Privacy in Data Publishing

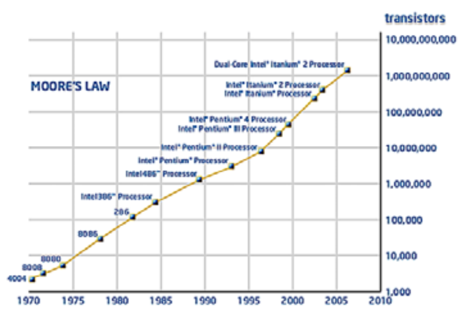
Johannes Gehrke, Cornell University
Ashwin Machanavajjhala, Yahoo! Research

An Abundance of Data

- Supermarket scanners
- Credit card transactions
- Call center records
- ATM machines
- Web server logs
- Customer web site trails
- Podcasts
- Blogs
- Closed caption
- Scientific experiments
- Sensors
- Cameras
- Interactions in social networks
- Facebook, Myspace
- Twitter
- Speech-to-text translation
- Email

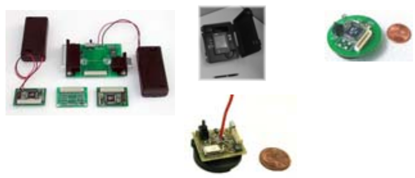
•Print, film, optical, and magnetic storage: 5 Exabytes (EB) of new information in 2002, doubled in the last three years
[How much Information 2003, UC Berkeley]

Driving Factors: A LARGE Hardware Revolution



[Intel Corporation]

Driving Factors: A small Hardware Revolution



- Experts on ants estimate that there are 10^{16} to 10^{17} ants on earth. In the year 1997, we produced one transistor per ant.
  [Gordon Moore]

Driving Factors: Analysis Capabilities

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Example pattern (Census Bureau Data):
If (relationship = husband), then (gender = male). 99.6%

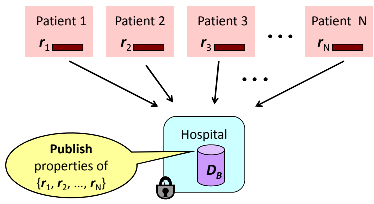
Driving Factors: Connectivity and Bandwidth

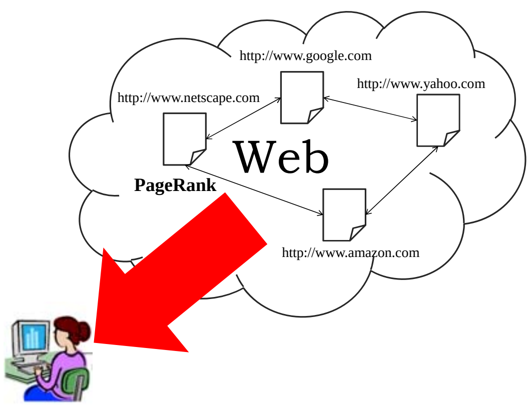
- Metcalf's law (network usefulness increases squared with the number of users)
- Gilder's law (bandwidth doubles every 6 months)

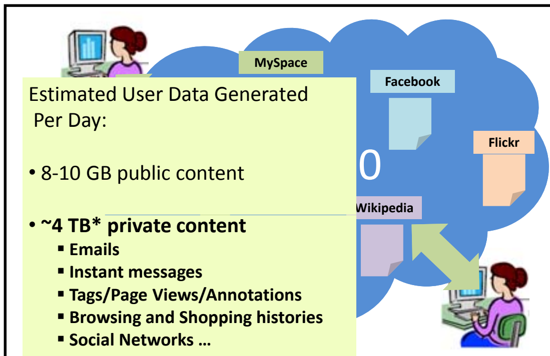
Data Collection Agencies Publish Sensitive Information to Facilitate Research.

Publish information that:

- Discloses as much statistical information as possible.
- Preserves the privacy of the individuals contributing the data.



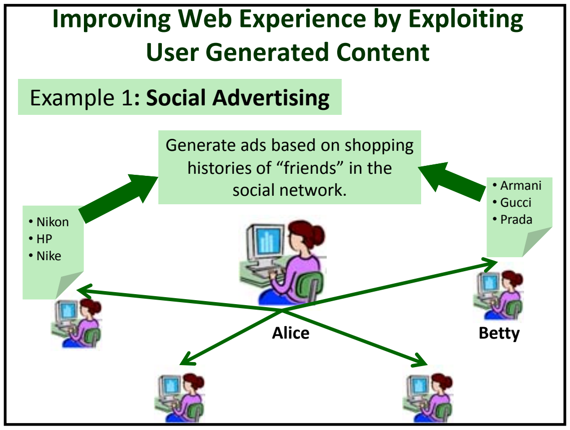


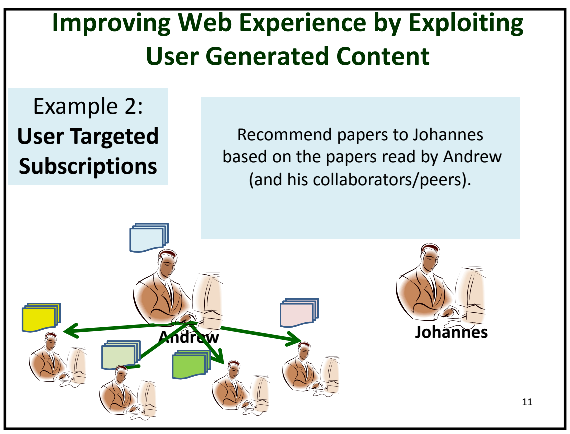


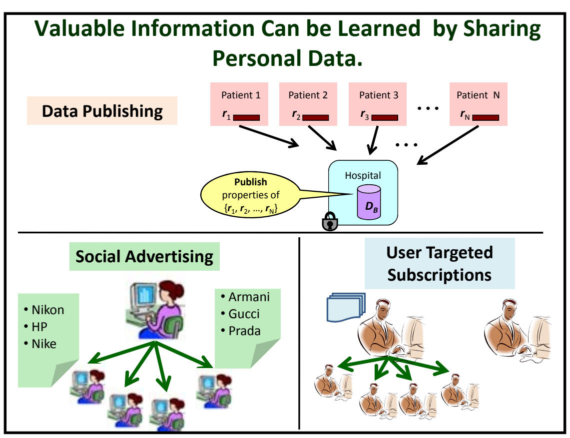
Estimated User Data Generated Per Day:

- 8-10 GB public content
- **~4 TB* private content**
  - **Emails**
  - **Instant messages**
  - **Tags/Page Views/Annotations**
  - **Browsing and Shopping histories**
  - **Social Networks …**

Ramakrishnan et al, IEEE Computer 2007


Improving Web Experience by Exploiting User Generated Content
Example 1: Social Advertising
Generate ads based on shopping histories of "friends" in the social network.
• Nikon
• HP
• Nike
• Armani
• Gucci
• Prada
Alice
Betty


Improving Web Experience by Exploiting User Generated Content
Example 2: User Targeted Subscriptions
Recommend papers to Johannes based on the papers read by Andrew (and his collaborators/peers).
Andrew
Johannes
11


Valuable Information Can be Learned by Sharing Personal Data.
Data Publishing
Patient 1
Patient 2
Patient 3
Patient N
Hospital
Publish properties of
Social Advertising
• Nikon
• HP
• Nike
• Armani
• Gucci
• Prada
User Targeted Subscriptions

What about Privacy?

"… Last week AOL did another stupid thing …
… but, at least it was in the name of science…"

Alternet, August 2006

AOL Data Release …

AOL "anonymously" released a list of 21 million web search queries.

UserIDs were replaced by random numbers …

[illegible]	Uefa cup
[illegible]	Uefa champions league
[illegible]	Champions league final
[illegible]	Champions league final 2008
[illegible]	exchangeability
[illegible]	Proof of deFinitti's theorem
[illegible]	Zombie games
[illegible]	Warcraft
[illegible]	Beatles anthology
[illegible]	Ubuntu breeze
[illegible]	Grammy 2008 nominees
[illegible]	Amy Winehouse rehab

A Face Is Exposed for AOL Searcher No. 4417749 [New York Times, August 9, 2006]

…

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from "numb fingers" to "60 single men" to "dog that urinates on everything."

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for "landscapers in Lilburn, Ga," several people with the last name Arnold and "homes sold in shadow lake subdivision gwinnett county georgia."

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. "Those are my searches," she said, after a reporter read part of the list to her.

…

A Face Is Exposed for AOL Searcher No. 4417749 [New York Times, August 9, 2006]

Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. In response, she plans to drop her AOL subscription. “We all have a right to privacy,” she said. “Nobody should have found this all out.”

http://data.aolsearchlogs.com

What is Privacy?

- “*The claim of individuals, groups, or institutions to determine for themselves* ***when, how and to what extent information about them is communicated to others***”

  Westin, Privacy and Freedom, 1967

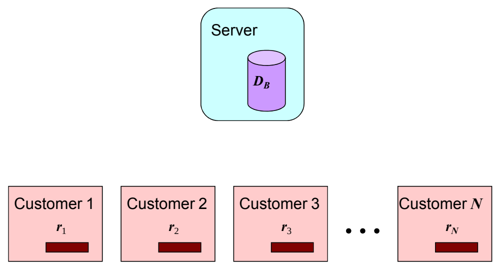
- But we need *quantifiable* notions of privacy …

What is Privacy?

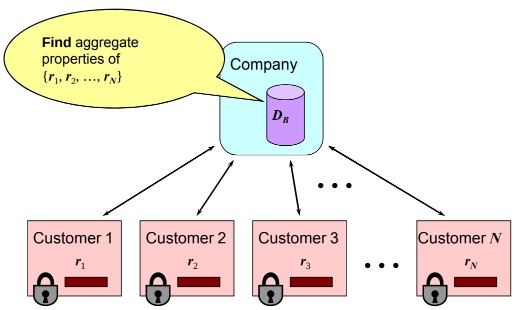
... nothing about an individual should be learnable from the database that cannot be learned without access to the database …

T. Dalenius, 1977

The Setup



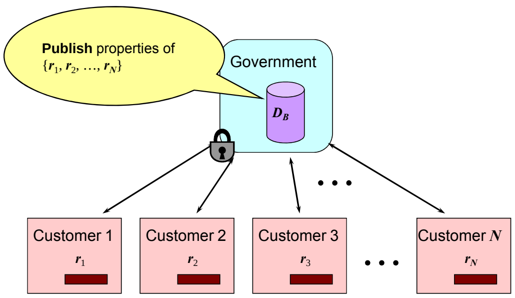
Model I: Untrusted Data Collector



Minimal Information Sharing

- Ideally, we want an algorithm that discloses only the query result, and only to the requesting party. (In practice, we need some extra disclosure.)
- How do we design algorithms that compute queries while preserving data privacy?
- How do we measure privacy (this extra disclosure)?

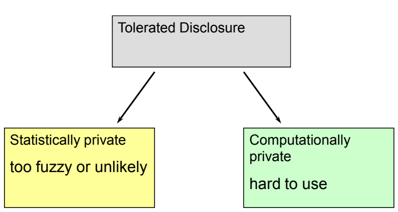
Model II: Trusted Data Collector



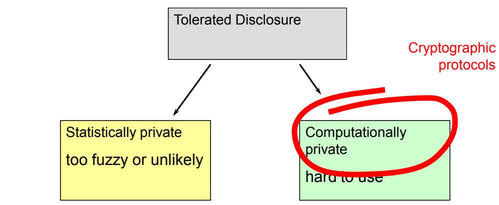
Disclosure Limitations

- Ideally, we want a solution that discloses as much statistical information as possible while preserving privacy of the individuals who contributed data.
- How do we design algorithms that allow the "largest" set of queries that can be disclosed while preserving data privacy?
- How do we measure disclosure?

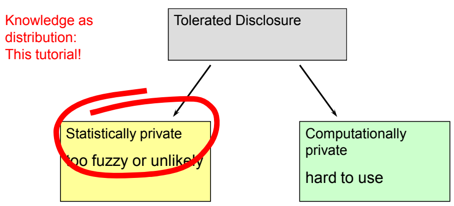
Types of Disclosure



Types of Disclosure



Types of Disclosure



This Tutorial

Privacy-preserving data publishing

- Untrusted data collector
- Trusted data collector

Caveat:

- Not a comprehensive survey

What is Left Out?

- Work on secure multi-party computation (secure join, secure intersection, homomorphic encryption, certificate revocation, etc.)
- Architectural and language issues (Hippocratic databases, P3P, etc.)
- Privacy through distributed data mining

And of course:

- Much more work on other privacy definition, attacks, motivating scenarios, etc.
- Check out www.cs.cornell.edu/bigreddata/privacy for updates.

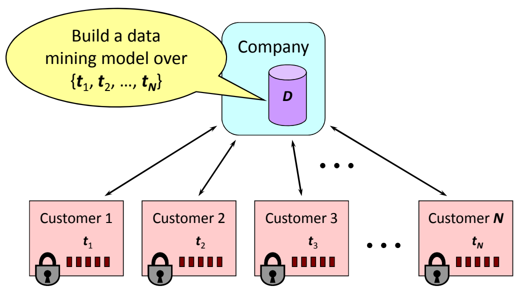
Tutorial Outline

- Untrusted Data Collector
- Trusted Data Collector
- A Success Story: OnTheMap

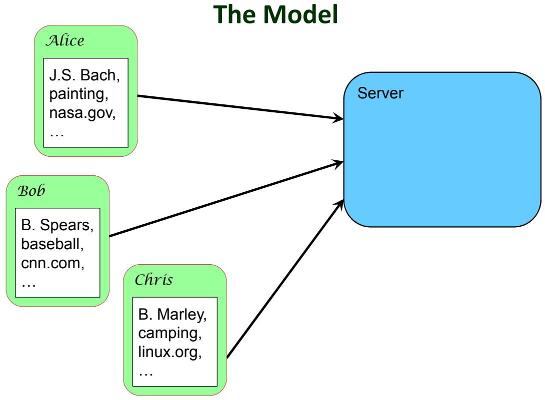
Tutorial Outline

- Untrusted Data Collector
  - Randomized response
  - Interval privacy
  - Entropy-based privacy
  - Alpha-beta privacy breaches
- Trusted Data Collector
- A Success Story: OnTheMap

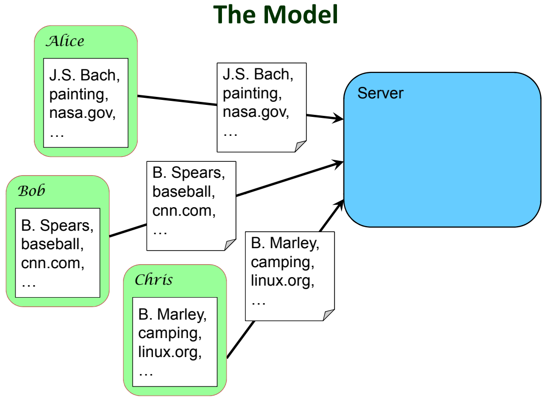
Untrusted Data Collector



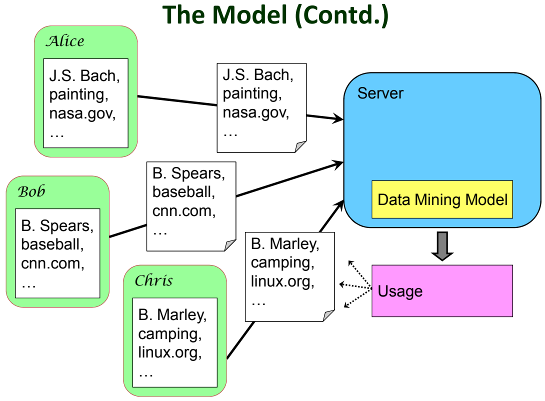
The Model



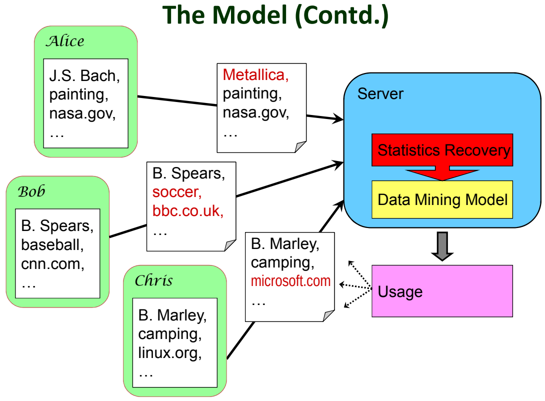
The Model



The Model (Contd.)



The Model (Contd.)



Problem

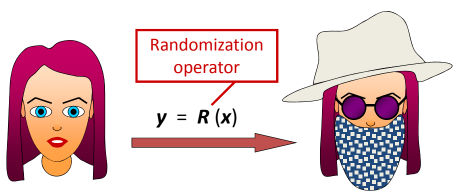
How to randomize the data such that

- We can build a good data mining model (utility)
  - Very simple model: Frequent itemsets (commonly occurring preferences)
- While preserving privacy at the record level (privacy)
  - What does privacy mean?

Motivation: A Social Survey

- Measures opinions, attitudes, behavior
- Problem: Questions of a sensitive nature
  - Examples: sexuality, incriminating questions, embarrassing questions, threatening questions, controversial issues, etc.
  - The "non-cooperative" group leads to errors in surveys and inaccurate data
  - Even though privacy is guaranteed, skepticism prevails

The Model



x
Original (private) data

Assumptions:
- Described by a random variable ***X***.
- Each individual client is independent.

y
Randomized data

Described by a random variable ***Y*** = ***R*** (***X***).

The Randomized Response Model

[Stanley Warner; JASA 1965]

- Respondents are given:
  1. A source of randomness (a biased coin)
  2. A statement: I am a member of the XYZ party.
- The procedure:
  - Flip the coin, associate Head with Yes, Tail with No
  - Answer YES if coin gives correct answer, answer NO otherwise

Randomized Response (Contd.)

- The procedure:
  - Flip the coin, associate Head with Yes, Tail with No
  - Answer YES if coin gives correct answer, Answer NO otherwise

	Yes	No
Head (Yes)	Yes	No
Tail (No)	No	Yes

Another View: Two Questions

- Respondents are given:
  1. A coin
  2. Two logically opposite statements:
     - S1: I am a member of the XYZ party.
     - S2: I am **not** a member of the XYZ party.
- The procedure:
  - Flip the coin
  - Answer either statement S1 or S2.

Randomized Response (Contd.)

- Version 1
  - Flip the coin, associate Head with Yes, Tail with No
  - Answer YES if coin gives correct answer, answer NO otherwise

	Yes	No
Head (Yes)	Yes	No
Tail (No)	No	Yes

- Version 2
  - Two logically opposite statements
  - Answers either statement S1 or S2.

	Yes	No
Head (S1)	Yes	No
Tail (S2)	No	Yes

Analysis

π = the true probability of S in the population.
p = the probability that the coin says YES.

Y_i = 1 if the i^{th} respondent says 'yes'.
0 if the i^{th} respondent reports 'no'.

- $P(Y_i=1) = \pi p + (1-\pi)(1-p) = p_{YES}$
- $P(Y_i=0) = (1-\pi)p + \pi(1-p) = p_{NO}$

	Yes	No
Head	Yes	No
Tail	No	Yes

Analysis (Contd.)

- Assume a sample with n records
  - n1 say YES, (n-n1) say NO
- Likelihood of this sample:
  - $L = p_{YES}^{n1}\, p_{NO}^{(n-n1)}$
    (Note: L is a function of π, p, n, n1)
  - This gives a maximum likelihood estimate for π of
    $\pi^{hat} = (p-1)/(2p-1) + n1/n(2p-1)$
- Easy to show:
  - $E(\pi^{hat}) = \pi$
  - $Var(\pi^{hat}) = \pi(1-\pi)]/n + [1/[16(p-0.5)^2]-0.25]/n$

Variance = Sampling + Coin Flips

Randomized Response: Extensions

- What we have seen so far is also called the "Related Question Procedure"
  - Q1: Do you have property P?
  - Q2: Do you have property P^{bar}?
- Unrelated Question Procedure
  - Q1: Do you use illegal drugs?
  - Q2: Were you born in January?
  - Two types of analyses, depending on whether "fraction of respondents who answer YES to Q2" is known.
- Sensitive attribute with several categories
- Quantitative sensitive attributes

Randomized Response Revisited

- However: Nothing about privacy.
- What is the privacy guaranteed by randomized response?

Interval Privacy

[Agrawal and Srikant; SIGMOD 2000]

Idea: Clients share randomized version of their data.

Randomization:
- For a numerical attribute value x, share value z=x+y, where y is drawn from some known distribution

Interval Privacy (Contd.)

Example:
- Add value drawn from a uniform distribution between -30 and +30 to age.
- If randomized age is 60
  - We know with 90% confidence that age is between 33 and 87.
  - We know with 100% confidence that age is between 30 and 90.

Width of interval to which adversary can localize x is the amount of privacy.
  - Example:
    - Interval width 54 with 90% confidence
    - Interval width 60 with 100% confidence

An Attack on Interval Privacy

[Agrawal and Aggarway; PODS 2001]
Example: Attribute X with the following density function $f_X(x)$:

- $f_X(x) = 0.5$, $0 \leq x \leq 1$
- $f_X(x) = 0.5$, $4 \leq x \leq 5$
- $f_X(x) = 0$, otherwise

Noise Y is distributed uniformly between [-1,1]
Claim: Privacy 2 at 100% confidence level

Reconstruction:
$Z \in [-1,2]$ gives $X \in [0,1]$, and $Z \in [3,6]$ gives $X \in [4,5]$

→ Privacy at 100% confidence level is at most 1.
- (X can be localized to even shorter intervals, e.g. Z=-0.5 gives $X \in [0,0.5]$, Z=-1 gives X=0!)

An Attack on Interval Privacy (Contd.)

- What went wrong with interval privacy? Original distribution of X was ignored!
  - Some values of X are highly unlikely
  - If we see “outlier” values of Z, they constrain the corresponding value of X
- Approach:
  - Quantify information content of distribution of randomized records compared to distribution of original records

Privacy Measure: Intuition

- A random variable distributed uniformly between [0,1] has half as much privacy as if it were distributed in [0,2]
- In general: If $f_B(x)=2f_A(2x)$ then B offers half as much privacy as A
  - Think of A as B stretched out at twice the length
- Need a privacy measure that captures this intuition

Differential Entropy

- Differential entropy h(X):

$$h(X) = -\int_{\Omega X} f_X(x) \log f_X(x) dx$$

- Examples:
  - X is uniformly distributed between 0 and 1: h(X)=0.
  - X is uniformly distributed between 0 and a: h(X)=$\log_2$(a).
- Random variables with less uncertainty than U[0,1] have negative differential entropy
- Random variables with more uncertainty than U[0,1] have positive differential entropy

Proposed Measure

- Propose $\Pi(X)=2^{h(X)}$ as measure of privacy for attribute X
- Examples:
  - Uniform U between 0 and 1: $\Pi(U)=2^{\log_2(1)}=2^0=1$
  - Uniform U between 0 and a: $\Pi(U)=2^{\log_2(a)}=a$
- In general, $\Pi(A)$ denotes the length of an interval over which a uniformly distributed random variable has as much uncertainty as A.
- Example:
  $\Pi(X)=2$: X has as much privacy as a random variable distributed uniformly in an interval of length 2

Conditional Privacy

- Conditional privacy takes the additional information in perturbed values into account:

$$h(X \mid Z) = -\int_{\Omega X,Z} f_{X,Z}(x,z) \log f_{X \mid Z=z}(x) dx\, dz$$

- Average conditional privacy of X given Z:
  $\Pi(X|Z)=2^{h(X|Z)}$

Privacy Loss Metric

- Conditional privacy loss of X given Z:

$Loss(X|Z)=1-\Pi(X|Z)/\Pi(X)=1-2^{-I(X;Z)}$, where

  - I(X;Z)=h(X)-h(X|Z), the mutual information between random variables X and Z

- Loss(X|Z) is the fraction of privacy of X which is lost by revealing Z

Recall the Attack

Example: Attribute X with the following density function $f_X(x)$:

- $f_X(x) = 0.5, 0 \leq x \leq 1$
- $f_X(x) = 0.5, \quad 4 \leq x \leq 5$
- $f_X(x) = 0$, otherwise

Noise Y is distributed uniformly between [-1,1]

- Claim: Privacy 2 at 100% confidence level

Reconstruction:

  - $Y \in [-1,2]$ gives $X \in [0,1]$, and $Y \in [3,6]$ gives $X \in [4,5]$

→ Privacy at 100% confidence level is at most 1.

  - (X can be localized to even shorter intervals, e.g. Z=-0.5 gives $X \in [0,0.5]$)

Loss Explains What Is Going On

- In the example: Privacy of X, P(X)=21=2
  → X has as much privacy as U[0, 2]
- We can calculate: I(X;Z) = h(Z) - h(Z|X) = … = 5/4
- Privacy loss of X after learning Z: Loss(X|Z)=1-2-5/4=0.5796
- Privacy of X after revealing Z:
  P(X|Z)=P(X)*(1-Loss(X|Z))=2*(1.0-0.5796)=0.8408
  → X has only as much privacy as U[0, 0.8408]

An Attack on Entropy-Based Privacy

Example:

- $f_X(x) = 0.5$, $0 \le x \le 1$
- $f_X(x) = 0.5$, $4 \le x \le 5$
- $f_X(x) = 0$, otherwise
- Uniform noise Y in [0,1]

- Assume sensitive property: “X<= 0.01.” (prior probability: 0.5%)

- If $Z \in [-1,-0.99]$, the posterior probability P[X <= 0.01 | Z = z] = 1.
- However, $Z \in [-1, -0.99]$ is unlikely (only one in 100,000 records) → not much privacy loss according to conditional differential entropy

An Attack (Contd.)

Recall Dalenius:

... nothing about an individual should be learnable from the database that cannot be learned without access to the database ...

- If $Z \in [-1,-0.99]$, the posterior probability P[X <= 0.01 | Z = z] = 1.
- Caveat:
  - Every time this occurs the property “X <= 0.01” is fully disclosed.
  - The mutual information, being an average measure, is not worried about this rare disclosure.

Randomized Response Revisited

Recall our question: What is the privacy guaranteed by randomized response?

- Interval privacy: No formal privacy definition
- Entropy privacy: Only protects privacy on average

Randomized Response Revisited

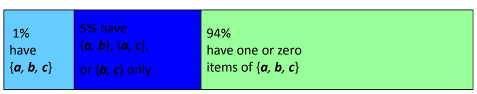
Return to our recommendation service. A "randomized response"-style algorithm:

Given a set of preferences:

- Keep (preference) item with 20% probability,
- Replace with a new random item with 80% probability.

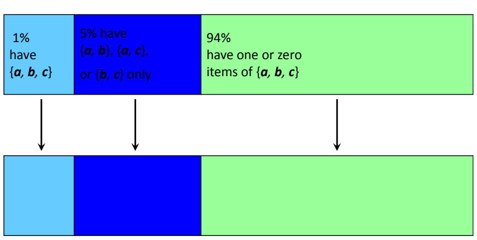
Example: {a, b, c}

10 M transactions of size 10 with 10 K items:



Example: {a, b, c}

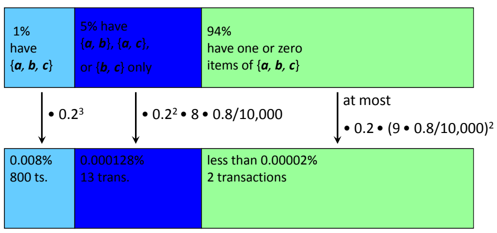
10 M transactions of size 10 with 10 K items:



After randomization: How many have {*a, b, c*} ?

Example: {a, b, c}

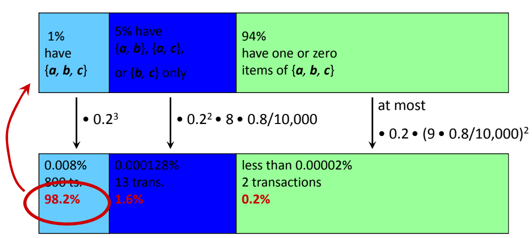
10 M transactions of size 10 with 10 K items:



After randomization: How many have {*a*, *b*, *c*} ?

Example: {a, b, c}

10 M transactions of size 10 with 10 K items:



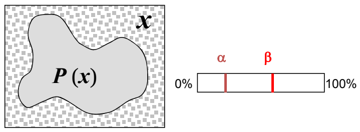
After randomization: How many have {*a*, *b*, *c*} ?

Example: {a, b, c}

- A-priori, we only know with 1% probability that {a, b, c} occurs in the original transaction
- Given {a, b, c} in the randomized transaction, we have about 98% certainty that {a, b, c} occurred in the original transaction.
- This is called a privacy breach.
- The example randomization preserves privacy "on average," but not "in the worst case."

α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



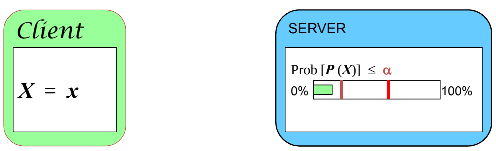
Example:

$P(x)$ = "transaction x contains $\{a, b, c\}$"

α = 1% and β = 50%

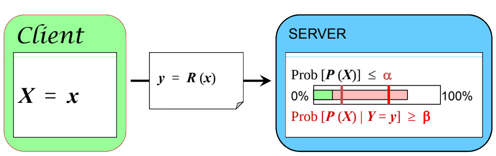
α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



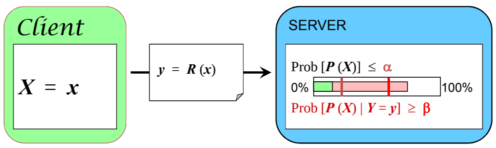
α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



α-to-β Privacy Breach

Let $P(x)$ be any property of client's private data;
Let $0 < \alpha < \beta < 1$ be two probability thresholds.



Disclosure of y causes an α-to-β privacy breach w.r.t. property $P(x)$.

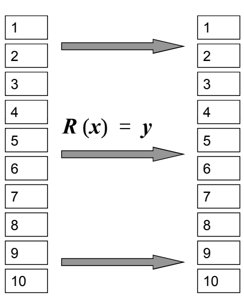
α-to-β Privacy Breach

Checking for α-to-β privacy breaches:

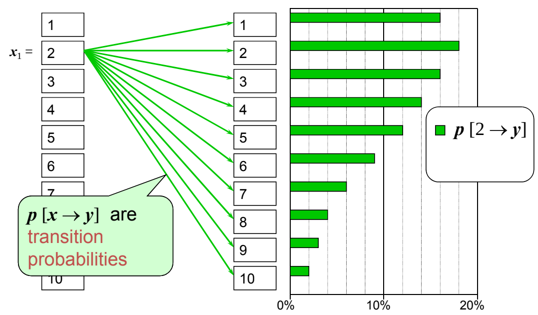
- There are exponentially many properties $P(x)$;
- We have to know the data distribution in advance in order to check whether
  Prob $[P(X)] \leq \alpha$ and Prob $[P(X) \mid Y = y] \geq \beta$

Is there a simple property of randomization operator R that limits privacy breaches?

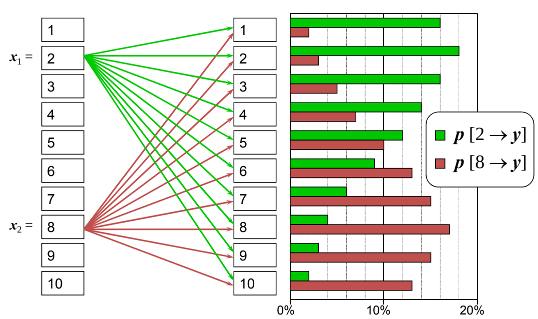
Amplification Condition



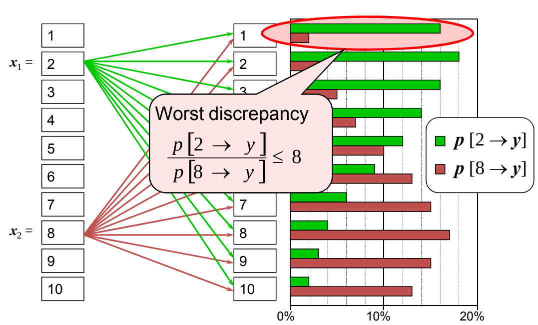
Amplification Condition



Amplification Condition



Amplification Condition



Amplification Condition

Definition:

- Randomization operator ***R*** is called "at most γ-amplifying" if:

$$\max_{x_1, x_2} \max_{y} \frac{p[x_1 \to y]}{p[x_2 \to y]} \leq \gamma$$

  - Transition probabilities ***p*** [***x*** → ***y***] = Prob [***R*** (***x***) = ***y***] depend only on the operator ***R*** and not on data.
  - We assume that all ***y*** have a nonzero probability.
- The bigger γ, the more may be revealed about ***x***.

The Bound on α-to-β Breaches

Theorem:

- If randomization operator ***R*** is at most γ-amplifying, and if:

$$\gamma < \frac{\beta}{\alpha} \cdot \frac{1-\alpha}{1-\beta}$$

- Then, revealing ***R*** (***X***) to the server will never cause an α-to-β privacy breach.

Amplification: Summary

- An α-to-β privacy breach w.r.t. property P (x) occurs when
  - Prob [P is true] $\leq \alpha$
  - Prob [P is true | Y = y] $\geq \beta$.
- Amplification methodology limits privacy breaches by just looking at transitional probabilities of randomization.
  - Does not use data distribution; only check:

$$\max_{x_1, x_2} \max_{y} \frac{p[x_1 \to y]}{p[x_2 \to y]} \leq \gamma$$

One Algorithm: Select-a-Size

- Given transaction t of size m, construct $t' = R(t)$:

t = *a, b, c, d, e, f, u, v, w*

t' =

Definition of Select-a-Size

- Given transaction t of size m, construct $t' = R(t)$:
  - Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $\{p[j]\}_{0..m}$;

t = *a, b, c, d, e, f, u, v, w*

t' =

$j = 4$

Definition of Select-a-Size

- Given transaction t of size m, construct $t' = R(t)$:
  - Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $\{p[j]\}_{0..m}$;
  - Include exactly j items of t into t';

t = *a, b, c, d, e, f, u, v, w*

t' = *b, e, u, w*

$j = 4$

Definition of Select-a-Size

- Given transaction t of size m, construct $t' = R(t)$:
  - Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $\{p[j]\}_{0..m}$;
  - Include exactly j items of t into t';
  - Each other item (not from t) goes into t' with probability ρ.

The choice of $\{p[j]\}_{0..m}$ and ρ is based on the desired privacy level.

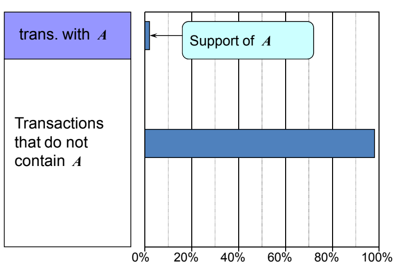
t = *a*, *b*, *c*, *d*, *e*, *f*, *u*, *v*, *w*

t' = *b*, *e*, *u*, *w* | *æ*, *å*, *β*, *§*, *ψ*, *€*, *ℵ*, *ъ*, *ђ*, …

$j = 4$ — items inserted with prob. ρ

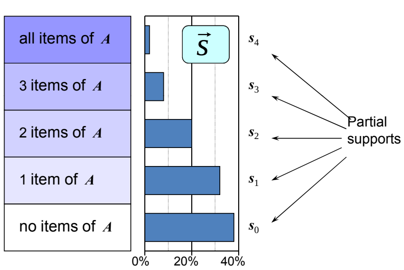
Support Recovery

Let itemset A have four items ($k = 4$).



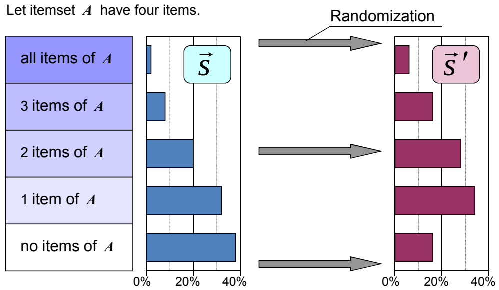
Support Recovery

Let itemset A have four items ($k = 4$).



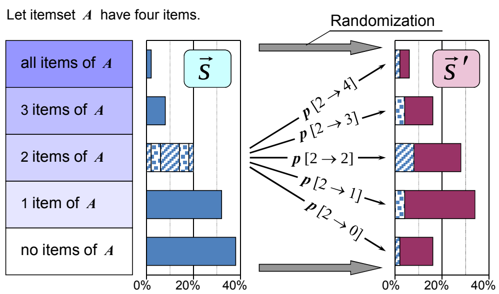
Support Recovery

Let itemset A have four items.



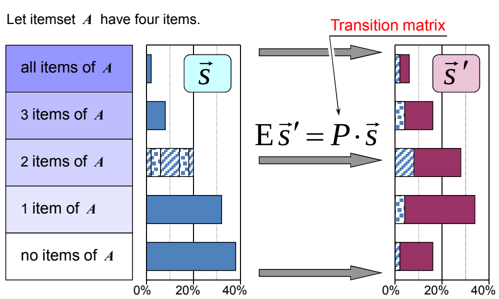
Support Recovery

Let itemset A have four items.



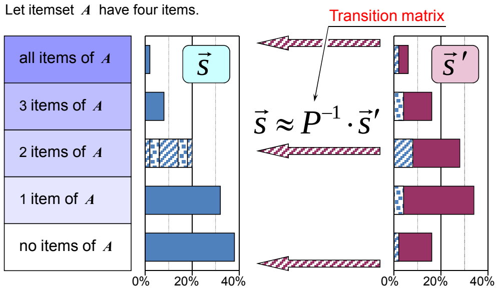
Support Recovery

Let itemset A have four items.



Support Recovery

Let itemset *A* have four items.



The Unbiased Estimators

- Given randomized partial supports, we can estimate the original partial supports:

$$\vec{s}_{\text{est}} = Q \cdot \vec{s}\,', \quad \text{where} \quad Q = P^{-1}$$

- Covariance matrix for this estimator:

$$\text{Cov}\ \vec{s}_{\text{est}} = \frac{1}{|T|} \sum_{l=0}^{k} s_l \cdot Q\, D[l]\, Q^T,$$

$$\text{where} \quad D[l]_{i,j} = P_{i,l} \cdot \delta_{i=j} - P_{i,l} \cdot P_{j,l}$$

- To estimate it, substitute s_l with $(s_{\text{est}})_l$.
  - Special case: estimators for support and its variance
- [RH02] reconstruct statistics similarly

Apriori

[Agrawal and Srikant, VLDB 1994]

Let ***k* = 1**, candidate sets = all 1-itemsets.

Repeat:

1. Count support for all candidate sets
2. Output the candidate sets with support ≥ $s_{\min}$
3. New candidate sets = all (***k* + 1**)-itemsets s.t. all their ***k***-subsets are candidate sets with support ≥ $s_{\min}$
4. Let ***k* = *k* + 1**

Stop when there are no more candidate sets.

The Modified Apriori

Let $k = 1$, candidate sets = all 1-itemsets.

Repeat:

1. Estimate support and variance (σ^2) for all candidate sets
2. Output the candidate sets with support $\geq s_{min}$
3. New candidate sets = all ($k + 1$)-itemsets s.t. all their k-subsets are candidate sets with support $\geq s_{min} - \sigma$
4. Let $k = k + 1$

Stop when there are no more candidate sets, or the estimator's precision becomes unsatisfactory.

Problems

- This did not take off
- No apps
- Does not extend to non-binary data: Show example

An Observation About Attribute Correlation

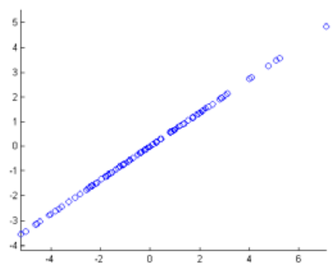
[Huang, Du, Chen; SIGMOD 2005]

- Correlation between attributes can thwart independent random noise
- Example:
  - Assume a dataset with with m attributes that have all the same value
  - We would now perturb the same
- If we do that, we can estimate the original data:
  - Let (t,t, …, t) be the original data,
  - Published data: $t + R_1, t + R_2, \ldots, t + R_m$
  - Let $Z = [(t+R_1)+ \ldots + (t+R_m)] / m$
  - Mean: $E(Z) = t$

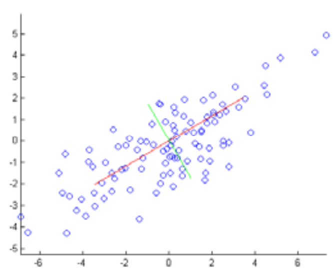
Intuition

- Observation:
  - Original data could be correlated.
  - Noise is not correlated.
- Similar observation by Kargupta and Datta [ICDM 2003]

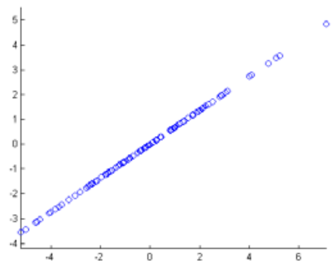
Data



After Randomization



After PCA and Removal of Second PC



What Happened?

Original data:

- Correlated.
- If we remove half the attributes, the actual information loss might be much smaller

Noise:

- Uncorrelated
- Variance evenly distributed across attributes
- If we remove half the attributes, the actual loss in noise should be 50%

Correlated Noise

- You need to add correlated noise, but we do not know what the correlation is.
- Easy to do in the trusted case.

Untrusted Data Collector: Summary

- Each person randomizes her data individually
- Server reconstructs distribution

Untrusted Data Collector: Summary (Contd.)

Importance:

- First setting that introduced a *formal* notion of privacy
  - Alpha-beta privacy
  - Strong semantic notion of privacy, satisfies Dalenius' desiderata

Unimportance:

- Untrusted data collector model has not found a good application (yet?)
  - Data currently mainly collected at servers (amazon, google, etc.)
  - Only statistically significant events can be discovered
  - Application thoughts: P2P file sharing, music recommendation services
  - Secure multi-party computations as alternative?

Untrusted Data Collector: Summary (Contd.)

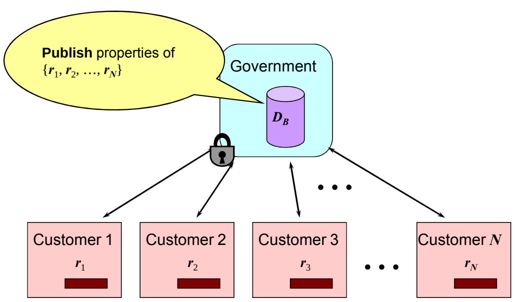
Open questions:

- Much work on privacy, but what about utility?
- What about repeated sharing of data?
- How can we in general analyze such personally randomized data?

Tutorial Outline

- Untrusted Data Collector
- Trusted Data Collector
- A Success Story: OnTheMap

Model II: Trusted Data Collector



Recall: Semantic Disclosure Risk

… nothing about an individual should be learnable from the database that cannot be learned without access to the database …

T. Dalenius, 1977

Untrusted data collector:

- Let x_i be individual *i*'s record.
- For every function/property $f(x_i)$: $dom(x_i) \rightarrow \{0,1\}$
  - Prior belief: $\alpha = Pr[f(x_i) = 1 \mid \text{prior distribution}]$
  - Posterior belief: $\beta = Pr[f(x_i) = 1 \mid \text{prior distribution and } y_i]$

α should be close to **β**

Semantic Disclosure Risk

... nothing about an individual should be learnable from the database that cannot be learned without access to the database ...

T. Dalenius, 1977

- Let x_i be individual *i*'s record.
- For every function $f(x_i)$: $dom(x_i) \rightarrow \{0,1\}$
  - Prior belief: $\boldsymbol{\alpha} = Pr[f(x_i) = 1 \mid \text{prior distribution}]$
  - Posterior belief: $\boldsymbol{\beta} = Pr[f(x_i) = 1 \mid \text{prior distribution} + \text{database}]$

α should be close to **β**

Can we Achieve Semantic Privacy ?

Impossibility of Semantic Privacy in Trusted Data Collector Model

[Dwork, ICALP 2006]

Given any algorithm San() that produces useful answers about the database, there exists some auxiliary information X such that for every prior distribution

$\alpha - \beta \geq \delta$

for a suitable choice of δ.

Impossibility of Semantic Disclosure Risk

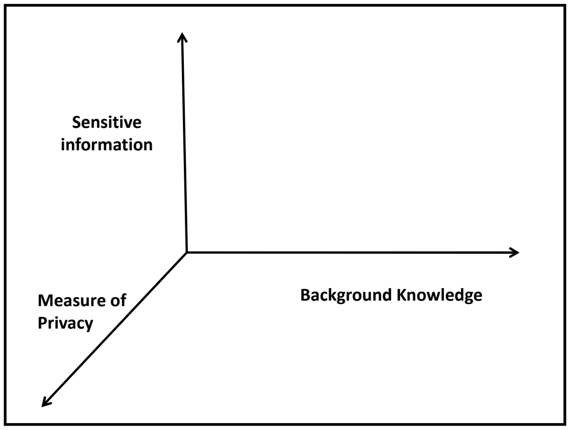
- Suppose:
  - *salary* is a sensitive attribute
  - Database *D* publishes average salaries of employees in different universities
  - Adversary knows: "Andrew earns $10 more than the average Cornell professor".
- Given background knowledge, adversary learns little.
- Given *D*, adversary knows exactly how much Andrew earns!!

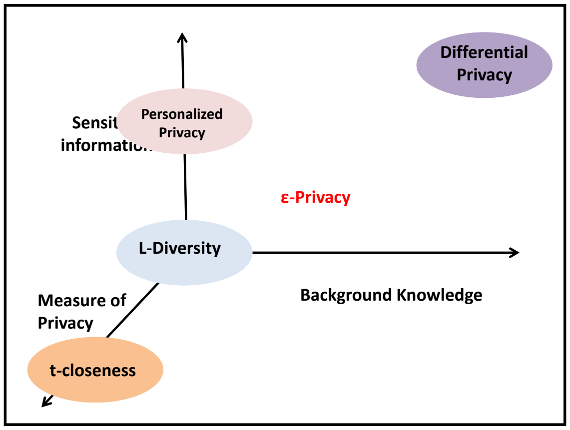
Relaxing Dalenius' Vision

- Three aspects to Dalenius' vision
  - Sensitive Information:
    Dalenius: Every property is secret
    Relaxation: Only some properties
  - Background Knowledge:
    Dalenius: Arbitrary prior and published database
    Relaxation: Only some classes of information
  - Measure of Privacy:
    Dalenius: Prior vs posterior
    Relaxation:
    - Bound posterior
    - Releasing information from a database D should not increase the privacy risk of an individual *x*, if *x* does not appear in D.

Implications

1. Any privacy definition must bound the amount of knowledge an adversary has.
   - *Weak adversaries: k-anonymity, l-diversity, t-closeness.*
2. Releasing information from a database D should not increase the privacy risk of an individual x_i, if x_i does not appear in D.
   - *Strong adversaries: Differential privacy, epsilon privacy*





Tutorial Outline

- Untrusted Data Collector
- Trusted Data Collector
- A Success Story: OnTheMap

Tutorial Outline

- Untrusted Data Collector
- Trusted Data Collector
  - Weak adversaries
  - Strong adversaries
  - Bridging the gap
- A Success Story: OnTheMap

Trusted Data Collector: Weak Adversaries

- Privacy and utility metrics
- Algorithms
- Increasing utility through release of additional data
- Releasing temporally changing data
- The minimality attack and simulatable auditing
- Privacy for social networks
- Active attacks on social networks

Offline Data Publishing

Algorithms:

- Generalization
- Synthetic Data Generation

Utility Metrics:

- Distance between published data and original data (e.g., KL-Divergence).
- Distance over a query workload.

Sample Microdata

SSN	Zip	Age	Nationality	Disease
631-35-1210	13053	28	Russian	Heart
051-34-1430	13068	29	American	Heart
120-30-1243	13068	21	Japanese	Viral
070-97-2432	13053	23	American	Viral
238-50-0890	14853	50	Indian	Cancer
265-04-1275	14853	55	Russian	Heart
574-22-0242	14850	47	American	Viral
388-32-1539	14850	59	American	Viral
005-24-3424	13053	31	American	Cancer
248-223-2956	13053	37	Indian	Cancer
221-22-9713	13068	36	Japanese	Cancer
615-84-1924	13068	32	American	Cancer

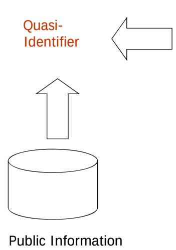
Removing SSN …

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

Medical Records of a hospital near Ithaca serving patients from

- Freeville (13068)
- Dryden (13053)
- Ithaca (14850, 14853)

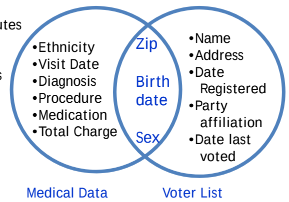
Linkage Attacks



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

Linkage Attacks (Contd.)

- Medical Data was considered anonymous, since identifying attributes were removed.
- Governor of Massachusetts, was uniquely identified by the attributes Zip, Birth Date, Sex
- Hence, his private medical records were out in the open
- {Zip, Birth Date, Sex} *Quasi-Identifier*
- 87 percent of US population uniquely identified using the above Quasi Identifier [S02]



Different Types of Disclosure

- Identity Disclosure
  - Should not disclose whether individual's record in the data.
- Attribute Disclosure
  - Should not disclose the value of sensitive attributes.

Quasi-Identifiers and Sensitive Attributes

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

- Base Table: Medical Records of a hospital near Ithaca serving patients from Freeville (13068), Dryden (13053), and Ithaca (14850, 14853)
- The combination {Zip, Age, Nationality} is the **quasi-identifier**
- Disease is the **sensitive attribute**

K-Anonymity

[Samarati et al, PODS 1998]

- 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 ≥ k

- Parameter k indicates the “degree” of anonymity

Generalization: Coarsen Attributes

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
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

Example Microdata

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

4-Anonymous Microdata

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

Attacks on K-Anonymity

- [Machanavajjhala, Gehrke, Kifer, Venkitasubramaniam; ICDE 2006]
- K-Anonymity does not protect against some simple attacks

Homogeneity Attack

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

- Alice's neighbor Bob is in the hospital.
- Alice knows Bob is 35 years old and is from Dryden (13053).
- Alice learns that Bob has cancer.

Alice

Background Knowledge Attack

Zip	Age	Occupation	Disease
130**	<30	*	~~Heart~~
130**	<30	*	~~Heart~~
130**	<30	*	Viral
130**	<30	*	Viral
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Viral
1485*	>40	*	Viral
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer
130**	30-40	*	Cancer

Alice

- Alice's friend Umeko is in the table.
- Alice knows Umeko is 24, a Japanese, living in Freeville (13068)
- Japanese have *extremely low incidence* of heart disease

Alice learns Umeko has a viral infection

Ensuring Diversity

- L-Diversity: Ensure that every group has at least L *well represented* groups of sensitive values"
  - "well represented" = roughly equal, non-negligible proportions

Two instantiations:

- Entropy l-diversity: Entropy(group) > log(l)
- Recursive (c,l)-diversity: $r_1 < c(r_\ell + r_{\ell+1} + \cdots + r_m)$, where r_i is the number of times the i^{th} most frequent sensitive value appears in the group.

3-Diverse Microdata

Zip	Age	Nationality	Disease
1306*	<=40	*	~~Heart~~
1306*	<=40	*	Viral
1306*	<=40	*	Cancer
1306*	<=40	*	Cancer
1485*	>40	*	Cancer
1485*	>40	*	Heart
1485*	>40	*	Viral
1485*	>40	*	Viral
1305*	<=40	*	Heart
1305*	<=40	*	Viral
1305*	<=40	*	Cancer
1305*	<=40	*	Cancer

- Bob is 35 years old and is from Dryden (13053).
- Umeko is 24, a Japanese from Freeville (13068)
- Japanese have *extremely low incidence* of heart disease

L-Diversity Revisited

- L -Diversity: Every group has at least L well represented groups

Q	S
q*	s
q*	~~s1~~
q*	~~s2~~
q*	~~s3~~
q*	~~s4~~

Note:

- L-diversity does not protect against adversaries having arbitrary background knowledge
- But: L-diversity increases the bar.

T-Closeness

[Li et al, ICDE 2007]

- Rationale: L-Diversity just looks at the posterior belief of an adversary who has linked an individual to a group.
- However, the adversary also learns the overall distribution of the sensitive attribute from the published table

Age	Zipcode	……	Gender	Disease
*	*	……	*	HIV
*	*	……	*	Flu
*	*	……	*	Viral
.	.	……	.	.
.	.	……	.	.
.	.	……	.	.
*	*	……	*	Flu

Background knowledge:

1. External knowledge
2. Distribution of disease in the table

T-Closeness (Contd.)

- So the adversary ends up with:
  - External knowledge (P0)
  - Distribution of sensitive attribute in database (P1)
  - Distribution of sensitive attribute for target individual (P2)
- T-closeness bounds the difference between P1 and P2 instead of P0 and P1
- Distance metric: Earth mover’s distance

Personalized Privacy

[Xiao et al, SIGMOD 2006]

- Goal:
  - a mechanism to capture personalized privacy requirements
  - criteria for measuring the degree of security provided by a generalized table
  - an algorithm for generating publishable tables

Motivation 1: Personalization

- Andy does not want anyone to know that he had a stomach problem
- Sarah does not mind at all if others find out that she had flu

A 2-diverse table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

An external database

Name	Age	Sex	Zipcode
Andy	4	M	12000
Bill	5	M	14000
Ken	6	M	18000
Nash	9	M	19000
Mike	*7*	*M*	*17000*
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

Motivation 2: SA Generalization

- How many female patients are there with age above 30?
- 4 · (60 – 30 + 1) / (60 – 21 + 1) = 3
- Real answer: 1

A generalized table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	gastritis
[21, 60]	F	[30001, 60000]	flu
[21, 60]	F	[30001, 60000]	flu

An external database

Name	Age	Sex	Zipcode
Andy	4	M	12000
Bill	5	M	14000
Ken	6	M	18000
Nash	9	M	19000
Mike	*7*	*M*	*17000*
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

Motivation 2: SA Generalization (Contd.)

- Generalization of the sensitive attribute is beneficial in this case

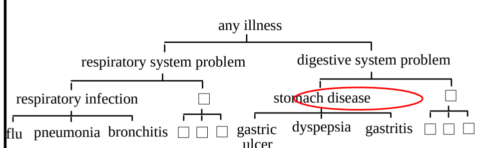
A better generalized table

Age	Sex	Zipcode	Disease
[1, 5]	M	[10001, 15000]	gastric ulcer
[1, 5]	M	[10001, 15000]	dyspepsia
[6, 10]	M	[15001, 20000]	pneumonia
[6, 10]	M	[15001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
[21, 30]	F	[30001, 40000]	gastritis
[21, 30]	F	[30001, 40000]	gastritis
[21, 30]	F	[30001, 40000]	flu
56	F	58000	respiratory infection

An external database

Name	Age	Sex	Zipcode
Andy	4	M	12000
Bill	5	M	14000
Ken	6	M	18000
Nash	9	M	19000
Mike	*7*	*M*	*17000*
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

Guarding Node

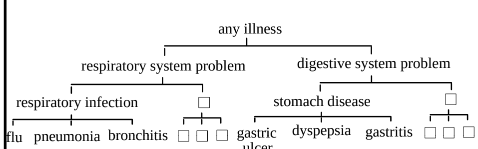


- Andy does not want anyone to know that he had a stomach problem
- He can specify "*stomach disease*" as the guarding node for his tuple

Name	Age	Sex	Zipcode	Disease	guarding node
Andy	4	M	12000	gastric ulcer	*stomach disease*

- The data publisher should prevent an adversary from associating Andy with "stomach disease"

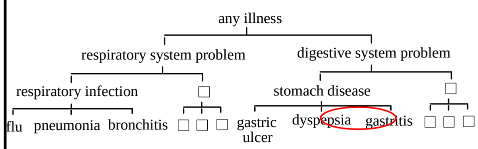
Guarding Node



- Sarah is willing to disclose her exact symptom
- She can specify $\emptyset$ as the guarding node for her tuple

Name	Age	Sex	Zipcode	Disease	guarding node
Sarah	28	F	37000	flu	$\emptyset$

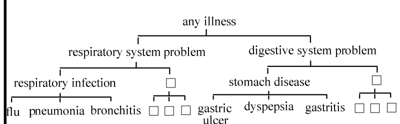
Guarding Node



- Bill does not have any special preference
- He can specify the guarding node for his tuple as the same with his sensitive value

Name	Age	Sex	Zipcode	Disease	guarding node
Bill	5	M	14000	dyspepsia	*dyspepsia*

A Personalized Approach



Name	Age	Sex	Zipcode	Disease	guarding node
Andy	4	M	12000	gastric ulcer	*stomach disease*
Bill	5	M	14000	dyspepsia	*dyspepsia*
Ken	6	M	18000	pneumonia	*respiratory infection*
Nash	9	M	19000	bronchitis	*bronchitis*
Alice	12	F	22000	flu	*flu*
Betty	19	F	24000	pneumonia	*pneumonia*
Linda	21	F	33000	gastritis	*gastritis*
Jane	25	F	34000	gastritis	Ø
Sarah	28	F	37000	flu	Ø
Mary	56	F	58000	flu	*flu*

Personalized Anonymity

Name	Age	Sex	Zipcode	Disease	guarding node
Andy	4	M	12000	gastric ulcer	*stomach disease*
Bill	5	M	14000	dyspepsia	*dyspepsia*
Ken	6	M	18000	pneumonia	*respiratory infection*
Nash	9	M	19000	bronchitis	*bronchitis*
Alice	12	F	22000	flu	*flu*
Betty	19	F	24000	pneumonia	*pneumonia*
Linda	21	F	33000	gastritis	*gastritis*
Jane	25	F	34000	gastritis	Ø
Sarah	28	F	37000	flu	Ø
Mary	56	F	58000	flu	*flu*

- A table satisfies *personalized anonymity* with a parameter p_{breach}
  - Iff no adversary can breach the privacy requirement of any tuple with a probability above p_{breach}
- If p_{breach} = 0.3, then any adversary should have no more than 30% probability to find out that:
  - Andy had a stomach disease
  - Bill had dyspepsia, etc

Personalized Anonymity

- Personalized anonymity with respect to a predefined parameter p_{breach}
  - an adversary can breach the privacy requirement of any tuple with a probability at most p_{breach}
- We need a method for calculating the breach probabilities

Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 20000]	gastric ulcer
[1, 10]	M	[10001, 20000]	dyspepsia
[1, 10]	M	[10001, 20000]	pneumonia
[1, 10]	M	[10001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
21	F	33000	stomach disease
25	F	34000	gastritis
28	F	37000	flu
56	F	58000	respiratory infection

What is the probability that Andy had some stomach problem?

Combinatorial Reconstruction

- Assumptions
  - the adversary has no prior knowledge about each individual
  - every individual involved in the microdata also appears in the external database

Combinatorial Reconstruction

- Andy does not want anyone to know that he had some stomach problem
- What is the probability that the adversary can find out that "Andy had a stomach disease"?

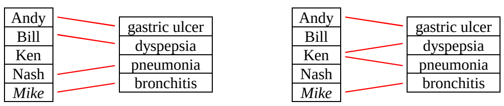
Name	Age	Sex	Zipcode
Andy	4	M	12000
Bill	5	M	14000
Ken	6	M	18000
Nash	9	M	19000
Mike	*7*	*M*	*17000*
Alice	12	F	22000
Betty	19	F	24000
Linda	21	F	33000
Jane	25	F	34000
Sarah	28	F	37000
Mary	56	F	58000

Age	Sex	Zipcode	Disease
[1, 10]	M	[10001, 20000]	gastric ulcer
[1, 10]	M	[10001, 20000]	dyspepsia
[1, 10]	M	[10001, 20000]	pneumonia
[1, 10]	M	[10001, 20000]	bronchitis
[11, 20]	F	[20001, 25000]	flu
[11, 20]	F	[20001, 25000]	pneumonia
21	F	33000	stomach disease
25	F	34000	gastritis
28	F	37000	flu
56	F	58000	respiratory infection

Combinatorial reconstruction (cont.)

- Can each individual appear more than once?
  - No = the primary case
  - Yes = the non-primary case
- Some possible reconstructions:

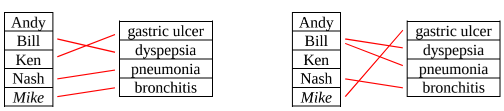
the primary case / the non-primary case



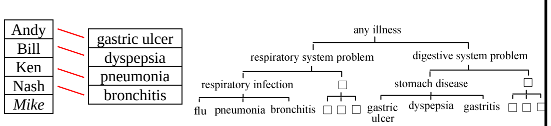
Combinatorial Reconstruction (cont.)

- Can each individual appear more than once?
  - No = the primary case
  - Yes = the non-primary case
- Some possible reconstructions:

the primary case / the non-primary case



Breach probability

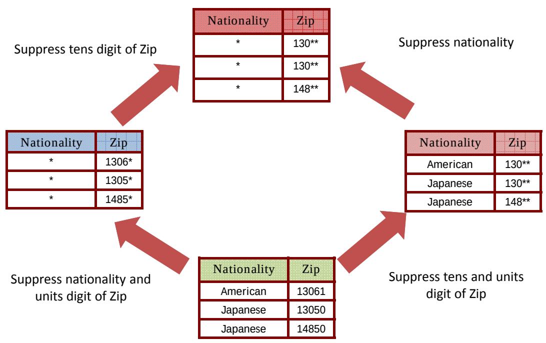


- Totally 120 possible reconstructions
- If Andy is associated with a stomach disease in n_b reconstructions
- The probability that the adversary should associate Andy with some stomach problem is n_b / 120
- Andy is associated with
  - gastric ulcer in 24 reconstructions
  - dyspepsia in 24 reconstructions
  - gastritis in 0 reconstructions
- n_b = 48
- The breach probability for Andy's tuple is 48 / 120 = 2 / 5

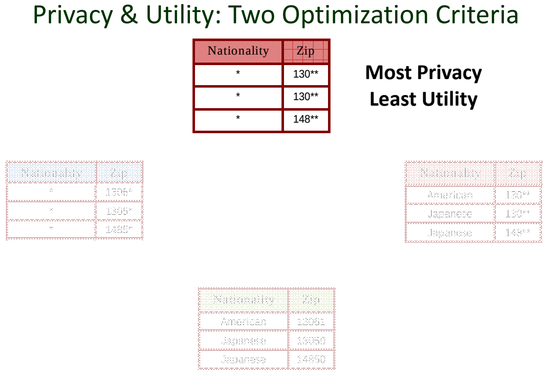
Trusted Data Collector: Weak Adversaries

- Privacy and utility metrics
- Algorithms
- Increasing utility through release of additional data
- Releasing temporally changing data
- The minimality attack and simulatable auditing
- Privacy for social networks
- Active attacks on social networks

Many Different Ways to Coarsen a Relation



Privacy & Utility: Two Optimization Criteria



Privacy & Utility: Two Optimization Criteria

Nationality	Zip
*	130**
*	130**
*	148**

Nationality	Zip
*	1306*
*	1305*
*	1485*

Nationality	Zip
American	130**
Japanese	130**
Japanese	148**

Nationality	Zip
American	13061
Japanese	13050
Japanese	14850

Least Privacy
Most Utility

Optimization problem: Find a table in the lattice with maximum utility & privacy

- Search space: Set of all generalizations
- Privacy metric: k-Anonymity, L-Diversity, …
- Utility metric: Avg. group size, Discernibility, KL-divergence

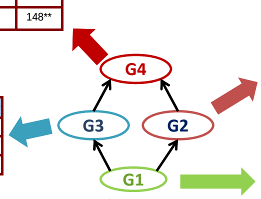
Challenge: Search is hard.

Generalizations Form a Lattice Structure

Nationality	Zip
*	130**
*	130**
*	148**

Nationality	Zip
American	130**
Japanese	130**
Japanese	148**

Nationality	Zip
*	1306*
*	1305*
*	1485*



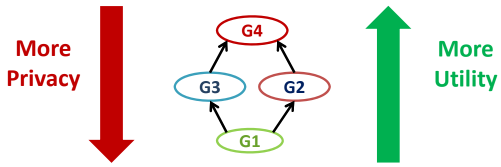
Generalization Lattice

Nationality	Zip
American	13061
Japanese	13050
Japanese	14850

⟶ **Suppress strictly more information**

Key Idea: Monotonicity

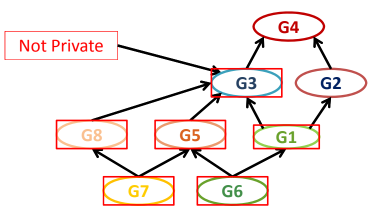
Many privacy and utility metrics are monotonic



Generalization Lattice

Key Idea: Monotonicity

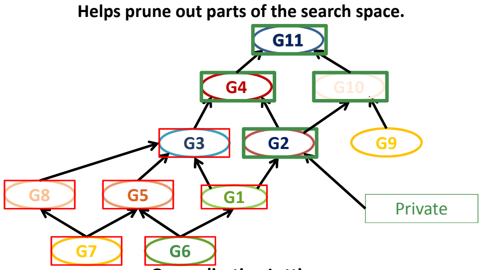
Many privacy and utility metrics are monotonic.
Helps prune out parts of the search space.



Generalization Lattice

Key Idea: Monotonicity

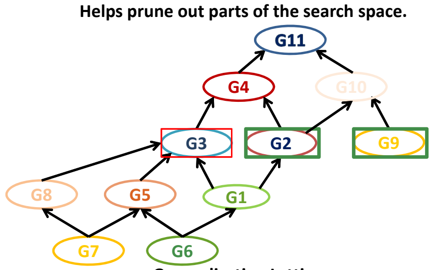
Many privacy and utility metrics are monotonic.
Helps prune out parts of the search space.



Generalization Lattice

Key Idea: Monotonicity

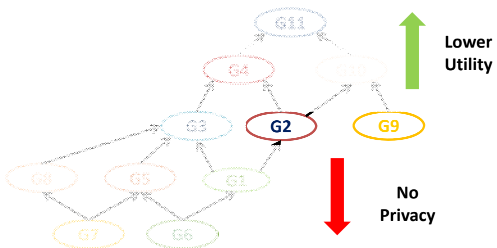
Many privacy and utility metrics are monotonic.
Helps prune out parts of the search space.



Generalization Lattice

Minimal Generalizations

Lowest generalizations on the lattice that guarantee privacy.



Generalization Lattice

Monotonicity Based Generalization Algorithms

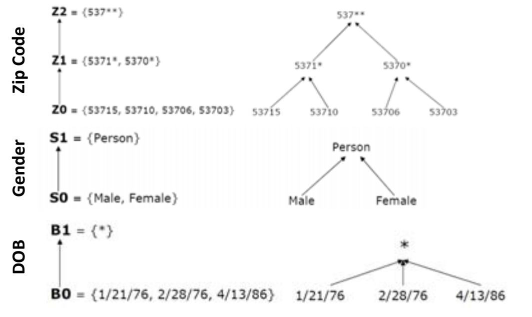
- Incognito [Levefre et al, SIGMOD 2006]
- Single Dimensional Recoding
- Mondrian
- …

Algorithms differ in

1. Construction of the generalization lattice.
2. Traversal of the search space.

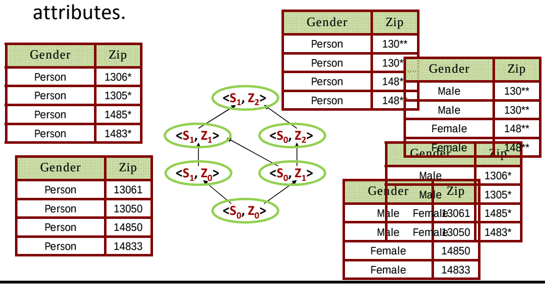
Incognito: Generalization Lattice

Generalization Hierarchy for every attribute …



Incognito: Generalization Lattice

- *Full Domain Generalization:*
  In each generalization step, all the tuples are coarsened one level higher in the hierarchy of one of the attributes.



Incognito: Efficient Full Domain Generalization

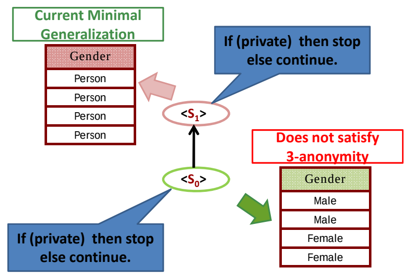
For *i* = 1 to # attributes.

- In each iteration consider a set A of *i* attributes.
- Search lattice of A for minimal generalizations.
- Prune the lattices from Level *i*+1.

Return all *minimal* tables in the lattice that are private.

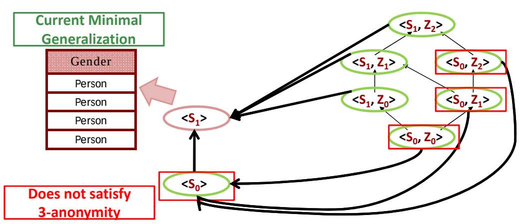
Incognito: Search for Minimal Generalizations

Iteration: *i* = 1, and A = {gender}



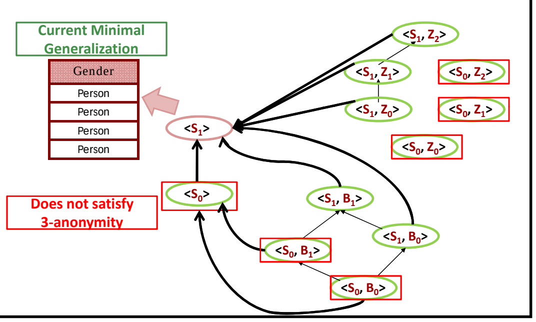
Incognito: Pruning Level *i*+1

Iteration: *i* = 1, and A = {gender}



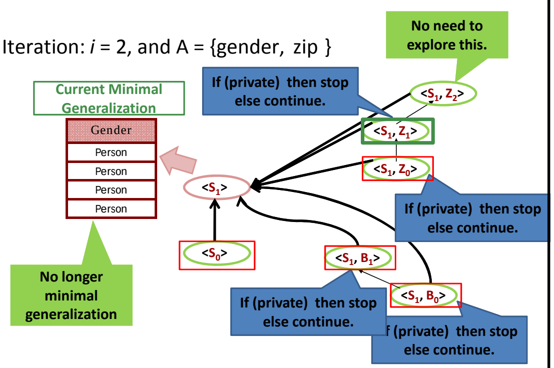
Incognito: Pruning Level *i*+1

Iteration: *i* = 1, and A = {gender}



Incognito

Iteration: *i* = 2, and A = {gender, zip }



Other Generalization Algorithms

- Single Dimension Recoding
  - No generalization hierarchies. Impose a total order on each attribute and look at all possible partitions.
  - Also uses a pruning algorithm like Incognito.
- Mondrian
  - Multidimensional splits like a kd-tree.
  - Uses a greedy traversal of the space.
- Hilbert
  - Converts multidimensional tuple into totally ordered 1-D space.
  - Generalizes by considering ranges on the total order.

Generalization: Summary

- Generalization is a simple technique by coarsening attributes.
- Leads to a lattice search problem that is intractable in general
- Monotonicity helps us to build efficient algorithms.

- Works only for monotonic privacy and utility metrics

Trusted Data Collector: Weak Adversaries

- Privacy and utility metrics
- Algorithms
- Increasing utility through release of additional data
- Releasing temporally changing data
- The minimality attack and simulatable auditing
- Privacy for social networks
- Active attacks on social networks

How Can We Increase Utility?

ID	Age	Height	Gender	Zip	Disease
1	28	5' 5"	F	13053	Heart
2	29	5' 8"	F	13068	Heart
3	21	6' 7"	M	13068	Flu
4	23	5' 9"	F	13053	Flu
5	50	3' 1"	M	14853	Cancer
6	55	6' 0"	M	14853	Heart
7	47	5' 7"	M	14850	Flu
8	49	5' 3"	F	14850	Flu
9	31	5' 6"	F	13053	Cancer
10	37	5' 5"	M	13053	Cancer
11	36	5' 11"	M	13068	Cancer
12	35	6' 1"	M	13068	Cancer
13	41	6' 0"	M	14850	Fracture

Loss of Utility Through Generalization

ID	Age	Height	Gender	Zip	Disease
1	≤ 40	*	*	13053	Heart
4	≤ 40	*	*	13053	Flu
9	≤ 40	*	*	13053	Cancer
10	≤ 40	*	*	13053	Cancer
5	> 40	*	*	1485*	Cancer
6	> 40	*	*	1485*	Heart
7	> 40	*	*	1485*	Flu
8	> 40	*	*	1485*	Flu
13	> 40	*	*	1485*	Fracture
2	≤ 40	*	*	13068	Heart
3	≤ 40	*	*	13068	Flu
11	≤ 40	*	*	13068	Cancer
12	≤ 40	*	*	13068	Cancer

Idea: Publish Additional Tables

[Kifer et al, SIGMOD 2006]

Marginal = GROUP BY View
Anonymized Marginal = GROUP BY View + Generalizations

Height	Count
≤ 5' 5"	4
5' 6" – 5' 11"	5
≥ 6' 0"	4

ID	Age	Height	Gender	Zip	Disease
1	28	5' 5"	F	13053	Heart
2	29	5' 8"	F	13068	Heart
3	21	6' 7"	M	13068	Flu
4	23	5' 9"	F	13053	Flu
5	50	3' 1"	M	14853	Cancer
6	55	6' 0"	M	14853	Heart
7	47	5' 7"	M	14850	Flu
8	49	5' 3"	F	14850	Flu
9	31	5' 6"	F	13053	Cancer
10	37	5' 5"	M	13053	Cancer
11	36	5' 11"	M	13068	Cancer
12	35	6' 1"	M	13068	Cancer
13	41	6' 0"	M	14850	Fracture

Gender	Disease	Count
F	Cancer	1
F	Heart	2
F	Flu	2
M	Cancer	4
M	Heart	1
M	Flu	2
M	Fracture	1

Idea: Publish Additional Tables

Gender	Disease	Count
F	Cancer	1
F	Heart	2
F	Flu	2
M	Cancer	4
M	Heart	1
M	Flu	2
M	Fracture	1

Zip	Disease	Count
13068	Cancer	2
13068	Flu	1
13068	Heart	1
13053	Flu	1
13053	Heart	1
13053	Cancer	2
1485*	Heart	1
1485*	Cancer	1
1485*	Flu	2
1485*	Fracture	1

Height	Count
≤ 5' 5"	4
5' 6" – 5' 11"	5
≥ 6' 0"	4

Age	Count
20-30	4
31-40	4
> 41	5

Recall: Utility of a Single Table

- Generalization height
- Measures based on group sizes:
  - Discernibility
  - Average group size
- Goal-oriented measures:
  - Classification metric
  - Information Gain/Privacy Loss
  - Workload aware metrics
- Needed:
  - General purpose utility metric
  - Aware of tuple distribution (measures information)

Table as a Probability Distribution

$P_T($ Age=33, Height=5' 5", Gender=F, Zip=14850, Disease=Measles $)=1/13$

ID	Age	Height	Gender	Zip	Disease
1	33	5' 5"	F	14850	Measles
2	26	5' 8"	M	14853	Allergy
3	22	6' 7"	M	14853	Gout
4	32	5' 9"	F	14853	Cancer
5	48	3' 1"	M	14850	Flu
6	47	6' 0"	M	14850	Heart
7	46	5' 7"	M	14850	Flu
8	53	5' 3"	F	14853	Cancer
9	51	5' 6"	F	14853	Heart
10	24	5' 5"	M	13063	Flu
11	38	5' 11"	M	13063	Cancer
12	38	6' 1"	M	13068	Cancer
13	30	6' 0"	F	13068	Heart

Marginals as Constraints

- PM(Zip=14850, Disease=Flu) = 2/13
- PM(Gender=M, Disease=Flu)=3/13
- PM: Maximum Entropy Distribution

Zip	Disease	Count
1306*	Cancer	2
1306*	Flu	1
1306*	Heart	1
14850	Flu	2
14850	Heart	1
14850	Measles	1
14853	Allergy	1
14853	Cancer	2
14853	Gout	1
14853	Heart	1

Gender	Disease	Count
F	Cancer	2
F	Heart	2
F	Measles	1
M	Allergy	1
M	Cancer	2
M	Flu	3
M	Gout	1
M	Heart	1

Background: Utility Measure

- PM – maximum entropy distribution consistent with marginals.
- PT – probability distribution represented by original table.
- Utility: distance between PT and PM
  - KL-divergence: $\sum_x P_T(x) \log\left(P_T(x)/P_M(x)\right)$
  - Additional interpretation in terms of likelihood and conditional independence in loglinear models.

Background: Loglinear Models

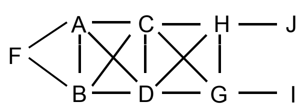
- Loglinear models
- Attributes: A, B, C, D, E, F, G
  - xABC is projection of x onto attributes A,B,C
- Expected count for cell x is modeled as:

$$\log m_x = u_{ABC}(x_{ABC}) + u_{ACE}(x_{ACE}) + u_{DEFG}(x_{DEFG})$$

- Interaction terms can be computed from corresponding marginals (ACE, ABC, DEFG).

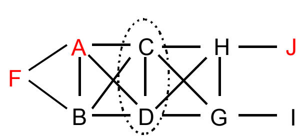
Conditional Independence

$$\log(m_x) = u_{ABF} + u_{ABCD} + u_{CDHG} + u_{HJ} + u_{GI}$$



Conditional Independence

$$\log(m_x) = u_{ABF} + u_{ABCD} + u_{CDHG} + u_{HJ} + u_{GI}$$



$$P(A,F,J \mid C,D) = P(A,F \mid C,D) * P(J \mid C,D)$$

Summary: Utility Measure

- P_M – maximum entropy distribution consistent with marginals.
- P_T – probability distribution represented by original table.
- Utility: distance between P_T and P_M
  - KL-divergence: $\sum_x P_T(x)\log\left(P_T(x)/P_M(x)\right)$
  - Equivalent to selecting marginals where loglinear model has highest likelihood.

Privacy: Example - Maxent

- Given these marginals, the maximum entropy distribution is ...

A	Disease	Count
a_1	Flu	2
a_1	Cancer	3
a_2	Flu	3
a_2	Cancer	4

B	Disease	Count
b_1	Flu	3
b_1	Cancer	5
b_2	Flu	2
b_2	Cancer	2

Example - Maxent

- Given these marginals, the maximum entropy distribution is:

A	Disease	Count
a_1	Flu	2
a_1	Cancer	3
a_2	Flu	3
a_2	Cancer	4

B	Disease	Count
b_1	Flu	3
b_1	Cancer	5
b_2	Flu	2
b_2	Cancer	2

A	B	Disease	Probability
a_1	b_1	Flu	0.1000
a_1	b_1	Cancer	0.1786
a_1	b_2	Flu	0.0667
a_1	b_2	Cancer	0.0714
a_2	b_1	Flu	0.1500
a_2	b_1	Cancer	0.2381
a_2	b_2	Flu	0.1000
a_2	b_2	Cancer	0.0952

Example (Contd.)

A	B	Disease	Probability
a_1	b_1	Flu	0.1000
a_1	b_1	Cancer	0.1786
a_1	b_2	Flu	0.0667
a_1	b_2	Cancer	0.0714
a_2	b_1	Flu	0.1500
a_2	b_1	Cancer	0.2381
a_2	b_2	Flu	0.1000
a_2	b_2	Cancer	0.0952

Conditioning on $A=a_1$, $B=b_1$

$$P(Flu \mid a_1, b_1) = 0.359$$

$$P(Cancer \mid a_1, b_1) = 0.641$$

Extending L-Diversity

- Option 1: For each point t in the domain of nonsensitive attributes
  - Maxent distribution is L-diverse.
  - Reflects the bias in selecting maxent distribution that best approximates the original data.
- Option 2: Random worlds:
  - Random world [Bacchus et al '93] is a possibility consistent with our knowledge.
  - Assume each consistent assignment of attributes (random world) is equally likely.
  - This gives a probability distribution over tuples.
  - Is resulting distribution L-diverse?

Example – Random Worlds

A	Disease	Count
a_1	Flu	2
a_1	Cancer	3
a_2	Flu	3
a_2	Cancer	4

B	Disease	Count
b_1	Flu	3
b_1	Cancer	5
b_2	Flu	2
b_2	Cancer	2

- Bob is in the table
- 58,212,000 random worlds
- 5,821,200 random worlds where Bob has (a_1, b_1, Flu).
- 10,395,500 random worlds where Bob has (a_1, b_1, Cancer)
- Given that Bob has $A=a_1$, $B=b_1$

$$P(Flu \mid a_1, b_1) = 0.359$$

$$P(Cancer \mid a_1, b_1) = 0.641$$

Maxent & Random Worlds

- Generally give different probability distributions.
- Asymptotically (as $N \to \infty$) the probabilities are the same. [Jaynes '82]
- Under certain conditions, answers are the same for finite N (depends on the structure of the marginals).

Algorithm

- For arbitrary collections of anonymized marginals
  - **Utility**: finding maxent distribution requires variants of iterative scaling (can be slow).
  - **Privacy**: checking for privacy is NP-hard.
    - Follows from [De Loera et al '04]
- But: Restrict allowable sets of anonymized marginals.
- Use ***decomposable*** marginals.
- Benefits:
  - **Utility**: closed form maxent probabilities.
  - **Privacy**: tractable.
  - Maxent and random worlds options are equivalent.

Summary: Releasing Additional Marginals

- Utility:
  - Maximum entropy
  - KL-divergence
- Privacy:
  - Extensions of L-diversity
    - Maximum entropy view
    - Random worlds view
- Works for decomposable marginals

Trusted Data Collector: Weak Adversaries

- Privacy and utility metrics
- Algorithms
- Increasing utility through release of additional data
- Releasing temporally changing data
- The minimality attack and simulatable auditing
- Privacy for social networks
- Active attacks on social networks

Multiple Releases: Motivating Example

- Bob was hospitalized in Mar. 2009

Name	Age	Zipcode
Bob	21	12000

G. ID	Age	Zipcode	Disease
1	[21, 22]	[12k, 14k]	dyspepsia
1	[21, 22]	[12k, 14k]	bronchitis
2	[23, 24]	[18k, 25k]	flu
2	[23, 24]	[18k, 25k]	gastritis
3	[36, 41]	[20k, 27k]	flu
3	[36, 41]	[20k, 27k]	gastritis
4	[37, 43]	[26k, 35k]	dyspepsia
4	[37, 43]	[26k, 35k]	flu
4	[37, 43]	[26k, 35k]	gastritis
5	[52, 56]	[33k, 34k]	dyspepsia
5	[52, 56]	[33k, 34k]	gastritis

2-diverse Generalization *T**(1)

Motivating Example

- One month later, in May 2009

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	18000	flu
David	23	25000	gastritis
Gary	41	20000	flu
Helen	36	27000	gastritis
Jane	37	33000	dyspepsia
Ken	40	35000	flu
Linda	43	26000	gastritis
Paul	52	33000	dyspepsia
Steve	56	34000	gastritis

Microdata *T*(1)

Motivating Example

- One month later, in May 2009
- Some obsolete tuples are deleted from the microdata.

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
~~Alice~~	~~22~~	~~14000~~	~~bronchitis~~
~~Andy~~	~~24~~	~~18000~~	~~flu~~
David	23	25000	gastritis
Gary	41	20000	flu
~~Helen~~	~~36~~	~~27000~~	~~gastritis~~
Jane	37	33000	dyspepsia
~~Ken~~	~~40~~	~~35000~~	~~flu~~
Linda	43	26000	gastritis
~~Paul~~	~~52~~	~~33000~~	~~dyspepsia~~
Steve	56	34000	gastritis

Microdata *T*(1)

Motivating Example

- Bob's tuple stays.

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Gary	41	20000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Steve	56	34000	gastritis

Microdata *T*(1)

Motivating Example

- Some new records are inserted.

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Gary	41	20000	flu
Mary	46	30000	gastritis
Ray	54	31000	dyspepsia
Steve	56	34000	gastritis
Tom	60	44000	gastritis
Vince	65	36000	flu

Microdata *T*(2)

Motivating Example

- May 2009: The hospital publishes *T**(2).

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Gary	41	20000	flu
Mary	46	30000	gastritis
Ray	54	31000	dyspepsia
Steve	56	34000	gastritis
Tom	60	44000	gastritis
Vince	65	36000	flu

Microdata *T*(2)

G. ID	Age	Zipcode	Disease
1	[21, 23]	[12k, 25k]	dyspepsia
1	[21, 23]	[12k, 25k]	gastritis
2	[25, 43]	[21k, 33k]	flu
2	[25, 43]	[21k, 33k]	dyspepsia
3	[25, 43]	[21k, 33k]	gastritis
3	[41, 46]	[20k, 30k]	flu
4	[41, 46]	[20k, 30k]	gastritis
4	[54, 56]	[31k, 34k]	dyspepsia
4	[54, 56]	[31k, 34k]	gastritis
5	[60, 65]	[36k, 44k]	gastritis
5	[60, 65]	[36k, 44k]	flu

2-diverse Generalization *T**(2)

Motivating Example

- Consider the previous adversary.

Name	Age	Zipcode
Bob	21	12000

G. ID	Age	Zipcode	Disease
1	[21, 23]	[12k, 25k]	dyspepsia
1	[21, 23]	[12k, 25k]	gastritis
2	[25, 43]	[21k, 33k]	flu
2	[25, 43]	[21k, 33k]	dyspepsia
3	[25, 43]	[21k, 33k]	gastritis
3	[41, 46]	[20k, 30k]	flu
4	[41, 46]	[20k, 30k]	gastritis
4	[54, 56]	[31k, 34k]	dyspepsia
4	[54, 56]	[31k, 34k]	gastritis
5	[60, 65]	[36k, 44k]	gastritis
5	[60, 65]	[36k, 44k]	flu

2-diverse Generalization *T**(2)

Motivating Example

- What the adversary learns from *T**(1).

Name	Age	Zipcode
Bob	21	12000

G. ID	Age	Zipcode	Disease
1	[21, 22]	[12k, 14k]	dyspepsia
1	[21, 22]	[12k, 14k]	bronchitis

......

- What the adversary learns from *T**(2).

Name	Age	Zipcode
Bob	21	12000

G. ID	Age	Zipcode	Disease
1	[21, 23]	[12k, 25k]	dyspepsia
1	[21, 23]	[12k, 25k]	gastritis

......

- So Bob must have contracted dyspepsia!

The *Critical Absence* Phenomenon

What the adversary learns from *T**(1)

Name	Age	Zipcode
Bob	21	12000

G. ID	Age	Zipcode	Disease
1	[21, 22]	[12k, 14k]	dyspepsia
1	[21, 22]	[12k, 14k]	bronchitis

……

Microdata *T*(2)

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Gary	41	20000	flu
Mary	46	30000	gastritis
Ray	54	31000	dyspepsia
Steve	56	34000	gastritis
Tom	60	44000	gastritis
Vince	65	36000	flu

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
Linda	43	26000	gastritis
Gary	41	20000	flu
Mary	46	30000	gastritis
Ray	54	31000	dyspepsia
Steve	56	34000	gastritis
Tom	60	44000	gastritis
Vince	65	36000	flu

Microdata *T*(2)

Name	Group-ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
*c*1	1	[21, 22]	[12k, 14k]	bronchitis
David	2	[23, 25]	[21k, 25k]	gastritis
Emily	2	[23, 25]	[21k, 25k]	flu
Jane	3	[37, 43]	[26k, 33k]	dyspepsia
*c*2	3	[37, 43]	[26k, 33k]	flu
Linda	3	[37, 43]	[26k, 33k]	gastritis
Gary	4	[41, 46]	[20k, 30k]	flu
Mary	4	[41, 46]	[20k, 30k]	gastritis
Ray	5	[54, 56]	[31k, 34k]	dyspepsia
Steve	5	[54, 56]	[31k, 34k]	gastritis
Tom	6	[60, 65]	[36k, 44k]	gastritis
Vince	6	[60, 65]	[36k, 44k]	flu

Counterfeited generalization *T**(2)

Group-ID	Count
1	1
3	1

The auxiliary relation *R*(2) for *T**(2)

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
Alice	1	[21, 22]	[12k, 14k]	bronchitis
Andy	2	[23, 24]	[18k, 25k]	flu
David	2	[23, 24]	[18k, 25k]	gastritis
Gary	3	[36, 41]	[20k, 27k]	flu
Helen	3	[36, 41]	[20k, 27k]	gastritis
Jane	4	[37, 43]	[26k, 35k]	dyspepsia
Ken	4	[37, 43]	[26k, 35k]	flu
Linda	4	[37, 43]	[26k, 35k]	gastritis
Paul	5	[52, 56]	[33k, 34k]	dyspepsia
Steve	5	[52, 56]	[33k, 34k]	gastritis

Generalization *T**(1)

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
*c*1	1	[21, 22]	[12k, 14k]	bronchitis
David	2	[23, 25]	[21k, 25k]	gastritis
Emily	2	[23, 25]	[21k, 25k]	flu
Jane	3	[37, 43]	[26k, 33k]	dyspepsia
*c*2	3	[37, 43]	[26k, 33k]	flu
Linda	3	[37, 43]	[26k, 33k]	gastritis
Gary	4	[41, 46]	[20k, 30k]	flu
Mary	4	[41, 46]	[20k, 30k]	gastritis
Ray	5	[54, 56]	[31k, 34k]	dyspepsia
Steve	5	[54, 56]	[31k, 34k]	gastritis
Tom	6	[60, 65]	[36k, 44k]	gastritis
Vince	6	[60, 65]	[36k, 44k]	flu

Counterfeited Generalization *T**(2)

Group-ID	Count
1	1
3	1

The auxiliary relation *R*(2) for *T**(2)

Name	Age	Zipcode
Bob	21	12000

Re-Publishing: Setting

- A dynamic microdata table *T*.
- Denote the snapshot of *T* at time *j* as *T*(*j*).
- *n* – 1 counterfeited generalizations {*T**(1), *R*(1)}, …, {*T**(*n*-1), *R*(*n*-1)} have been published.
- Problem: given *T*(*n*), to compute a counterfeited generalization {*T**(*n*), *R*(*n*)} of *T*(*n*), such that
  the publication of {*T**(*n*), *R*(*n*)} incurs a small risk of privacy disclosure.

Adversary Model

- The adversary has the following background knowledge:
  - the identity and the QI values of each individual, as well as the time his/her tuple is inserted into (deleted from) *T*;

Adversary Model

- The adversary has the following background knowledge:
  - the identity and the QI values of each individual, as well as the time his/her tuple is inserted into (deleted from) *T*;
- For instance, in our running example

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	18000	flu
David	23	25000	gastritis
...	…	…	…

Microdata *T*(1)

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
...	…	…	…

Microdata *T*(2)

Adversary Model

- The adversary has the following background knowledge:
  - the identity and the QI values of each individual, as well as the time his/her tuple is inserted into (deleted from) *T*;
- For instance, in our running example

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
Alice	22	14000	bronchitis
Andy	24	18000	flu
David	23	25000	gastritis
...	...	...	...

Microdata *T*(1)

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia
David	23	25000	gastritis
Emily	25	21000	flu
Jane	37	33000	dyspepsia
...	...	...	...

Microdata *T*(2)

Adversary Model

- The adversary has the following background knowledge:
  - the identity and the QI values of each individual, as well as the time his/her tuple is inserted into (deleted from) *T*;
  - the generalization principle adopted by the data publisher.

Evaluation of Disclosure Risk

- Let *B* denote the background knowledge of the adversary.
- Let *o* be an individual with a sensitive value *v*.
- $risk(o) = \Pr(o \text{ has } v \mid T^*(1), R(1)\ldots, T^*(n), R(n), B)$.

Name	Age	Zipcode	Disease
Bob	21	12000	dyspepsia

- The disclosure risk for Bob:
- $risk(\text{Bob}) = \Pr(\text{Bob has dyspepsia} \mid T^*(1), R(1), T^*(2), R(2), B)$
- Objective: for each individual *o*, *risk*(*o*) <= a threshold

m-Uniqueness

- A generalized table $T^*(j)$ is *m*-unique, if and only if
  - each QI-group in $T^*(j)$ contains at least *m* tuples
  - all tuples in the same QI-group have different sensitive values.

G. ID	Age	Zipcode	Disease
1	[21, 22]	[12k, 14k]	dyspepsia
1	[21, 22]	[12k, 14k]	bronchitis
2	[23, 24]	[18k, 25k]	flu
2	[23, 24]	[18k, 25k]	gastritis
3	[36, 41]	[20k, 27k]	flu
3	[36, 41]	[20k, 27k]	gastritis
4	[37, 43]	[26k, 35k]	dyspepsia
4	[37, 43]	[26k, 35k]	flu
4	[37, 43]	[26k, 35k]	gastritis
5	[52, 56]	[33k, 34k]	dyspepsia
5	[52, 56]	[33k, 34k]	gastritis

A 2-unique generalized table

Signature

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
Alice	1	[21, 22]	[12k, 14k]	bronchitis
…	…	…	…	…
Jane	4	[37, 43]	[26k, 35k]	dyspepsia
Ken	4	[37, 43]	[26k, 35k]	flu
Linda	4	[37, 43]	[26k, 35k]	gastritis
…	…	…	…	…

$T^*(1)$

- The signature of Bob in $T^*(1)$ is {dyspepsia, bronchitis}
- The signature of Jane in $T^*(1)$ is {dyspepsia, flu, gastritis}

The *m*-invariance Principle

- A sequence of generalized tables $T^*(1)$, …, $T^*(n)$ is m-invariant, if and only if
  - $T^*(1)$, …, $T^*(n)$ are *m*-unique, and
  - each individual has the same signature in every generalized table s/he is involved.

- A sequence of generalized tables $T^*(1), \ldots, T^*(n)$ is m-invariant, if and only if
  - $T^*(1), \ldots, T^*(n)$ are *m*-unique, and
  - each individual has the same signature in every generalized table s/he is involved.

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
Alice	1	[21, 22]	[12k, 14k]	bronchitis
Andy	2	[23, 24]	[18k, 25k]	flu
David	2	[23, 24]	[18k, 25k]	gastritis
Gary	3	[36, 41]	[20k, 27k]	flu
Helen	3	[36, 41]	[20k, 27k]	gastritis
Jane	4	[37, 43]	[26k, 35k]	dyspepsia
Ken	4	[37, 43]	[26k, 35k]	flu
Linda	4	[37, 43]	[26k, 35k]	gastritis
Paul	5	[52, 56]	[33k, 34k]	dyspepsia
Steve	5	[52, 56]	[33k, 34k]	gastritis

Generalization $T^*(1)$

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
c1	1	[21, 22]	[12k, 14k]	bronchitis
David	2	[23, 25]	[21k, 25k]	gastritis
Emily	2	[23, 25]	[21k, 25k]	flu
Jane	3	[37, 43]	[26k, 33k]	dyspepsia
c2	3	[37, 43]	[26k, 33k]	flu
Linda	3	[37, 43]	[26k, 33k]	gastritis
Gary	4	[41, 46]	[20k, 30k]	flu
Mary	4	[41, 46]	[20k, 30k]	gastritis
Ray	5	[54, 56]	[31k, 34k]	dyspepsia
Steve	5	[54, 56]	[31k, 34k]	gastritis
Tom	6	[60, 65]	[36k, 44k]	gastritis
Vince	6	[60, 65]	[36k, 44k]	flu

Generalization $T^*(2)$

- A sequence of generalized tables $T^*(1), \ldots, T^*(n)$ is m-invariant, if and only if
  - $T^*(1), \ldots, T^*(n)$ are *m*-unique, and
  - each individual has the same signature in every generalized table s/he is involved.

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
Alice	1	[21, 22]	[12k, 14k]	bronchitis
Andy	2	[23, 24]	[18k, 25k]	flu
David	2	[23, 24]	[18k, 25k]	gastritis
Gary	3	[36, 41]	[20k, 27k]	flu
Helen	3	[36, 41]	[20k, 27k]	gastritis
Jane	4	[37, 43]	[26k, 35k]	dyspepsia
Ken	4	[37, 43]	[26k, 35k]	flu
Linda	4	[37, 43]	[26k, 35k]	gastritis
Paul	5	[52, 56]	[33k, 34k]	dyspepsia
Steve	5	[52, 56]	[33k, 34k]	gastritis

Generalization $T^*(1)$

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
c1	1	[21, 22]	[12k, 14k]	bronchitis
David	2	[23, 25]	[21k, 25k]	gastritis
Emily	2	[23, 25]	[21k, 25k]	flu
Jane	3	[37, 43]	[26k, 33k]	dyspepsia
c2	3	[37, 43]	[26k, 33k]	flu
Linda	3	[37, 43]	[26k, 33k]	gastritis
Gary	4	[41, 46]	[20k, 30k]	flu
Mary	4	[41, 46]	[20k, 30k]	gastritis
Ray	5	[54, 56]	[31k, 34k]	dyspepsia
Steve	5	[54, 56]	[31k, 34k]	gastritis
Tom	6	[60, 65]	[36k, 44k]	gastritis
Vince	6	[60, 65]	[36k, 44k]	flu

Generalization $T^*(2)$

- A sequence of generalized tables $T^*(1), \ldots, T^*(n)$ is m-invariant, if and only if
  - $T^*(1), \ldots, T^*(n)$ are *m*-unique, and
  - each individual has the same signature in every generalized table s/he is involved.

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
Alice	1	[21, 22]	[12k, 14k]	bronchitis
Andy	2	[23, 24]	[18k, 25k]	flu
David	2	[23, 24]	[18k, 25k]	gastritis
Gary	3	[36, 41]	[20k, 27k]	flu
Helen	3	[36, 41]	[20k, 27k]	gastritis
Jane	4	[37, 43]	[26k, 35k]	dyspepsia
Ken	4	[37, 43]	[26k, 35k]	flu
Linda	4	[37, 43]	[26k, 35k]	gastritis
Paul	5	[52, 56]	[33k, 34k]	dyspepsia
Steve	5	[52, 56]	[33k, 34k]	gastritis

Generalization $T^*(1)$

Name	G.ID	Age	Zipcode	Disease
Bob	1	[21, 22]	[12k, 14k]	dyspepsia
c1	1	[21, 22]	[12k, 14k]	bronchitis
David	2	[23, 25]	[21k, 25k]	gastritis
Emily	2	[23, 25]	[21k, 25k]	flu
Jane	3	[37, 43]	[26k, 33k]	dyspepsia
c2	3	[37, 43]	[26k, 33k]	flu
Linda	3	[37, 43]	[26k, 33k]	gastritis
Gary	4	[41, 46]	[20k, 30k]	flu
Mary	4	[41, 46]	[20k, 30k]	gastritis
Ray	5	[54, 56]	[31k, 34k]	dyspepsia
Steve	5	[54, 56]	[31k, 34k]	gastritis
Tom	6	[60, 65]	[36k, 44k]	gastritis
Vince	6	[60, 65]	[36k, 44k]	flu

Generalization $T^*(2)$

The *m*-Invariance principle

- Lemma: if a sequence of generalized tables {*T**(1), …, *T**(*n*)} is *m*-invariant, then for any individual *o* involved in any of these tables, we have

  $risk(o) <= 1/m$

The *m*-invariance principle

- Lemma: if {*T**(1), …, *T**(*n*-1)} is *m*-invariant, then {*T**(1), …, *T**(*n*-1), *T**(*n*)} is also *m*-invariant, if and only if {*T**(*n*-1), *T**(*n*)} is *m*-invariant
- Only *T**(*n* - 1) is needed for the generation of *T**(*n*).

*T**(1), *T**(2), …, *T**(*n*-2), *T**(*n*-1), *T**(*n*)

Can be discarded

Algorithm

- Given *T*(*n*), *T**(*n*-1) and a parameter *m*, our algorithm generates a counterfeited generalization *T**(*n*) of *T*(*n*), such that {*T**(1), …, *T**(*n*)} is *m*-invariant.
- Optimization goal: to impose as little amount of generalization as possible.

Trusted Data Collector: Weak Adversaries

- Privacy and utility metrics
- Algorithms
- Increasing utility through release of additional data
- Releasing temporally changing data
- The minimality attack and simulatable auditing
- Privacy for social networks
- Active attacks on social networks