Robust De-anonymization of Large Sparse Datasets

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

Presented by Youngjin Jin (20204570)

NETFLIX (now)

- World’s most famous OTT streaming platform

(15 years ago)

- World’s largest online DVD rental service

Netflix Prize

- Competition for the *best* **user rating prediction algorithm**
  - Grand prize of US$1M!
- Includes a dataset from Netflix
  - Contains *anonymous* movie ratings of ~500,000 users

```
<user, movie, date of rating, rating>

<1023812, 104, 2004-11-24, 4>
<2125493, 132, 2005-07-25, 5>
< 723945, 375, 2005-09-11, 1>
```

… but is the data *really* anonymous?

- What if we can trace the record back to the user?
- Is it possible to ***de-anonymize*** datasets?

<user, movie, date of rating, rating>

Alice A. ← <1023812, 104, 2004-11-24, 4>

Bob B. ← <2125493, 132, 2005-07-25, 5>

Elon M. ← < 723945, 375, 2005-09-11, 1>

… and does it really matter?

- What can we infer from ratings data?
  - Viewing history
  - Non-public sensitive information

```
<user, movie, date of rating, rating>

Alice A.  <1023812, 104, 2004-11-24, 4>
          <1023812,  93, 2004-11-29, 2>
          <1023812, 442, 2004-12-01, 5>
```

Infer…

Introduction (Main Idea)

Data containing info on specific individuals

- Privacy implications on anonymized data
  - Formal model for privacy breaches in anonymized *microdata*
- De-anonymization algorithm
  - Very little background knowledge needed for sparse datasets
  - *Robust* to the imprecision of data
- Netflix Prize dataset analysis
  - Adversary can identify records of specific users

Most data points are dissimilar to each other

Related 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*

vs. this paper

- MovieLens dataset vs. Netflix Prize dataset
  - Utilizing internal datasets vs. solely using public data

Background

Collaborative Filtering: predict **future** choices from **past** behavior

Alice	
Movie	**Rating**
Lord of the Rings	4
Star Wars	5
Harry Potter	?

Bob	
Movie	**Rating**
Lord of the Rings	4
Star Wars	4
Harry Potter	**2**

Charlie	
Movie	**Rating**
Lord of the Rings	5
Star Wars	5
Harry Potter	**3**

David	
Movie	**Rating**
Lord of the Rings	4
Star Wars	5
Harry Potter	**3**

Emily	
Movie	**Rating**
Lord of the Rings	5
Star Wars	4
Harry Potter	**3**

Alice	
Movie	**Rating**
Lord of the Rings	4
Star Wars	5
Harry Potter	3

Background

Database

Attributes

Record of user 123456		
Movie ID	**Date of Rating**	**Rating**
13	2004-03-25	5
9824	2005-01-12	4
7913	2005-02-11	5
825	2003-12-09	1
10012	2002-09-30	3

Background

Similarity Measure - $\mathbf{Sim}(\boldsymbol{r_1}, \boldsymbol{r_2})$: how *similar* are two records r_1 and r_2?

r_1: Record of user 123456		
Movie ID	**Date of Rating**	**Rating**
13	2004-03-25	5
9824	2005-01-12	4
7913	**2005-02-11**	**5**
825	2003-12-09	1
10012	2002-09-30	3

r_2: Record of user 234567		
Movie ID	**Date of Rating**	**Rating**
1923	2003-07-13	3
1132	2004-04-09	3
7913	**2005-02-10**	**4**
993	2006-11-20	5
10104	2003-01-02	3

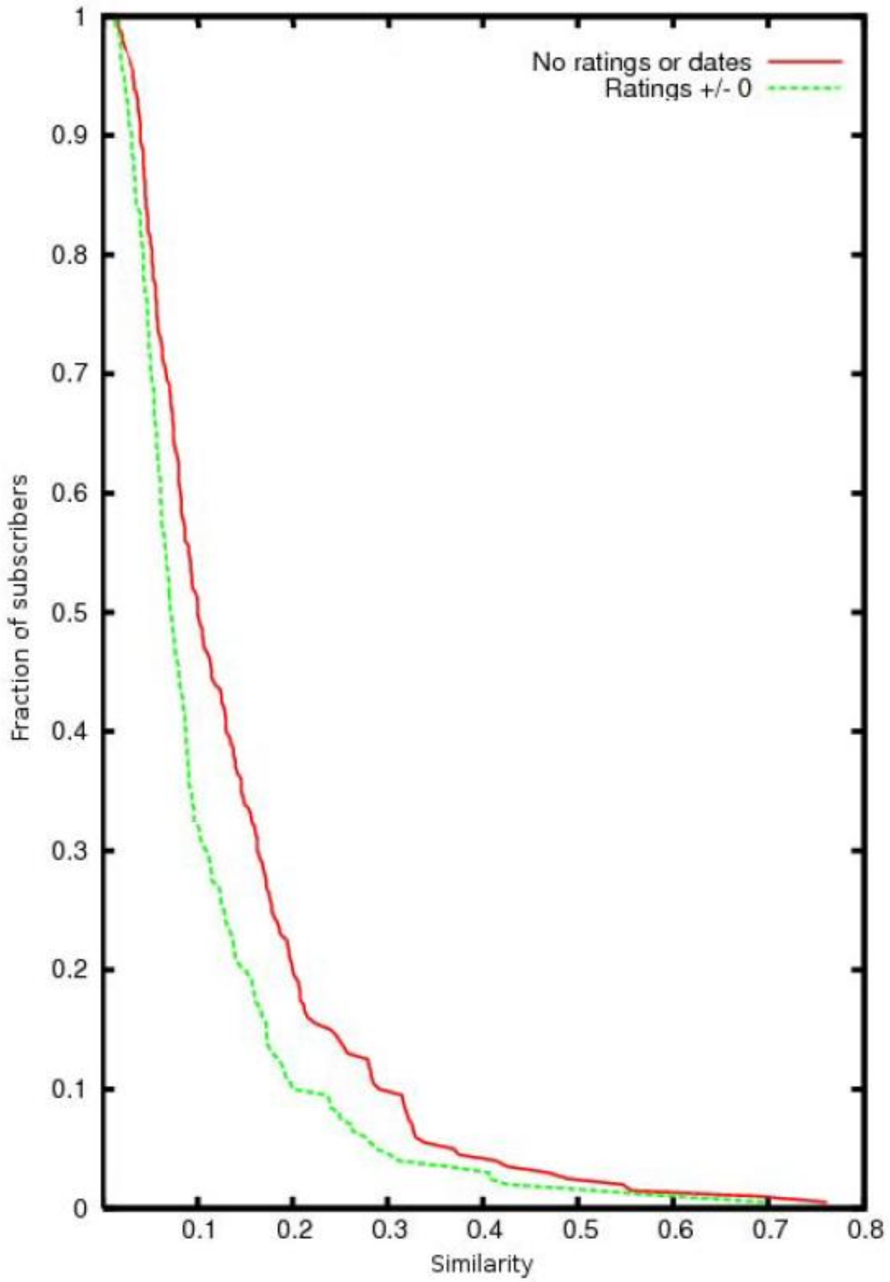
0: no similarity
1: complete similarity

$$\mathrm{Sim}(r_1, r_2) = 0.184$$

Background

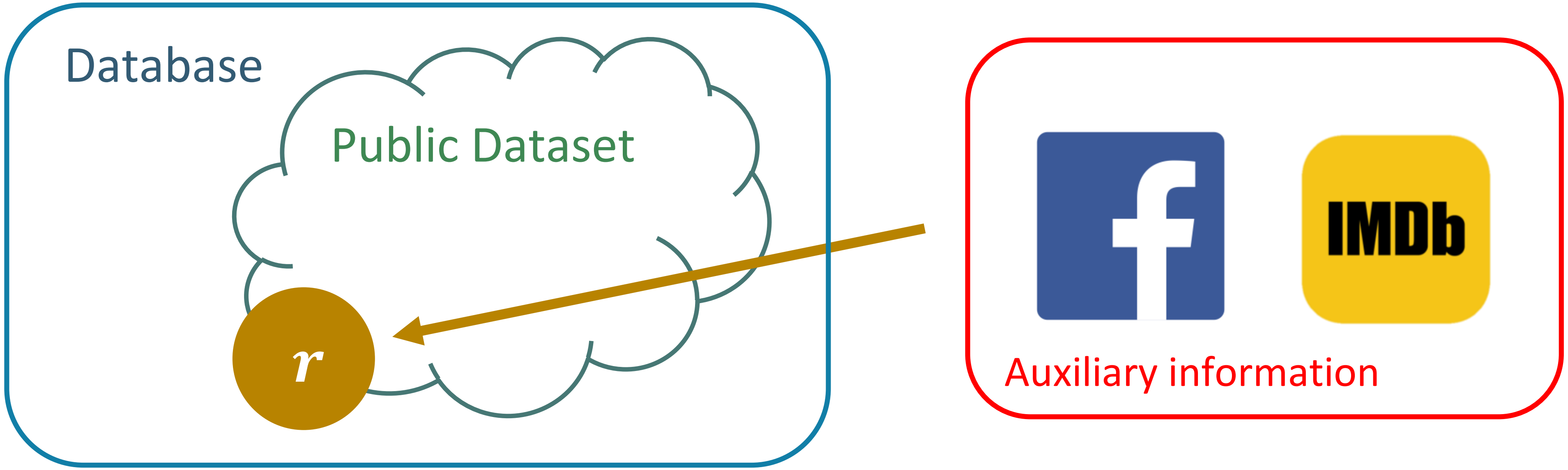
Sparsity: measure of how much few *similar* records there are

Low fraction of users with high *max similarity*
→ **Netflix Prize dataset is very *sparse***



Adversary Model

- Fix a target record r in the original dataset
- Given: *auxiliary information* (background knowledge)
- **Goal: learn as much about r's attributes as possible**



Adversary Model

- Goal: learn as much about r's attributes as possible

r: Record of user 123456		
Movie ID	**Date of Rating**	**Rating**
13	2004-03-25	5
9824	2005-01-12	4
7913	2005-02-11	5
825	2003-12-09	1
10012	2002-09-30	3

	r: Record of user 123456		
	Movie ID	**Date of Rating**	**Rating**
In the dataset (public)	**13**	**2004-03-25**	**5**
	9824	**2005-01-12**	**4**
Only in the DB (not public)	7913	2005-02-11	5
	825	2003-12-09	1
	10012	2002-09-30	3

Privacy Breach Scenarios

- Scenario 1: automated large-scale de-anonymization
  → Produce a single “best-guess” output
- Scenario 2: produce a “lineup” of candidate records
  - Not enough auxiliary info / need additional analysis

  → Produce a probability distribution over the candidate records

Challenges

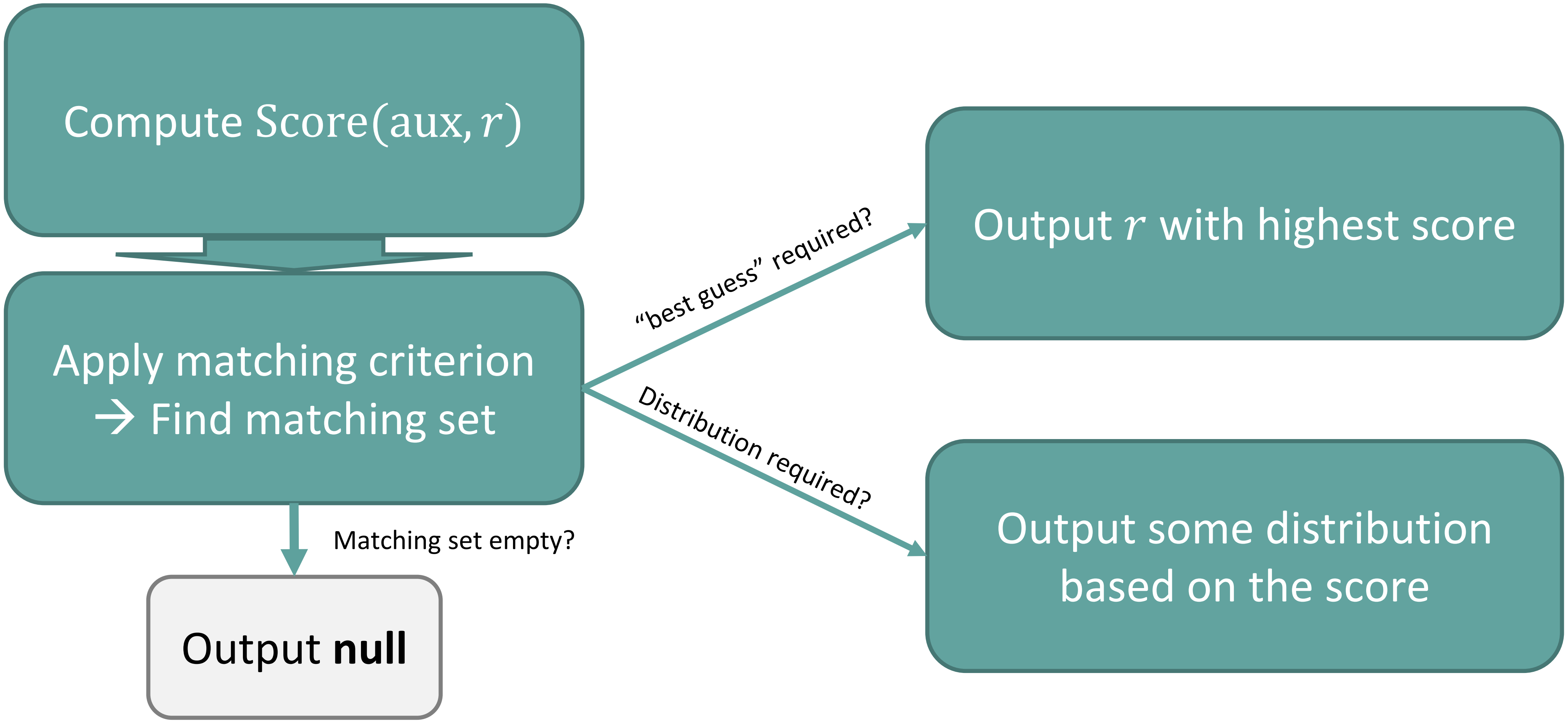
- Auxiliary information is noisy
- Released dataset may be altered / perturbed
- Only a fraction of record data available
  - Netflix Prize dataset → *less than 1/10* of the original database

De-anonymization Algorithm

Consists of three main components

- **Scoring function**: how well does the data match the auxiliary information?
- **Matching criterion**: determine match using scores
- **Record selection**: select one “best-guess” record

De-anonymization Algorithm



Algorithm: Scoreboard

Compute $\text{Score}(\text{aux}, r)$:
$$\min_{i \in \text{supp}(\text{aux})} \text{Sim}(\text{aux}_i, r_i)$$

Find matching set
$$\{r \in D : \text{Score}(\text{aux}, r) > \alpha\}$$

User	…	*Rating_n*	*Rating_{n+1}*	*Rating_{n+2}*	*Rating_{n+3}*
…		…	…	…	…
aux	…	4	-	**3**	5
r	…	3	5	**3**	2

$$\text{Score}(\text{aux}, r) = \min\{\text{Sim}(4, 3), \text{Sim}\,(3, 3), \text{Sim}(5, 2)\}$$
$$= \text{Sim}(5,2)$$

Limitations

- **Not sufficiently robust:** fails if *any of the attributes* in the auxiliary information are completely incorrect

Improvement: Scoreboard-RH

- **Add weight to score:** *higher weights* to statistically rare attributes

$$\text{Score}(\text{aux}, r) = \sum_{i \in \text{supp}(\text{aux})} \frac{1}{\log|\text{supp}(i)|} \text{Sim}(\text{aux}_i, r_i)$$

User	...	Rating$_n$ (Finding Nemo)	Rating$_{n+1}$ (Star Wars)	Rating$_{n+2}$ (A not-so famous film)	Rating$_{n+3}$ (Kill Bill)
...		...	...	...	...
aux	...	4	-	**3**	5
r	...	3	5	**3**	2

$$\text{Score}(\text{aux}, r) = \frac{1}{\log 10^6} \text{Sim}(4, 3) + \frac{1}{\log 30} \text{Sim}(3, 3) + \frac{1}{\log 10^5} \text{Sim}(5, 2)$$

Improvement: Scoreboard-RH

- **Best guess should be "really good":** *top guess* score should be significantly above the second-best guess
- Only output the best guess if…

$$\frac{\text{difference between the } \textbf{best} \text{ and } \textbf{second best} \text{ scores}}{\text{standard deviation of scores}} > \phi$$

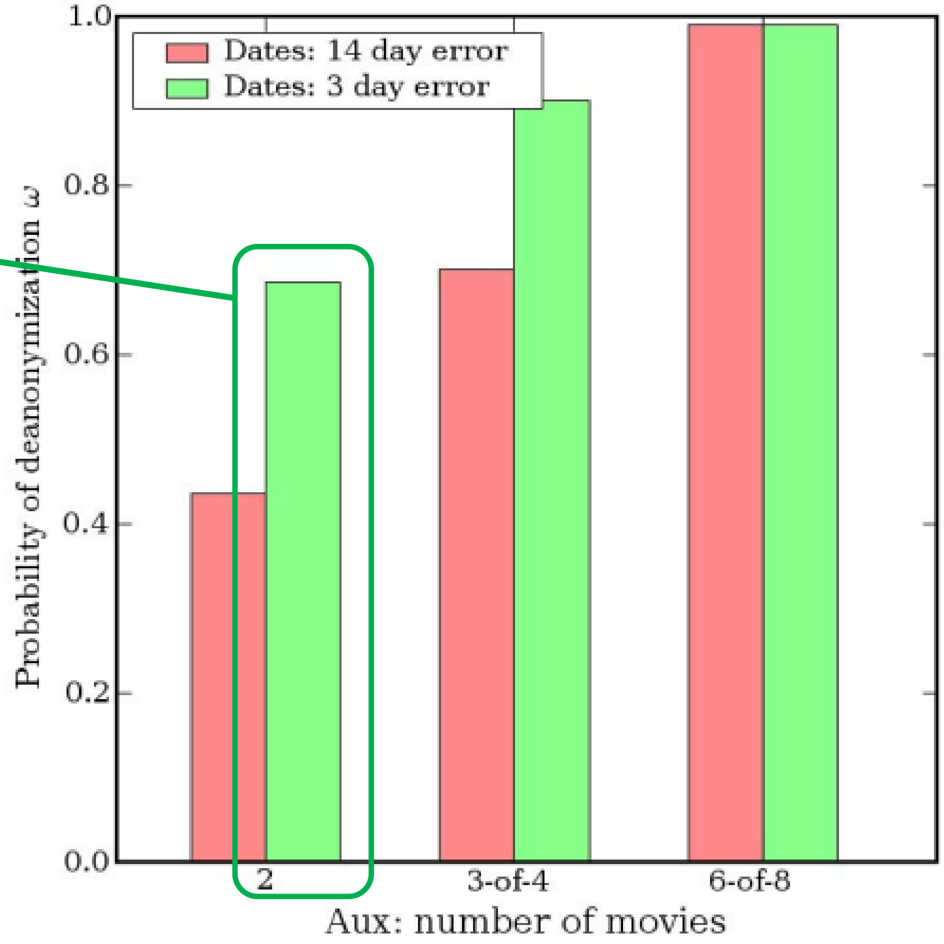
Netflix Prize Data Analysis

"[You shouldn't worry about privacy because] all customer identifying information has been removed…"

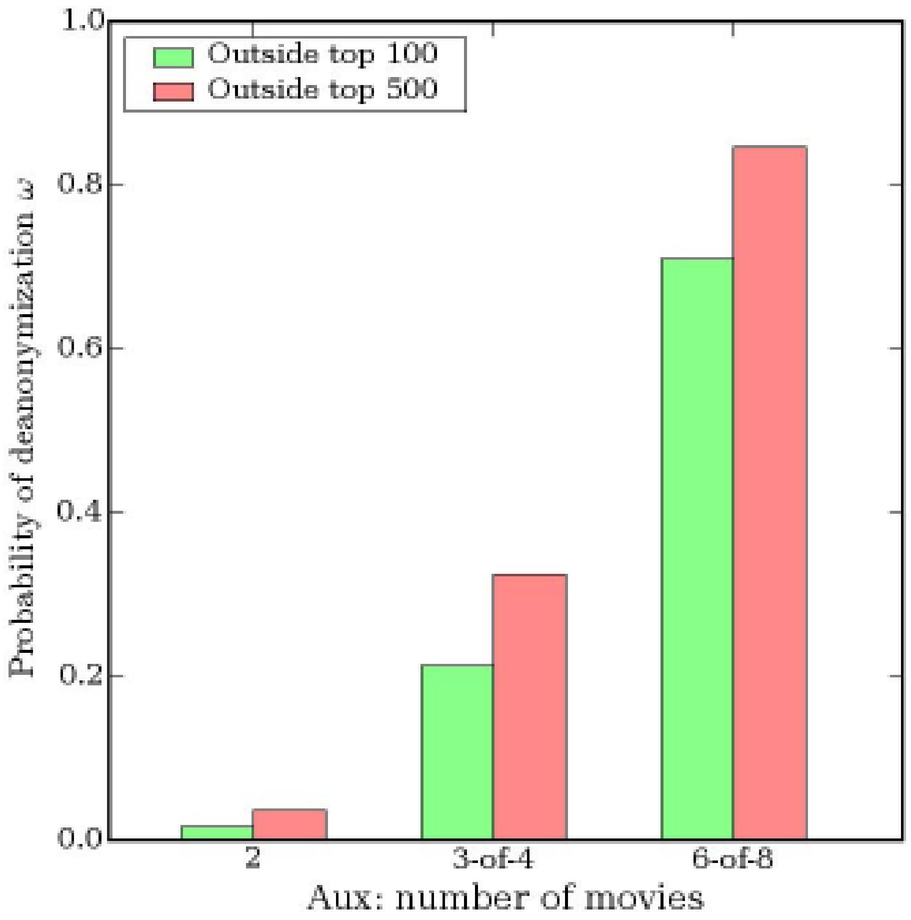
- Netflix, 2006

De-anonymization: Exact ratings & approximate dates

68% probability of deanonymization
Even only with 2 movies as auxiliary information



De-anonymization: Dates are unknown to the adversary



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

Mar 12, 2010, 12:35pm EST

Netflix Settles Privacy Lawsuit, Cancels Prize Sequel

Taylor Buley Contributor
Firewall Contributor Group
er, in all senses of the phrase

On Friday, N
a lawsuit related to it
challenged machine learning
better recommendations than the movi
themselves.

The New York Times

Netflix Cancels Contest After Concerns Are Raised About Privacy

By Steve Lohr
March 12, 2010

Netflix's $1 million contest for the best method to improve its movie recommendations was such a research and business hit that, when the winners were declared last September, th
immediately annou

, which included movie recommendations and choices made by hundreds of thousands of customers, could in fact be used to identify them.

NetFlix Cancels Recommendation Contest After Privacy Lawsuit

Netflix is canceling its second $1 million Netflix Prize to settle a legal challenge that it breached customer privacy as part of the first contest's race for a better movie-recommendation engine. Friday's announcement came five months after Netflix had announced a successor to its algorithm-improvement contest. The company at the time said it intended to [...]

Countermeasures

- Release the records without the column identifiers
  - Not a perfect countermeasure, but makes de-anonymization harder
- Interactive mechanisms for privacy-protecting data mining
  - Blum et al. “Practical Privacy: the SuLQ Framework” (2005)
    - Introduces noise to the query responses in a database with a goal of maintaining privacy
  - Dwork et al. “Calibrating Noises to Sensitivity in Private Data Analysis” (2006)
    - Perturbs true answer by the addition of random noise generated according to some distribution

Cou easures

Release the recor ithout the column identifiers

- N t e f t ou e m de-a ny ati arde

NOW IRRELEVANT!

- In a ti e cha s rot ct g d m ning
  - Blum et al. "Practical Privacy: the SuLQ Frame (2005)
    - Introduces noise to the query responses in a database with a g taining privacy
  - ork et al. "Calibrating Noises to Sensitivity in Private D lysi 06)
    - true answer by the addition of random noise generated according t tribution

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)
  - Closely related to the field of **Differential Privacy**
  - How to learn useful information about a population while learning nothing about individuals in the data?
  - Now adopted in many real-world use cases
    - Research, US Census Bureau, Google, Apple, Microsoft, LinkedIn

Test of Time Award (2019)

Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets" (2008)

- 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

- Very little background required for de-anonymization
  - Scoreboard-RH works under very general assumptions
  - Worked for the Netflix Prize data (with IMDb auxiliary data), which has less than 1/10 of the entire database
- Cannot be prevented through data perturbation
  - A new method of data privacy is necessary! → differential privacy
- Age of big data + “big brother” companies that possess everyone’s information
  - ***What’s the worst that can happen?***

Q&A

From 정현식 (Best Question #1): Is there any research on mathematical upper/lower limit of de-anonymization? For example, the lower and upper bound limits of time complexity for the sorting algorithm is well-known, so there's no further room for improvement.

A: Interesting question, but it’s a very difficult question to answer: we have to define what a “successful” de-anonymization is, and there are many factors that affect de-anonymization.

Q&A

From 이태현 (Best Question #2): I think the algorithm introduced in the paper is a statistical or linear algebraic version of the KNN-algorithm (best-guess part). Is there any good anonymization method for preventing a machine learning-based approach of de-anonymization?

A: There have been efforts in de-identification of data through deep learning methods, but data anonymization research has sort of declined after this work.

- Yadav, Shweta, et al. "Deep learning architecture for patient data de-identification in clinical records." *Proceedings of the clinical natural language processing workshop (ClinicalNLP)*. 2016.

- Shin, Hoo-Chang, et al. "Medical image synthesis for data augmentation and anonymization using generative adversarial networks." *International workshop on simulation and synthesis in medical imaging*. Springer, Cham, 2018.

Q&A

From 정현식 & 김성중: Are there studies that have applied this algorithm to other domains or data?

A: This algorithm hasn’t been extensively studied, but many de-anonymization efforts were made in the past decade and even well before this work.

(refer to the AOL search data leak in 2006: AOL search data leak - Wikipedia)

Q&A

From 이태화 & Tuan: In the paper, releasing the dataset without column identifiers (i.e., names of movies) is mentioned to protect privacy. Is this a good solution?

A: Nope :(

Thank You

Any questions?