Randomization based Privacy Preserving Data Mining

Xintao Wu
Department of Computer Science
University of North Carolina at Charlotte

ECML/PKDD06, Berlin

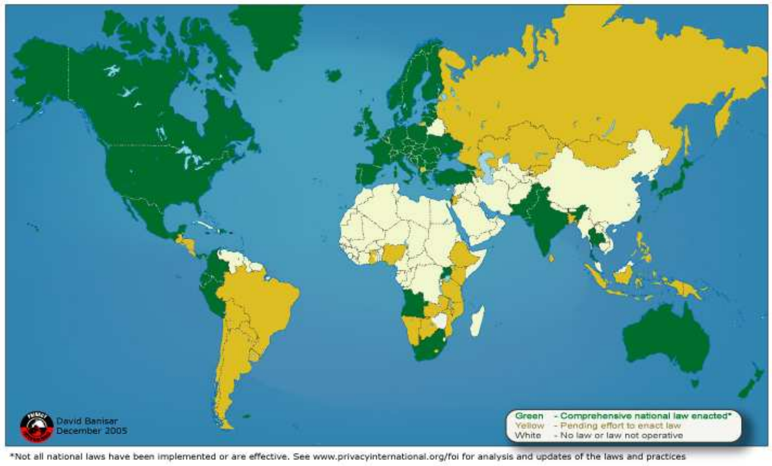
Privacy Case

- **Nydia Velázquez** (1982)

Three weeks after Nydia Velázquez won the New York Democratic Party's nomination to serve in the U.S. House of Representatives, somebody at St. Claire Hospital in New York faxed Velázquez's medical records to the New York Post. The records detailed the care that Velázquez had received at the hospital after a suicide attempt--an attempt that had happened several years before the election.

Database Nation: The Death of Privacy in the 21st Century, Simson Garfinkel, Jan 2000, 1-56592-653-6

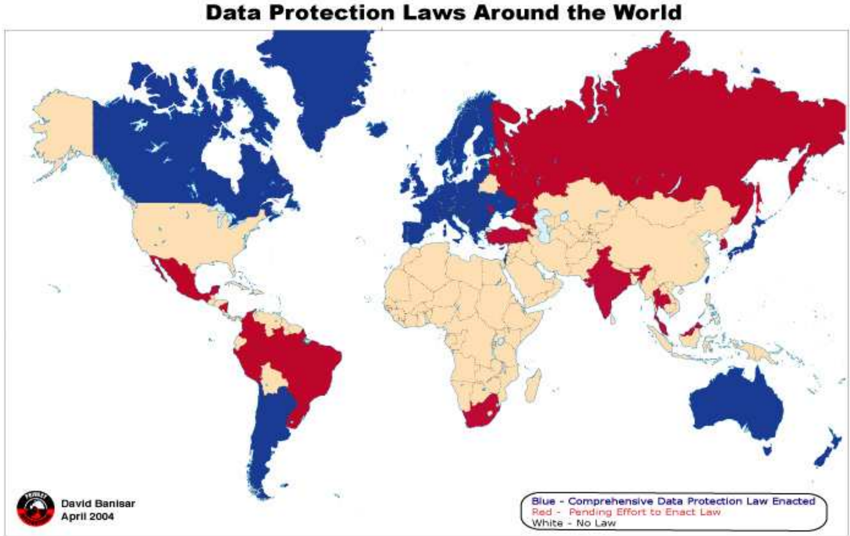
National Freedom of Information Laws 2005



*Not all national laws have been implemented or are effective. See www.privacyinternational.org/foi for analysis and updates of the laws and practices

Source: http://www.privacyinternational.org/issues/foia/foia-laws.jpg

Data Protection Laws Around the World



Source: http://www.privacyinternational.org/survey/dpmap.jpg

National Laws

- USA
  - HIPAA for health care
    - Passed August 21, 96
    - lowest bar and the States are welcome to enact more stringent rules
      - California State Bill 1386
  - Grann-Leach-Bliley Act of 1999 for financial institutions
  - COPPA for childern's online privacy
  - etc.
- Canada
  - PIPEDA 2000
    - Personal Information Protection and Electronic Documents Act
    - Effective from Jan 2004
- European Union (Directive 94/46/EC)
  - Passed by European Parliament Oct 95 and Effective from Oct 98.
  - Provides guidelines for member state legislation
  - Forbids sharing data with states that do not protect privacy

Mining vs. Privacy

- Data mining
  - The goal of data mining is summary results (e.g., classification, cluster, association rules etc.) from the data (distribution)
- Individual Privacy
  - Individual values in database must not be disclosed, or at least no close estimation can be got by attackers
  - Contractual limitations: privacy policies, corporate agreements
- Privacy Preserving Data Mining (PPDM)
  - How to transform data such that
    - we can build a good data mining model (data utility)
    - while preserving privacy at the record level (privacy)?

Two Approaches

- Distributed
  - Suitable for multi-party platforms
  - Secure multi-party computation
  - Tolerated disclosure: computationally private
- See some other excellent tutorials

- Randomization
  - Perturb data to protect privacy of individual records.
  - Preserve intrinsic distributions necessary for modeling.
  - Tolerated disclosure: statistically private
- Our focus

Other Tutorials on PPDM

- *Privacy in data system,* Rakesh Agrawal, PODS03
- *Privacy preserving data mining*, Chris Clifton, PKDD02, KDD03
- *Preserving privacy in database systems*, Johann-Chrostoph Freytag, WAIM06
- *Models and methods for privacy preserving data publishing and analysis*, Johannes Gehrke, ICDM05, ICDE06, KDD06
- *Cryptographic techniques in privacy preserving data mining*, Helger Lipmaa, PKDD06

Scope

	ssn	name	zip	race	…	age	Sex	Bal	income	…	IntP
1			28223	Asian	…	20	M	10k	85k	…	2k
2			28223	Asian	…	30	F	15k	70k	…	18k
3			28262	Black	…	20	M	50k	120k	…	35k
4			28261	White	…	26	M	45k	23k	…	134k
.			.	.	…	.	.	.	.	…	.
N			28223	Asian	…	20	M	80k	110k	…	15k

Part II:
focus on Random Response
.

Part I: focus

Outline (Part I)

Randomization based PPDM

- ***Additive noise***
- Rotation
- General Linear Transformation
- *Condensation or modeling based*

Attacking Method

- Additive noise
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
- Rotation and General Linear Transformation
  - ICA
  - A-priori Knowledge Based Attack

Additive Noise Randomization Example

	Bal	income	…	IntP
1	10k	85k	…	2k
2	15k	70k	…	18k
3	50k	120k	…	35k
4	45k	23k	…	134k
.	.	.	…	.
N	80k	110k	…	15k

$$\begin{pmatrix} 17.334 & 88.759 & 2.099 \\ 19.199 & 77.537 & 25.939 \\ 59.199 & 128.447 & 38.678 \\ 51.208 & 30.313 & 135.939 \\ 89.048 & 115.692 & 21.318 \end{pmatrix} = \begin{pmatrix} 10 & 85 & 2 \\ 15 & 70 & 18 \\ 50 & 120 & 35 \\ 45 & 23 & 134 \\ 80 & 110 & 15 \end{pmatrix} + \begin{pmatrix} 7.334 & 3.759 & 0.099 \\ 4.199 & 7.537 & 7.939 \\ 9.199 & 8.447 & 3.678 \\ 6.208 & 7.313 & 1.939 \\ 9.048 & 5.692 & 6.318 \end{pmatrix}$$

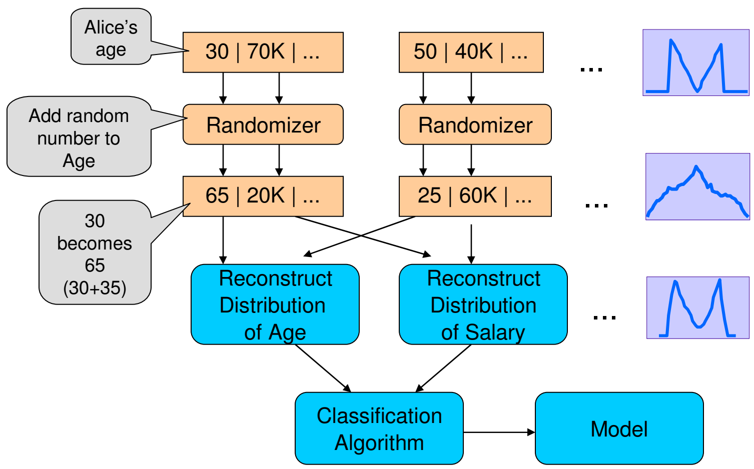
$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$$

Assumption of Additive Noise

- In this tutorial, we assume the additive noise E is independent with the original data X.
- If E is correlated with X, it may significantly affect data utility although it may better preserve privacy.
  - See Huang, Du and Chen SIGMOD05

Additive Randomization (*Z=X+Y*)

- *R.Agrawal and R.Srikant SIGMOD 00*



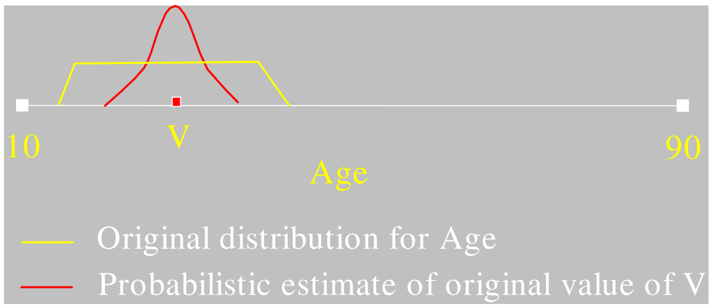
Reconstruction Problem

- 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.

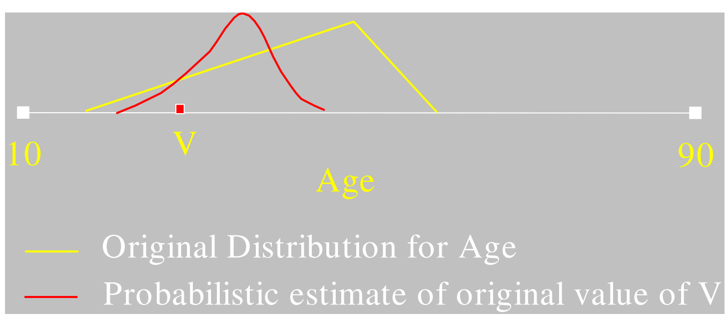
Intuition (Reconstruct single point)

- Use Bayes' rule for density functions



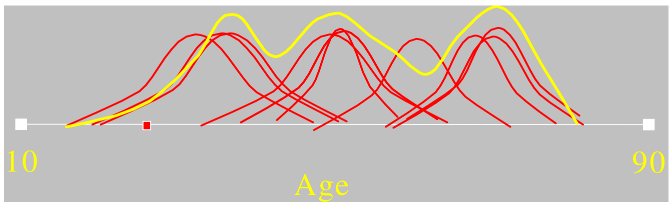
Intuition (Reconstruct single point)

- Use Bayes' rule for density functions



Reconstruct the Distribution

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



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

Distribution Reconstruction Alg.

- Bootstrapping Algorithm

$f_X^0 =$ uniform distribution

$j = 0$

repeat

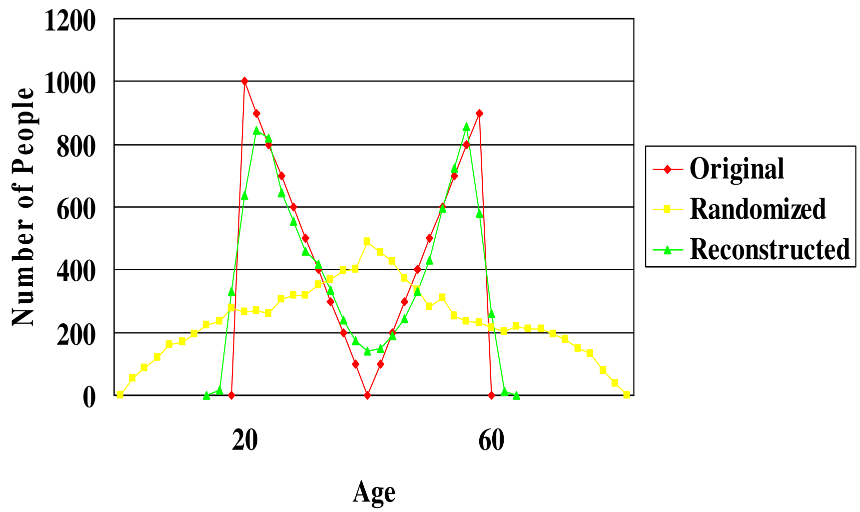
$$f_X^{j+1}(a) = \frac{1}{n}\sum_{i=1}^{n}\frac{f_Y((x_i+y_i)-a)f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i+y_i)-a)f_X^j(a)}$$

$j = j+1$

until (stopping criterion met)

- Converges to maximum likelihood estimate
  - Agrawal and Aggarwal PODS 01
- Extension to muti-variate case
  - Domingo-Ferrer et al. PSD04

Works well



ECML/PKDD06 Tutorial

19

UNC CHARLOTTE

More Example

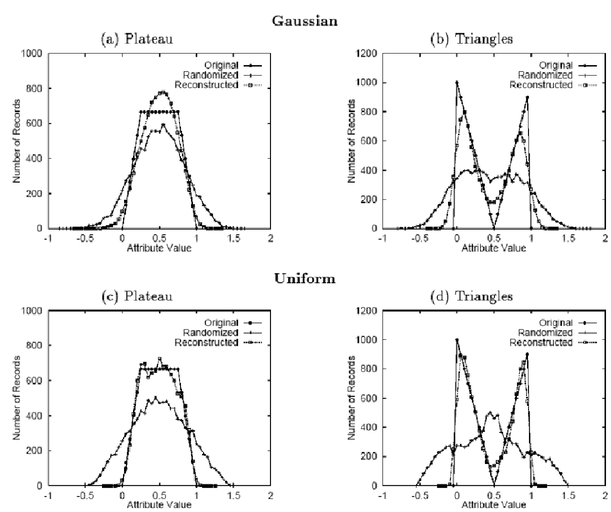


Figure 2: Reconstructing the Original Distribution

ECML/PKDD06 Tutorial

20

UNC CHARLOTTE

Why privacy is preserved?

- Cannot reconstruct individual values accurately.
- Can only reconstruct distributions.
- Noise $\uparrow \Rightarrow$ quality of the distribution reconstruction $\downarrow$

Individual value reconstruction methods EXIST

Simple Multiplicative Noise

$$y_{ij} = x_{ij} e_{ij} \qquad i = 1, \cdots, N \qquad j = 1, \cdots, p$$

- x_{ij} the value for the i-th individual's j-th attribute.
- e_{ij} denotes the noise where all e_{ij}'s for a given j follow the same distribution $N(\mu_j, \sigma_j^2)$
- It turns out to be additive noise by taking logarithms on both sides.

More details, See Kim and Winkler, 2003

Outline

Part I: Randomization based PPDM

- Additive noise
- ***Rotation***
- General Linear Transformation

Part II: Attacking Method

- Additive noise
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
- Rotation and General Linear Transformation
  - ICA
  - A-priori Knowledge Based Attack

Rotation Randomization Example

	Bal	income	…	IntP
1	10k	85k	…	2k
2	15k	70k	…	18k
3	50k	120k	…	35k
4	45k	23k	…	134k
.	.	.	…	.
N	80k	110k	…	15k

$$\begin{bmatrix} 61.33 & 63.67 & 110.00 & 119.67 & 63.33 \\ 49.33 & 30.67 & 55.00 & -59.33 & -31.67 \\ -33.67 & -21.33 & -30.00 & 51.67 & -51.67 \end{bmatrix} = \begin{bmatrix} 0.3333 & 0.6667 & 0.6667 \\ -0.6667 & 0.6667 & -0.3333 \\ -0.6667 & -0.3333 & 0.6667 \end{bmatrix} \begin{bmatrix} 10 & 15 & 50 & 45 & 80 \\ 85 & 70 & 120 & 23 & 110 \\ 2 & 18 & 35 & 134 & 15 \end{bmatrix}$$

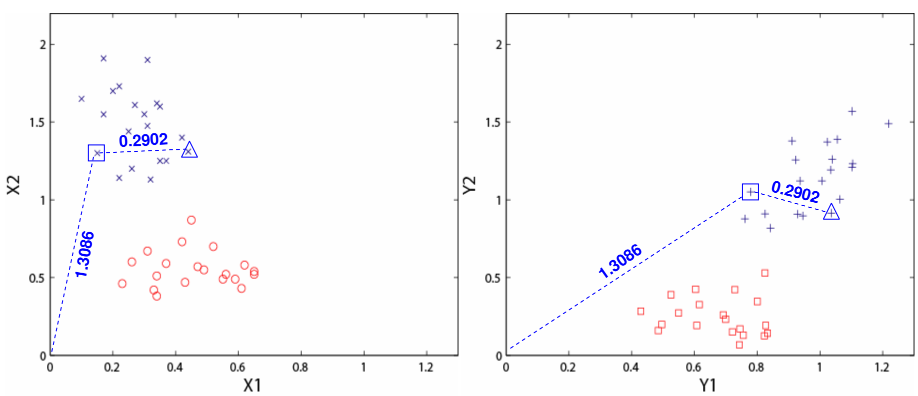
$$\boldsymbol{Y} = \boldsymbol{R} \quad \boldsymbol{X}$$

$$RR^T = R^T R = I$$

Why R is orthonormal?

- When R is an orthonormal matrix ($R^T R = RR^T = I$)
  - Vector length: $|Rx| = |x|$
  - Euclidean distance: $|Rx_i - Rx_j| = |x_i - x_j|$
  - Inner product : $\langle Rx_i, Rx_j \rangle = \langle x_i, x_j \rangle$
- Many clustering and classification methods are invariant to this rotation perturbation.
  - Classification, Chen and Liu, ICDM 05
  - Distributed data mining, Liu and Kargupta, TKDE 06

Rotation Example



$$Y = RX$$

$$R = \begin{pmatrix} 0.866 & 0.500 \\ -0.500 & 0.866 \end{pmatrix}$$

Rotation-Invariant Classifiers

- The classifier trained with the rotation perturbed dataset delivers the same accuracy as that trained with the the original dataset
- Examples
  - KNN, Kernel methods (distance)
  - SVM classifiers with three popular kernels
    - Polynomial kernel & neural network kernel (inner product)
    - Radial basis kernel (distance)
  - Hyperplane-based classifiers (hyperplane)

Chen and Liu, ICDM 05

Is *Y=RX* Secure?

- Can we get *X* from *Y=RX* when only *Y* is available?
  - It seems Independent Component Analysis can help.
  - $X = AS$ ICA noise-free Model

Outline

Part I: Randomization based PPDM

- Additive noise
- Rotation
- ***General Linear Transformation***

Part II: Attacking Method

- Additive noise
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
  - Challenging problems
- Rotation and General Linear Transformation
  - ICA
  - A-priori Knowledge Based Attack

UNC CHARLOTTE

Linear Transformation Example

$$\begin{pmatrix} 265.95 & 286.63 & 475.68 & 581.71 & 520.53 \\ 394.30 & 338.49 & 569.58 & 174.22 & 277.79 \\ 362.55 & 394.11 & 665.37 & 776.46 & 463.08 \end{pmatrix} =$$

$$\begin{pmatrix} 4.751 & 2.429 & 2.282 \\ 1.156 & 4.457 & 0.093 \\ 3.034 & 3.811 & 4.107 \end{pmatrix} \begin{pmatrix} 10 & 15 & 50 & 45 & 80 \\ 85 & 70 & 120 & 23 & 110 \\ 2 & 18 & 35 & 134 & 15 \end{pmatrix} + \begin{pmatrix} 7.334 & 4.199 & 9.199 & 6.208 & 9.048 \\ 3.759 & 7.537 & 8.447 & 7.313 & 5.692 \\ 0.099 & 7.939 & 3.678 & 1.939 & 6.318 \end{pmatrix}$$

$$Y = R \quad X + E$$

R can be any random matrix

UNC CHARLOTTE

General Linear Transformation

- $Y = R\,X + E$
  - When $R = I$: $Y = X + E$ (Additive Noise Model)
  - When $RR^T = R^TR = I$ and $E = 0$: $Y = RX$ (Rotation Model)
  - R can be an arbitrary rotation matrix

- Is $Y=RX+E$ more secure?
  - Will be discussed in part II

ECML/PKDD06 Tutorial 31 UNC CHARLOTTE

Condensation/modeling

- Idea
  - Build some statistical model from the original data
  - Apply model built to generate data

- Skipped due to time constraints

ECML/PKDD06 Tutorial 32 UNC CHARLOTTE

Outline

Part I: Randomization based PPDM

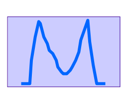
- Additive noise
- Rotation
- General Linear Transformation

Part II: Attacking Methods

- ***Additive noise***
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
- Rotation and General Linear Transformation
  - ICA
  - A-priori Knowledge Based Attack

Motivation

- The goal of additive randomization-based perturbation
  - To hide the sensitive data by randomly modifying the data values using some additive noise
  - To keep the aggregate characteristics or distribution remain unchanged or recoverable
- Do those aggregate characteristics or distribution contain confidential information which may be exploited by snoopers to derive individual's sensitive data?



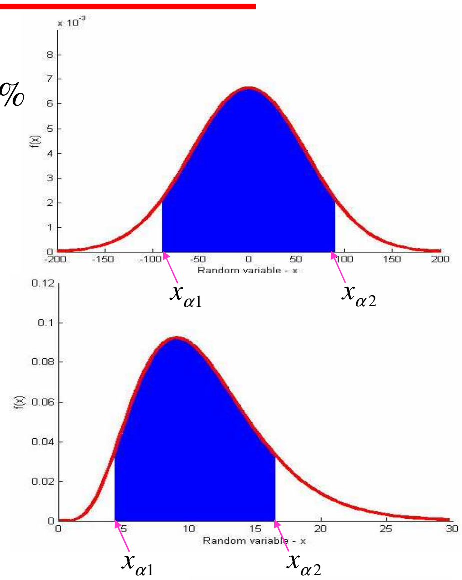
private information

More details, See Guo, Wu and Li, PDM 2006

Inter-Quantile Range (IQR)

- Inter-Quantile Range $[x_{\alpha 1}, x_{\alpha 2}]$ is defined as $P(x_{\alpha 1} \le x \le x_{\alpha 2}) \ge c\%$ while $c = \alpha 2 - \alpha 1$ denotes the confidence.

- IQR measures the amount of spread and variability of the variable. Hence it can be used by attackers to estimate the range of each individual value.



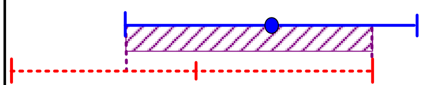
Comparison with other Privacy definitions

- Interval privacy (Agrawal and Srikant, SIGMOD00)
  - If the original value can be estimated with *c%* confidence to lie in the interval *[a, b]*, then the interval width *(b-a)* defines the amount of privacy at *c%* confidence level

- Mutual Information (Aggarwal and Agrawal, PODS01)
- Reconstruction privacy (Rizvi & Haritsa, VLDB02)
- α-to-β privacy breach (Evfimievski et al. PODS03)

Disclosure Measure

Individual's privacy interval
$[u_i^l, u_i^u]$



Attacker's estimated range
$[x_{(1-c)/2}, x_{(1+c)/2}]$

Measure Similarity

$$d_i = \frac{[u_i^l, u_i^u] \bigcap [x_{(1-c)/2}, x_{(1+c)/2}]}{[u_i^l, u_i^u] \bigcup [x_{(1-c)/2}, x_{(1+c)/2}]}$$

$$D = 1/n \sum_{i=1}^{n} d_i$$

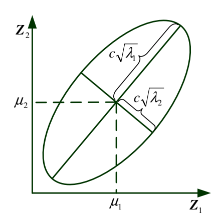
Complete disclosed point if its estimated range

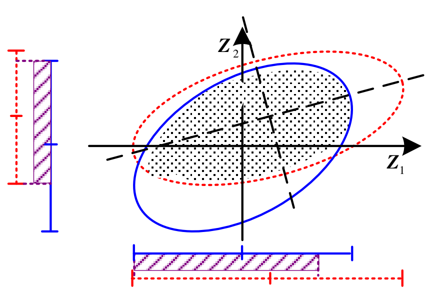
- contains the original value
- fully falls within the pre-specified privacy interval

Extend to Multivariate Cases

- In practice, the distribution of multiple numerical attributes are often modeled by one multivariate normal distribution, $N(\mu, \Sigma)$
- The ellipsoid $\{z : (z-\mu)^T \Sigma^{-1} (z-\mu) \le \chi_p^2(a)\}$ contains a fixed percentage, $(1-a)100\%$ of data values.
- The projection of this ellipsoid on axis z_i has bound:

$$[\mu_i - \sqrt{\chi_p^2(\alpha)\sigma_{ii}}, \mu_i + \sqrt{\chi_p^2(\alpha)\sigma_{ii}}]$$





Outline(Part I)

Randomization based PPDM

- Additive noise
- Rotation
- General Linear Transformation

Attacking Methods

- Additive noise
  - IQR (from distribution)
  - ***Spectral Filtering, PCA, SVD***
- Rotation and General Linear Transformation
  - ICA
  - AK-ICA

Individual Value Reconstruction (Additive Noise)

- Methods
  - Spectral Filtering, Kargupta et al. ICDM03
  - PCA, Huang, Du, and Chen SIGMOD05
  - SVD, Guo, Wu and Li, PKDD06
- All aim to remove noise by projecting on lower dimensional space.
  - PCA is similar to SF except SF focus more on small data sets
  - PCA is equivalent to SVD in some sense
- Let us focus on how PCA works first

Preliminary

- F-norm and 2-norm

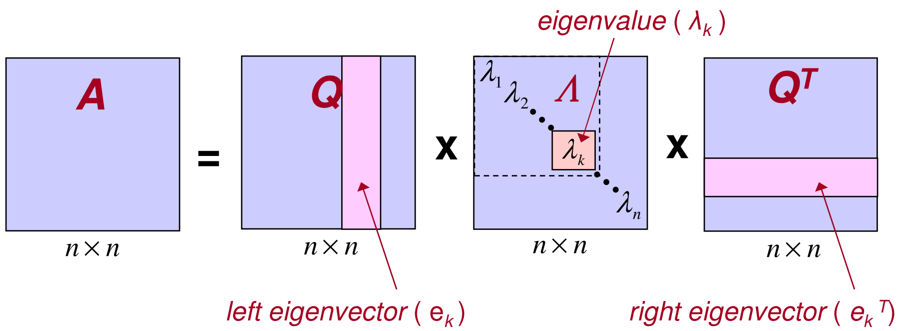
$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2} \qquad \|A\|_2 = \underbrace{\max}_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

- Some properties
  - $\|AB\|_F \leq \|A\|_F \|B\|_F$ and $\|AB\|_2 \leq \|A\|_2 \|B\|_2$
  - $\|A\|_2 \leq \|A\|_F \leq \sqrt{n}\,\|A\|_2$
  - $\|A\|_2 \leq \sqrt{\lambda_{\max}(A^T A)}$,the square root of the largest eigenvalue of $A^T A$
  - If A is symmetric, then $\|A\|_2 \leq \lambda_{\max}(A)$,the largest eigenvalue of A

PCA Revisited

- Idea
  - Given data points in n-dimensional space, project into lower *k*-dimensional space while preserving as much information as possible
  - In particular, choose projection that minimizes the squared error in reconstructing original data

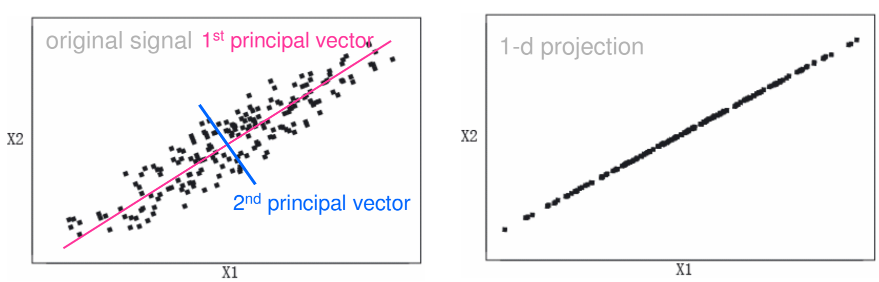
EVD



$$A = Q \Lambda Q^T$$

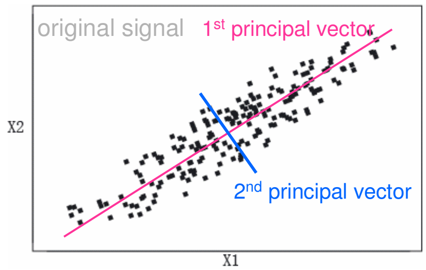
$$= \sum_{k=1}^{n} \lambda_k e_k e_k^T$$

PCA Example

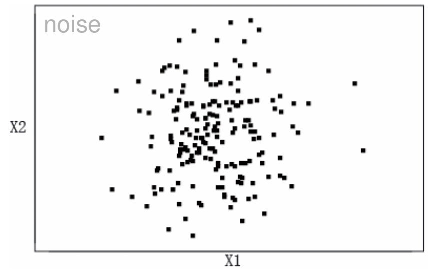


Data vs. Noise

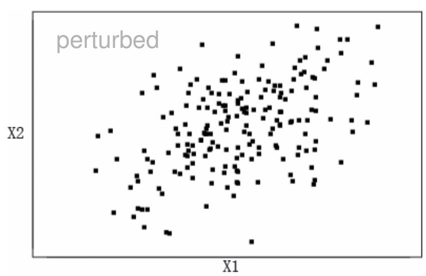
Original data are correlated

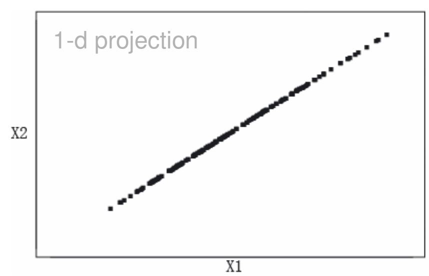


Noise are not correlated

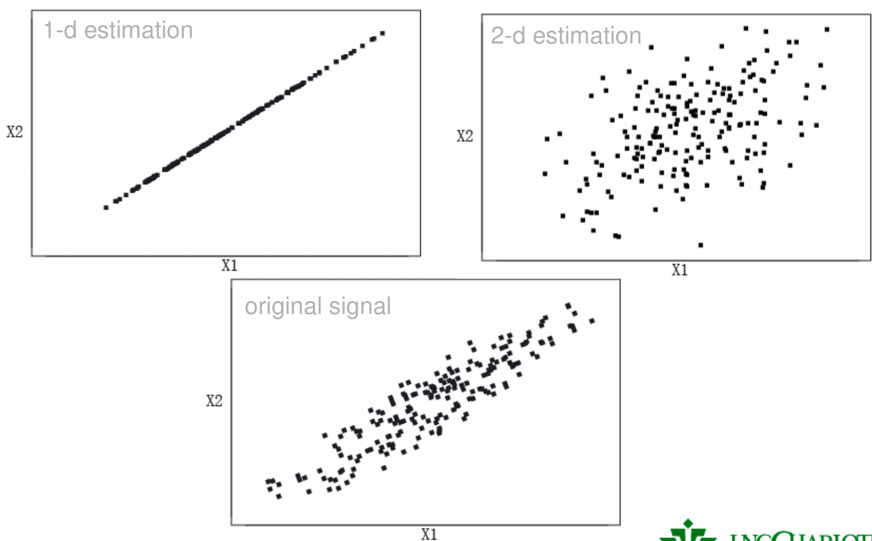


PCA on Randomized Data





1-d vs. 2-d Estimation



Why it works?

Original Data

- Correlated
- When we remove the 2nd component, the actual information loss is smaller

Noise

- Uncorrelated
- Variance evenly distributed in the space
- When remove the 2nd component, 50% noise is removed

PCA Reconstruction Alg. (Up=U+V)

- Applying PCA on covariance matrices of *Up* and *V*

$$U_p^T U_p = Q_p \Lambda_p Q_p^T \qquad V^T V = Q_v \Lambda_v Q_v^T$$

- Determining the first *k* components Q_{p_k} based on

$$\Lambda_p \approx \Lambda_u + \Lambda_v$$

- Reconstructing the data:

$$\hat{U} = U_p P_\chi = U_p Q_{p_k} Q_{p_k}^T$$

Two Problems

- Problem 1
  - How to determine *k*?
- Problem 2
  - How accurate can we achieve?

P1: Determining k

- Strategy 1:
  - $k = \max\{ i \mid \tilde{\lambda}_i \geq \lambda_V \}$
- Strategy 2:
  - $k = \min\{ i \mid \tilde{\lambda}_i < 2\lambda_V \} - 1$
  - The estimated data using $\hat{U} = \tilde{U} P_{\tilde{\chi}} = \tilde{U} \tilde{Q}_k \tilde{Q}_k^T$ is approximate optimal

Why strategy 2 is better?

- Strategy 2:
  - The estimated data using $\hat{U} = \tilde{U} P_{\tilde{\chi}} = \tilde{U} \tilde{Q}_k \tilde{Q}_k^T$ is approximate optimal when $k = \min\{ i \mid \tilde{\lambda}_i < 2\lambda_V \} - 1$

$$f(k) = \hat{U} - U$$

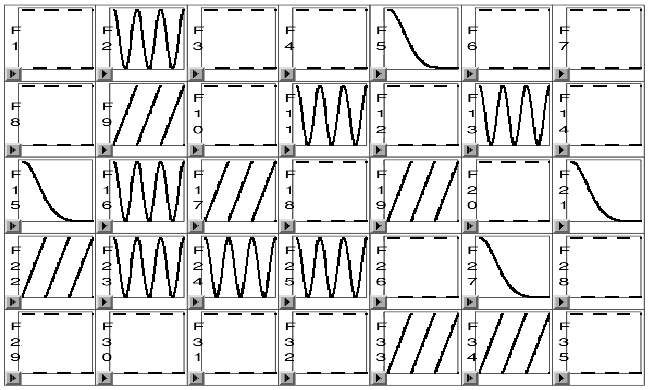
$$f(k+1) - f(k) = V\tilde{e}_{k+1}\tilde{e}_{k+1}^T + U\tilde{e}_{k+1}\tilde{e}_{k+1}^T$$

$$\| V\tilde{e}_{k+1}\tilde{e}_{k+1}^T \|_F^2 \approx \lambda_V \qquad \| U\tilde{e}_{k+1}\tilde{e}_{k+1}^T \|_F^2 \approx \lambda_{k+1}$$

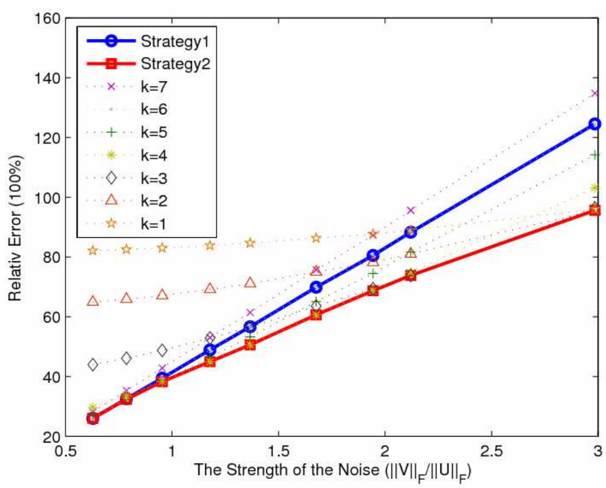
More details, See Guo, Wu and Li, PKDD 2006

Experimental Results

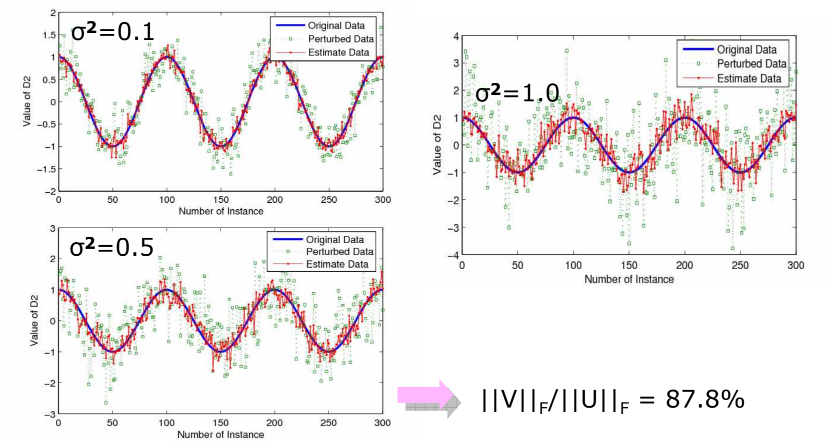
- Artificial Dataset
- 35 correlated variables
- 30,000 tuples



Strategy 1 vs. 2



Effect of varying noise



$||V||_F/||U||_F$ = 87.8%

P2: How accurate we can achieve?

- $U_p^T U_p = (U+V)^T(U+V) = U^TU + V^TU + U^TV + V^TV$

- When signal and noise are uncorrelated, for large number of observations: $V^TU \sim 0$ and $U^TV \sim 0$,

$$U_p^T U_p = U^TU + V^TV$$

- Hence,

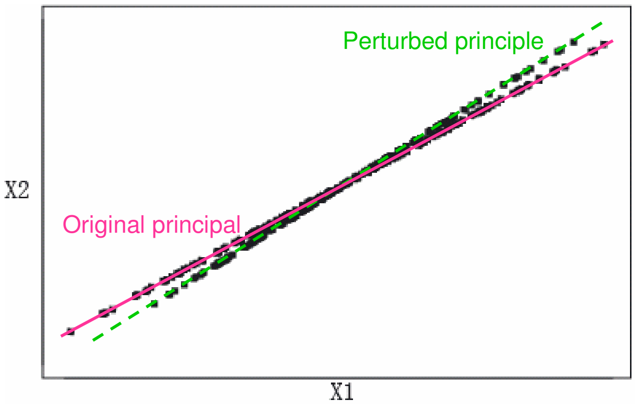
$$\hat{U} = U_p Q_{p_k} Q_{p_k}^T = (U+V) Q_{p_k} Q_{p_k}^T = U Q_{p_k} Q_{p_k}^T + V Q_{p_k} Q_{p_k}^T$$

- Result from Huang et al. SIGMOD05

$$Var(VQ_{p_k}Q_{p_k}^T) = Var(V)\frac{k}{n}$$

Other Factor

- Q_{p_k} from U_p is not the same as Q_k from U



- More theoretical analysis from matrix perturbation is needed

Spectral Filtering

- It can handle small data sets better.
- SF uses random matrix theory.
- When data size is small, eigenvalues of normal random matrix have semi-circle distribution with a thin range given by $\lambda_{\min}$ and $\lambda_{\max}$ (**Wigner's law**)

Distribution of Noise Eigenvalues

- Let $V_{m\times n}$ be a random matrix whose entries are i.i.d. random variables with zero mean and variance σ^2 . Let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be the eigenvalues of covariance matrix $V^T V$

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} U(x-\lambda_i), \quad where \quad U(x) = \begin{cases} 1 & x \ge 0 \\ 0 & x < 0 \end{cases}$$

- The empirical c.d.f. of the eigenvalues λ_i

$$\lambda_{\min} = \sigma^2 (1 - 1/\sqrt{Q})^2 \qquad \frac{m(N)}{n(N)} \to Q$$
$$\lambda_{\max} = \sigma^2 (1 + 1/\sqrt{Q})^2$$

SF vs. PCA

- In most data sets, the number of tuples is much larger than that of attributes.
- Hence, $Q = \frac{m(N)}{n(N)}$ tends to be large
- So $\lambda_{\min} \approx \lambda_{\max} \approx \lambda_V$

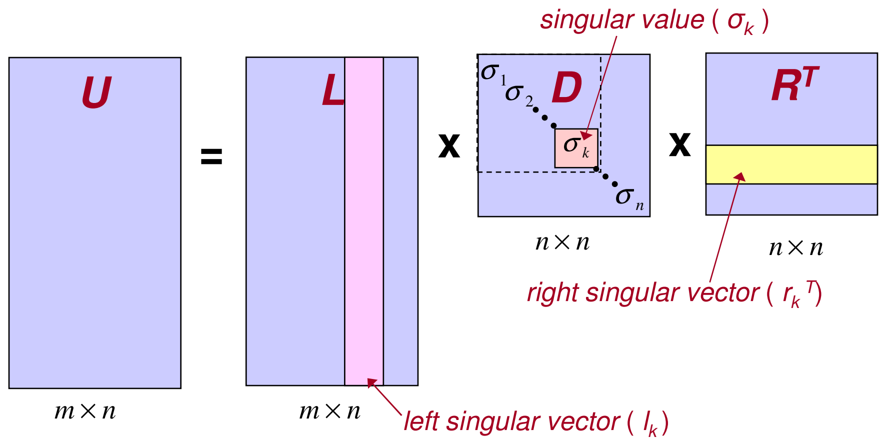
$$\lambda_{\min} = \sigma^2 (1 - 1/\sqrt{Q})^2$$
$$\lambda_{\max} = \sigma^2 (1 + 1/\sqrt{Q})^2$$

Why SVD?

- SVD is equivalent to PCA
  - SVD works directly on data
  - PCA works on covariance matrix of data
- Its Schmidt Theorem can help determine the lower bound of reconstruction accuracy

ECML/PKDD06 Tutorial

SVD Revisited



$$U = \sum_{k=1}^{n} \sigma_k l_k r_k^T$$

ECML/PKDD06 Tutorial

Theorem

$$U + V = \tilde{U}$$

Perturbation

- Theorem (Weyl)

$$|\tilde{\sigma}_i - \sigma_i| \le \|V\|_2, \qquad i = 1, \ldots, n$$

- Theorem (Mirsky)

$$\sqrt{\sum_i (\tilde{\sigma}_i - \sigma_i)^2} \le \|V\|_F$$

- Theorem (Schmidt)
  - The matrix U_k is a matrix of rank k that is nearest U

$$\|\hat{U} - U\|_F^2 \ge \sum_{i=1}^{n} (\tilde{\sigma}_i - \sigma_i)^2 \ge \sigma_{k+1}^2 + \cdots + \sigma_n^2 \ge \|U_k - U\|_F^2$$

SVD vs. PCA

$$A = U^T U = Q \Lambda Q^T \qquad U = LDR^T$$

$$U^T U = (LDR^T)^T (LDR^T) = RD^T L^T LDR^T = R(D^T D)R^T$$

- The singular values of U are the square roots of the eigenvalues of $U^T U$ or UU^T
- Rows of R are eigenvectors of $U^T U$
- LD gives coordinates of rows of U in the space of principal components

SVD Reconstruction

Input: $\tilde{U}$, a given perturbed data set
V, a noise data set
Output: $\hat{U}$, a reconstructed data

BEGIN

1 Apply SVD on $\tilde{U}$ to get $\tilde{U} = \tilde{L}\tilde{D}\tilde{R}^T$

2 Apply SVD on V and assume $\sigma_{V_{\max}}$ is the largest singular value

3 Determine the first k components of $\tilde{U}$ by $k = \min\{i \mid \tilde{\sigma}_i < \sqrt{2}\sigma_V\} - 1$

4 Reconstructing the data as

$$\hat{U} = \tilde{U}_k = \sum_{i=1}^{k} \tilde{\sigma}_i \tilde{l}_i \tilde{r}_i^T \, (k \le \rho)$$

END

UNC CHARLOTTE

SVD vs. PCA Reconstruction

- Equivalence of methods

$$\hat{U}_{SF} = \tilde{U}P_{\tilde{\chi}} = \tilde{U}\tilde{Q}_k\tilde{Q}_k^T \longleftrightarrow \hat{U}_{SVD} = \tilde{L}_k\tilde{D}_k\tilde{R}_k^T$$

$$\tilde{Q} = \tilde{R}$$

$$\tilde{U}\tilde{R}_k = \tilde{U}\tilde{R}\begin{pmatrix} I_k \\ 0 \end{pmatrix} = (\tilde{L}\tilde{D}\tilde{R}^T)\tilde{R}\begin{pmatrix} I_k \\ 0 \end{pmatrix} = \tilde{L}\tilde{D}\begin{pmatrix} I_k \\ 0 \end{pmatrix} = \tilde{L}_k\tilde{D}_k$$

- Equivalence of determining k

$$k = \min\{i \mid \tilde{\lambda}_i < 2\lambda_V\} - 1 \longleftrightarrow k = \min\{i \mid \tilde{\sigma}_i < \sqrt{2}\sigma_V\} - 1$$

$$\tilde{\sigma}_i = \sqrt{\lambda_i(\tilde{U}^T\tilde{U})} = \sqrt{\tilde{\lambda}_i}$$

$$\sqrt{2}\sigma_V = \sqrt{2\lambda(V^TV)} = \sqrt{2\lambda_V}$$

UNC CHARLOTTE

Challenging Questions

- Previous work on individual reconstruction are only empirical
  - Attacker question: How close the estimated data using SF is from the original one?
  - Data owner question: How much noise should be added to preserve privacy at a given tolerated level?

$$\tau_1 \leq \| \hat{U} - U \| \leq \tau_2$$

Upper Bound

- The upper bound of $\| \hat{U} - U \|_F$ in terms of V
  - The upper bound determines how close the estimated data achieved by attackers is from the original one
  - It imposes a serious threat of privacy breaches

Upper Bound

- Given $\tilde{U} = U + V$

  Let $\hat{U}$ be the estimation obtained from the SF/PCA, then

$$\|\hat{U} - U\|_F \le \|\tilde{U}\|_F \frac{2\|E\|_F}{(\tilde{\lambda}_k - \|E\|_2) - \sqrt{2}\|E\|_F} + \|VP_\chi\|_F$$

  where $E = V^T U + U^T V + V^T V$ is the derived perturbation on the original covariance matrix $A = U^T U$

UNC CHARLOTTE

Special Cases

- When the noise matrix is generated by i.i.d. Gaussian distribution with zero mean and known variance

$$\|\hat{U} - U\|_F \le \|U_P\|_F \frac{2\|V\|_F^2}{(\tilde{\lambda}_k - \|E\|_2) - \sqrt{2}\|V\|_F^2} + \sqrt{k/n}\|V\|_F$$

- When the noise is completely correlated with data

$$\|\hat{U} - U\|_F \le \|U_P\|_F \frac{2\|V\|_F^2}{(\tilde{\lambda}_k - \|E\|_2) - \sqrt{2}\|V\|_F^2} + \|V\|_F$$

More details, See Guo and Wu, SAC06

UNC CHARLOTTE

Lower Bound

- The lower bound represents the best estimate the attacker can achieve by the spectral filtering technique
- Hard problem --- it could not be derived by matrix perturbation theory

Lower Bound

- The lower bound of SVD reconstruction $\hat{U} = \tilde{U}_k = \tilde{L}_k \tilde{D}_k \tilde{R}_k^T$ is $\| U_k - U \|_F \leq \| \hat{U} - U \|_F$
  where $k = \min\{i \mid \tilde{\sigma}_i < \sqrt{2}\sigma_V\} - 1$
- The lower bound of SVD is the lower bound of PCA since SVD reconstruction is proved to be equivalent to PCA.

More details, See Guo, Wu and Li, PKDD 2006

How much noise be added?

Input: privacy threshold τ
Output: variance of the noise
BEGIN

- Determine k: $\tau \| U \|_F \leq \| U_k - U \|_F = \sigma_{k+1}^2 + \cdots + \sigma_n^2$

$$k = \max\{i \mid \tau \leq (\sigma_{i+1}^2 + \cdots + \sigma_n^2) / \| U \|_F\}$$

- Add i.i.d noise and let the eigenvalues of $V^T V$ satisfy:

$$\lambda_{k+1} < \lambda_V \leq \lambda_k$$

- Output $Var(V) = \lambda_V / (m-1)$

END

Outline (Part I)

Randomization based PPDM

- Additive noise
- Rotation
- General Linear Transformation

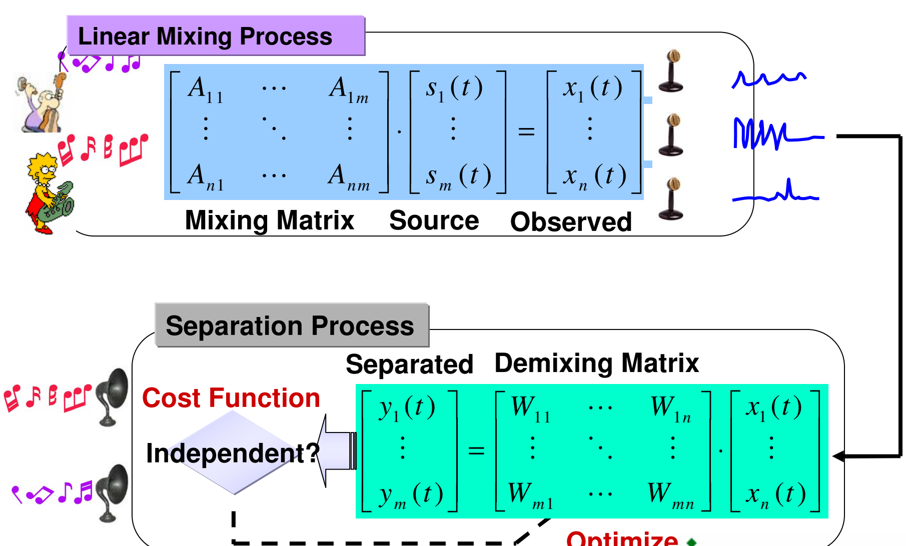
Attacking Methods

- Additive noise
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
- ***Rotation and General Linear Transformation***
  - ICA
  - A-priori Knowledge Based Attack

ICA Revisited

- ICA Motivation
  - Blind source separation: separating unobservable or latent independent source signals when mixed signals are observed
    - Cocktail-party problem
- What is ICA
  - ICA is a statistical technique which aims to represent a set of random variables as linear combinations of statistically independent component variables
  - ICA is a process for determining the structure that produced a signal

ICA



$$\begin{bmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nm} \end{bmatrix} \cdot \begin{bmatrix} s_1(t) \\ \vdots \\ s_m(t) \end{bmatrix} = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

$$\begin{bmatrix} y_1(t) \\ \vdots \\ y_m(t) \end{bmatrix} = \begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{m1} & \cdots & W_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix}$$

ICA Direct Attack?

- Can we get X from $Y = RX$ when only Y is available?
  - It seems Independent Component Analysis can help.

$Y = R\ X$ Rotation Model

$X = A\ S$ Noise-free ICA Model

Restriction of ICA

- Restrictions:
  - All the components s_i should be independent;
    They must be non-Gaussian with the possible exception of one component.
  - The number of observed linear mixtures m must be at least as large as the number of independent components n.
  - The matrix A must be of full column rank

$X = AS$

$Y = RX$

- Can we apply the ICA directly? No
  - Correlations among attributes of X
  - More than one attributes may have Gaussian distributions

General Linear Transformation

- We can not apply noisy ICA direct attack either

$Y = R \; X + E$ General Linear Perturbation Model

$X = A \; S + N$ Noisy ICA Model

Outline(Part I)

Randomization based PPDM

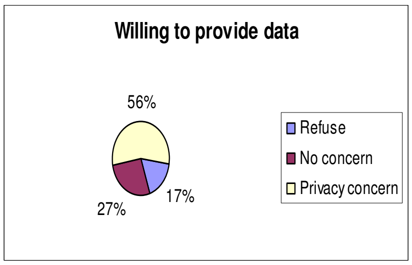
- Additive noise
- Rotation
- General Linear Transformation

Attacking Methods

- Additive noise
  - IQR (from distribution)
  - Spectral Filtering, PCA, SVD
- Rotation and General Linear Transformation
  - ICA
  - ***A-priori Knowledge Based Attack***

A-priori Knowledge Attack

- Privacy can be breached when a small subset of the original data *X* , is available to attackers
- Assumption is reasonable!



	Bal	income	…	IntP
1	10k	85k	…	2k
2	***15k***	***70k***	***…***	***18k***
3	***50k***	***120k***	***…***	***35k***
4	45k	23k	…	134k
.	.	.	…	.
N	80k	110k	…	15k

Understanding net users' attitude about online privacy, April 99

A-priori Knowledge Based Attack

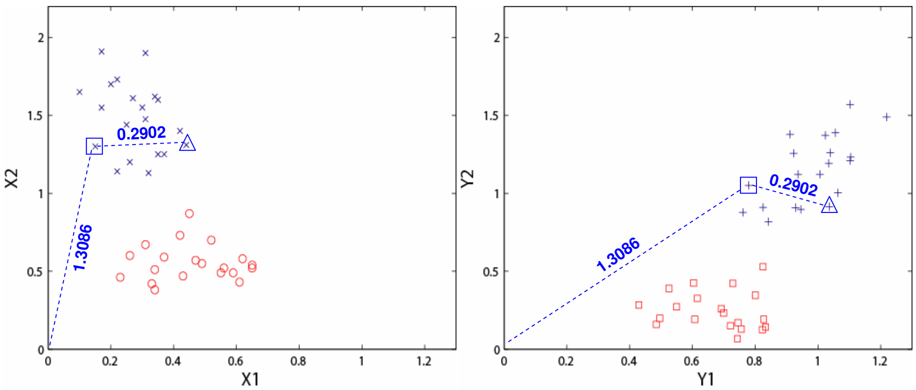
- For rotation $Y = RX, RR^T = I$, it is straightforward

1. If *Rank(X)* = *d*, we assume at least *d* points in *X* are known by attackers, denoting this subset as $\tilde{X}$
2. Locate the corresponding points in *Y* based on their preserved lengths, denoting this subset as $\tilde{Y}$
3. Find the linear transformation (*R*) by applying multivariate linear regression analysis on $\tilde{X}$ and $\tilde{Y}$

$$B^T = (\tilde{Y}\tilde{Y}^T)^{-1}\tilde{Y}\tilde{X}^T$$

$$\hat{X} = BY$$

Why easy for rotation?



$$Y = RX$$

$$R = \begin{pmatrix} 0.866 & 0.500 \\ -0.500 & 0.866 \end{pmatrix}$$

A-priori Knowledge Based Attack

- **For General Linear Transformation** ***Y = RX + E***
  - **It is not straightforward since we can not find matched pairs (due to distance is not preserved).**
  - **AK-ICA**

A-priori Knowledge based ICA (AK-ICA) Attack

input Y, a given perturbed data set
$\tilde{X}$, a given subset of original data
output $\hat{X}$, a reconstructed data set

BEGIN
1 Applying ICA on $\tilde{X}$ and Y to get
$$\tilde{X} = A_{\tilde{x}} S_{\tilde{x}}$$
$$Y = A_y S_y$$
2 Deriving the transformation matrix J by comparing the distributions of $S_{\tilde{x}}$ and S_y
3 Reconstructing X approximately as
$$\hat{X} = A_{\tilde{x}} J S_y$$
END

Correctness of AK-ICA

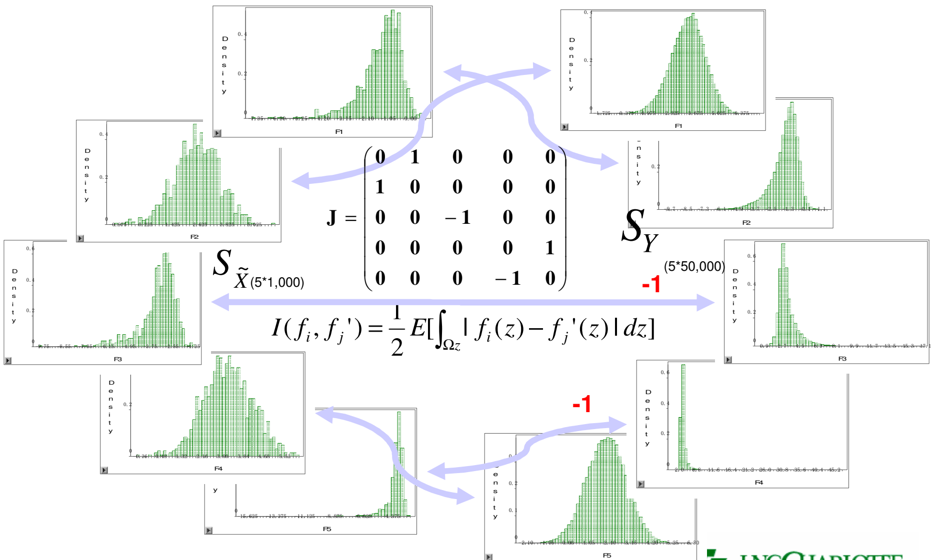
- We prove that *J* exists such that

$$\hat{X} = A_{\tilde{x}} J S_y \approx X$$

  - *J* represents the connection between the distributions of $S_{\tilde{X}}$ and S_Y

More details, See Guo and Wu, 2006

How to determine J ?



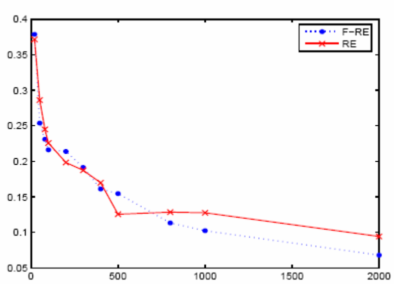
$$\mathbf{J} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 \end{pmatrix}$$

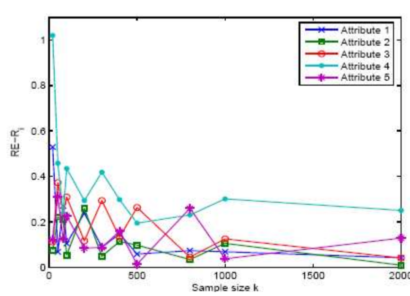
$$I(f_i, f_j') = \frac{1}{2} E[\int_{\Omega z} | f_i(z) - f_j'(z) | dz]$$

Experimental Results

- Reconstruction error vs. Sample size

$$F-RE = \frac{\| \hat{X} - X \|_F}{\| X \|_F} \qquad RE = \frac{1}{d \times n} \sum_{i=1}^{d} \sum_{j=1}^{n} \left| \frac{x_{ij} - \hat{x}_{ij}}{x_{ij}} \right| \qquad RE-R_i = \frac{1}{n} \sum_{j=1}^{n} \left| \frac{x_{ij} - \hat{x}_{ij}}{x_{ij}} \right|$$

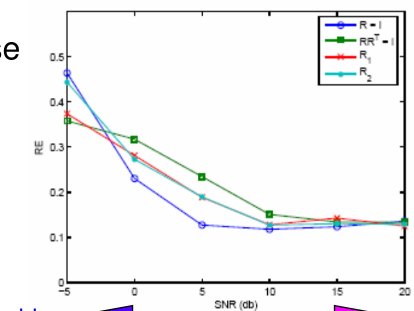




Bank data set (5*50,000)

Experimental Results (cont'd)

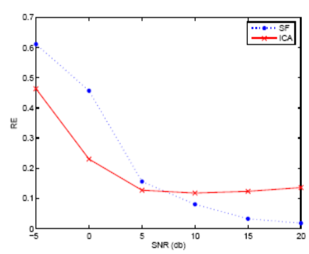
- $Y = RX + E$
  - $R = I$
  - $RR^T = R^T R = I$
  - R_1: Random matrix $det(R_1) = 0.444$ and $||R_1||_F = 3.167$
  - R_2: Random matrix $det(R_2) = 2.48 \times 10^9$ and $||R_2||_F = 281.8$
- Reconstruction error vs. Noise
- Robust with R



big noise small

UNC CHARLOTTE

AK-ICA vs SF



big noise small

	AK-ICA	Spectral Filter
Applicable Model	$Y = RX + E$	$Y = X + E$
Assumption	A small subset of original data	Covariance matrix of the noise
Cumulants	fourth-order	second-order
Small Noise		Better
Big Noise	Better	

UNC CHARLOTTE

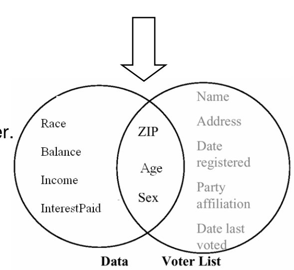
Challenging Questions (Part I)

- Additive noise vs. Reconstruction accuracy
  - How much noise should be at least added to preserve privacy?
  - What kind of noise is better for a given dataset?
  - How does additive noise affect accuracy of data mining results?
- Sample vs. Reconstruction accuracy
  - How many records are sufficient to allow attackers breach the security?
  - What kind of datasets are more prone to privacy breaching?
  - What happens if sample is biased?
- How about some attributes are known to attackers?
- How about combinations of confidential attributes?
  - Is total income = Bal + Income + IntP secure?

Scope

	ssn	name	zip	race	…	age	Sex	Bal	income	…	IntP
1			28223	Asian	…	20	M	10k	85k	…	2k
2			28223	Asian	…	30	F	15k	70k	…	18k
3			28262	Black	…	20	M	50k	120k	…	35k
4			28261	White	…	26	M	45k	23k	…	134k
.			.	.	…	.	.	.	.	…	.
N			28223	Asian	…	20	M	80k	110k	…	15k

69% unique on zip and birth date
87% with zip, birth date and gender.
k-anonymity, L-diversity
SDC etc.



Outline(Part II)

Randomization Response (Survey)

- Model
  - One Dichotomous Attribute
  - One Polychotomous Attribute
  - Multi Attribute
- Disclosure Analysis

What is left out?

- Microdata generalization
  - K-anonymity, L-diversity, (alpha,k)-anonymity
  - Some recent developments (skipped, see Gehrke's tutorial)

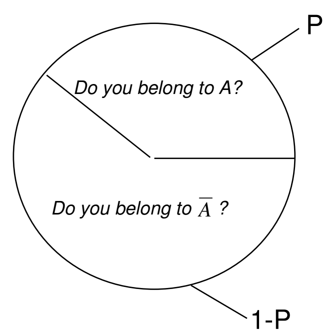
- Privacy in market basket 0-1 data
  - Data swapping
  - Item randomization
  - Frequent itemsets or rule hiding
  - Inverse frequent itemset mining

Randomized Response History

- Survey sampling
  - The Warner Model, 1965
  - Extension to polychotomous attribute
  - Extension to multi-attribute
- Statistical databases
  - Reinvented as PRAM by Statistics Netherland (1997)
  - Applied to microdata
- Introduced to KDD community
  - Du kdd04

One Dichotomous Attribute

- Warner Model
  - Each respondent belongs to a sensitive group A or its complement $\bar{A}$
  - A randomization device is given to the respondent to choose one of the two questions.
- Problem
  - Given λ ,the prob, of observing a "yes" answer and P, estimate π_A ,the true prob. of respondents in A



Randomization device

$$\lambda = \pi_A P + (1-\pi_A)(1-P)$$

Analysis

Randomized response (RR) Warner Model

$$\hat{\pi}_{AW} = \frac{P-1}{2P-1} + \frac{n_1}{(2P-1)n}$$

n_1 is the number of "yes" response in sample n

$$\text{var}(\hat{\pi}_{AW}) = \frac{\pi_A(1-\pi_A)}{n} + \frac{1}{n}\left[\frac{1}{16(P-0.5)^2} - \frac{1}{4}\right]$$

Direct response (DR) Model:

$$\hat{\pi}_{AD} = n_1 / n$$

$$\text{var}(\hat{\pi}_{AD}) = \frac{\pi_A(1-\pi_A)}{n}$$

One Polychotomous Attribute

- A sensitive attribute has t classes
- A respondent belonging to the *i* th class $(i = 1,\dots,t)$ will report j $(j = 1,\dots,t)$ with respective probabilities $p_{1i}, p_{2i},\dots, p_{ti}$, $\sum_{j=1}^{t} p_{ji} = 1$

$$\begin{array}{cc} & i \\ & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ j \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0.60 & 0.20 & 0.00 & 0.10 \\ 0.20 & 0.50 & 0.20 & 0.10 \\ 0.15 & 0.15 & 0.70 & 0.30 \\ 0.05 & 0.15 & 0.10 & 0.50 \end{pmatrix} \end{array} \Longrightarrow P = ((p_{ji}))$$

- E.g. If a respondent belonging to the *2nd* category , he will report category 3 with 0.15 probability .

Vector Response

- $\pi = (\pi_1, ..., \pi_t)'$ is the true proportions of the population
- $\lambda = (\lambda_1, ..., \lambda_t)'$ is the observed proportions in the survey
- $P = ((p_{ji}))$ is the randomization device set by the interviewer.

$$\begin{array}{cc} & i \\ & \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ j \; \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0.60 & 0.20 & 0.00 & 0.10 \\ 0.20 & 0.50 & 0.20 & 0.10 \\ 0.15 & 0.15 & 0.70 & 0.30 \\ 0.05 & 0.15 & 0.10 & 0.50 \end{pmatrix} \end{array} \begin{pmatrix} 0.10 \\ 0.30 \\ 0.20 \\ 0.40 \end{pmatrix} = \begin{pmatrix} 0.16 \\ 0.25 \\ 0.32 \\ 0.27 \end{pmatrix}$$

$$P \quad \pi = \lambda$$

Analysis

$$\lambda = P\pi \quad \Longrightarrow \quad \hat{\pi} = P^{-1}\hat{\lambda}$$

$$disp(\hat{\lambda}) = n^{-1}(\lambda^{\delta} - \lambda\lambda') \quad \Longrightarrow \quad disp(\hat{\pi}) = n^{-1}P^{-1}(\lambda^{\delta} - \lambda\lambda')P'^{-1}$$

diagonal matrix with elements λ

$$= n^{-1}(P^{-1}\lambda^{\delta}P'^{-1} - \pi\pi')$$

$$= \Sigma_1 + \Sigma_2$$

$\Sigma_1 = n^{-1}(\pi^{\delta} - \pi\pi')$ the dispersion matrix of the regular survey estimation

$\Sigma_2 = n^{-1}P^{-1}(\lambda^{\delta} - P\pi^{\delta}P')P'^{-1}$ nonnegative definite, represents the components of dispersion associated with RR experiment

PRAM

- Post RAandomisation Method
  - Developed by Statistics Netherland (1997)
  - Applied to categorical microdata
- Similar to vector response model except
  - PRAM assumes micro data is a-priori known
  - It may also incorporate invariant matrix P

$$\begin{array}{cc} & i \\ j & \begin{array}{c|cccc} & 1 & 2 & 3 & 4 \\ \hline 1 & 0.95 & 0.00 & 0.00 & 0.10 \\ 2 & 0.05 & 0.98 & 0.00 & 0.00 \\ 3 & 0.00 & 0.02 & 0.95 & 0.00 \\ 4 & 0.00 & 0.00 & 0.05 & 0.90 \end{array} \end{array} \begin{pmatrix} 0.20 \\ 0.50 \\ 0.20 \\ 0.10 \end{pmatrix} = \begin{pmatrix} 0.20 \\ 0.50 \\ 0.20 \\ 0.10 \end{pmatrix}$$

Multi Polychotomous Attributes

- m sensitive attributes: $A_1, A_2, ..., A_m$, each has t_j categories: $A_{j1}, ..., A_{jt_j}$
- $\pi_{i_1,...,i_m}$ denote the true proportion corresponding to the combination $(A_{1i_1}, ..., A_{mi_m})$, π be vector with elements $\pi_{i_1,...,i_m}$ $(i_j = 1,..,t_j)$,arranged lexicographically.
- e.g., if m =2, t1 =2 and t2=3

$$\pi = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{23})'$$

Simultaneous Model

- Consider all variables as one compounded variable and apply the regular vector response RR technique
- Can preserve structural zeros, e.g., a man can not be pregnant.

Sequential Model

§ Each respondent makes m independent RR trials, one for each attribute.

§ $\lambda_{\mu_1,\ldots,\mu_m}$ is the probability of getting a response $(\mu_1,\ldots,\mu_m)$

$$\lambda = (P_1 \otimes P_2 \otimes \ldots \otimes P_m)\pi$$

$\otimes$ stands for Kronecker product

§ A unbiased estimate of π is

$$\hat{\pi} = (P_1^{-1} \otimes \ldots \otimes P_m^{-1})\hat{\lambda}$$

Kronecker Product Example

$$P_1: \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \qquad P_2: \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix}$$

$$P_1 \otimes P_2 = \begin{bmatrix} a_{11}P_2 & a_{12}P_2 \\ a_{21}P_2 & a_{22}P_2 \end{bmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{12}b_{11} & a_{12}b_{12} & a_{12}b_{13} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} \\ a_{11}b_{31} & a_{11}b_{32} & a_{11}b_{33} & a_{12}b_{31} & a_{12}b_{32} & a_{12}b_{33} \\ a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{22}b_{11} & a_{22}b_{12} & a_{22}b_{13} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} \\ a_{21}b_{31} & a_{21}b_{32} & a_{21}b_{33} & a_{22}b_{31} & a_{22}b_{32} & a_{22}b_{33} \end{pmatrix}$$

Analysis

$$\lambda = (P_1 \otimes \cdots \otimes P_m)\pi \quad \Longrightarrow \quad \hat{\pi} = (P_1^{-1} \otimes \cdots \otimes P_m^{-1})\hat{\lambda}$$

$$disp(\hat{\pi}) = n^{-1}(P^{-1}\lambda^{\delta}P'^{-1} - \pi\pi')$$

Similarly, the dispersion matrix can be decomposed to two parts: one corresponds to that of the regular survey estimation and the other corresponds to the components of dispersion associated with RR experiment

Attributes Correlation

- Association is preserved
- The problem of testing the null hypothesis of complete independence of the sensitive attributes $A_1, A_2, \ldots, A_m$,

$$H_0 : \pi = \pi^{(1)} \times \cdots \times \pi^{(m)}$$

where $\pi^{(j)}$ is the vector of marginal proportion corresponding to A_j

is equivalent to $H_0' : \lambda = \lambda^{(1)} \times \cdots \times \lambda^{(m)}$ where $\lambda^{(j)} = P_j \pi^{(j)}$

Disclosure Analysis with RR

R: Typical response which is "yes" (y) or "no" ($\bar{y}$)
$P(R|A)$, $P(R|\bar{A})$ are conditional probabilities set by investigators

Posterior probabilities:

$$P(A|R) = \frac{\pi_A P(R|A)}{\pi_A P(R|A) + (1-\pi_A) P(R|\bar{A})} \qquad P(\bar{A}|R) = 1 - P(A|R)$$

R is regarded as jeopardizing with respect A or $\bar{A}$ if:

$$P(A|R) > \pi_A \quad \text{or} \quad P(\bar{A}|R) > 1 - \pi_A$$

Efficient Estimation vs. Privacy Protection

Since $$\frac{P(A|R)}{P(\bar{A}|R)}\frac{1-\pi_A}{\pi_A}=\frac{P(R|A)}{P(R|\bar{A})}$$

Disclosure measure (Warner 1976):

$$g(R|A)=\frac{P(R|A)}{P(R|\bar{A})} \qquad g(R|\bar{A})=\frac{1}{g(R|A)}$$

No disclosure if and only if:

$$g(R|A)=1$$

But it's impossible to get an Unbiased Estimate for π_A

$$\hat{\pi}_A=\frac{\hat{\lambda}-P(y|\bar{A})}{P(y|A)-P(y|\bar{A})} \quad (P(y|A)-P(y|\bar{A})\neq 0)$$

Disclosure Analysis

Risks of suspicion:

$$P(A|R)\leq \xi_2<1 \qquad P(\bar{A}|R)\leq 1-\xi_1<1$$

$$\xi_1\leq P(A|R)\leq \xi_2$$

So:

$$g(y|A)\leq \frac{1-\pi_A}{\pi_A}\frac{\xi_2}{1-\xi_2}$$

$$g(\bar{y}|\bar{A})\leq \frac{\pi_A}{1-\pi_A}\frac{1-\xi_1}{\xi_1}$$

Risk Disclosure with RR

- Some Generalities
  - X be a variable of interest with unknown distribution parameter θ
  - Y be a response variable through an RR device
  - $I(Y)$ be Fisher's information per unit observation on Y
  - $I(X)$ be Fisher's information per unit observation on X if X is observable

$$I(X) > I(Y)$$

  - The loss of information is defined as $L = 1 - \frac{I(Y)}{I(X)}$
  - $\mathrm{var}(X \mid Y = y)$ as a measure of protection of privacy for a given y
  - $E(\mathrm{var}(X \mid Y))$ as an overall measure of protection of privacy

Summary

- We only touched the surface. PPDM is
  - Practical
  - Fun
  - Challenging!
- Some general open problems
  - Tradeoff of utility vs. privacy
  - How to handle mixing attributes together?
  - Theory and application of various reconstruction methods
  - Formalization of background knowledge
  - Optional randomization (horizontal, vertical)
  - Etc.

Acknowledgement

- USA NSF CCR-0310974 and IIS-0546027

- Discussions and Slides
  - Keke Chen, Chris Clifton, Kevin Du, Johannes Gehrke, Ling Guo, Songtao Guo, Hillo Kargupta, Yingjiu Li

- For a more complete list of references
  - http://www.cs.uncc.edu/~xwu/ppdm-bibl.htm

References

[1] R. Agrawal and R. Srikant. *"Privacy-preserving data mining", SIGMOD00*

[2] D. Agrawal and C. Agrawal. *"On the design and quantification of privacy preserving data mining algorithms"*, PODS01

[3] C. Aggarwal and P. Yu. "A condensation approach to privacy preserving data mining", EDBT04.

[4] A. Chaudhuri and R. Mukerjee. *"Randomized Response Theory and Techniques",* Marcel Dekker, 1998.

[5] K. Chen and L. Liu. *"Privacy preserving data classification with rotation perturbation"*, ICDM05

[6] J. Domingo-Ferrer, F. Seb and J. Castell. *"On the security of noise addition for privacy in statistical databases", PSD04*

[7] W. Du and Z. Zhan. *"Using randomized response techniques for privacy preserving data mining",* KDD03

References

[8] A. Evfimievski and J. Gehrke and R. Srikant. *"Limiting privacy breaches in privacy preserving data mining"*, PODS03

[9] A. Evfimievski. "*Randomization in privacy preserving data mining*", SIGKDD Explorations 4(2),2002.

[10] S.E. Fienberg, U.E. Markov, and R. J. Steele. "*Disclosure limitation using perturbation and related methods for categorical data"*. Journal of Official Statistics, 14(4), 1998.

[11] J.M. Gouweleeuw, P. Kooiman, L. Willenborg, and P.P. de Wolf. *"Post randomisation for statistical disclosure control: Theory and implementation",* Journal of Official Statistics, 14(4):463-478, 1998.

[12]S. Guo and X. Wu. *"On the Use of Spectral Filtering for Privacy Preserving Data Mining",* SAC06.

[13]S. Guo, X. Wu and Y. Li. "*On the Lower Bound of Reconstruction Error for Spectral Filtering based Privacy Preserving Data Mining*",PKDD06.

[14] S. Guo, X. Wu and Y. Li. *"Deriving Private Information from Perturbed Data Using IQR based Approach",* PDM06.

[

References

[15]S. Guo, X. Wu. *"Deriving Private Information from General Linear Transformation Perturbed Data",* In Submission

[16] Z. Huang and W. Du and B. Chen. *"Deriving private information from randomized data",* SIGMOD05

[17] V. Iyengar. *"Transforming data to satisfy privacy constraints",* SIGMOD02

[18] H. Kargupta and S. Datta and Q. Wang and K. Sivakumar. *"On the Privacy Preserving Properties of Random Data Perturbation Techniques",* ICDM03

[19] J.J. Kim and W.E. Winkler. *"Multiplicative noise for masking continuous data ",* Statistics #2003-01, Technical Report, U.S. Bureau of the Census.

[20] K. LeFevre, D. DeWitt, and R. Ramakrishnan. "*Incognito: efficient full domain k-anonymity"*. SIGMOD05.

[21] K. LeFevre, D. DeWitt, and R. Ramakrishnan. "*Mondrian multidimensional k-anonymity"*. ICDE06.

References

[22] K. Liu, C. Giannella, and H. Kargupta. "Preserving maps for privacy preserving data mining", PKDD06.

[23] K. Liu and H. Kargupta and J. Ryan. *"Random projection based multiplicative data perturbation for privacy preserving distributed data mining"*, TKDE06

[24] A. Machanavajjhala, J. Gehrke, and D. Kifer. *"L-diversity: privacy beyond k-anonymity ",* ICDE06.

[25] S. Rizvi and J. Haritsa. *"Privacy preserving association rule mining",* VLDB02

[26] P. Samarati. *"Protecting respondents' identities in microdata release",* TKDE 2001.

[27] P.Samarati and L. Sweeney. "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression", IEEE S&P 1998.

[28] M. Trottini, S.E. Fienberg, U.E. Makov, and M.M Meyer. "*Additive noise and multiplicative bias as disclosure limitation techniques for continuous microdata: a simulation study*", Journal of Computational Methods in Sciences and Engineering 4(2004).

References

[29] R.C. Wong, J. Li, A.W. Fu, and K. Wangf. *"(alpha, k)-Anonimity: an enhanced k-anonimity model for privacy preserving data publishing",* KDD06.

[30] X. Wu, S. Guo, and Y. Li. *"Towards Value Disclosure Analysis in Modeling General Databases",* SAC06.

[31] X. Xiao and Y. Tao. "Personalized privacy preservation". SIGMOD06

For a more complete list of references

http://www.cs.uncc.edu/~xwu/ppdm-bibl.htm

Q&A