Curs 09
Privacy Enhancing Techniques

Data sanitization

Confidentiality

1/27/2024

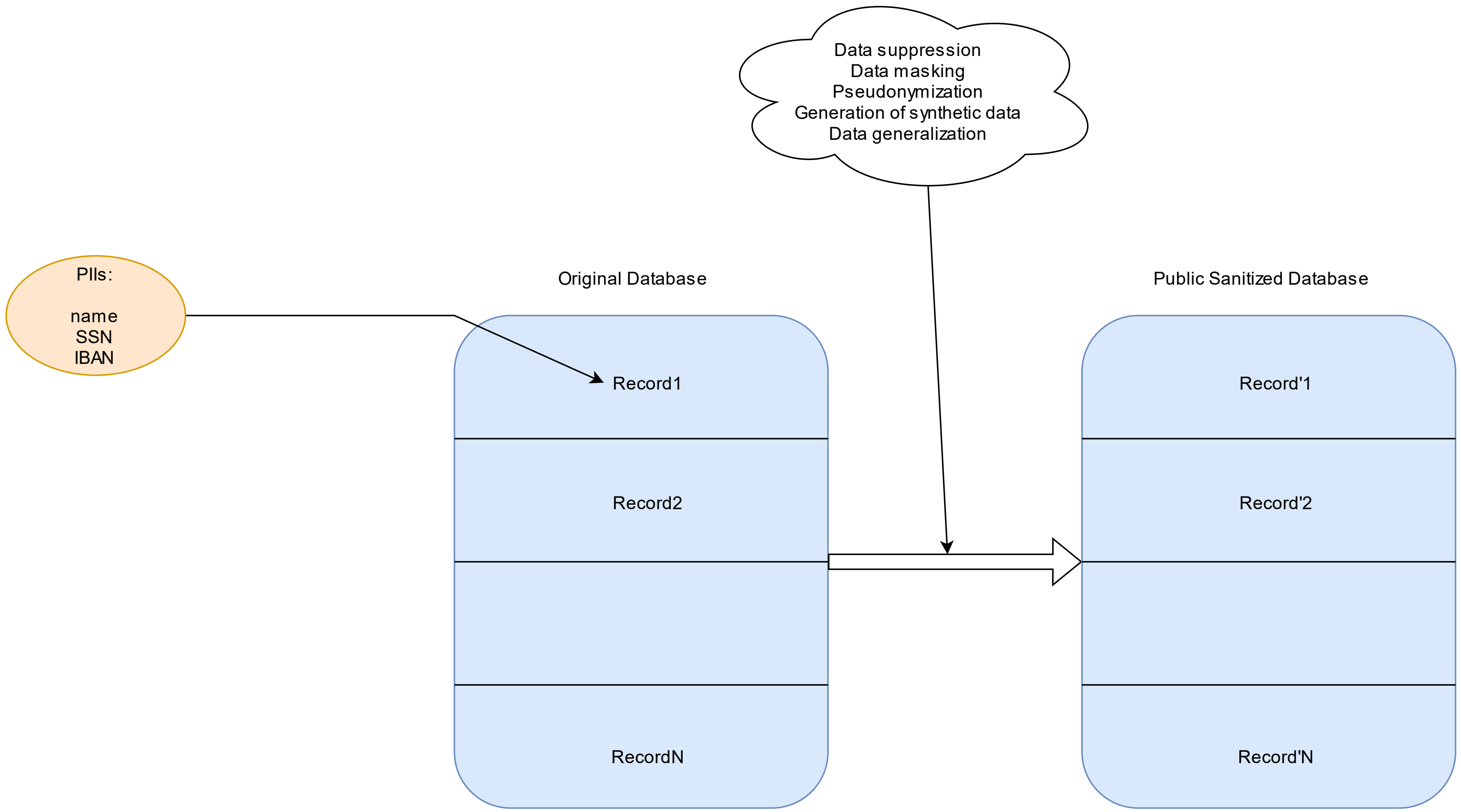
Course schedule

1. Why?
2. Cauzalitate
3. Măsurare
4. Modelare și eșantionare
5. Tehnici de analiză
   - Analiza factorială
   - Analiza cluster
   - Analiza de regresie
   - Analiza de rețea
   - Serii de timp
6. Predicție
7. Programare și ML
8. Why Privacy?
9. Privacy Enhancing Techniques
10. Differential Privacy
11. Homomorphic Encryption. PIR
12. Membership Inference Attacks
13. Federated Architecture. Multi-party computation
14. Explainable AI
15. Zero knowledge proof. Blockchain architecture

Data sanitization

Data suppression
Data masking
Pseudonymisation
Generation of synthetic data
Data generalization

What is data sanitization?



Data suppression

Data Suppression

- Strongest method of data anonymization
- Based on removing information from the dataset
- Two types:
  - Attribute Suppression
  - Cell/Record Suppression
- Still susceptible to re-identification attack and background knowledge attack

Attribute suppression

Student	Tutor	Test Score
John	Teddy	87
Stella	Teddy	56
Ming	Teddy	92
Poh	Song	83
Jake	Song	67
Yong	Song	45

Tutor	Test Score
Teddy	87
Teddy	56
Teddy	92
Song	83
Song	67
Song	45

Source: https://libguides.ntu.edu.sg/c.php?g=927336&p=6698844

Cell/Record Suppression

Student	Tutor	Test Score
John	Teddy	87
Stella	Teddy	56
Ming	Teddy	92
Poh	Song	83
Jake	Song	67
Yong	Song	45

Student	Tutor	Test Score
John	Teddy	87
Stella	Teddy	56
Ming	Teddy	92
Poh	Song	83
Jake	Song	67

Source: https://libguides.ntu.edu.sg/c.php?g=927336&p=6698844

Data masking

Data masking

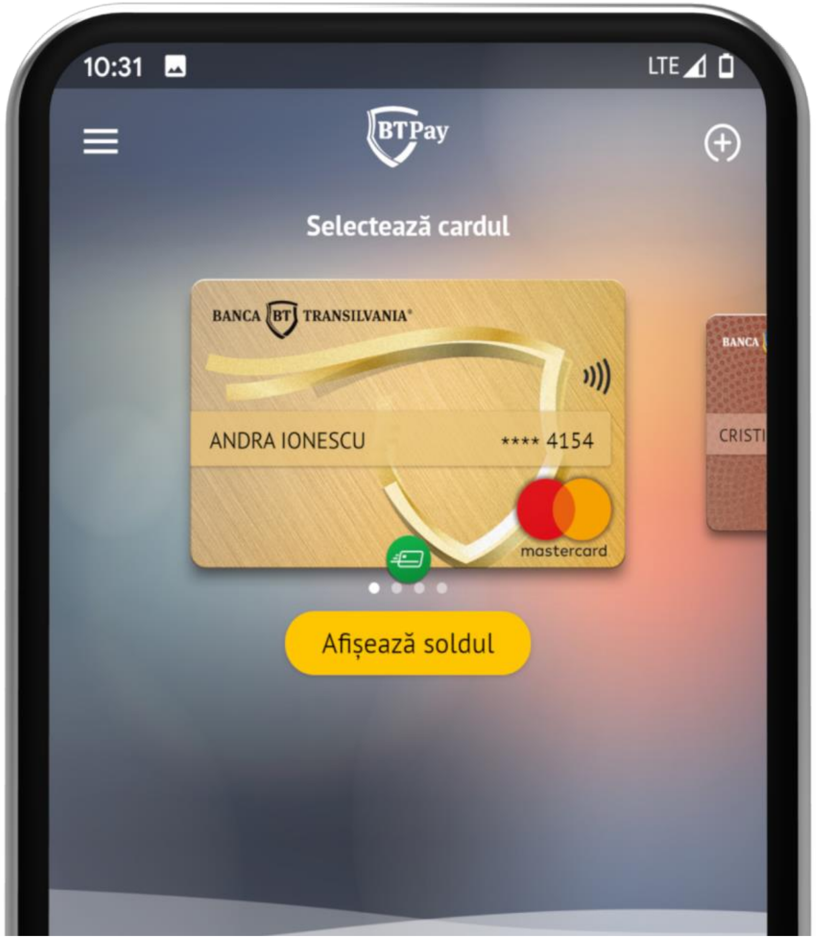
- Based on the replacing of sensitive data in the dataset
- 'X' or random generated characters are used "to mask" data
- Two types:
  - Partial Data masking
  - Fully Data masking
- Still susceptible to re-identification attack and background knowledge attack

Partial Data masking (1)

Account Number	Partial Masking
20085466123	20XXXXXX123
14875123654	14XXXXXX654
84569226644	84XXXXXX644

Source: https://www.sqlshack.com/understanding-dynamic-data-masking-in-sql-server/

Partial Data masking (2)



Source: https://www.bancatransilvania.ro/wallet-bt-pay/

Fully Data masking

last_name	first_name	ssn	gender	state
Smith	Bob	123-45-6789	M	CA
Doe	Jane	098-76-5432	F	PA
King	Stephen	888-67-5309	M	WI
Savage	Randal;	135-24-6789	M	FL
Downer	Debbie	918-55-4680	F	NC

→

last_name	first_name	ssn	gender	state
Smith	Bob	xxx-xx-xxxx	M	CA
Doe	Jane	xxx-xx-xxxx	F	PA
King	Stephen	xxx-xx-xxxx	M	WI
Savage	Randy	xxx-xx-xxxx	M	FL
Downer	Debbie	xxx-xx-xxxx	F	NC

Source: https://www.tibco.com/reference-center/what-is-data-masking

Pseudonymization

Pseudonymization

- Based on the replacement of sensitive data with made up values
- Also known as "coding"
- Can be irreversible or reversible
- Used when data values in the dataset needs to be uniquely identified
- Similar to how some UPB classbooks are made to respect GDPR

Pseudonymization

Person	Pre-Assessment Result	Hours of Lessons Taken
John Rohit	B	25
Stella Campbell	D	26
Ming Siew Lee	A	30
Poh Boon	B	32
Siva Vasanth	C	29
Siti Raudhah	A	25

Person	Pre-Assessment Result	Hours of Lessons Taken
4135891	B	25
3229873	D	26
4398642	A	30
783127	B	32
583419	C	29
983429	A	25

Source: https://libguides.ntu.edu.sg/c.php?g=927336&p=6698844

Reversibility?

- Possible only if we securely keep an identity database

Pseudonym	Person
4135891	John Rohit
3229873	Stella Campbell
4398642	Ming Siew Lee
783127	Poh Boon
583419	Siva Vasanth
983429	Siti Raudhah

Source: https://libguides.ntu.edu.sg/c.php?g=927336&p=6698844

Generation of synthetic data

Generation of synthetic data

- Synthetic data = '"fake data", artificially or programmatically generated

- Why?
  - Retains the underlying structure and statistical distribution of the original data
  - Does not rely on masking or omitting of the original data
  - Provides a strong privacy guarantee to prevent sensitive user information from being disclosed

Generation of synthetic data (2)

Name	Age	Gender	SIN	Chest pain location
Rylie Bradford	72	M	100 709 112	0
Karyn Polley	54	F	722 260 965	1
Gordie Quincy	53	M	795 635 739	1

Name	Age	Gender	SIN	Chest pain location
Simone Peacock	75	F	970 440 905	1
Allyson Wortham	69	M	748 665 544	1
Cyprian Traylor	46	M	265 183 491	0

Source: https://towardsdatascience.com/synthetic-data-applications-in-data-privacy-and-machine-learning-1078bb5dc1a7

Data generalization

Data generalization

- Technique that allows to replace sensitive data values with less accurate ones

- The utility of the data for analysis should be kept

- Works on both categorical and ordinal data

- Performs better when used alongside suppression and masking

Data generalization (2)

	ORIGINAL DATA	GENERALIZED DATA
AGES	16 18 21 23 27 32 32 36 38 39 44 47 47 48 49	10-19 (2) 20-29 (3) 30-39 (5) 40-49 (5)

Source: https://satoricyber.com/data-masking/data-generalization/

Data generalization (3)

- Before generalization:

  {(1, Radiologist), (2, Internist), (3, Dancer), (4, Singer)}

- After generalization:

  {(1, Physician), (2, Physician), (3, Artist), (4, Artist)}

Data aggregation

Data aggregation

Donor	Monthly income ($)	Amount donated in 2018 ($)
Donor 1	4000	200
Donor 2	4900	400
Donor 3	2200	150
Donor 4	4200	100
Donor 5	5500	250
Donor 6	2600	50
Donor 7	3300	100
Donor 8	5500	200
Donor 9	1600	50
Donor 10	3200	50

Monthly income ($)	No. of donations received (2018)	Sum of amount donated in 2018 ($)
1000-1999	1	50
2000-2999	2	200
3000-3999	2	150
4000-4999	3	700
5000-5999	2	450
Total	**10**	**1550**

Source: https://satoricyber.com/data-masking/data-generalization/

Confidentiality

K-Anonymity

L-Diversity

K-Anonymity

K-Anonymity

- Model of data confidentiality
  - Constraint: at least K individuals in the dataset share the set of attributes that can become identifying for each individual
- Based on "hide in the crowd" idea
- Use data masking or data generalization
- It is done only on quasi-identifiers

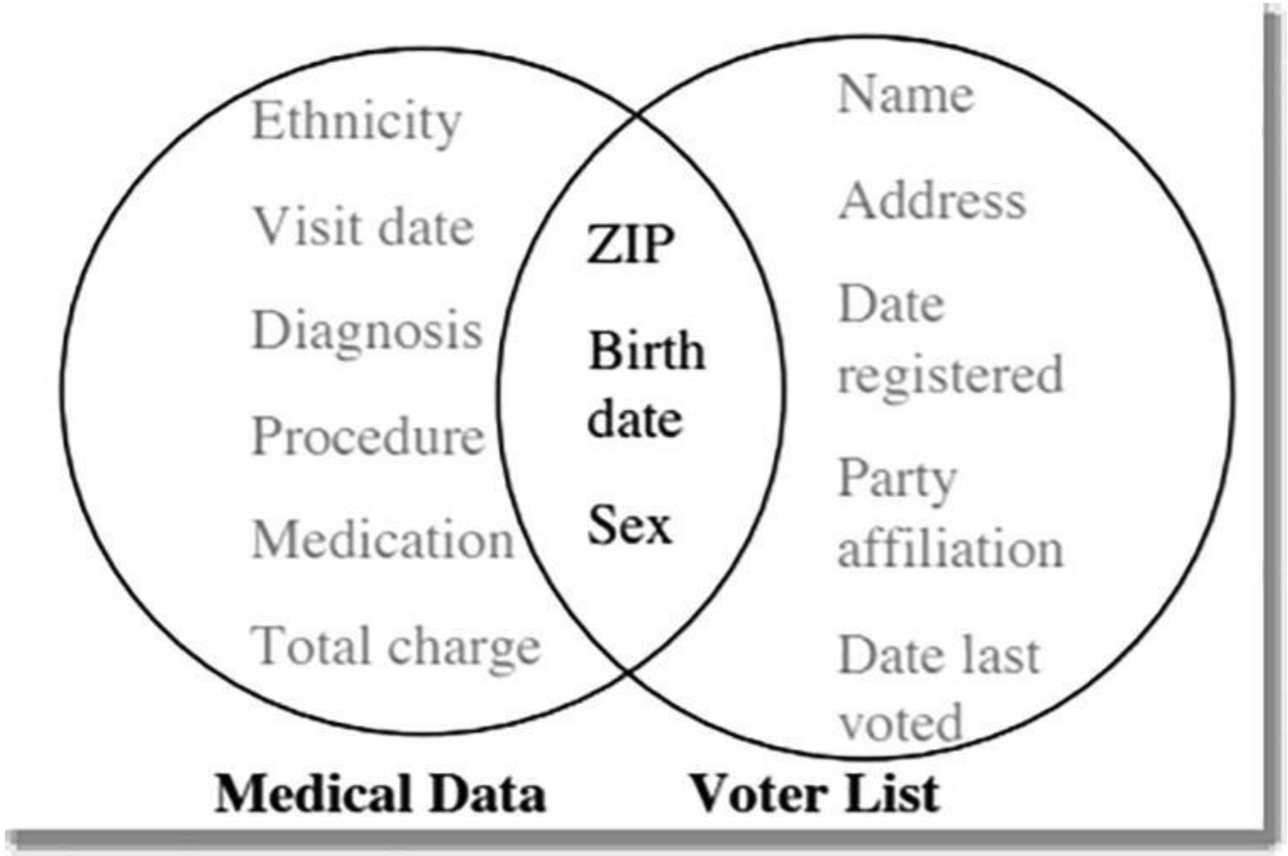
K-Anonymity (2.1)

- Why is relevant?

Identifying the governor of Massachusetts

Source: https://campus.datacamp.com/courses/data-privacy-and-anonymization-in-python/more-on-privacy-preserving-techniques?ex=7

K-Anonymity (2.2)



Source: https://ars.els-cdn.com/content/image/1-s2.0-S1319157821001002-gr1_lrg.jpg

K-Anonymity (3)

ID	Age	Zip	Disease
1	7	53715	Flu
2	9	55410	Diarrhea
3	13	52121	Flu
4	19	56421	Fever
5	29	02263	Diarrhea
6	34	02296	Fever
7	39	02278	Flue
8	33	02254	Diarrhea

Sensitive Table

ID	Age	Zip	Disease
1	0-20	5****	Flu
2	0-20	5****	Diarrhea
3	0-20	5****	Flu
4	0-20	5****	Fever
5	20-40	022**	Diarrhea
6	20-40	022**	Fever
7	20-40	022**	Flue
8	20-40	022**	Diarrhea

2 – Anonymized Table

Source: https://thamindur.medium.com/k-anonymity-privacy-preservation-in-data-mining-8d5b5ad19d45

Limitations

- Works on datasets with low dimensional data
- Finding the optimal value for K is NP-hard
- Still susceptible to homogeneity attack and background knowledge attack

Attacks on k-anonymized data

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

Source: https://www.chegg.com/homework-help/questions-and-answers/question-table-disease-sensitive-value-edited-table-order-enhance-protection-homogeneity-a-q50377779

L-Diversity

L-Diversity

- Reduces the risk of attacks on k-anonymized data

- Constraint:
  - At least L distinct values for the sensitive data in each susbset generated after applying k-anonymity

L-Diversity (2)

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Source: https://elf11.github.io/2017/04/22/kanonymity.html

Case study – Narayanan et. al, 2008

Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)

Arvind Narayanan and Vitaly Shmatikov

The University of Texas at Austin

February 5, 2008

Abstract

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as individual preferences, recommendations, transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes in the adversary's background knowledge.

We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie ratings of 500,000 subscribers of Netflix, the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Does the privacy of this movie ratings matter?

YES!

Basic Premises

- Two databases:
  - Anonymized Netflix database
    - Records containing movie ratings created by ~500 thousand users
  - Public IMDb database
    - Small Db with records containing movie ratings
    - Used as side information
    - Noisy data

Basic Premises (2)

- Instead of using a second database:
  - Background knowledge consisting of what are the people preferences in terms of movies
- Propose a statistics model based on:
  - Can similarity between two different ratings in the two databases imply that they belong to the same person?

Results – De-anonymization with auxiliary info

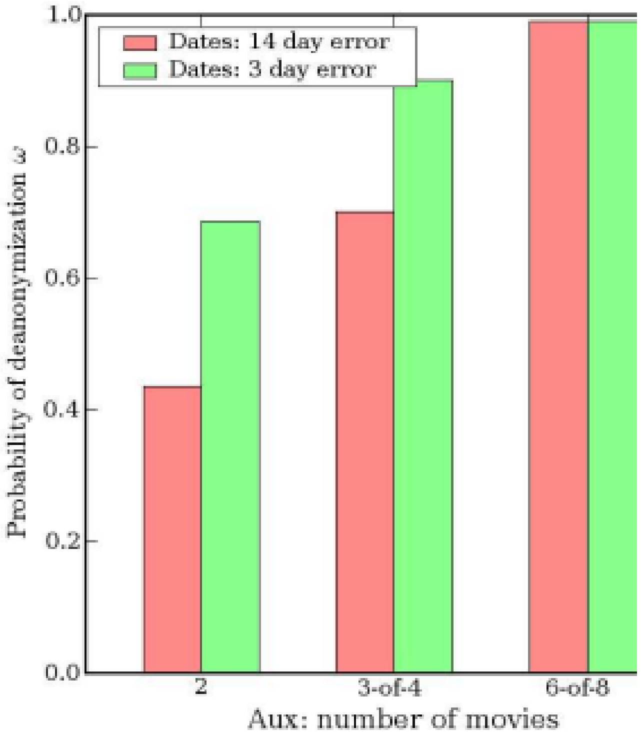


Figure 1: De-anonymization: adversary knows exact ratings and approximate dates.

Source: Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

Results – De-anonymization with auxiliary info (2)

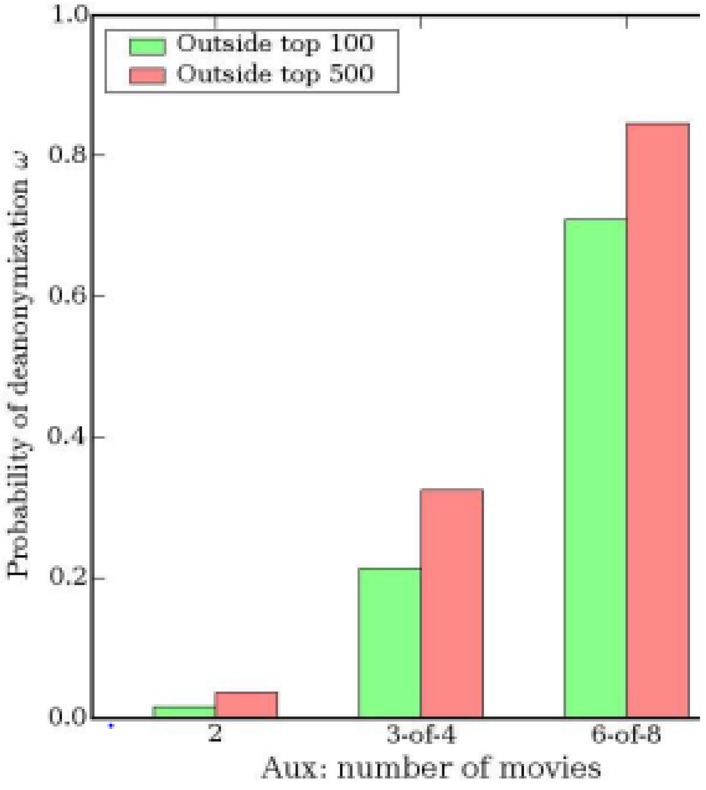


Figure 4: Adversary knows exact ratings but does not know dates at all.

Source: Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

Results – De-anonymization using IMDb

- 2 users were identified with high-confidence:
  - One from the ratings
  - One from the dates

- A small number though it raises privacy concerns:
  - Movies watched → Political orientation, Religious Views

Conclusions

1/27/2024

Conclusions

- These anonymization techniques does not offer enough privacy guarantees
- Still susceptible to attacks as Background Knowledge Attack, Reidentification Attack
- Even noisy data can be used to breach the techniques discussed in this course

References

1. https://course.ece.cmu.edu/~ece734/lectures/lecture-2018-10-08-deanonymization.pdf

2. https://www.immuta.com/blog/k-anonymity-everything-you-need-to-know-2021-guide/

3. https://satoricyber.com/data-masking/how-k-anonymity-preserves-data-privacy/

4. Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.