

Privacy, Security, and 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, Erik Herskedal, Murat Kantarcioglu, and Jaideep Vaidya for assistance in slide preparation



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



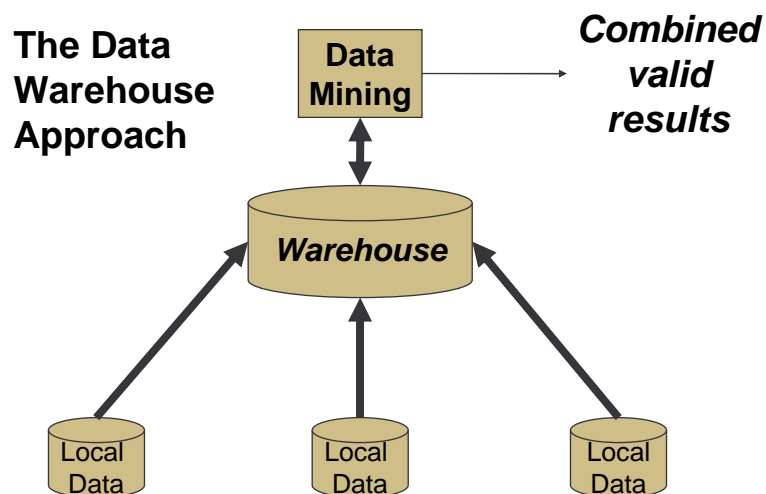
Privacy constraints don't prevent data mining

- 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!



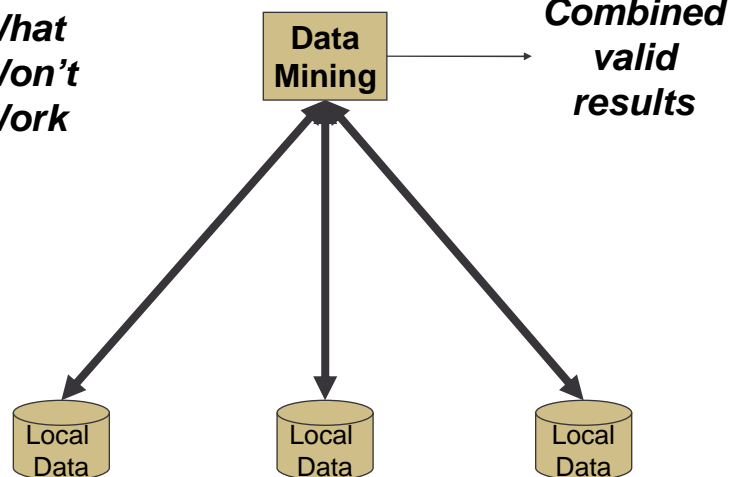
Distributed Data Mining: The “Standard” Method





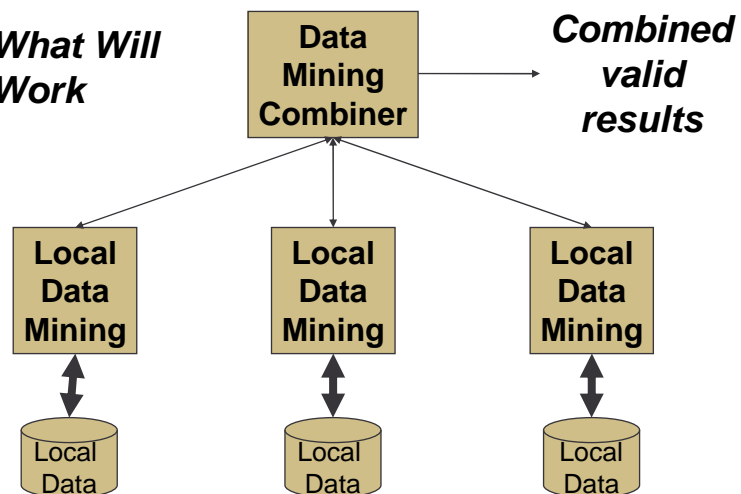
Private Distributed Mining: What is it?

**What
Won't
Work**



Private Distributed Mining: What is it?

**What Will
Work**



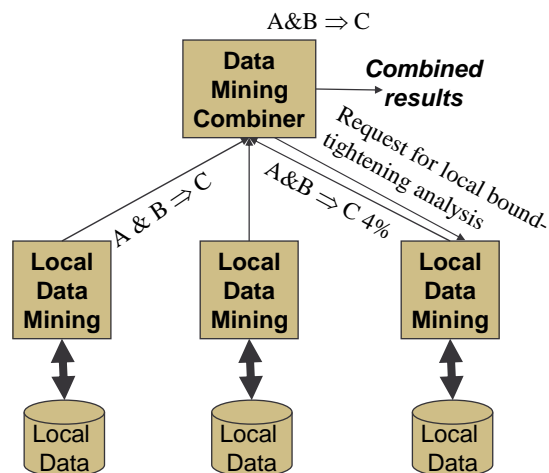


Example: *Association Rules*

- 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 \Rightarrow 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

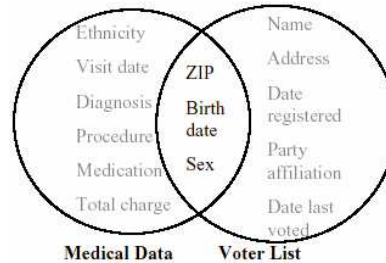
- Data that can't be traced to an individual not viewed as private
 - Remove “identifiers”
- But can we ensure it can't be traced?
 - Candidate Key in non-identifier information
 - Unique values for **some** individuals

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



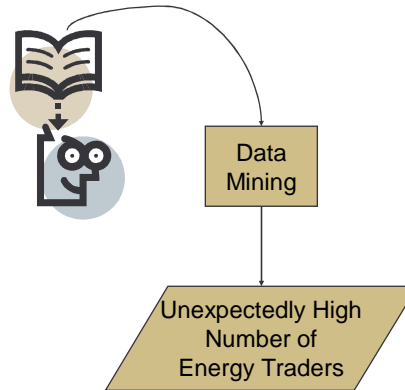
Collection Privacy

- 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



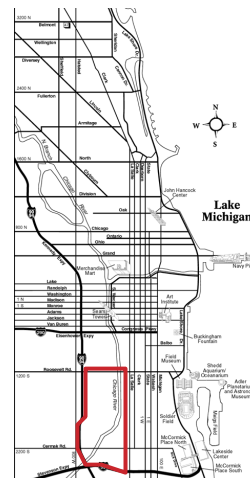
Collection Privacy Example: Corporate Phone Book

- 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 94/46/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 \rightarrow 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



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 $\mu = 0, \sigma$



Privacy Metric

If, from perturbed data, the original value x can be estimated to lie between $[x_1, x_2]$ with $c\%$ confidence, then the privacy at $c\%$ confidence level is related to $x_2 - x_1$

	Confidence		
	50%	95%	99.9%
Discretization	$0.5 \times W$	$0.95 \times W$	$0.999 \times W$
Uniform	$0.5 \times 2\alpha$	$0.95 \times 2\alpha$	$0.999 \times 2\alpha$
Gaussian	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$

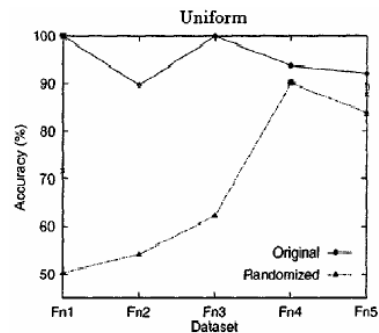
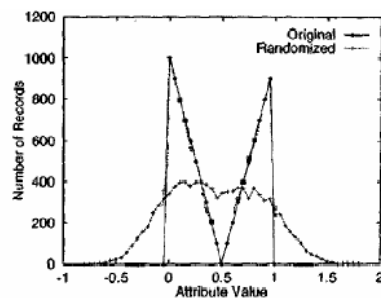
Example

- Salary 20K - 150K
- 95% Confidence
- 50% Privacy in Uniform
- $2\alpha = 0.5 \times 130K / 0.95$
 $= 68K$

- Issues
 - For very high privacy, discretization will lead to a poor model
 - Gaussian provides more privacy at higher confidence levels



Problem: Reconstructing the Original Distribution



- Selecting split points obvious in original distribution
- Hidden in randomized (perturbed) data



Original Distribution Reconstruction: Formal Definition

- x_1, x_2, \dots, x_n are the n original data values
 - Drawn from n iid random variables X_1, X_2, \dots, X_n similar to X
- Using *value distortion*,
 - The given values are $w_1 = x_1 + y_1, w_2 = x_2 + y_2, \dots, w_n = x_n + y_n$
 - y_i 's are from n iid random variables Y_1, Y_2, \dots, Y_n similar to Y
- Reconstruction Problem:
 - Given F_Y and w_i 's, estimate F_X



Original Distribution Reconstruction: Method

- The estimated density function:

$$f'_X(a) = \frac{1}{n} \sum_{i=1}^n \frac{f_Y(w_i - a) f_X(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(z) dz}$$

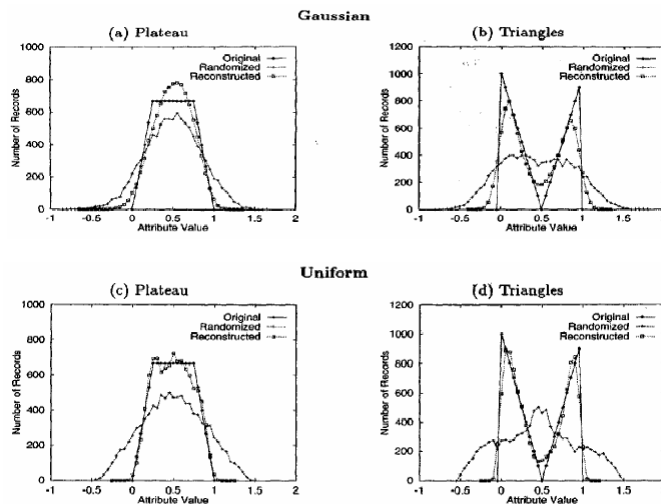
- Given a sufficiently large number of samples, it is expected to be close to the real density function
 - However, f_X is not known

- Iteratively,
$$f_X^{j+1}(a) = \frac{1}{n} \sum_{i=1}^n \frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z) dz}$$

- The initial estimate for f_X at $j=0$ is the uniform distribution



Original Distribution Reconstruction: Empirical Evaluation



Decision-Tree Classification Overview

- Recursively partition the data until each partition contains mostly examples from the same class
- Two phase development
 - Growth Phase (building the tree):
 - Evaluate each split for each attribute (gini, entropy, ...)
 - Use the best split to partition the data into two nodes
 - Repeat for each node, if points are not of the same class
 - Prune Phase
 - Generalize the tree
 - Remove statistical noise



Decision-Tree Classification Using the Perturbed Data

- **Split Points**
 - During reconstruction phase, attribute values are partitioned into intervals.
 - So, interval boundaries are the candidate split points.
- **Data Partitioning**
 - Reconstruction phase gives an estimate of the number of points in each interval.
 - Partition data into S_1 and S_2 where S_1 contains the lower valued data points.



Decision-Tree Classification Using the Perturbed Data (cont.)

- **When and How are the distributions reconstructed?**
 - **Global**
 - Reconstruct for each attribute once at the beginning
 - Build the decision tree using the reconstructed data
 - **ByClass**
 - First split the training data
 - Reconstruct for each class separately
 - Build the decision tree using the reconstructed data
 - **Local**
 - First split the training data
 - Reconstruct for each class separately
 - Reconstruct at each node while building the tree

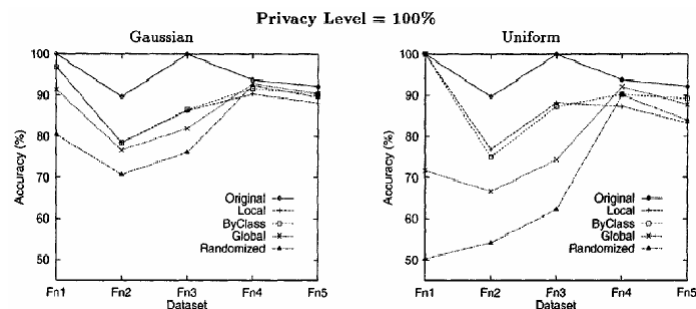


Experiments Methodology

- What to compare?
 - Classification accuracy of decision trees induced from
 - Global, ByClass, and Local algorithms vs.
 - Original and Randomized Distributions
- Synthetic Data
 - 100,000 training and 5,000 test records
 - All attributes are uniformly distributed ?
 - Equally split between two classes
 - Five classification functions (five datasets)
 - $(age < 40 \text{ and } 50K \leq salary \leq 100K)$ or $(40 \leq age < 60 \text{ and } 75K \leq salary \leq 125K)$ or $(age \geq 60 \text{ and } 25K \leq salary \leq 75K)$
- Issues
 - 95% confidence level
 - Privacy ranges from 25% to 200%
 - k (number of intervals) is selected, heuristically, such that
 - $10 < k < 100$ and average number of points in each is 100



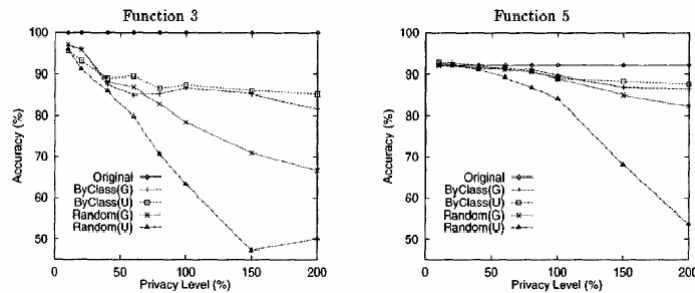
Experiments Results



- Global performs worse than ByClass and Local
- ByClass and Local have accuracy within 5% to 15% (absolute error) of the Original accuracy
- Overall, all are much better than the Randomized accuracy



Experiments Results (cont.)



- The effect of both Uniform and Gaussian are quite comparable after reconstruction
- The accuracy is fairly close to the Original



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 2α
 - confidence level 50% gives privacy α
 - confidence level 100% gives privacy 2α
- Incomplete in some situations



Quantification of privacy II

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

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

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 $f_B(x) = 2f_A(2x)$ then B offers half as much privacy as A
- Also: if a sequence of random variable A_n , $n=1, 2, \dots$ converges to random variable B, then privacy inherent in A_n should converge to the privacy inherent in B



Differential entropy

- Based on differential entropy $h(A)$:

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

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

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

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



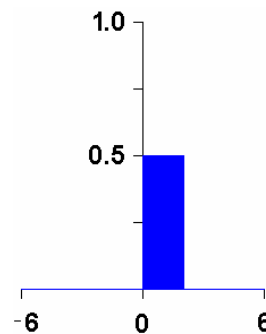
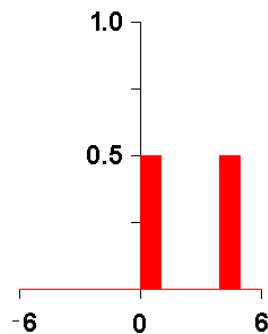
Privacy loss

- 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:
- $f_X(x) = 0.5, 0 \leq x \leq 1$
- $f_X(x) = 0.5, 4 \leq x \leq 5$
- $f_X(x) = 0, \text{ otherwise}$
- 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]

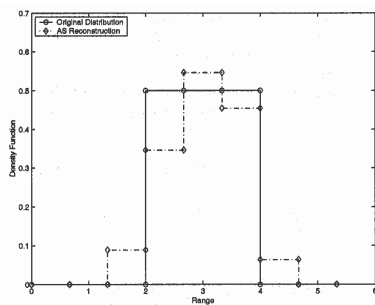


Figure 2: Reconstructed Uniform Distribution (AS Algorithm)

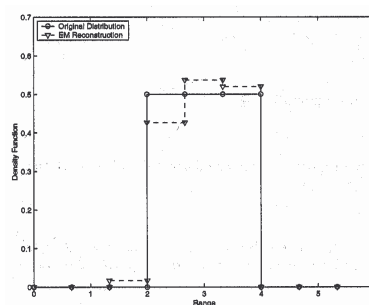


Figure 3: Reconstructed Uniform Distribution (EM Algorithm)



Gaussian distribution

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

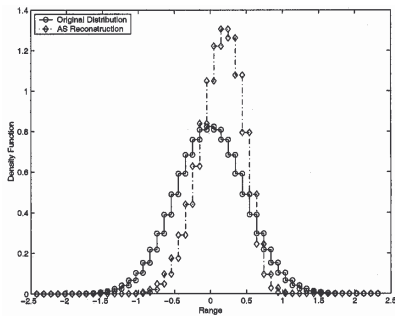


Figure 4: Reconstructed Gaussian Distribution (AS Algorithm)

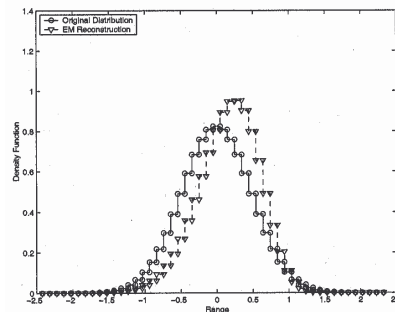


Figure 5: Reconstructed Gaussian Distribution (EM Algorithm)



Information loss / privacy loss

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

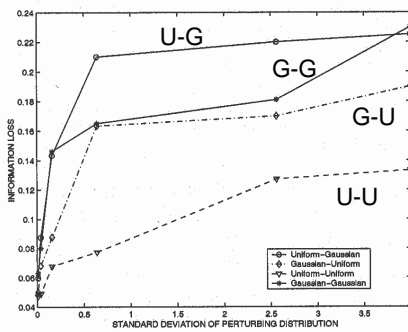


Figure 6: Information Loss with Standard Deviation of Perturbing Distribution

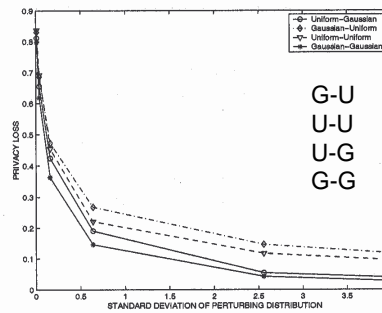


Figure 7: Privacy Loss with Standard Deviation of Perturbing Distribution

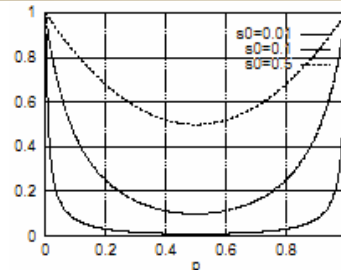


Association Rules

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: \exists polynomial time S such that $\{S(x, f(x, y))\} \equiv \{\text{View}(x, y)\}$
- Semi-Honest model: follow protocol, but remember intermediate exchanges
- Malicious: “cheat” to find something out

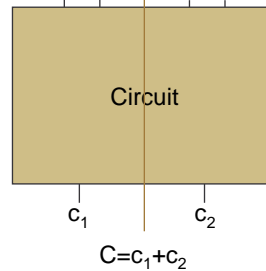


How does it work?

- 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

$$A = a_1 + a_2$$

$$B = b_1 + b_2$$



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



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 $\langle X_0, X_1 \rangle$ and the receiver has an input $\alpha \in \{0, 1\}$. At the end of the protocol the receiver should get X_α 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, \dots, a_m be the values of attribute A and let $T(a_1), \dots, T(a_m)$ 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, \dots, 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 $(|T_1(c_1)|, \dots, |T_1(c_L)|), (|T_2(c_1)|, \dots, |T_2(c_L)|)$
 - 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 $T(a_1), \dots, T(a_m)$ where a_1, \dots, 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, \dots, a_m ,
 C have l possible values c_1, \dots, 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:

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



X In X

- Taylor Series of natural logarithm:

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

- 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

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



Comparison

- Fully Generic solution $|R| \cdot |T| \cdot \log m$ 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 - $O(k \log |T| \cdot |S|)$ 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

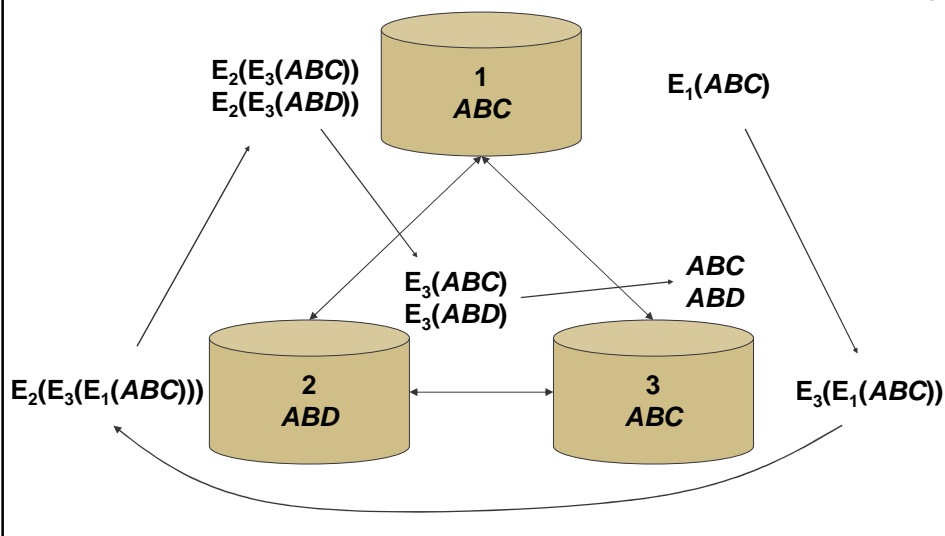
- Key: Commutative Encryption ($E_a(E_b(x)) = E_b(E_a(x))$)
 - Compute local candidate set
 - Encrypt and send to next site
 - Continue until all sites have encrypted all rules
 - Eliminate duplicates
 - Commutative encryption ensures if rules the same, encrypted rules the same, regardless of order
 - 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

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

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

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

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

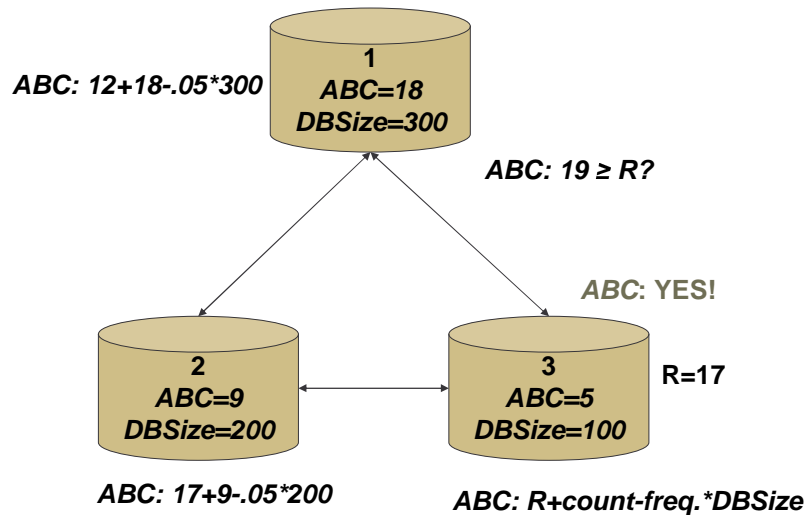


Which Candidates Are Globally Supported? (Continued)

- Now securely compute $\text{Sum} \geq 0$:
 - Site₀ generates random R
Sends $R + \text{count}_0 - \text{frequency} * \text{dbsize}_0$ to site₁
 - Site_k adds $\text{count}_k - \text{frequency} * \text{dbsize}_k$, sends to site_{k+1}
- Final result: Is sum at site_n - $R \geq 0$?
 - Use Secure Two-Party Computation
- This protocol is secure in the semi-honest model



Computing Frequent: Is $ABC \geq 5\%$?



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, $A_1 \dots A_p$
- B has q attributes, $B_1 \dots B_q$
- 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. $L_1 = \{\text{large 1-itemsets}\}$
2. for ($k=2$; $L_{k-1} \neq \emptyset$; $k++$) do begin
3. $C_k = \text{apriori-gen}(L_{k-1})$;
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 $c.count = \vec{X} \cdot \vec{Y} = \sum_{i=1}^n x_i * y_i$
11. endif
12. $L_k = L_k \cup c | c.count \geq \text{minsup}$
13. end
14. end
15. Answer = $\cup_k L_k$



Protocol

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

$$\langle x_1 + a_{1,1} * R_1 + a_{1,2} * R_2 + \dots + a_{1,n/2} * R_{n/2} \rangle$$

$$\langle x_2 + a_{2,1} * R_1 + a_{2,2} * R_2 + \dots + a_{2,n/2} * R_{n/2} \rangle$$

⋮

$$\langle x_n + a_{n,1} * R_1 + a_{n,2} * R_2 + \dots + a_{n,n/2} * R_{n/2} \rangle$$

- 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.

$S =$

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

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



Protocol (cont)

$S =$

$$x_1 * y_1 + x_2 * y_2 + \dots + x_n * y_n \rightarrow \sum_{i=1}^n x_i * y_i$$

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

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

$$\vdots$$

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

Grouping
components
vertically
and
factoring out
 R_i



Protocol (complete)

$S =$

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

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

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

\vdots

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

- A already knows $R_1 \dots R_{n/2}$
- 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



EM Clustering (Lin & Clifton '02)

- 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:

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



EM Clustering: Process

- Estimate μ_i , π , and σ^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!



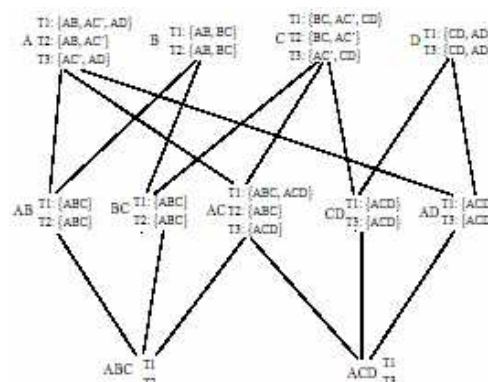
Disclosure Limitation of Sensitive Rules (*Atallah et. al. '99*)

- 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

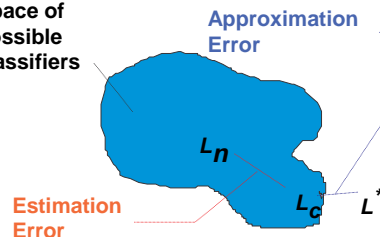
- Build graph of all supported itemsets
- To hide large itemset:
 - Go up tree to find item with lowest support
 - Select transaction affecting fewest 2-itemsets
 - Remove item from that transaction





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

Space of possible classifiers



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 $\epsilon < 1/2$

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

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 $L \in (0, \frac{1}{2})$

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

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

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

and also, for $\varepsilon \leq L\varepsilon^{\frac{1}{4}}$,

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

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
 - When the secrets are in the results and we want to share the data
- *Still lots of work to do!*



ECML/PKDD-2002

References

- [1] Latanya Sweeney. *Computational Disclosure Control: A Primer on Data Privacy Protection*. PhD thesis, Massachusetts Institute of Technology, 2001.
- [2] Directive 94/46/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*, No I.(281):31–50, October 24 1995. URL http://europa.eu.int/comm/internal_market/en/dataprot/law/.
- [3] Standard for privacy of individually identifiable health information. *Federal Register*, 66(40), February 28 2001. URL <http://www.hhs.gov/ocr/hipaa/finalreg.html>.
- [4] The platform for privacy preferences 1.0 (p3p1.0) specification, April 16 2002. URL <http://www.w3.org/TR/P3P/>.
- [5] Privacy bird, July 2002. URL <http://www.privacybird.com>.
- [6] Richard A. Moore, Jr. Controlled data-swapping techniques for masking public use microdata sets. Statistical Research Division Report Series RR 96-04, U.S. Bureau of the Census, Washington, DC., 1996. URL <http://www.census.gov/srd/papers/pdf/rr96-4.pdf>.
- [7] Dorothy E. Denning. Secure statistical databases with random sample queries. *ACM Transactions on Database Systems*, 5(3):291–315, September 1980. URL <http://doi.acm.org/10.1145/320613.320616>.
- [8] David Dobkin, Anita K. Jones, and Richard J. Lipton. Secure databases: Protection against user influence. *ACM Transactions on Database Systems*, 4(1):97–106, March 1979. URL <http://doi.acm.org/10.1145/320064.320068>.
- [9] Nabil R. Adam and John C. Wortmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys*, 21(4):515–556, December 1989. URL <http://doi.acm.org/10.1145/76894.76895>.
- [10] Michael A. Palley and Jeffrey S. Simonoff. The use of regression methodology for the compromise of confidential information in statistical databases. *ACM Transactions on Database Systems*, 12(4):593–608, December 1987. URL <http://doi.acm.org/10.1145/32204.42174>.
- [11] CRISP-DM. Cross industry standard process for data mining. <http://www.crisp-dm.org>, December 1999. URL <http://www.crisp-dm.org>.
- [12] Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD Conference on Management of Data*, pages 439–450, Dallas, TX, May 14-19 2000. ACM. URL <http://doi.acm.org/10.1145/342009.335438>.
- [13] Dakshi Agrawal and Charu 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*, pages 247–255, Santa Barbara, California, USA, May 21-23 2001. ACM. URL <http://doi.acm.org/10.1145/375551.375602>.
- [14] Shariq J. Rizvi and Jayant R. Haritsa. Maintaining data privacy in association rule mining. In *Proceedings of 28th International Conference on Very Large Data Bases. VLDB*, August 20-23 2002. URL <http://www.vldb.org>.
- [15] Andrew C. Yao. How to generate and exchange secrets. In *Proceedings of the 27th IEEE Symposium on Foundations of Computer Science*, pages 162–167. IEEE, 1986.
- [16] 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*, pages 218–229, 1987. URL <http://doi.acm.org/10.1145/28395.28420>.
- [17] Oded Goldreich. Secure multi-party computation, September 1998. URL <http://www.wisdom.weizmann.ac.il/~oded/pp.html>. (working draft).
- [18] Yehuda Lindell and Benny Pinkas. Privacy preserving data mining. In *Advances in Cryptology – CRYPTO 2000*, pages 36–54. Springer-Verlag, August 20-24 2000. URL <http://link.springer.de/link/service/series/0558/bibs/1880/18800036.htm>.
- [19] Murat Kantarcioglu and Chris 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)*, pages 24–31, June 2 2002. URL <http://www.bell-labs.com/user/minos/DMKD02/Papers/kantarcioglu.pdf>.
- [20] Jaideep Shrikant Vaidya and Chris Clifton. Privacy preserving association rule mining in vertically partitioned data. In *The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 23-26 2002.
- [21] Xiaodong Lin and Chris Clifton. Distributed EM clustering without sharing local information. *Journal of Information Science*, February 2003. Submitted to Special Issue on Knowledge Discovery from Distributed Information Sources.
- [22] 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)*, pages 25–32, Chicago, Illinois, November 8 1999. URL <http://ieeexplore.ieee.org/iel5/6764/18077/00836532.pdf?isNumber=18077&prod=CNF&arnumber=00836532prod=CNF&arnumber=00836532>.
- [23] Yücel Saygin, Vassilios S. Verykios, and Chris Clifton. Using unknowns to prevent discovery of association rules. *SIGMOD Record*, 30(4):45–54, December 2001. URL <http://www.acm.org/sigmod/record/issues/0112/SPECIAL/5.pdf>.
- [24] Chris Clifton. Using sample size to limit exposure to data mining. *Journal of Computer Security*, 8(4):281–307, November 2000. URL <http://iospress.metapress.com/openurl.asp?genre=article&issn=0926-227X&volume=8&issue=4&spage=281>.