

18734: Foundations of Privacy

Database Privacy: k-anonymity and de-anonymization attacks

Piotr Mardziel or Anupam Datta

CMU

Fall 2018

Publicly Released Large Datasets

- ▶ Useful for improving recommendation systems, collaborative research
- ▶ Contain personal information
- ▶ Mechanisms to protect privacy, e.g. anonymization by removing names

▶ Yet, private information leaked by attacks on anonymization mechanisms



m o v i e l e n s
helping you find the *right* movies



amazon.com.



WIKIPEDIA
The Free Encyclopedia

Article [Discussion](#)

AOL search data leak

From Wikipedia, the free encyclopedia

Non-Interactive Linking

**Background/
Auxiliary
Information**

DB1

DB2



Algorithm to link information



De-identified record

Roadmap

▶ ~~Motivation~~

▶ Privacy definitions



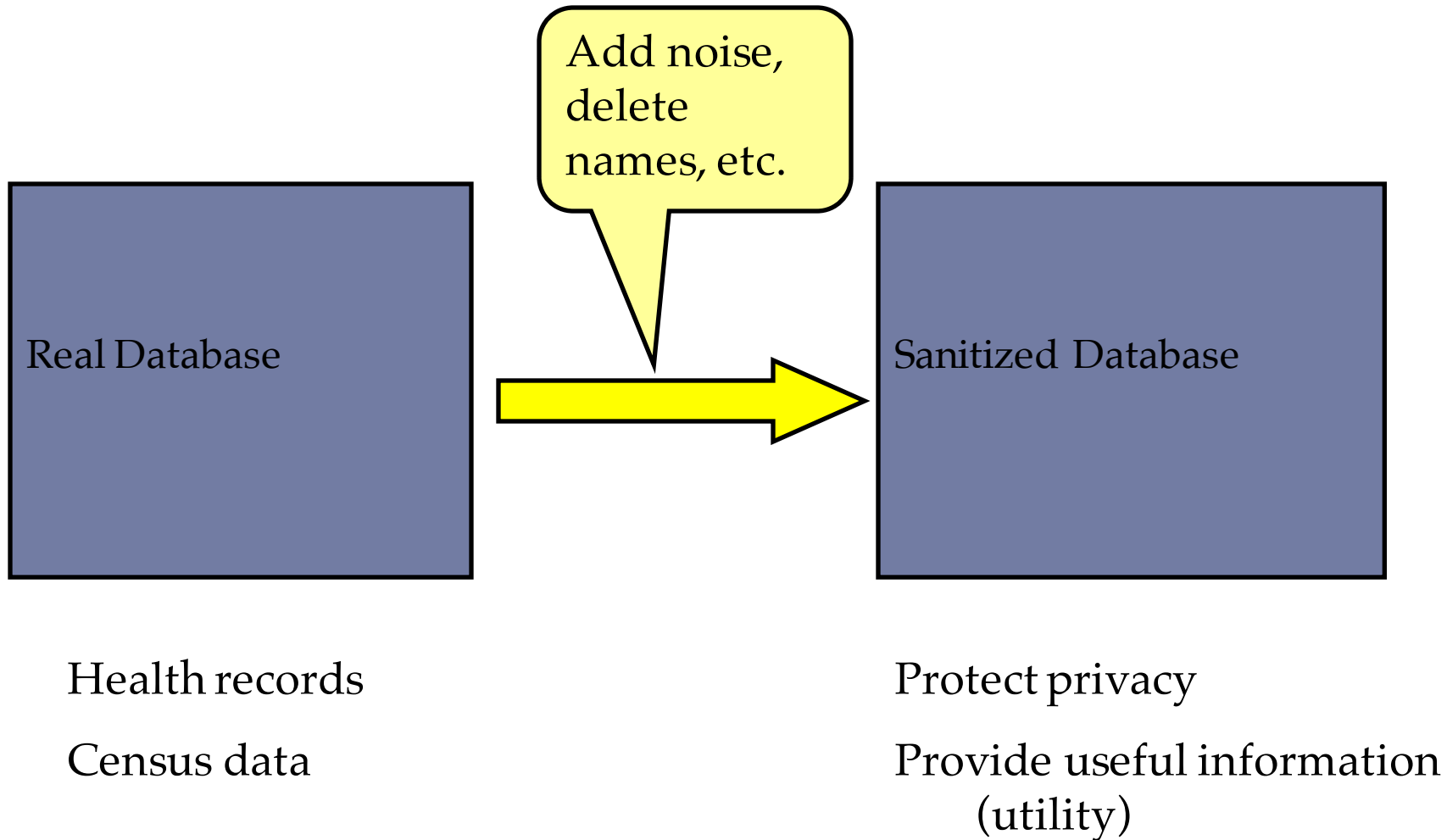
▶ Netflix-IMDb attack

▶ Theoretical analysis

▶ Empirical verification of assumptions

▶ Conclusion

Sanitization of Databases

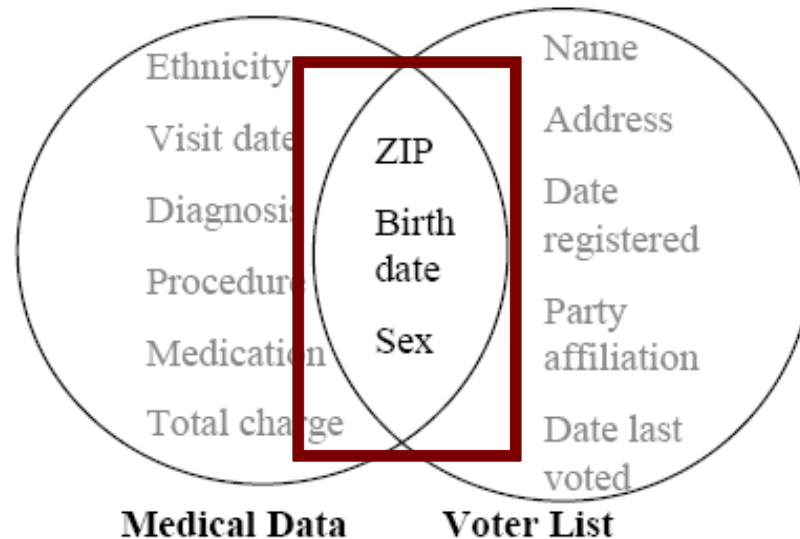


Database Privacy

- ▶ Releasing sanitized databases
 1. k-anonymity [Samarati 2001; Sweeney 2002]
 2. Differential privacy [Dwork et al. 2006] (*future lecture*)

Re-identification by linking

Linking two sets of data on shared attributes may uniquely identify some individuals:



87 % of US population uniquely identifiable by 5-digit ZIP, gender, DOB

K-anonymity

- ▶ Quasi-identifier: Set of attributes that can be linked with external data to uniquely identify individuals
- ▶ Make every record in the table indistinguishable from at least $k-1$ other records with respect to quasi-identifiers
- ▶ Linking on quasi-identifiers yields at least k records for each possible value of the quasi-identifier

K-anonymity and beyond

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

Figure 1. Inpatient Microdata

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	130**	< 30	*	Heart Disease
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Figure 2. 4-anonymous Inpatient Microdata

Provides some protection: linking on ZIP, age, nationality yields 4 records

Limitations: lack of diversity in sensitive attributes, background knowledge, subsequent releases on the same data set

Re-identification Attacks in Practice

Examples:

- ▶ Netflix-IMDB
- ▶ Movielens attack
- ▶ Twitter-Flicker
- ▶ Recommendation systems – Amazon, Hunch,..

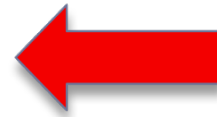
Goal of De-anonymization: To find information about a record in the released dataset

Roadmap

▶ ~~Motivation~~

▶ ~~Privacy definitions~~

▶ Netflix-IMDb attack



▶ Theoretical analysis

▶ Empirical verification of assumptions

▶ Conclusion

Anonymization Mechanism


	Gladiator	Titanic	Heidi
 Bob	5	2	1
Alice	3	2.5	2
Charlie	1.5	2	2

Each row corresponds to an individual

Each column corresponds to an attribute, e.g. movie

Delete name identifiers and add noise



	Gladiator	Titanic	Heidi
 r ₁	4	1	0
r ₂	2	1.5	1
r ₃	0.5	1	1

Anonymized Netflix DB

De-anonymization Attacks Still Possible

▶ Isolation Attacks

- ▶ Recover individual's record from anonymized database
- ▶ E.g., find user's record in anonymized Netflix movie database

▶ Information Amplification Attacks

- ▶ Find more information about individual in anonymized database
- ▶ E.g. find ratings for specific movie for user in Netflix database

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings
(noisy)

	Titanic	Heidi
 Bob	2	1

Used as auxiliary information



Weighted Scoring Algorithm



Isolation Attack!

	r_1	4	1	0
---	-------	---	---	---

Problem Statement

Anonymized database

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Auxiliary information about a record (noisy)

	Titanic	Heidi
 Bob	2	1

Attacker uses algorithm to find record

Attacker's goal: Find r_1 or record similar to Bob's record

Enhance theoretical understanding of why empirical de-anonymization attacks work

Research Goal

Characterize classes of auxiliary information and properties of database for which re-identification is possible

Roadmap

▶ ~~Motivation~~

▶ ~~Privacy definitions~~

▶ ~~Netflix-IMDb attack~~

▶ Theoretical analysis



▶ Empirical verification of assumptions

▶ Conclusion

Netflix-IMDb Empirical Attack [Narayanan et al 2008]

Anonymized Netflix DB

	Gladiator	Titanic	Heidi
r_1	4	1	0
r_2	2	1.5	1
r_3	0.5	1	1

Publicly available IMDb ratings
(noisy)

	Titanic	Heidi
 Bob	2	1

Used as auxiliary information

Weighted Scoring Algorithm

How do you measure similarity of this record with Bob's record?
(Similarity Metric)

What does **auxiliary information** about a record mean?

	r_1	4	1	0
---	-------	---	---	---

Definition: Asymmetric Similarity Metric

	Gladiator	Titanic	Heidi
	v_1	v_2	v_3
y	5	0	-
r	0	2	3

Individual Attribute Similarity

$$T(y(i), r(i)) = 1 - \frac{|y(i) - r(i)|}{p(i)}$$

$$T(y(v_1), r(v_1)) = 1 - \frac{|5 - 0|}{5} = 0$$

Intuition: Measures how closely two people's ratings match on one movie

Movie (i)	$T(y(i), r(i))$
Gladiator	0
Titanic	0.6
Heidi	0

$p(i)$: range of attribute i

Intuition: Measures how closely two people's ratings match overall

$S(y, r)$	$0.6 / 2 = 0.3$
-----------	-----------------

Similarity Metric

$$S(y, r) = \sum_{i \in \text{supp}(y)} \frac{T(y(i), r(i))}{|\text{supp}(y)|}$$

$\text{supp}(y)$: non null attributes in y

Definition: Auxiliary Information

Intuition:

aux about y should be a subset of record y
aux can be noisy

aux captures information available outside normal data release process

e.g. Netflix



sample

e.g. IMDb

aux



perturb

Bound level of perturbation in *aux*

$$\gamma \in [0, 1]$$

(m, γ) -perturbed auxiliary information

$$\forall i \in \text{supp}(aux) . T(y(i), aux(i)) \geq 1 - \gamma$$

$|\text{supp}(aux)| = m = \text{no. of non null attributes in } aux$

Weighted Scoring [Narayanan et al 2008, Frankowski et al 2006]

Intuition: The fewer the number of people who watched a movie, the rarer it is

Weight of an attribute i

$$w(i) = \frac{1}{\log(|\text{supp}(i)|)}$$

$|\text{supp}(i)|$ = no. of non null entries in column i

Use weight as an indicator of rarity

Score gives a weighted average of how closely two people match on every movie, giving higher weight to rare movies

Scoring Methodology

$$\text{Score}(aux, r_j) = \frac{\sum_{i \in \text{supp}(aux)} w(i) * T(aux(i), r_j(i))}{|\text{supp}(aux)|}$$

$|\text{supp}(aux)|$ = m = no. of non null attributes in aux

Compute *Score* for every record r in anonymized DB to find out which one is closest to target record y

Weighted Scoring Algorithm [Narayanan et al 2008]

Compute *Score* for every r in D

$$Score(aux, r_j) = \sum_{i \in \text{supp}(aux)} \frac{w(i) * T(aux(i), r_j(i))}{|\text{supp}(aux)|}$$

w_i	0.63	0.5	0.63
	v_1	v_2	v_3
r_1	5	2	-
r_2	3	1	4
r_3	-	2	4

Score(aux, r_j)
0.52
0.40
0.23

v_1	v_2
4.5	2.3

aux

One of the records r in anonymized database is y , which row is it?

Eccentricity measure > threshold

$$e(aux, D) = \max_{r \in D} (Score(aux, r)) - \max_{2, r \in D} (Score(aux, r))$$

Output record with max Score

r_1	5	2	-
-------	---	---	---

Score(aux, r) used to predict $S(y, r)$



Where do Theorems Fit?



Computed:
Score of all records r in D with aux



→
Theorems help bridge the gap




Desired:
Guarantee about *Similarity*

r_1	5	2	-
-------	---	---	---



r_1	5	2	-
-------	---	---	---

Theorems

- ▶ Theorem 1: When Isolation Attacks work? 
- ▶ Theorem 2: Why Information Amplification Attacks work?

Theorem 1: When Isolation Attacks work?

Intuition: If eccentricity is high, algorithm always finds the record corresponding to auxiliary information!

If

aux is (m, γ) -perturbed

Eccentricity threshold $> \gamma M$

Eccentricity:
Highest score -
Second highest
score

γ : Indicator of perturbation in aux
 M : Average of weights in aux
 r : Record output by algorithm
 y : Target record

then

$$Score(aux, r) = Score(aux, y)$$

If r is the only record with the highest score then $r = y$

Isolation Attack: Theorem

Theorem IV.1 *Let y denote the target record from a given database D . Let aux_y denote (m, γ) -perturbed auxiliary information about record y . If the eccentricity measure $e(aux_y, D) > \gamma M$ where $M = \frac{\sum_{i \in \text{supp}(aux_y)} w_i}{|\text{supp}(aux_y)|}$ is the scaled sum of weights of attributes in aux_y , then*

- 1) $\max_{r \in D} (\text{Score}(aux_y, r)) = \text{Score}(aux_y, y)$.
- 2) Additionally, if only one record has maximum score value $= \text{Score}(aux_y, y)$, then the record o returned by the algorithm is the same as target record y .

Theorems

- ▶ ~~Theorem 1: When Isolation Attacks work?~~
- ▶ Theorem 2: Why Information Amplification Attacks work?



Intuition: Why Information Amplification Attacks work?

- ▶ If two records agree on rare attributes, then with high probability they agree on other attributes too
- ▶ Use intuition to find record r similar to aux on many rare attributes (using aux as 'proxy' for y)

Intuition: Why Information Amplification Attacks work?

For > 90%
of records

> 0.75

- ▶ If a high **fraction** of attributes in *aux* are **rare**, then any record *r* that is **similar to *aux***, is **similar to *y***

Similarity
> 0.75

Similarity
> 0.65

Theorem 2: Why Information Amplification Attacks work?

Define Function

$$f_D(\eta_1, \eta_2, \eta_3)$$

If a high **fraction** of attributes in *aux* are **rare**, then any record *r* **similar to *aux***, is **similar to *y***

- Measure overall similarity between target record *y* and *r* that depends on:

η_1 : Fraction of rare attributes in *aux*

η_2 : Lower bound on similarity between *r* and *aux*

η_3 : Fraction of target records for which guarantee holds

$$S(y, r) \geq f_D(\eta_1, \eta_2, \eta_3)$$

Theorem 2: Why Information Amplification Attacks work?

Using Function

$$f_D(\eta_1, \eta_2, \eta_3)$$

$$S(y, r) \geq f_D(\eta_1, \eta_2, \eta_3)$$

Theorem gives guarantee about similarity of record output by algorithm with target record

Roadmap

▶ ~~Motivation~~

▶ ~~Privacy definitions~~

▶ ~~Netflix-IMDb attack~~

▶ ~~Theoretical analysis~~

▶ Empirical verification of assumptions

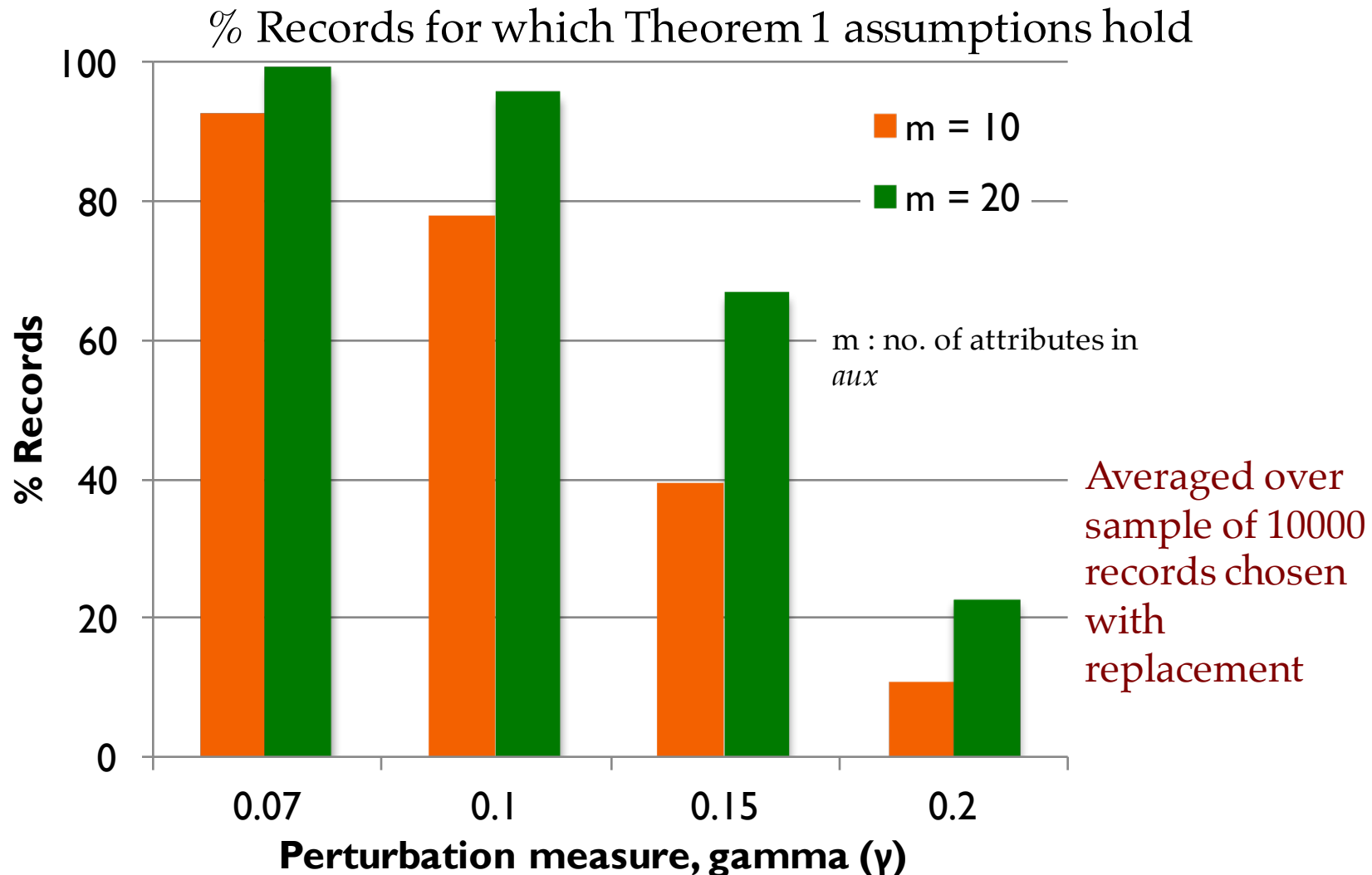


▶ Conclusion

Empirical verification

- ▶ Use `anonymized' Netflix database with 480,189 users and 17,770 movies
- ▶ Percentage values claimed in our results = percentage of records not filtered out because of
 - ▶ insufficient attributes required to form aux OR
 - ▶ insufficient rare or non-rare attributes required to form aux

Do Assumptions hold over Netflix Database?

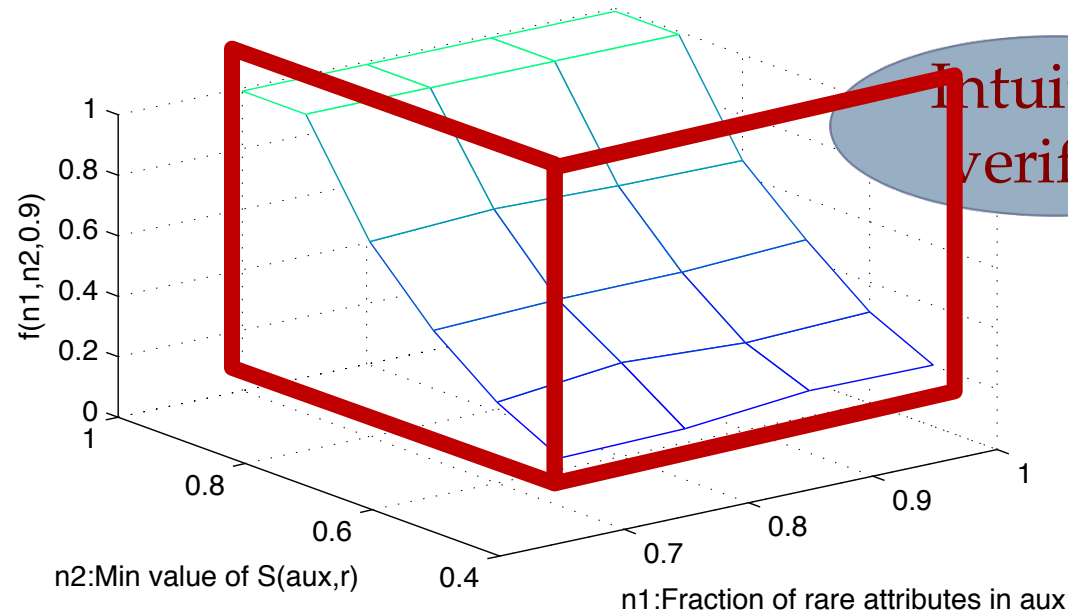


Does Intuition about f_D hold for Netflix Database?

$f_D(\eta_1, \eta_2, \eta_3)$ can be evaluated given D

$$S(y, r) \geq f_D(\eta_1, \eta_2, \eta_3)$$

- η_1 : Fraction of rare attributes in *aux*
- η_2 : Lower bound on similarity between *r* and *aux*
- η_3 : Fraction of target records for which guarantee holds

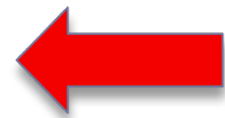


For Netflix DB,

$f_D(\eta_1, \eta_2, \eta_3)$ is monotonically increasing in η_1 and η_2 and tends to 1 as η_2 increases

Roadmap

- ▶ Motivation
- ▶ Privacy definitions
- ▶ Netflix-IMDb attack
- ▶ Theoretical analysis
- ▶ Empirical verification of assumptions
- ▶ Conclusion



Conclusion

- ▶ Naïve anonymization mechanisms do not work
- ▶ We obtain **provable** bounds about, and **verify empirically**, why some de-anonymization attacks work in practice
- ▶ Even perturbed auxiliary information can be used to launch de-anonymization attacks if:
 - ▶ *Database* has many **rare dimensions** and
 - ▶ *Auxiliary information* has information about these rare dimensions

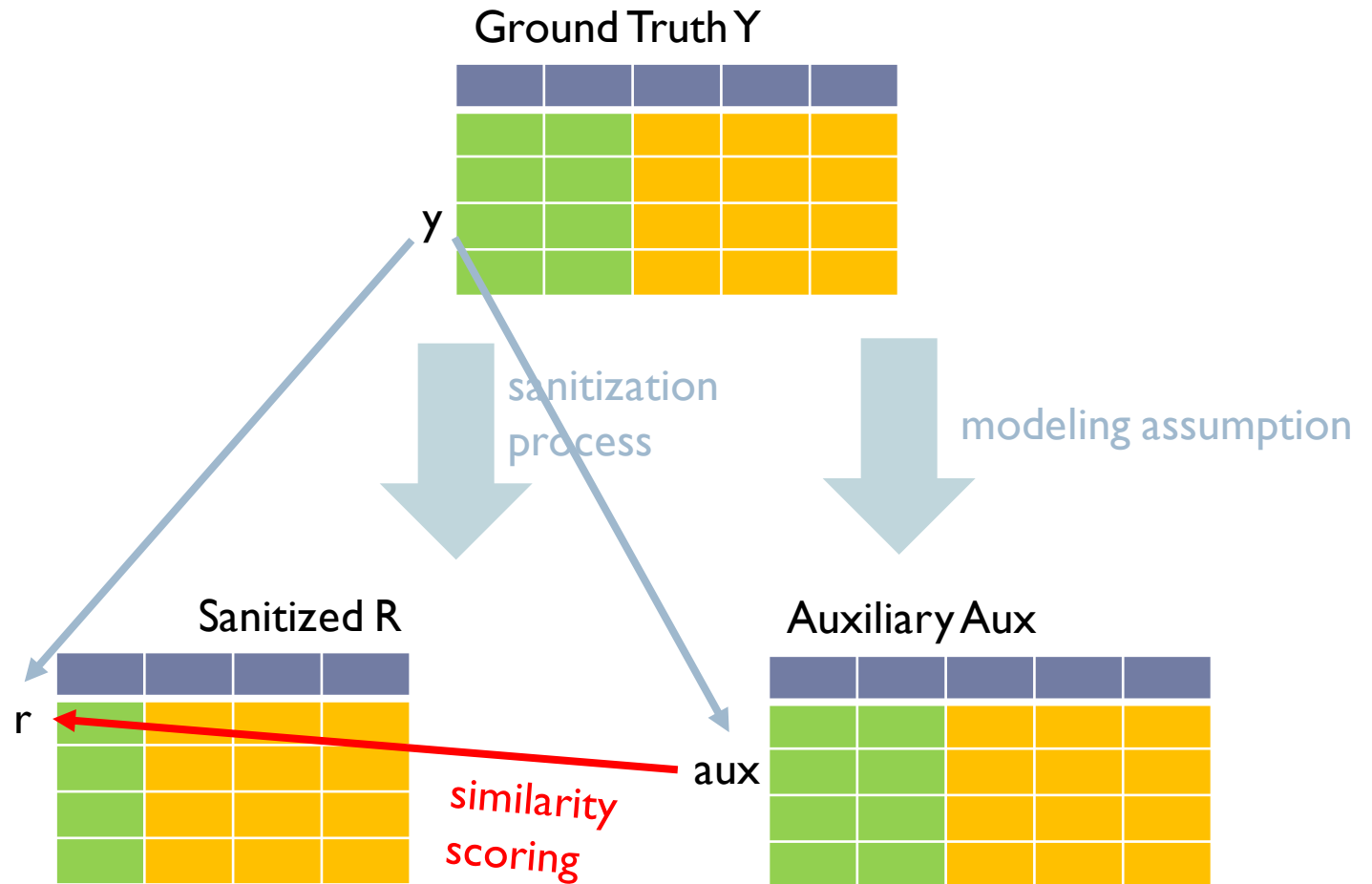
Summary

- ▶ **Anonymity via sanitization**
 - ▶ Offline sanitization
 - ▶ Online sanitization (next lecture)
- ▶ **Privacy definitions**
 - ▶ k-anonymity
 - ▶ l-diversity
 - ▶ ...

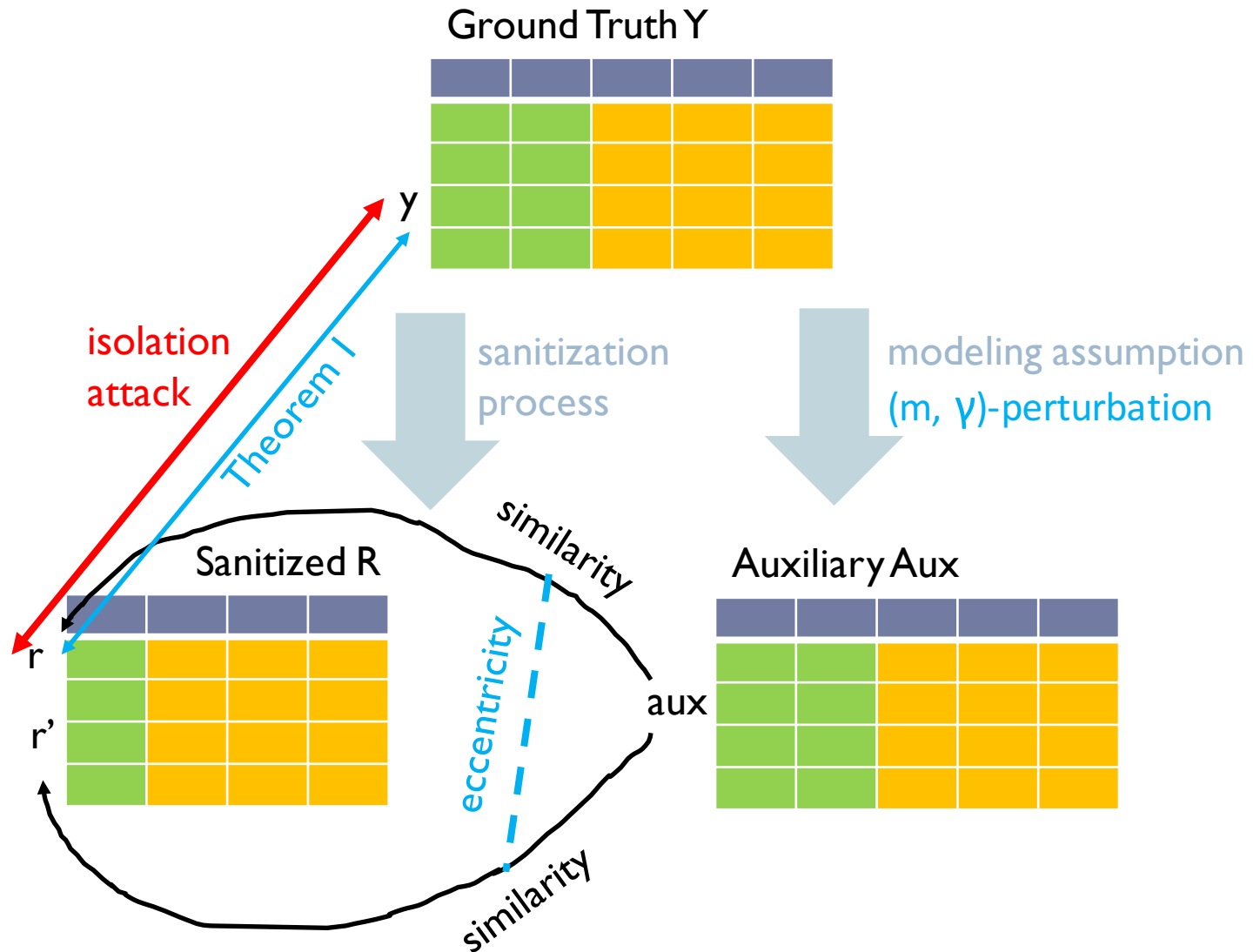
Summary

- ▶ **Deanonmyzation attacks**
 - ▶ Isolation
 - ▶ Amplification
- ▶ **Measuring attack success without ground truth**
 - ▶ Assumptions
 - ▶ (m, γ) -perturbation
 - ▶ Measurables
 - ▶ similarity
 - ▶ eccentricity
 - ▶ η_1, η_2, η_3

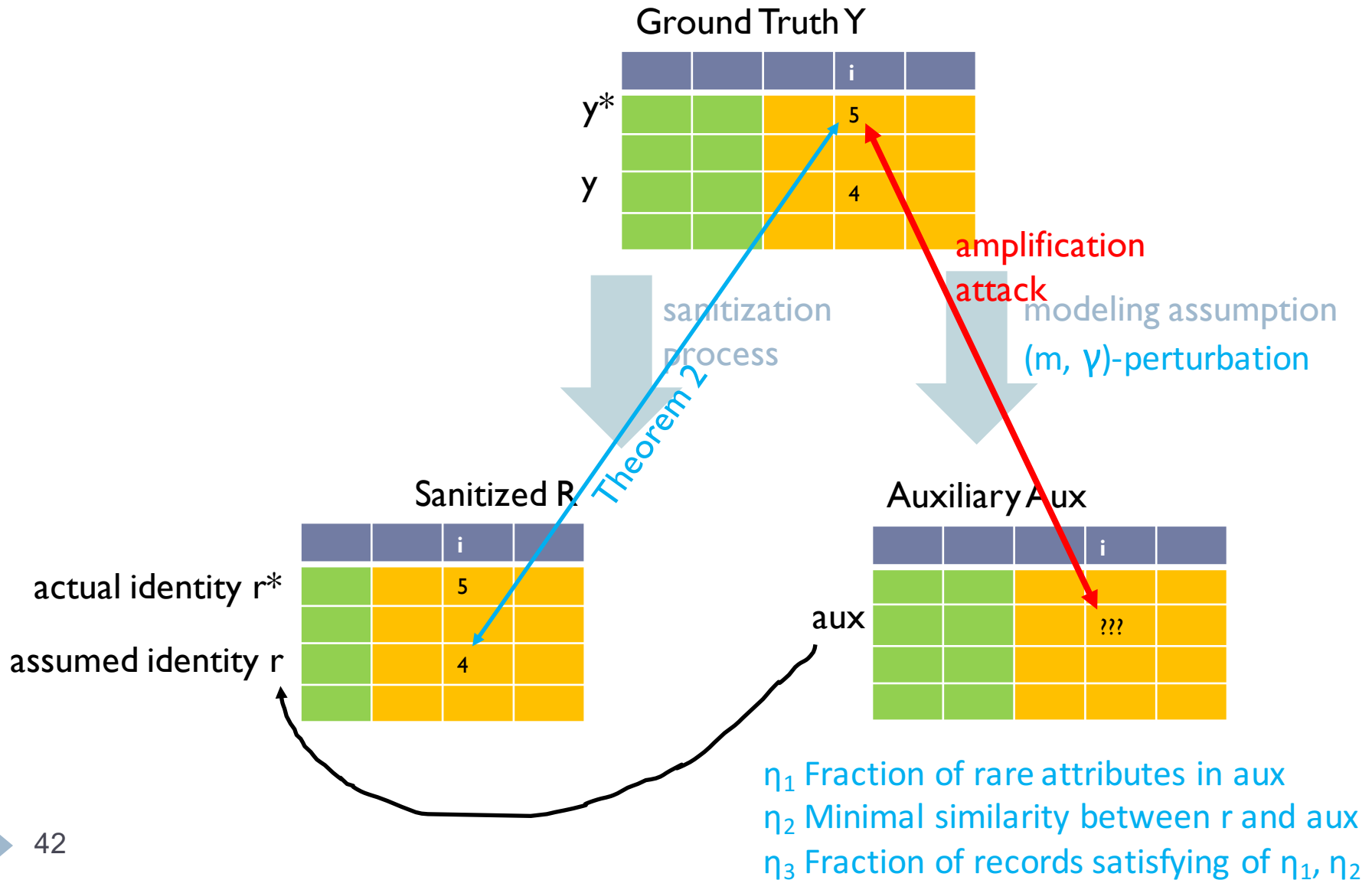
Deanononymization



Isolation attack



Amplification attack



Anonymization settings

Offline/non-interactive
release sanitized dataset

Online/interactive
sanitize queries

Privacy definitions

k-anonymity
Minimum anonymity set size

l-diversity
Minimum sensitive range size

Assumptions and Experimental Measurements

Given aux in Aux, isolate r in R closest to it

Modeling

Y -- Ground Truth records (**NOT KNOWN**)

R -- Sanitized records = Obf(Y)

Aux -- Auxiliary records

(m, γ)-perturbation – g.t./aux relationship

Measurements

e – eccentricity

best isolate r vs second best r'

η_1 Fraction of rare attributes in aux

η_2 Minimal similarity between r and aux

η_3 Fraction of records satisfying of η_1, η_2

Deanonimization attacks

Isolation

Link auxiliary aux in A to r in R.

Is aux is same identity as g.t. $y \rightarrow r$?

Amplification

Use R to find values of fields not in aux

Are predicted values close to g.t. y ?

Success estimation

Theorem 1: Isolation success

(m, γ), eccentricity \rightarrow successful isolation

Theorem 2: Amplification success

$\eta_1, \eta_2, \eta_3 \rightarrow$ isolated r is close to g.t. y