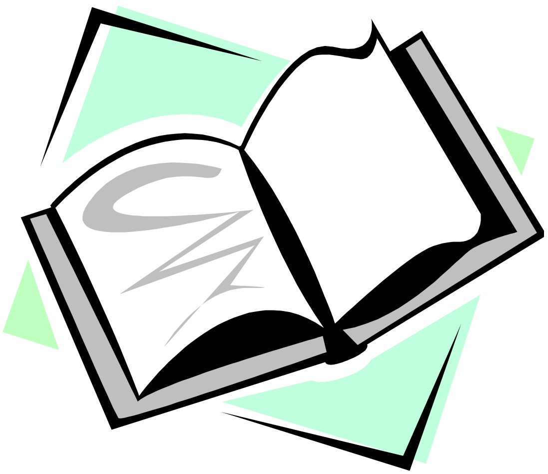
Data Security and Privacy

Topic 18: k-Anonymity, l-Diversity, and t-Closeness

Optional Readings for This Lecture

- *t*-Closeness: Privacy Beyond *k*-Anonymity and l-Diversity. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. In ICDE, April 2007.



All Kinds of Privacy Concerns

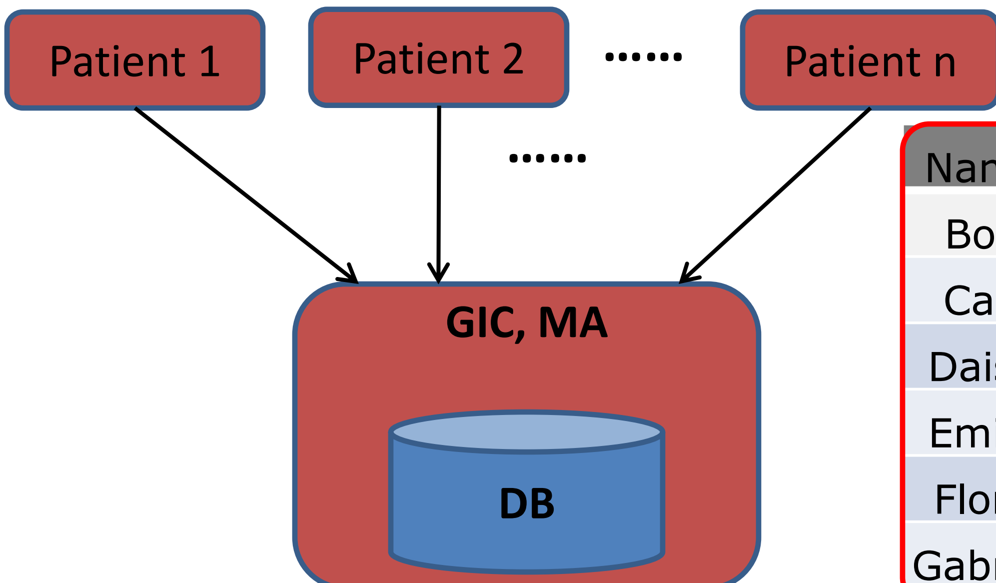
- Deciding what data to collect and why, how to use the data, and with whom to share data
- Communicate privacy policies to end users
- Ensure that data are used in ways consistent with privacy policies
- Protect collected data (security)
- Anonymity in communications
- Sharing data or using data for purposes in a way not allowed by privacy policies
  - How?

Privacy Preserving Data Sharing

- It is often necessary to share 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.

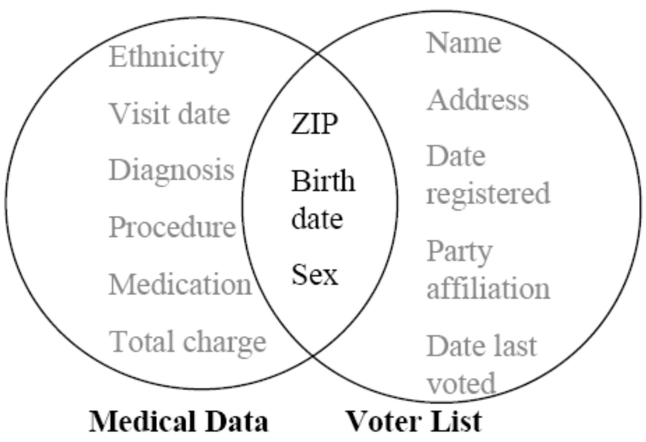


Name	DoB	Gender	Zip code	Disease
Bob	1/3/45	M	47906	Cancer
Carl	4/7/64	M	47907	Cancer
Daisy	9/3/69	F	47902	Flu
Emily	6/2/71	F	46204	Gastritis
Flora	2/7/80	F	46208	Hepatitis
Gabriel	5/5/68	F	46203	Bronchitis

Re-identification occurs!

Real Threats of Linking Attacks

- Fact: **87%** of the US citizens can be uniquely linked using only three attributes **<Zipcode, DOB, Sex>**
- Sweeney [Sweeney, 2002] managed to re-identify the medical record of the government of Massachusetts.



- Census data (income), medical data, transaction data, tax data, etc.

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

GWAS Privacy Issue

Published Data

	Disease Group Avg	Control Group Avg
SNP1=A	43%	…
SNP2=A	11%	…
SNP3=A	58%	…
SNP4=A	23%	…
…		

Adv. Info & Inference

Population Avg	Target individual Info	Target in Disease Group
42%	yes	+
10%	no	-
59%	no	+
24%	yes	-

Membership disclosure occurs!

Main Research Problems

- How to define privacy for sharing data?
- How to publish/anonymize data to satisfy privacy while providing utility?

Attempts at Defining Privacy

- Preventing the following disclosures
  - Identification disclosure
  - Attribute disclosure
  - Membership disclosure

k-Anonymity [Sweeney, Samarati]

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

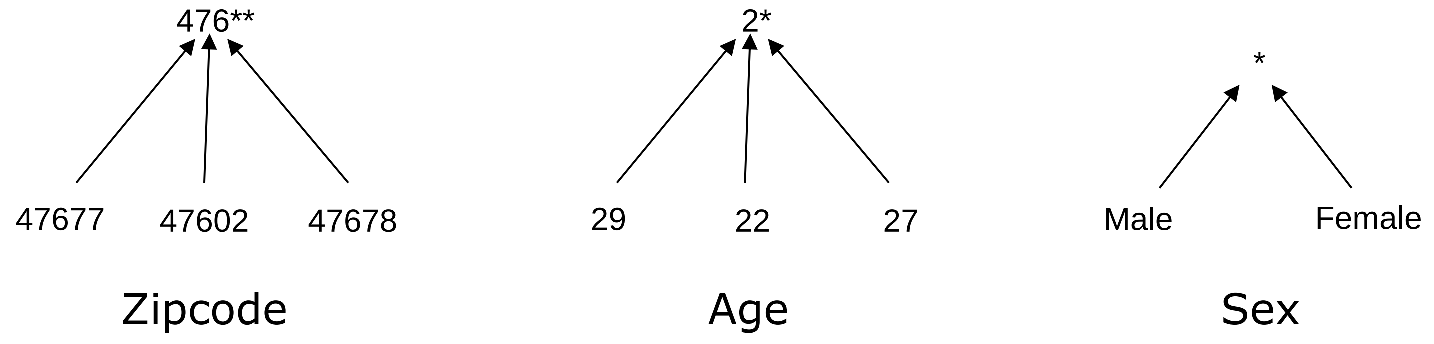
A 3-Anonymous Table

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

- k-Anonymity
  - Attributes are separated into Quasi-identifiers (QIDs) and Sensitive Attributes (SAs)
  - Each record is indistinguishable from $\geq$ k-1 other records when only "quasi-identifiers" are considered
  - These k records form an equivalence class

k-Anonymity & Generalization

- *k*-Anonymity
  - Each record is indistinguishable from at least *k-1* other records
  - These *k* records form an *equivalent class*
  - *k*-Anonymity ensures that linking cannot be performed with confidence > 1/k.
- Generalization
  - Replace with less-specific but semantically-consistent values



Data Publishing Methods

- Generalization
  - Make data less precise
- Perturbation
  - Add noise/errors
- Suppression
  - Remove certain data
- Data generating
  - Generate similar data
- Segmentation
  - Divide data up before publishing
- ???

Attacks on *k*-Anonymity

- k-anonymity does not prevent attribute disclosure 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]

- The l-diversity principle
  - Each equivalent 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)≥*$\log_2(l)$

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

Limitations of *l*-Diversity

l-diversity may be difficult and unnecessary to achieve.

- Consider a single sensitive attribute
  - Two values: HIV positive (1%) and HIV negative (99%)
  - Very different degrees of sensitivity
    - One would not mind being known to be tested negative but one would not want to be known/considered to be tested positive.
- l-diversity is unnecessary to achieve
  - 2-diversity is unnecessary for an equi-class that contains only negative records.
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total.
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equi-classes.

The Skewness Attack: An Example

- 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.
- Using diversity and entropy does not differentiate:
  - Equi-class 1: 49 positive + 1 negative
  - Equi-class 2: 1 positive + 49 negative

The overall distribution of sensitive values matters.

The Similarity Attack: An Example

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.

The semantic meanings of attribute values matters.

How to Prevent These Attacks?

- Goal is to quantify/limit amount of information leakage through data publication.

- Looking only at the final output is inherently problematic because it cannot measure information gain.

Our Main Insight

- Revealing the overall distribution of the sensitive attribute in the whole dataset should be considered to have no privacy leakage (is an ideal world for privacy)
  - In other words, we assume that removing all quasi-identifier attributes preserves privacy
  - Seems unavoidable unless willing to destroy utility
  - Also seems desirable from utility perspective
- Goal is to simulate this ideal world.

t-Closeness [Li et al. 2007]

- Rationale

Belief	Knowledge
B_0	External Knowledge
B_1	Overall distribution Q of sensitive values
B_2	Distribution P_i of sensitive values in each equi-class

A completely generalized table

Age	Zipcode	……	Gender	Disease
2*	479**	……	Male	Flu
2*	479**	……	Male	Heart Disease
2*	479**	……	Male	Cancer
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
≥50	4766*	……	*	Gastritis

- Q should be public information
  - The distribution **Q** is always available to the attacker as long as one wants to release the data at all.
- We separate knowledge gain into two parts:
  - About the whole population (from B_0 to B_1)
  - About specific individuals (from B_1 to B_2)
- We bound knowledge gain between B_1 and B_2 instead

- Principle
  - The distance between **Q** and $\mathbf{P_i}$ should be bounded by a threshold t.

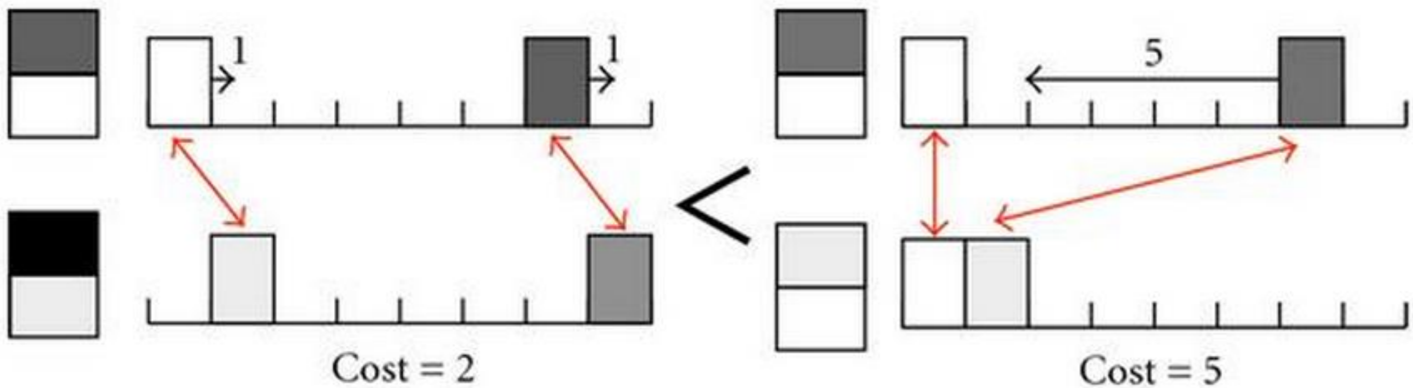
t-Closeness

- Principle: Distribution of sensitive attribute value in each equi-class should be close to that of the overall dataset (distance $\leq$ t)

- How to measure distance between two distributions so that semantic relationship among sensitive attribute values is captured.
  - Assume distribution of income is (10K, 20K, 30K, …, 90K); intuitively (20K,50K,80K) is closer to it than (10K,20K,30K).

The Earth Mover Distance

- We use Earth Mover Distance.



- Distance between (10K, 20K, 30K, …, 90K) and (20K,50K,80K) is $0.1 \times \frac{1}{9} \times 6 = \frac{2}{30} \approx 0.0067$
- Distance between (10K, 20K, 30K, …, 90K) and (10K,20K,30K) is $\frac{1}{9} \times (0.3 + 0.4 + 0.4 + 0.5 + 0.5 + 0.6) = 0.3$

Limitations of t-Closeness

- Utility may suffer too much, since interesting and significant deviation from global distribution cannot be learned.

- **(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
  - Okay to learn information about a large group.

(n,t)-Closeness

- One may argue that requiring t-closeness may destroy data utility
- The notion of (n,t)-closeness requires distribution close to a large-enough natural group of size at least n

- Intuition:
  - It is okay to learn information about the a big group
  - It is not okay to learn information about one individual

Other Limitations

- Requires the distinction between Quasi-identifiers and sensitive attributes

- The t-closeness notion is a property of input dataset and output dataset, not that of the algorithm; thus additional information leakage is possible when the algorithm is known

Limitation of These Privacy Notions

- Limitation of previous privacy notions:
  - Requires identifying which attributes are quasi-identifier or sensitive, not always possible
  - Difficult to pin down adversary's background knowledge
    - There are many adversaries when publishing data
  - Syntactic in nature (property of anonymized dataset)

Privacy Notions: Syntactic versus Algorithmic

- Syntactic: Privacy is a property of only the final output
- Algorithmic: Privacy is a property of the algorithm
- Syntactic notions are typically justified by considering a particular inferencing strategy; however, adversaries may consider other sources of information
  - E.g., Minimality Attack

Illustrating the Syntactic Nature of k-Anonymity

- Method 1 for achieving k anonymity: Duplicating each record k times
- Method 2: clusters records into groups of at least k, use one record from each group to replace all other records in the group
  - Privacy of some individuals are violated
- Method 3: cluster records into groups, then use generalized values to replace the specific values (e.g., consider a 2-D space)
  - Record with extraordinary values are revealed/re-identified

Differential Privacy [Dwork et al. 2006]

- Definition: A mechanism *A* satisfies ε-Differential Privacy if and only if
  - for any neighboring datasets D and D'
  - and any possible transcript $t \in$ Range(A),
    $\Pr[A(D) = t] \le e^{\epsilon} \Pr[A(D') = t]$
  - For relational datasets, typically, datasets are said to be neighboring if they differ by a single record.

Next Lecture

- Local Differential Privacy (By Tianhao Wang)