Privacy and Background Knowledge

Johannes Gehrke
Department of Computer Science

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

Driving Factors: A LARGE Hardware Revolution

![Diagram showing the trend of transistors over time, possibly illustrating Moore's Law.](image)
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.

Other Driving Factors

- Gilder’s law (bandwidth doubles every 6 months)

- Metcalf’s law (network usefulness increases squared with the number of users)
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
Pulsar Surveys

- Most demanding of the ALFA surveys
  - ~100 MB/s to disk
  - ~1 PB for entire survey (3-5 yr @ 6-10% duty cycle)
- Requires coarsely parallel processing of raw data in discrete, local data chunks
  - processing time ~ 50-200x data acquisition time on single processor (Intel 2.5 GHz 512k cached with 1GB ram)
  - depends on data set details, algorithms, code
  - Distributed initial processing (Cornell + 5 sites)
- Requires meta-analysis of data products of the initial analysis
  - Database and data mining research problems

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 to the astronomy research community
- Processing
  - Extremely processor intensive
  - Currently just exhaustive search over a large parameter space (periodicity, dispersion, time)
  - Find new pulsars --- and other interesting phenomena

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%
And Even the Popular Press Caught On

Concerns About Privacy

- S. Garfinkel, "Database Nation: The Death of Privacy in 21st Century", O’ Reilly, Jan 2000

The Setup

![Database Diagram]
Model I: Untrusted Data Collector

Find aggregate properties of \{r_1, r_2, \ldots, r_N\}

Customer 1 \(r_1\)
Customer 2 \(r_2\)
Customer 3 \(r_3\)
Customer N \(r_N\)

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

Publish properties of \{r_1, r_2, \ldots, r_N\}

Customer 1 \(r_1\)
Customer 2 \(r_2\)
Customer 3 \(r_3\)
Customer N \(r_N\)
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 privacy?

Types of Disclosure

- Tolerated Disclosure
  - Statistically private: too fuzzy or unlikely
  - Computationally private: hard to use

- Cryptographic protocols
Types of Disclosure

Knowledge as distribution: This talk!

- Tolerated Disclosure
  - Statistically private (too fuzzy or unlikely)
  - Computationally private (hard to use)

Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions

Privacy Preserving Associations

Find associations among items in \( \{t_1, t_2, \ldots, t_N\} \)

An association is an itemset with frequency \( \geq s_{\text{min}} \)
Problem Introduction

Abstract:
- A set of items \( \{1,2,\ldots,k\} \)
- A database of transactions \( D = \{t_1, t_2, \ldots, t_n\} \), a subset of \( \{1,2,\ldots,k\}\)

**GOAL:**
Find all itemsets that appear in at least \( s_{min} \) transactions

(“appear in” == “are subsets of”)

For an itemset \( I \), the number of transactions it appears in is called the support of \( I \).

\( s_{min} \) is called the minimum support.

Concrete:
- \( I = \{\text{milk, bread, cheese, ...}\} \)
- \( D = \{\{\text{milk,bread,cheese}\}, \{\text{bread,cheese,juice}\}, \ldots\} \)

**GOAL:**
Find all itemsets that appear in at least 1000 transactions

Transaction (milk,bread,cheese) supports itemset (milk,bread)

Example

**Example:**
- \( I = \{1,2,3,4\} \)
- \( D = \{(1,2,3,4), (1,2,3,4,1,2), (1,2,3,4,1,2), (1,2,3,4,1,2)\} \)

**Questions:**
- What is the support of \( \{1\} \)? \( \{1,2\} \)?
- Given a minimum support of 5, what is the output? 4? 3?

**Observations:**
- \( \text{Support}(\{1,2\}) \leq \text{Support}(\{1\}) \)
- \( \text{Support}(\{1,2\}) \leq \text{Support}(\{2\}) \)

The Itemset Lattice

![Itemset Lattice Diagram]

- (1)
- (2)
- (3)
- (4)
- (1,2)
- (1,3)
- (1,4)
- (2,3)
- (2,4)
- (3,4)
- (1,2,3)
- (1,2,4)
- (1,3,4)
- (2,3,4)
- (1,2,3,4)
Frequent Itemsets

Breath First Search: 1-Itemsets

The Apriori Principle:
\[ \text{infrequent} \Rightarrow (I \cup \{x\}) \text{ infrequent} \]
**Breath First Search: 1-Itemsets**

<table>
<thead>
<tr>
<th>Infrequent</th>
<th>Frequent</th>
<th>Currently examined</th>
<th>Don't know</th>
</tr>
</thead>
<tbody>
<tr>
<td>{}</td>
<td>[1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2]</td>
<td>[3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>[1,2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,4]</td>
<td>[2,3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3,4]</td>
<td>[1,2,4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,3,4]</td>
<td>[2,3,4]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Apriori Principle:

\[ I \text{ infrequent} \Rightarrow (I \cup \{x\}) \text{ infrequent} \]

**Breath First Search: 2-Itemsets**

<table>
<thead>
<tr>
<th>Infrequent</th>
<th>Frequent</th>
<th>Currently examined</th>
<th>Don't know</th>
</tr>
</thead>
<tbody>
<tr>
<td>{}</td>
<td>[1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2]</td>
<td>[3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>[1,2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,4]</td>
<td>[2,3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3,4]</td>
<td>[1,2,4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,3,4]</td>
<td>[2,3,4]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Breath First Search: 3-Itemsets**

<table>
<thead>
<tr>
<th>Infrequent</th>
<th>Frequent</th>
<th>Currently examined</th>
<th>Don't know</th>
</tr>
</thead>
<tbody>
<tr>
<td>{}</td>
<td>[1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[2]</td>
<td>[3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[4]</td>
<td>[1,2]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,4]</td>
<td>[2,3]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[3,4]</td>
<td>[1,2,4]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>[1,3,4]</td>
<td>[2,3,4]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Infrequent
Frequent
Currently examined
Don't know
Breath First Search: 3-Itemsets

- Infrequent
- Frequent
- Currently examined
- Don't know

Breadth First Search: Remarks

- We prune infrequent itemsets and avoid to count them
- To find an itemset with $k$ items, we first need to count all $2^k$ subsets

Depth First Search (1): Start

- Infrequent
- Frequent
- Currently examined
- Don't know
Graphical representation of Depth First Search (DFS) proceeds through nodes, exploring each branch before backtracking. The root node is empty, and each child node represents the addition of an item to the current set. Branches are labeled 'Infrequent,' 'Frequent,' 'Currently examined,' or 'Don’t know.'

**Depth First Search: Remarks**

- We prune frequent itemsets and avoid counting them.
- To find an itemset with k items, we need to count k prefixes.
Associations Recap

- A transaction \( t \) is a set of items
- All transactions form a set \( T \) of transactions
- Any itemset \( A \) has support \( s \) in \( T \) if
  \[
  s = \text{supp}(A) = \frac{|\{t \in T | A \subseteq t\}|}{|T|}
  \]
- Itemset \( A \) is frequent if \( s \geq s_{\min} \)
- If \( A \subseteq B \), then \( \text{supp}(A) \geq \text{supp}(B) \).
- Association rule: \( A \Rightarrow B \) holds when the union \( A \cup B \)
  is frequent and: \( \text{supp}(A \cup B) \geq \text{supp}(A) \cdot \text{conf}_{\min} \)

Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions

Our Model

- **Alice**: J.S. Bach, painting, nasa.gov, ...
- **Bob**: B. Spears, baseball, cnx.com, ...
- **Chris**: B. Marley, camping, linux.org, ...
- Recommendation Service
Our Model (Contd.)

Alice
- J.S. Bach, painting, nasa.gov, ...
- J.S. Bach, painting, nasa.gov, ...

Bob
- B. Spears, baseball, cnn.com, ...
- B. Marley, camping, linux.org, ...

Chris
- B. Marley, camping, linux.org, ...

Recommendation Service

Find associations among items in \( \{t_1, t_2, \ldots, t_n\} \)

An association is an itemset with frequency \( \geq s_{\text{min}} \)

Privacy Preserving Associations
Minimal Information Sharing

- Ideally, we want an algorithm that discloses only the association rules.
- However, in practice, we need some extra disclosure.
Our Model: Another View

\[ y = R(x) \]

Randomized data
Described by a random variable \( Y = R(X) \)

Original (private) data
Assumptions:
- Described by a random variable \( X \)
- Each client is independent.

The Problem

- How to randomize transactions so that
  - we can find frequent itemsets
  - while preserving privacy at transaction level?

The Randomized Response Model

[Stanley Warner, JASA 1965]

- Respondents are given:
  1. A source of randomness for YES and NO answers (a biased coin)

- The procedure:
  - Respondent flips the coin
  - Answers YES iff coin gives correct answer, answers NO otherwise
Another View: Two Questions

- Respondents are given:
  1. The coin
  2. Two logically opposite statements:
     - \( S: \) I am teaching database systems with R+G DBMS.
     - \( S^\text{bar}: \) I am not teaching database systems with R+G DBMS.

- The procedure:
  - Respondent flips the coin
  - Answers either statement \( S_1 \) or \( S_2 \).

Analysis

- \( n \) = 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’.

- \( P(Y_i=1) = np + (1-n)(1-p) = p\text{YES} \)
- \( P(Y_i=0) = (1-n)p + n(1-p) = p\text{NO} \)

Analysis (Contd.)

- Assume a sample with \( n \) records, \( n_1 \) say YES, \( n-n_1 \) say NO
- Likelihood of this sample:
  - \( L = p_{\text{YES}}^{n_1} p_{\text{NO}}^{(n-n_1)} \)
  - This gives a maximum likelihood estimate of
    \( \pi = (p-1)/(2p-1) + n_1/n(2p-1) \)
- Easy to show:
  - \( E(n^{\text{est}}) = n \)
  - \( \text{Var}(n^{\text{est}}) = n(1-n)/n + [1/[16(p-0.5)^2]-0.25]/n \)

Source of Variance  Sampling  Coin Flips
Interval Privacy

- Agrawal & Srikant, SIGMOD 2000

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

Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions
Background Knowledge

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:

<table>
<thead>
<tr>
<th>1% have ({a, b, c})</th>
<th>5% have ({a, b}, {a, c}, ) or ({b, c}) only</th>
<th>94% have one or zero items of ({a, b, c})</th>
</tr>
</thead>
</table>

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

\[ \cdot 0.2^3 \cdot 0.2^3 \cdot 0.8/10,000 \leq (0.2 \cdot (9 \cdot 0.8/10,000))^2 \]

<table>
<thead>
<tr>
<th>After randomization: How many have {a, b, c}?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.008% 800 ts. 0.000128% 13 trans. less than 0.000002% 2 transactions</td>
</tr>
</tbody>
</table>

Example: \{a, b\}

10 M transactions of size 10 with 10 K items:

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

\[ \cdot 0.2^3 \cdot 0.2^3 \cdot 0.8/10,000 \leq (0.2 \cdot (9 \cdot 0.8/10,000))^2 \]

<table>
<thead>
<tr>
<th>After randomization: How many have {a, b}?</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.008% 800 ts. 0.000128% 13 trans. less than 0.000002% 2 transactions</td>
</tr>
</tbody>
</table>

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 what we call a privacy breach.

- The example randomization preserves privacy “on average,” but not “in the worst case.”
Privacy Breaches

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

- Note: Interval privacy has the same problem
  - Assume interval privacy [-30,30]
  - Assume you see an age value $y = 130$

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 $\beta$ (e.g. 50%) if:
  \[
  \Pr[z \in t | A \subseteq t'] \geq \beta
  \]

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

Privacy Breaches: Goals

- We want a bound for all privacy breaches
  - not only for: item $e \in t$ versus itemset $\subseteq t'$

- No knowledge of data distribution is required in advance
  - We should not need to know $\Pr[item \in t]$

- Applicable to numerical data as well

- Easy to work with, even for complex randomizations
\( \alpha \)-to-\( \beta \) 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 \( \beta = 50\% \)

\( \alpha \)-to-\( \beta \) Privacy Breach

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

Client

\( X = x \)

SERVER

\[ \text{Prob} \left[ P(X) \right] \leq \alpha \]

\( \alpha \) to \( \beta \) Privacy Breach

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

Client

\( X = x \)

SERVER

\[ \text{Prob} \left[ P(X) \leq \alpha \right] \]

\[ \text{Prob} \left[ P(X) \mid Y = y \right] \geq \beta \]
\( \alpha \)-to-\( \beta \) 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 \( \alpha \)-to-\( \beta \) privacy breach w.r.t. property \( P(x) \).

\( \alpha \)-to-\( \beta \) Privacy Breach

Checking for \( \alpha \)-to-\( \beta \) 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

\[
\begin{array}{c|c}
1 & 1 \\
2 & 2 \\
3 & 3 \\
4 & 4 \\
5 & 5 \\
6 & 6 \\
7 & 7 \\
8 & 8 \\
9 & 9 \\
10 & 10 \\
\end{array}
\]

\[ R(x) = y \]
Amplification Condition

\[ p [x \rightarrow y \mid x \rightarrow y \text{ are transition probabilities} ] \]

Amplification Condition

\[ p [x \rightarrow y \mid x \rightarrow y \text{ are transition probabilities} ] \]

Amplification Condition

\[ p [x \rightarrow y \mid x \rightarrow y \text{ are transition probabilities} ] \]
Amplification Condition

Definition:
- Randomization operator \( R \) is called "at most \( \gamma \)-amplifying" if:
  \[
  \max_{x \in A} \max_{y \in B} \frac{p[x \rightarrow y]}{p[x_0 \rightarrow y]} \leq \gamma
  \]
  
  - Transition probabilities \( p[x \rightarrow y] = \text{Prob}[R(x) = y] \) depend only on the operator \( R \) and not on data.
  - We assume that all \( y \) have a nonzero probability.
  - The bigger \( \gamma \) is, the more may be revealed about \( x \).

The Bound on \( \alpha \)-to-\( \beta \) Breaches

Theorem:
- If randomization operator \( R \) is at most \( \gamma \)-amplifying, and if:
  \[
  \gamma < \frac{\beta}{\alpha} \cdot \frac{1 - \alpha}{1 - \beta}
  \]
  
  Then, revealing \( R(X) \) to the server will never cause an \( \alpha \)-to-\( \beta \) privacy breach.

Examples:
- 5%-to-50% privacy breaches do not occur for \( \gamma < 19 \):
  \[
  \frac{0.5}{0.05} = \frac{1 - 0.05}{1 - 0.5} = 19
  \]
- 1%-to-98% privacy breaches do not occur for \( \gamma < 4851 \):
  \[
  \frac{0.98}{0.01} = \frac{1 - 0.01}{1 - 0.98} = 4851
  \]
- 50%-to-100% privacy breaches do not occur for any finite \( \gamma \).
Amplification: Summary

- An \( \alpha \)-to-\( \beta \) privacy breach w.r.t. property \( P(x) \) occurs when
  - \( \text{Prob} \{ P \text{ is true} \} \leq \alpha \)
  - \( \text{Prob} \{ P \text{ is true} \mid Y = y \} \geq \beta \).

- Amplification methodology limits privacy breaches by just looking at transitional probabilities of randomization.
  - Does not use data distribution:
    \[
    \max_{x,y} \max_{v,w} \frac{P[x \to y]}{P[x \to y]} \leq \gamma
    \]

Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions

Definition of select-a-size

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

\[
\begin{align*}
  t &= \{a, b, c, d, e, f, u, v, w\} \\
  t' &= \text{(result after randomization)}
\end{align*}
\]
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 $\{p[j]\}_{0 \leq j \leq m}$;
  - Include exactly $j$ items of $t$ into $t'$;
  - Each other item (not from $t$) goes into $t'$ with probability $\rho$.

The choice of $\{p[j]\}_{0 \leq j \leq m}$ and $\rho$ is based on the desired privacy level.
Let itemset $A$ have four items ($k = 4$).

1. Partial supports
2. Randomization

Transactions that do not contain $A$
Support Recovery

Let itemset $\mathcal{A}$ have four items.

Transition matrix

$$E\tilde{\mathbf{s}} = P\tilde{\mathbf{s}}$$

\[ \begin{array}{c|ccc} & \text{1 item of } \mathcal{A} & \text{2 items of } \mathcal{A} & \text{3 items of } \mathcal{A} & \text{all items of } \mathcal{A} \\ \hline \text{0%} & \text{20%} & \text{40%} & \text{0%} & \text{20%} \end{array} \]

Support Recovery

Randomization
The Unbiased Estimators

- Given randomized partial supports, we can estimate original partial supports:
  \[ \hat{x}_\text{est} = Q \cdot \hat{x} \], where \( Q = P^{-1} \)
- Covariance matrix for this estimator:
  \[ \text{Cov} \hat{x}_\text{est} = \frac{1}{P-1} \sum_{i=1}^{P-1} \hat{x}_i \cdot Q D[i] Q^T, \]
  where \( D[i] = P_{i,:} - \delta_{i,:} - P_{i,:} \cdot P_{i,:} \)
- To estimate it, substitute \( \hat{x}_i \) with \( (\hat{x}\text{est})_i \).
- 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 \( \geq x_{\text{min}} \)
3. New candidate sets = all \((k + 1)\)-itemsets s.t. all their \(k\)-subsets are candidate sets with support \( \geq x_{\text{min}} \)
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 (\(\sigma^2\)) for all candidate sets
2. Output the candidate sets with support \( \geq x_{\text{min}} \)
3. New candidate sets = all \((k + 1)\)-itemsets s.t. all their \(k\)-subsets are candidate sets with support \( \geq x_{\text{min}} - \sigma \)
4. Let \( k = k + 1 \)
Stop when there are no more candidate sets, or the estimator’s precision becomes unsatisfactory.
Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions

Select-A-Size Revisited

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

The choice of \( \{p[j]\}_{0 \leq j \leq m} \) and \( \rho \) is based on the desired privacy level.

\[
x = a, b, c, d, e, f, u, v, w
\]
\[
y = \begin{align*}
  & b, e, u, w \quad \text{\textit{j items}} \\
  & \quad \text{\textit{items inserted with prob. } \rho}
\end{align*}
\]

Often \( \rho \approx 0.5 \) and \( n \approx 10 \ldots 100 \) K items, making \( y \) HUGE!
Select-A-Size Revisited

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

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

Let $m = 10$, $n = 100,000$, mining itemsets of size $\leq 5$.
For $\rho = 0.5$ : 100,000 bits per transaction.

The Idea of Compression

- The idea: Let the items in $y$ be computed by a pseudorandom number generator.

The Idea of Compression

- The idea: Let the items in $y$ be computed by a pseudorandom number generator.
Pseudorandom Generator

- Definition: Item number $i$

Seed $\xi \rightarrow G \rightarrow \text{bit } \in \{0, 1\}$

- If seed $\xi$ is uniformly random, $\text{Prob}[G(\xi, i) = 1] = \rho$

- For any $q$ integers $1 \leq i_1 < i_2 < \ldots < i_q \leq n$:
  - Uniformly random seed $\xi \rightarrow G$ Statistically independent
Itemset Compression

Given transaction $x$ of size $m$, construct $y = R(x)$:
- Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $\{p[j]\}_{0..m}$.
- Choose exactly $j$ items of $x$ (to include into $y$).

\[ x = \{a, b, c, d, e, f, u, v, w\} \]
\[ y = \{b, c, u, w\} \]
Itemset Compression

- Given transaction $x$ of size $m$, construct $y = R(x)$:
  - Choose a number $j \in \{0, 1, \ldots, m\}$ with distribution $(p\j)$:
  - Choose exactly $j$ items of $x$ (to include into $y$):
  - Choose a seed $\xi$ uniformly at random conditioned by:
    - For all items in $x$, $G(\xi, \text{item}) = 1$ if the item is chosen above.

$x = \{a, b, c, d, e, f, u, v, w\}$
$y = \{b, c, e, u, w\}$

"Transparency" of Compression

New transactions — the same old algorithms:
- We can do support recovery the same way as if there is no compression (for small itemsets);
- We can check amplification condition and select randomization parameters the same way as if there is no compression.

The bits produced by the pseudorandom generator must be $q$-wise independent, where $q = \text{max. transaction size} + \text{max. association size}$

Compression in Practice

- Suppose we use Bose-Chaudhuri-Hocquenghem (BCH) error-correcting codes for pseudorandom generators.
- Let $m = 10$, $n = 100\,000$, mining itemsets of size $\leq 5$.
- For $\rho = 0.5$:
  - "Ordinary" way: 100,000 bits per transaction;
  - "Compressed" way: 136 bits per transaction.
- For $\rho = 1/16$:
  - "Ordinary" way: 100,000 \cdot $H(1/16) = 33\,729$ bits per transaction;
  - "Compressed" way: 570 bits per transaction.
Talk Outline

- Introduction
- Privacy-preserving data mining
  - Association rules
  - Problem definition
  - Privacy breaches
  - Select-A-Size randomization
  - Itemset compression
  - Experimental results
- Privacy-preserving data publishing
- Conclusions

Lowest Discoverable Support

- LDS is s.t., when predicted, it is $\pm \sigma$ away from zero.
- Roughly, LDS is proportional to $\frac{1}{\sqrt{|t|}}$

\[ |t| = 5, \rho = 50\% \]

LDS vs. number of transactions

\begin{align*}
\text{Number of transactions, millions} & \quad 0 & \quad 0.2 & \quad 0.4 & \quad 0.6 & \quad 0.8 & \quad 1 & \quad 1.2 \\
\text{LDS, \%} & \quad |1-itemsets| & \quad |2-itemsets| & \quad |3-itemsets| \\
\end{align*}

LDS vs. Breach Level

- Reminder: breach level is the limit on $\Pr [z \in t | A \subseteq t']$
LDS vs. Transaction Size

\[ \rho = 50\%, |T| = 5 \text{ M} \]

- Very long transactions cannot be used for prediction

Real datasets: Soccer, Mailorder
- Soccer is the clickstream log of WorldCup’98 web site, split into sessions of HTML requests.
  - 11 K items (HTMLs), 6.5 M transactions
  - Available at [http://www.acm.org/sigcomm/ITA/](http://www.acm.org/sigcomm/ITA/)
- Mailorder is a purchase dataset from an on-line store
  - Products are replaced with their categories
  - 96 items (categories), 2.9 M transactions

A small fraction of transactions are discarded as too long.
- longer than 10 (for soccer) or 7 (for mailorder)

Modified Apriori on Real Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Itemset Size</th>
<th>True 1-itemsets</th>
<th>True 2-itemsets</th>
<th>False Drops</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer:</td>
<td>1</td>
<td>266</td>
<td>254</td>
<td>12</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>217</td>
<td>195</td>
<td>22</td>
<td>-45</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>48</td>
<td>43</td>
<td>5</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Itemset Size</th>
<th>True 1-itemsets</th>
<th>True 2-itemsets</th>
<th>False Drops</th>
<th>False Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mailorder:</td>
<td>1</td>
<td>65</td>
<td>65</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>228</td>
<td>212</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>22</td>
<td>18</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Breach level = 50%. Inserted 20-50% items to each transaction.
False Drops       False Positives

<table>
<thead>
<tr>
<th>Size</th>
<th>≤ 0.1</th>
<th>0.1-0.15</th>
<th>0.15-0.2</th>
<th>≥0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soccer</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>≤ 0.1</th>
<th>0.1-0.15</th>
<th>0.15-0.2</th>
<th>≥0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True supp%, when pred. supp ≥ 0.2%</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>10</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>≤ 0.1</th>
<th>0.1-0.15</th>
<th>0.15-0.2</th>
<th>≥0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mailorder</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>≤ 0.1</th>
<th>0.1-0.15</th>
<th>0.15-0.2</th>
<th>≥0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>True supp%, when pred. supp ≥ 0.2%</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>18</td>
</tr>
</tbody>
</table>

Talk Outline

- Introduction
- Privacy-preserving data mining
- Privacy-preserving data publishing
  - K-Anonymity
  - Attacks
  - L-Diversity
- Conclusions

Trusted Data Collector

Publish properties of \{r_1, r_2, \ldots, r_N\}
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?

Sample Microdata

<table>
<thead>
<tr>
<th>SSN</th>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>630-35-1230</td>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>053-30-1230</td>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>070-30-2432</td>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Viral</td>
</tr>
<tr>
<td>238-30-0890</td>
<td>14853</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>295-05-1275</td>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>574-92-0242</td>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>386-92-1378</td>
<td>14850</td>
<td>50</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>005-94-3444</td>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>248-22-2456</td>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>221-22-9773</td>
<td>12068</td>
<td>32</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>615-84-1924</td>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Removing SSN …

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>Japanese</td>
<td>Viral</td>
</tr>
<tr>
<td>14853</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14850</td>
<td>59</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Medical Records of a hospital near Ithaca serving patients from:
- Freeville (13068)
- Dryden (13053)
- Ithaca (14850, 14853)
Linkage Attacks

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14R50</td>
<td>39</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14R50</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14R50</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14R50</td>
<td>59</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

Quasi-Identifiers and Sensitive Attributes

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14R50</td>
<td>39</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14R50</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14R50</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14R50</td>
<td>59</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
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 \( \geq k \).
- 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
## Example Microdata

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>28</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>21</td>
<td>Japanese</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>28</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14853</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14850</td>
<td>50</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13053</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13053</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

## 4-Anonymous Microdata

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>148**</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>148**</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>148**</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>148**</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

## K-Anonymity Algorithms

- **Optimal Full-Domain Algorithms**
  - Binary Search [Sa01] of the lattice finds solution of minimum height
- **Optimal Algorithms:**
  - Bayardo-Agrawal [BA05]
  - Leveque-Devitt-Ramakrishnan [LDR05]
- **Heuristic Algorithms**
  - Greedy Heuristic Search [Sw02-2, FWY05, WYC04]
  - No guarantees about optimality
- **Stochastic Search**
  - Genetic Algorithms [Iy02]
  - Simulated Annealing [Wid02]
  - Long run times to convergence; do not guarantee optimality
- **Approximation Algorithms**
  - Cell-suppression [MW04, AFKH+05]
  - Have not been implemented
Talk Outline

- Introduction
- Privacy-preserving data mining
- Privacy-preserving data publishing
  - K-Anonymity
  - Attacks
  - L-Diversity
- Conclusions

Example Microdata

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13008</td>
<td>29</td>
<td>American</td>
<td>Heart</td>
</tr>
<tr>
<td>13005</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13005</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13005</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13005</td>
<td>23</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14853</td>
<td>50</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>14853</td>
<td>55</td>
<td>Russian</td>
<td>Heart</td>
</tr>
<tr>
<td>14850</td>
<td>47</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>14850</td>
<td>49</td>
<td>American</td>
<td>Viral</td>
</tr>
<tr>
<td>13005</td>
<td>31</td>
<td>American</td>
<td>Cancer</td>
</tr>
<tr>
<td>13005</td>
<td>37</td>
<td>Indian</td>
<td>Cancer</td>
</tr>
<tr>
<td>13008</td>
<td>36</td>
<td>Japanese</td>
<td>Cancer</td>
</tr>
<tr>
<td>13008</td>
<td>32</td>
<td>American</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

4-Anonymous Microdata

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>1485*</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>
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'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.

---

**Homogeneity Attack**

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

**Background Knowledge Attack**

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Occupation</th>
<th>Salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&lt;30</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>130**</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>130**</td>
<td>30-40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

**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:

$P(t[S] = s | t[Q] = q & t^* \in T^*) = \sum_{s'} n_{s'q} \frac{f(s'|q)}{f(s'|q^*q)}$

Privacy Definition (1)

- Positive Disclosure: Posterior Belief $> 1 - \delta$
- Negative Disclosure: Posterior Belief $< \delta$

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
  \[ \text{Posterior belief} \sim \text{prior belief} \]

- Example instantiation: \( \alpha \)-to-\( \beta \) privacy breach definition
  \[
  \begin{align*}
  \text{Prior Belief} < \alpha & \quad \text{and} \quad \text{posterior Belief} > \beta \quad \text{OR} \\
  \text{Prior Belief} > 1 - \alpha & \quad \text{and} \quad \text{posterior Belief} < 1 - \beta
  \end{align*}
  \]

- Automatically eliminates homogeneity attack
  - Homogeneity \( \rightarrow \) 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 =
  \[
  \frac{n_{q,s}}{\sum_{s'} n_{q',s'}} \frac{f(s,q)}{f(s',q')}
  \]

- Homegeneity attack
  \[ \forall s' \neq s, n_{q,s} \gg n_{q',s'} \]
Towards A Practical Definition (2)

- Posterior belief =

\[
\frac{\sum_q n_q f(v, q)}{\sum_{q'} n_{q'} f(v, q')}
\]

- Background knowledge attack

\[\forall s' \neq s, \frac{f(v, q)}{f(v', q')} \approx 0\]

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: \( \text{Entropy}(\text{group}) > \log(1) \)

\[-\sum_s p_s \log(p_s) \geq \log(\text{l}), \quad p_s = \frac{n_s}{\sum n_{s'}}\]

- Recursive (c,l)-diversity

3-Diverse Microdata

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Nationality</th>
<th>Disease</th>
</tr>
</thead>
<tbody>
<tr>
<td>13053</td>
<td>&lt;=40</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>13067</td>
<td>&lt;=40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>13067</td>
<td>&gt;40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>13068</td>
<td>&lt;=40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>&gt;40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>13068</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>13068</td>
<td>&gt;40</td>
<td>*</td>
<td>Viral</td>
</tr>
<tr>
<td>13068</td>
<td>&lt;=40</td>
<td>*</td>
<td>Cancer</td>
</tr>
<tr>
<td>13067</td>
<td>&lt;=40</td>
<td>*</td>
<td>Heart</td>
</tr>
<tr>
<td>13067</td>
<td>&lt;=40</td>
<td>*</td>
<td>Cancer</td>
</tr>
</tbody>
</table>

- 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.

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

Talk Outline

- Introduction
- Privacy-preserving data mining
- Privacy-preserving data publishing
- Conclusions
What I Talked About

- Privacy-preserving association rule mining
  - $\alpha$-to-$\beta$ privacy breaches
  - Amplification condition
  - Select-a-size randomization, itemset compression
- Privacy-preserving data publishing
  - Attacks due to background knowledge
  - L-diversity

What I Talked About (Only Useful Stuff)

The primary purpose of the DATA statement is to give names to constants; instead of referring to pi as 3.141592653589793 at every appearance, the variable PI can be given that value with a DATA statement and used instead of the longer form of the constant. This also simplifies modifying the program, should the value of pi change.

-- FORTRAN manual for Xerox Computers

Open Problems

- We only scratched the surface
- Selected future topics:
  - Tradeoff of utility versus privacy
  - Re-publication
  - Theory of learning from summaries
  - Multi-round protocols
  - Combination of randomization with other techniques (secure-multiparty computation, sketching, etc.)
  - Formalization of classes of background knowledge
Modeling Belief

- We model background knowledge (i.e. prior belief) as a background distribution plus simple existential statements about tuples (e.g. Person X is in the original table).

Knowing a Tuple-Level Pattern

- Person X is a 35 yr old Male living in 14850.
- Males do not usually have miscarriages, or ovarian or breast cancer.

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Sex</th>
<th>Malady</th>
</tr>
</thead>
<tbody>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Miscarriage</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Ovarian cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Miscarriage</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Malaria</td>
</tr>
</tbody>
</table>

Knowing a Table-Level Pattern

- Person X is a 38 yr old Female living in 14850.
- Person Y is a 42 yr old Male living in 14850.
- X and Y are married.
- Viruses often attack spouses together.

<table>
<thead>
<tr>
<th>Zip</th>
<th>Age</th>
<th>Sex</th>
<th>Malady</th>
</tr>
</thead>
<tbody>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Miscarriage</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Ovarian cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Breast cancer</td>
</tr>
<tr>
<td>1485*</td>
<td>3*</td>
<td>*</td>
<td>Miscarriage</td>
</tr>
<tr>
<td>4*</td>
<td>*</td>
<td>*</td>
<td>Malaria</td>
</tr>
<tr>
<td>14850</td>
<td>4*</td>
<td>*</td>
<td>Malaria</td>
</tr>
<tr>
<td>14850</td>
<td>4*</td>
<td>*</td>
<td>Malaria</td>
</tr>
<tr>
<td>14850</td>
<td>4*</td>
<td>*</td>
<td>Malaria</td>
</tr>
<tr>
<td>14850</td>
<td>4*</td>
<td>*</td>
<td>Influenza</td>
</tr>
<tr>
<td>14850</td>
<td>4*</td>
<td>*</td>
<td>Malaria</td>
</tr>
</tbody>
</table>
Issue

- Representing prior knowledge by a distribution on tuples plus simple tuple-level existential statements is not sufficient → Distribution over tables.
- Distribution over distributions? Hidden variables? Encompassing framework?

Thanks

Current Students:
- Ashwin Machanavajjhala
- Daniel Kifer (graduating summer 2006)
- David Martin
- Muthuramakrishnan Venkitasubramaniam

Former students:
- Alexandre Evfimievski (now IBM Almaden)

Collaborators:
- Rakesh Agrawal (IBM Almaden)
- Ramakrishnan Srikant (IBM Almaden)

But Of Course We Have More Confidence Than Scott Adams …
Questions?

http://www.cs.cornell.edu/johannes
johannes@cs.cornell.edu

Generalization

- Originally defined by Samarati and Sweeney [Sa01, Sw02-1, Sw02-2]
- Each attribute has a domain of values
- Many-to-one (user-defined) generalization functions map the domains of each quasi-identifier attribute to successively more general domains