Models and Methods for Privacy-Preserving Data Publishing and Analysis

Johannes Gehrke
Department of Computer Science
http://www.cs.cornell.edu/johannes

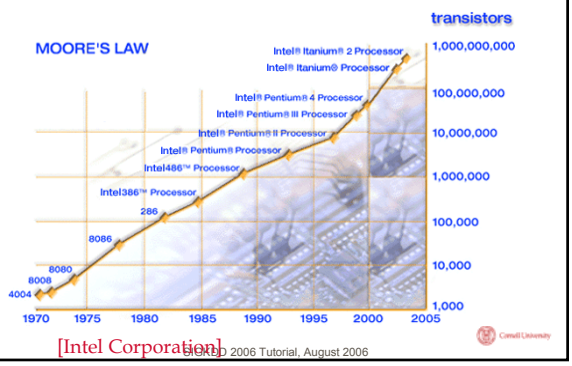
Cornell University

An Abundance of Data

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

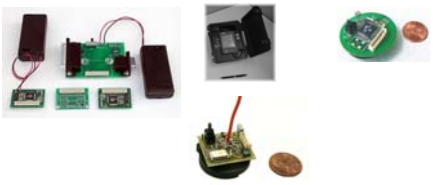
•Print, film, optical, and magnetic storage: 5 Exabytes (EB) of new information in 2002, doubled in the last three years [How much Information 2003, UC Berkeley]

Driving Factors: A LARGE Hardware Revolution

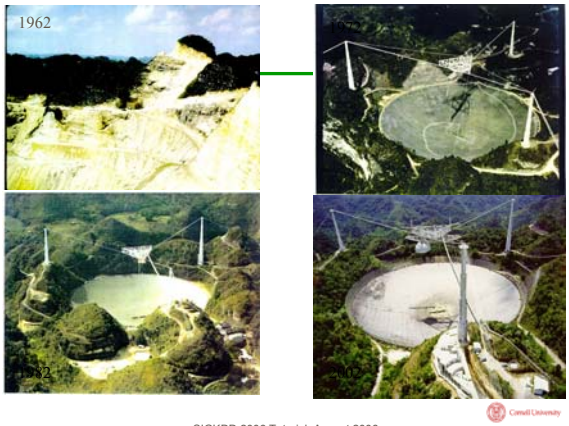


[Intel Corporation]

Driving Factors: A small Hardware Revolution

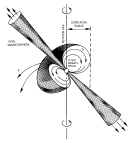


- Experts on ants estimate that there are 10^{16} to 10^{17} ants on earth. In the year 1997, we produced one transistor per ant. [Gordon Moore]



Pulsars

- Pulsars are rotating stars
- Of interest are
  - Millisecond pulsars
  - Compact binaries
- Example:
  - Hulse-Taylor binary
  - Used to infer gravitational waves in support of Einstein's General Theory of Relativity
  - Nobel price in physics in 1993



Project Requirements

- Data
  - 14 TB every 2 weeks
  - Shipped on USB-2 disk drives
  - Need to archive raw data 5+ years
  - Need to make data products available to the astronomy research community
- Processing
  - Extremely processor intensive
  - Find new pulsars --- and other *interesting* phenomena

[Calimlim, Cordes, Demers, Gehrke, Lifka;
http://arecibo.tc.cornell.edu]

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)

Concerns About Privacy

Recent example:

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

[Annalee Newitz, AlterNet, August 15, 2006]

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.

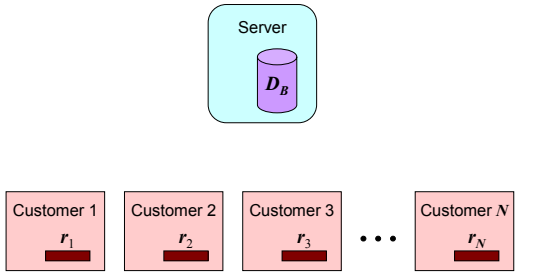
…

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

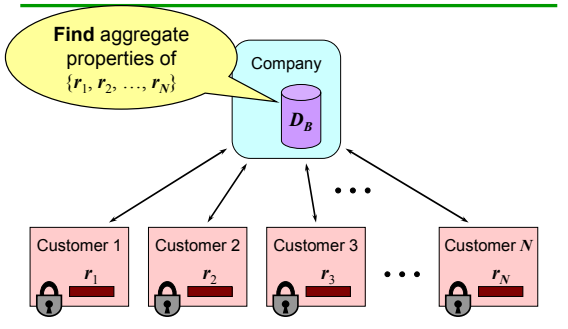
Ms. Arnold says she loves online research, but the disclosure of her searches has left her disillusioned. In response, she plans to drop her AOL subscription. "We all have a right to privacy," she said. "Nobody should have found this all out."

http://data.aolsearchlogs.com

The Setup



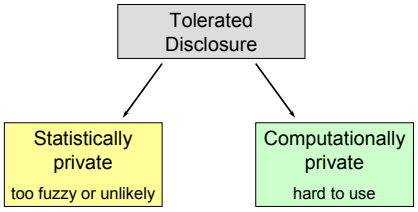
Model I: Untrusted Data Collector



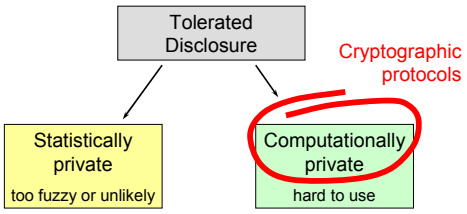
Minimal Information Sharing

- 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)?

Types of Disclosure

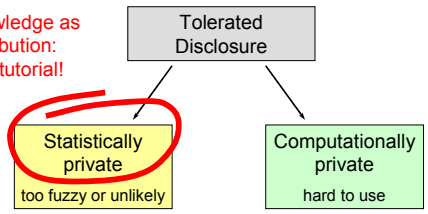


Types of Disclosure

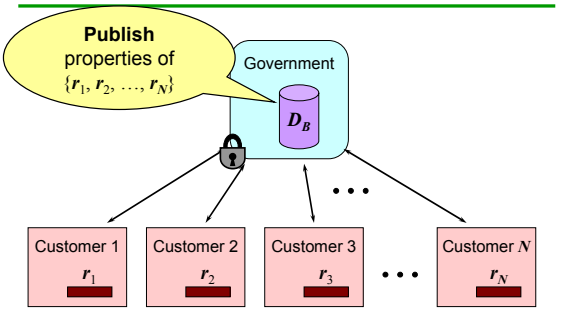


Types of Disclosure

Knowledge as distribution: This tutorial!



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?

This Tutorial: Statistical Methods

- Privacy-preserving data analysis
- Privacy-preserving data publishing

Goal:

- Rather than talk about everything superficially, but nothing in-depth, make hard choices

Caveats:

- 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.)
- Disclosure control (statistical databases, auditing, database queries, etc.)
- Privacy through distributed data mining

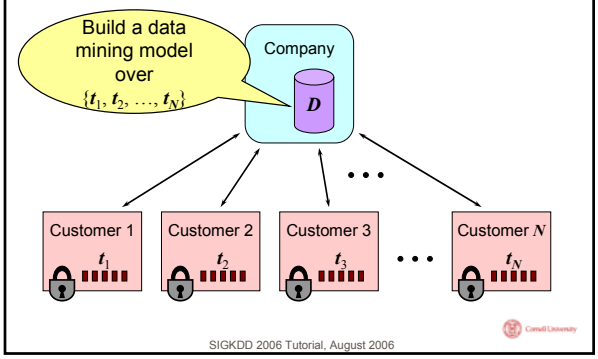
- Resources
  - See excellent tutorials by Rakesh Agrawal and Chris Clifton; keynote talk by Srikant Ramakrishnan at this conference.

Tutorial Outline

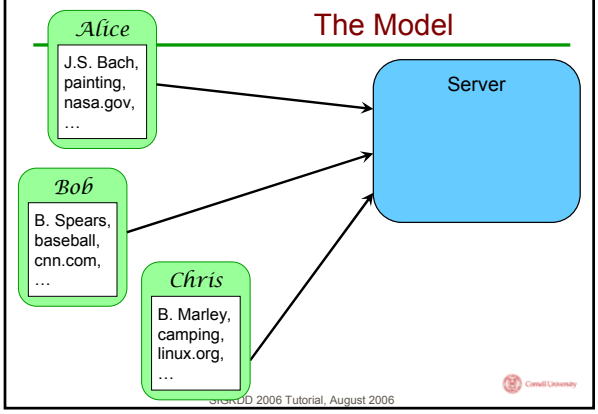
- Untrusted data collector

- Trusted data collector

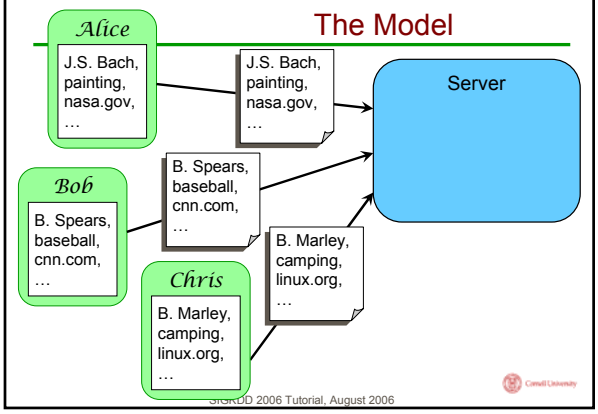
Privacy Preserving Data Mining

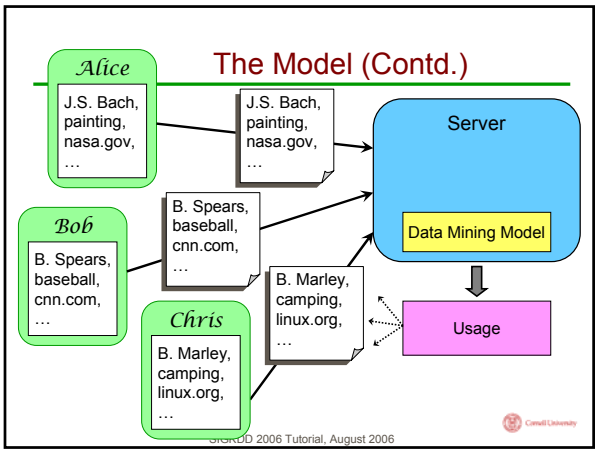


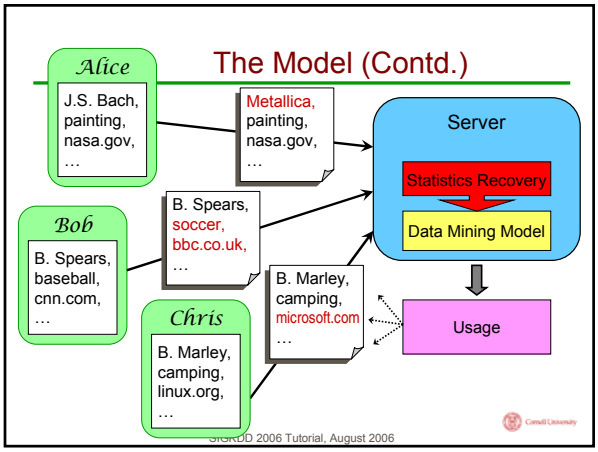
The Model



The Model







The Problem

- How to randomize data such that
  - we can build a good data mining model (utility)
  - while preserving privacy at the record level (privacy)?

SIGKDD 2006 Tutorial, August 2006

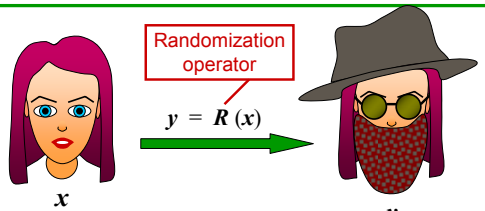
Tutorial Outline

- Untrusted data collector
  - Randomized response [W65]
  - The search for a good privacy definition
    - Interval privacy [AS00]
    - Mutual information [AA01]
    - (α,β) privacy breach [EGS03]
  - Comments
- Trusted data collector

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



Original (private) data

Assumptions:

- Described by a random variable X.
- Each client is independent.

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

Extensions To Randomized Response

- 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

- What is the privacy guaranteed by randomized response?

Tutorial Outline

- Untrusted data collector
  - Randomized response [W65]
  - The search for a good privacy definition
    - Interval privacy [AS00]
    - Mutual information [AA01]
    - (α,β) privacy breach [EGS03]
  - Comments
- Trusted data collector

Interval Privacy [AS00]

[Agrawal and Srikant; SIGMOD 2000]

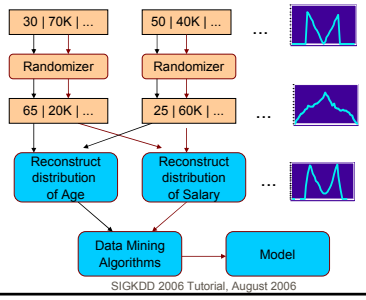
Idea: Clients share randomized version of their data.

Intuition: Randomized response.

Randomization:

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

Interval Privacy: Illustration



Interval Privacy: Example

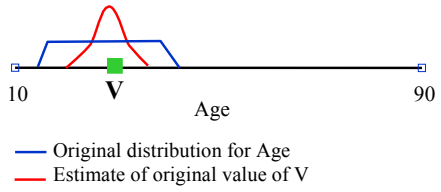
- Add a random value between -30 and +30 to age.
- If randomized value is 60
  - We know with 90% confidence that age is between 33 and 87.
- Interval width is the amount of privacy.
  - Example:
    - Interval width 54 with 90% confidence
    - Interval width 60 with 100% confidence

Interval Privacy: Reconstruction

- Original values $x_1, x_2, \ldots, x_n$
  - Unknown distribution F
- To hide these values, we use $y_1, y_2, \ldots, y_n$
  - From known distribution G
- The problem: Given
  - $z_1=x_1+y_1, z_2=x_2+y_2, \ldots, z_n=x_n+y_n$
  - Distribution G
- Problem:
  - Reconstruct F
  - (Note that F does not have a parametric form → reconstruct $x_1^{hat}, x_2^{hat}, \ldots, x_n^{hat}$?)

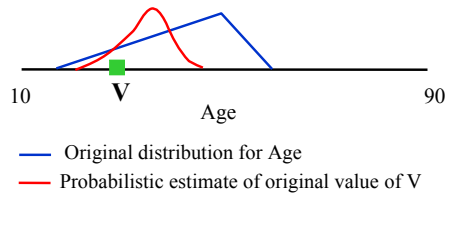
Intuition (Reconstruct Single Point)

- Use Bayes' rule



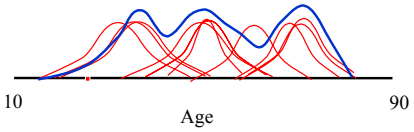
Intuition (Reconstruct Single Point)

- Use Bayes' rule



Reconstructing the Distribution

- Combine estimates of where point came from for all the points
  - Gives estimate of original distribution
  - Similar to a kernel-density estimator



Reconstruction: Iterative Algorithm

f_X^0 := Uniform density
j := 0 // Iteration number
repeat
  $f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y((x_i+y_i)-a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i+y_i)-a) f_X^j(a)}$
  j := j+1
until (stopping criterion met)

- Other approach:
  - Assume parametric distribution
  - Perform MLE of distribution parameters through the EM Algorithm [AA01]

Tutorial Outline

- Untrusted data collector
  - Randomized response [W65]
  - The search for a good privacy definition
    - Interval privacy [AS00]
    - Mutual information [AA01]
    - (α,β) privacy breach [EGS03]
  - Comments
- Trusted data collector

Recall Interval Privacy

- We know that the original value lies within an interval of size w with confidence c.
- Example:
  Add uniform distribution [-30,30]
  - Interval width 54 with 90% confidence
  - Interval width 60 with 100% confidence

An Attack on Interval Privacy [AA01]

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

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

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 $f_X(x)$:

- $f_X(x) = 0.5$, $0 \le x \le 1$
- $f_X(x) = 0.5$, $4 \le x \le 5$
- $f_X(x) = 0$, 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, $\Pi(X)=2^1=2$
  → 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:
  $\Pi(X|Z)=\Pi(X)*(1-Loss(X|Z))=2*(1.0-0.5796)=0.8408$
  → X has only as much privacy as U[0, 0.8408]

Caveat: Privacy Preserved Only On Average

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
- Caveat:
  - Every time this occurs the property "X <= 0.01" is fully disclosed.
  - The mutual information, being an average measure, does not notice this rare disclosure.

Tutorial Outline

- Untrusted data collector
  - Randomized response [W65]
  - The search for a good privacy definition
    - Interval privacy [AS00]
    - Mutual information [AA01]
    - (α,β) privacy breach [EGS03]
  - Comments
- Trusted data collector

Privacy Breaches

[Evfimievski, Gehrke, Srikant; PODS 2003]

A randomization may "look strong" but sometimes fail to hide some items of an individual transaction.

- Simple randomization example: Given a transaction
  - Keep item with 20% probability,
  - Replace with a new random item with 80% probability.

Example: $\{a, b, c\}$

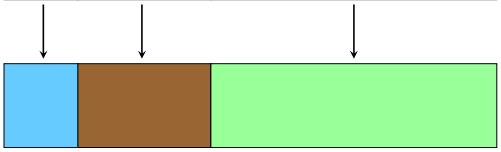
10 M transactions of size 10 with 10 K items:

1% have $\{a, b, c\}$	5% have $\{a, b\}$, $\{a, c\}$, or $\{b, c\}$ only	94% have one or zero items of $\{a, b, c\}$

Example: $\{a, b, c\}$

10 M transactions of size 10 with 10 K items:

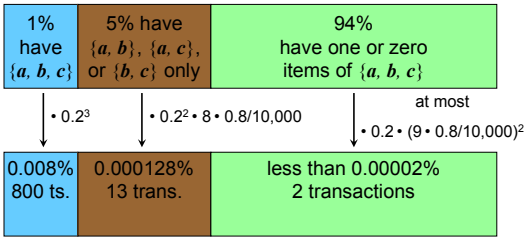
1% have $\{a, b, c\}$	5% have $\{a, b\}$, $\{a, c\}$, or $\{b, c\}$ only	94% have one or zero items of $\{a, b, c\}$



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

Example: $\{a, b, c\}$

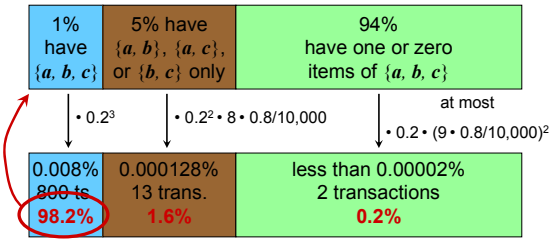
10 M transactions of size 10 with 10 K items:



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}

- Given nothing, we have only 1% probability that {a, b, c} occurs in the original transaction
- Given {a, b, c} in the randomized transaction, we have about 98% certainty of {a, b, c} in the original transaction.
- This is called a privacy breach.
- The example randomization preserves privacy "on average," but not "in the worst case."

An Observation

- A randomization may "look strong" but sometimes fails to hide properties of an individual transaction.

Simple Privacy Breaches

- Suppose the "adversary" wants to know if $z \in t$, where
  - t is an original transaction;
  - t′ is the corresponding randomized transaction;
  - A is an itemset
- Itemset A causes a privacy breach of level β (e.g. 50%) if:

$$\text{Prob}\left[z \in t \mid A \subseteq t'\right] \geq \beta$$

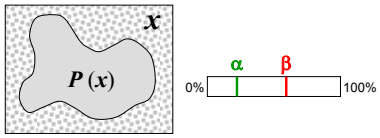
- Knowledge of $A \subseteq t'$ makes a jump from Prob $[z \in t]$ to Prob $[z \in t \mid A \subseteq t']$ (in the adversary's viewpoint).

Privacy Breaches: Goals

- We need a bound for all privacy breaches
  - not only for: item $\in$ t versus itemset $\subseteq$ t′
- No knowledge of data distribution is required in advance
  - We should not need to know Prob [item $\in$ t]
- Applicable to numerical data as well
- Easy to work with, even for complex randomizations

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.



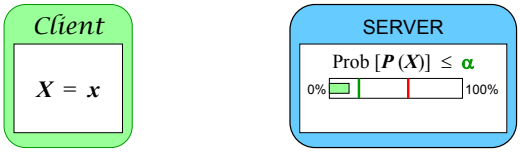
Example:

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

$\alpha = 1\%$ and $\beta = 50\%$

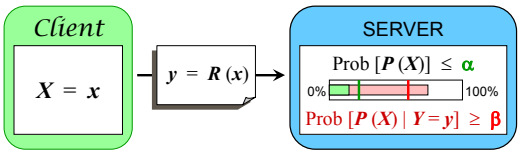
α-to-β Privacy Breach

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



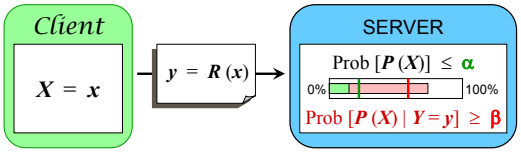
α-to-β Privacy Breach

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 $P(x)$.

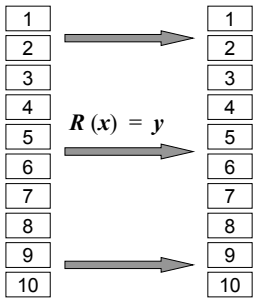
α-to-β Privacy Breach

Checking for α-to-β privacy breaches:

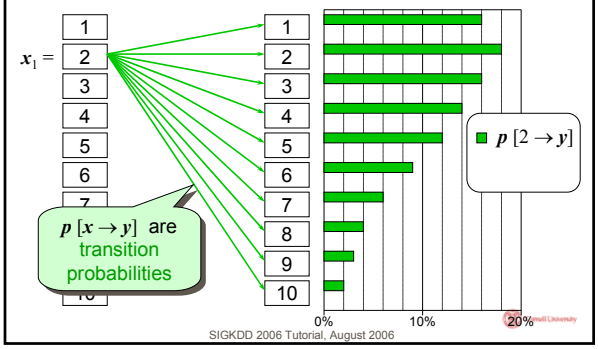
- There are exponentially many properties $P(x)$;
- We have to know the data distribution in order to check whether $\text{Prob}[P(X)] \leq \alpha$ and $\text{Prob}[P(X) \mid Y = y] \geq \beta$

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

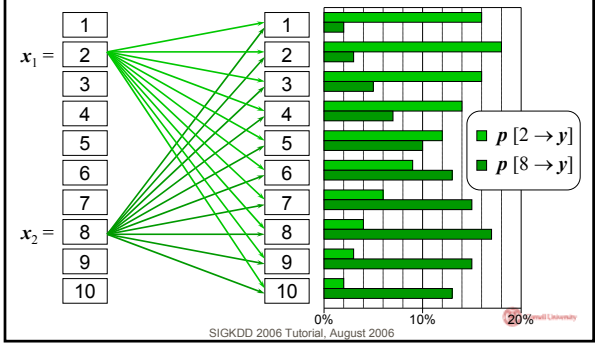
Amplification Condition



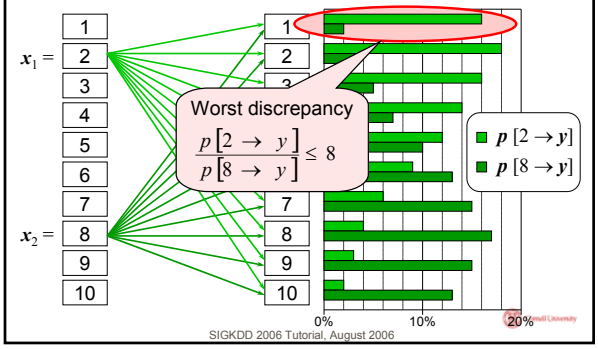
Amplification Condition



Amplification Condition



Amplification Condition



Amplification Condition

Definition:

- Randomization operator $\boldsymbol{R}$ 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 $\boldsymbol{p}[\boldsymbol{x} \to \boldsymbol{y}] = \text{Prob}[\boldsymbol{R}(\boldsymbol{x}) = \boldsymbol{y}]$ depend only on the operator $\boldsymbol{R}$ and not on data.
  - We assume that all $\boldsymbol{y}$ have a nonzero probability.
- The bigger γ is, the more may be revealed about $\boldsymbol{x}$.

The Bound on α-to-β Breaches

Theorem:

- If randomization operator $\boldsymbol{R}$ is at most γ-amplifying, and if:

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

- Then, revealing $\boldsymbol{R}(\boldsymbol{X})$ to the server will never cause an α-to-β privacy breach.

The Bound on α-to-β Breaches

Examples:

- 5%-to-50% privacy breaches do not occur for $\gamma < 19$:

$$\frac{0.5}{0.05} \cdot \frac{1-0.05}{1-0.5} = 19$$

- 1%-to-98% privacy breaches do not occur for $\gamma < 4851$:

$$\frac{0.98}{0.01} \cdot \frac{1-0.01}{1-0.98} = 4851$$

- 50%-to-100% privacy breaches do not occur for any finite γ.

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:

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

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.

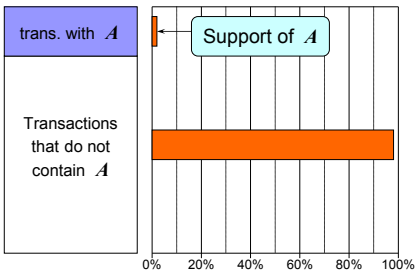
t = a, b, c, d, e, f, u, v, w

t' = b, e, u, w, æ, å, β, §, ψ, €, ϰ, ɔ, ħ, …

$j = 4$ items inserted with prob. ρ

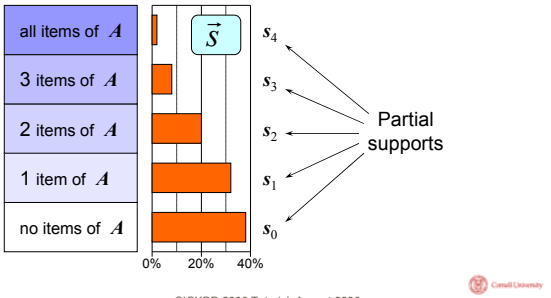
Support Recovery

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



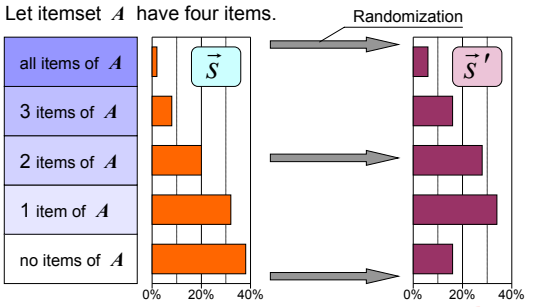
Support Recovery

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



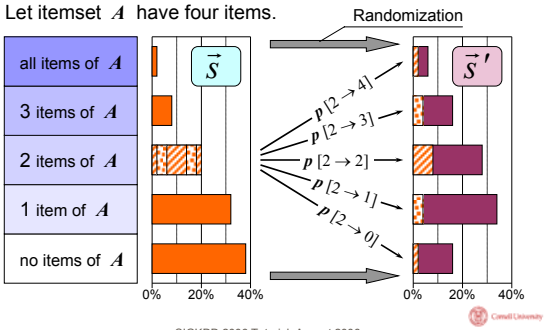
Support Recovery

Let itemset A have four items.



Support Recovery

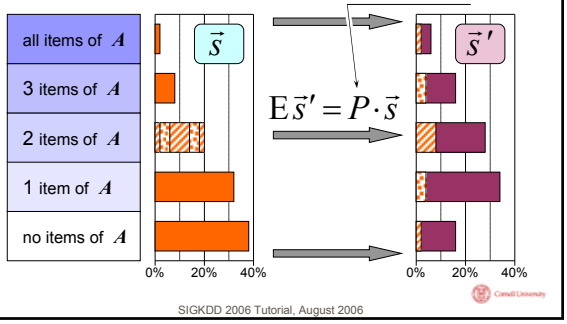
Let itemset A have four items.



Support Recovery

Let itemset A have four items.

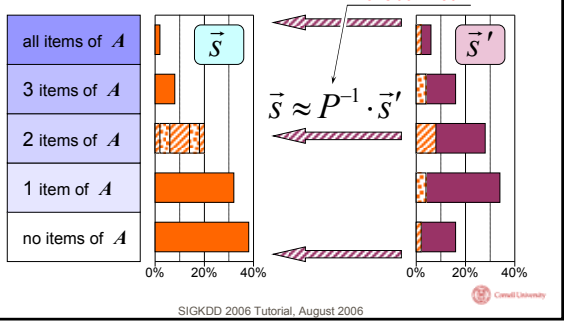
Transition matrix



Support Recovery

Let itemset A have four items.

Transition matrix



The Unbiased Estimators

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

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

- Covariance matrix for this estimator:

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

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

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

Apriori [AS94]

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

Repeat:

1. Count support 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 $\geq s_{min}$
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.

Tutorial Outline

- Untrusted data collector
  - Randomized response [W65]
  - The search for a good privacy definition
    - Interval privacy [AS00]
    - Mutual information [AA01]
    - (α,β) privacy breach [EGS03]
  - Comments
- Trusted data collector

Extensions: FRAPP [AH05]

[Agrawal and Haritsa, ICDE 2005]

- Examines randomization methods based on transition matrices
- Two main ideas:
  - How to design "good" transition matrices
    - Well conditioned matrices
  - Randomize the transition matrix

Extensions: (s,α,β) Privacy Breach [AST05]

[Agrawal, Srikant, Thomas; SIGMOD 2005]

- Consider the following class of randomization operators:
  - Each attribute value is retained with probability p and replaced with probability (1-p) with a value selected from a replacing distribution

Example: Uniform perturbation

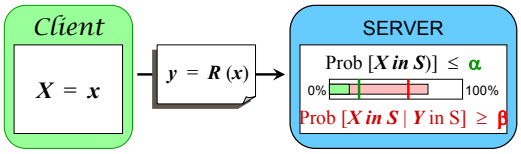
- Replacing distribution is the uniform distribution on the domain

(s,α,β) Privacy Breach (Contd.)

- Consider the following probabilities:
  - $P_f[X \text{ in } S] = p_S$, where P_f is the a priori distribution
  - $P_g[Y \text{ in } S] = m_S$, where P_g is the replacing distribution.
- Define the *relative a priori probability* of event S as p_s/m_S.
- Intuition: How frequent is S in its a priori distribution compared to the replacing distribution?

(s,α,β) Privacy Breach (Contd.)

Let $P(x)$ be any property of a client's private data;
Let $0 < \alpha < \beta < 1$ be two probability thresholds.
If $p_S/m_S < s$ and if the following holds:



Then disclosure of y causes an (s,α,β) privacy breach w.r.t. set S .

(s,α,β) Privacy Breach (Contd.)

- Theorem [AST05]: Uniform perturbation applied to a single column is secure against a (s,α,β) privacy breach if

$$s < \frac{(\beta - \alpha)(1-p)}{(1-\beta)p}$$

(Recall: p is probability not to pick from randomizing distribution)

(s,α,β) Privacy Breach: Issues

- Problem (due to Sasha Evfimievski): Correlation between attributes
- Example:
  - Age (0..99) and YOB (1900..1999)
  - Assume we replace with uniform distribution 90% of the time, and we leave original value 10% of the time
  - P[Age = 30] = 0.01
  - P[Age = 30 | randomized Age = 30] = 0.109
    → 1% to 11% privacy breach
  - P[Age = 30 | randAge = 30 and randYOB = 1976] = 30] ~ 0.6!
    → 1% to 60% privacy breach
- Thus maybe need Prob[X in S | Y in S'] ≥ β?

An Observation About Attribute Correlation

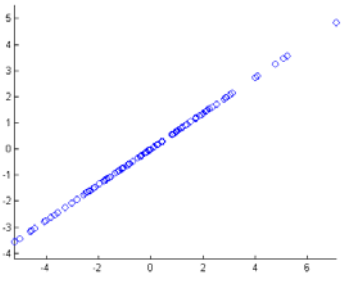
[Huang, Du, Chen; SIGMOD 2005]

- Correlation between attributes can thwart independent random noise
- Example: We cannot perturb the same number for several times.
- If we do that, we can estimate the original data:
  - Let (t,t, …, t) be the original data,
  - Published data: $t + R_1$, $t + R_2$, …, $t + R_m$
  - Let $Z = [(t+R_1)+ \ldots + (t+R_m)] / m$
  - Mean: $E(Z) = t$

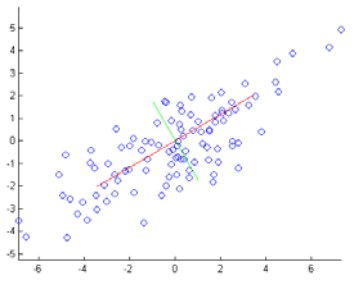
Intuition

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

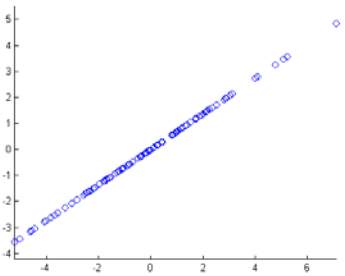
Data



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%

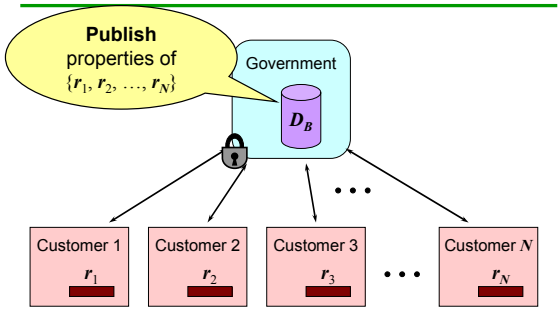
Other Comments

- 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
- Much work on privacy, but what about utility? What about repeated sharing of data? What do I need to do to analyze such data?
- Comparison with secure multiparty computation protocols
- Questions about this model?

Tutorial Outline

- Untrusted data collector
- Trusted data collector
  - Limiting disclosure
  - K-Anonymity
  - L-Diversity

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 compute the "largest" set of queries that can be disclosed while preserving data privacy?
- How do we measure privacy?

Goals

- Safe from attackers who try to learn customers' identities or sensitive information
- Useful for a wide range of statistical analyses
- Easy for users to analyze with standard statistical methods
  - Just load the published dataset into your favorite analysis tool

[Reiter, Chance 17(3), 2004]

Why is Disclosure Bad?

- Violation of laws and thus subject to legal action
- Lose the trust of the public (no future participants)
- Data of dubious quality (since participants are afraid that their privacy is threatened)

[Reiter, Chance 17(3), 2004]

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

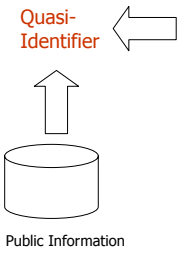
Removing SSN …

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

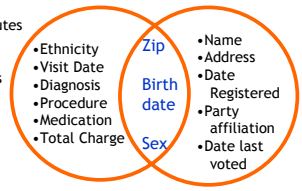


Zip	Age	Nationality	Disease
13053	28	Russian	Heart
13068	29	American	Heart
13068	21	Japanese	Viral
13053	23	American	Viral
14853	50	Indian	Cancer
14853	55	Russian	Heart
14850	47	American	Viral
14850	59	American	Viral
13053	31	American	Cancer
13053	37	Indian	Cancer
13068	36	Japanese	Cancer
13068	32	American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



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 [Sweeney02]

- 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
```

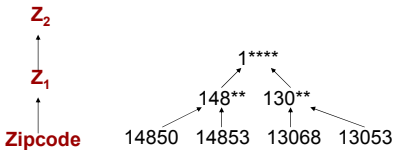
  is ≥ k
- Parameter k indicates the "degree" of anonymity

K-Anonymity

- There are at least k tuples sharing the same values for each combination of the quasi-identifiers.
- Techniques
  - Generalizing non-sensitive attributes
  - Tuple Suppression
  - Data Swapping
  - Randomization

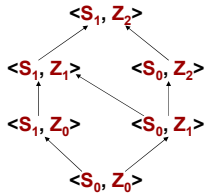
K-Anonymity Through Generalization

- Generalization functions induce value generalization hierarchies
- Corresponding domain generalization hierarchies



Generalization: Multiple Attributes

- Cross-product lattice

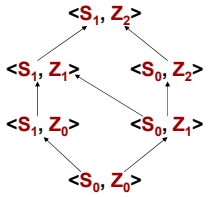


K-Anonymity Algorithms

- Optimal Full-Domain Algorithms [Sa01]
  - Binary Search of the lattice finds solution of minimum height
- Optimal Algorithms:
  - Bayardo-Agrawal [BA05]
  - LeFevre-DeWitt-Ramakrishnan [LDR05, LDR06]
- Heuristic Algorithms
  - Greedy Heuristic Search [Sw02-2, FWY05, WYC04]
  - No guarantees about optimality
- Stochastic Search
  - Genetic Algorithms [Iy02]
  - Simulated Annealing [Wi02]
  - Long run times to convergence; do not guarantee optimality
- Approximation Algorithms
  - Cell-suppression [MW04, AFKM+05]
  - Have not been implemented

Incognito [LDR05]

- Intuition: Apriori-Algorithm/Cube computation
- We discuss Incognito [LeFevre, DeWitt, Ramakrishnan; SIGMOD 2005]



Some Simple Observations

- Generalization Property: If T is k-anonymous with respect to a set of attributes, then it is k-anonymous with respect to any generalization of these attributes.

Hospital Patients

DOB	Sex	Zipcode	Disease
1/21/76	Male	5371*	Flu
1/21/76	Male	5370*	Broken Arm
2/28/76	Male	5370*	Bronchitis
4/13/86	Female	5371*	Hepatitis
4/13/86	Female	5370*	Sprained Ankle
2/28/86	Female	5370*	Hang Nail

Some Simple Observations

- Generalization Property
- Rollup Property: If attribute set P is a generalization of Q, counts grouped by P can be computed directly from the counts grouped by Q.

Hospital Patients

DOB	Sex	Zipcode	Disease
1/21/76	Male	5371*	Flu
1/21/76	Male	5370*	Broken Arm
2/28/76	Male	5370*	Bronchitis
4/13/86	Female	5371*	Hepatitis
4/13/86	Female	5370*	Sprained Ankle
2/28/86	Female	5370*	Hang Nail

Some Simple Observations

- Generalization Property
- Rollup Property
- Subset Property: If T is k-anonymous with respect to attribute set Q, then T is k-anonymous with respect to P ⊆ Q.

Hospital Patients

DOB	Sex	Zipcode	Disease
1/21/76	Male	537**	Flu
1/21/76	Male	537**	Broken Arm
2/28/76	Male	537**	Bronchitis
4/13/86	Female	537**	Hepatitis
4/13/86	Female	537**	Sprained Ankle
2/28/86	Female	537**	Hang Nail

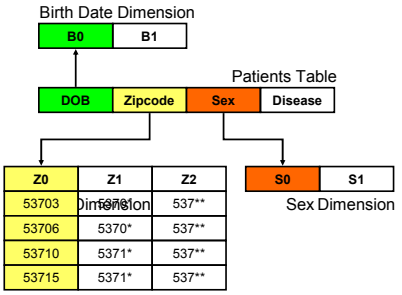
Some Simple Observations

- Generalization Property
- Rollup Property

→ **Data Cube**

- Subset Property → **Frequent Itemsets**

Star-Schema Representation for Generalization Hierarchies



Basic Incognito Algorithm

- Finds all k-anonymous full-domain generalizations
- Begins by checking k-anonymity with respect to single-attribute subsets of quasi-identifier. Then iteratively checks larger subsets. (*Subset Property*)
- Each iteration has two phases:
  - Breadth-first search (*Rollup Property*)
  - Candidate graph construction

Condensation Approach to Privacy [AY04]

[Aggarwal and Yu; EDBT 2004]

- K-indistinguishability: For any given record, there are at least k records in the dataset from which it cannot be distinguished.
- Idea:
  - Cluster the data into groups of k records
  - Compute sufficient statistics for some distribution for each cluster
  - Sample from this distribution

Taxonomy of K-Anonymization [LDR05]

- Generalization versus suppression versus data swapping
- Global versus local recoding
- Hierarchy versus partition-based generalization
- Other very interesting theoretical work (just in its beginning).
- However, k-anonymity has its problems…

Tutorial Outline

- Untrusted data collector
- Trusted data collector
  - K-Anonymity
  - L-Diversity

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

[Ohrn, Ohno-Machado; Artif Intell Med. 15(3), 1999]
[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

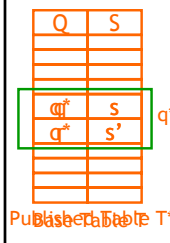
- Alice's friend Umeko is in the table.
- Alice knows Umeko is 24, a Japanese, living in Freeville (13068)
- Japanese have *extremely low incidence* of heart disease

Alice learns Umeko has a viral infection

Data Publishing Desiderata

- Need to defend against attacks based on background knowledge
- Need to permit efficient sanitization algorithms
- Guarantee understood by a lay person

Incorporating Background Knowledge



- Worst-case assumption: Adversary has *full knowledge of the joint distribution of the attributes*.
- Prior Belief:
  P[t[S] = s | t[Q] = q] = f(s|q)
- Posterior Belief:
  P[t[S] = s | t[Q] = q & t* ε T*]

$$= \frac{n_{sq^*}\frac{f(s,q)}{f(s,q^*)}}{\sum_{s'} n_{s'q^*}\frac{f(s',q)}{f(s',q^*)}}$$

Privacy Definition (1)

- Positive Disclosure: Posterior Belief > 1-δ
- Negative Disclosure: Posterior Belief < δ

BUT:

- Not all positive disclosures are bad
  - OK to disclose Bob is healthy
- Not all negative disclosures are bad
  - OK to disclose Bob does not have Ebola

Privacy Definition (2)

- Bayes-optimal privacy: After publishing we have
  Posterior belief ~ prior belief
- Example instantiation: α-to-β privacy breach definition
  Prior Belief < α and posterior Belief > β **OR**
  Prior Belief >1- α and posterior Belief <1-β
- Automatically eliminates homogeneity attack
  - Homogeneity → Posterior belief = 1

Bayes-Optimal Privacy– Drawbacks

- Insufficient knowledge
  - Nobody knows the complete joint distribution
- Adversary's knowledge unknown
  - Data publisher does not know how much the adversary knows
- Computational intractability
  - Checking for every (q,s) pair ...

Towards A Practical Definition (1)

- Posterior belief = $$\frac{n_{sq^*}\frac{f(s,q)}{f(s,q^*)}}{\sum_{s'} n_{s'q^*}\frac{f(s',q)}{f(s',q^*)}}$$
- Homegeneity attack

Q	S
q*	s
q*	s
q*	s
q*	s
q*	s'

$$\forall s' \neq s, n_{sq^*} >> n_{s'q^*}$$

Towards A Practical Definition (2)

- Posterior belief = $$\frac{n_{sq^*}\frac{f(s,q)}{f(s,q^*)}}{\sum_{s'} n_{s'q^*}\frac{f(s',q)}{f(s',q^*)}}$$
- Background knowledge attack

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

$$\forall s' \neq s, \frac{f(s',q)}{f(s',q^*)} \approx 0$$

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)

$$-\sum_{s\in S} p_{sq^*}\log(p_{sq^*}) \geq \log(\ell), \qquad p_{sq^*} = \frac{n_{sq^*}}{\sum_{s'\in S} n_{s'q^*}}$$

- Recursive (c,l)-diversity

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

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

L-Diversity: Summary

- Defends against background knowledge attacks and homogeneity attacks
  - L-Diversity ensures diversity
  - Gives guarantees against "unknown" background knowledge
  - Can model don't care values ("person is healthy")
- Guarantee understood by a lay person
  - "At least L different values"
- Permits efficient sanitization algorithms
  - Bayes-optimal definition is not monotone
  - L-Diversity and (c,k)-recursive L-Diversity are monotone
- Experiments show that little utility is lost compared to k-anonymity

(α,k)-Anonymization

[Wong, Li, Fu, and Wang; KDD 2006]
Defends against homogeneity attacks

- Dataset is α-deassociative for a value s: Relative frequency of s within its group is <= α.
- (α,k)-anonymity: Dataset is k-anonymous and α-deassociative for all values in the domain of a sensitive attribute

What About Other Knowledge?

- If Carol and David are both sick and if Carol has the flu, then David also has the flu:
  t_{Carol}[Disease] = Influenza →
  t_{David}[Disease] = Influenza
- Other types of knowledge?
- Language for background knowledge?
- Complexity, guarding against worst-case disclosure?

The Curse of Dimensionality [A05]

[Aggarwal; VLDB 2005]

- Curse of dimensionality
- Formal analysis that shows with increasing dimensionality all information in the data is lost in order to achieve k-anonymity

K-Anonymity and the Curse of Dimensionality (Contd.)

- M(S): Maximum Euclidean distance between any pair of points in S
- M(D): Maximum Euclidean distance between any pair of points in whole database S
- Relative condensation loss L(S) through k-anonymization L(S) = M(S)/M(D)

- Theorem [A05]: For any set S of points to be k-anonymous, the relative condensation loss goes to 1 with increasing dimensionality:

$$\lim_{d \to \infty} E[M(S)/M(D)] = 1$$

Protection Against An Adversary

[Aggarwal, Pei, and Zhang; KDD 2006]

- Problem: Any attribute might be sensitive; need to defend against inference attacks based on rules learned from the data
  - Example:
    [Type = Manager and DEP = Toy] → Salary > 100k;
    Confidence of rule: 100%
    Simple suppression of private values insufficient.
  - Approach: Make strong rules weaker

Open Problems

- Tradeoff of utility versus privacy
  - See Kifer et al, SIGMOD 2006,
    Levefre et al, KDD 2006, Xu et al., KDD 2006
- Re-publication
- Theory of learning from summaries
- Multi-round protocols
- Formalization of classes of background knowledge
- Location privacy

Tutorial Summary

- Untrusted data collector
- Trusted data collector
- Many interesting open problems!

Thanks

Rakesh Agrawal, Chris Clifton, Wenliang Du, Cynthia Dwork, Alexandre Evfimievski, Ashwin Machanavajjhala, Daniel Kifer, Lucja Kot, Kristen Lefevre, David Martin, Kobbi Nissim, Muthuramakrishnan Venkitasubramaniam, Ramakrishnan Srikant, Walker White

For an annotated list of references for all the topics see (soon :-)
http://www.cs.cornell.edu/database/privacy

Questions?

johannes@cs.cornell.edu.

CORNELL

Data Mining