Distributed Data Mining: Current Pleasures and Emerging Applications

Hillol Kargupta

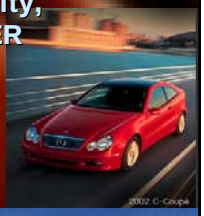
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250, USA
http://www.cs.umbc.edu/~hillol
hillol@cs.umbc.edu
&
AGNIK, LLC
Columbia, MD 21045
http://www.agnik.com
hillol@agnik.com

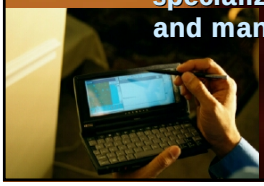
Roadmap

- Introduction
- Distributed Data Mining (DDM): An Overview
- DDM Applications
- A Taste of Algorithmic Issues
- Conclusions

Research & Development at UMBC DIADIC Laboratory and AGNIK, LLC

- **Distributed and mobile data mining.**
- **Supported by Department of Homeland Security, NASA, US National Science Foundation CAREER award and other grants, US Air Force, TRW Research Foundation, Maryland Technology Development Council, and others.**
- **Agnik, LLC: A Spin-off from DIADIC Lab, specializing on mobile and distributed data mining and management.**





Data Mining and Distributed Data Mining (DDM)

- Data Mining: Scalable analysis of data by paying careful attention to the resources:
  - computing,
  - communication,
  - storage, and
  - human-computer interaction.
- Distributed data mining (DDM): Mining data using distributed resources.

Early Days of the Community

- ACM SIGKDD Workshop on Distributed Data Mining, 1998.
- ACM SIGKDD Workshop on Distributed Data Mining, 2000.
- PKDD Workshop on Ubiquitous Data Mining for Mobile and Distributed Environments, 2001.
- SIAM International Data Mining Conference Workshop on High Performance and Distributed Mining (2001, 2002, 2003, 2004, 2005, 2006)

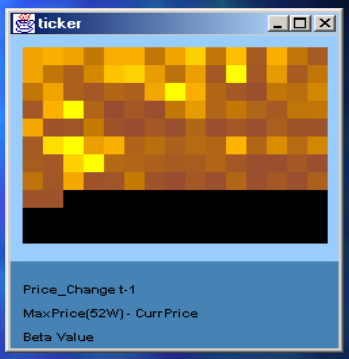
Data Mining in Distributed and Mobile Environments

- **Mining Databases from distributed sites**
  - Earth Science, Astronomy, Counter-terrorism, Bioinformatics
- **Monitoring Multiple time critical data streams**
  - Monitoring vehicle data streams in real-time
  - Monitoring physiological data streams
- **Analyzing data in Lightweight Sensor Networks and Mobile devices**
  - Limited network bandwidth
  - Limited power supply
- **Preserving privacy**
  - Security/Safety related applications
- **Peer-to-peer data mining**
  - Large decentralized asynchronous environments

Early Applications

- Work on multi-agent learning, ensemble learning
- Columbia University---Meta-learning-based system for distributed intrusion detection, Sal Stolfo, 1997.
- Los Alamos National Laboratory, PADMA system for distributed text data mining, Kargupta, 1996.

MobiMiner: A Mobile Data Stream Mining System for Stock Market Data

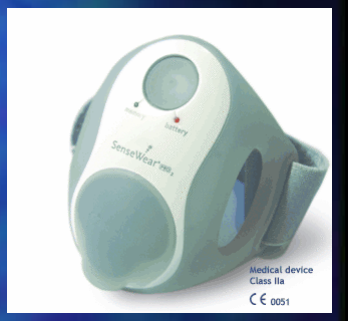


- An interactive PDA-based data mining system for stock-market data (IEEE TKDE, Kargupta et al., 2000)
- Visualize the spectrum of the decision tree

Resource-Constrained Real-time Physiological Data Stream Monitoring

- Wearable sensors available in the market
  - SenseWear Armband from BodyMedia
  - Wearable West[1]
  - LifeShirt Garment from Vivometrics
- SenseWear armband can measure heat flux, accelerometer, galvanic skin response, skin temperature, near body temperature
- Arm band can store up to about 5 days of data.

1. www.smartextiles.info



http://www.armband.it/

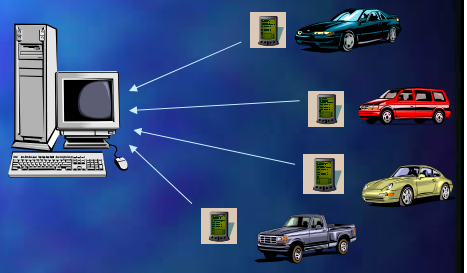
http://www.vivometrics.com

A Network of Physiological Data Stream Monitoring Devices

Detecting emerging patterns in a group of health workers, soldiers, elderly individuals, animals.

MineFleet: A Vehicle Data Stream Management and Mining Software System

- On-board Module:
  - Continuous data streams from the vehicle data bus
  - Onboard data stream mining
  - Communicates with a remote control station
  - Privacy management
- Central control station:
  - Data Management
  - Data mining
  - Communicates with the on-board modules over wireless networks
  - Privacy management



Funded by US Air Force. A commercial product to be released in Q1, 2006.

Data Collection Module: Components

Optional GPS module

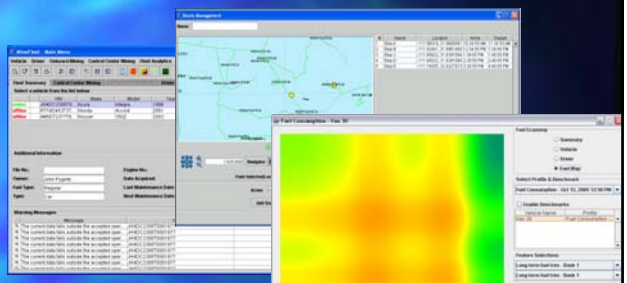
Closer Look

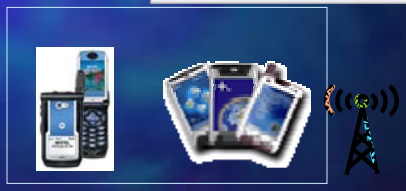
OBD-II adapter connected to a Ford Van OBD-II port

Power supply

Modes of Operations

- Passive Mode: Collect data and analyze later offline.
- Active Mode: Analyze data in real-time either on-board (cell-phone, PDA, embedded device) or remote desktop connected over wireless network.





Onboard devices

Vehicle Data Stream Mining

- **Vehicle Health Monitoring and Maintenance:**
  - Several model and data driven fault-tests
  - Detecting unusual behavior for a subsystem and accessing the data producing this behavior
- **Fuel Consumption Analysis:**
  - Is the vehicle burning fuel efficiently? Identify influencing factors and optimize
  - Detect influence of driver behavior on gas mileage and eliminate inefficient driving practices
- **Driver Behavior Monitoring:**
  - Route monitoring: Fixed and variable routes
  - Direct Cost Issues: e.g. Idling, braking habits
  - Safety Issues: e.g. speeding, trajectory monitoring (e.g. stopping, turns)
- **Vehicle location related services**
- **Vehicular network security and privacy management**

washingtonpost.com

Hackers Target U.S. Power Grid

Government Quietly Warns Utilities To Beef Up Their Computer Security

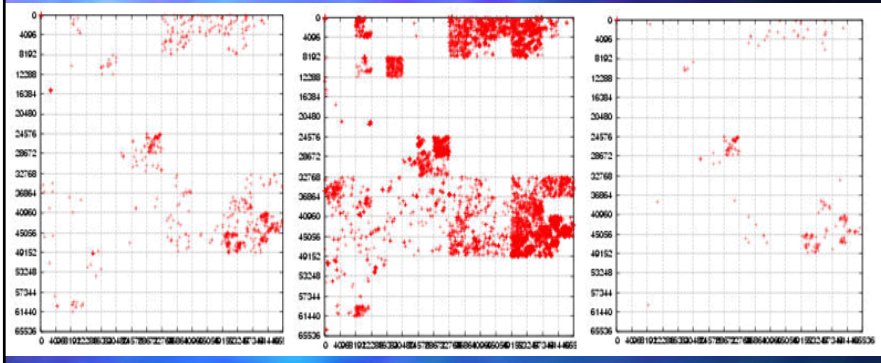
By Justin Blum
Washington Post Staff Writer
Friday, March 11, 2005; Page E01

Hundreds of times a day, hackers try to slip past cyber-security into the computer network of Constellation Energy Group Inc., a Baltimore power company with customers around the country.

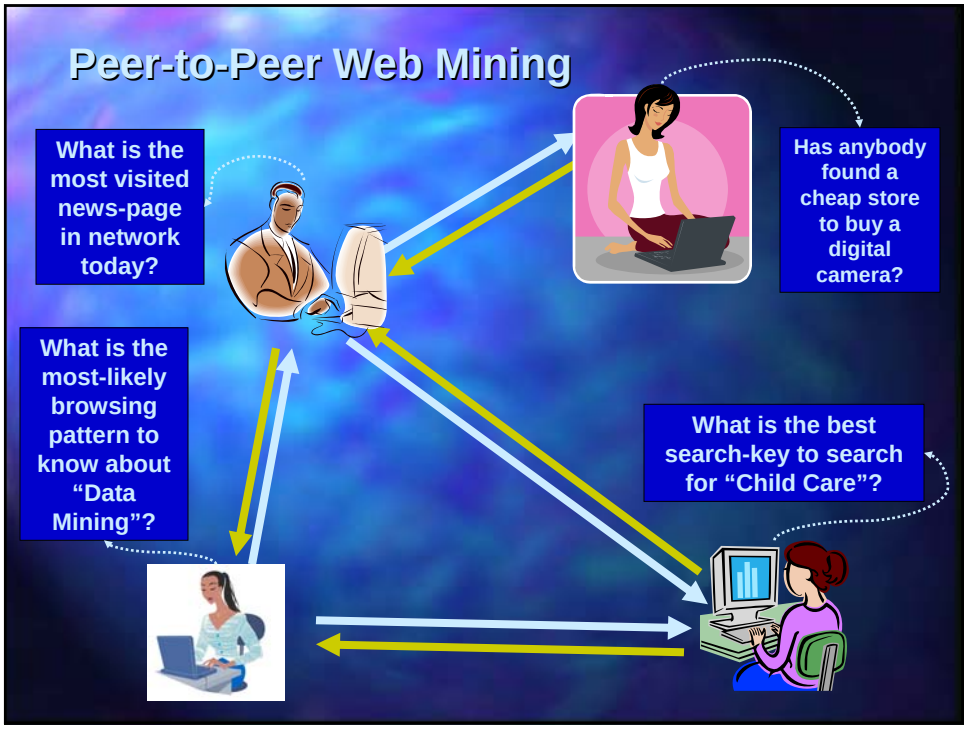
"We have no discernable way of knowing who is trying to hit our system," said John R. Collins, chief risk officer for Constellation, which operates Baltimore Gas and Electric. "We just know it's being hit."

PURSUIT: Privacy-Sensitive Cross-Domain Intrusion Detection

- Cross-Domain Network Attack Detection system using Privacy-Preserving Distributed Data Mining
  - Detecting stealth attacks
  - Identifying botnets
  - Identifying cross-domain attack patterns, worm classification
- Sponsor: US Department of Homeland Security
- Partners:
  - Agnik, Army High Performance Research Center, University of Minnesota, and Tresys Inc.
- PURSUIT Consortium:
  - Purdue University
  - Ohio State University
  - Stevens University
  - SRI International
  - University of Illinois at Urbana-Champaign



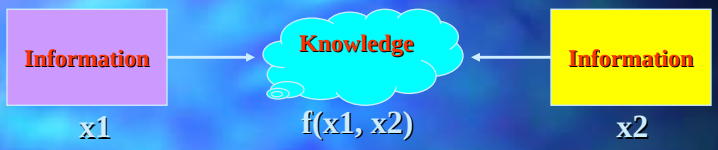
- Spatial Attack Distribution of IPs on the Same Day: (Left) IPs attacking the UFL network on 12/09/04 (712 scanners). (Middle) IPs attacking the UMN network on 12/09/04 (14,938 scanners). (Right) Intersection of the IPs attacking UFL and UMN (201 scanners). *Courtesy: Vipin Kumar, UMN*



Useful Browser Data

1. Web-browser history
2. Browser cache
3. Click-stream data stored at browser (browsing pattern)
4. Search queries typed in the search engine
5. User profile
6. Bookmarks

DDM Algorithms

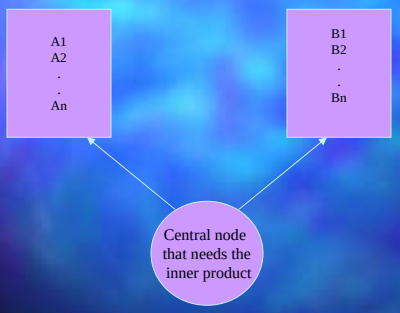


- Hundreds of DDM algorithms exist
  - Distributed association rule learning
  - Distributed PCA and other dimension reduction techniques
  - Distributed clustering and Bayesian network learning
  - Distributed multi-variate regression and other statistical algorithms
  - Distributed construction of ensemble models
  - http://www.cs.umbc.edu/~hillol/DDMBIB
- Environments:
  - Homogeneous Sites: Sites observing a common set of features.
  - Heterogeneous Sites: Sites observing different feature sets.

Functions and Inner Products

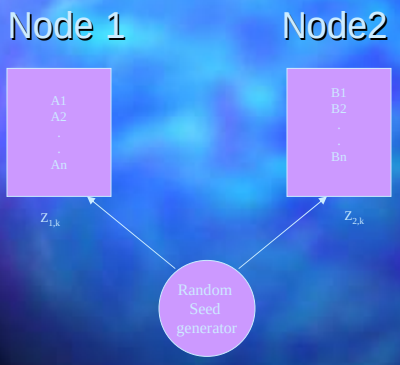
- Function representation using Inner products
- Inner product is a useful primitive
  - Correlation matrix and Euclidean distance computation
  - Clustering
  - Principal component analysis
  - Decision tree construction
  - Bayesian network construction
- Computing Inner Product
  - Deterministic
  - Probabilistic

Inner Product Computation: Deterministic Techniques



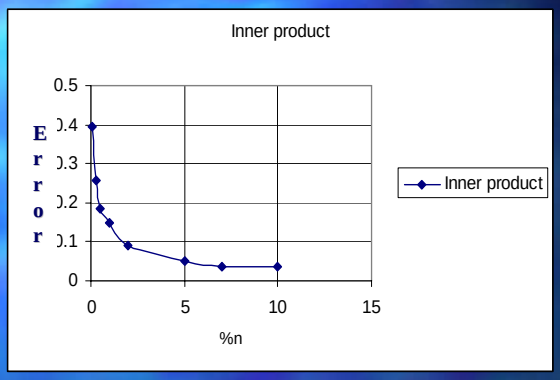
- Exact computation
- Orthogonal transformations
  - Fourier
  - Wavelet
  - Eigenvectors
- Transform the data vectors, communicate the sufficient statistics, compute the inner product

Inner Product Computation: A Probabilistic Technique



- Node 1 computes $Z_{1,k}$
  - $Z_{1k}=A1.J_1+..+An.J_n$
  - $J_i \in \{+1,-1\}$ with uniform probability
- Node 2 calculates $Z_{2,k}$
  - $Z_{2k}=B1.J_1+..+Bn.J_n$
- Compute $z_{1,k}.z_{2,k}$ for a few times and take the average

Inner Product Results



- Variation of the mean relative error in distributed inner product with respect to *n* (in percentage) using uniformly distributed data [0,1]. The mean is computed over 10 independent runs.

Inner Products: An Ordinal Approach

- Not interested in the value of the inner products
- Find the ones that rank high

Continued

- $A_1, A_2, \cdots A_n$ be n random samples; Cdf F(x);
- Bound the probability

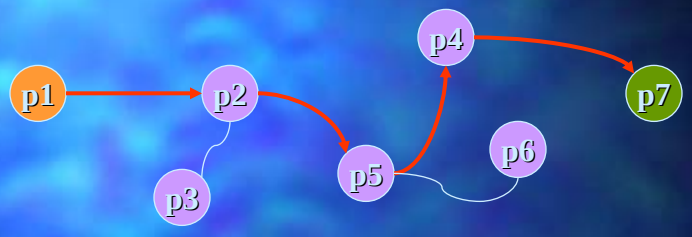
$$A_{[1]} < A_{[2]} < \cdots < A_{[n]} \qquad P(A_{[n]} > \zeta_p) > q$$

$$1 - p^n > q$$

$$n > \frac{\lg(1-q)}{\lg p}$$

p=0.95, q=0.95
n>=59

Ordinal Sampling in a P2P Environment

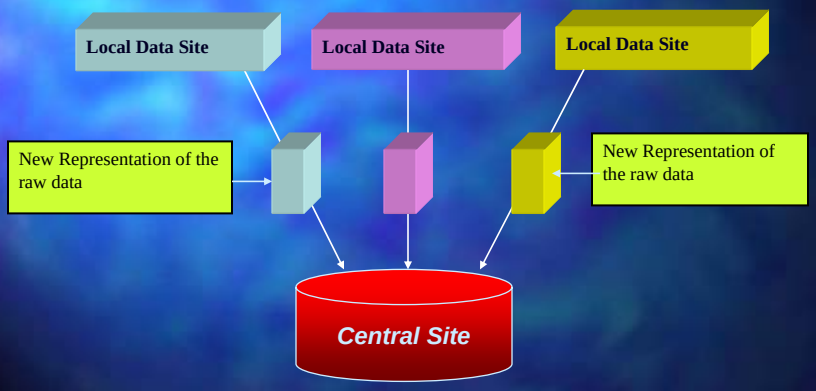


- Node p_i contains a data vector x_{p_i}
- Initiate a random walk, select p_7 and compute the inner product between x_{p_1} and x_{p_7}
- Metropollis-Hasting algorithm for random walk

Blending Privacy-Preserving Techniques

- Data sanitization
- Random perturbation (Agrawal and Srikant, 2001)
- Random multiplicative noise
- Secured Multi-Party Computation (Goldreich, 1998)
- K-Anonymity (Sweeney, 2002)
- K-Ring of Privacy (Kargupta, et al., 2005)

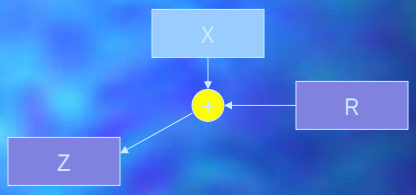
Data Perturbation-based Approach: The Idea



Random Additive Noise

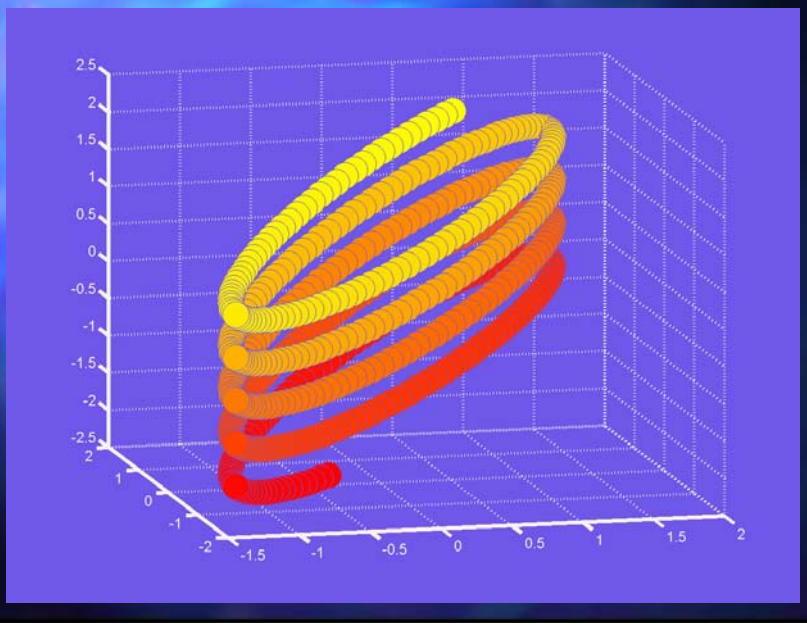
- Perturbed Data (U1) = Original Data (U) + Noise (R)
- Entries of noise matrix R are i.i.d.
- References:
  - Agrawal and Srikant, SIGMOD, 2000
  - Evfimievski, December, 2002 SIGKDD Explorations
  - Evfimievski, Srikant, Agrawal, Gehrke, ACM SIGKDD Conference, 2002
  - Rizvi and Haritsa, 2002
  - Others....

Random Additive Perturbation

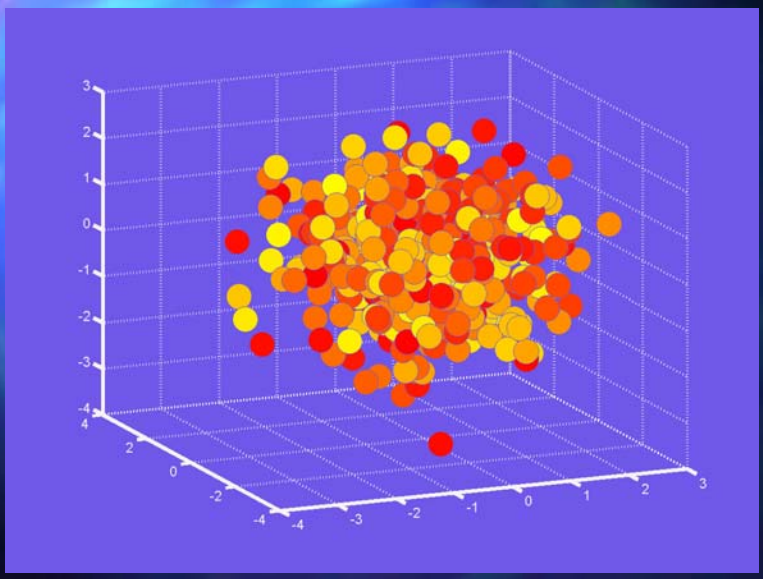


- Given m×n dimensional data set X and
- m×n dimensional noise matrix with i.i.d. entries.
- Compute the perturbed data Z, where Z = X + R
- Release Z to the data miner for estimating patterns.
- Agrawal and Srikant, 2001.

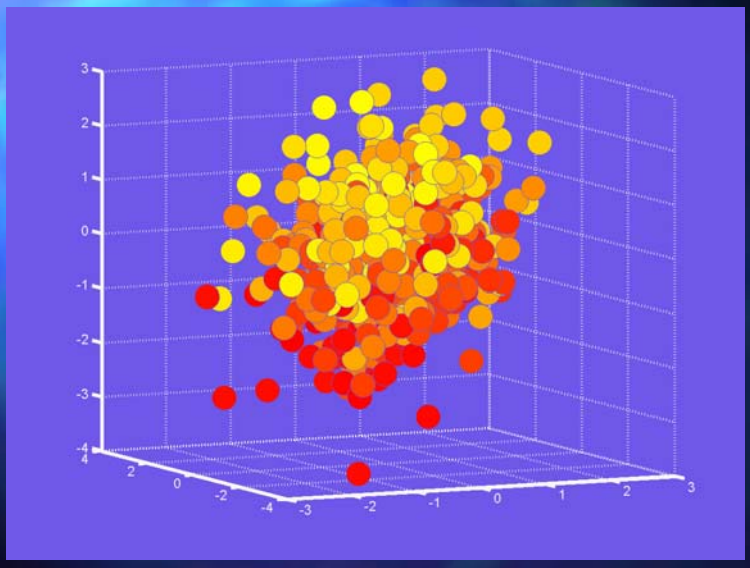
Structured Data



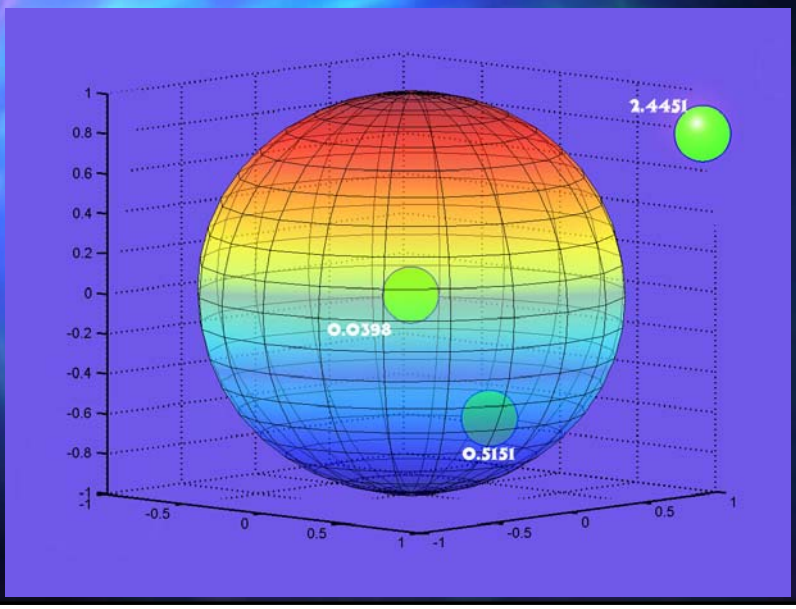
Random Noise



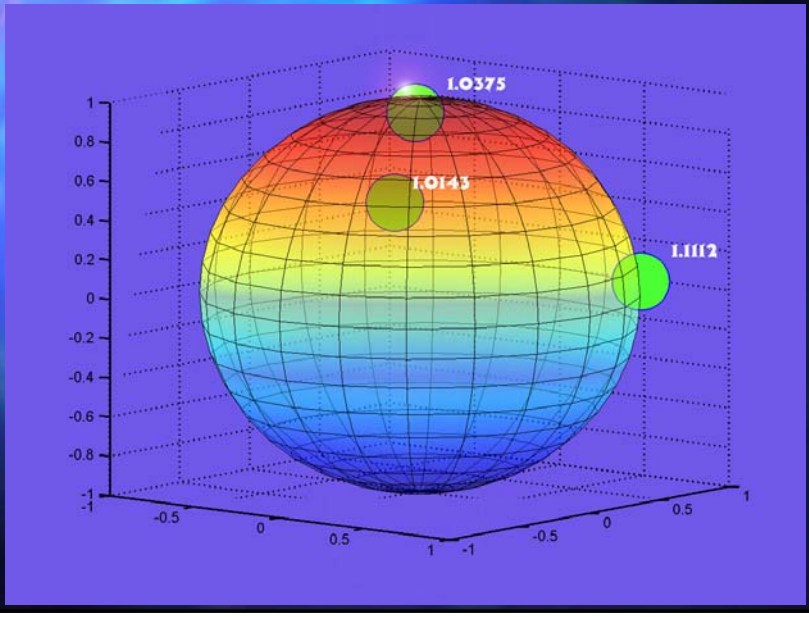
Perturbed Version (Data + Noise)



Eigenstates of Original Data



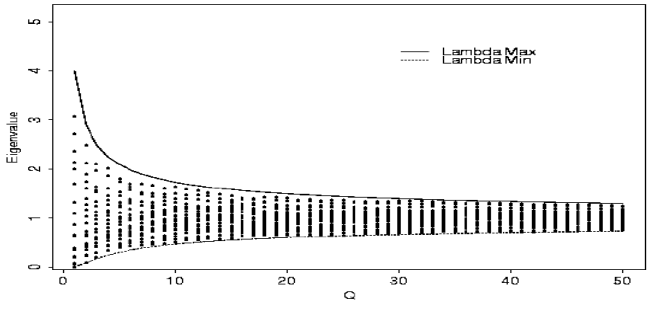
Eigenstates of the Random Noise



Eigenvalues of Random Matrices

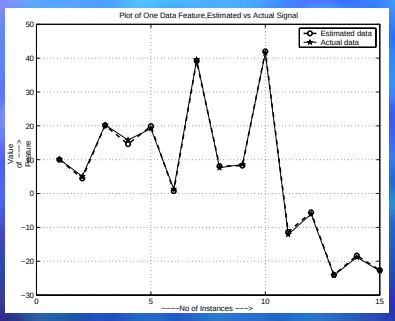
Bounds of the eigenvalues:

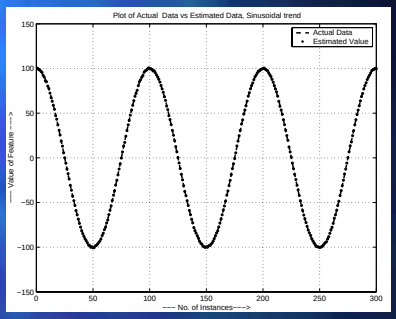
$$\lambda_{\min} = \sigma^2(1 - 1/\sqrt{Q})^2.$$
$$\lambda_{\max} = \sigma^2(1 + 1/\sqrt{Q})^2.$$



Q=number of rows/number of columns.
σ^2 = variance of entries in the noise matrix.

Random Value Perturbation & Eigenvalues of Random Matrices





- Spectral filtering based on properties of eigenvalues of random matrices.
- **Random additive perturbation may not preserve a whole lot of privacy in many cases**

References

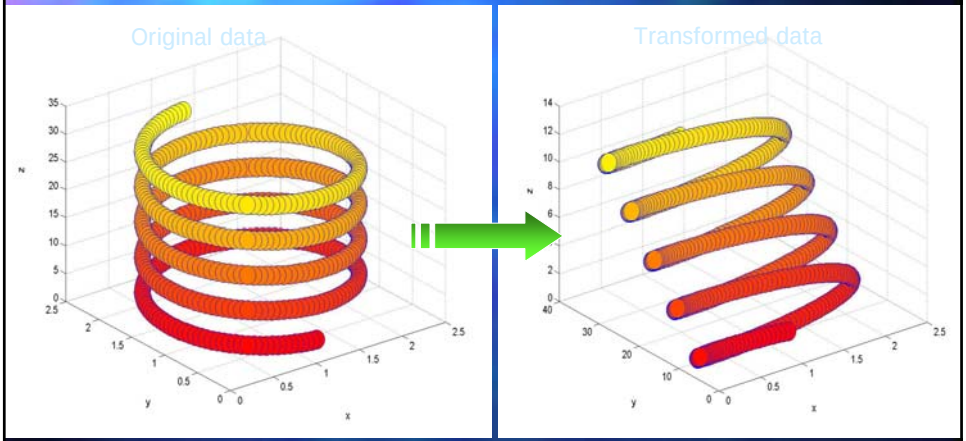
- D. Meng, K. Sivakumar, and H. Kargupta. (2004). Privacy Sensitive Bayesian Network Parameter Learning. Proceedings of the Fourth IEEE International Conference on Data Mining. Brighton, UK, pages 427-430.

- H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. (2005). Random Data Perturbation Techniques and Privacy Preserving Data Mining. *Knowledge and Information Systems Journal*, volume 7, number 4, pages 387--414.

Multiplicative Noise

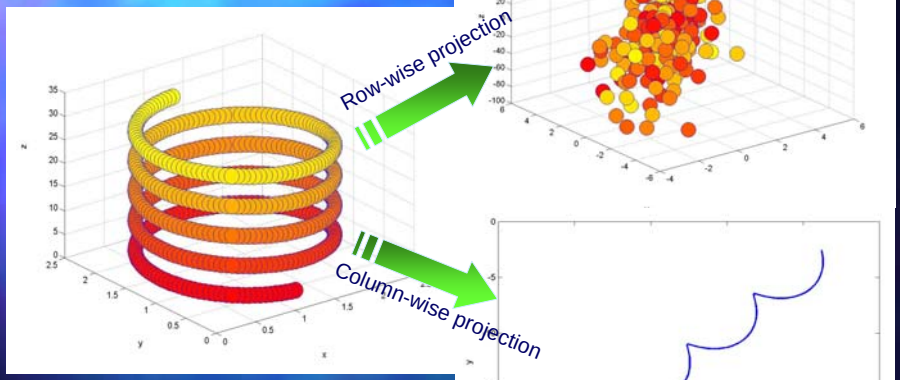
- Perturbed Data (U1) = Original Data (U)*Noise (R)
- U1 = U R
- Can U1 be used for privacy preserving applications?

Random Orthogonal Transformation



Preserves inner product and Euclidean distance.

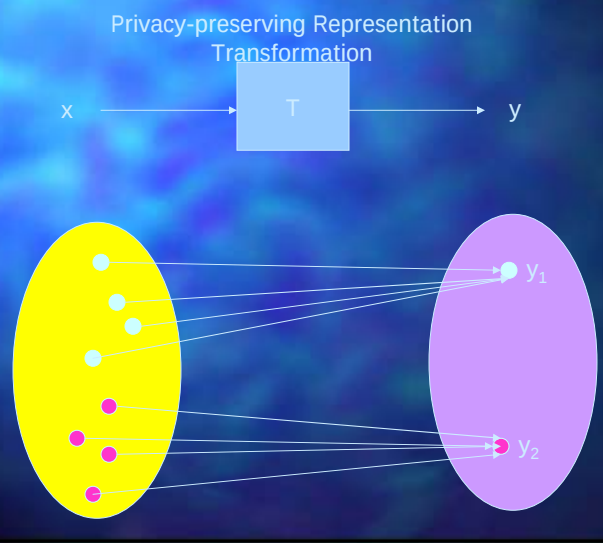
Random Projection



Reference

- K. Liu, H. Kargupta, and J. Ryan. (2005). Multiplicative Noise, Random Projection, and Privacy Preserving Data Mining from Distributed Multi-Party Data. Accepted for publication in the *IEEE Transactions on Knowledge and Data Engineering*. (In Press)

K-Ring of Privacy



Formal Definition

$S_T = \{(x_i, y_i)\}$ *and* $X_{y_i} = \{x_i \mid (x_i, y_i) \in S_T\}$

$k = \min_i \| X_{y_i} \|$

If for all y_i we can guarantee $\frac{P[y_i \mid x_1]}{P[y_i \mid x_2]} \leq \gamma \; \forall x_1, x_2 \in X_{y_i}$

then transformation T offers a (k, γ) – Ring of Privacy

A Simplified Definition

For all y_i we guarantee $P[y_i \mid x_1] = P[y_i \mid x_2] \quad \forall x_1, x_2 \in X_{y_i}$

Therefore, $\gamma = 1$

- A Two-Channel Plan
  - Noise free pattern-channel
  - Noisy channel privacy-presrvation

A Functionally Complete Representation

- Consider a basis set
- A target function

$$G(\mathbf{x}) = \sum_j W_j \Psi_j(\mathbf{x})$$

$$G_X = \Psi_X W$$

Example: Multi-Variate Fourier Representation

- Fourier representation $f(x) = \sum_{k \in J} w_k \psi_k(x)$
- where,
  - J is an indexed set
  - w_k is the k-th coefficient; $w_k = \sum_x f(x) \psi_k(x)$
  - $\psi_k(x)$ is the k-th basis function.
- In binary domain
  $\psi_k(x) = (-1)^{k \cdot x}$

Random Mixing

$$G_X = \Psi_X W$$
$$= \Psi_X P P^{-1} W = (\Psi_X P)(P^{-1} W) = \Psi_X' W'$$

- Where P is a random invertible matrix
- Data owner releases Ψ_X' and W'

Illustration: Nearest Neighbor Computation

- The pair-wise similarity matrix

$$G_{x_p,x_q} = \Psi_{x_p} W \Psi_{x_q}^T$$

- Where W is a diagonal matrix

- In Fourier representation entries are from the set:

$$\left\{\frac{1}{2}, \frac{-1}{2}, \frac{n}{2}, 0\right\}$$

Random Mixing

$$\begin{aligned} G_{x_p,x_q} &= \Psi_{x_p} W \Psi_{x_q}^T = \Psi_{x_p} P P^{-1} W (P^T)^{-1} P^T \Psi_{x_q}^T \\ &= [\Psi_{x_p} P][P^{-1} W (P^T)^{-1}][P^T \Psi_{x_q}^T] \\ &= \Psi'_{x_p} W' \Psi'^T_{x_q} \end{aligned}$$

- Release Ψ'_{x_p} and W'

Example

- Two bit domain {00, 01, 10, 11}
- Multi-variate Fourier basis set

$$\Psi_x = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad W = \begin{bmatrix} \frac{n}{2} & 0 & 0 & 0 \\ 0 & \frac{-1}{2} & 0 & 0 \\ 0 & 0 & \frac{-1}{2} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Conclusions

- Increasing number of data rich distributed applications
  - Pervasive wireless environments
  - Grid
  - P2P file sharing networks
  - Cross-domain multi-organizational environments
- Interesting Algorithmic Challenges

Future Work

- Current directions of the field of DDM:
  - Resource constrained data stream management and mining
  - P2P data mining
  - Privacy preserving data mining
  - Large-scale grid-based DDM
  - Human-computer interaction issues
  - Communication & collaboration management, reasoning capabilities---Multi-agent systems