PURDUE UNIVERSITY


# Privacy-Preserving Data Mining

*How do we mine data when we can't even look at it?*

Chris Clifton
clifton@cs.purdue.edu
www.cs.purdue.edu/people/clifton

*Thanks to Mohamed Elfeky, Eirik Herskedal, Murat Kantarcioglu, Ramakrishnan Srikant, and Jaideep Vaidya for assistance in slide preparation*

CERIAS Center for Education and Research in Information Assurance and Security

Indiana Center for Database Systems

## Privacy and Security Constraints

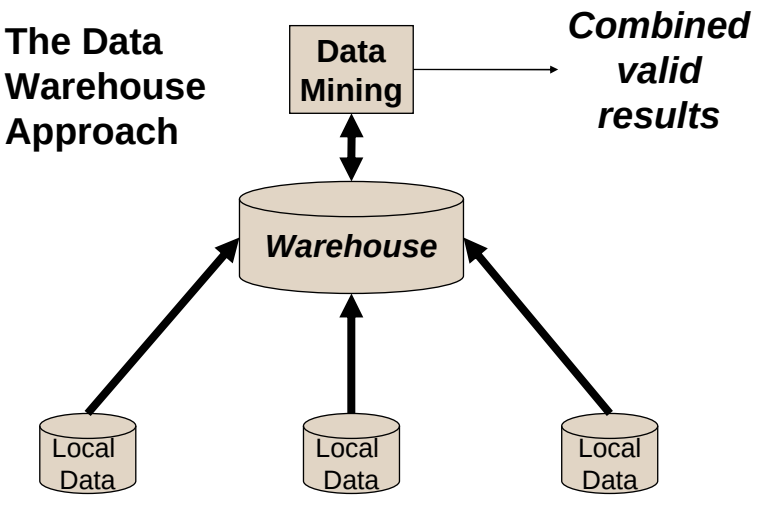
- Individual Privacy
  - Nobody should know more about any entity after the data mining than they did before
  - Approaches: Data Obfuscation, Value swapping
- Organization Privacy
  - Protect knowledge about a collection of entities
    - Individual entity values may be known to all parties
    - Which entities are at which site may be secret

# Privacy constraints don't prevent data mining

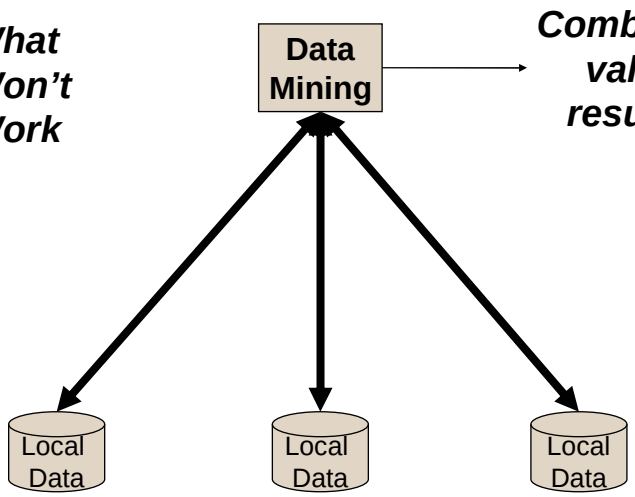
- Goal of data mining is summary results
  - Association rules
  - Classifiers
  - Clusters
- The results alone need not violate privacy
  - Contain no individually identifiable values
  - Reflect overall results, not individual organizations

*The problem is computing the results without access to the data!*

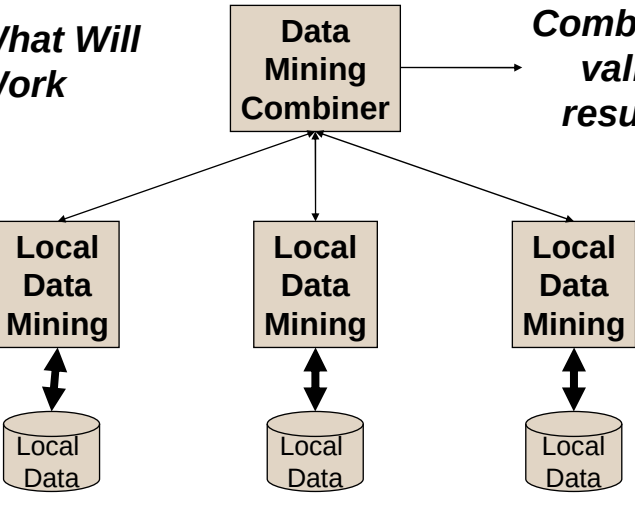
# Distributed Data Mining: The "Standard" Method



# Private Distributed Mining: What is it?



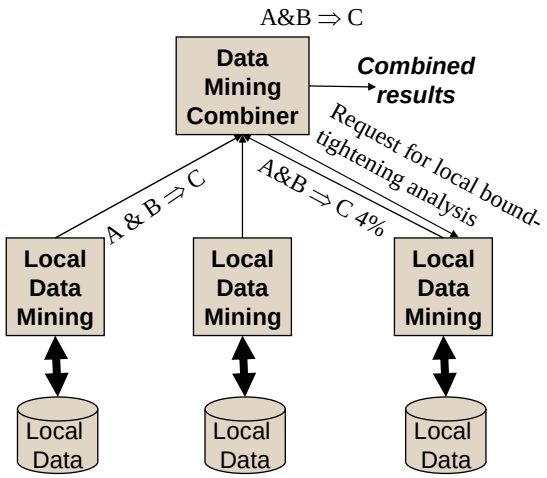

# Private Distributed Mining: What is it?



# Example: *Association Rules*

- Assume data is horizontally partitioned
  - Each site has complete information on a set of entities
  - Same attributes at each site
- If goal is to avoid disclosing entities, problem is easy
- Basic idea: Two-Phase Algorithm
  - First phase: Compute candidate rules
    - Frequent globally ⇒ frequent at some site
  - Second phase: Compute frequency of candidates

# Association Rules in Horizontally Partitioned Data



# Privacy-Preserving Data Mining: Who?

- Government / public agencies. Example:
  - The Centers for Disease Control want to identify disease outbreaks
  - Insurance companies have data on disease incidents, seriousness, patient background, etc.
  - But can/should they release this information?
- Industry Collaborations / Trade Groups. Example:
  - An industry trade group may want to identify best practices to help members
  - But some practices are trade secrets
  - How do we provide "commodity" results to all (Manufacturing using chemical supplies from supplier X have high failure rates), while still preserving secrets (manufacturing process Y gives low failure rates)?

# Privacy-Preserving Data Mining: Who?

- Multinational Corporations
  - A company would like to mine its data for globally valid results
  - But national laws may prevent transborder data sharing
- Public use of private data
  - Data mining enables research studies of large populations
  - But these populations are reluctant to release personal information

# Outline

- Privacy and Security Constraints
  - Types: Individual, collection, result limitation
  - Sources: Regulatory, Contractual, Secrecy
- Classes of solutions
  - Data obfuscation
  - Summarization
  - Data separation
- When do we address these issues?

*Break*

# Outline (after the break): Technical Solutions

- Data Obfuscation based techniques
  - Reconstructing distributions for developing classifiers
  - Association rules from modified data
- Data Separation based techniques
  - Overview of Secure Multiparty Computation
  - Secure decision tree construction
  - Secure association rules
  - Secure clustering
- What if the secrets are in the results?

## Individual Privacy: Protect the "record"

- Individual item in database must not be disclosed
- Not necessarily a person
  - Information about a corporation
  - Transaction record
- Disclosure of parts of record may be allowed
  - Individually identifiable information

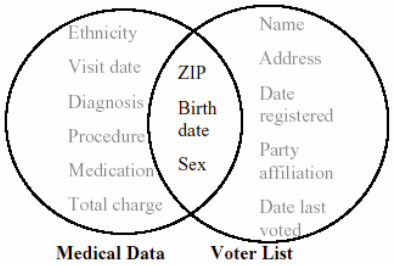
## Individually Identifiable Information

- Data that can't be traced to an individual not viewed as private
  - Remove "identifiers"
- But can we ensure it can't be traced?
  - Candidate Key in non-identifier information
  - Unique values for some individuals

*Data Mining enables such tracing!*

# Re-identifying "anonymous" data (Sweeney '01)

- 37 US states mandate collection of information
- She purchased the voter registration list for Cambridge Massachusetts
  - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



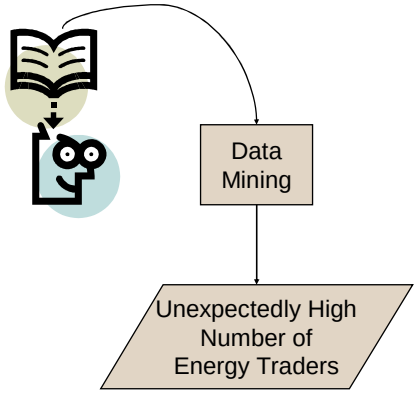

- Solution: *k*-anonymity
  - Any combination of values appears at least *k* times
- Developed systems that guarantee k-anonymity
  - Minimize distortion of results

# Collection Privacy

- Disclosure of individual data may be okay
  - Telephone book
  - De-identified records
- Releasing the whole collection may cause problems
  - Trade secrets – corporate plans
  - Rules that reveal knowledge about the holder of data

# Collection Privacy Example: Corporate Phone Book

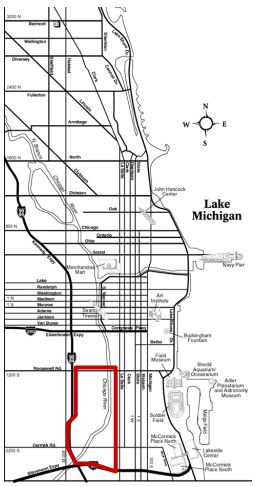
- Telephone Directory discloses how to contact an individual
  - *Intended use*
- Data Mining can find more
  - Relative sizes of departments
  - *Use to predict corporate plans?*
- Possible Solution: Obfuscation
  - Fake entries in phone book
  - *Doesn't prevent intended use*
- Key: Define Intended Use
  - *Not always easy!*




# Restrictions on Results

- Use of Call Records for Fraud Detection vs. Marketing
  - FCC § 222(c)(1) restricted use of individually identifiable information
    *Until overturned by US Appeals Court*
  - 222(d)(2) allows use for fraud detection
- Mortgage Redlining
  - Racial discrimination in home loans prohibited in US
  - Banks drew lines around high risk neighborhoods!!!
  - These were often minority neighborhoods
  - Result: Discrimination (redlining outlawed)

*What about data mining that "singles out" minorities?*



## Sources of Constraints

- Regulatory requirements
- Contractual constraints
  - Posted privacy policy
  - Corporate agreements
- Secrecy concerns
  - Secrets whose release could jeopardize plans
  - Public Relations – “bad press”

## Regulatory Constraints: Privacy Rules

- Primarily national laws
  - European Union
  - US HIPAA rules (www.hipaadvisory.com)
  - Many others: (www.privacyexchange.org)
- Often control transborder use of data
- Focus on intent
  - Limited guidance on implementation

# European Union Data Protection Directives

- Directive 94/46/EC
  - Passed European Parliament 24 October 1995
  - Goal is to ensure free flow of information
    - *Must preserve privacy needs of member states*
  - Effective October 1998
- Effect
  - Provides guidelines for member state legislation
    - Not directly enforceable
  - Forbids sharing data with states that don't protect privacy
    - Non-member state must provide adequate protection,
    - Sharing must be for "allowed use", or
    - Contracts ensure adequate protection
  - US "Safe Harbor" rules provide means of sharing (July 2000)
    - Adequate protection
    - But voluntary compliance
- Enforcement is happening
  - Microsoft under investigation for Passport (May 2002)
  - Already fined by Spanish Authorities (2001)

# EU 95/46/EC: Meeting the Rules

- Personal data is any information that can be traced directly *or indirectly* to a specific person
- Use allowed if:
  - Unambiguous consent given
  - Required to perform contract with subject
  - Legally required
  - Necessary to protect vital interests of subject
  - In the public interest, or
  - Necessary for legitimate interests of processor and doesn't violate privacy
- Some uses specifically proscribed
  - Can't reveal racial/ethnic origin, political/religious beliefs, trade union membership, health/sex life
- Must make data available to subject
  - Allowed to object to such use
  - Must give advance notice / right to refuse direct marketing use
- Limits use for automated decisions
  - Onus on processor to show use is legitimate

europa.eu.int/comm/internal_market/en/dataprot/law

## US Healthcare Information Portability and Accountability Act (HIPAA)

- Governs use of patient information
  - Goal is to protect the patient
  - Basic idea: Disclosure okay if anonymity preserved
- Regulations focus on outcome
  - A covered entity may not use or disclose protected health information, except as permitted or required…
    - To individual
    - For treatment (generally requires consent)
    - To public health / legal authorities
  - Use permitted where "there is no reasonable basis to believe that the information can be used to identify an individual"
- Safe Harbor Rules
  - Data presumed not identifiable if 19 identifiers removed (§ 164.514(b)(2)), e.g.:
    - Name, location smaller than 3 digit postal code, dates finer than year, identifying numbers
  - Shown not to be sufficient (Sweeney)
  - Also not necessary

*Moral: Get Involved in the Regulatory Process!*

## Regulatory Constraints: Use of Results

- Patchwork of Regulations
  - US Telecom (Fraud, not marketing)
    - Federal Communications Commission rules
    - Rooted in antitrust law
  - US Mortgage "redlining"
    - Financial regulations
    - Comes from civil rights legislation
- Evaluate on a per-project basis
  - Domain experts should know the rules
  - You'll need the domain experts anyway – ask the right questions

# Contractual Limitations

- Web site privacy policies
  - "Contract" between browser and web site
  - Groups support voluntary enforcement
    - TrustE – requires that web site DISCLOSE policy on collection and use of personal information
    - BBBonline
      - posting of an online privacy notice meeting rigorous privacy principles
      - completion of a comprehensive privacy assessment
      - monitoring and review by a trusted organization, and
      - participation in the programs consumer dispute resolution system
    - Unknown legal "teeth"
      - Example of customer information viewed as salable property in court!!!
  - P3P: Supports browser checking of user-specific requirements
    - Internet Explorer 6 – disallow cookies if non-matching privacy policy
    - PrivacyBird – Internet Explorer plug-in from AT&T Research
- Corporate agreements
  - Stronger teeth/enforceability
  - But rarely protect the individual

# Secrecy

- Governmental sharing
  - Clear rules on sharing of classified information
  - Often err on the side of caution
    - Touching classified data "taints" everything
    - Prevents sharing that wouldn't disclose classified information
- Corporate secrets
  - Room for cost/benefit tradeoff
  - Authorization often a single office
    - Convince the right person that secrets aren't disclosed and work can proceed
- Bad Press
  - Lotus proposed "household marketplace" CD (1990)
    - Contained information on US households from public records
    - Public outcry forced withdrawal
  - Credit agencies maintain public and private information
    - Make money from using information for marketing purposes
  - Key difference? *Personal information isn't disclosed*
    - Credit agencies do the mining
    - "Purchasers" of information don't see public data

## Classes of Solutions

- Data Obfuscation
  - Nobody sees the *real* data
- Summarization
  - Only the needed facts are exposed
- Data Separation
  - Data remains with trusted parties

## Data Obfuscation

- Goal:  Hide the protected information
- Approaches
  - Randomly modify data
  - Swap values between records
  - Controlled modification of data to hide secrets
- Problems
  - Does it really protect the data?
  - Can we learn from the results?

# Example: US Census Bureau Public Use Microdata

- US Census Bureau summarizes by census block
  - Minimum 300 people
  - Ranges rather than values
- For research, "complete" data provided for sample populations
  - Identifying information removed
    - Limitation of detail: geographic distinction, continuous → interval
    - Top/bottom coding (eliminate sparse/sensitive values)
  - Swap data values among similar individuals (Moore '96)
    - Eliminates link between potential key and corresponding values
    - If individual determined, sensitive values likely incorrect

    Preserves the privacy of the individuals, as no entity in the data contains actual values for any real individual.
  - Careful swapping preserves multivariate statistics
    - Rank-based: swap similar values (randomly chosen within max distance)

    Preserves dependencies with (provably) high probability
  - Adversary can estimate sensitive values if individual identified

    *But data mining results enable this anyway!*

# Summarization

- Goal: Make only innocuous summaries of data available
- Approaches:
  - Overall collection statistics
  - Limited query functionality
- Problems:
  - Can we deduce data from statistics?
  - Is the information sufficient?

# Example: Statistical Queries

- User is allowed to query protected data
  - Queries must use statistical operators that summarize results
    - Example: Summation of total income for a group doesn't disclose individual income
  - Multiple queries can be a problem
    - Request total salary for all employees of a company
    - Request the total salary for all employees but the president
    - Now we know the president's salary
- Query restriction – Identify when a set of queries is safe (Denning '80)
  - *query set overlap control* (Dobkin, Jones, and Lipton '79)
    - Result generated from at least *k* items
    - Items used to generate result have at most *r* items in common with those used for previous queries
    - At least 1+(*k*-1)/*r* queries needed to compromise data
  - Data perturbation: introducing noise into the original data
  - Output perturbation: leaving the original data intact, but introducing noise into the results

# Example: Statistical Queries

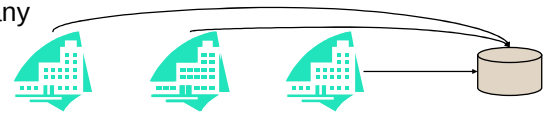
- Problem: Can approximate real values from multiple queries (Palley and Simonoff '87)
  - Create histograms for unprotected independent variables (e.g., job title)
  - Run statistical queries on the protected value (e.g., average salary)
  - Create a synthetic database capturing relationships between the unprotected and protected values
  - Data mining on the synthetic database approximate real values
- Problem with statistical queries is that the adversary creates the queries
  - Such manipulation likely to be obvious in a data mining situation
  - Problem: *Proving* that individual data not released

# Data Separation

- Goal: Only trusted parties see the data
- Approaches:
  - Data held by owner/creator
  - Limited release to trusted third party
  - Operations/analysis performed by trusted party
- Problems:
  - Will the trusted party be willing to do the analysis?
  - Do the analysis results disclose private information?

# Example: Patient Records

- My health records split among providers
  - Insurance company
  - Pharmacy
  - Doctor
  - Hospital



- Each agrees not to release the data without my consent
- Medical study wants correlations across providers
  - Rules relating complaints/procedures to "unrelated" drugs
- Does this need my consent?
  - *And that of every other patient!*
- It shouldn't!
  - Rules don't disclose my individual data

## When do we address these concerns?

- Must articulate that
  - A problem exists
    - There will be problems if we don't worry about privacy
  - We need to know the issues
    - Domain-specific constraints
  - A technical solution is feasible
    - Results valid
    - Constraints (provably) met

## What we need to know

- Constraints on release of data
  - Define in terms of Disclosure, not Privacy
  - What can be released, what mustn't
- Ownership/control of data
  - Nobody allowed access to "real" data
  - Data distributed across organizations
    - Horizontally partitioned: Each entity at a separate site
    - Vertically partitioned: Some attributes of each entity at each site
- Desired results: Rules? Classifier? Clusters?

# When to Address: CRISP-DM Stages

- Phase 1.2: Assess Situation
  - Capture privacy requirements while determining constraints
    *You've got the domain experts now – use them!*
- Phase 1.3: Determining data mining goals
  - Do the expected results violate constraints?
- Phase 2: Data understanding
  - Possible with non-private subset of data – Permission given or locally owned?
- Phase 3: Data preparation
  - 3.3: Will actual or derived (obfuscated) data be needed?
  - 3.4: Will warehouse-style integration be possible?
- Phase 4.1: Select modeling technique
  - Identify *(develop?)* technical solution
  - Document how solution meets constraints
- Phase 6.1: Plan deployment
  - Does the deployment satisfy constraints on use of results?

*CRoss Industry Standard Process for Data Mining*: www.crisp-dm.org

# Goal: Technical Solutions *that*

- Preserve privacy and security constraints
  - Disclosure Prevention that is
  - Provable, or
  - Disclosed data can be human-vetted
- Generate correct models: Results are
  - Equivalent to non-privacy preserving approach,
  - Bounded approximation to non-private result, or
  - Probabilistic approximation
- Efficient

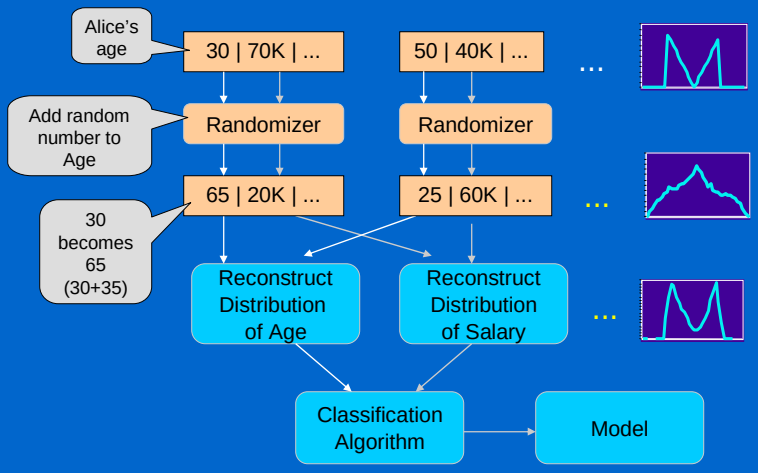
## Data Obfuscation Techniques

- Miner doesn't see the real data
  - Some knowledge of how data obscured
  - Can't reconstruct real values
- Results still valid
  - CAN reconstruct enough information to identify patterns
  - But not entities
- Example – *Agrawal & Srikant '00*

## Decision Trees
## *Agrawal and Srikant '00*

- Assume users are willing to
  - Give true values of certain fields
  - Give modified values of certain fields
- Practicality
  - 17% refuse to provide data at all
  - 56% are willing, as long as privacy is maintained
  - 27% are willing, with mild concern about privacy
- Perturb Data with Value Distortion
  - User provides $x_i+r$ instead of $x_i$
  - $r$ is a random value
    - Uniform, uniform distribution between $[-\alpha, \alpha]$
    - Gaussian, normal distribution with $\mu = 0, \sigma$

# Randomization Approach Overview



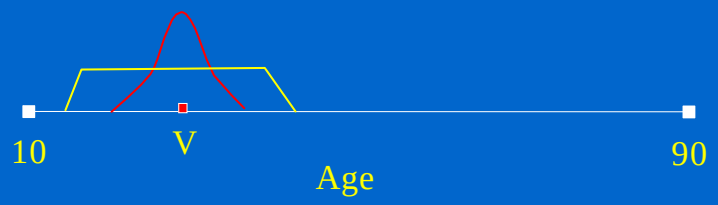

# Reconstruction Problem

- Original values $x_1, x_2, ..., x_n$
  - from probability distribution X (unknown)
- To hide these values, we use $y_1, y_2, ..., y_n$
  - from probability distribution Y
- Given
  - $x_1+y_1, x_2+y_2, ..., x_n+y_n$
  - the probability distribution of Y

  Estimate the probability distribution of X.

## Intuition (Reconstruct single point)

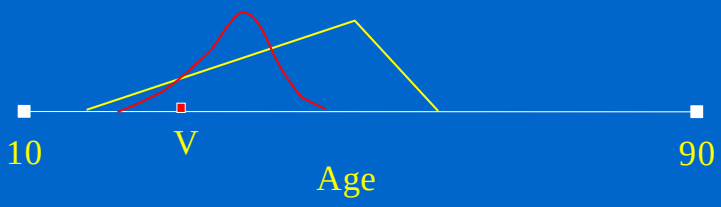
- Use Bayes' rule for density functions




Original distribution for Age
Probabilistic estimate of original value of V

## Intuition (Reconstruct single point)

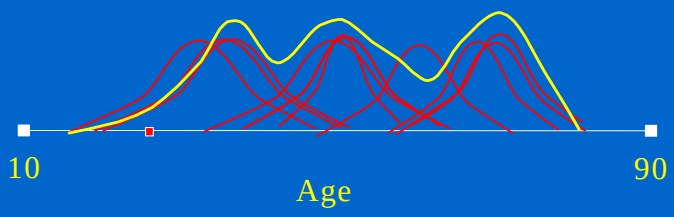
- Use Bayes' rule for density functions




Original Distribution for Age
Probabilistic estimate of original value of V

## Reconstructing the Distribution

- Combine estimates of where point came from for all the points:
  - Gives estimate of original distribution.



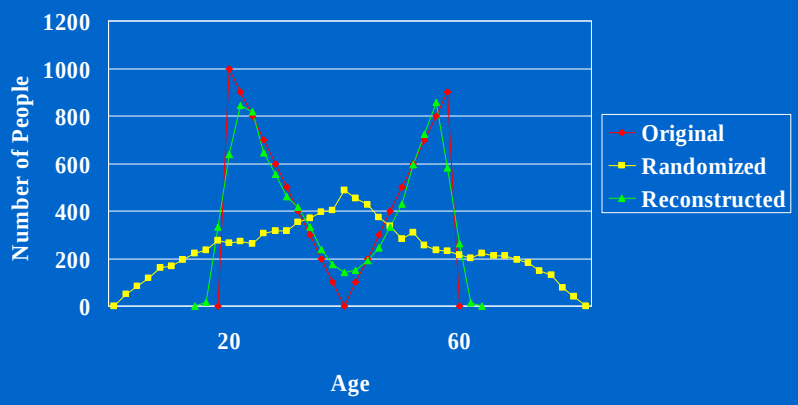

$$f_X = \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y((x_i + y_i) - a)\, f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a)\, f_X^j(a)}$$

## Reconstruction: Bootstrapping

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

- Converges to maximum likelihood estimate.
  - D. Agrawal & C.C. Aggarwal, PODS 2001.

## Works well




## Recap: Why is privacy preserved?

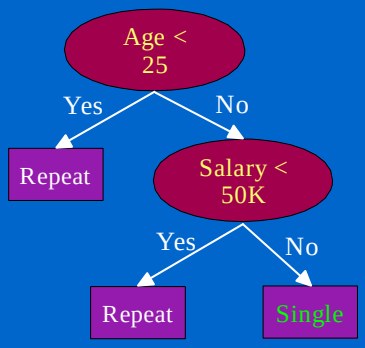
- Cannot reconstruct individual values accurately.
- Can only reconstruct distributions.

# Classification

- Naïve Bayes
  - Assumes independence between attributes.
- Decision Tree
  - Correlations are weakened by randomization, not destroyed.

# Decision Tree Example

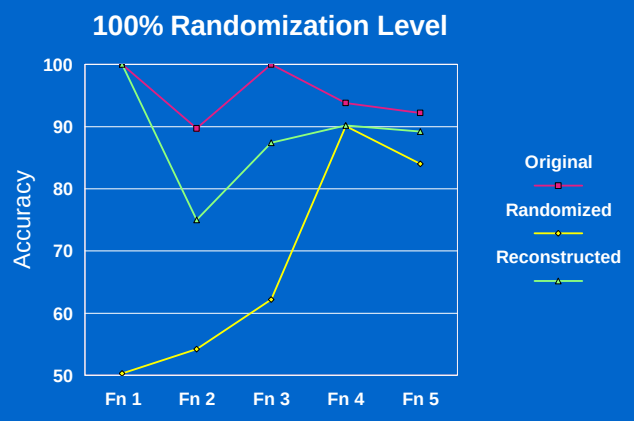
| Age | Salary | Repeat Visitor? |
|---|---|---|
| 23 | 50K | Repeat |
| 17 | 30K | Repeat |
| 43 | 40K | Repeat |
| 68 | 50K | Single |
| 32 | 70K | Single |
| 20 | 20K | Repeat |



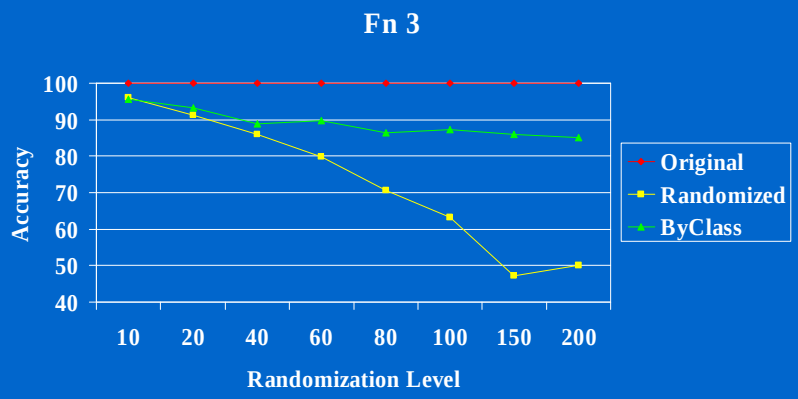
# Randomization Level

- 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) / (Range of Age: 100) ♏ 54% randomization level @ 90% confidence

# Decision Tree Experiments



## Accuracy vs. Randomization Level




## Quantification of Privacy
### *Agrawal and Aggarwal '01*

- Previous definition:

  If the original value can be estimated with c% confidence to lie in the interval [$\alpha_1$, $\alpha_2$], then the interval width ($\alpha_2$-$\alpha_1$) defines the amount of privacy at c% confidence level
- Ex: Interval width 2$\alpha$
  - confidence level 50% gives privacy $\alpha$
  - confidence level 100% gives privacy 2$\alpha$
- Incomplete in some situations

# Quantification of privacy II

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

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

Perturbing attribute Y is distributed uniformly between [-1,1]

- Privacy 2 at 100% confidence level
- Reconstruction with enough data, and Y-distribution public: $Z \in [-1,2]$ gives $X \in [0,1]$ and $Z \in [3,6]$ gives $X \in [4,5]$
- This means privacy offered by Y at 100% confidence level is at most 1. (X can be localized to even shorter intervals, e.g. Z=-0.5 gives $X \in [0,0.5]$ )

# Intuition

- Intuition: A random variable distributed uniformly between [0,1] has half as much privacy as if it were in [0,2]
- In general: If fB(x)=2fA(2x) then B offers half as much privacy as A
- Also: if a sequence of random variable An, n=1, 2, … converges to random variable B, then privacy inherent in An should converge to the privacy inherent in B

## Differential entropy

- Based on differential entropy h(A):

$h(A) = -\int_{\Omega A} f_A(a) \log_2 f_A(a) da$ where $\Omega_A$ is the domain of A

- Random variable *U* distributed between 0 and a, h(*U*)=$\log_2$(a). For a=1, h(*U*)=0
- Random variables with less uncertainty then uniform distribution on [0,1] have negative differential entropy, more uncertainty →positive differential entropy

## Proposed metric

- Propose $\Pi(A)=2^{h(A)}$ as measure of privacy for attribute A
- Uniform *U* between 0 and a: $\Pi(U)=2^{\log_2(a)}=a$
- General random variable A, $\Pi$(A) denote length of interval, over which a uniformly distributed random variable has equal uncertainty as A
- Ex: $\Pi(A)=2$ means A has as much privacy as a random variable distributed uniformly in an interval of length 2

# Conditional privacy

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

$$h(A \mid B) = -\int_{\Omega A,B} f_{A,B}(a,b) \log_2 f_{A|B=b}(a) da\, db$$

- Average conditional privacy of A given B:

$\Pi(A|B)=2^{h(A|B)}$

# Privacy loss

- Conditional privacy loss of A given B:

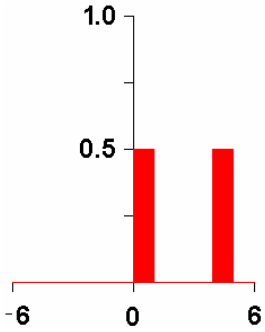
$P(A|B)=1-\Pi(A|B)/\Pi(A)=1-2^{h(A|B)}/2^{h(A)}=1-2^{-I(A;B)}$

Where $I(A;B)=h(A)-h(A|B)=h(B)-h(B|A)$

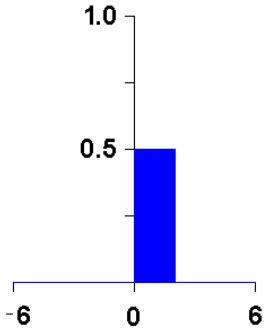
- I(A;B) is known as mutual information between random variables A and B
- $P(A|B)$ is the fraction of privacy of A which is lost by revealing B

# Example

- Look at earlier 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




- Intuition from figures: X has as much privacy as a uniform variable over an interval of length 2 –
- Areas are the same:




# Distribution Reconstruction: Agrawal and Aggarwal

- Expectation Maximization-based algorithm for Distribution Reconstruction
  - Generalizes Agrawal-Srikant algorithm
  - Better worst-case performance

*500 data points, uniform on [2,4], perturbed from [-1,1]*

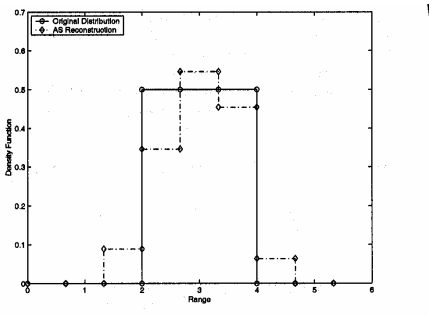



**Figure 2: Reconstructed Uniform Distribution (AS Algorithm)**

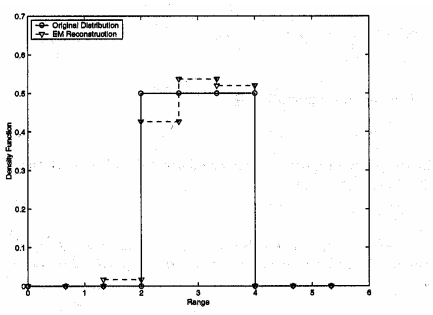



**Figure 3: Reconstructed Uniform Distribution (EM Algorithm)**

# Gaussian distribution

- Gaussian distribution, 500 data points, standard deviation of 2/πe
- Perturbing distribution – Gaussian, variance 1

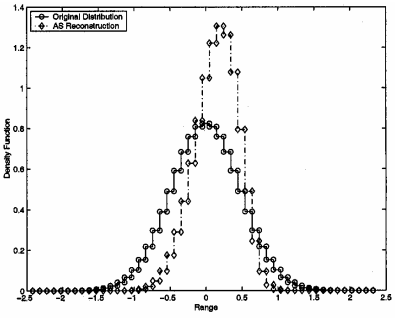



Figure 4: Reconstructed Gaussian Distribution (AS Algorithm)

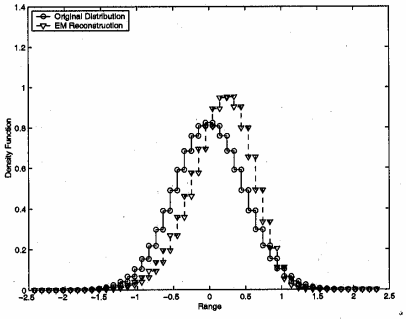



Figure 5: Reconstructed Gaussian Distribution (EM Algorithm)

# Information loss / privacy loss

- Gaussian – standard deviation of 2/pe
- Uniform distribution [-1,1] – same inherent privacy
- 500 data points

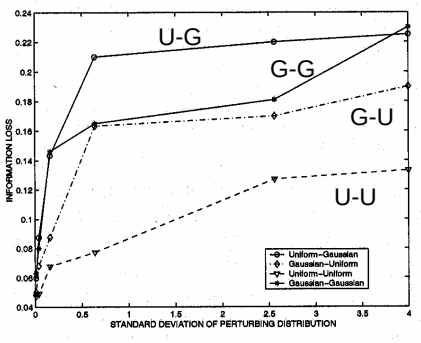



Figure 6: Information Loss with Standard Deviation of Perturbing Distribution

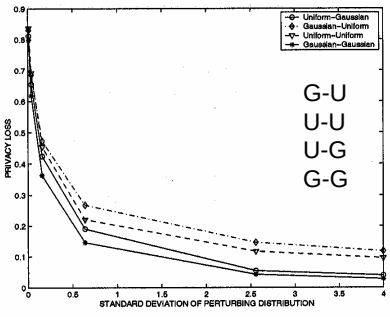


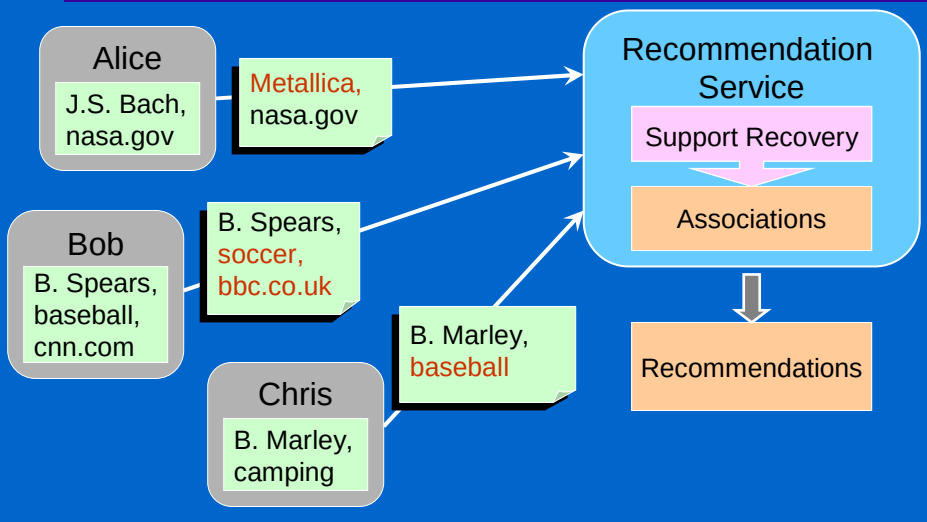

Figure 7: Privacy Loss with Standard Deviation of Perturbing Distribution

## Discovering Associations Over Privacy Preserved Categorical Data

A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, "Privacy Preserving Mining of Association Rules", KDD 2002.

- A transaction ***t*** is a set of items
- Support ***s*** for an itemset ***A*** is the number of transactions in which ***A*** appears
- Itemset ***A*** is frequent if $s \geq s_{min}$
- Task: Find all frequent itemsets, while preserving the privacy of individual transaction.

## Recommendation Service



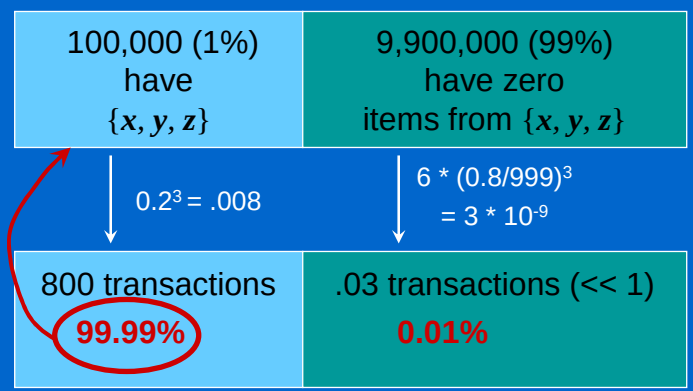
## Uniform Randomization

- Given a transaction,
  - keep item with 20% probability,
  - replace with a new random item with 80% probability.

Is there a problem?

## Example: $\{x, y, z\}$

10 M transactions of size 3 with 1000 items:




Uniform randomization: How many have $\{x, y, z\}$ ?

## Solution

"Where does a wise man hide a leaf? In the forest.
But what does he do if there is no forest?"
"He grows a forest to hide it in."

G.K. Chesterton

- Insert many false items into each transaction
- Hide true itemsets among false ones

## Cut and Paste Randomization

- Given transaction $t$ of size $m$, construct $t'$:
  - Choose a number $j$ between 0 and $K_m$ (cutoff);
  - Include $j$ items of $t$ into $t'$;
  - Each other item is included into $t'$ with probability $p_m$.

The choice of $K_m$ and $p_m$ is based on the desired level of privacy.

$t$ = a, b, c, u, v, w, x, y, z

$t'$ = b, v, x, z, œ, å, ß, ξ, ψ, €, ℵ, ъ, ħ, …

$j = 4$

## Partial Supports

To recover original support of an itemset, we need randomized supports of its subsets.

- Given an itemset ***A*** of size ***k*** and transaction size ***m***,
- A vector of partial supports of ***A*** is

$$\vec{s} = (s_0, s_1, \ldots, s_k), \text{ where}$$

$$s_l = \frac{1}{|T|} \cdot \#\{t \in T \mid \#(t \cap A) = l\}$$

  - Here $s_k$ is the same as the support of ***A***.
  - Randomized partial supports are denoted by $\vec{s}'$.

## Transition Matrix

- Let ***k*** = |***A***|, ***m*** = |***t***|.
- Transition matrix ***P*** = ***P*** (***k, m***) connects randomized partial supports with original ones:

$$\mathrm{E}\,\vec{s}' = P \cdot \vec{s}, \text{ where}$$

$$P_{l',l} = \Pr\left[\#(t' \cap A) = l' \mid \#(t \cap A) = l\right]$$

## The Estimators

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

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

- Covariance matrix for this estimator:

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

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

- To estimate it, substitute $\boldsymbol{s}_l$ with $(\boldsymbol{s}_{\text{est}})_l$.
  - Special case: estimators for support and its variance

## Privacy Breach Analysis

- How many added items are enough to protect privacy?
  - Have to satisfy Pr [$\boldsymbol{z} \in \boldsymbol{t} \mid \boldsymbol{A} \subseteq \boldsymbol{t'}$] < $\boldsymbol{\rho}$ ($\Leftrightarrow$ no privacy breaches)
  - Select parameters so that it holds for all itemsets.
  - Use formula ( $s_l^+ = \Pr\left[\#(t \cap A) = l, z \in t\right]$, $s_0^+ = 0$ ):

$$\Pr\left[z \in t \mid A \subseteq t'\right] = \sum_{l=0}^{k} s_l^+ \cdot P_{k,l} \Big/ \sum_{l=0}^{k} s_l \cdot P_{k,l}$$

- Parameters are to be selected in advance!
  - Enough to know maximal support of an itemset for each size.
  - Other parameters chosen for worst-case impact on privacy breaches.

## Can we still find frequent itemsets?

Privacy Breach level = 50%.

Soccer:

$s_{min} = 0.2\%$

| Itemset Size | True Itemsets | True Positives | False Drops | False Positives |
|---|---|---|---|---|
| 1 | 266 | 254 | 12 | 31 |
| 2 | 217 | 195 | 22 | 45 |
| 3 | 48 | 43 | 5 | 26 |

Mailorder:

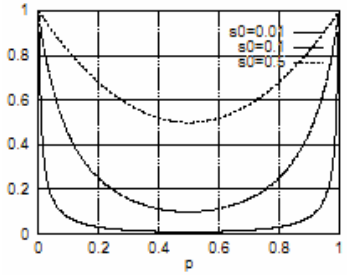
$s_{min} = 0.2\%$

| Itemset Size | True Itemsets | True Positives | False Drops | False Positives |
|---|---|---|---|---|
| 1 | 65 | 65 | 0 | 0 |
| 2 | 228 | 212 | 16 | 28 |
| 3 | 22 | 18 | 4 | 5 |

## Association Rules
### *Rizvi and Haritsa '02*

- "Market Basket" problem
  - Presence/absence of attributes in transactions
  - Few positive examples per transaction
- Bits "flipped" with probability *p*
  - Goal is low probability of knowing true value
  - Sparseness helps
- Mining the data
  - Get distorted data and *p*
  - $C^T = M^{-1}C^D$

$$M = \begin{bmatrix} p & 1-p \\ 1-p & p \end{bmatrix}, C = \begin{bmatrix} C_1 \\ C_0 \end{bmatrix}$$




Test: *p=0.9*, support=.25%

| *Length* | *Rules* | *Support Error* | *Missing* | *Extras* |
|---|---|---|---|---|
| 1 | 249 | 5.89 | 4.02 | 2.81 |
| 2 | 239 | 3.87 | 6.89 | 9.59 |
| 3 | 73 | 2.60 | 10.96 | 9.59 |
| 4 | 4 | 1.41 | 0 | 25.0 |

## Data Separation

- Data holders trusted with content
  - But only their own
- Mustn't share
  - But this doesn't prevent global models

## Secure Multiparty Computation
*It can be done!*

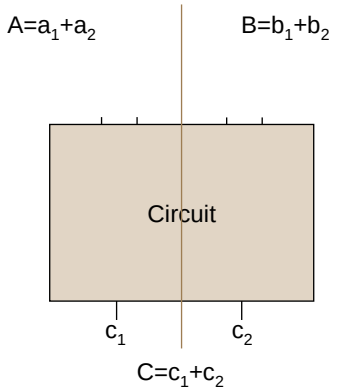
- Goal: Compute function when each party has some of the inputs
- Yao's Millionaire's problem *(Yao '86)*
  - Secure computation possible if function can be represented as a circuit
  - Idea: Securely compute gate
    - Continue to evaluate circuit
- Works for multiple parties as well *(Goldreich, Micali, and Wigderson '87)*

# Secure Multiparty Computation: Definitions

- Secure
  - Nobody knows anything but their own input and the results
  - Formally: ∃ polynomial time S such that {S(x,f(x,y))} ≡ {View(x,y)}
- Semi-Honest model: follow protocol, but remember intermediate exchanges
- Malicious: "cheat" to find something out

# How does it work?

- Each side has input, knows circuit to compute function
- Add random value to your input, give to other side
  - Each side has *share* of all inputs
- Compute share of output
  - Add results at end
- XOR gate: just add locally
- AND gate: send your share encoded in truth table
  - Oblivious transfer allows other side to get only correct value out of truth table




| value of $(a_2,b_2)$ | (0,0) | (0,1) | (1,0) | (1,1) |
|---|---|---|---|---|
| OT-input | 1 | 2 | 3 | 4 |
| value of output | $c_1+a_1b_1$ | $c_1+a_1(b_1+1)$ | $c_1+(a_1+1)b_1$ | $c_1+(a_1+1)(b_1+1)$ |

# Oblivious Transfer

- What is it?
  - *A* has inputs $a_i$
  - *B* makes choice
  - *A* doesn't know choice, *B* only sees chosen value.
- How?
  - *A* sends public key *p* to B
  - B selects 4 random values *b*
    - encrypts (only) $b_{choice}$ with $f_p$, sends all to *A*
  - *A* decrypts all with private key, sends to B:
    $c_i = a_i \oplus e(f_p^{-1}(b_i))$
  - B outputs $c_{choice} \oplus e(b_{choice}) = a_{choice} \oplus e(f_p^{-1}(f_p(b_{choice}))) \oplus e(b_{choice})$

# Decision Tree Construction *(Lindell & Pinkas '00)*

- Two-party horizontal partitioning
  - Each site has same schema
  - Attribute set known
  - Individual entities private
- Learn a decision tree classifier
  - ID3
- Essentially ID3 meeting Secure Multiparty Computation Definitions

# Key Assumptions/Limitations

- Protocol takes place in the semi-honest model
- Only Two-party case considered
  - Extension to multiple parties is not trivial
- Computes an ID3 approximation
  - Protocol for computation of $ID3_\delta \in ID3_\delta$
  - δ-approximation of ID3
  - δ has implications on efficiency
- Deals only with categorical attributes

# Cryptographic Tools

- Oblivious Transfer
  - 1-out-of-2 oblivious transfer. Two parties, sender and receiver. Sender has two inputs $<X_0,X_1>$ and the receiver has an input $\alpha \in (0,1)$. At the end of the protocol the receiver should get $X_\alpha$ and nothing else and the sender should learn nothing.
- Oblivious Evaluation of Polynomials
  - Sender has polynomial P of degree k over some finite field F and a receiver with an element z in F (the degree k is public). The receiver obtains P(z) without learning anything about the polynomial P and the sender learns nothing about z.
- Oblivious Circuit Evaluation
  - Two party Yao's protocol. A has input x and B has a function f and a combinatorial circuit that computes f. At the end of the protocol A outputs f(x) and learns no other information about f while B learns nothing at all.

# ID3

- R – the set of attributes
- C – the class attribute
- T – the set of transactions

**ID3**$(R, C, T)$

1. If $R$ is empty, return a leaf-node with the class value assigned to the most transactions in $T$.
2. If $T$ consists of transactions which all have the same value $c$ for the class attribute, return a leaf-node with the value $c$ (finished classification path).
3. Otherwise,
   (a) Determine the attribute that *best* classifies the transactions in $T$, let it be $A$.
   (b) Let $a_1, ..., a_m$ be the values of attribute $A$ and let $T(a_1), ..., T(a_m)$ be a partition of $T$ such that every transaction in $T(a_i)$ has the attribute value $a_i$.
   (c) Return a tree whose root is labeled $A$ (this is the test attribute) and has edges labeled $a_1, ..., a_m$ such that for every $i$, the edge $a_i$ goes to the tree ID3$(R - \{A\}, C, T(a_i))$.

# Privacy Preserving ID3

Step 1: *If R is empty, return a leaf-node with the class value assigned to the most transactions in T*

- Set of attributes is public
  – Both know if R is empty
- Run Yao's protocol for the following functionality:
  – Inputs $(|T_1(c_1)|,\ldots,|T_1(c_L)|)$, $(|T_2(c_1)|,\ldots,|T_2(c_L)|)$
  – Output *i* where $|T_1(c_i)|+|T_2(c_i)|$ is largest

# Privacy Preserving ID3

Step 2: *If T consists of transactions which have all the same value c for the class attribute, return a leaf node with the value c*

- Represent having more than one class (in the transaction set), by a fixed symbol different from $c_i$,
- Force the parties to input either this fixed symbol or $c_i$
- Check equality to decide if at leaf node for class $c_i$
- Various approaches for equality checking
  - Yao'86
  - Fagin, Naor '96
  - Naor, Pinkas '01

# Privacy Preserving ID3

- Step 3:(a) *Determine the attribute that best classifies the transactions in T, let it be A*
  - Essentially done by securely computing x*(ln x)
- (b,c) *Recursively call $ID3_\delta$ for the remaining attributes on the transaction sets $T(a_1),\ldots,T(a_m)$ where $a_1,\ldots, a_m$ are the values of the attribute A*
  - Since the results of 3(a) and the attribute values are public, both parties can individually partition the database and prepare their inputs for the recursive calls

# Determining the best attribute

- Let A have m possible values $a_1,\ldots,a_m$,
  C have l possible values $c_1,\ldots,c_l$
- $T(a_j)$ is transactions with attribute A set to $a_j$
  $T(a_j, c_i)$ is transactions with A set to $a_j$ and class $c_i$
- Conditional entropy is the weighted sum of entropies, which is simplified as follows:

$$\begin{aligned} H_C(T|A) &= \sum_{j=1}^{m} \frac{|T(a_j)|}{|T|} H_C(T(a_j)) \\ &= \frac{1}{|T|} \sum_{j=1}^{m} |T(a_j)| \sum_{i=1}^{\ell} -\frac{|T(a_j,c_i)|}{|T(a_j)|} \cdot \log(\frac{|T(a_j,c_i)|}{|T(a_j)|}) \\ &= \frac{1}{|T|} \left( -\sum_{j=1}^{m} \sum_{i=1}^{\ell} |T(a_j,c_i)| \log(|T(a_j,c_i)|) + \sum_{j=1}^{m} |T(a_j)| \log(|T(a_j)|) \right) \end{aligned}$$

# X ln X

- Taylor Series of natural logarithm:

$$\ln(1+\varepsilon) = \sum_{i=1}^{\infty} \frac{(-1)^{i-1}\varepsilon^i}{i} = \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \frac{\varepsilon^4}{4} + \cdots \qquad \text{for } -1 < \varepsilon < 1$$

- Error for partial evaluation:

$$\left| \ln(1+\varepsilon) - \sum_{i=1}^{k} \frac{(-1)^{i-1}\varepsilon^i}{i} \right| < \frac{|\varepsilon|^{k+1}}{k+1} \cdot \frac{1}{1-|\varepsilon|}$$

- Error shrinks exponentially as k grows

$$\ln(x) = \ln(2^n(1+\varepsilon)) = n\ln 2 + \varepsilon - \frac{\varepsilon^2}{2} + \frac{\varepsilon^3}{3} - \frac{\varepsilon^4}{4} + \cdots$$

# Comparison

- Fully Generic solution |R|*|T|*logm oblivious transfers (for every bit)
- Semi generic protocol (uses circuit evaluation for x ln x)
  - Computes Taylor series (k multiplications)
  - $O(k^3 \log^2|T||S|)$ since multiplication is quadratic in terms of input size
- Their solution - O(klog|T|*|S|) bits
  - Order $O(k^2 \log|T|)$ more efficient

# Association Rule Mining: Horizontal Partitioning

- Distributed Association Rule Mining: Easy without sharing the individual data *[Cheung+'96] (Exchanging support counts is enough)*
- What if we do not want to reveal which rule is supported at which site, the support count of each rule, or database sizes?
  - Hospitals want to participate in a medical study
  - But rules only occurring at one hospital may be a result of bad practices
    - *Is the potential public relations / liability cost worth it?*

## Overview of the Method *(Kantarcioglu and Clifton '02)*

- Find the union of the locally large candidate itemsets securely
- After the local pruning, compute the globally supported large itemsets securely
- At the end check the confidence of the potential rules securely

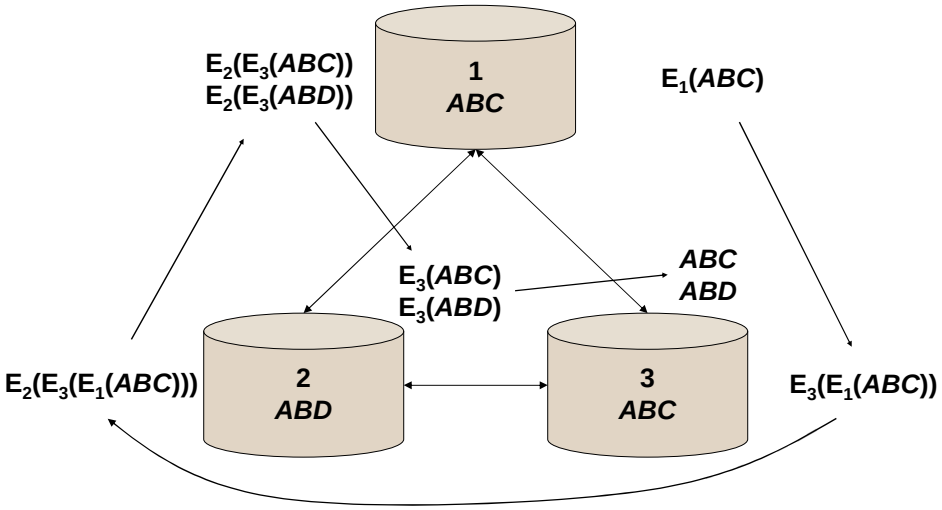
## Securely Computing Candidates

- Key: Commutative Encryption ($E_a(E_b(x)) = E_b(E_a(x))$)
  - Compute local candidate set
  - Encrypt and send to next site
    - Continue until all sites have encrypted all rules
  - Eliminate duplicates
    - Commutative encryption ensures if rules the same, encrypted rules the same, regardless of order
  - Each site decrypts
    - After all sites have decrypted, rules left
- Care needed to avoid giving away information through ordering/etc.

Redundancy maybe added in order to increase the security.

*Not fully secure according to definitions of secure multi-party*

# Computing Candidate Sets




# Compute Which Candidates Are Globally Supported?

- Goal: To check whether

$$X.\sup \geq s * \sum_{i=1}^{n} |DB_i| \quad (1)$$

$$\sum_{i=1}^{n} X.\sup_i \geq \sum_{i=1}^{n} s^* | DB_i | \quad (2)$$

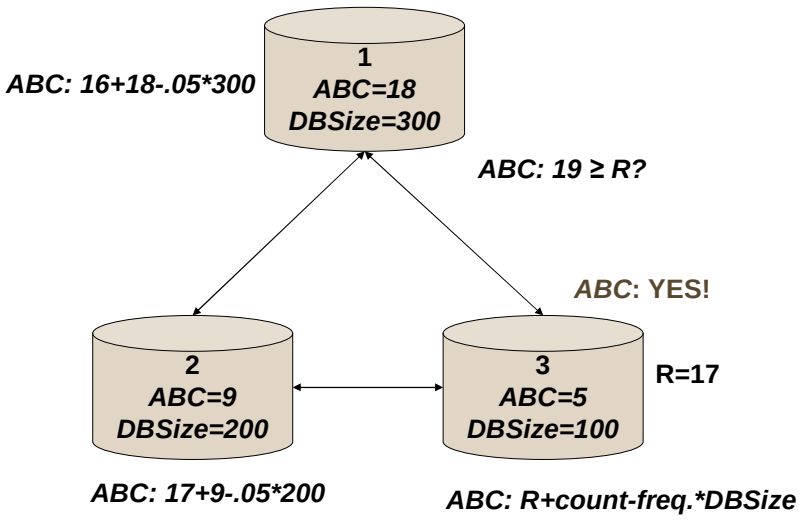
$$\sum_{i=1}^{n} (X.\sup_i - s^* | DB_i |) \geq 0 \quad (3)$$

Note that checking inequality (1) is equivalent to checking inequality (3)

## Which Candidates Are Globally Supported? (Continued)

- Now securely compute Sum ≥ 0:
  - Site$_0$ generates random *R*
    Sends *R+count*$_0$ *– frequency*dbsize*$_0$ to site$_1$
  - Site$_k$ adds *count*$_k$ *– frequency*dbsize*$_k$, sends to site$_{k+1}$
- Final result: Is sum at site$_n$ - *R* ≥ 0?
  - Use Secure Two-Party Computation
- This protocol is secure in the semi-honest model

## Computing Frequent: Is *ABC* ≥ 5%?



# Computing Confidence

- Checking confidence can be done by the previous protocol. Note that checking confidence for X $\Rightarrow$ Y

$$\frac{\{X \cup Y\}.\sup}{X.\sup} \geq c \Rightarrow \frac{\sum_{i=1}^{n} XY.\sup_i}{\sum_{i=1}^{n} X.\sup_i} \geq c$$

$$\Rightarrow \sum_{i=1}^{n} (XY.\sup_i - c * X.\sup_i) \geq 0$$

# Association Rules in Vertically Partitioned Data

- Two parties – Alice (A) and Bob (B)
- Same set of entities (data cleansing, join assumed done)
- A has *p* attributes, $A_1 \ldots A_p$
- B has *q* attributes, $B_1 \ldots B_q$
- Total number of transactions, *n*
- Support Threshold, *k*

| | | | | | | |
|---|---|---|---|---|---|---|
| JSV | Brain Tumor | Diabetic | *JSV* | *5210* | *Li/Ion* | *Piezo* |

# Vertically Partitioned Data *(Vaidya and Clifton '02)*

- Learn globally valid association rules
- Prevent disclosure of individual relationships
  - Join key revealed
  - Universe of attribute values revealed
- Many real-world examples
  - Ford / Firestone
  - FBI / IRS
  - Medical records

# Basic idea

- Find out if itemset $\{A_1, B_1\}$ is frequent (i.e., If support of $\{A_1, B_1\} \geq k$)

A

| Key | $A_1$ |
|---|---|
| $k_1$ | 1 |
| $k_2$ | 0 |
| $k_3$ | 0 |
| $k_4$ | 1 |
| $k_5$ | 1 |

B

| Key | $B_1$ |
|---|---|
| $k_1$ | 0 |
| $k_2$ | 1 |
| $k_3$ | 0 |
| $k_4$ | 1 |
| $k_5$ | 1 |

- Support of itemset is defined as number of transactions in which all attributes of the itemset are present
- For binary data, support $=|A_i \wedge B_i|$
- Boolean AND can be replaced by normal (arithmetic) multiplication.

# Basic idea

- Thus, $Support = \sum_{i=1}^{n} A_i \times B_i$
- This is the scalar (dot) product of two vectors
- To find out if an arbitrary (shared) itemset is frequent, create a vector on each side consisting of the component multiplication of all attribute vectors on that side (contained in the itemset)
- E.g., to find out if $\{A_1, A_3, A_5, B_2, B_3\}$ is frequent
  - A forms the vector X = $\prod A_1 A_3 A_5$
  - B forms the vector Y = $\prod B_2 B_3$
  - Securely compute the dot product of X and Y

# The algorithm

1. $L_1$ = {large 1-itemsets}
2. for (k=2; $L_{k-1} \neq \phi$; k++) do begin
3. $C_k$ = apriori-gen($L_{k-1}$);
4. for all candidates $c \in C_k$ do begin
5. if all the attributes in $c$ are entirely at A or B
6. that party independently calculates $c.count$
7. else
8. let A have $l$ of the attributes and B have the remaining $m$ attributes
9. construct $\vec{X}$ on A's side and $\vec{Y}$ on B's side where $\vec{X} = \prod_{i=1}^{l} \vec{A}_i$ and $\vec{Y} = \prod_{i=1}^{m} \vec{B}_i$
10. compute $c.count = \vec{X}.\vec{Y} = \sum_{i=1}^{n} x_i * y_i$
11. endif
12. $L_k = L_k \cup c | c.count \geq minsup$
13. end
14. end
15. Answer = $\cup_k L_k$

# Protocol

- A generates $n/2$ randoms, $R_1 \ldots R_{n/2}$
- A sends the following $n$ values to B

$$\left\langle x_1 + a_{1,1} * R_1 + a_{1,2} * R_2 + \cdots + a_{1,n/2} * R_{n/2} \right\rangle$$
$$\left\langle x_2 + a_{2,1} * R_1 + a_{2,2} * R_2 + \cdots + a_{2,n/2} * R_{n/2} \right\rangle$$
$$\vdots$$
$$\left\langle x_n + a_{n,1} * R_1 + a_{n,2} * R_2 + \cdots + a_{n,n/2} * R_{n/2} \right\rangle$$

- The $(n^2/2)$ $a_{i,j}$ values are known to both A and B

# Protocol (cont.)

- B multiplies each value he gets with the corresponding y value he has and adds all of them up to get a sum S, which he sends to A.

$$S = \left[ \begin{array}{l} y_1 * \{x_1 + (a_{1,1} * R_1 + a_{1,2} * R_2 + \cdots + a_{1,n/2} * R_{n/2})\} \\ + y_2 * \{x_2 + (a_{2,1} * R_1 + a_{2,2} * R_2 + \cdots + a_{2,n/2} * R_{n/2})\} \\ \vdots \\ + y_n * \{x_n + (a_{n,1} * R_1 + a_{n,2} * R_2 + \cdots + a_{n,n/2} * R_{n/2})\} \end{array} \right]$$

- Group the $x_i * y_i$ terms, and expand the equations

# Protocol (cont)

$$S = \sum_{i=1}^{n} x_i * y_i$$

$$x_1 * y_1 + x_2 * y_2 + \cdots + x_n * y_n$$

$$+\left( a_{1,1} * y_1 * R_1 + a_{1,2} * y_1 * R_2 + \cdots + a_{1,n/2} * y_1 * R_{n/2} \right)$$

$$+\left( a_{2,1} * y_2 * R_1 + a_{2,2} * y_2 * R_2 + \cdots + a_{2,n/2} * y_2 * R_{n/2} \right)$$

$$\vdots$$

$$+\left( a_{n,1} * y_n * R_1 + a_{n,2} * y_n * R_2 + \cdots + a_{n,n/2} * y_n * R_{n/2} \right)$$

Grouping components vertically and factoring out $R_i$

# Protocol (complete)

$$S = \sum_{i=1}^{n} x_i * y_i$$

$$+ R_1 * \left( a_{1,1} * y_1 + a_{2,1} * y_2 + \cdots + a_{n,1} * y_n \right)$$

$$+ R_2 * \left( a_{1,2} * y_1 + a_{2,2} * y_2 + \cdots + a_{n,2} * y_n \right)$$

$$\vdots$$

$$+ R_{n/2} * \left( a_{1,n/2} * y_1 + a_{2,n/2} * y_2 + \cdots + a_{n,n/2} * y_n \right)$$

- A already knows $R_1 \ldots R_{n/2}$
- Now, if B sends these n/2 values to A,
- *A can remove the baggage and get the scalar product*

# Security Analysis

- A sends to B
  - n values (which are linear equations in 3n/2 unknowns – the n x-values and n/2 R-values)
  - The final result (which reveals another linear equation in the n/2 R-values) (Note – this can be avoided by allowing A to only report if scalar product exceeds threshold)
- B sends to A
  - The sum, S (which is one linear equation in the n y-values)
  - n/2 values (which are linear equations in n unknowns – the n y-values)

# Security Analysis

- Security based on the premise of revealing less equations than the number of unknowns – possible solutions infinite!
- Security of both is symmetrical
- Just from the protocol, nothing can be found out
- Everything is revealed *only* when about half the values are revealed

## The Trouble with {0,1}

- Input values are restricted only to 0 or 1
- Parties reveal linear equation in values
  - Adversary could try all combinations of {0,1} and see which fits
- Solution: Eliminate unique solution
  - Create $a_{i,j}$ values so 0's and 1's paired
  - No way of knowing *which* is 0 or 1
- *Completely different approach for three or more parties*

## EM Clustering *(Lin & Clifton '03)*

- Goal: EM Clustering in Horizontally Partitioned Data
  - Avoid sharing individual values
  - Nothing should be attributable to individual site
- Solution: Partition estimation update
  - Each site computes portion based on it's values
  - Securely combine these to complete iteration

# Expectation Maximization

- log $L_c(\Psi)$ = log $f_c(\mathbf{x}; \Psi)$:
- E-Step: On the (t+1)st step, calculate the expected complete data log likelihood given observed data values.
  - $G(\Psi; \Psi^{(t)}) = E_{\Psi(t)}\{\log L_c(\Psi)||y\}$
- M-Step: Find $\Psi^{(t+1)}$ to maximize $G(\Psi; \Psi^{(t)})$
- For finite normal mixtures:

$$f(y,\Psi) = \sum_{i=1}^{k} \pi_i f_i(y;\theta_i) \text{ where } f_i(y;\theta_i) = (2\pi_i\sigma_i^2)^{-1/2} \exp\{k - \frac{(y-\mu_i)^2}{2\sigma_i^2}\}$$

# EM Clustering: Process

- Estimate $\mu_i$, $\pi$, and $\sigma^2$ at each iteration
  - $\mu_i^{(t+1)} = \sum_{j=1}^{n} z_{ij}^{(t)} y_j \,/\, \sum_{j=1}^{n} z_{ij}^{(t)}$
  - $\sigma_i^{2(t+1)} = \sum_{j=1}^{n} z_{ij}^{(t)} (y_j - \mu_i^{(t+1)})^2 \,/\, n$
  - $\pi_i^{(t+1)} = \sum_{j=1}^{n} z_{ij}^{(t)} \,/\, n$
- Each Sum can be partitioned across sites
  - Compute global sum securely *(Kantarcioglu and Clifton '02)*

## What if the Secrets are in the Results?

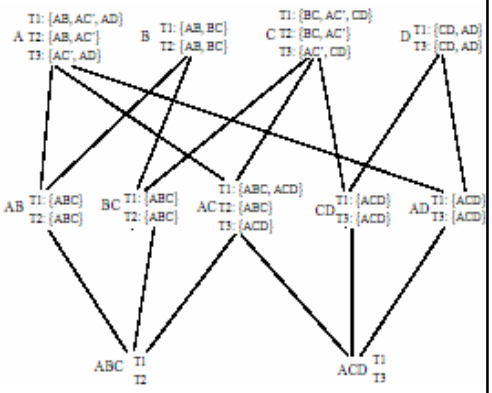
- Assume we want to make data available
  - Example: Telephone directory
- But the data contains rules we don't want people to learn
  - Areas of corporate expertise
- How do we hide the rules?
  - While minimizing effect on the data!

## Disclosure Limitation of Sensitive Rules *(Atallah et. al. '99)*

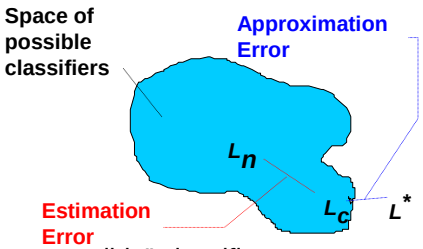
- Given a database and a set of "secret" rules, modify database to hide rules
  Change 1's to 0's and vice-versa to
  - Lower support
  - Lower confidence
- Goal: Minimize effect on non-sensitive rules
  - Problem shown to be NP-Hard!

# Heuristic Solution: Minimize effect on small itemsets

- Build graph of all supported itemsets
- To hide large itemset:
  - Go up tree to find item with lowest support
  - Select transaction affecting fewest 2-itemsets
  - Remove item from that transaction




# What if we don't know what we want to hide? *(Clifton '00)*




Total (mean-squared) error

$$\int ( c - L(a) )^2 P(a, c)\, da\, dc$$

- $L^*$ : "best possible" classifier
- $L_n$ : classifier learned from the sample
- $L_c$ : best classifier from those that can be described by the given classification model (e.g decision tree, neural network)

Goal: determine sample size so expected error is sufficiently large *regardless of technique used to learn classifier.*

## Sample size needed to classify when zero approximation error

- Let *C* be a class of discrimination functions with VC dimension $V \geq 2$. Let *X* be the set of all random variables *(X,Y)* for which $L_C = 0$. For $\delta \leq 1/10$ and $\varepsilon < 1/2$

$$N(\varepsilon,\delta) \geq \frac{V-1}{12\varepsilon}$$

*Intuition: This is difficult, because we must have a lot of possible classifiers for one to be perfect.*

## Sample size needed to classify when no perfect classifier exists

- Let *C* be a class of discrimination functions with VC dimension $V \geq 2$. Let *X* be the set of all random variables *(X,Y)* for which for fixed $L \in (0, ½)$

$$L = \inf_{g \in \boldsymbol{C}} \mathrm{P}\{g(X) \neq Y\}.$$

Then for every discrimination rule $g_n$ based on $X_1, Y_1, \ldots, X_n, Y_n$,

$$N(\varepsilon,\delta) \geq \frac{L(V-1)\mathrm{e}^{-10}}{32} \times \min\left(\frac{1}{\delta^2}, \frac{1}{\varepsilon^2}\right)$$

and also, for $\varepsilon \leq L$ ε ¼,

$$N(\varepsilon,\delta) \geq \frac{L}{4\varepsilon^2} \log \frac{1}{4\delta}$$

*Intuition: If the space of classifiers is small, getting the best one is easy (but it isn't likely to be very good).*

# Summary

- Privacy and Security Constraints can be impediments to data mining
  - Problems with access to data
  - Restrictions on sharing
  - Limitations on use of results
- Technical solutions possible
  - Randomizing / swapping data doesn't prevent learning good models
  - We don't need to share data to learn global results
  - References to more solutions in the tutorial notes
- *Still lots of work to do!*

## References


[1] C. Clifton, *Handbook of Data Mining.* Lawrence Erlbaum Associates, Apr. 2003, ch. 18: Security and Privacy, pp. 441–452.

[2] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: $k$-anonymity and its enforcement through generalization and suppression," in *Proceedings of the IEEE Symposium on Research in Security and Privacy*, Oakland, CA, May 1998.

[3] L. Sweeney, "Computational disclosure control: A primer on data privacy protection," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[4] "Directive 95/46/EC of the european parliament and of the council of 24 october 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Communities*, vol. No I., no. 281, pp. 31–50, Oct. 24 1995. [Online]. Available: http://europa.eu.int/comm/internal_market/en/dataprot/law/

[5] "Standard for privacy of individually identifiable health information," *Federal Register*, vol. 66, no. 40, Feb. 28 2001. [Online]. Available: http://www.hhs.gov/ocr/hipaa/finalreg.html

[6] S. Blackmer and Wilmer, Cutler & Pickering, "Transborder personal data flows: Administrative practice," in *The Privacy and American Business Meeting on Model Data Protection Contracts and Laws*, Washington, DC, Feb. 24-25 1998. [Online]. Available: http://www.privacyexchange.org/tbdi/pdataflow.html

[7] "The platform for privacy preferences 1.0 (P3P1.0) specification," Apr. 16 2002. [Online]. Available: http://www.w3.org/TR/P3P/

[8] "Privacy bird," July 2002. [Online]. Available: http://www.privacybird.com

[9] R. A. Moore, Jr., "Controlled data-swapping techniques for masking public use microdata sets," U.S. Bureau of the Census, Washington, DC., Statistical Research Division Report Series RR 96-04, 1996. [Online]. Available: http://www.census.gov/srd/papers/pdf/rr96-4.pdf

[10] K. Muralidhar, R. Sarathy, and R. A. Parsa, "A general additive perturbation method for database security," *Management Science*, vol. 45, no. 10, pp. 1399–1415, 1999.

[11] K. Muralidhar, R. Sarathy, and R. A. Parsa, "An improved security requirement for data perturbation with implications for e-commerce," *Decision Science*, vol. 32, no. 4, pp. 683–698, Fall 2001.

[12] D. E. Denning, "Secure statistical databases with random sample queries," *ACM Transactions on Database Systems*, vol. 5, no. 3, pp. 291–315, Sept. 1980. [Online]. Available: http://doi.acm.org/10.1145/320613.320616

[13] D. Dobkin, A. K. Jones, and R. J. Lipton, "Secure databases: Protection against user influence," *ACM Transactions on Database Systems*, vol. 4, no. 1, pp. 97–106, Mar. 1979. [Online]. Available: http://doi.acm.org/10.1145/320064.320068

[14] N. R. Adam and J. C. Wortmann, "Security-control methods for statistical databases: A comparative study," *ACM Computing Surveys*, vol. 21, no. 4, pp. 515–556, Dec. 1989. [Online]. Available: http://doi.acm.org/10.1145/76894.76895

[15] M. A. Palley and J. S. Simonoff, "The use of regression methodology for the compromise of confidential information in statistical databases," *ACM Transactions on Database Systems*, vol. 12, no. 4, pp. 593–608, Dec. 1987. [Online]. Available: http://doi.acm.org/10.1145/32204.42174

[16] "Cross industry standard process for data mining," http://www.crisp-dm.org, Dec. 1999. [Online]. Available: http://www.crisp-dm.org

[17] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*. Dallas, TX: ACM, May 14-19 2000, pp. 439–450. [Online]. Available: http://doi.acm.org/10.1145/342009.335438

[18] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. Santa Barbara, California, USA: ACM, May 21-23 2001, pp. 247–255. [Online]. Available: http://doi.acm.org/10.1145/375551.375602

[19] C. Clifton and V. Estivill-Castro, Eds., *IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining*, vol. 14. Maebashi City, Japan: Australian Computer Society, Dec. 9 2002. [Online]. Available: http://crpit.com/Vol14.html

[20] "Special section on privacy and security," *SIGKDD Explorations*, vol. 4, no. 2, pp. i–48, Jan. 2003. [Online]. Available: http://www.acm.org/sigs/sigkdd/explorations/issue4-2/contents.htm

[21] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 217–228. [Online]. Available: http://doi.acm.org/10.1145/775047.775080

[22] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proceedings of 28th International Conference on Very Large Data Bases*. Hong Kong: VLDB, Aug. 20-23 2002, pp. 682–693. [Online]. Available: http://www.vldb.org/conf/2002/S19P03.pdf

[23] R. Agrawal, A. Evfimievski, and R. Srikant, "Information sharing across private databases," in *Proceedings of ACM SIGMOD International Conference on Management of Data*, San Diego, California, June 9-12 2003.

[24] A. C. Yao, "How to generate and exchange secrets," in *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*. IEEE, 1986, pp. 162–167.

[25] O. Goldreich, S. Micali, and A. Wigderson, "How to play any mental game - a completeness theorem for protocols with honest majority," in *19th ACM Symposium on the Theory of Computing*, 1987, pp. 218–229. [Online]. Available: http://doi.acm.org/10.1145/28395.28420

[26] O. Goldreich, "Secure multi-party computation," Sept. 1998, (working draft). [Online]. Available: http://www.wisdom.weizmann.ac.il/~oded/pp.html

[27] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in *Advances in Cryptology – CRYPTO 2000*. Springer-Verlag, Aug. 20-24 2000, pp. 36–54. [Online]. Available: http://link.springer.de/link/service/series/0558/bibs/1880/18800036.htm

[28] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in *The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02)*, Madison, Wisconsin, June 2 2002, pp. 24–31. [Online]. Available: http://www.bell-labs.com/user/minos/DMKD02/Papers/kantarcioglu.pdf

[29] M. Kantarcıoğlu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," *IEEE Transactions on Knowledge and Data Engineering*, to appear.

[30] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Alberta, Canada, July 23-26 2002, pp. 639–644. [Online]. Available: http://doi.acm.org/10.1145/775047.775142

[31] J. Vaidya and C. Clifton, "Privacy-preserving $k$-means clustering over vertically partitioned data," in *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Washington, DC, Aug. 24-27 2003.

[32] B. Rozenberg and E. Gudes, "Privacy preserving frequent item-set mining in vertically partitioned databases," in *Seventeenth Annual IFIP WG 11.3 Working Conference on Data and Applications Security*, Estes Park, Colorado, U.S.A., Aug. 4-6 2003.

[33] X. Lin, C. Clifton, and M. Zhu, "Privacy preserving clustering with distributed EM mixture modeling," *Knowledge and Information Systems*, to appear.

[34] H. Kargupta, "Distributed data mining for pervasive and privacy-sensitive applications," in *National Science Foundation Workshop on Next Generation Data Mining*, H. Kargupta, A. Joshi, and K. Sivakumar, Eds., Baltimore, MD, Nov. 1-3 2002, pp. 109–117.

[35] C. Clifton, M. Kantarcioglu, and J. Vaidya, "Defining privacy for data mining," in *National Science Foundation Workshop on Next Generation Data Mining*, H. Kargupta, A. Joshi, and K. Sivakumar, Eds., Baltimore, MD, Nov. 1-3 2002, pp. 126–133.

[36] M. Kantarcıoğlu and C. Clifton, "Assuring privacy when big brother is watching," in *The 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'2003)*, San Diego, California, June 13 2003.

[37] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, "Disclosure limitation of sensitive rules," in *Knowledge and Data Engineering Exchange Workshop (KDEX'99)*, Chicago, Illinois, Nov. 8 1999, pp. 25–32. [Online]. Available: http://ieeexplore.ieee.org/iel5/6764/18077/00836532.pdf?isNumber=18077&%prod=CNF&arnumber=00836532

[38] Y. Saygin, V. S. Verykios, and C. Clifton, "Using unknowns to prevent discovery of association rules," *SIGMOD Record*, vol. 30, no. 4, pp. 45–54, Dec. 2001. [Online]. Available: http://www.acm.org/sigmod/record/issues/0112/SPECIAL/5.pdf

[39] T. Johnsten and V. Raghavan, "A methodology for hiding knowledge in databases," in *Proceedings of the IEEE ICDM Workshop on Privacy, Security and Data Mining*, ser. Conferences in Research and Practice in Information Technology, vol. 14. Maebashi City, Japan: Australian Computer Society, Dec. 9 2002, pp. 9–17. [Online]. Available: http://crpit.com/confpapers/CRPITV14Johnsten.pdf

[40] C. Clifton, "Using sample size to limit exposure to data mining," *Journal of Computer Security*, vol. 8, no. 4, pp. 281–307, Nov. 2000. [Online]. Available: http://iospress.metapress.com/openurl.asp?genre=article&issn=0926-227X&%volume=8&issue=4&spage=281

[41] W. Du and M. J. Atallah, "Privacy-preserving statistical analysis," in *Proceeding of the 17th Annual Computer Security Applications Conference*, New Orleans, Louisiana, USA, December 10-14 2001. [Online]. Available: http://www.cerias.purdue.edu/homes/duw/research/paper/acsac2001.ps