

Module 1: Empirical Privacy

Privacy for Data Analysis and ML

CS848 Fall 2024



UNIVERSITY OF
WATERLOO



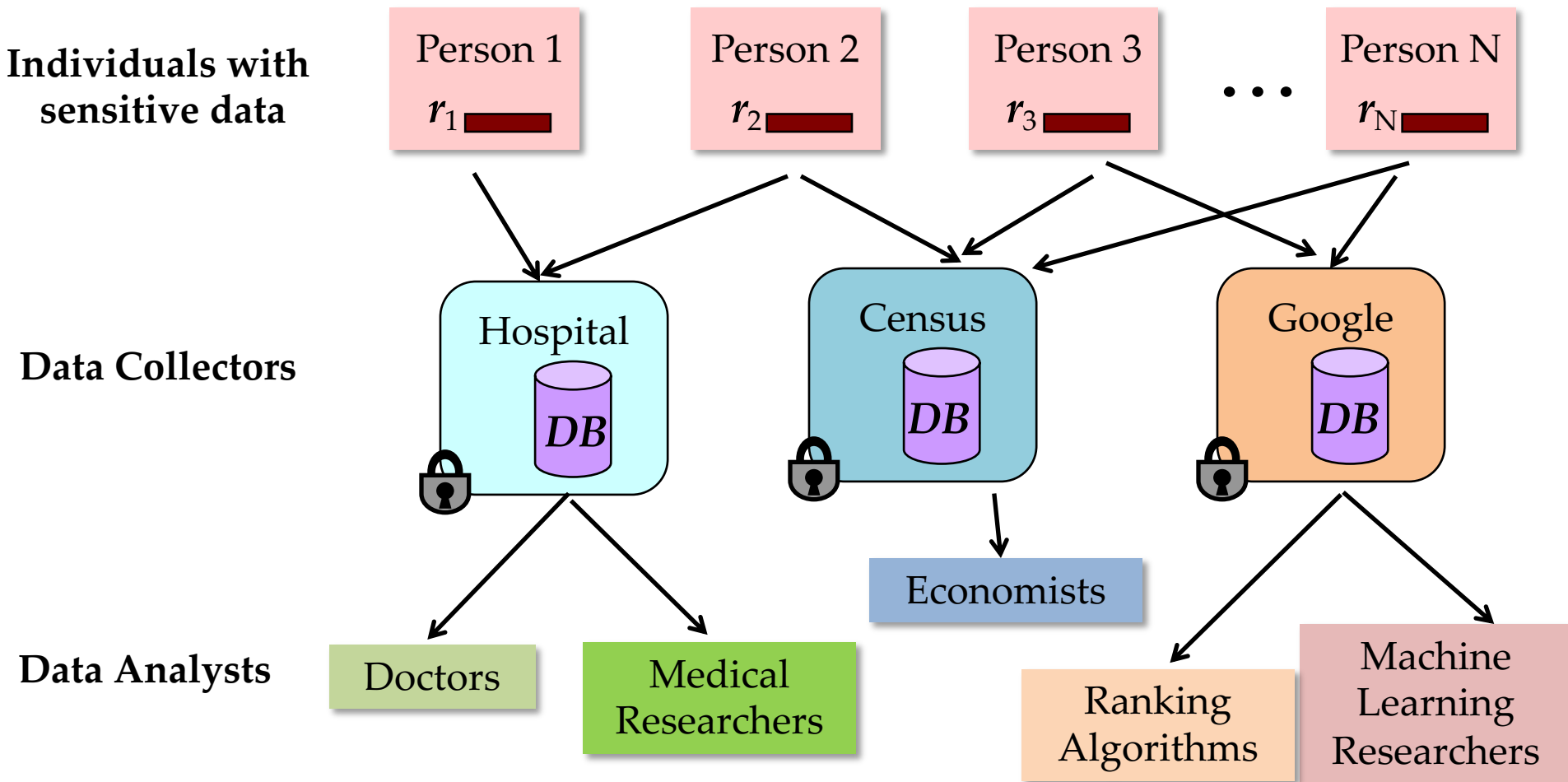
Module 1: Empirical Privacy

1. De-anonymizing Data: (30 mins)
A case study on de-anonymizing Netflix data
2. Measures of Anonymity/Privacy: (30 mins)
k-Anonymity, l-Diversity, t-Closeness
3. Privacy Attacks Practicum: (30 mins)
In-class exercises
4. Privacy Risks in ML: (20 mins)
Membership inference attacks

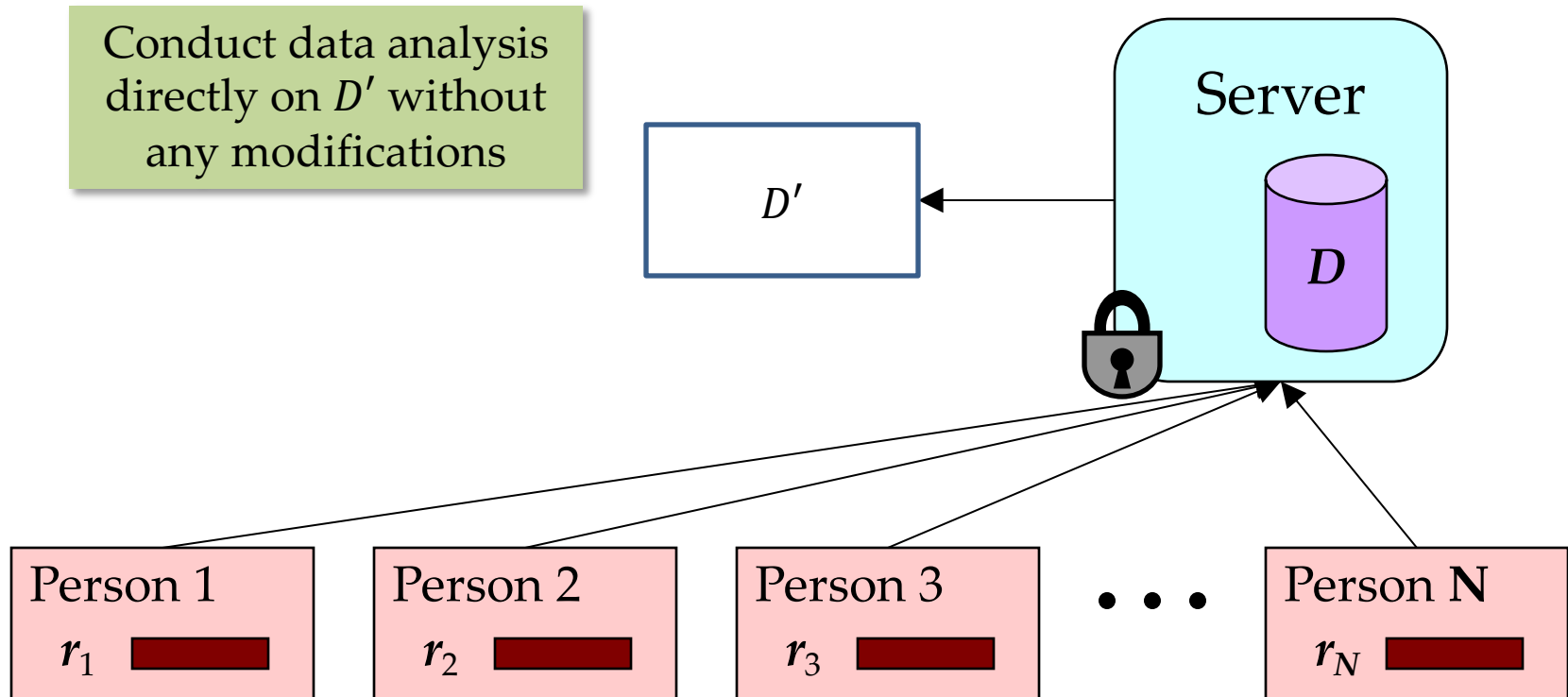
A case study on de-anonymizing Netflix data

1. DE-ANONYMIZING DATA

Statistical Databases



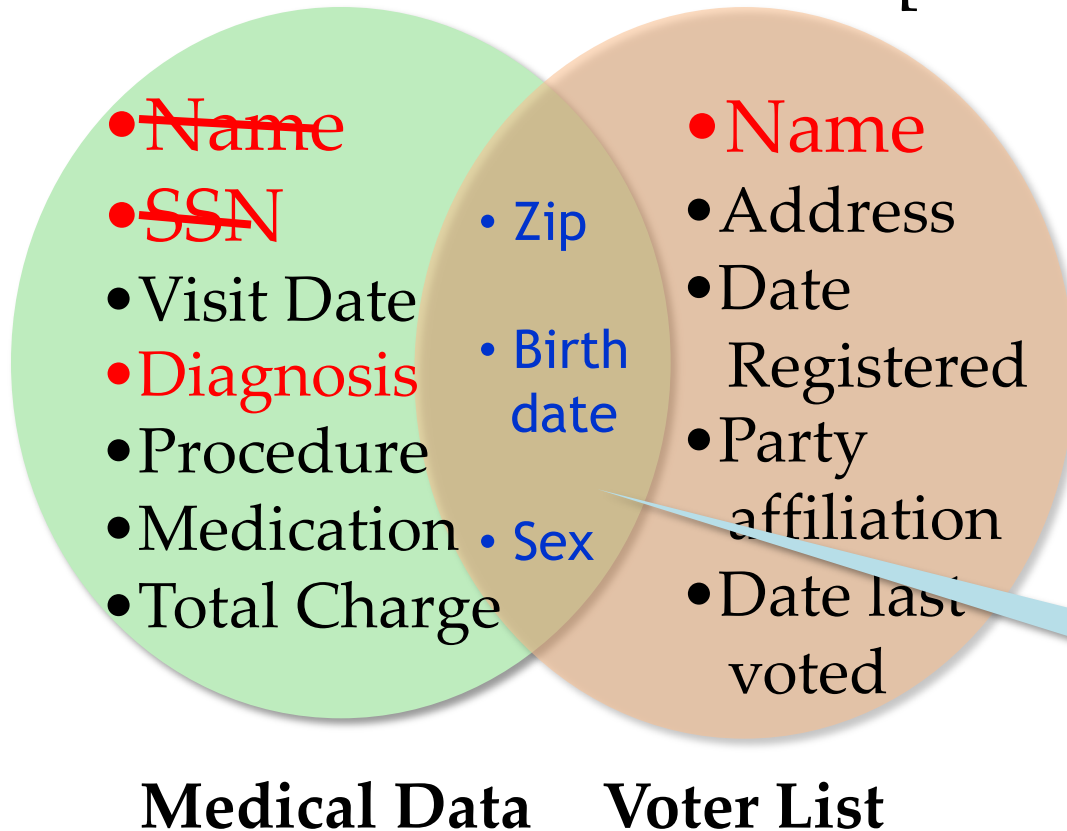
Anonymous/Sanitized Data Publishing



Naïve Anonymization

- Remove identifying attributes from the data
 - E.g., Health Insurance Portability and Protection Act (HIPPA): remove 18 attributes regarded as Personally Identifying Information (PII)
 - Name
 - Geography smaller than state
 - Date (more detailed than year)
 - Tel/Fax/Email
 - SSN
 - IDs (Medical record/Health insurance/Accounts/Certificates/Devices)
 - Vehicle ID/License plate
 - URLs/IP addresses
 - Full face photos/biometrics/genetic code

Can re-identify individuals using other datasets ... [Sweeney IJUFKS 2002]



- 87 % of US population **uniquely identified** using ZipCode, Birth Date, and Sex.

Quasi Identifier

De-anonymization

- What is it?
 - “Algorithms for identifying individual records and their sensitive values from naively anonymized data using background knowledge (usually other public datasets)”
- Also called
 - Record linkage
 - Entity resolution
 - Fuzzy matching
 - ...
- Case study in this class
 - Algorithmically de-anonymizing Netflix data

Netflix Dataset

Column/Attribute

Record (r)

Rating + TimeStamp

Support: Set (or number) of non-null attributes in a record or column

	Movies									
Users	3			4			2	1		5
			1		1			1		
		5		5				1		
	5				2		2	1		
		4			2			1	4	
			3			3				5
	4			3	1					
	3						2		4	

De-anonymization

- Suppose we have a table AUX
 - $\langle \text{name/id} \rangle$, set of known movie ratings
 - E.g., a single record about someone you know
 - IMDb ratings which are public
- Goal:
 - Match individuals in the Netflix data to individuals in the AUX

General Strategy for De-Anonymization

- Inputs:
 - Private database D , and auxiliary information AUX
- Pairwise Matching:
 - Compute the similarity between candidate matching pairs
 - Based on attributes of the individuals
- Record Linkage:
 - For each record in AUX , find the best matching record in D (or no match) ... or vice versa
- Blocking:
 - Identity obvious non-matches (and exclude them ...)
 - Remaining set of pairs are candidates matches

Pairwise Matching Features

- Comparison vector:
 - For two records x and y , compute a *vector similarity scores* of component attribute
 - [Same rating for movie X ,
same rating for movie Y ,
Number of Drama movies rated in both records,
...]
- Similarity scores
 - Boolean (match or not-match)
 - Real values based on distance functions
 - Real values based on set or vector similarity

Summary of Matching Features

Permit efficient scalable implementation

- Equality on a Boolean predicate
- Edit distance
 - Levenstein, Smith-Waterman, Affine

- Set similarity
 - Jaccard, Dice
- Vector Based
 - Cosine similarity, TFIDF

Good for Text, sets, class membership, ...

Handle typographical errors

Good for Names

- Alignment-based or Two-tiered
 - Jaro-Winkler, Soft-TFIDF, Monge-Elkan
- Phonetic Similarity
 - Soundex

Translation-based

- Numeric distance between values
- Domain-specific

Useful for abbreviations, alternate names.

Useful packages:

- Second string: <https://secondstring.sourceforge.net/>
- Simmetrics: <https://sourceforge.net/projects/simmetrics/>

Netflix Paper Comparison Vector

Given 2 records r in D and r' in AUX

For each movie m ,

$\text{Sim}(r[m], r'[m]) = 1$ if m was rated in both records with *similar* values and at *similar* times

$= 0$ otherwise

Pairwise Match Score

- Problem: Given a vector of component-wise similarities for a pair of records (x,y) , compute $P(x \text{ and } y \text{ match})$.

SCHEME USED in NETFLIX ATTACK

$$w = \frac{1}{\log(\text{support}(m))}$$

- Solutions:

1. Weighted sum or average of component-wise similarity scores.

$$0.05 * \text{Sim}[m1] + 0.02 * \text{Sim}[m2] + 0.03 * \text{Sim}[m3] + \dots$$

- How to pick weights?
 - Similarity on rare attribute (rate movie) is more predictive of match than similarity on common attribute (blockbuster)

Threshold determines match or non-match.

- Hard to tune a threshold.

Pairwise Match Score

- Problem: Given a vector of component-wise similarities for a pair of records (x,y) , compute $P(x \text{ and } y \text{ match})$.
- Solutions:
 1. Weighted sum or average of component-wise similarity scores.
Threshold determines match or non-match.
 2. Formulate rules about what constitutes a match.
($\text{Sim}[m1] > 0.7 \text{ AND } \text{Sim}[m2] > 0.8$) OR ($\text{Sim}[m1] > 0.9 \text{ AND } \text{Sim}[m3] > 0.9$)
 - Manually formulating the right set of rules is hard

Many methods to compute pairwise matching

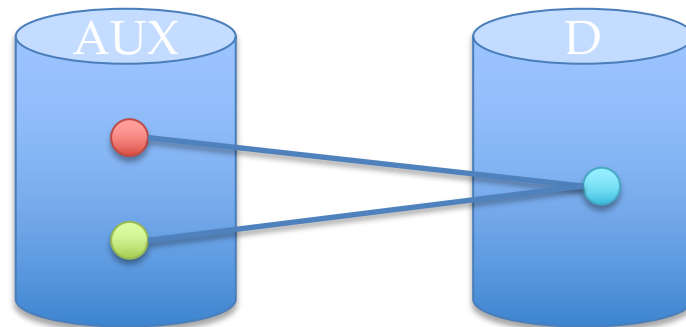
- Fellegi & Sunter Model [FS, Science'69]
 - Assume attributes are independent (Naïve Bayes Assumption) to simplify the problem
 - Use Training datasets to compute
 - Record pair: $r=(x,y)$ in $A \times B$
 - $C(r)$ is a comparison vector
 - E.g., $C = ["Is x.name=y.name?", "Is x.address=y.address?", \dots]$
 - Assume binary vector for simplicity
 - M : set of matching pairs of records
 - U : set of non-matching pairs of records
- Think of this as a machine learning classification problem
 - Given some training data
 - Classify pairs of records as matches or non-matches

General Strategy for De-Anonymization

- Inputs:
 - Private database D , and auxiliary information AUX
- Pairwise Matching:
 - Compute the similarity between candidate matching pairs
 - Based on attributes of the individuals
- Record Linkage:
 - For each record in AUX , find the best matching record in D (or no match) ... or vice versa
- Blocking:
 - Identity obvious non-matches (and exclude them ...)
 - Remaining set of pairs are candidates matches

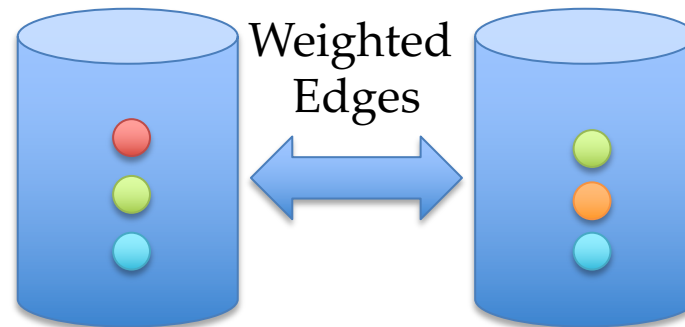
Record Linkage

- Want to find the best matching between AUX and D ...
- ... but pairwise matching may result in 2 records in AUX having a high probability of matching the same record in D



Record Linkage

- Solutions:
 - Pick the best match such that second best match has a very low score ... (Netflix attack solution)
 - Bipartite Matching
 - Edge weights: log odd of matching

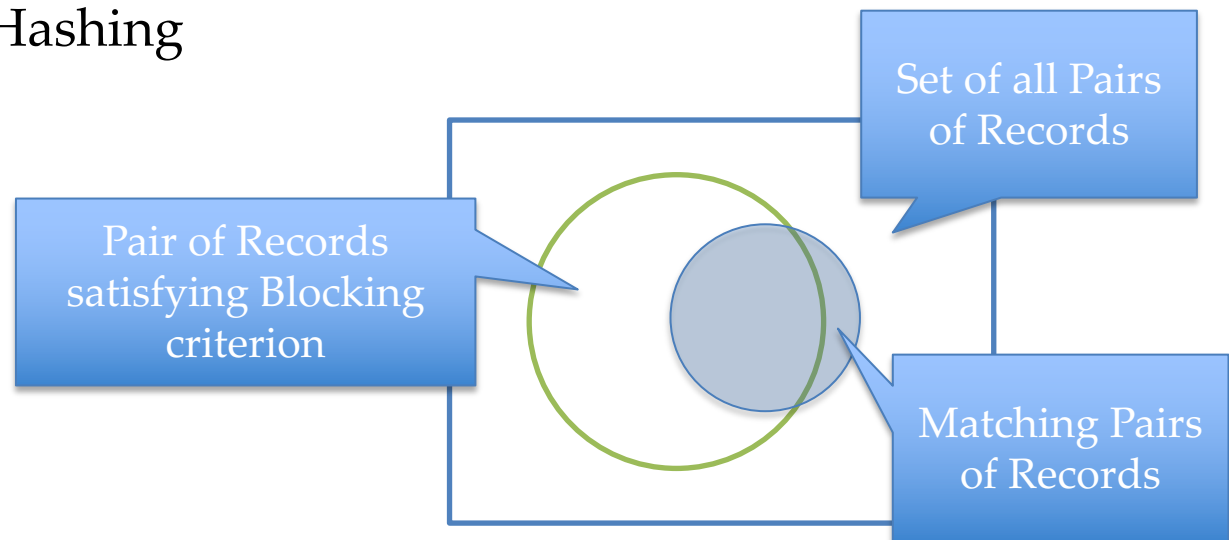


General Strategy for De-Anonymization

- Inputs:
 - Private database D , and auxiliary information AUX
- Pairwise Matching:
 - Compute the similarity between candidate matching pairs
 - Based on attributes of the individuals
- Record Linkage:
 - For each record in AUX , find the best matching record in D (or no match) ... or vice versa
- Blocking (optional):
 - Identity obvious non-matches (and exclude them ...)
 - Remaining set of pairs are candidates matches

Blocking

- Number of pairs of records = $|AUX| \times |D|$
 - Techniques can be inefficient when these databases are very large
- Blocking
 - Identify pairs of records that don't match (with very high probability)
 - Example: minHashing



Back to Netflix Attack

- Pairwise Matching:
 - Comparison vector: for each movie, 1 if similar ratings at similar time in both records
 - Weighted sum: weights inversely proportional to popularity of movie
 - Threshold: prespecified α
- Record Linkage:
 - Best score: pick the record in D with highest score such that second highest score is much smaller
- Blocking: NONE

Analysis

- Theorem 1: Consider a matching threshold $\alpha = 1 - \epsilon$. If the auxiliary record r contains m randomly chosen attributes s.t.

$$m \geq \frac{\log N - \log \epsilon}{-\log(1-\delta)},$$

then the best matching record r' in D is s.t.

$$\Pr[\text{Sim}(r, r') > 1 - \epsilon - \delta] > 1 - \epsilon$$

With high probability, there are no false matches in the matching set.

Summary of Netflix Paper

- Adversary can use a subset of ratings made by a user to uniquely identify the user's record from the "anonymized" dataset with high probability
- Simple algorithm provably guarantees identification of records in the Netflix dataset
- Identification is possible even if records in AUX do not exactly match records in D

k-Anonymity, l-Diversity, t-Closeness

2. MEASURES OF ANONYMITY/PRIVACY

Naïve Anonymization is susceptible to Linkage Attacks

Quasi-identifier



Public information

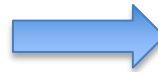
Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer

K-Anonymity [Samarati et al, PODS 1988]

- Generalize, modify, or distort quasi-identifier values so that no individual is uniquely identifiable from a group of k
- In SQL, table T is k -anonymous if each
 - `SELECT COUNT(*)`
`FROM T`
`GROUP BY Quasi-Identifier`
is $\geq k$
- Parameter k indicates the “degree” of anonymity

Example 1: Generalization (Coarsening)

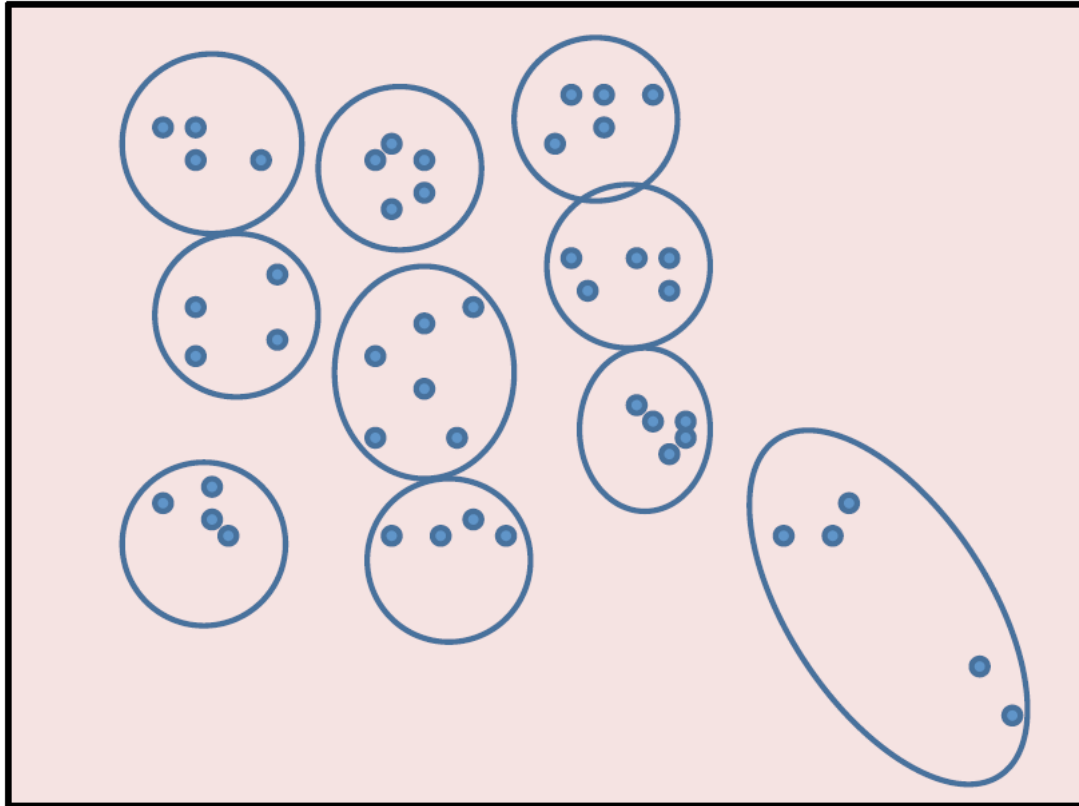
Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37		
13068	36		
13068	32		



Zip	Age	Nationality	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

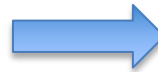
Equivalence Class: Group of k-anonymous records that share the same value for the Quasi-identifier attributes

Example 2: Clustering



Example 3: Microaggregation

Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Flu
13053	23	American	Flu
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Flu
14850	59	American	Flu
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American	Cancer



Zip	Age	Nationality	Disease
4 tuples Zip code = 130** $23 < \text{Age} < 29$ Average(age) = 25			2 Heart and 2 Flu
4 tuples Zip = 1485* $47 < \text{Age} < 59$ Average(age) = 53			1 Cancer, 1 Heart and 2 Flu
4 tuples Zip = 130** $31 < \text{Age} < 37$ Average(age) = 34			All Cancer patients

K-Anonymity

- Joining the published data to an external dataset using quasi-identifiers results in **at least k records** per quasi-identifier combination.
- What is a quasi-identifier?
 - Combination of attributes (that an adversary may know) that uniquely identify a large fraction of the population.
 - There can be many sets of quasi-identifiers
 - If $Q=\{B,Z,S\}$ is a quasi-identifier, thane $Q+\{N\}$ is also a quasi-identifier.
 - Need to guarantee k-anonymity against the largest set of quasi-identifiers

Does k -Anonymity guarantee
sufficient privacy?

Attack 1: Homogeneity

Bob has Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Attack 2: Background knowledge

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

Umeko has Flu

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Recall the attacks on k-Anonymity

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

Umeko has Flu

Bob has Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Heart
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Flu
1485*	>40	*	Flu
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

3-Diverse Table

Name	Zip	Age	Nat.
Umeko	13068	24	Japan

Japanese have a very low incidence of Heart disease.

Umeko has Flu

Bob has Cancer

Name	Zip	Age	Nat.
Bob	13053	35	??

Zip	Age	Nat.	Disease
130**	<30	*	Heart
130**	<30	*	Cancer
130**	<30	*	Flu
130**	<30	*	Flu
1485*	>40	*	Cancer
1485*	>40	*	Heart
		*	Flu
		*	Flu
		*	Heart
		*	Flu
		*	Cancer
130**	30-40	*	Cancer

L-Diversity Principle:
Every group of tuples with the same Q-ID values has $\geq L$ distinct sensitive values of roughly equal proportions

L-Diversity: Privacy Beyond K-Anonymity

[Machanavajjhala et al. ICDE 2006]

- L-Diversity Principle:
 - Every group of tuples with the same Q-ID values has $\geq L$ distinct “well represented” sensitive values
- Questions:
 - What kind of adversarial attacks do we guard against?
 - Why is this the right definition of privacy?
 - What does the parameter L signify?

Privacy Specification for L-Diversity

- The link between identify and attribute value is the sensitive information
 - “Does Bob have Cancer? Heart disease? Flu?”
 - “Does Umeko have Cancer? Heart disease? Flu?”
- Adversary knows $\leq L - 2$ negation statements
 - “Umeko does not have Heart Disease”
 - Data Publisher may not know exact adversarial knowledge
- Privacy is breached when the adversary learns the sensitive attribute value with high probability
 - $\Pr[\text{“Bob has Cancer”} \mid \text{published table, adv. knowledge}] > t$

Individual u does not have a specific disease s

Calculating Probabilities

Every world represents a unique assignment of diseases to individuals

	World 1	World 2	World 3	World 4	World 5
Sasha	Cancer	Heart	Heart	Flu	Heart	
Tom	Cancer	Heart	Flu	Heart	Flu	
Umeko	Cancer	Flu	Flu	Heart	Heart	
Van	Cancer	Flu	Heart	Flu	Flu	
Amar	Cancer	Cancer	Heart	Cancer	Flu	
Boris	Cancer	Heart	Cancer	Flu	Heart	
Carol	Cancer	Flu	Flu	Heart	Flu	
Dave	Cancer	Flu	Flu	Flu	Cancer	
Bob	Cancer	Cancer	Cancer	Cancer	Cancer	
Charan	Cancer	Cancer	Cancer	Cancer	Cancer	
Daiki	Cancer	Cancer	Cancer	Cancer	Cancer	
Ellen	Cancer	Cancer	Cancer	Cancer	Cancer	

Set of all possible worlds

Calculating Probabilities

Every world represents a unique assignment of diseases to individuals

	T*	World 1	World 2	World 3	World 4	World 5
Sasha	Cancer 0 Heart 2 Flu 2	Cancer	Heart	Heart	Flu	Heart	
Tom		Cancer	Heart	Flu	Heart	Flu	
Umeko		Cancer	Flu	Flu	Heart	Heart	
Van		Cancer	Flu	Heart	Flu	Flu	
Amar	Cancer 1 Heart 1 Flu 2	Cancer	Cancer	Heart	Cancer	Flu	
Boris		Cancer	Heart	Cancer	Flu	Heart	
Carol		Cancer	Flu	Flu	Heart	Flu	
Dave		Cancer	Flu	Flu	Flu	Cancer	
Bob	Cancer 4 Heart 0 Flu 0	Cancer	Cancer	Cancer	Cancer	Cancer	
Charan		Cancer	Cancer	Cancer	Cancer	Cancer	
Daiki		Cancer	Cancer	Cancer	Cancer	Cancer	
Ellen		Cancer	Cancer	Cancer	Cancer	Cancer	

Set of all possible worlds consistent with T*

Calculating Probabilities

B: Umeko.Disease
≠ Heart

Every world represents
a unique assignment of
diseases to individuals

T*		World 2	World 3	World 4	World 5
Sasha	Cancer 0 Heart 2 Flu 2	Heart	Heart	Flu	Heart	
Tom		Heart	Flu	Heart	Flu	
Umeko		Flu	Flu	Heart	Heart	
Van		Flu	Heart	Flu	Flu	
Amar	Cancer 1 Heart 1 Flu 2	Cancer	Heart	Cancer	Flu	
Boris		Heart	Cancer	Flu	Heart	
Carol		Flu	Flu	Heart	Flu	
Dave		Flu	Flu	Flu	Cancer	
Bob	Cancer 4 Heart 0 Flu 0	Cancer	Cancer	Cancer	Cancer	
Charan		Cancer	Cancer	Cancer	Cancer	
Daiki		Cancer	Cancer	Cancer	Cancer	
Ellen		Cancer	Cancer	Cancer	Cancer	

Set of all possible worlds consistent with (T*,B) with T*

Calculating Probabilities

B: Umeko.Disease
≠ Heart

$$\Pr[\text{Umeko has Flu} \mid B, T^*] =$$

$$\frac{\# \text{ worlds consistent with } B, T^* \text{ where Umeko has Flu}}{\# \text{ worlds consistent with } B, T^*}$$

	T*
Sasha	Cancer 0 Heart 2 Flu 2
Tom	
Umeko	
Van	
Amar	
Boris	Cancer 1 Heart 1 Flu 2
Carol	
Dave	
Bob	
Charan	Cancer 4 Heart 0 Flu 0
Daiki	
Ellen	

World 2 World 3 World 4 World 5

Heart	Heart	Flu	Heart
Heart	Flu	Heart	Flu
Flu	Flu	Heart	Heart
Flu	Flu	Flu	Flu

Counting the # worlds consistent with B, T* is tedious
(and is intractable for more complex forms of B)

Theorem: # worlds consistent with B, T* where Umeko
has Flu is (where B has negation statements)
proportion to
#tupels in Umeko's group who have Flu.

Cancer Cancer Cancer Cancer

Set of all possible worlds consistent with (T*,B) with T*

Data publisher does not know the adversary's knowledge about u

- *Different adversaries have varying amount of knowledge.*
- *Adversaries may have different knowledge about different individuals.*

Therefore, in order for privacy,
check for each individual u , and each disease s

$$\Pr["u \text{ has disease } s" \mid T^*, \text{ adv. knowledge about } u] < t$$

And we are done ...??

NO

L-Diversity: Guarding against unknown adversarial knowledge.

- Limit adversarial knowledge
 - Knows $\leq (L - 2)$ negation statements of the form
 - “Umeko does not have Heart Disease”
- Consider the worst case
 - Consider all possible conjunctions of $\leq (L - 2)$ statements

At least L sensitive values should appear in every group

$L=5$

Cancer	10
Heart	5
Hepatitis	2
Jaundice	1

$\Pr[\text{Bob has Cancer}] = 1$

Guarding against unknown adversarial knowledge

- Limit adversarial knowledge
 - Knows $\leq (L - 2)$ negation statements of the form
 - “Umeko does not have Heart Disease”
- Consider the worst case
 - Consider all possible conjunctions of $\leq (L - 2)$ statements

The L distinct sensitive values in each group should be roughly of equal proportions

Let $t = 0.75$. Privacy of individuals in this group is ensured if, $\frac{\# \text{Cancer}}{\# \text{Cancer} + \# \text{Malaria}} < 0.75$

$L=5$

Cancer	10
Heart	5
Hepatitis	2
Jaundice	1
Malaria	1

$\Pr[\text{Bob has Cancer}] \approx 1$

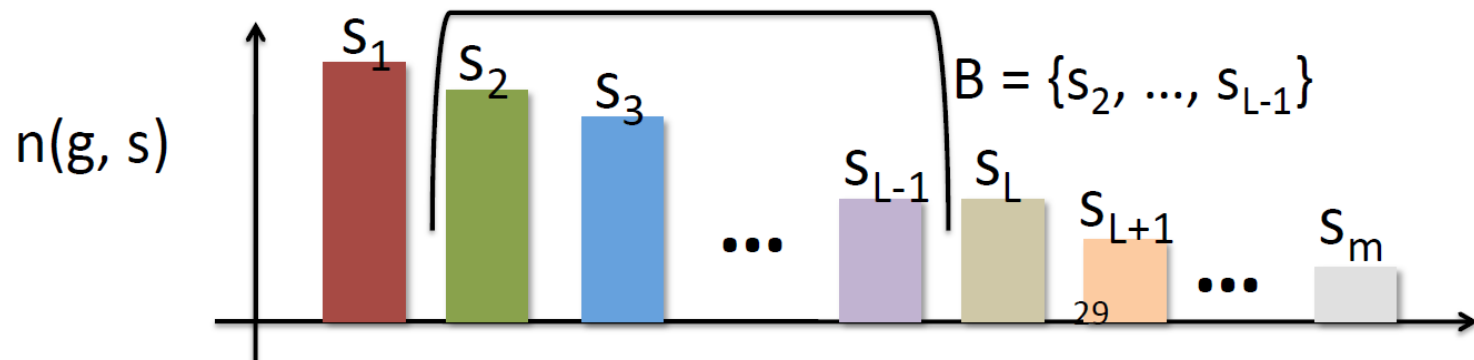
T-closeness [Li et al. ICDE 2007]

- Theorem: For all groups g , for all $s \in S$, and for all B , $|B| \leq (L - 2)$

$$\frac{n(g, s)}{\sum_{s' \in (S \setminus B)} n(g, s')} \leq t$$

is equivalent to

$$\frac{n(g, s_1)}{n(g, s_1) + n(g, s_L) + n(g, s_{L+1}) + \dots + n(g, s_m)} \leq t$$



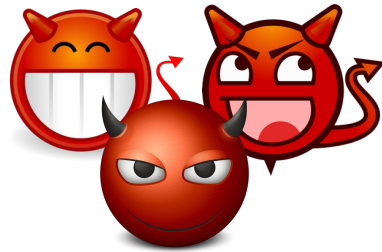
Can we de-anonymize published data that satisfy k -anonymity/ l -diversity/ t -closeness?

In-class exercises

3. PRIVACY ATTACKS PRACTICUM

Your Turn!

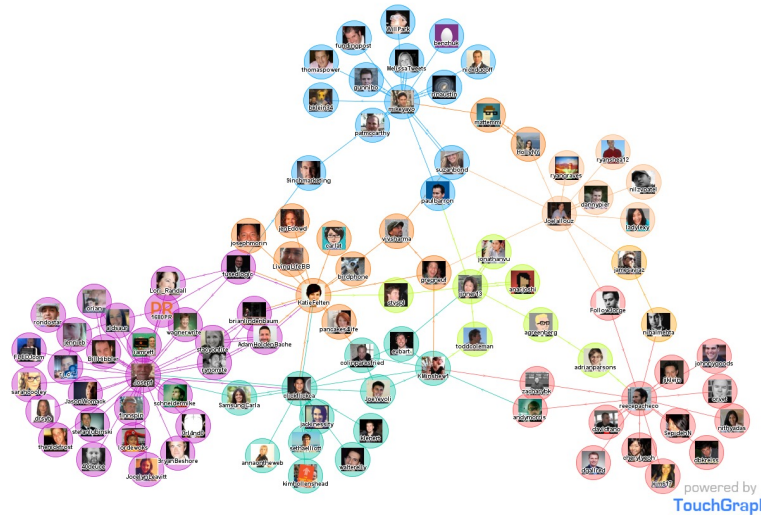
- Divide into groups of 3



- Attack 4 problems as a group (15 mins)

Problem 1

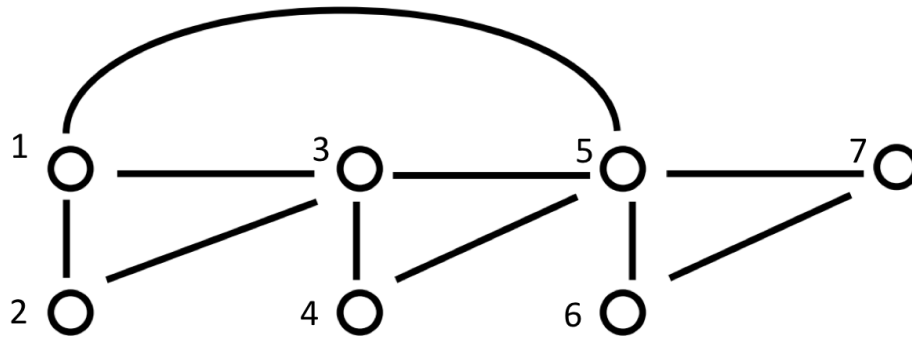
- Social networks: graphs where each node represents a social entity, and each edge represents certain relationship between two entities



- Example: email communication graphs, social interactions like in Facebook, Yahoo! Messenger, etc.

Problem 1

- Anonymized email communication graph



- Unfortunately for the email service providers, investigative journalists **Alice** and **Cathy** are part of this graph. What can they deduce?

Problem 2

- The email service provider also released perturbed records as per a linear function, but with *secret* parameters.

Node ID	Age (perturbed)
1	40
2	34
3	52
4	28
5	48
6	22
7	92

- What can Alice and Cathy deduce now?

Problem 3

- Releasing tables that achieve k-anonymity
 - At least k records share the same quasi-identifier
 - E.g. 4-anonymous table by generalization

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

(a)

Problem 3

- 2 tables of k-anonymous patient records

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

Hospital A (4-anonymous)

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

Hospital B (6-anonymous)


- If Alice visited both hospitals, can you deduce Alice's medical condition?

Problem 4

 U.S. Department of Health & Human Services

[About Us](#) [Careers](#) [Contact Us](#) [Español](#) [FAQ](#) [✉ Email Updates](#)

 **Agency for Healthcare Research and Quality**
Advancing Excellence in Health Care


HCUPnet

Healthcare Cost and Utilization Project

[Home](#)

[Glossary](#)

[Methodology](#)

[Our Partners](#)

[Tutorial](#)

Free Health Care Statistics

HCUPnet is a free, on-line query system based on data from the Healthcare Cost and Utilization Project (HCUP)

The system provides health care statistics and information for hospital inpatient, emergency department, and ambulatory settings, as well as population-based health care data on counties

[Create a New Analysis](#) 

[Get Quick Statistics Tables](#) 

[Find out more about HCUP](#)

[What's new with HCUPnet](#)

The HCUPnet Web site has been redesigned. The new site has a modernized look and feel, a simplified process for querying data, fewer clicks to reach the same information, and more flexibility in changing the content and display of data you are viewing.

Problem 4

- Publishes tables of counts, for counts that are less than 10, they are suppressed as *

Manage Analysis ▾



Analysis Type: Descriptive Statistics
 Setting of Care: Hospital Inpatient
 Geographic Settings: State
 Years: 2009
Categorization Type: Diagnoses--Clinical Classification Software (CCS)
Diagnoses--Clinical Classification Software (CCS): Cancer of ovary
Principal or All-Listed: Principal
Outcome and Measures: Number
Patient Characteristics: Age groups | Sex | Race/ethnicity | Payer | Location of patient's residence
State: New Jersey

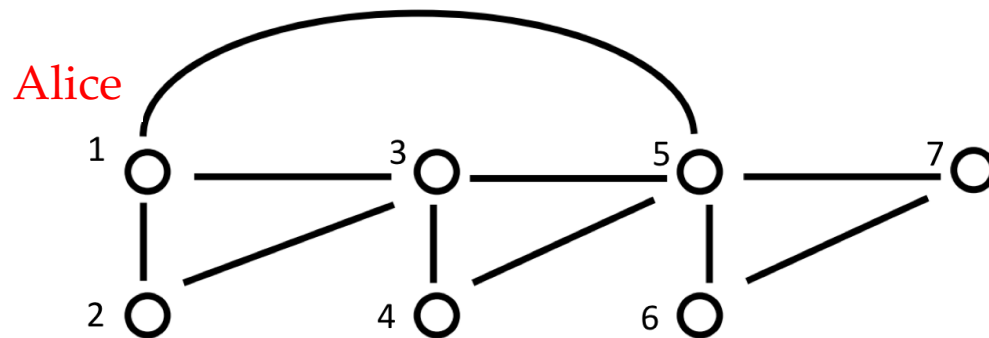
- Can you tell their values?

Let's begin! (15 mins)



Problem 1: Naïve Anonymization

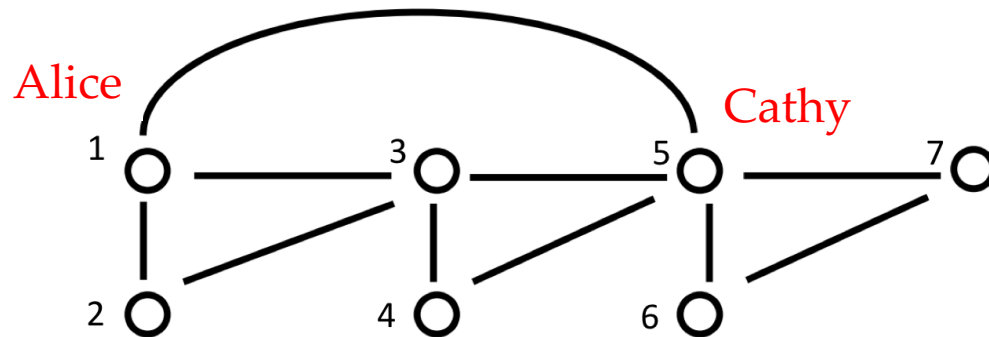
- Auxiliary knowledge:
 - Alice has sent emails to Bob, Cathy, and Ed
 - Cathy has sent emails to everyone, except Ed



- Only one node has a degree 3 \rightarrow node 1: Alice

Problem 1: Naïve Anonymization

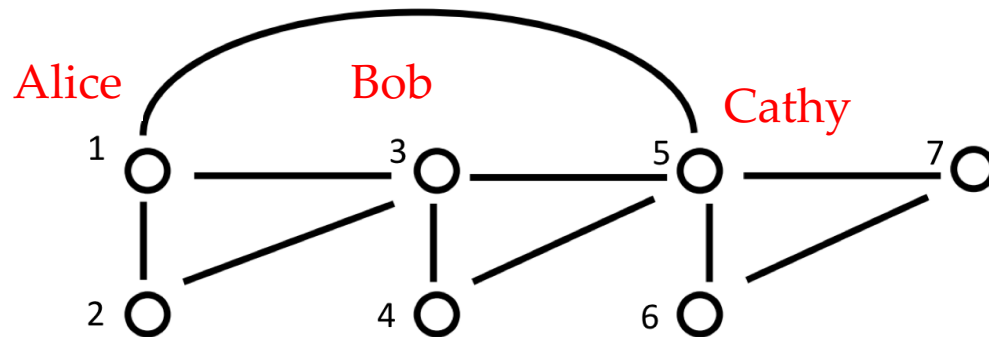
- Auxiliary knowledge:
 - Alice has sent emails to Bob, Cathy, and Ed
 - Cathy has sent emails to everyone, except Ed



- Only one node has a degree 5 \rightarrow node 5: Cathy

Problem 1: Naïve Anonymization

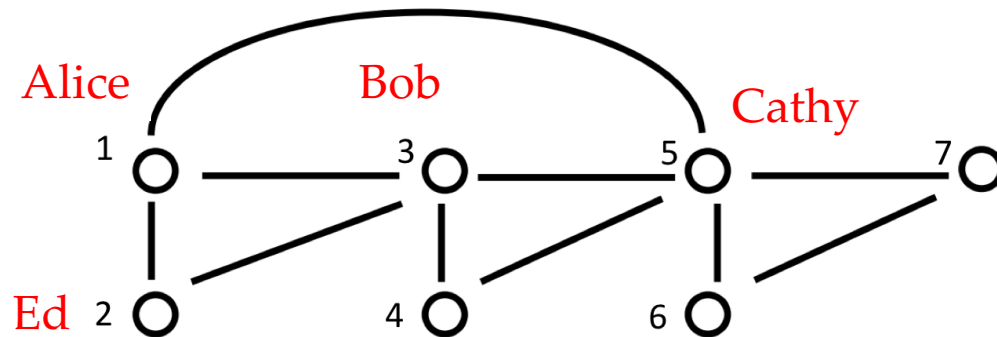
- Auxiliary knowledge:
 - Alice has sent emails to Bob, Cathy, and Ed
 - Cathy has sent emails to everyone, except Ed



- Alice and Cathy know that only Bob has sent emails to both of them → node 3: Bob

Problem 1: Naïve Anonymization

- Auxiliary knowledge:
 - Alice has sent emails to Bob, Cathy, and Ed
 - Cathy has sent emails to everyone, except Ed



- Alice has sent emails to Bob, Cathy, and Ed only
→ node 2: Ed

Attacks using Background Knowledge

- Degrees of nodes [Liu and Terzi, SIGMOD 2008]
- The network structure, e.g., a subgraph of the network. [Zhou and Pei, ICDE 2008, Hay et al., VLDB 2008]
- Anonymized graph with labeled nodes [Pang et al., SIGCOMM CCR 2006]

Desiderata for a Privacy Definition

1. Resilience to background knowledge

- A privacy mechanism must be able to protect individuals' privacy from attackers who may possess background knowledge



Problem 2: Privacy by Obscurity

- Many organization think their data are private because they perturb the data and make the parameters of perturbation secret.

Problem 2: Privacy by Obscurity

Node ID	Name	Age ($\alpha x + \beta$)	True Age
1	Alice	40	25
2	Ed	34	
3	Bob	52	
4		28	
5	Cathy	48	29
6		22	
7		92	


$$\alpha = 2, \beta = -10$$

Problem 2: Privacy by Obscurity

Node ID	Name	Age ($\alpha x + \beta$)	True Age
1	Alice	40	25
2	Ed	34	22
3	Bob	52	31
4		28	19
5	Cathy	48	29
6		22	16
7		92	51


$$\alpha = 2, \beta = -10$$

Desiderata for a Privacy Definition

1. Resilience to background knowledge

- A privacy mechanism must be able to protect individuals' privacy from attackers who may possess background knowledge

2. Privacy without obscurity

- Attacker must be assumed to know the algorithm used as well as all parameters [MK15]

Problem 4: Post-processing

Counts less than k are suppressed achieving k-anonymity

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	*	19	22
1-17	*	*	*	*	*	*	*	*
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

Problem 4: Post-processing

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	*	*	*	*	*	*
18-44	70	40	13	*				*
45-64	330	236	31	32			1	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

$$= 535 - (40+236+229+29)$$

Problem 4: Post-processing

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	*	*	*	*	*	*

Problem 4: Post-processing

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]	[0-2]
18-44	70	40	13	*	*	*	*	*
45-64	330	236	31	32	*	*	11	*
65-84	298	229	35	13	*	*	*	*
85+	34	29	[1-3]	*	*	*	*	*

Can Construct Tight Bounds on Rest of Data

[VSJO 13]

Age	#discharges	White	Black	Hispanic	Asian/ Pcf Hlnder	Native American	Other	Missing
#discharges	735	535	82	58	18	1	19	22
1-17	3	1	[0-2]	[0-2]	[0-1]	[0]	[0-1]	[0-1]
18-44	70	40	13	[9-10]	[0-6]	[0]	[0-6]	[1-8]
45-64	330	236	31	32	[10]	[0]	11	[10]
65-84	298	229	35	13	[2-8]	[1]	[2-8]	[4-10]
85+	34	29	[1-3]	[1-4]	[0-1]	[0]	[0-1]	[0-1]

Can Construct Tight Bounds on Rest of Data

[VSJO 13]

In fact, when linked with queries giving other statistics, we can figure out that exactly 1 Native American woman diagnosed with ovarian cancer went to a privately owned, not for profit, teaching hospital in new Jersey with more than 435 beds in 2009.

Furthermore, the woman did not pay by private insurance, had a routine discharge, with a stay in the hospital of 33.5 days, with her home residence being in a county with 1 million plus residents (large fringe metro, suburbs), and her age was exactly 75 years.

Desiderata for a Privacy Definition

1. Resilience to background knowledge

- A privacy mechanism must be able to protect individuals' privacy from attackers who may possess background knowledge

2. Privacy without obscurity

- Attacker must be assumed to know the algorithm used as well as all parameters [MK15]

3. Post-processing

- Post-processing the output of a privacy mechanism must not change the privacy guarantee [KL10, MK15]

Problem 3: Multiple Releases

- 2 tables of k-anonymous patient records [GKS08]

Non-Sensitive				Sensitive	Non-Sensitive				Sensitive
	Zip code	Age	Nationality	Condition		Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS	1	130**	<35	*	AIDS
2	130**	<30	*	Heart Disease	2	130**	<35	*	Tuberculosis
3	130**	<30	*	Viral Infection	3	130**	<35	*	Flu
4	130**	<30	*	Viral Infection	4	130**	<35	*	Tuberculosis
5	130**	≥40	*	Cancer	5	130**	<35	*	Cancer
6	130**	≥40	*	Heart Disease	6	130**	<35	*	Cancer
7	130**	≥40	*	Viral Infection	7	130**	≥35	*	Cancer
8	130**	≥40	*	Viral Infection	8	130**	≥35	*	Cancer
9	130**	3*	*	Cancer	9	130**	≥35	*	Cancer
10	130**	3*	*	Cancer	10	130**	≥35	*	Tuberculosis
11	130**	3*	*	Cancer	11	130**	≥35	*	Viral Infection
12	130**	3*	*	Cancer	12	130**	≥35	*	Viral Infection

Hospital A (4-anonymous)

Hospital B (6-anonymous)

- Alice is 28 and she visits both hospitals

Problem 3: Multiple Releases

- 2 tables of k-anonymous patient records [GKS08]

Non-Sensitive				Sensitive	Non-Sensitive				Sensitive
	Zip code	Age	Nationality	Condition		Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS	1	130**	<35	*	AIDS
2	130**	<30	*	Heart Disease	2	130**	<35	*	Tuberculosis
3	130**	<30	*	Viral Infection	3	130**	<35	*	Flu
4	130**	<30	*	Viral Infection	4	130**	<35	*	Tuberculosis
5	130**	≥40	*	Cancer	5	130**	<35	*	Cancer
6	130**	≥40	*	Heart Disease	6	130**	<35	*	Cancer
7	130**	≥40	*	Viral Infection	7	130**	≥35	*	Cancer
8	130**	≥40	*	Viral Infection	8	130**	≥35	*	Cancer
9	130**	3*	*	Cancer	9	130**	≥35	*	Cancer
10	130**	3*	*	Cancer	10	130**	≥35	*	Tuberculosis
11	130**	3*	*	Cancer	11	130**	≥35	*	Viral Infection
12	130**	3*	*	Cancer	12	130**	≥35	*	Viral Infection

Hospital A (4-anonymous)

Hospital B (6-anonymous)

- 4-anonymity + 6-anonymity $\not\Rightarrow$ k-anonymity, for any k

Desiderata for a Privacy Definition

1. Resilience to background knowledge
 - A privacy mechanism must be able to protect individuals' privacy from attackers who may possess background knowledge
2. Privacy without obscurity
 - Attacker must be assumed to know the algorithm used as well as all parameters [MK15]
3. Post-processing
 - Post-processing the output of a privacy mechanism must not change the privacy guarantee [KL10, MK15]
4. Composition over multiple releases
 - Allow a graceful degradation of privacy with multiple invocations on the same data [DN03, GKS08]

Why Composition?

- Reasoning about privacy of a complex algorithm is hard.
- Helps software design
 - If building blocks are proven to be private, it would be easy to reason about privacy of a complex algorithm built entirely using these building blocks.



Dinur Nissim Result [DN03]

- A vast majority of records in a database of size n can be reconstructed when $n \log(n)^2$ queries are answered by a statistical database ...

... even if each answer has been arbitrarily altered to have up to $o(\sqrt{n})$ error

A Bound on the Number of Queries

- In order to ensure utility, a statistical database must leak some information about each individual
- We can only hope to bound the amount of disclosure
- Hence, there is a limit on number of queries that can be answered



Desiderata for a Privacy Definition

1. Resilience to background knowledge

- A privacy mechanism must be able to protect individuals' privacy from attackers who may possess background knowledge

2. Privacy without obscurity

- Attacker must be assumed to know the algorithm used as well as all parameters [MK15]

3. Post-processing

- Post-processing the output of a privacy mechanism must not change the privacy guarantee [KL10, MK15]

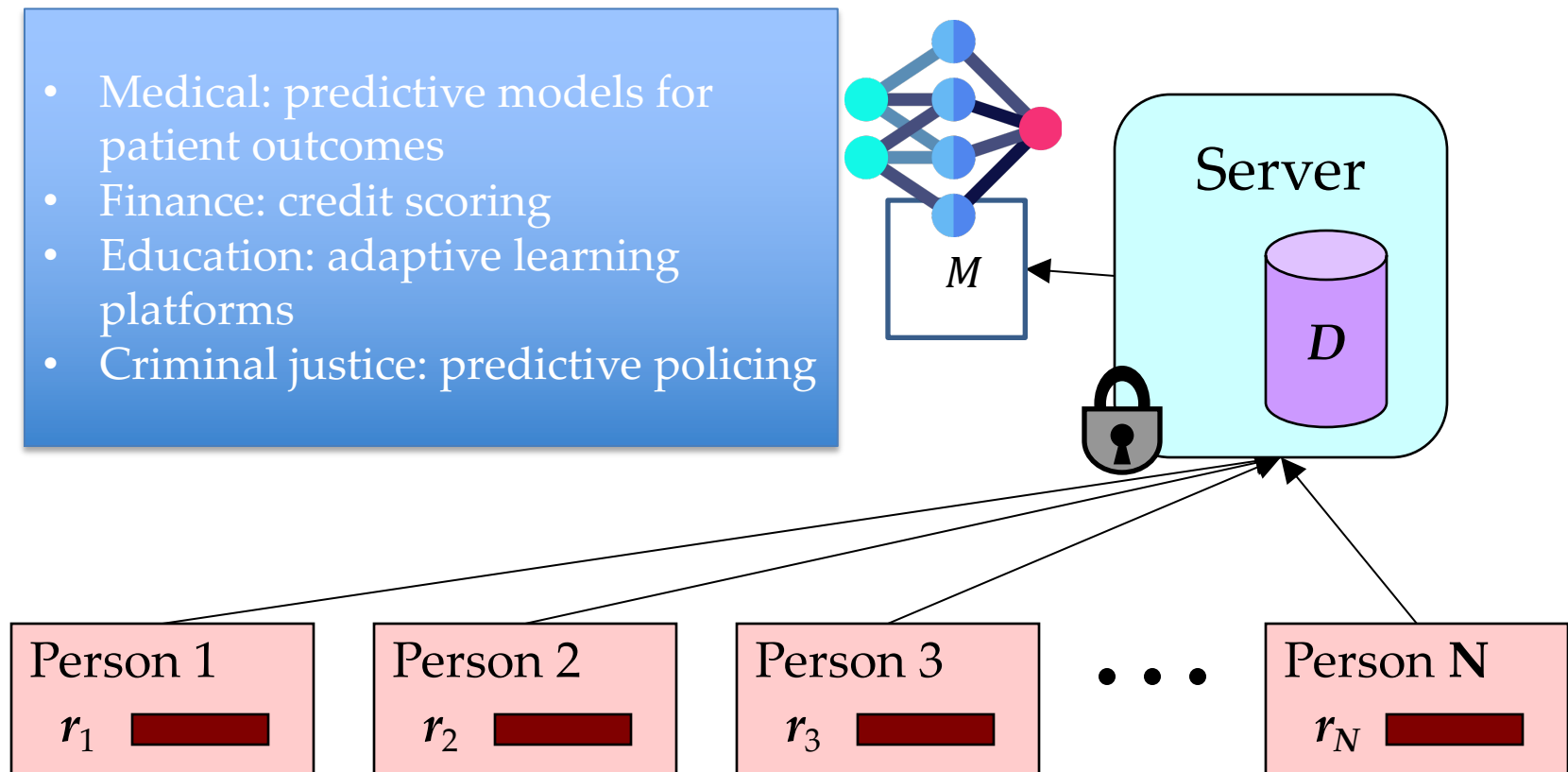
4. Composition over multiple releases

- Allow a graceful degradation of privacy with multiple invocations on the same data [DN03, GKS08]

Membership inference attacks (MIAs)

4. PRIVACY RISKS IN ML

Train Model on Sensitive Data



Privacy Risks in ML

- Membership Inference
 - From $f(x)=y$, determine whether or not x is in training set
- Model Inversion
 - Reconstruct the training data
- Training Example Extraction
 - Extract some # training samples (e.g., 1% training samples)
- Attribute Inference
 - Infer sensitive attribute (e.g., race)
- Model Extraction
 - From public data, train $f' \sim f$

Model Inversion [Fredrikson et al., CCS 2015]



**Original face from
training data**

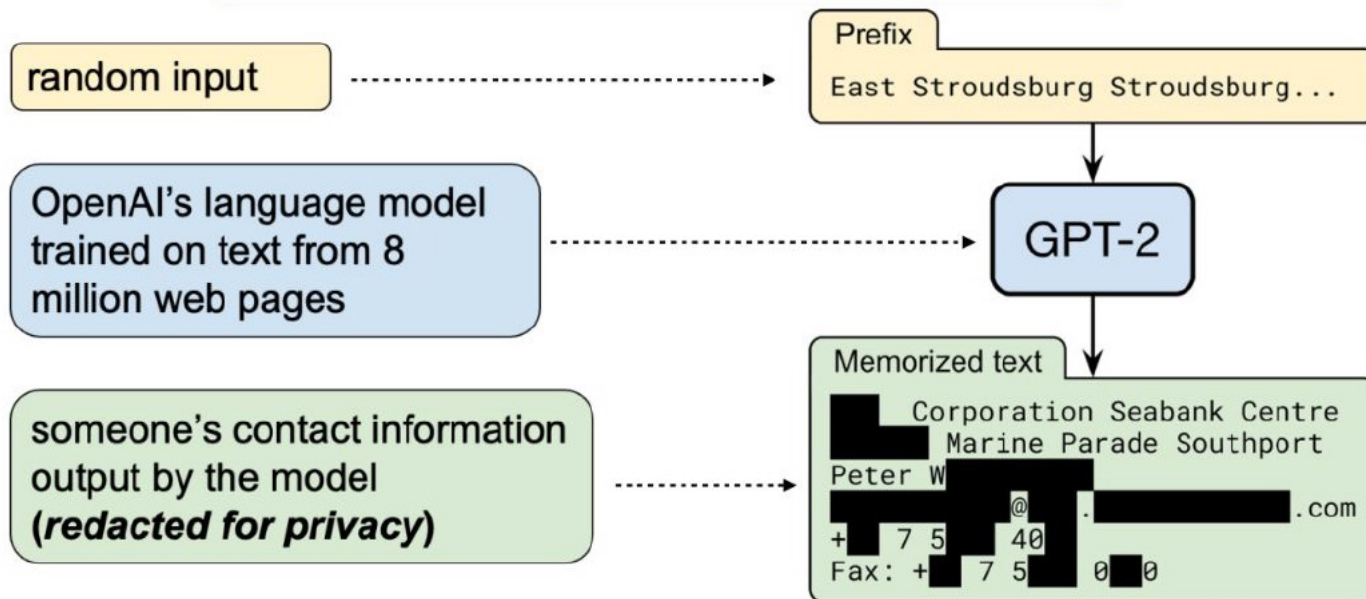


**Reconstructed face
after inversion attack**

“The attacker is given only the person’s name and access to a facial recognition system that returns a class confidence score”

Training Data Inference from LLM

[Carlini et al., USENIX2021]



“Given query access to a neural network language model, we extract an individual person’s name, email address, phone number, fax number, and physical address. The example in this figure shows information that is all accurate so we redact it to protect privacy.”

Training Data Extraction from Diffusion Models

[Carlini et al., USENIX 2023]

Training Set



Caption: Living in the light with Ann Graham Lotz

Generated Image



Prompt: Ann Graham Lotz

“Diffusion models memorize individual training examples and generate them at test time.

Left: an image from Stable Diffusion’s training set (licensed CC BY-SA 3.0).

Right: a Stable Diffusion generation when prompted with “Ann Graham Lotz”. The reconstruction is nearly identical”

Original:

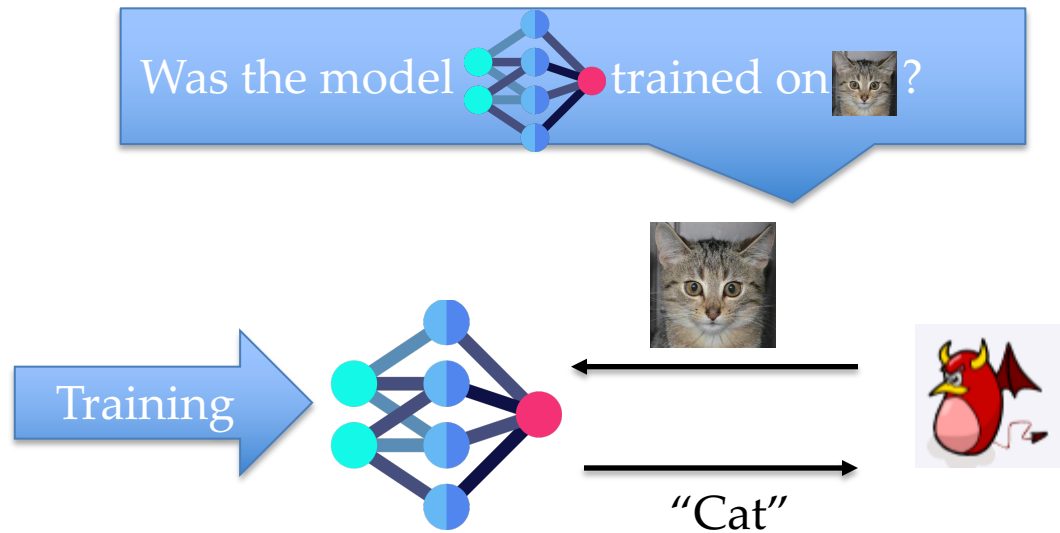
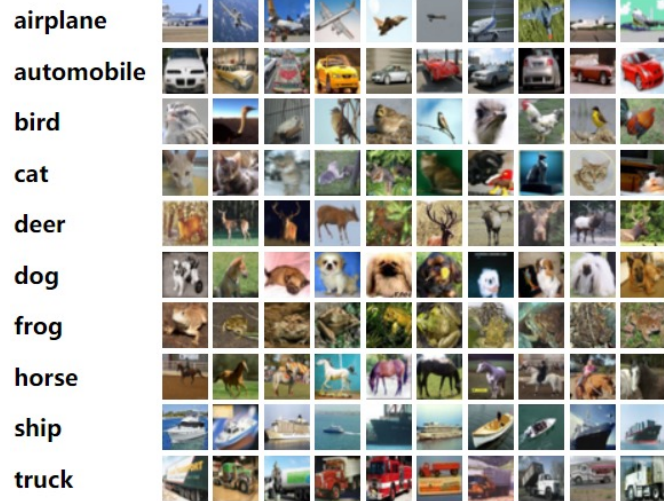


Generated:



Membership Inference Attack

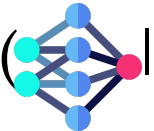


[Shokri et al. SP2017]



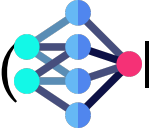


Why Care About MIA?

- **Curiosity** (e.g., did hospital x use my data?)
- **Gather intelligence** for future attacks
- **As building block** for data extraction attacks
- **Auditing** correctness of privacy mechanisms

Why is MIA possible?

Confidence( |  is from the training data of )

>

Confidence( |  is not from the training data of )

- Low-hanging fruit:
 - Statistical distinguishability of model's confidence on members vs. non-members
 - Root cause: overfitting --- models are more confident on members of their training set than on non-members

Threat Model Considerations: Adversary

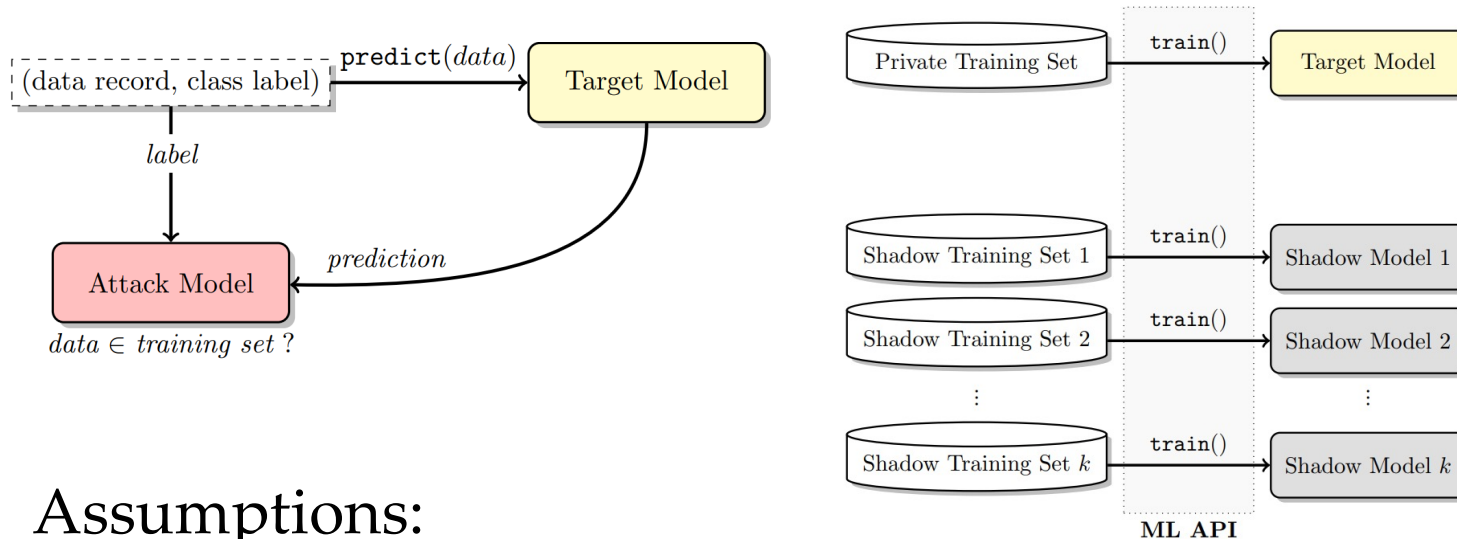
- Probability-based:
 - Has access to some set of samples from same distribution as training data
 - Oracle access to prediction API with label probability scores and predicted label
 - Model architecture (e.g., MLaaS provider published model details)
- Label Only:
 - Oracle access with only label access
 - Can do augmentation and/or perturbation of target inputs to observe model's sensitivity

Threat Model Considerations: Defender

- MLaaS API exposes only predictions, not intermediate activations
- Adversary can't compute gradients through the model
- MLaaS API is rate-limited: e.g., adversary can't make more than 5K queries/hour
- Adversary can't perform more than \$50K worth of compute (limits model extraction adversaries)

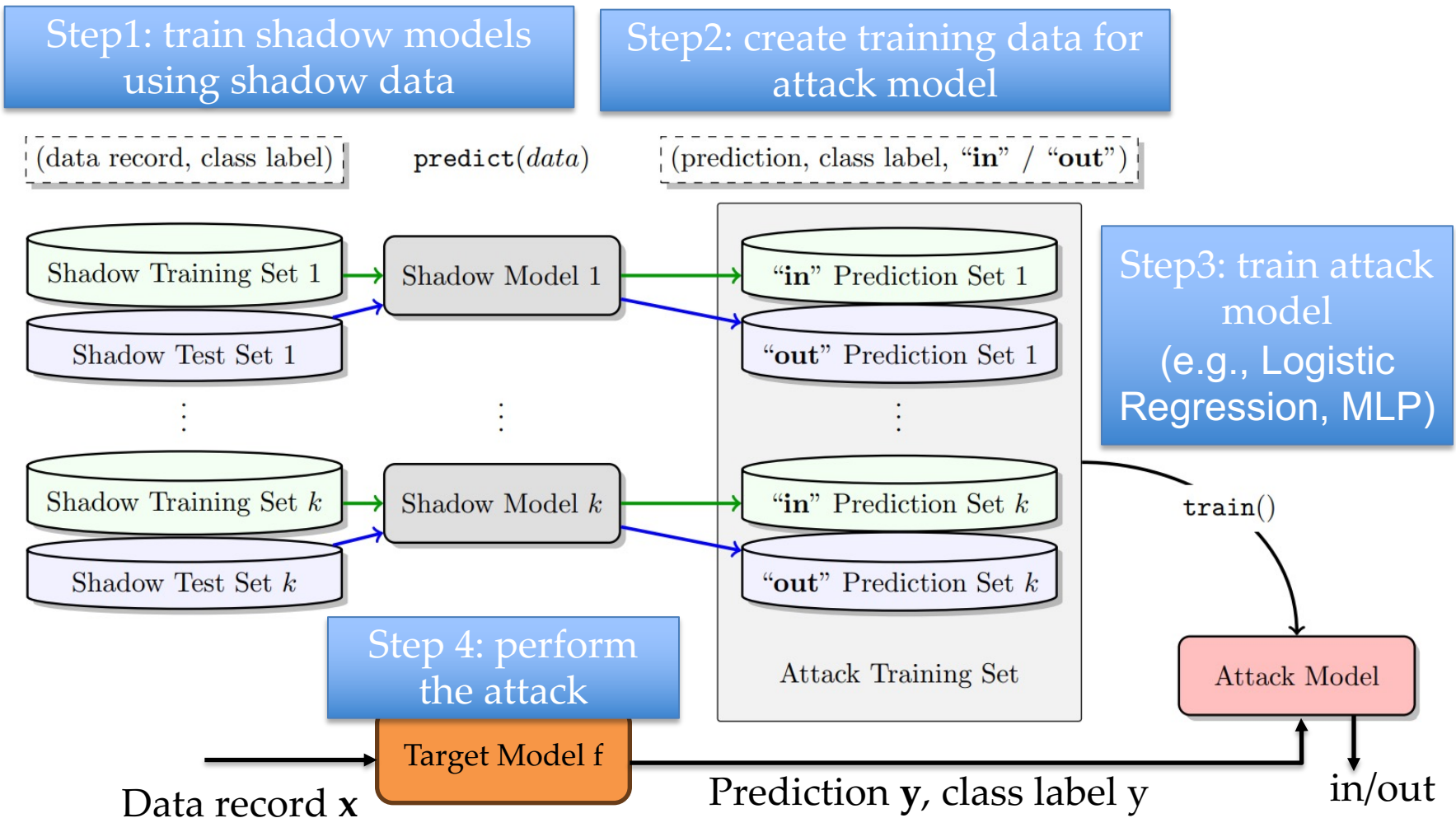
The Shadow Model Attack

- Goal: training a model to infer whether or not a data point is in the training set of the target model.



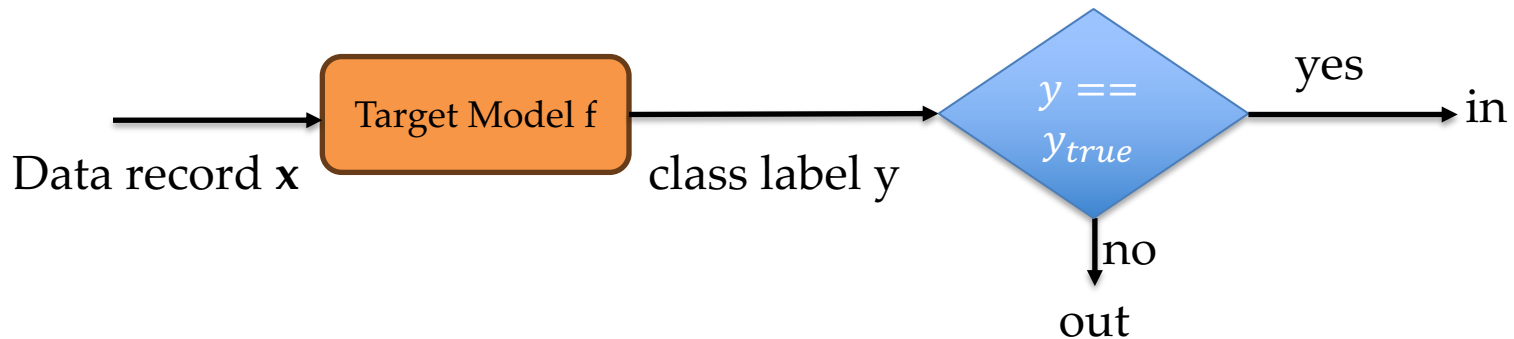
- Assumptions:
 - Attacker has access to a **shadow dataset** that comes from same distribution as the training set
 - Attacker has probability distributions of each prediction

The Shadow Model Attack



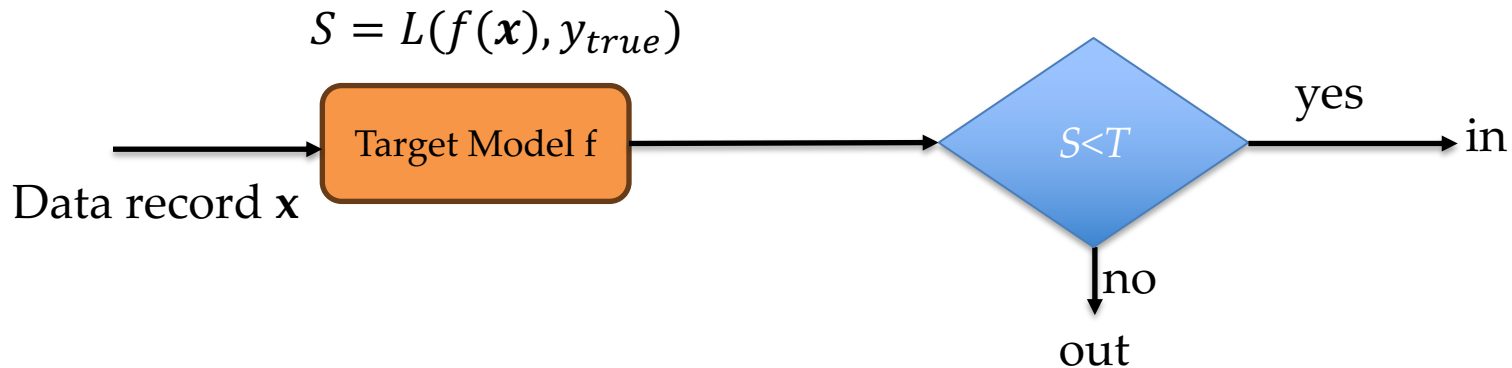
Prediction Correctness-based Attack

- **Intuition:** an overfit model makes more mistakes on non-members than members.



Prediction Loss-based Attack

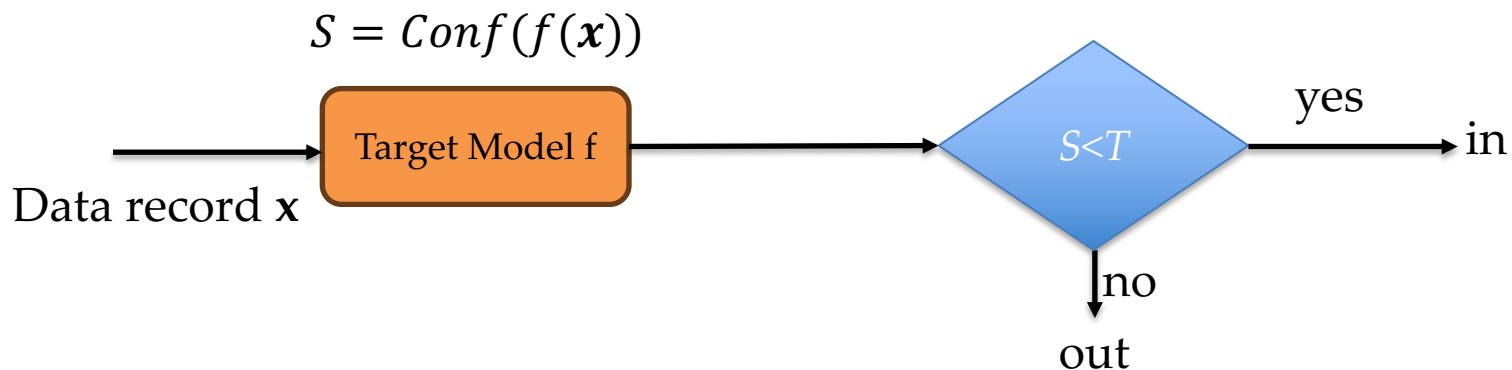
- Assign a score to a data-point x using the target model's loss as S
- T : can be estimated as the average of maximum loss of f on training data (which could be public information when reported by the model owner)



- **Intuition:** models are trained to minimize loss and can achieve zero loss for training data.

Prediction Confidence-based Attack

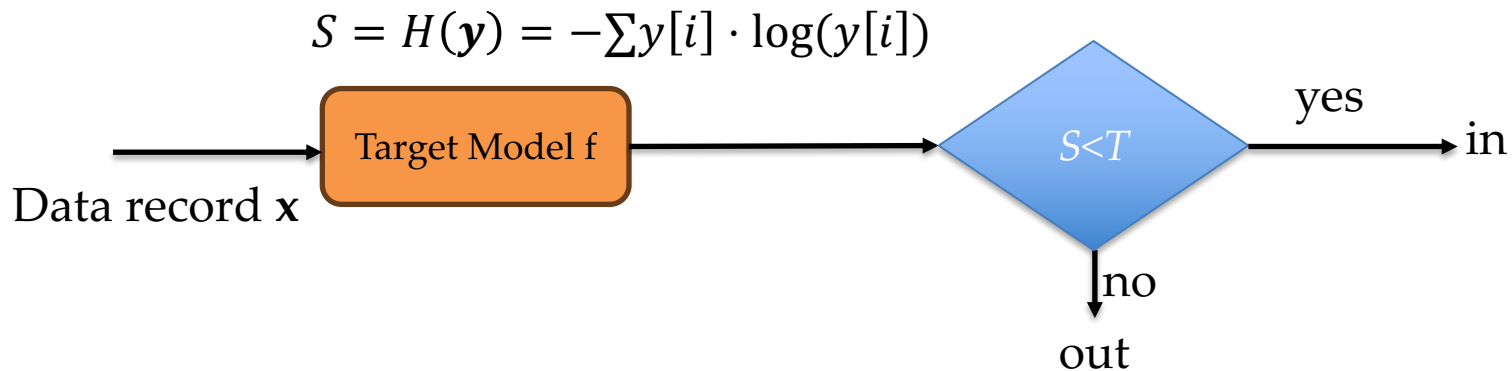
- Assign a score to a data-point x based on the model's confidence in its predicted class
- T can be chosen based on the average/minimum on training data if available or can be determined on observations from a batch of predictions



- **Intuition:** models are often more confident on training examples (even when predictions are incorrect)

Prediction Entropy-based Attack

- Assign a score to a data-point x using the entropy of the probability distribution of the classes as: $H(f(x) = y[1], \dots, y[k])$ for k # of classes
- T can be estimated as the average entropy f on training data if available or can be determined based on an attacker-supplied batch of predictions



- **Intuition:** models are often more confident on training examples (even when predictions are incorrect)

Summary of Privacy for ML

- Privacy risks against ML:
 - membership inference, training data extraction, sensitive information reconstruction, sensitive attribute inference, ...
- Threat models:
 - probability-based, label-only
- Membership inference attack:
 - shadow model attack, threshold-based attacks, ...

Module 1: Summary

- Privacy attacks on naïve approaches, anonymized data, ML models.
- Privacy desiderata include resilience to background knowledge, privacy without obscurity, closure under post-processing, and composition.
- Next, how to define privacy and design privacy-preserving mechanism that achieve these desiderata?
 - Differential Privacy
 - Basic Algorithms and Composition

References

- [S02] Sweeney, “K-anonymity”, IJFUKS 2010
- [LT08] Liu and Terzi, “Towards Identity Anonymization on Graphs”, SIGMOD 2008
- [ZP08] Zhou and Pei, “Preserving Privacy in Social Networks Against Neighborhood Attacks”, ICDE 2008
- [HMJTW08] Hay et al, “Resisting Structural Reidentification Anonymized Social Networks”, VLDB 2008
- [PAPL06] Pang et al , “The devil and packet trace anonymization”, SIGCOMM CCR 2006
- [VSJO13] Vaidya et al., “Identifying inference attacks against healthcare data repositories”, AMIA 2013
- [GKS08] Ganta et al. “Composition Attacks and Auxiliary Information in Data Privacy”, KDD 2008
- [DN03] Dinur, Nissim, “Revealing information while preserving privacy”, PODS 2003
- [KL10] Kifer, Lin, “Towards an Axiomatization of Statistical Privacy and Utility.”, PODS 2010
- [MK15] Machanavajjhala, Kifer, “Designing statistical privacy for your data”, CACM 2015
- Xi He’s previous course on “Privacy & Fairness in Data Science”
- Ashwin Machnanavajjhala’s course on “Privacy in a Mobile-Social World”
- Birhanu Eshete’s lecture on “Membership inference attacks”