A Survey of Randomization Techniques for Privacy-Preserving Data Mining

Presented by

Amiya Kumar Maji

Data Mining

- Data Mining is the process of extracting patterns from data.
- Data Mining Models
  - Exploratory Data Analysis
    - Interactive visualization of data
  - Descriptive Modeling
    - Density estimation
    - Cluster analysis
    - Dependency modeling
  - Predictive Modeling
    - Classification and regression

- Data Mining Models
  - Discovering Patterns and Rules
    - Association rules
    - Outlier analysis
  - Retrieval by Content
    - Given a pattern find similar items from the data set

What is Privacy?

- No standard (unified) definition
- Varies with contexts
  - From a set of records an adversary should not identify the person associated with a record
  - If multiple parties hold different data segments, one party cannot see the data of other parties
    - May apply to aggregates
  - The data mining algorithm cannot read the original data values
    - Perturbation
  - The results of data mining operations are sensitive
    - A perfect classifier may accurately predict an individual's "class" even if the "class" is hidden

Examples of Privacy-Preserving Data Mining

- *n* Drug Manufacturers each with drug consumption/reaction data
  - Want to generate association rules across all data sets
  - One company will not release either its own share of data or its partial aggregate data
  - Add a random value to local aggregate and initiate protocol
- Secure Multiparty Computation (SMC) using commutative encryption
  - Costly
  - Building 408 node decision tree from 1728 items took 29 hours [Clifton 2004]

Categorizing Privacy-Preserving Data Mining

- The Randomization Method
  - Release perturbed data
    - Add noise to the original data
    - Perturbed data maintain statistical properties, but hard to guess original data from perturbed value
- *k*-Anonymity and *l*-Diversity
  - De-identification
  - Pseudo-identifiers
  - *k*-anonymity
  - Generalization and Suppression
  - Usability of Data
  - *l*-Diversity
- Distributed Privacy Preservation
- Downgrading Application Effectiveness

Examples of Privacy-Preserving Data Mining

- Release of anonymous data
  - De-identification
  - Pseudo-identifiers
  - *k*-anonymity
  - Generalization and Suppression
  - Usability of Data

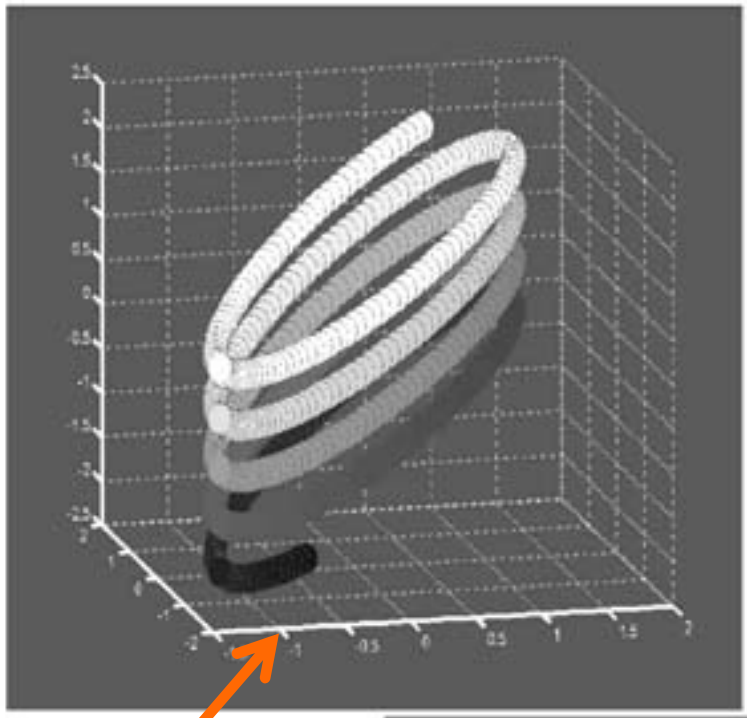
- Release perturbed data
  - Add noise to the original data
  - Perturbed data maintain statistical properties, but hard to guess original data from perturbed value

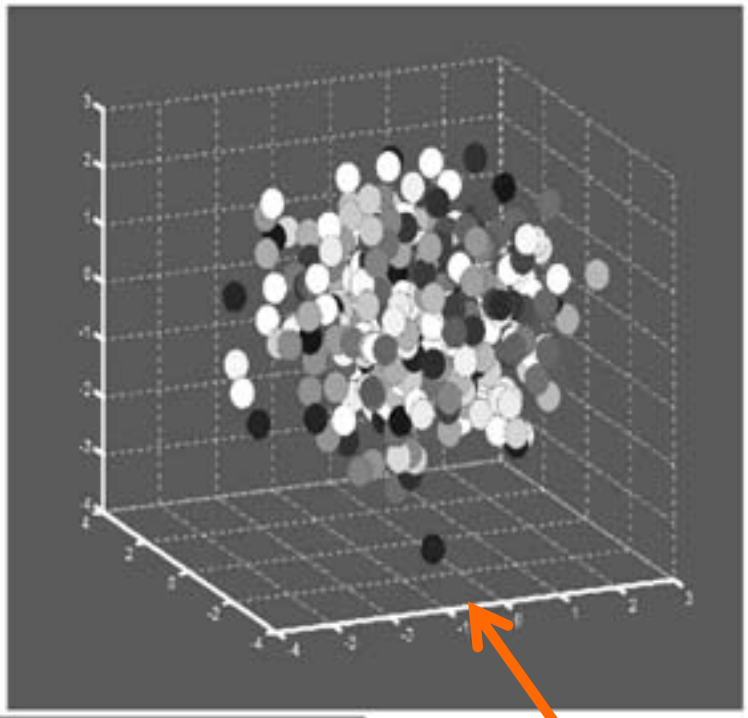
Randomization Techniques

- Additive Randomization
  - Consider a set of data records

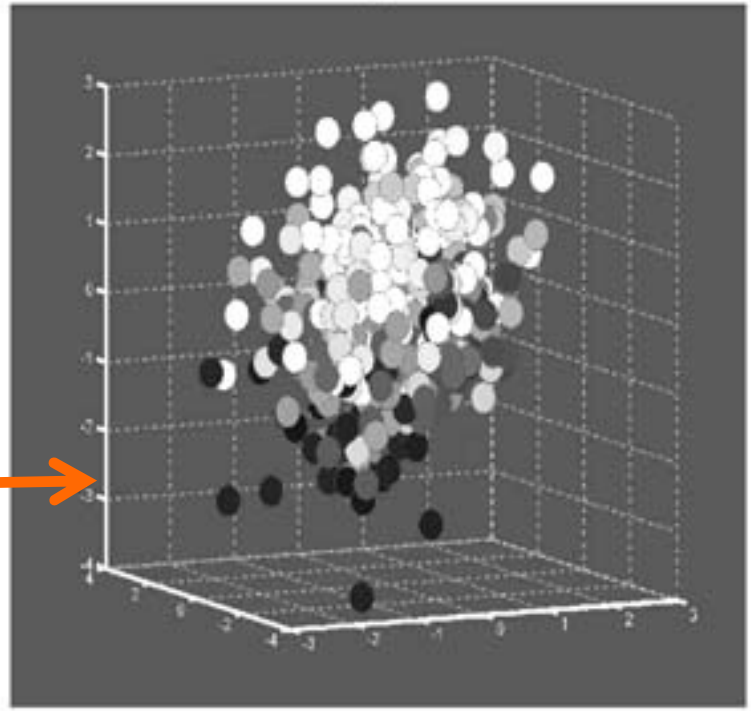
    $$X = \{x_1, x_2, \ldots, x_n\}$$

  - Draw *independent* noise components $Y = \{y_1, y_2, \ldots, y_n\}$ with probability distribution $f_Y(y)$
  - Let the distorted records be $z_1=x_1+y_1$, $z_2=x_2+y_2$, …, $z_n=x_n+y_n$
  - The distorted values $Z = \{z_1, z_2, \ldots, z_n\}$ are released
  - The distribution of noise $f_Y(y)$ is publicly known





Nonrandom data in 3D space



Random Noise (Uniform)

Perturbed Data

Kargupta 2004

The Reconstruction Problem

- Given a cumulative distribution function F_Y and the realizations of n i.i.d. random samples x_1+y_1, x_2+y_2, ..., x_n+y_n, estimate F_X
- Assume f' and F' are the estimated density functions
- Using Bayes' formula

$$F'(a) = \int_{-\infty}^{a} f_{X_1}(w|X_1 + Y_1 = z_1)dw$$

The Reconstruction Problem

$$F'(a) = \frac{\int_{-\infty}^{a} f_Y(z_1 - w) \cdot f_X(w) dw}{\int_{-\infty}^{\infty} f_Y(z_1 - w) \cdot f_X(w) dw}$$

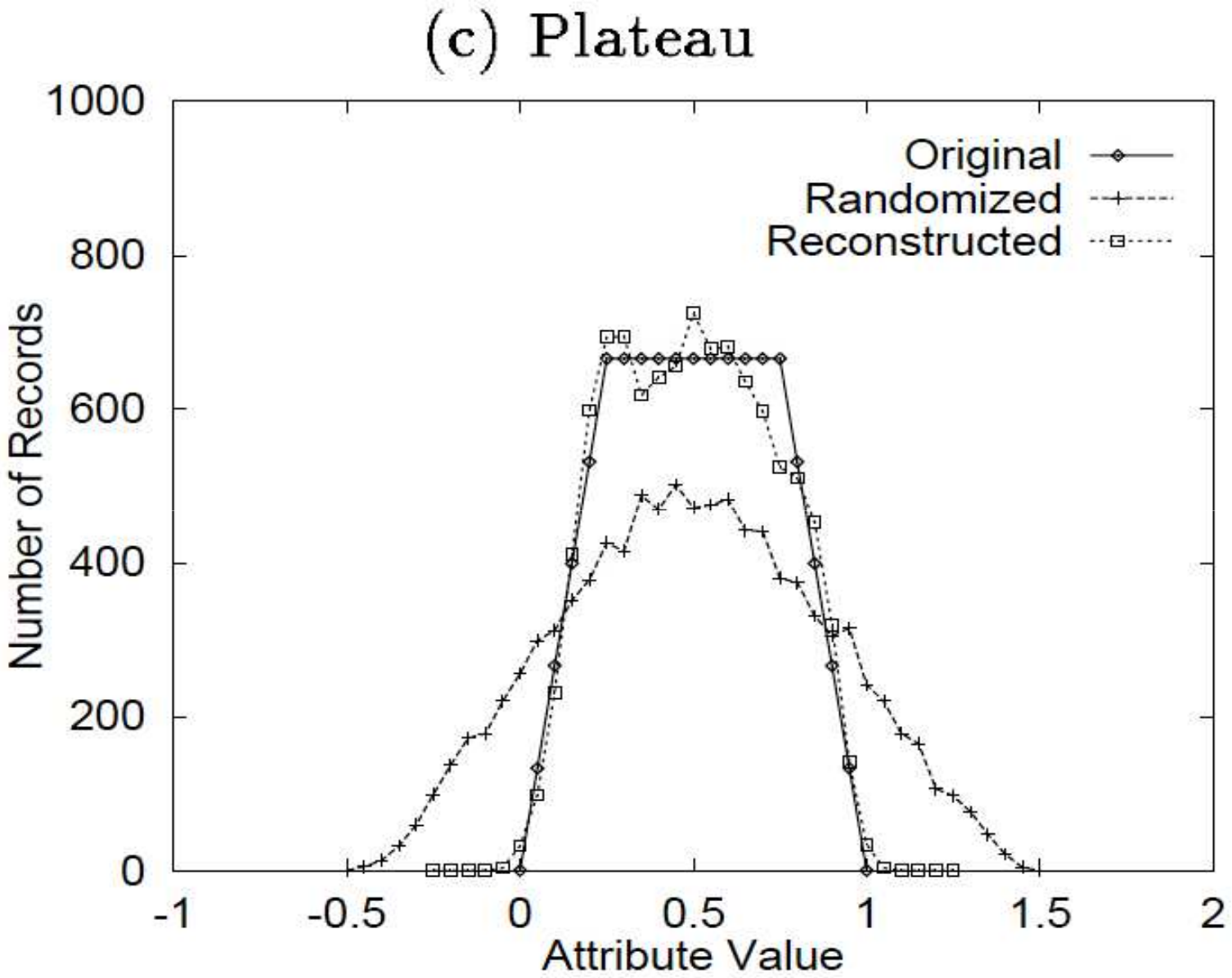
- Finally

$$f'(a) = (1/N) \cdot \sum_{i=1}^{N} \frac{f_Y(z_i - a) \cdot f_X(a)}{\int_{-\infty}^{\infty} f_Y(z_i - w) \cdot f_X(w) dw}$$

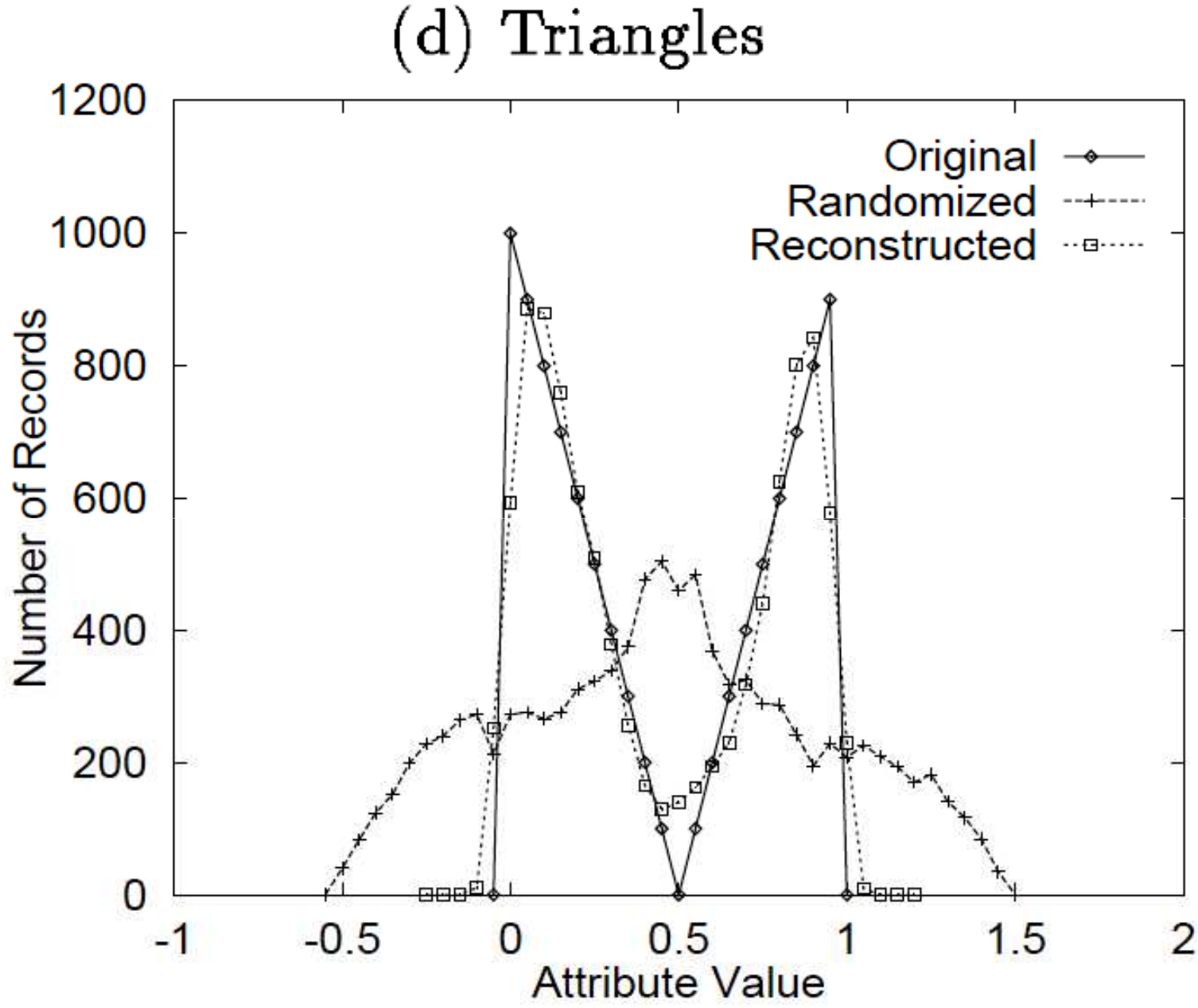
The Reconstruction Problem

- Note that $f_X(x)$ is unknown
- Use an iterative solution
- Start with f_X = *uniform distribution*
  - At each iteration refine f_X using earlier formula
  - Repeat until stopping condition is met

- Other reconstruction technique
  - EM (Expectation Maximization) based approach
  - Bayesian reconstruction is a special instance of EM approach



Agrawal and Srikant 2000



Agrawal and Srikant 2000

Quantifying Privacy

- If we can estimate with $c\%$ confidence that x lies in the interval $[x_1, x_2]$ then $(x_2 - x_1)$ defines the amount of privacy at $c\%$ confidence level.
- For example, if noise is *uniform* between $[-y, y]$ then, amount of privacy at 50% confidence is $0.5*2y = y$
- Insufficient if we consider properties of data

Quantifying Privacy

- Assume

$$f_X(x) = \begin{cases} 0.5 & 0 \le x \le 1 \\ 0.5 & 4 \le x \le 5 \\ 0 & otherwise \end{cases}$$

- Noise is uniform in [-1, 1]
- At 100% confidence privacy interval must be 2.
- If perturbed data lies in [-1, 2], then original data must lie in [0, 1]
- If perturbed data lies in [3, 6], then original data must lie in [4, 5]
- Privacy interval of 1

Quantifying Privacy

- Alternate approach
  - Use conditional differential entropy

$$h(A) = -\int_{\Omega_A} f_A(a) \log_2 f_A(a)\, da$$

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

- Average conditional privacy

$$2^{h(A|B)}$$

Multiplicative Randomization

- Multiply the records with random vectors
- Transform the data so that inter-record distances are preserved approximately
- Applications
  - Privacy-preserving clustering
  - Classification
- Attacks
  - Known input-output attack
    - Attacker knows some linearly independent collection of records and their perturbed versions
  - Known sample attack
    - Attacker has some independent samples from the original distribution

Randomization for Association Rule Mining

- Randomization through deletion and addition of items in transactions
- *Select-a-size* operator
  - Assume transaction size = m and a probability distribution $p[0], p[1], \ldots, p[m]$, over $\{0, 1, \ldots, m\}$
  - Given a transaction t of size m, generate randomized transaction t' as:
    - Select j at random from $0, .., m$ using above distribution
    - Select j items from t (unifromly without replacement) and place in t'
    - For each item a not in t, place a in t' with probability ρ
    - ρ is the randomization level

Microdata Randomization

- Swap the data values across rows
- Must leave a given number of records unchanged
- Can be very accurate for certain types of statistics

Some Conflicting Issues

- Privacy vs Utility

Curse of Dimensionality

- High dimensional data is difficult to anonymize
  - Number of pseudo-identifiers is very large
  - Data becomes sparse in high-dimensional space
  - Preserving k-anonymity requires generalization beyond tolerable limit for data mining purposes
  - Data loses utility

Curse of Dimensionality

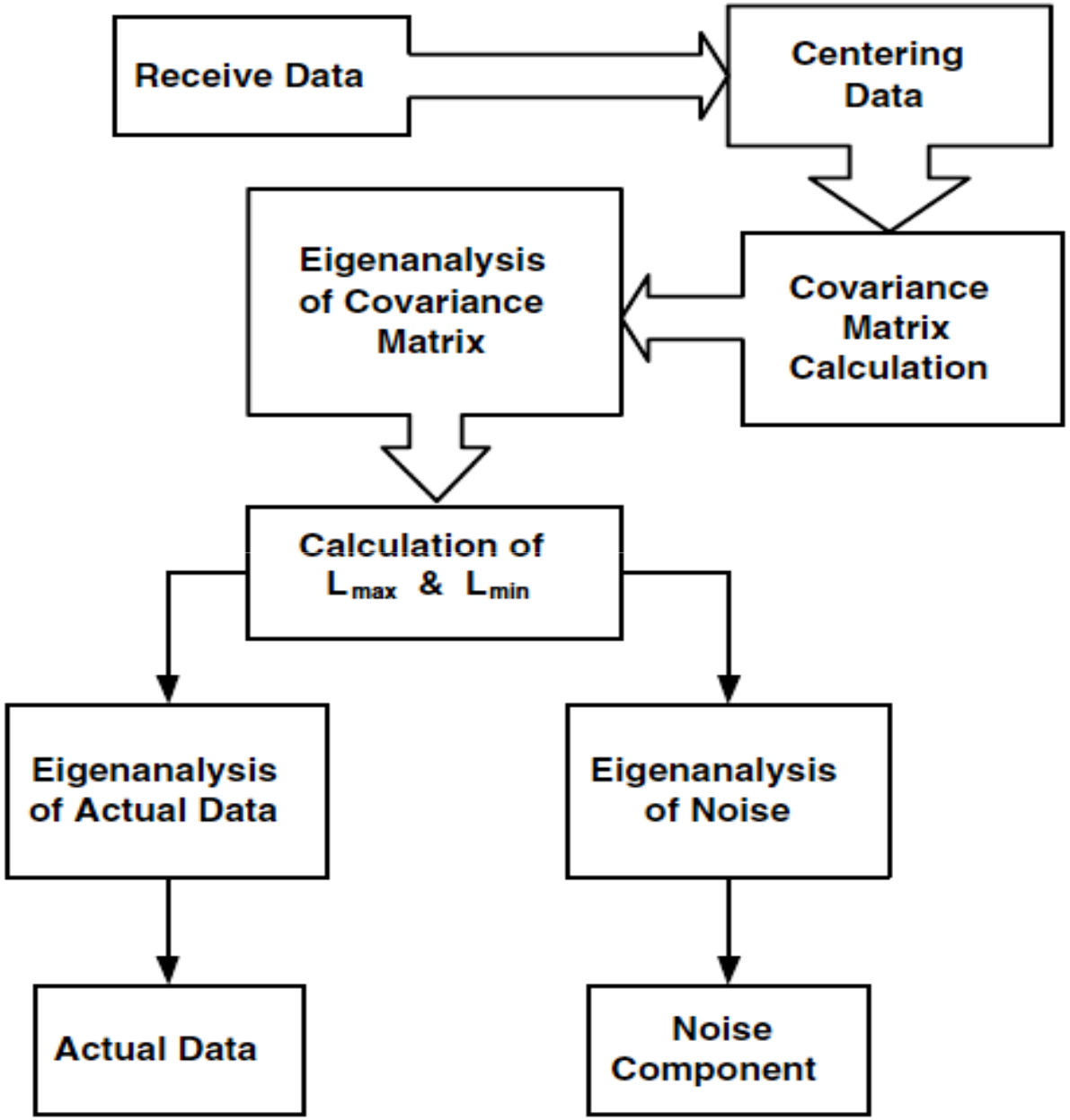
- Dimensionality and Randomization
  - Assume a *d*-dimensional record

    $\mathbf{X} = \{x_1\ x_2\ \ldots\ x_n\}$
  - Perturbed record

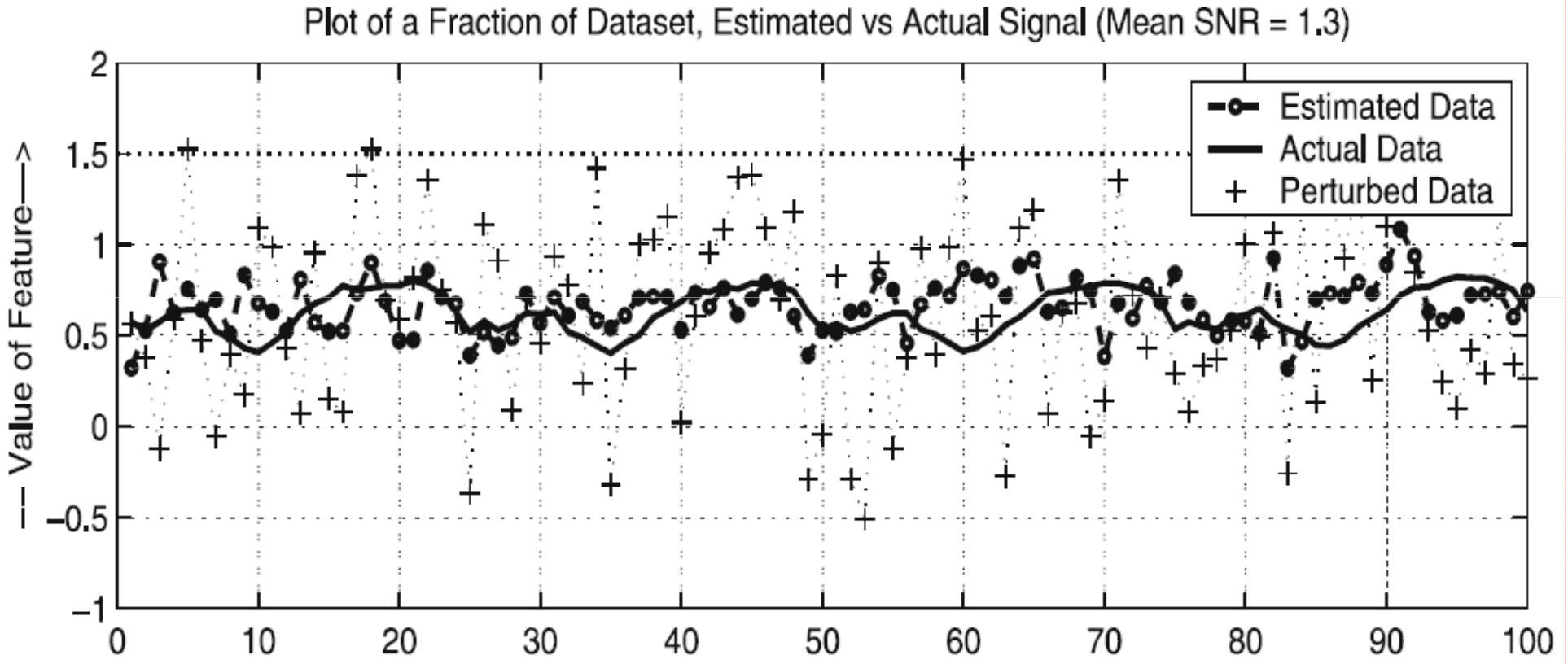
    $\mathbf{Z} = \{z_1\ z_2\ \ldots\ z_n\}$
  - For a given public record $\mathbf{W} = \{w_1\ w_2\ \ldots\ w_n\}$ we want to compute the probability that **Z** is computed from **W** using **Y**
  - Estimate the likelihood that $\{z_1–w_1, z_2–w_2, \ldots, z_n–w_n\}$ fits the distribution of **Y**
  - To preserve privacy need greater perturbations so that spurious records fit more accurately
  - The probability that a spurious record fits a perturbed data reduces with no. of dimensions

Attacks against Randomization

- The original data represents "*signal*"
- Randomization adds "*noise*"
- Use noise filtering techniques to estimate the original data values
- PCA-based analysis
- Spectral noise reduction [Kargupta 2004]
  - Randomness may not necessarily imply uncertainty
  - Randomness does have structure
  - Applies properties of random matrix
  - Effective for SNR upto 1

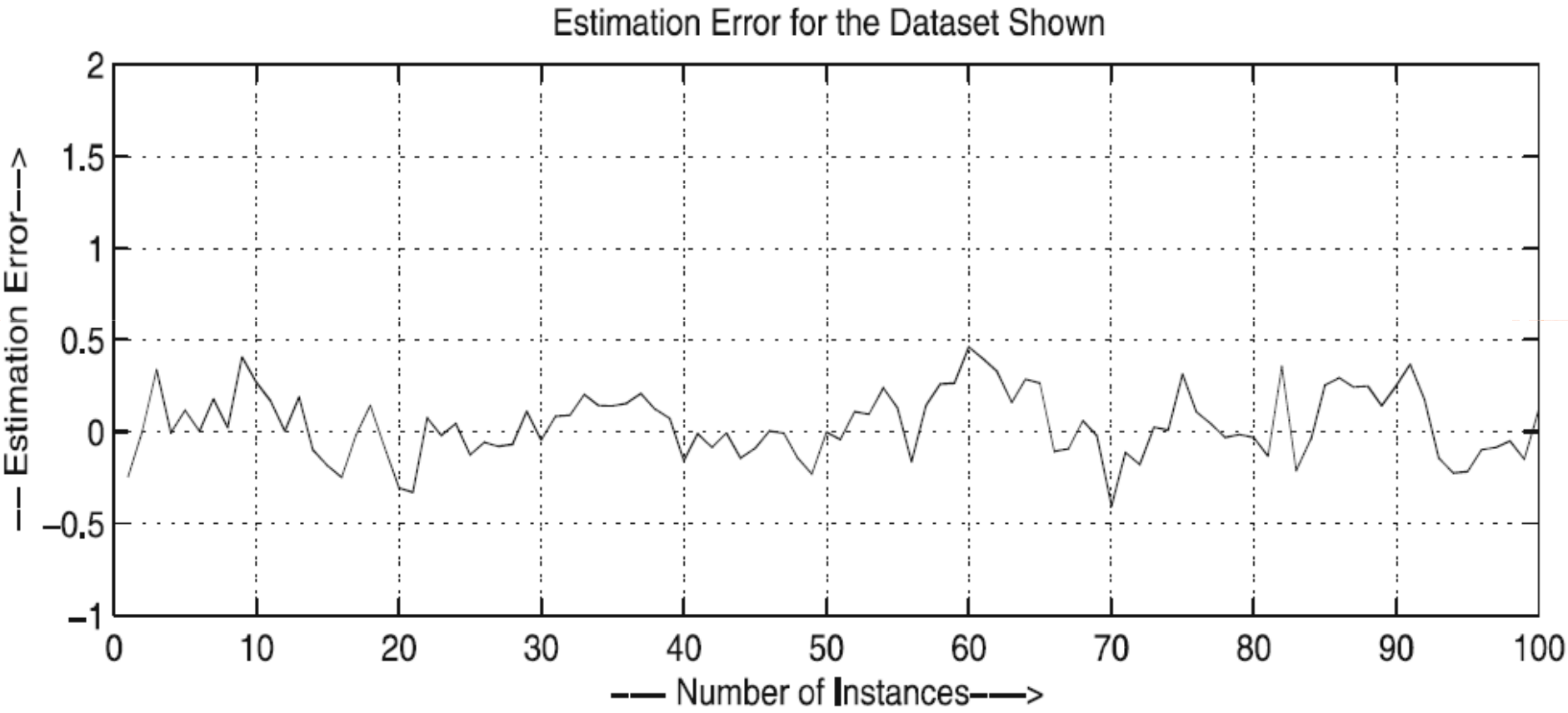

Receive Data
Centering Data
Covariance Matrix Calculation
Eigenanalysis of Covariance Matrix
Calculation of L_{max} & L_{min}
Eigenanalysis of Actual Data
Eigenanalysis of Noise
Actual Data
Noise Component

Attacks against Randomization



Kargupta 2004

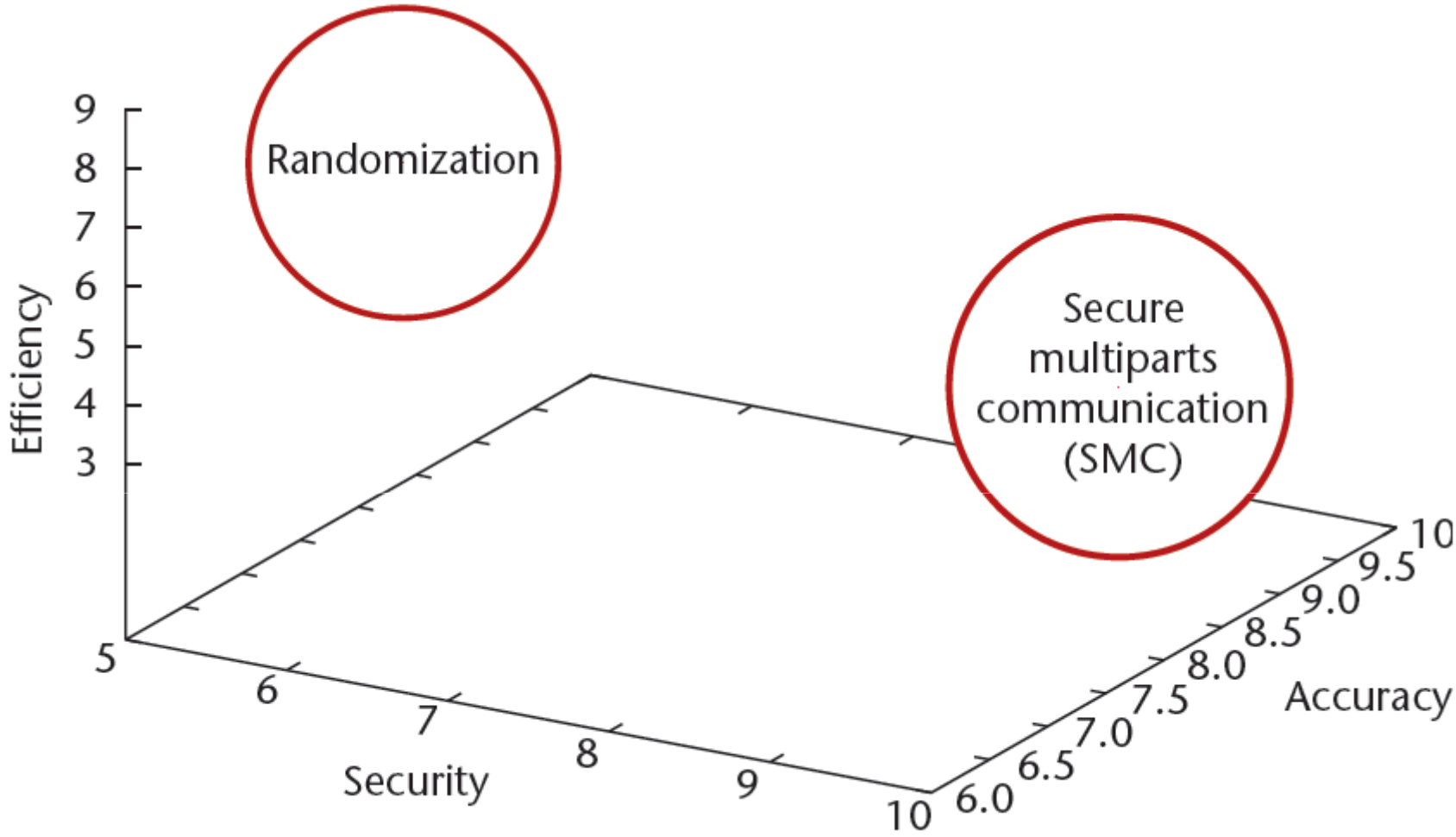
Attacks against Randomization



Kargupta 2004

Advantages of Randomization

- Noise is independent of data
- Do not need entire dataset for perturbation
- Can be applied at data collection time
- Does not require a trusted server with all the records
- Much faster compared to SMC



Clifton 2004

Applications of Privacy-Preserving Data Mining

- Medical Databases
  - Datafly system
- Protection against Bioterrorism
  - Rapidly detect outbreak of anthrax against common respiratory diseases
- Homeland Security
  - Watchlist problem
- Genomic Privacy

Questions

REFERENCES

1. Jaideep Vaidya, Chris Clifton, "Privacy-Preserving Data Mining: Why, How, and When," *IEEE Security and Privacy*, vol. 2, no. 6, pp. 19-27, Nov. 2004
2. Charu C. Aggarwal and Philip S. Yu. "A General Survey of Privacy-Preserving Data Mining Models and Algorithms," In *Privacy-Preserving Data Mining: Models and Algorithms,* P. 11-52, 2008.
3. Charu C. Aggarwal and Philip S. Yu. "A Survey of Randomization Methods for Privacy-Preserving Data Mining," In *Privacy-Preserving Data Mining: Models and Algorithms,* P. 137-156, 2008.
4. Agrawal, R. and Srikant, R. 2000. "Privacy-preserving data mining." *SIGMOD Rec.* 29, 2 (Jun. 2000), 439-450.
5. Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. 2005. "Random-data perturbation techniques and privacy-preserving data mining." *Knowl. Inf. Syst.* 7, 4 (May. 2005), 387-414.
6. Evfimievski, A. 2002. "Randomization in privacy preserving data mining." *SIGKDD Explor. Newsl.* 4, 2 (Dec. 2002), 43-48.