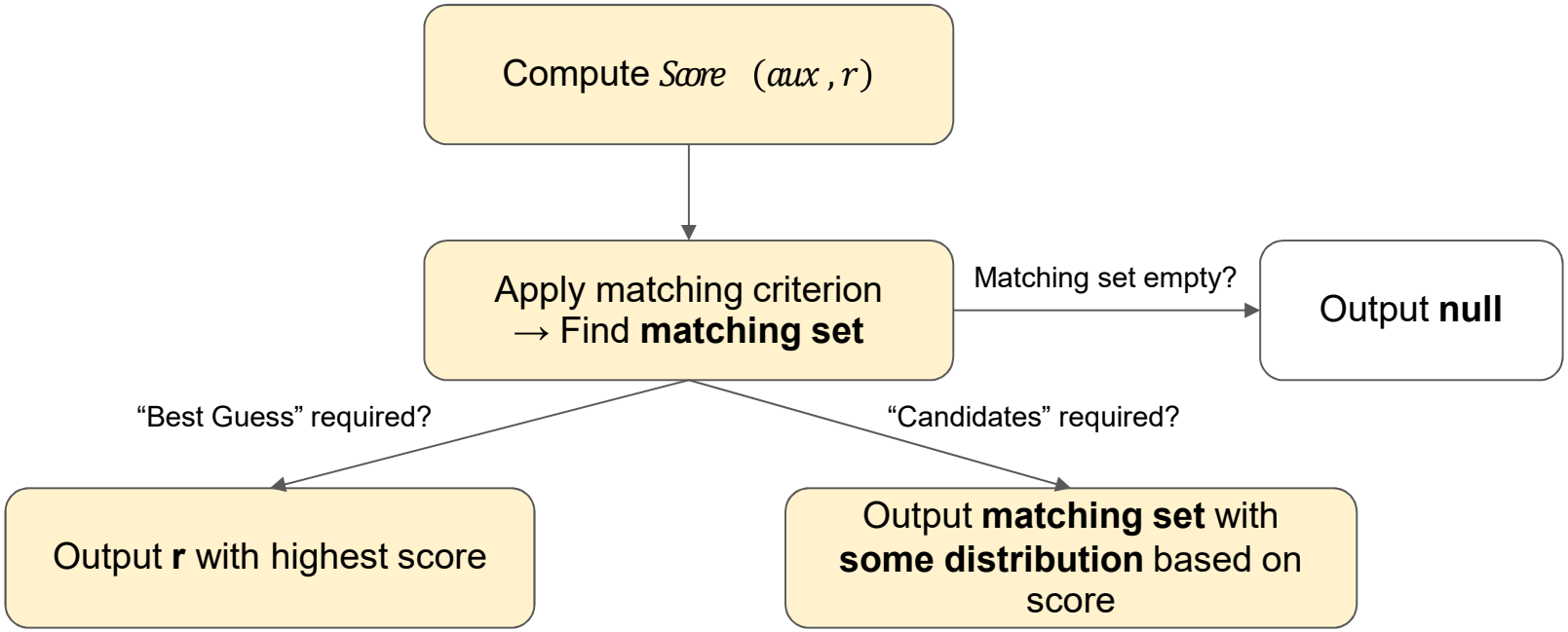
Model


Original
Database
D
(N x M)
Sanitization
Sampling
Released
Database
D´
(N x M)
input
De-anonymization
Algorithm
output
r´
Aux(·)
r
aux
Public

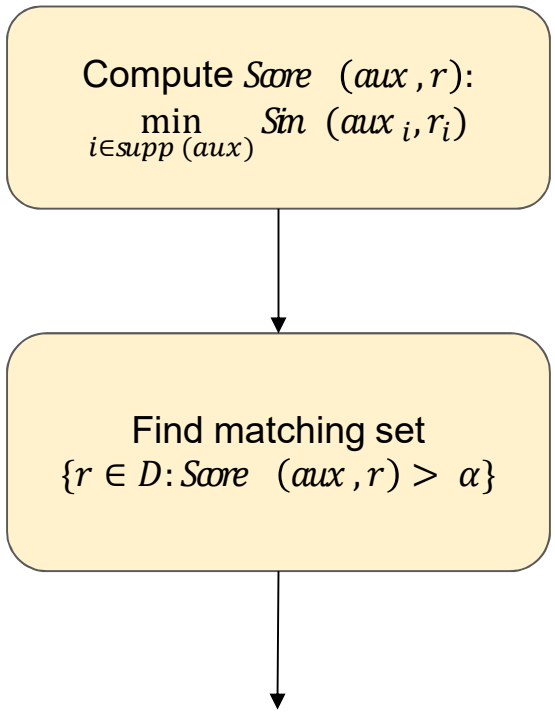
Privacy Breach Scenarios

- Scenario 1: Best Guess (1)
  - Automated large-scale de-anonymization
- Scenario 2: Candidates (>1) with probability distribution
  - Not enough auxiliary info / need additional analysis

De-anonymization Algorithm



Algorithm Scoreboard



r →

User	Movie 1	Movie 2	Movie 3	…
…				
10001	3	1	2	…
…				

aux →

4	-	4	…

$$Score\ (aux, r) = \min\{Sim\ (3,4), Sim\ (2,4)\}$$
$$= \mathrm{Sim}(3,4)$$

Algorithm Scoreboard-RH

- Heuristic 1. Rare attributes help de-anonymization
  - Weighted sum in scoring function
  - $Score\ (aux, r) = \sum_{i \in supp\ (aux)} \frac{1}{\log |supp\ (i)|} Sim\ (aux_i, r_i)$

	User	Movie 1 (Harry Potter)	Movie 2	Movie 3 *(Not famous)*	…
	…				
r →	10001	3	1	2	…
	…				
aux →		4	-	4	…

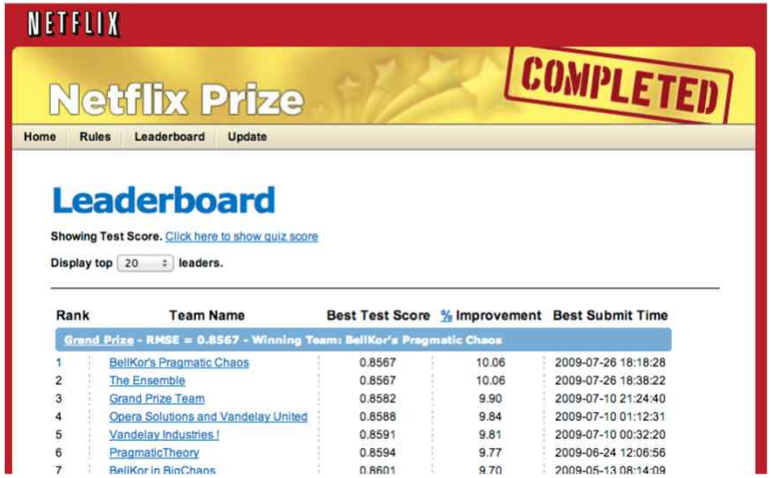
$$Score\ (aux, r) = \frac{1}{\log 10^6} Sim\ (3,4) + \frac{1}{\log 10^1} Sim\ (2,4)$$

Algorithm Scoreboard-RH

- Heuristic 2. Best guess should be “really good”
  - If a “best guess” is required, $\frac{\textbf{top score} - \textbf{second best score}}{\text{standard deviation of scores}} > \phi$
  - If a “candidates” is required, return the matching set with distribution which is **exponentially proportional to score**

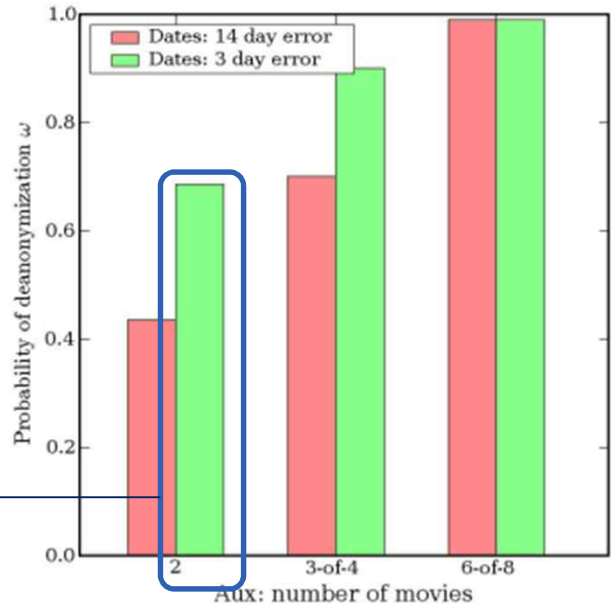
Netflix Dataset

- Ratings for 17,700 movies by 2,649,430 users
- Attributes: Movie ratings & dates



Evaluation: Best Guess

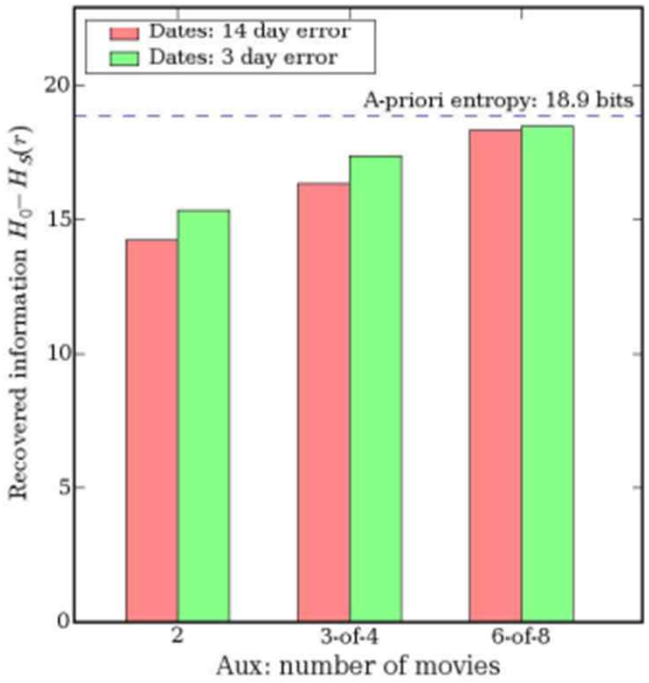
Only 2 exact movies as auxiliary information can de-anonymize with **68% probability**



Experiment settings

Ratings	Exact
Dates	±3 / ±14
Type	Best Guess
Aux selection	Uniform

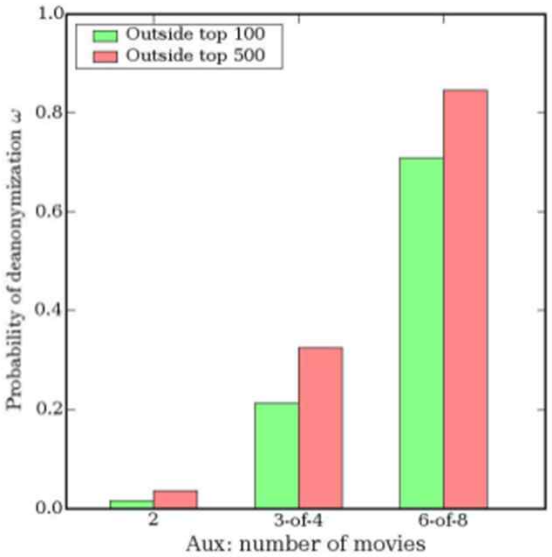
Evaluation: Candidates



Experiment settings

Ratings	Exact
Dates	±3 / ±14
Type	Candidates
Aux selection	Uniform

Evaluation: Rare movies



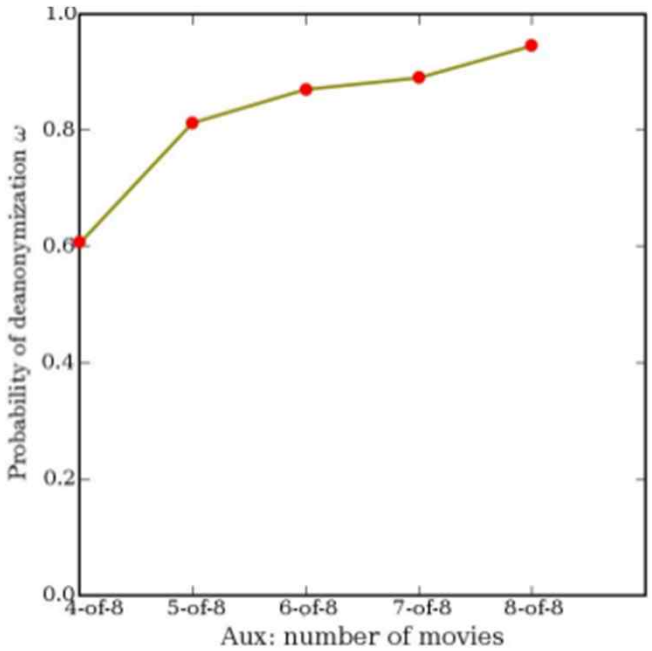
Not in X most rated	% of subscribers who rated ...		
	≥ 1 movie	≥ 5	≥ 10
$X = 100$	100%	97%	93%
$X = 500$	99%	90%	80%
$X = 1000$	97%	83%	70%

Many subscribers rate rare movies

Experiment settings

Ratings	Exact
Dates	No Info
Type	Best Guess
Aux selection	Not 100/500

Evaluation: Robustness



More error in Aux

Experiment settings

Ratings	±1
Dates	±14
Type	Best Guess
Aux selection	Uniform

Future Works

- Narayanan, et al. “De-anonymizing Social Networks” (IEEE S&P 2009)
  - Develop a re-identification algorithm based purely on the network topology targeting anonymized social network graphs
- Dwork, et al. “Differential Privacy” (ICALP 2006)
  - The removal or addition of a single database item does not (substantially) affect the outcome of any analysis
  - Now adopted in many real-world use cases
    - Research, US Census Bureau, Google, Apple, Microsoft, LinkedIn

Test of Time Award (2019)

- Narayanan and Shmatikov, “Robust de-anonymization of large sparse datasets: a decade later” (2019)
  - Reflects on the field of de-anonymization of the past decade
  - Discusses many impacts in the world after this work

Conclusion

- A robust de-anonymization methodology for sparse micro-dataset
  - Algorithm Scoreboard-RH
- Practical applicability in the Netflix Prize Dataset
  - 2 ~ 8 movies are sufficient for de-anonymization

Thank you!

Questions

Q. (Best Question) **Synthetic data** is algorithmically generated information with no relation to any actual case. Is it a suitable way to anonymize our dataset in this case?

A. Synthetic data generation is one of anonymization algorithm.
It depends on how you synthesize it.

PATE-GAN: Generating synthetic data with differential privacy guarantees (ICLR, ‘19)

Questions

Q. (Best Question) Have de-anonymization attacks on databases with **important consequences** (political, economic, ...) already been recorded?

A. Credit card metadata, geolocated data, medical patient data, …

Unique in the shopping mall: On the reidentifiability of credit card metadata (‘15)
De-anonymization attack on geolocated data (‘15)

Questions

Q. (Best Question) Can we keep the **efficiency of the algorithm** while **de-identifying** the data?

A. Trade-off between utility vs. privacy
- Privacy model : k-anonymity, l-diversity, differential privacy, …
- Utility measurement: accuracy, …

On the tradeoff between privacy and utility in data publishing ('09)
Optimization of privacy-utility trade-offs under informational self-determination ('20)

Questions

Q. It seems that the sure way is not to disclose the data set itself. Do you have any other **alternatives**?
Q. The countermeasures mentioned in paper are mentioned not irrelevant. Are there any other **countermeasures**?
Q. What are the current best-practice **anonymization algorithms** when publishing datasets?

A. Synthetic data generation, Data masking, Data encryption, …

Questions

Q. Will **vendors** stop releasing public datasets because of these privacy concerns? Or, will they keep releasing for their original purposes?
Q. As I understand it, **Netflix** publicly released the **Netflix** Prized database because they thought they had anonymized the micro-data in it. Thanks to this paper, which shows that it is possible, with a high probability, to de-anonymize the data, wouldn't **Netflix** be obliged to ask for the users' consent before? Or did they publish it because according to the rules of use that we accepted; it is possible for them?

A. There's a law called GDPR (General Data Protection Regulation) in EU
There's a committee in Korea

개인정보보호위원회

Thank you!