



Privacy Preserving Data Mining

Moheeb Rajab



Agenda

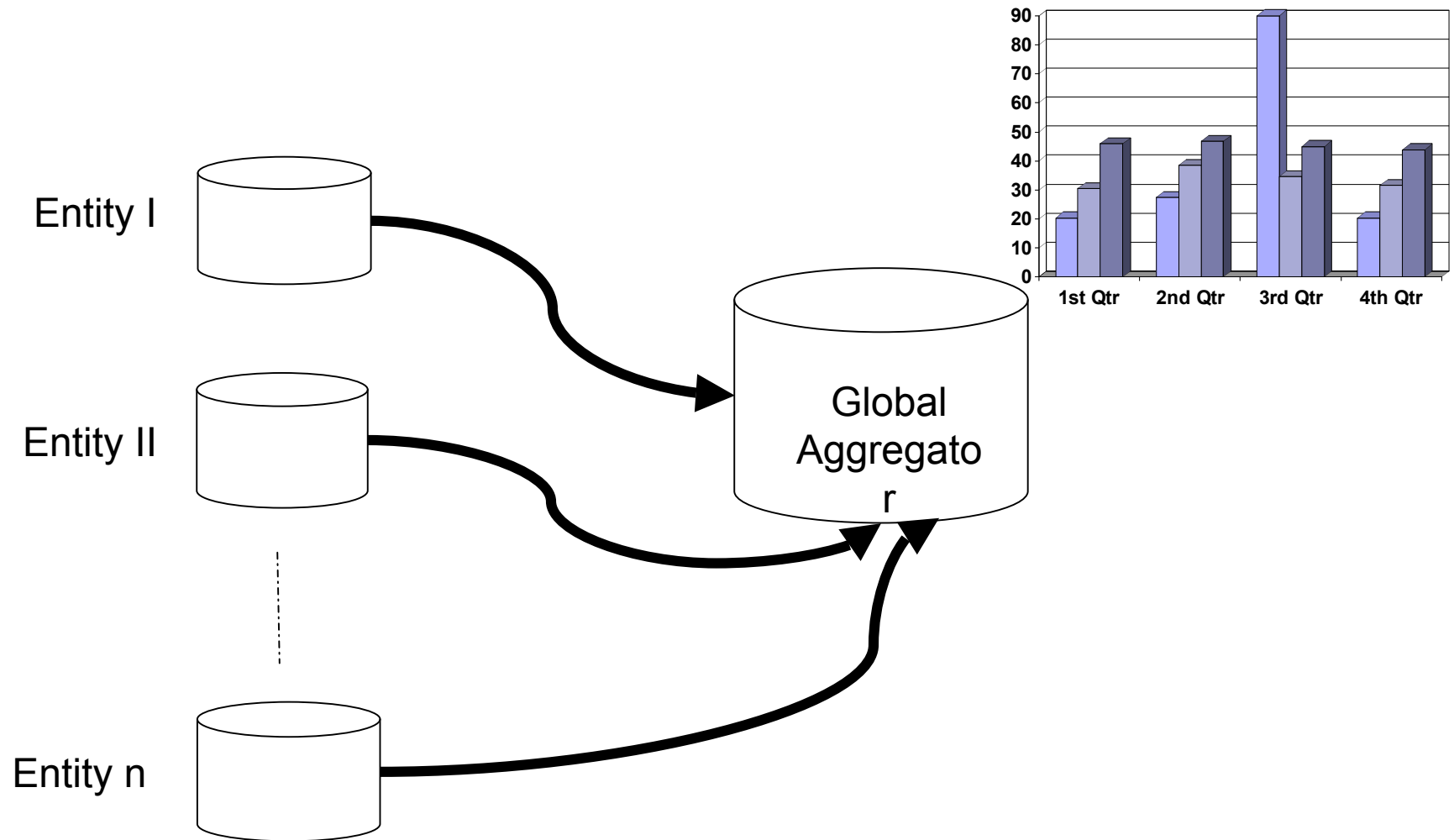
- Overview and Terminology
- Motivation
- Active Research Areas
 - Secure Multi-party Computation (SMC)
 - Randomization approach
- Limitations
- Summary and Insights



Overview

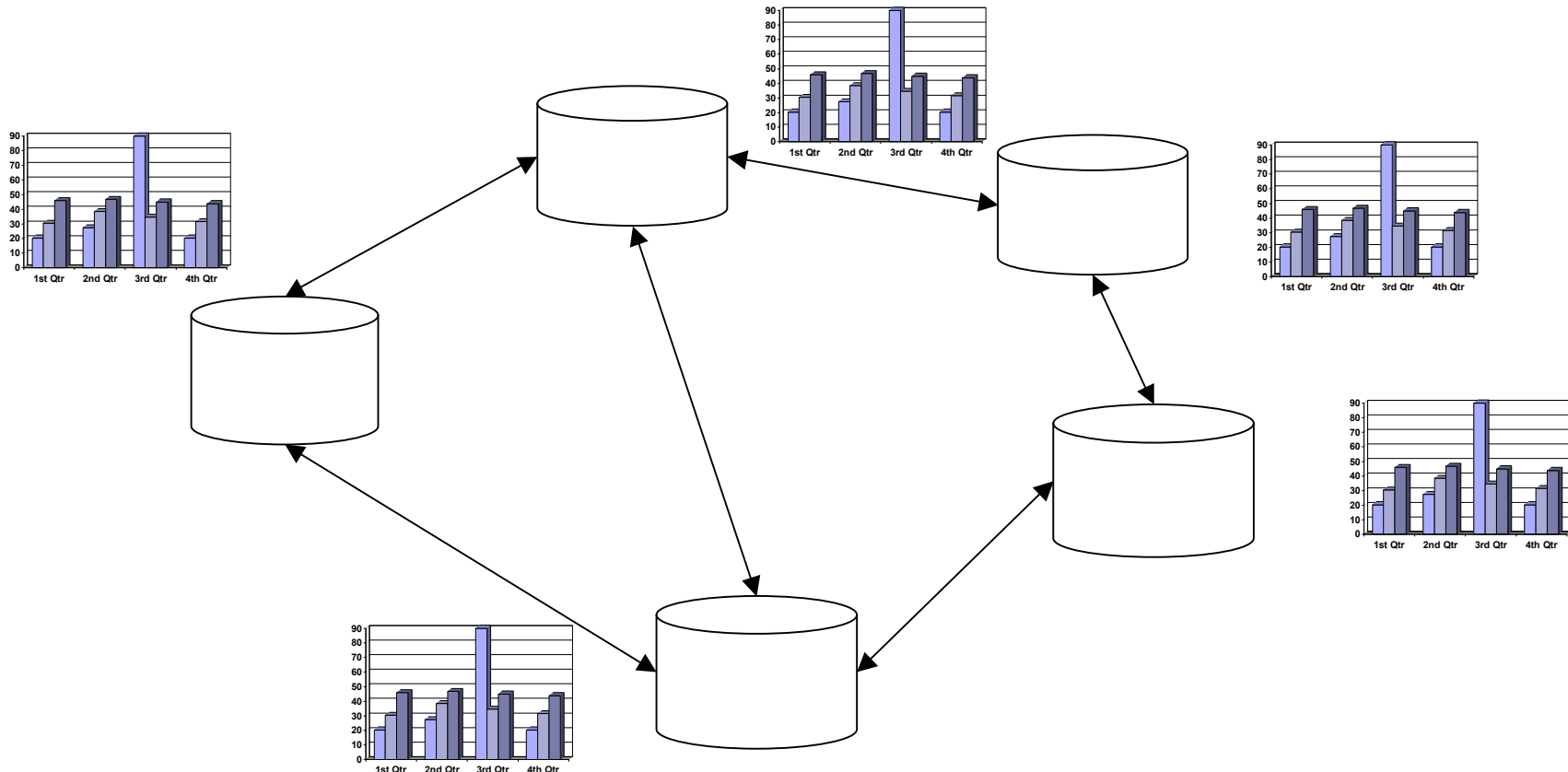
- What is Data Mining?
 - Extracting implicit un-obvious patterns and relationships from a warehoused of data sets.
- This information can be useful to increase the efficiency of the organization and aids future plans.
- Can be done at an organizational level.
 - By Establishing a data Warehouse
- Can be done also at a global Scale.

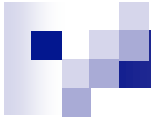
Data Mining System Architecture



Distributed Data Mining Architecture

■ Lower scale Mining





Challenges

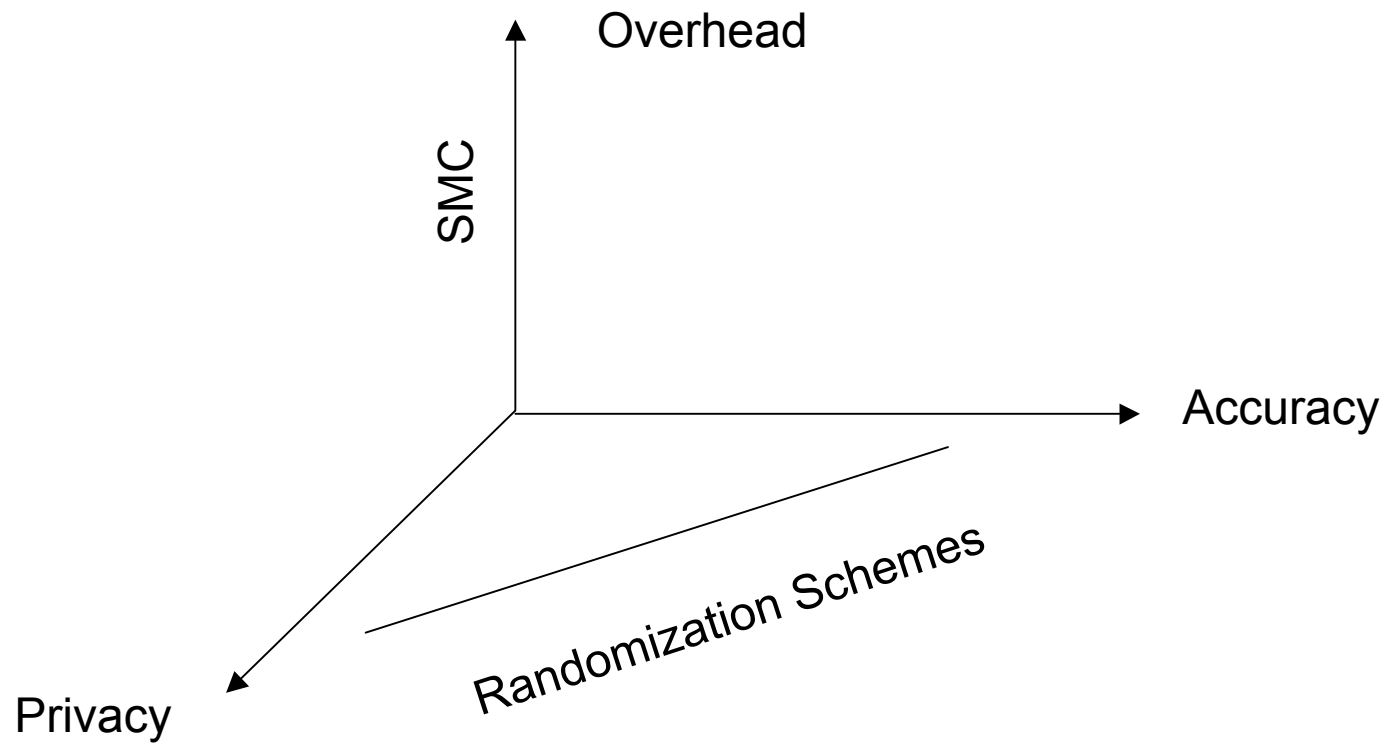
- Privacy Concerns
- Proprietary information disclosure
- Concerns about Association breaches
- Misuse of mining
- These Concerns provide the motivation for **privacy preserving data mining solutions**



Approaches to preserve privacy

- Restrict Access to data (Protect Individual records)
- Protect both the data and its source:
 - Secure Multi-party computation (SMC)
 - Input Data Randomization
- There is no such one solution that fits all purposes

SMC vs Randomization

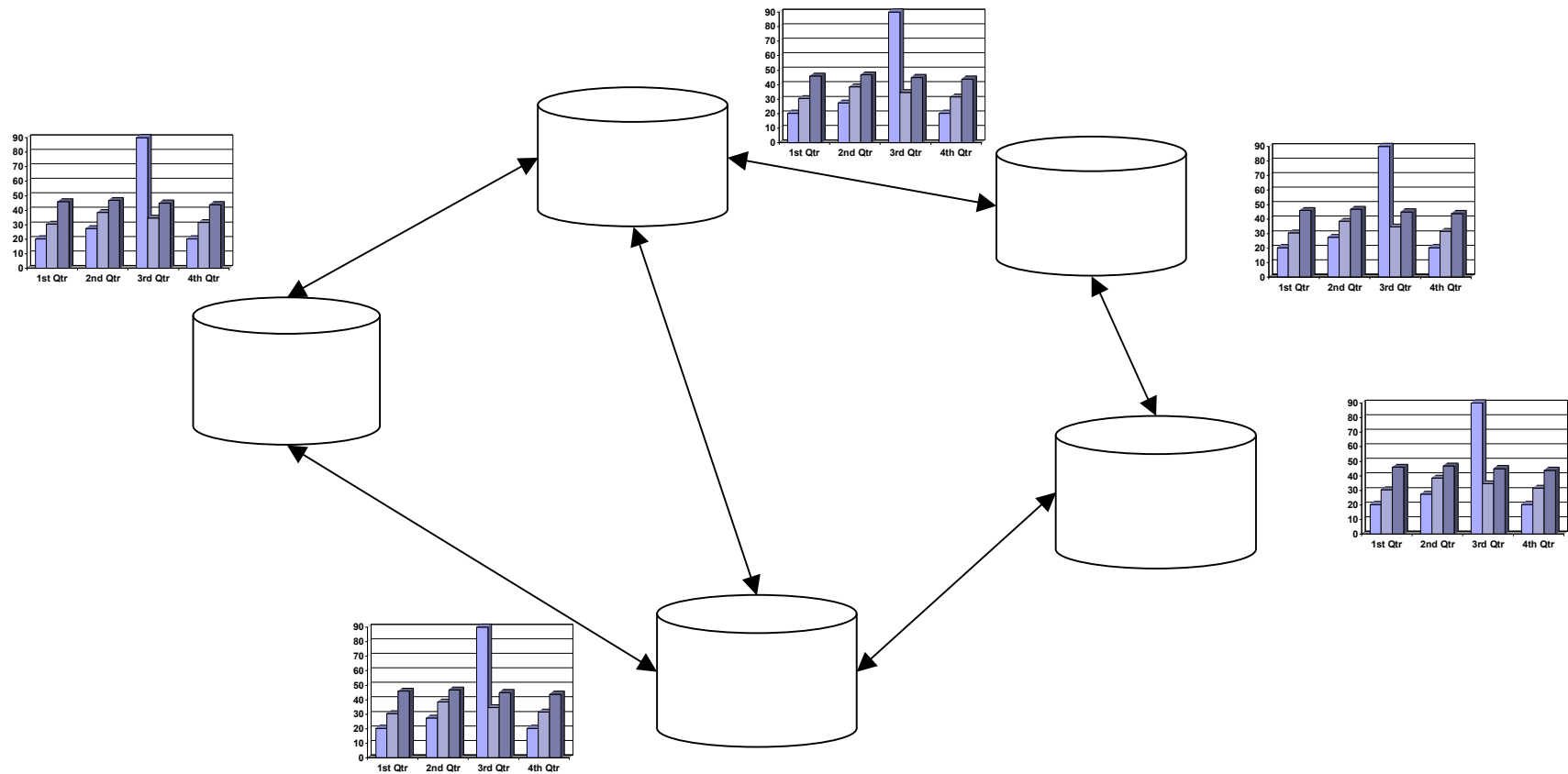




Secure Multi-party Computation

- Multiple parties sharing the burden of creating the data aggregate.
- Final processing if needed can be delegated to any party.
- Computation is considered secure if each party only knows its input and the result of its computation.

SMC



Each Party Knows its input and the result of the operation and nothing else



Key Assumptions

- The ONLY information that can be leaked is the information that we can get as an overall output from the computation (aggregation) process
- Users are not Malicious but can honestly curious
 - All users are supposed to abide to the SMC protocol
- Otherwise, for the case of having malicious participants is not easy to model! [Penkas et al, Argawal]



“Tools for Privacy Preserving Distributed Data Mining” *Clifton et al [SIGKDD]*

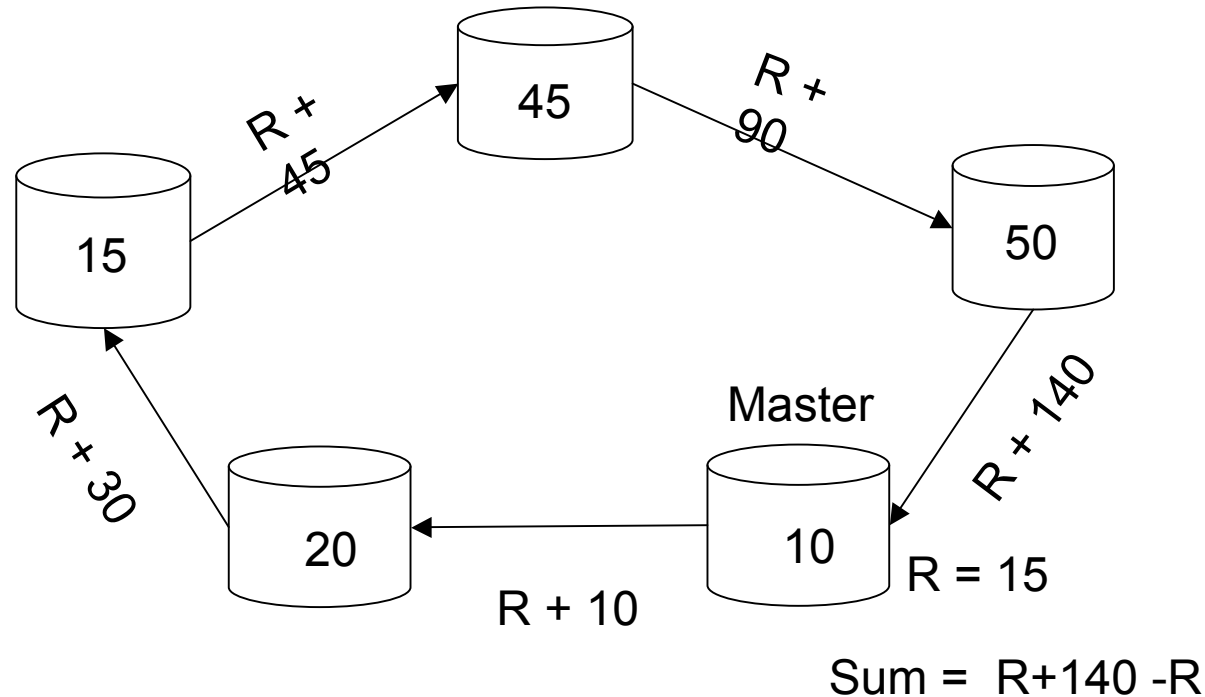
- **Secure Sum**

- Given a number of values x_1, x_2, \dots, x_n belonging to n entities

- We need to compute $\sum_{i=1}^n x_i$

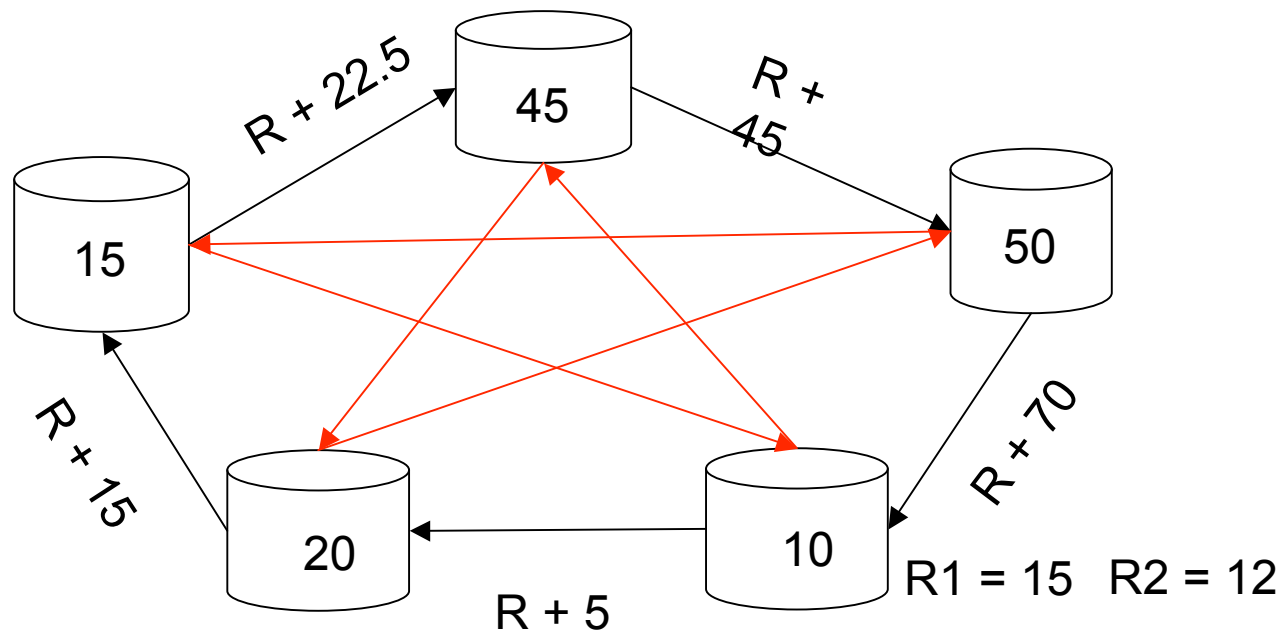
- Such that each entity ONLY knows its input and the result of the computation (The aggregate sum of the data)

Examples (Secure Sum)



- Problem:
 - Colluding members
- Solution
 - Divide values into shares and have each share permute a disjoint path (no site has the same neighbor twice)

Split path solution



$$\text{Sum} = R_1 + 70 - R_1 + R_2 + 70 - R_2 = 140$$



Secure Set Union

- Consider n sets S_1, S_2, \dots, S_n
Compute,

$$U = S_1 \cup S_2 \cup S_3, \dots, \cup S_n$$

Such that each entity **ONLY** knows U and nothing else.



Secure Union Set

- Using the properties of Commutative Encryption
- For any permutation i, j the following holds

$$E_{K_{i_1}} (...E_{K_{i_n}} (M)...) = E_{K_{j_1}} (...E_{K_{j_n}} (M)...)$$

$$P(E_{K_{i_1}} (...E_{K_{i_n}} (M_1)...) == E_{K_{j_1}} (...E_{K_{j_n}} (M_2)...)) < \varepsilon$$

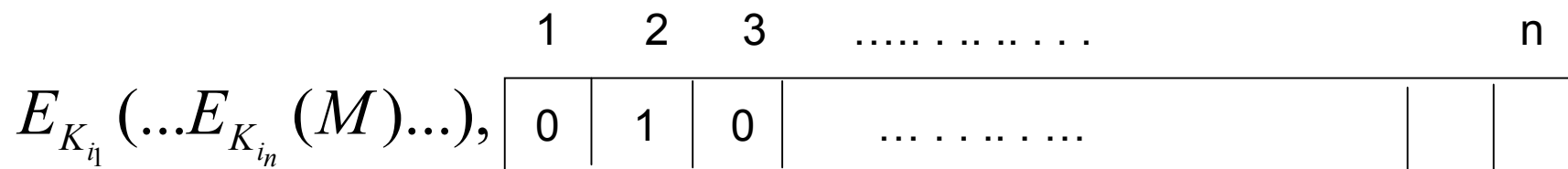


Secure Set Union

- Global Union Set U .
- Each site:
 - Encrypts its items
 - Creates an array $M[n]$ and adds it to U
- Upon receiving U an entity should encrypt all items in U that it did not encrypt before.
- In the end: all entries are encrypted with all keys K_1, K_2, \dots, K_n
- Remove the duplicates:
 - Identical plain text will result the same cipher text regardless of the order of the use of encryption keys.
- Decryption U :
 - Done by all entities in any order.

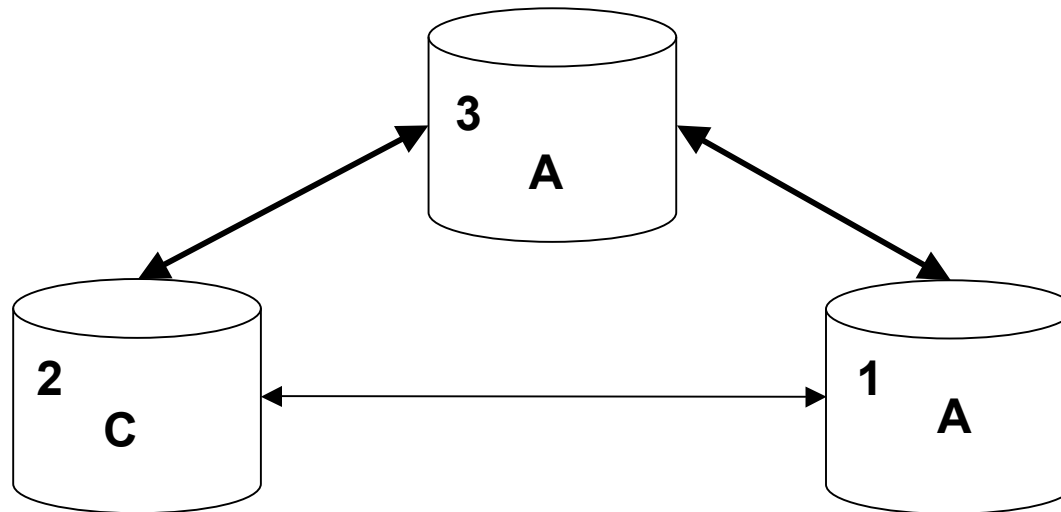


Secure Union Set





$U = \{E3(E2(E1(A))), E3(E2(C)), E3(A)\}$



$U = \{E1(A)\}$

$U = \{E2(E1(A)), E2(C)\}$

$U = \{E3(E2(E1(A))), E1(E3(E2(C))), E1(E3(A))\}$

$U = \{E3(E2(E1(A))), E1(E3(E2(C))), E2(E1(E3(A)))\}$

■ Problem:

- Computation Overhead, number of exchanged messages $O(n*m)$



Problems with SMC

- Scalability
- High Overhead
- Details of the trust model assumptions
 - Users are honest and follow the protocol



Randomization Approach

- **“Privacy Preserving Data Mining”, *Argawal et. al* [SIKDD]**
- Applied generally to provide estimates for data distributions rather than single point estimates
- A user is allowed to alter the value provided to the aggregator
- The alteration scheme should be known to the aggregator
- The aggregator estimates the overall global distribution of input by removing the randomization from the aggregate data



Randomization Approach (ctnd.)

- Assumptions:

- ☐ Users are willing to divulge some form of their data
- ☐ The aggregator is not malicious but may honestly curious (they follow the protocol)

- Two main data perturbation schemes

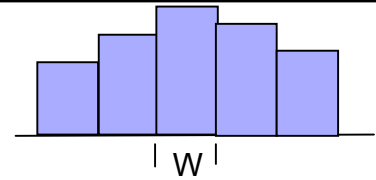
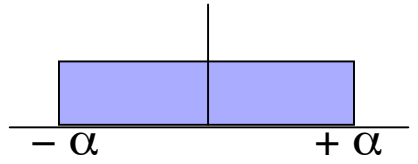
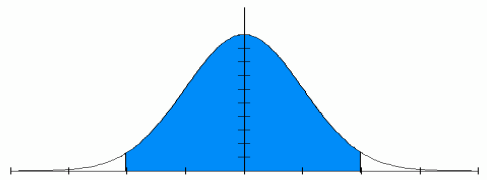
- ☐ Value- class membership (Discretization)
- ☐ Value distortion



Randomization Methods

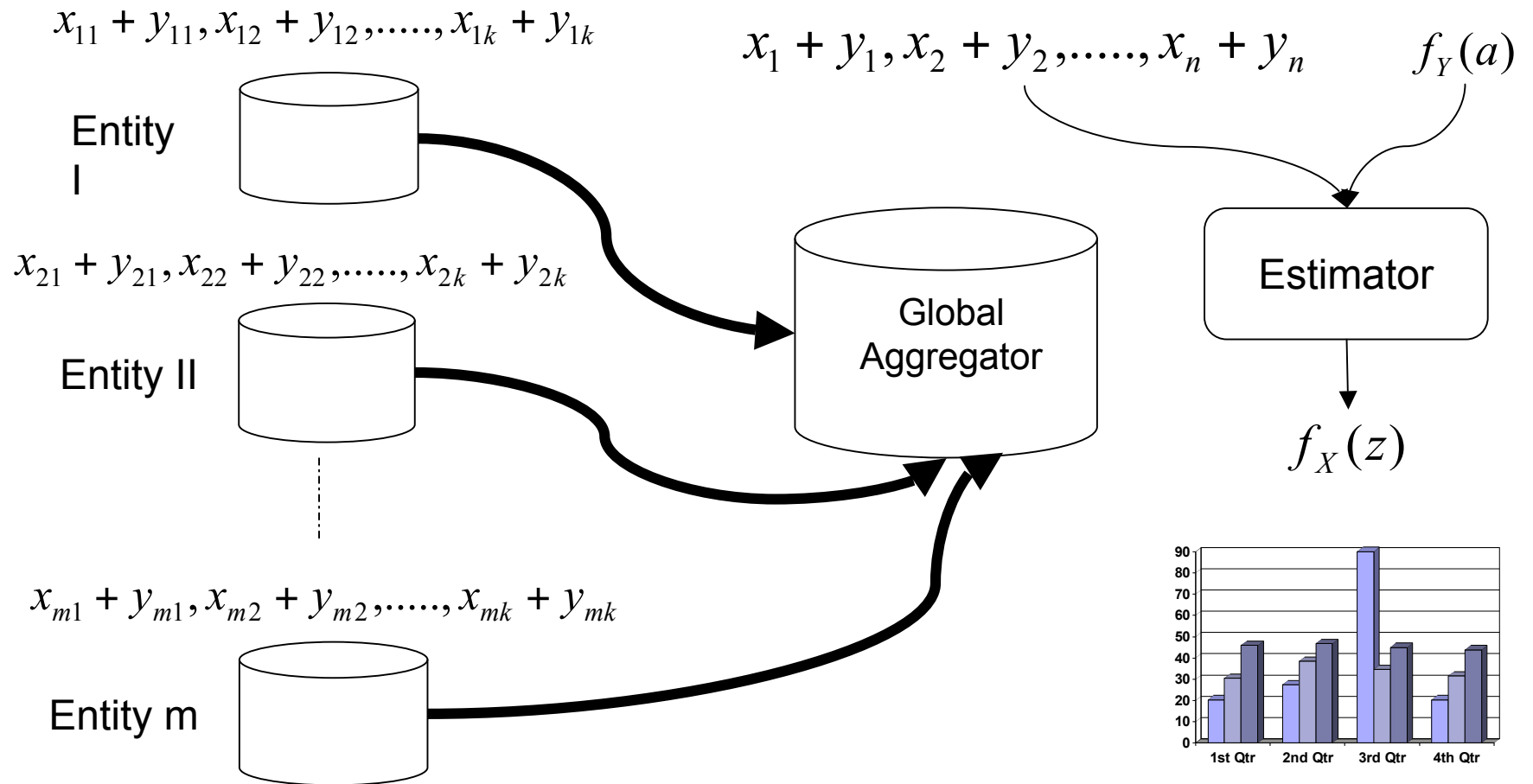
- Value Distortion Method
- Given a value x_i the client is allowed to report a distorted value $(x_i + r)$ where r is a random variable drawn from a known distribution
 - Uniform Distribution: $\mu = 0, [-\alpha, +\alpha]$
 - Gaussian Distribution: $\mu = 0, \sigma$

Quantifying the privacy of different randomization Schemes

Confidence (α)	50 %	95 %	99.9 %	Distribution
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	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$	

Gaussian Distribution provides the best accuracy at higher confidence levels

Problem Statement





Reconstruction of the Original Distribution

- Reconstruction problem can be viewed in in the general framework of the “Inverse Problems”
- Inverse Problems: describing system internal structure from indirect noisy data.
- Bayesian Estimation is an Effective tools for such settings



Formal problem statement

- Given one dimensional array of randomized data

$$x_1 + y_1, x_2 + y_2, \dots, x_n + y_n$$

- Where x_i 's are iid random variables each with the same distribution as the random variable X
- And y_i 's are realizations of a globally known random distribution with CDF F_Y
- Purpose: Estimate F_X



Background: Bayesian Inference

- An Estimation method that involves collecting observational data and use it a tool to adjust (either support of refute) a prior belief.
- The previous knowledge (hypothesis) has an established probability called (prior probability)
- The adjusted hypothesis given the new observational data is called (posterior probability)




Bayesian Inference

- Let $P(H_0)$ the prior probability, then Bayes' rule states that the posterior probability of (H_0) given an observation (D) is given by:

$$P(H_0 | D) = \frac{P(D | H_0)P(H_0)}{P(D)}$$

- Bayes rule is a cyclic application of the general form of the joint probability theorem:

$$P(D, H_0) = P(H_0 | D)P(D)$$



Bayesian Inference (Classical Example)

- Two Boxes:
 - Box-I : 30 Red balls and 10 White Balls
 - Box-II: 20 Red balls and 20 White Balls
- A Person draws a Red Ball, what is the probability that the Ball is from Box-I
- Prior Probability $P(\text{Box-I}) = 0.5$
- From the data we know that:
 - $P(\text{Red}|\text{Box-I}) = 30/40 = 0.75$
 - $P(\text{Red}|\text{Box-II}) = 20/40 = 0.5$



Example (cntd.)

- Now, given the new observation (The Red Ball) we want to know the posterior probability of Box-I (i.e $P(\text{Box-I} \mid \text{Red})$)

$$P(\text{Box} - I \mid \text{RED}) = \frac{P(\text{RED} \mid \text{Box} - I)P(\text{Box} - I)}{P(\text{RED})}$$

$$P(\text{RED}) = P(\text{RED}, \text{Box} - I) + P(\text{RED}, \text{Box} - II)$$

$$P(\text{RED}) = P(\text{RED} \mid \text{Box} - I)P(\text{Box} - I) + P(\text{RED} \mid \text{Box} - II)P(\text{Box} - II)$$

$$P(\text{RED}) = 0.5 \times 0.75 + 0.5 \times 0.5$$



Example (cntd)

- Computing the joint probability:

$$P(RED) = P(RED \mid Box - I)P(Box - I) + P(RED \mid Box - II)P(Box - II)$$

$$P(RED) = 0.5 \times 0.75 + 0.5 \times 0.5$$

- Substituting,

$$P(Box - I \mid RED) = \frac{0.75 \times 0.5}{0.5 \times 0.75 + 0.5 \times 0.5} = 0.6$$

- The posterior probability of Box-I is amplified by the observation of the Red Ball



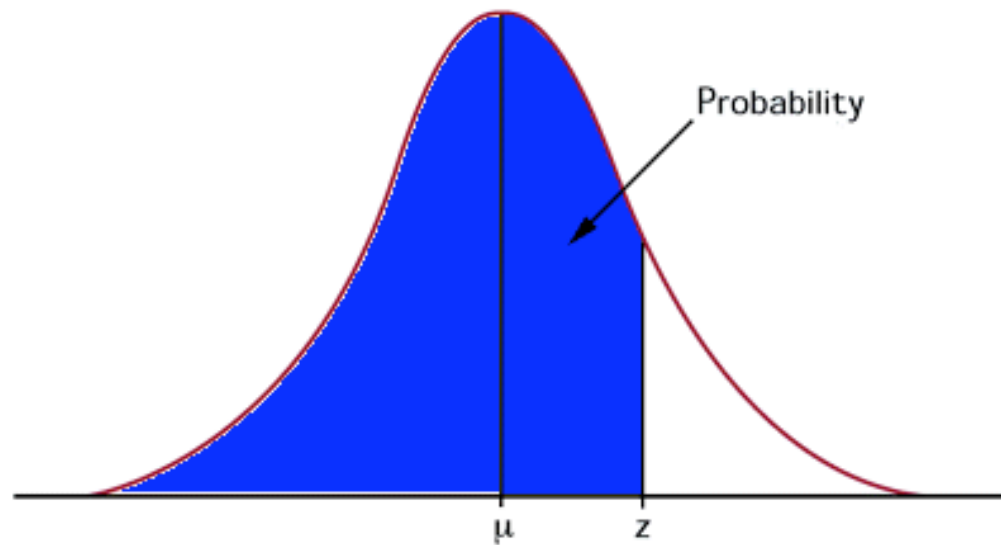
Back: Formal problem statement

- Given one dimensional array of randomized data

$$x_1 + y_1, x_2 + y_2, \dots, x_n + y_n$$

- Where x_i 's are iid random variables each with the same distribution as the random variable X
- And y_i 's are realizations of a globally known random distribution with CDF F_Y
- Purpose: Estimate F_X

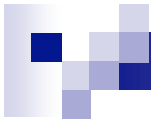
Continuous probability distributions



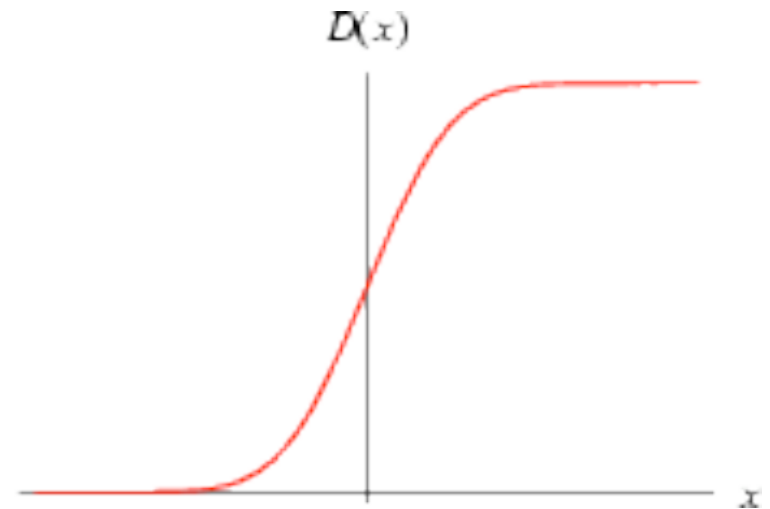
$$P\{r \leq z\} = \int_{-\infty}^z f_X(k)dk = CDF(z) = F_X(z)$$

$$P\{r = z\} = 0$$

$$\int_{-\infty}^{+\infty} f_X(k)dk = 1$$



CDF and PDF





Estimation of F_X

- Bayes Rule:
$$P(H_0 | D) = \frac{P(D | H_0)P(H_0)}{P(D)}$$

- Posterior Probability

$$F_X(a) \equiv \int_{z=-\infty}^{z=a} f_X(z | X_1 + Y_1 = w_1) dz$$

- Applying Bayes rule

$$F_X(a) = \int_{-\infty}^a \frac{f_{X_1+Y_1}(w_1 | X_1 = z) f_{X_1}(z) dz}{f_{X_1+Y_1}(w_1)}$$



Estimation of F_X

- We want to evaluate $f_{X_1+Y_1}(w_1)$

$$f_{X_1+Y_1}(w_1) = \int_{-\infty}^{\infty} f_{X_1+Y_1}(w_1 | X_1 = k) f_{X_1}(k) dk$$

- Substituting:

$$F_X(a) = \frac{\int_{-\infty}^a f_{X_1+Y_1}(w_1 | X_1 = z) f_{X_1}(z) dz}{\int_{-\infty}^{\infty} f_{X_1+Y_1}(w_1 | X_1 = k) f_{X_1}(k) dk}$$



Estimation of F_X

- Simplification (independence):

$$F_X(a) = \frac{\int_{-\infty}^a f_Y(w_i - z) f_X(z) dz}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(k) dk}$$



Estimation of F_X

- For all n observations:

$$F_X(a) = \frac{1}{n} \sum_{i=1}^n \frac{\int_{-\infty}^a f_Y(w_i - z) f_X(z) dz}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X(k) dk}$$



Estimation of the PDF f_X

- f_X Is just the derivative of the CDF

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



Algorithm

$$f_X^0 := \text{Uniform Distribution}$$

$$j := 0$$

While (not Stopping Condition):

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

$$j := j + 1$$



Stopping Criteria

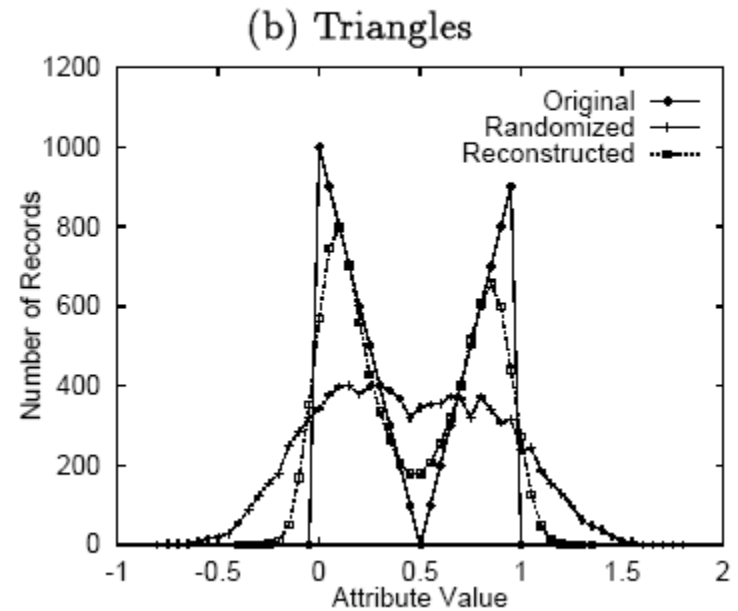
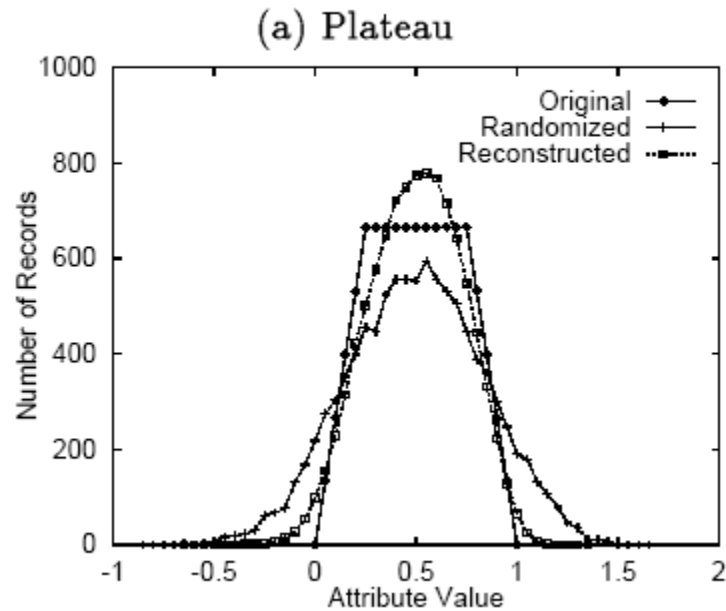
- The algorithm should terminate if:

$$f_X^{j+1}(a) \cong f_X^j(a)$$

- For each round a χ^2 goodness of fit test is performed.
- Iteration is stopped when the difference between the two estimates is too small (lower than a certain threshold)

Evaluation

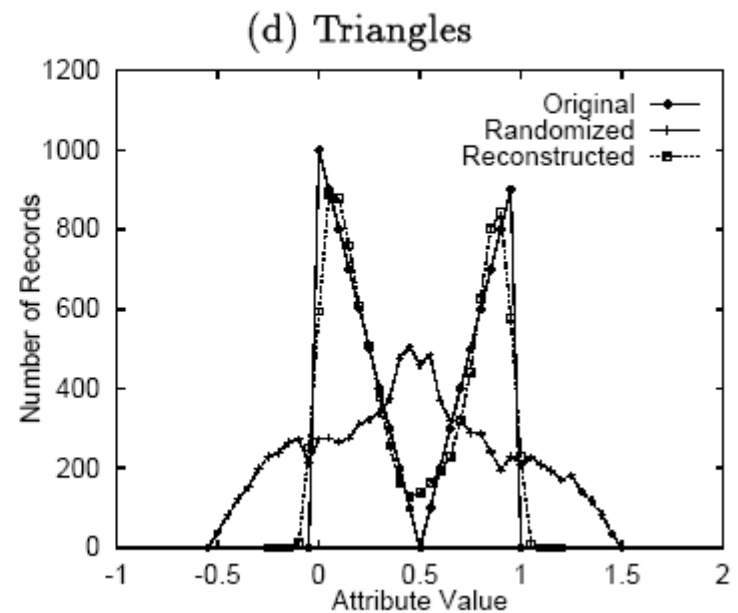
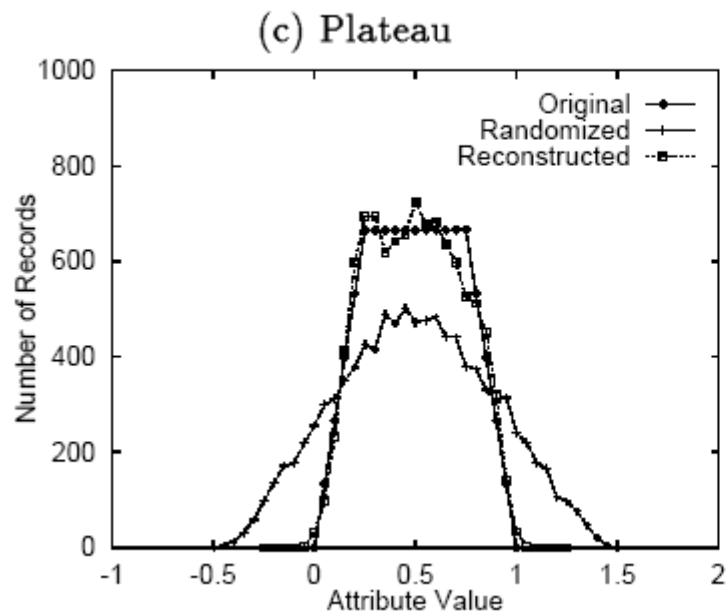
Gaussian



Gaussian Randomizing Function $\mu=0$, $\sigma = 0.25$

Evaluation

Uniform



Uniform Randomizing Function $[-0.5, 0.5]$



How is this different from Kalman Estimator?

- Both are estimation techniques
- Kalman is stateless
- In Kalman filter case we knew the distribution and estimation is used to validate whether the trend of the data matches that distribution
- In Bayesian Inference the observation data is used to adjust the prior hypothesis (probability distribution)



Is the Problem Solved?

- Suppose a client randomizes Age records using a uniform random variable $[-50, 50]$
- If the aggregator receives value 120, with 100% confidence it knows that actual age is ≥ 70
- Simply randomization does not guarantee absolute privacy



How to achieve better randomization scheme

- **“Limiting Privacy Breaches in Privacy Preserving Data Mining” Evfimievski et al**
- Define an evaluation metric of how privacy preserving a scheme is.
- Based on the developed metric, develop a randomization scheme that abides to this metric



How Privacy preserving is a scheme?

- Information Theoretic Approach:
 - Computes the average information disclose in a randomized attribute by computing the mutual information between the actual and the randomized attribute
- Privacy breach
 - Defines a criteria that should be satisfied for a randomization scheme to be privacy preserving



What is a privacy breach?

- A privacy breach occurs when the disclosure of a randomized value y_i to the aggregator reveals that a certain property $Q(x)$ of the “individual” input x_i holds with high probability



Privacy Breach

- Back to Bayes'
- Prior Probability $P(Q(x))$ where $Q(x)$ is the property
- Posterior probability: $P(Q(x) | y_i)$



Amplification

- Is defined in terms of the transitive probability $P[x \rightarrow y]$ where y is a fixed randomized output value
- Intuitive definition:
 - if there are many x_i 's that can be mapped to y by the randomizing scheme then disclosing y have gives little information about x_i
 - We say we amplify the probability that $P[x \rightarrow y]$



Amplification factor

Let,

R a randomization operator

$y \in V_y$ a randomized value of x .

Revealing R will not cause privacy breach if :

$$\frac{p_2}{p_1} \frac{(1 - p_1)}{(1 - p_2)} > \gamma$$



Summary

- No one solution can fit all.
- Which area looks more promising?
- Can we create robust randomization schemes to a wide scale of applications and different distributions of data?
- How to deal with the case of Malicious participants?