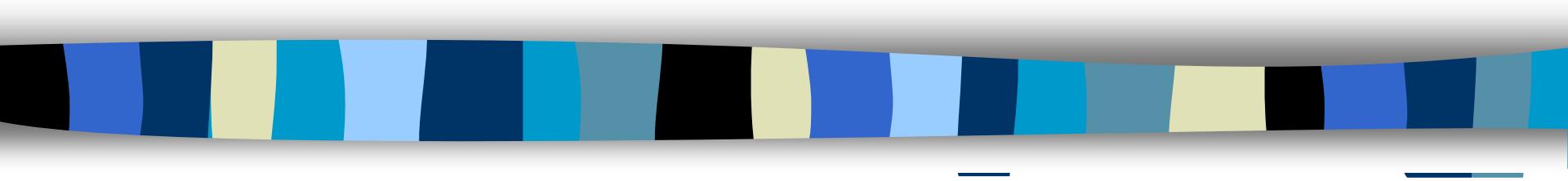
Information Security
CS 526



Topic 21: Data Privacy

What is Privacy?

- Privacy is the protection of an individual's personal information.

- Privacy is the rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure and disposal of personal information.

- Privacy ≠ Confidentiality

OECD Privacy Principles

- **1. Collection Limitation Principle**
  - There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

- **2. Data Quality Principle**
  - Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

OECD Privacy Principles

- **3. Purpose Specification Principle**
  - The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

- **4. Use Limitation Principle**
  - Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3 except:
  - a) with the consent of the data subject; or
  - b) by the authority of law.

OECD Privacy Principles

- **5. Security Safeguards Principle**
  - Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

- **6. Openness Principle**
  - There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

OECD Privacy Principles

- **7. Individual Participation Principle**
  - An individual should have the right:
  - a) to request to know whether or not the data controller has data relating to him;
  - b) to request data relating to him, …
  - c) to be given reasons if a request is denied; and
  - d) to request the data to be rectified, completed or amended.

- **8. Accountability Principle**
  - A data controller should be accountable for complying with measures which give effect to the principles stated above.

Areas of Privacy

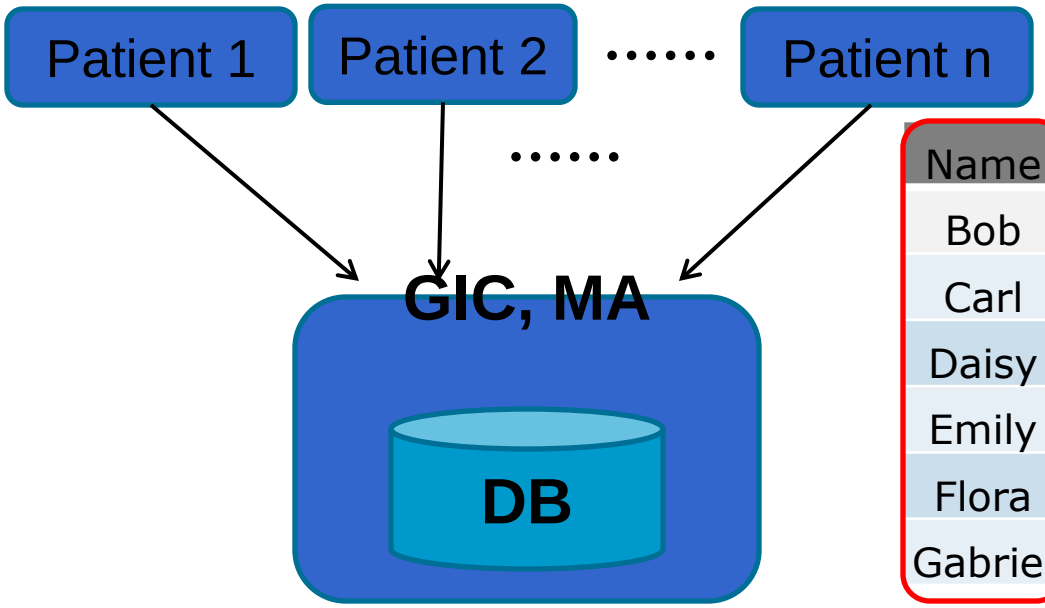
- Anonymity
  - Anonymous communication:
    - e.g., The TOR software to defend against traffic analysis
- Web privacy
  - Understand/control what web sites collect, maintain regarding personal data
- Mobile data privacy, e.g., location privacy
- Privacy-preserving data usage

Privacy Preserving Data Sharing

- The need to sharing data
  - For research purposes
    - E.g., social, medical, technological, etc.
  - Mandated by laws and regulations
    - E.g., census
  - For security/business decision making
    - E.g., network flow data for Internet-scale alert correlation
  - For system testing before deployment
  - …
- However, publishing data may result in privacy violations

GIC Incidence [Sweeny 2002]

- Group Insurance Commissions (GIC, Massachusetts)
  - Collected patient data for ~135,000 state employees.
  - Gave to researchers and sold to industry.
  - Medical record of the former state governor is identified.



Name	Age	Sex	Zip code	Disease
Bob	69	M	47906	Cancer
Carl	65	M	47907	Cancer
Daisy	52	F	47902	Flu
Emily	43	F	46204	Gastritis
Flora	42	F	46208	Hepatitis
Gabriel	47	F	46203	Bronchitis

Re-identification occurs!

AOL Data Release [NYTimes 2006]

- In August 2006, AOL Released search keywords of 650,000 users over a 3-month period.
  - User IDs are replaced by random numbers.
  - 3 days later, pulled the data from public access.

AOL searcher # 4417749

"landscapers in Lilburn, GA"
queries on last name "Arnold"
"homes sold in shadow lake subdivision Gwinnett County, GA"
"num fingers"
"60 single men"
"dog that urinates on everything"

NYT

Thelman Arnold, a 62 year old widow who lives in Liburn GA, has three dogs, frequently searches her friends' medical ailments.

Re-identification occurs!

Netflix Movie Rating Data [Narayanan and Shmatikov 2009]

- Netflix released anonymized movie rating data for its Netflix challenge
  - With date and value of movie ratings
- Knowing 6-8 approximate movie ratings and dates is able to uniquely identify a record with over 90% probability
  - Correlating with a set of 50 users from imdb.com yields two records
- Netflix cancels second phase of the challenge

Re-identification occurs!

Genome-Wide Association Study (GWAS) [Homer et al. 2008]

- A typical study examines thousands of singe-nucleotide polymorphism locations (SNPs) in a given population of patients for statistical links to a disease.
- From aggregated statistics, one individual's genome, and knowledge of SNP frequency in background population, one can infer participation in the study.
  - The frequency of every SNP gives a very noisy signal of participation; combining thousands of such signals give high-confidence prediction

Membership disclosure occurs!

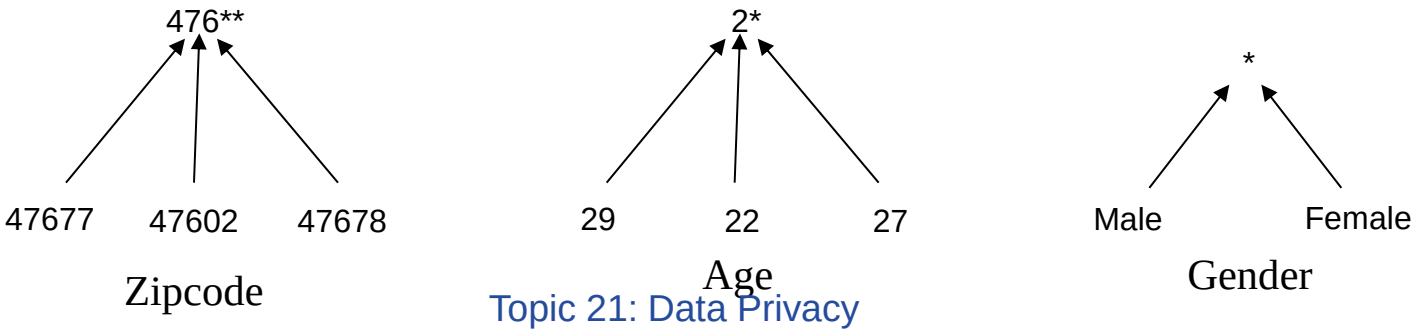
	Study group Avg	Population Avg	Target individual
SNP 1=A	43%	42%	yes
SNP 2=A	11%	11%	no
SNP 3=A	58%	59%	no
SNP 4=A	23%	22%	yes
…			
…			

Need for Data Privacy Research

- Identification Disclosure (GIC, AOL, Netflix)
  - Leaks the subject individual of one record
- Attribute Disclosure
  - leaks more precise information about the attribute values of some individual
- Membership Disclosure (GWAS)
  - leaks an individual's participation is in the dataset
- Research Program: Develop theory and techniques to anonymize data so that they can be beneficially used without privacy violations.
- How to define privacy for anonymized data?
- How to publish data to satisfy privacy while providing utility?

k-Anonymity [Sweeney, Samarati]

- Privacy is "protection from being **brought to the attention of others**."
- k-Anonymity
  - Each record is indistinguishable from ≥ k-1 other records when only "quasi-identifiers" are considered
  - These k records form an equivalence class
- To achieve k-Anonymity, uses
  - Generalization: Replace with less-specific values
  - Suppression: Remove outliers



Example of *k*-Anonymity & Generalization

The Microdata

QID			SA
Zipcode	Age	Gen	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

The Generalized Table

QID			SA
Zipcode	Age	Gen	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

- 3-Anonymous table
  - The adversary knows Alice's QI values (47677, 29, F)
  - The adversary does not know which one of the first 3 records corresponds to Alice's record.

Attacks on *k*-Anonymity

- k-anonymity does not provide privacy if:
  - Sensitive values lack diversity
  - The attacker has background knowledge

Homogeneity Attack

Bob	
Zipcode	***Age***
47678	27

Background Knowledge Attack

Carl does not have heart disease

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

l–Diversity: [Machanavajjhala et al. 2006]

- Principle
  - Each equi-class contains at least l well-represented sensitive values
- Instantiation
  - Distinct l-diversity
    - Each equi-class contains *l* distinct sensitive values
  - Entropy l-diversity
    - *entropy(equi-class)*≥*log*$_2$(l)

$$H(X) = \mathrm{E}(I(X)) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$$

The Skewness Attack on *l*-Diversity

- Two values for the sensitive attribute
  - HIV positive (1%) and HIV negative (99%)
- Highest diversity still has serious privacy risk
  - Consider an equi-class that contains an equal number of positive records and negative records.
- l-diversity does not differentiate:
  - Equi-class 1: 49 positive + 1 negative
  - Equi-class 2: 1 positive + 49 negative

l-diversity does not consider the overall distribution of sensitive values

The Similarity Attack on *l*-Diversity

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

Conclusion

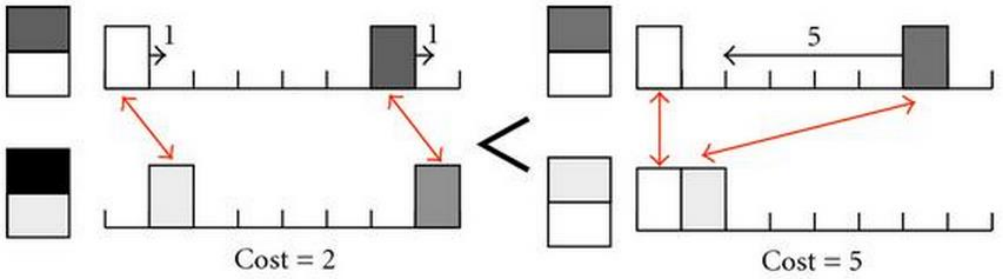
1. Bob's salary is in [20k,40k], which is relative low.
2. Bob has some stomach-related disease.

l-diversity does not consider semantic meanings of sensitive values

t-Closeness

- Principle: Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (distance $\leq$ t)
  - Assuming that publishing a completely generalized table is always acceptable
  - We use Earth Mover Distance to capture semantic relationship among sensitive attribute values
- **(n,t)-closeness:** Distribution of sensitive attribute value in each equi-class should be close to that of some natural super-group consisting at least n tuples

N. Li, T. Li, S. Venkatasubramanian: t-Closeness: Privacy Beyond k-Anonymity and *l*-diversity. In ICDE 2007. Journal version in TKDE 2010.



From Syntactical Privacy Notions to Differential Privacy

- Limitation of previous privacy notions:
  - Requires identifying which attributes are quasi-identifier or sensitive, not always possible
  - Difficult to pin down due to background knowledge
  - Syntactic in nature (property of anonymized dataset)
    - Not exhaustive in inference prevented
- Differential Privacy [Dwork et al. 2006]
  - Privacy is not violated is one's information is not included
  - Output does not overly depend on any single tuple

Definition (ε-Differential Privacy)

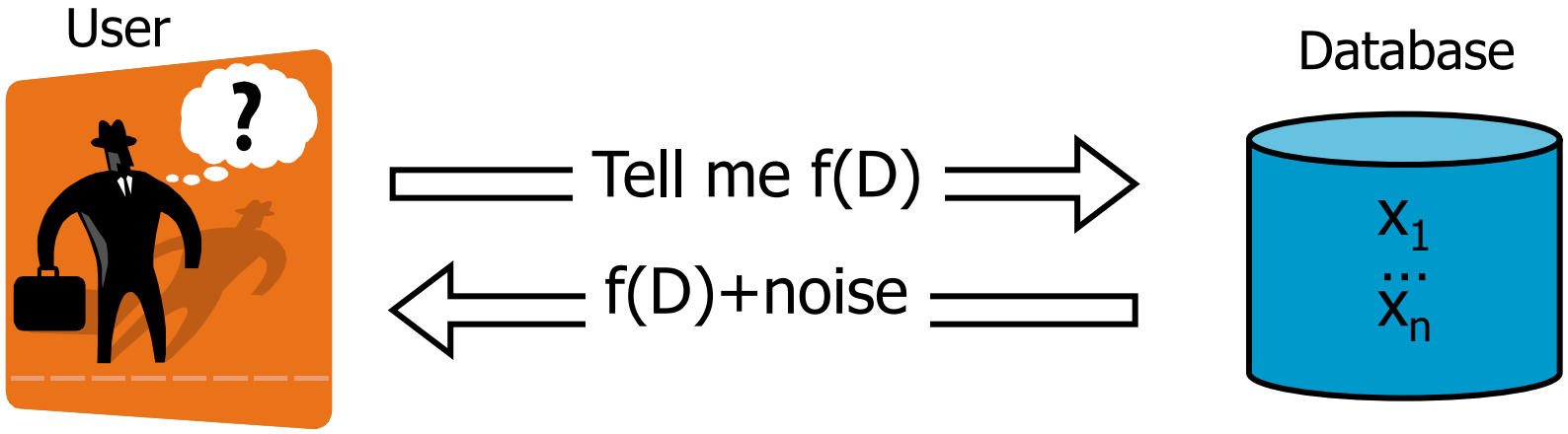
A randomized algorithm $\mathcal{A}$ satisfies ε-differential privacy, if for any pair of neighboring datasets D and D' and for any $O \subseteq \text{Range}(\mathcal{A})$:

$$e^{-\varepsilon} Pr[\mathcal{A}(D') \in O] \leq Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in O]$$

Variants of Differential Privacy

- Bounded Differential Privacy: D and D' are neighbors if and only D' can be obtained from D by replacing one tuple with another tuple
  - D and D' have the same number of tuples
  - Revealing size of dataset does not affect privacy
- Unbounded Differential Privacy: D and D' are neighbors if and only D' can be obtained from D by adding or removing one tuple
  - The numbers of tuples in D and D' differ by 1
- In most cases, can use either one.
- Other definitions of neighboring datasets have also been considered

One Primitive to Satisfy Differential Privacy: Add Noise to Output



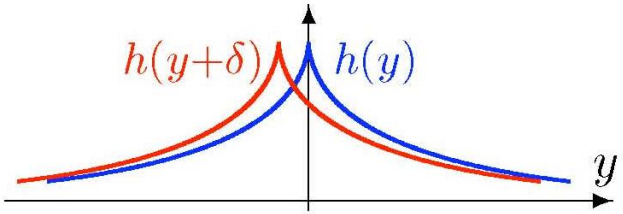
- Intuition: f(D) can be released accurately when f is insensitive to individual entries x_1, … x_n
- Global sensitivity $GS_f = \max_{\text{neighbors } D,D'} ||f(D) - f(D')||_1$
  - Example: $GS_{average}$ = 1/n for sets of numbers between 0 and 1
- Theorem: $f(x) + Lap(GS_f / \varepsilon)$ is ε-indistinguishable
  - Noise generated from Laplace distribution

Sensitivity with Laplace Noise

Theorem

If $A(x) = f(x) + \mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$ *then* A *is* ε*-indistinguishable.*

Laplace distribution $\mathsf{Lap}(\lambda)$ has density $h(y) \propto e^{-\frac{\|y\|_1}{\lambda}}$



Sliding property of $\mathsf{Lap}\left(\frac{\mathsf{GS}_f}{\varepsilon}\right)$: $\frac{h(y)}{h(y+\delta)} \leq e^{\varepsilon \cdot \frac{\|\delta\|}{\mathsf{GS}_f}}$ for all y, δ

Proof idea:

$A(x)$: blue curve

$A(x')$: red curve

$\delta = f(x) - f(x') \leq \mathsf{GS}_f$

Another Primitive for Satisfying Differential Privacy: Exponential Mechanism

- The goal is to output f(D); $f(D) \in R$
  - E.g., which item is purchased the most frequently
- Define a quality function $q(D, r \in R)$
  - which gives a real number describing the desirability of outputting r on input dataset D
- Compute the sensitivity of the quality function
  - $\Delta q = \max_r \max_{D,D'} |q(D,r) - q(D',r)|$
- Returns r with probability proportional to $\exp(q(D,r) / 2\varepsilon \Delta q)$ satisfies ε-DP

Properties of Differential Privacy

- Composability
  - If A_1 satisfies ε_1-DP, and A_2 satisfies ε_2-DP, then outputting both A_1 and A_2 satisfies $(\varepsilon_1+\varepsilon_2)$-DP
- Some queries, such as counting queries, can be answered relatively accurately
  - Since one tuple affects the result by at most 1
  - A small amount of noise (following the Laplace distribution) can be added to achieve DP
- Some queries are hard to answer
  - E.g., max, since it can be greatly affected by a single tuple
- Challenge in using it
  - Find suitable queries to ask so that noisy answers provide most utility

Relaxing Differential Privacy

- Satisfying differential privacy can result in too much distortion in many applications
  - Several efforts to relax differential privacy
- Differential Privacy means that one cannot distinguish between $D \cup \{t\}$ and D.
  - With precise knowledge of D and t
- Relaxation: cannot distinguish between $D \cup \{t\}$ and D, with precise knowledge of t, but only statistical knowledge of D
  - I.e., the setting of GWAS
  - GIC, AOL, Netflix requires only precise knowledge of t

Differential Privacy under Sampling

Definition

[(ε, δ)-Differential Privacy ((ε, δ)-DP)]: A randomized algorithm $\mathcal{A}$ satisfies (ε, δ)-differential privacy, if for any pair of neighboring datasets D and D' and for any $O \subseteq \text{Range}(\mathcal{A})$:

$$Pr[\mathcal{A}(D) \in O] \leq e^{\varepsilon} Pr[\mathcal{A}(D') \in O] + \delta$$

Definition (Differential privacy under sampling)

An algorithm $\mathcal{A}$ gives $(\beta, \varepsilon, \delta)$-DPS if and only if the algorithm $\mathcal{A}^{\beta}$ gives (ε, δ)-DP, where $\mathcal{A}^{\beta}$ denotes the algorithm to first sample with probability β (include each tuple in the input dataset with probability β), and then apply $\mathcal{A}$ to the sampled dataset.

N. Li, W. Qardaji, D. Su: On sampling, anonymization, and differential privacy: or, *k*-anonymization meets differential privacy. ASIACCS 2012:

The Amplification Effect of Sampling

- A smaller sampling rate can achieve a stronger privacy protection

Theorem

Any algorithm that satisfies $(\beta_1, \varepsilon_1, \delta_1)$-DPS *also satisfies* $(\beta_2, \varepsilon_2, \delta_2)$-DPS *for any* $\beta_2 < \beta_1$, *where*
$\varepsilon_2 = \ln\left(1 + \left(\frac{\beta_2}{\beta_1}(e^{\varepsilon_1} - 1)\right)\right)$, *and* $\delta_2 = \frac{\beta_2}{\beta_1}\delta_1$.

β	e^{ε}	ε	δ
1	11	$\ln 11 \approx 2.40$	10^{-5}
0.1	2	$\ln 2 \approx 0.69$	10^{-6}
0.01	1.1	$\ln 1.1 \approx 0.095$	10^{-7}

β	ε
1	1
0.1	0.159
0.01	0.017

Safe k-Anonymization Meets Differential Privacy

- k-Anonymization has two steps
  1. Output how the domain is partitioned
     - strongly safe if this step does not depend on the dataset
  2. Output how many tuples there are in each cell, outputting 0 when a partition contains fewer than k tuples

Theorem

Any strongly-safe k-anonymization algorithm satisfies $(\beta, \varepsilon, \delta)$-DPS *for any* $0 < \beta < 1$, $\varepsilon \geq -\ln(1-\beta)$, *and* $\delta = d(k, \beta, \varepsilon)$, *where the function* d *is defined as*

$$d(k, \beta, \varepsilon) = \max_{n : n \geq \left\lceil \frac{k}{\gamma} - 1 \right\rceil} \sum_{j > \gamma n}^{n} f(j; n, \beta),$$

where $\gamma = \frac{(e^{\varepsilon} - 1 + \beta)}{e^{\varepsilon}}$.

Towards Membership Privacy

- A series of papers challenging differential privacy
  - Differential privacy is not robust to arbitrary background knowledge (Kifer and Machanavajjhala)
  - Difficult to choose ε in DP, propose differential identifiability (Lee and Clifton)
  - Differential privacy does not prevent attribute disclosure (Cormode)

- Needs a framework to examine these privacy notions
  - Privacy violation = positive membership disclosure
  - No membership disclosure means no attribute disclosure and no re-idenification disclosure

Intuition for Membership Privacy

- Adversary has some prior belief about what the input dataset (a prob. distribution over all possible datasets)
  - Gives the prior probability of any t's membership
- Adversary updates belief after observing output of the algorithm, via Bayes rule
  - Obtains posterior probability of any t's membership
- For any t, posterior belief should not change too much from prior
  - Whether this holds depends on the prior distribution
- Membership privacy is relative to the family of prior distributions the adversary is allowed to have

The Membership Privacy Framework

Definition (Positive Membership Privacy ($(\mathbb{D}, \gamma)$-PMP))

We say that a mechanism $\mathcal{A}$ provides γ-positive membership privacy under a family $\mathbb{D}$ of distributions over $2^{\mathcal{U}}$, i.e., $((\mathbb{D}, \gamma)$-PMP), where $\gamma \geq 1$, if and only if for any $S \subseteq \text{range}(\mathcal{A})$, any distribution $\mathcal{D} \in \mathbb{D}$, and any entity $t \in \mathcal{U}$, we have

$$\Pr_{\mathcal{D},\mathcal{A}}[t \in \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \leq \gamma \Pr_{\mathcal{D}}[t \in \mathbf{T}] \quad (1)$$

$$\text{and } \Pr_{\mathcal{D},\mathcal{A}}[t \notin \mathbf{T} \mid \mathcal{A}(\mathbf{T}) \in S] \geq \frac{\Pr_{\mathcal{D}}[t \notin \mathbf{T}]}{\gamma} \quad (2)$$

where $\mathbf{T}$ is a random variable drawn according to the distribution $\mathcal{D}$.

N. Li, W. Qardaji, D. Su,: Y. Wu, W. Yang: **Membership Privacy: A Unifying Framework For Privacy Definitions.** ACM CCS 2013.

Results from Membership Privacy

- Membership privacy for the family of all possible distributions is infeasible
  - Requires publishing similar output distributions for two completely different datasets
  - Output has (almost) no utility
- Moral: One has to make some assumptions about the adversary's prior belief
  - Assumptions need to be clearly specified and reasonable

Results from Membership Privacy

- Differential privacy is equivalent to membership privacy under the family of all distributions where the entities/tuples are mutually independent

- Differential privacy insufficient for membership privacy without independence assumption

Instances of Membership Privacy Notions

All distributions:
Privacy with Almost no Utility

Mutually Independent (MI) dist:
- **Pr[T] = $\Pi_{t \in T} p_t \quad \Pi_{t \notin T} (1-p_t)$**
- **Unbounded Differential Privacy**

Bounded Mutually Independent dist.:
- **MI distributions conditioned on all datasets have the same size**
- **Bounded Differential Privacy**

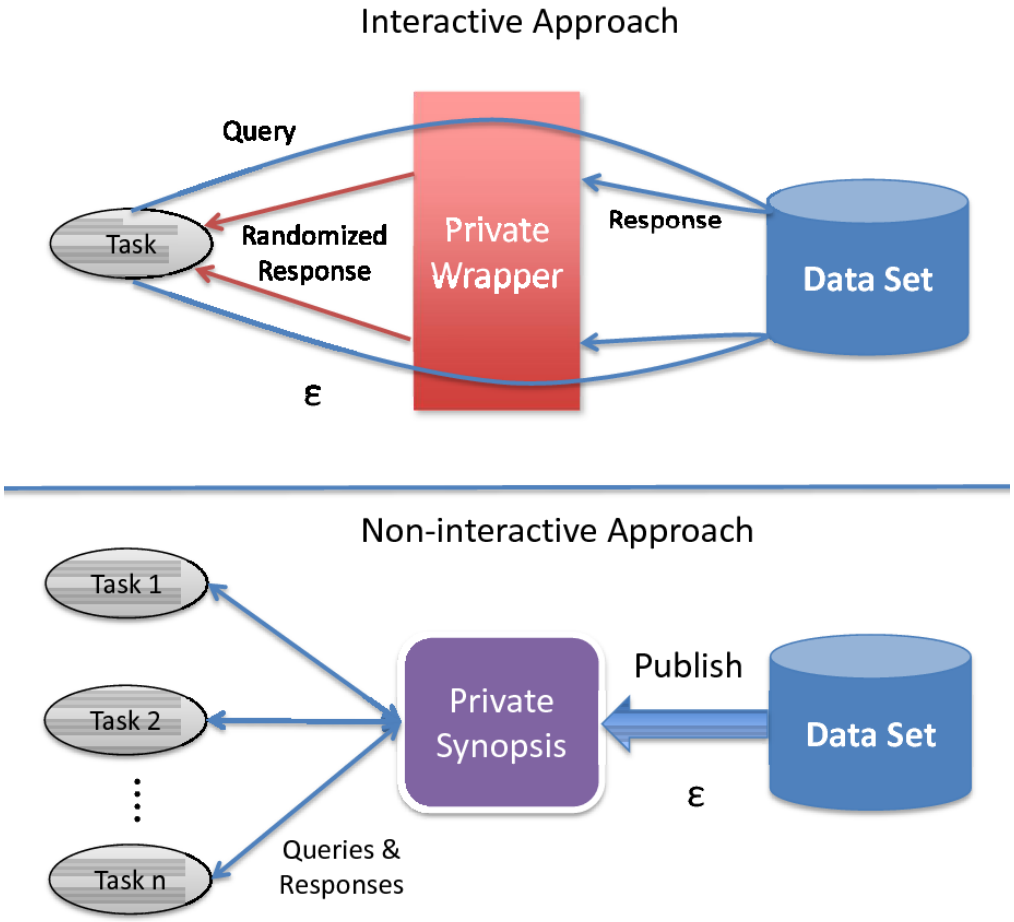
- **MI distributions where each p_t is either 0 or β**
- **Differential Privacy Under Sampling**

- **MI distributions where each p_t is either 1 or 1/m, where m is number of t's have probability ≠1**
- **Differential Identifiability**

- **MI distributions where each p_t is 1/2**
- **New privacy notion**

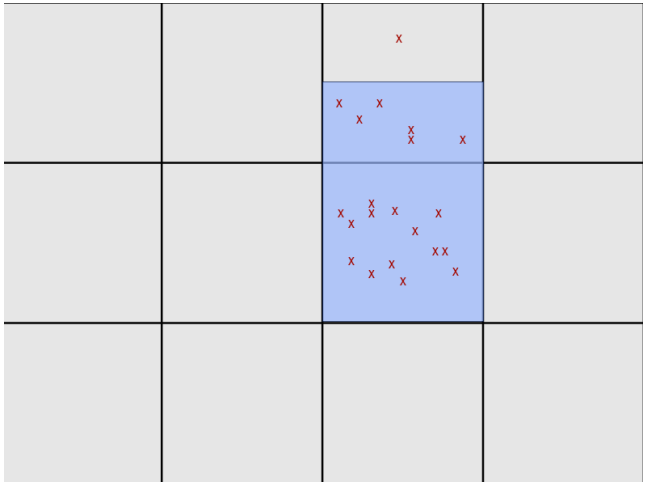
Provable Private Publishing Datasets While Preserving Utility

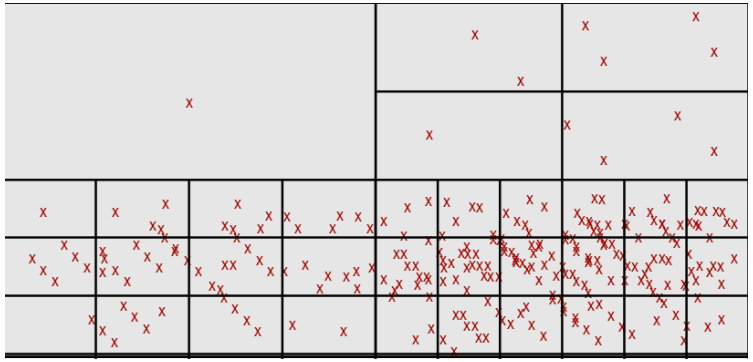
- Non-interactive data publishing rather than interactive query answering
- Diverse problem domains require different methods
  - e.g., number of dimensions
- Methodology: Combining analysis of how algorithm performs with experimental validation



An Example: Differentially Private Publishing of Geospatial Data

- Publishing synopsis of 2-D dataset while ensuring high accuracy for range queries
- Uniform Grid method:
  - Add noise to points count of each cell
  - Choose grid size to balance
    - Error due to added noise
    - Error due to non-uniformity
- Adaptive Grid method:
  - Two-level partitioning





W. Qardaji, W. Yang, N. Li: Differentially private grids for geospatial data. ICDE 2013.

Coming Attractions …

- Role-Based and Attribute-based Access Control