



Privacy Preserving Data Mining

A Randomization Approach



Ramakrishnan Srikant



Data Mining and Privacy

- The primary task in data mining: development of models about aggregated data.
- What if we randomize individual data records to protect privacy?
- Can we still develop accurate models?



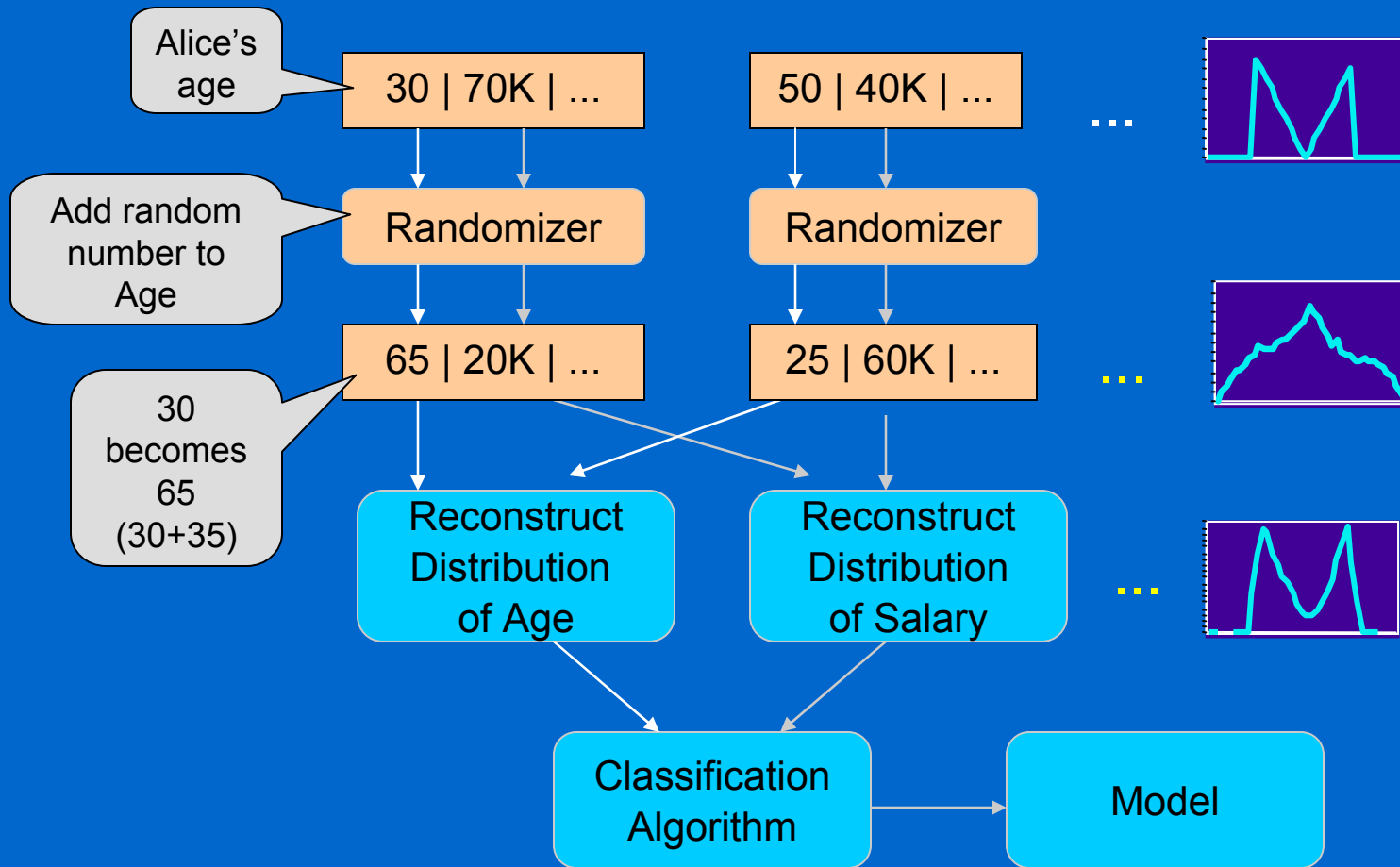
Talk Outline

- Classification
 - R. Agrawal and R. Srikant, “Privacy Preserving Data Mining”, SIGMOD 2000.
- Association Rules
- Open Problems

Y. Lindell, B. Pinkas. Privacy Preserving Data Mining. Crypto 2000.



Randomization Approach Overview



Reconstruction Problem

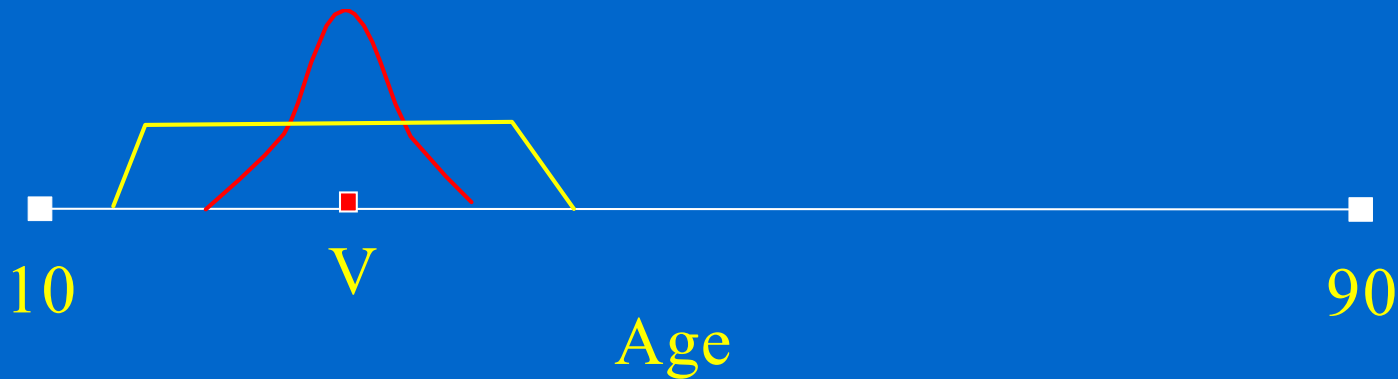
- Original values x_1, x_2, \dots, x_n
 - from probability distribution X (unknown)
- To hide these values, we use y_1, y_2, \dots, y_n
 - from probability distribution Y
- Given
 - $x_1+y_1, x_2+y_2, \dots, x_n+y_n$
 - the probability distribution of Y

Estimate the probability distribution of X .

-
-
-

Intuition (Reconstruct single point)

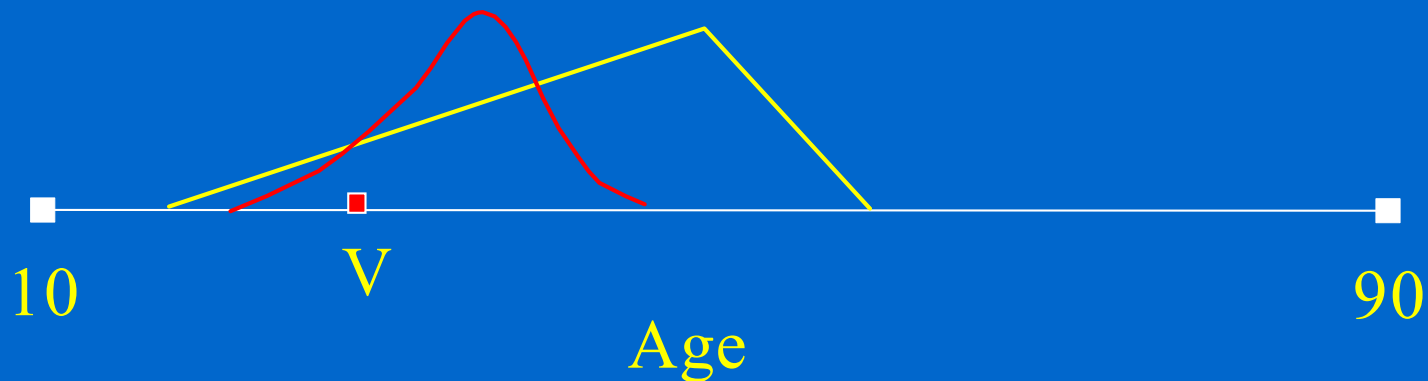
- Use Bayes' rule for density functions



- Original distribution for Age
- Probabilistic estimate of original value of V

Intuition (Reconstruct single point)

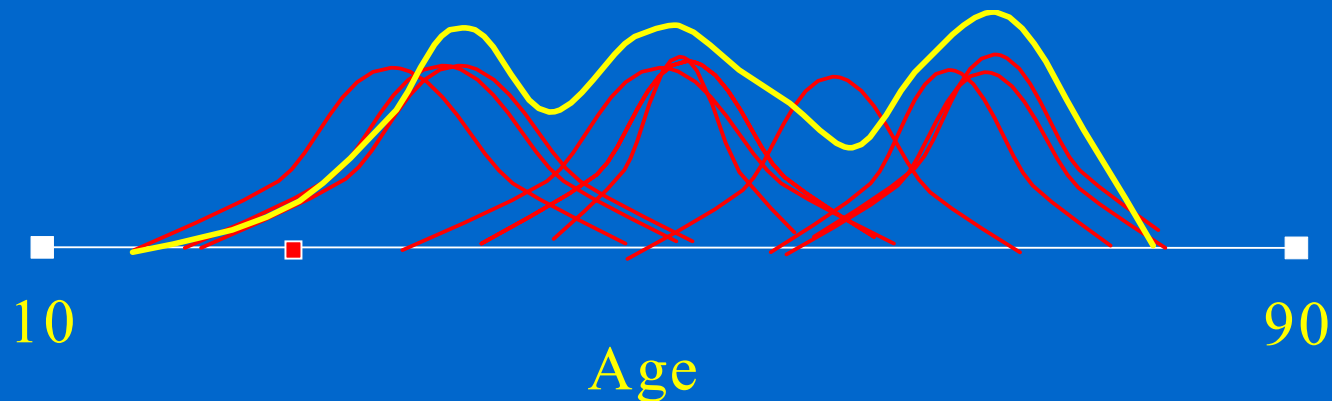
- Use Bayes' rule for density functions



- Original Distribution for Age
- Probabilistic estimate of original value of V

Reconstructing the Distribution

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



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

Reconstruction: Bootstrapping

f_X^0 := Uniform distribution

$j := 0$ // Iteration number

repeat

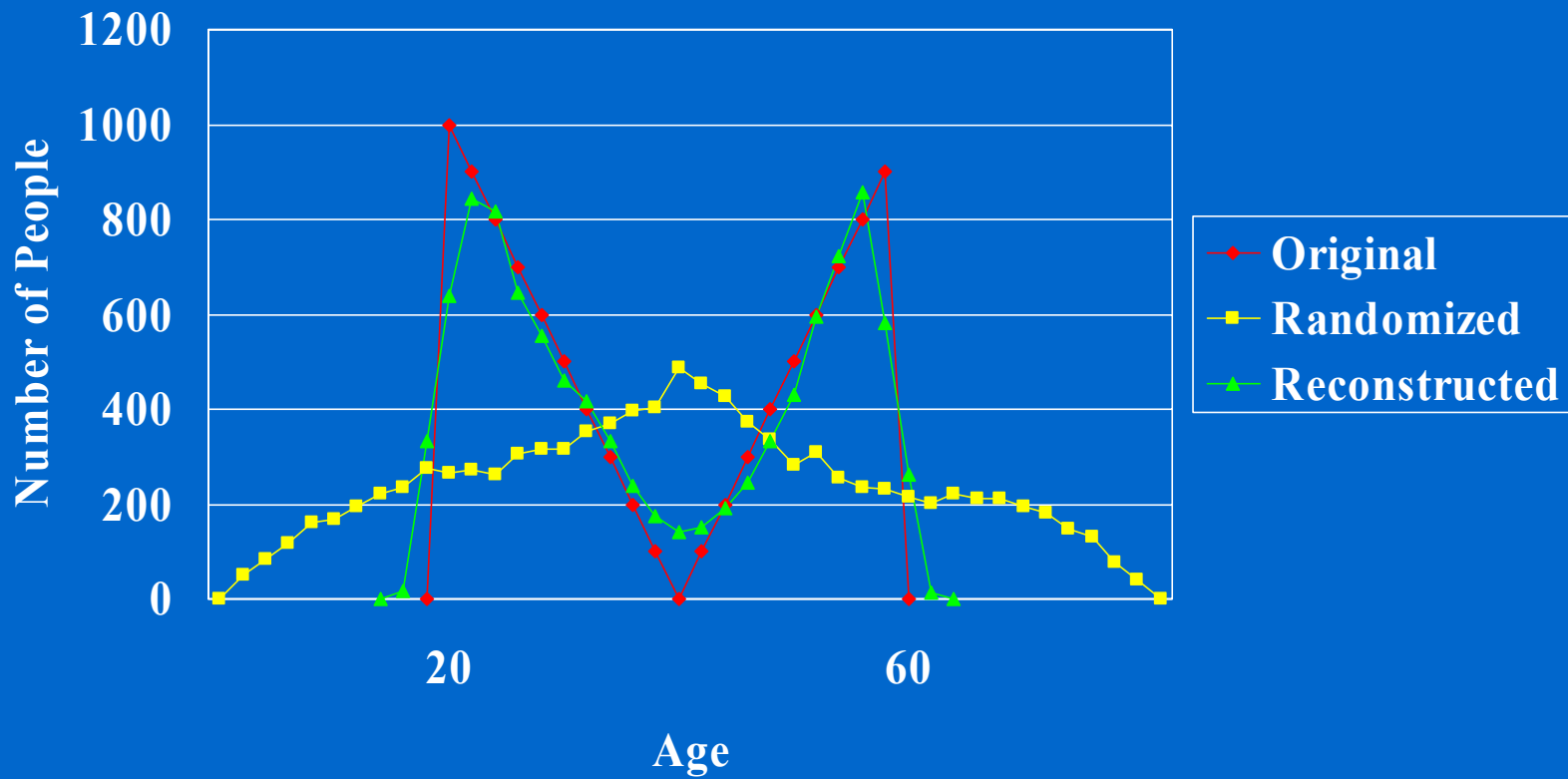
$$f_X^{j+1}(a) := \frac{1}{n} \sum_{i=1}^n \frac{f_Y((x_i + y_i) - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y((x_i + y_i) - a) f_X^j(a)} \quad (\text{Bayes' rule})$$

$j := j+1$

until (stopping criterion met)

- Converges to maximum likelihood estimate.
 - D. Agrawal & C.C. Aggarwal, PODS 2001.

Works well



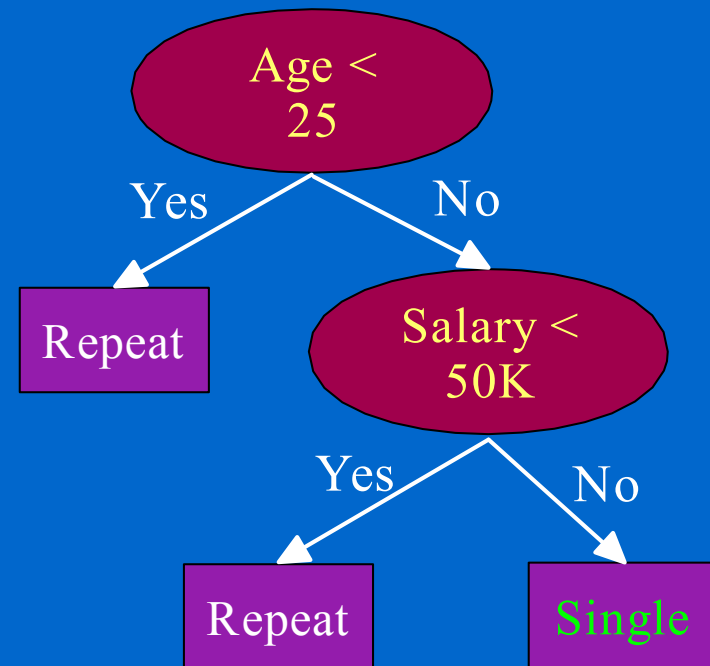
-
-
-

Classification

- Naïve Bayes
 - Assumes independence between attributes.
- Decision Tree
 - Correlations are weakened by randomization, not destroyed.

Decision Tree Example

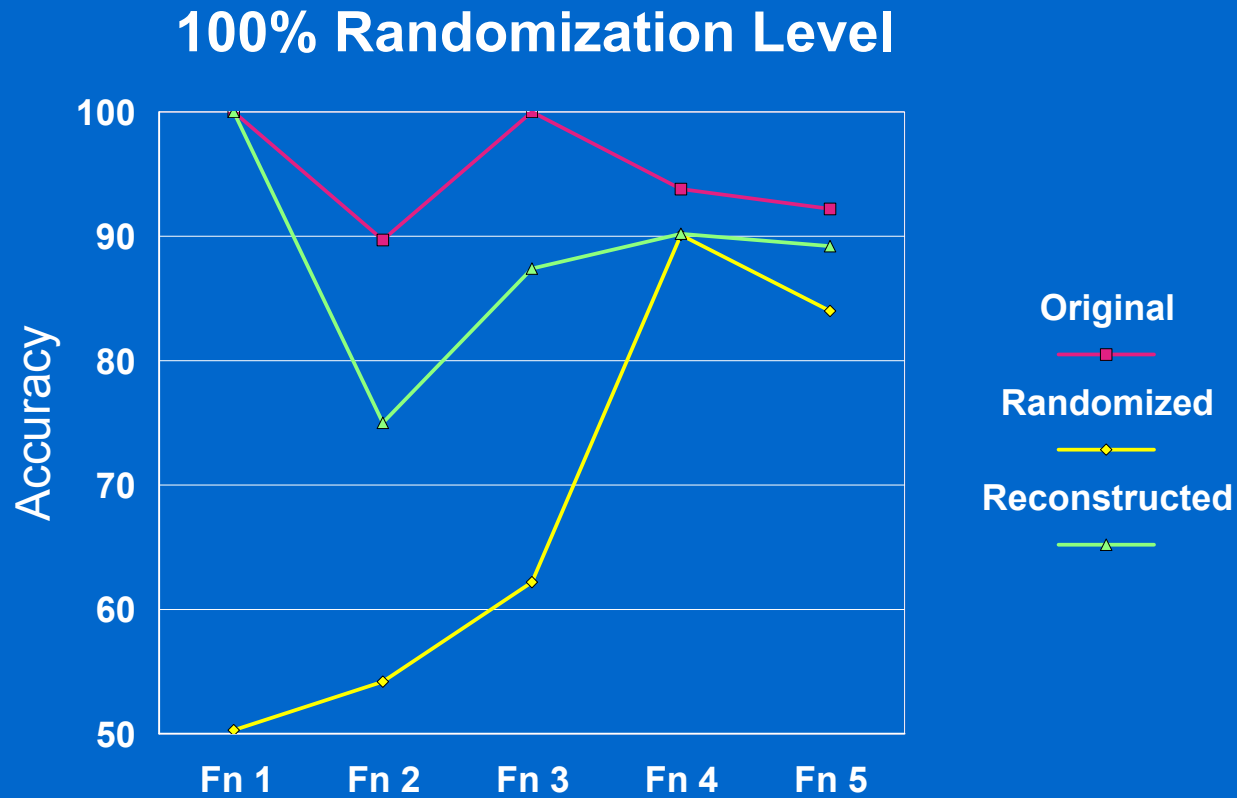
Age	Salary	Repeat Visitor?
23	50K	Repeat
17	30K	Repeat
43	40K	Repeat
68	50K	Single
32	70K	Single
20	20K	Repeat



Randomization Level

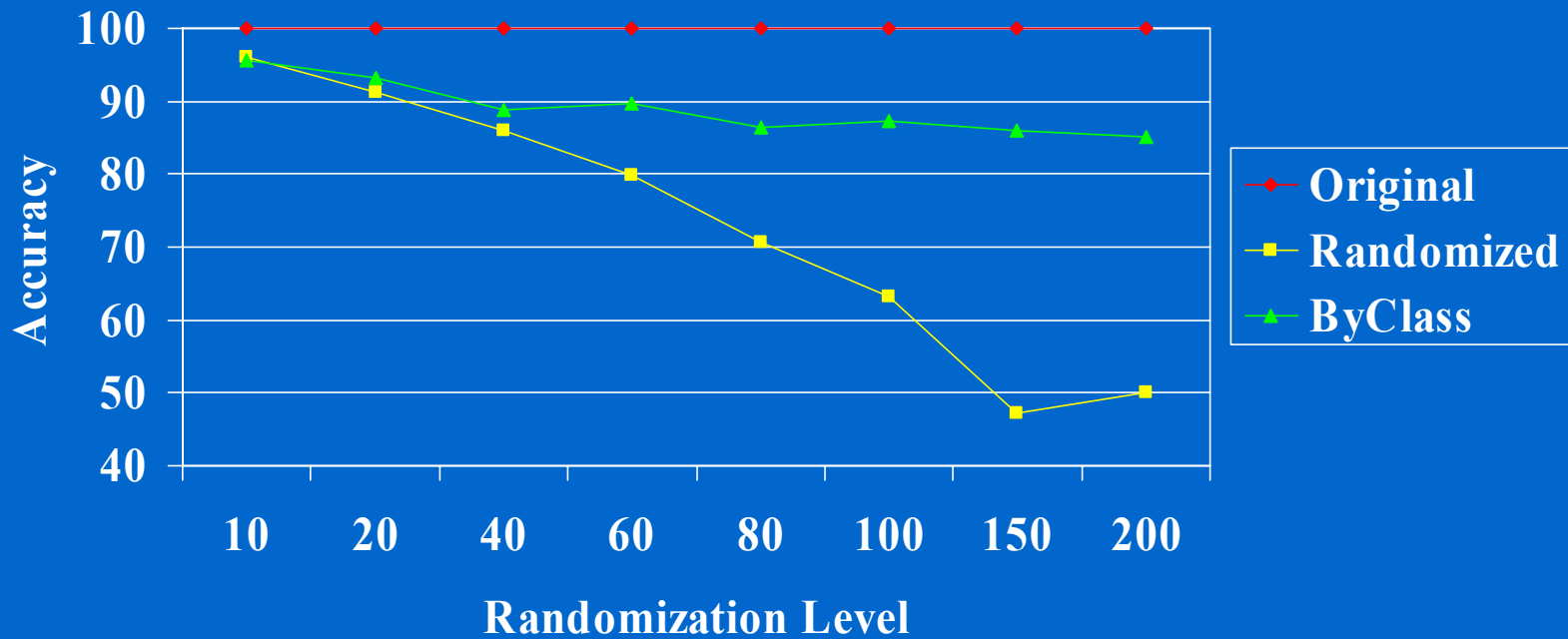
- Add a random value between -30 and +30 to age.
- If randomized value is 60
 - know with 90% confidence that age is between 33 and 87.
- Interval width \propto amount of privacy.
 - Example: (Interval Width : 54) / (Range of Age: 100) \Rightarrow 54% randomization level @ 90% confidence

Decision Tree Experiments



Accuracy vs. Randomization Level

Fn 3



Talk Outline

- Motivation
- Association Rules
 - A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, “Privacy Preserving Mining of Association Rules”, KDD 2002.
- Open Problems

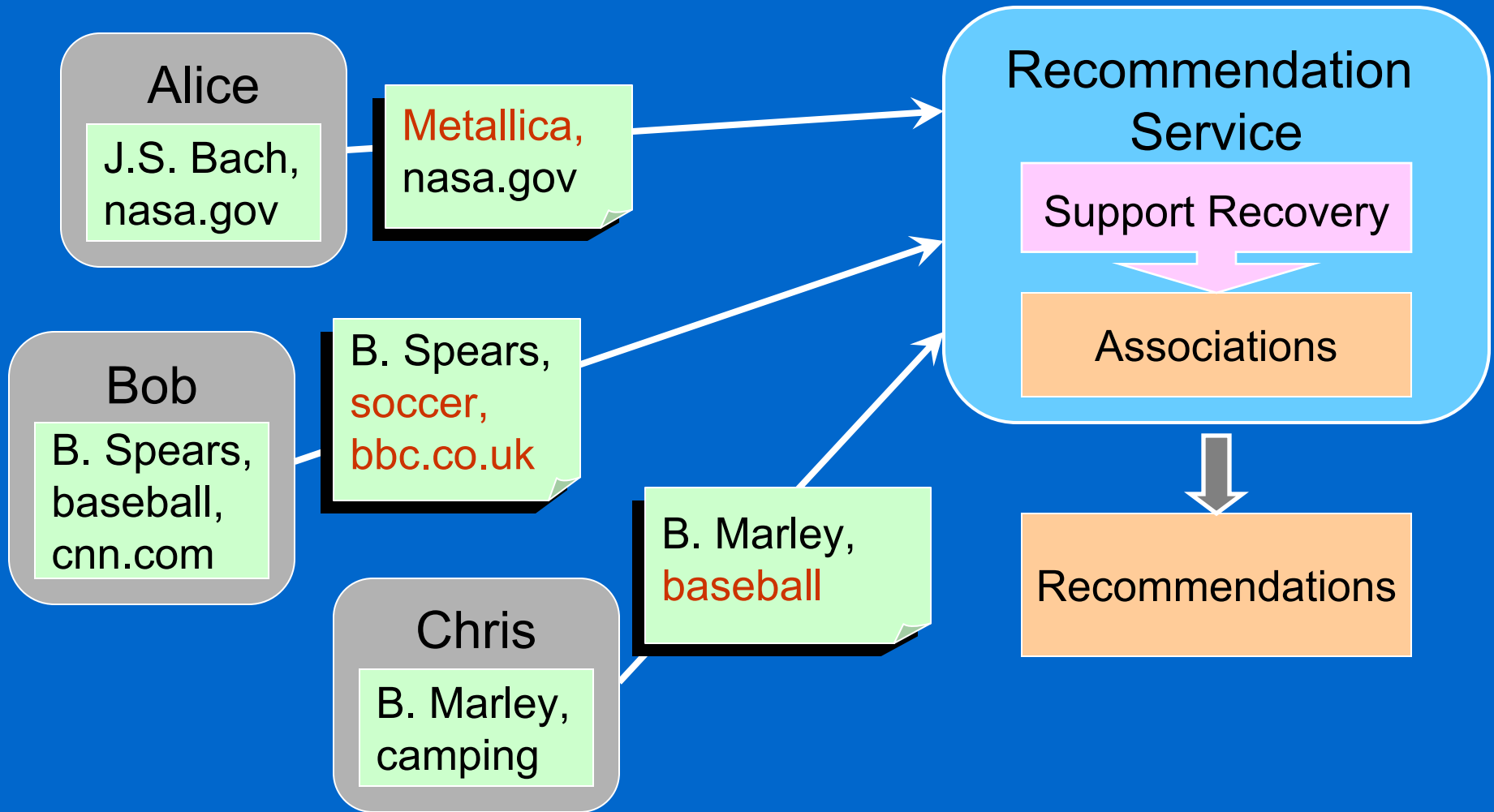
S. Rizvi, J. Haritsa. Privacy-Preserving Association Rule Mining.
VLDB 2002

J. Vaidya, C.W. Clifton. Privacy Preserving Association Rule Mining in
Vertically Partitioned Data. KDD 2002.

Discovering Associations Over Privacy Preserved Categorical Data

- A transaction t is a set of items
- Support s for an itemset A is the number of transactions in which A appears
- Itemset A is frequent if $s \geq s_{\min}$
- Task: Find all frequent itemsets, while preserving the privacy of individual transaction.

Recommendation Service



-
-
-

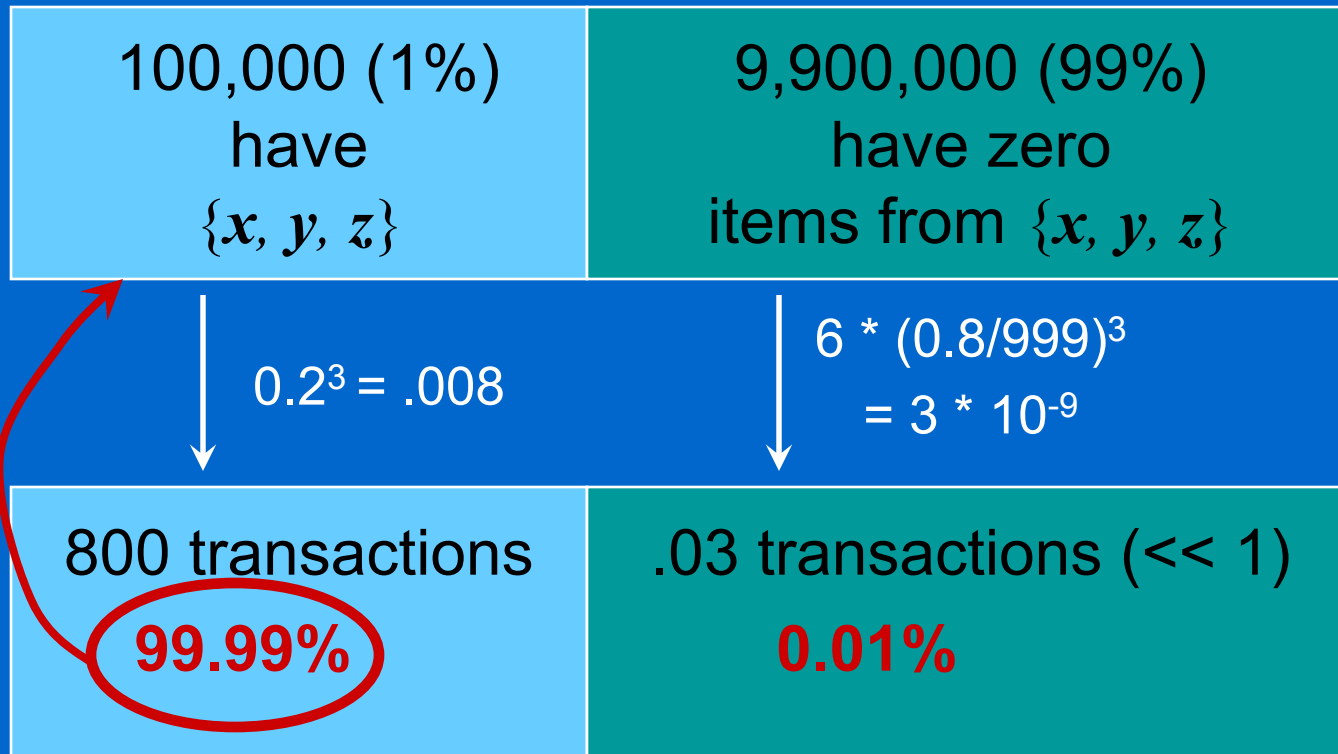
Uniform Randomization

- Given a transaction,
 - keep item with 20% probability,
 - replace with a new random item with 80% probability.

Is there a problem?

Example: $\{x, y, z\}$

10 M transactions of size 3 with 1000 items:



Uniform randomization: How many have $\{x, y, z\}$?

-
-
-

Our Solution

“Where does a wise man hide a leaf? In the forest.
But what does he do if there is no forest?”
“He grows a forest to hide it in.”

G.K. Chesterton

- Insert many false items into each transaction
- Hide true itemsets among false ones

Cut and Paste Randomization

- Given transaction t of size m , construct t' :
 - Choose a number j between 0 and K_m (cutoff);
 - Include j items of t into t' ;
 - Each other item is included into t' with probability p_m .

The choice of K_m and p_m is based on the desired level of privacy.

$t =$ $a, b, c, u, v, w, x, y, z$

$t' =$ b, v, x, z $\alpha, \acute{a}, \beta, \xi, \psi, \epsilon, \aleph, \upsilon, \grave{h}, \dots$

\longleftrightarrow
 $j = 4$

Partial Supports

To recover original support of an itemset, we need randomized supports of its subsets.

- Given an itemset A of size k and transaction size m ,
- A vector of partial supports of A is

$$\vec{s} = (s_0, s_1, \dots, s_k), \text{ where}$$

$$s_l = \frac{1}{|T|} \cdot \# \{t \in T \mid \#(t \cap A) = l\}$$

- Here s_k is the same as the support of A .
- Randomized partial supports are denoted by \vec{s}' .

Transition Matrix

- Let $k = |\mathbf{A}|$, $m = |\mathbf{t}|$.
- Transition matrix $\mathbf{P} = \mathbf{P}(k, m)$ connects randomized partial supports with original ones:

$$\mathbb{E} \vec{s}' = \mathbf{P} \cdot \vec{s}, \text{ where}$$

$$P_{l',l} = \Pr \left[\#(t' \cap A) = l' \mid \#(t \cap A) = l \right]$$

The Estimators

- Given randomized partial supports, we can estimate original partial supports:

$$\vec{s}_{\text{est}} = Q \cdot \vec{s}', \quad \text{where} \quad Q = P^{-1}$$

- Covariance matrix for this estimator:

$$\text{Cov } \vec{s}_{\text{est}} = \frac{1}{|T|} \sum_{l=0}^k s_l \cdot Q D[l] Q^T,$$

$$\text{where } D[l]_{i,j} = P_{i,l} \cdot \delta_{i=j} - P_{i,l} \cdot P_{j,l}$$

- To estimate it, substitute s_l with $(s_{\text{est}})_l$.
 - Special case: estimators for support and its variance

Privacy Breach Analysis

- How many added items are enough to protect privacy?
 - Have to satisfy $\Pr [z \in t \mid A \subseteq t'] < \rho$ (\Leftrightarrow no privacy breaches)
 - Select parameters so that it holds for all itemsets.
 - Use formula ($s_l^+ = \Pr [\#(t \cap A) = l, z \in t]$, $s_0^+ = 0$):

$$\Pr[z \in t \mid A \subseteq t'] = \frac{\sum_{l=0}^k s_l^+ \cdot P_{k,l}}{\sum_{l=0}^k s_l \cdot P_{k,l}}$$

- Parameters are to be selected in advance!
 - Enough to know maximal support of an itemset for each size.
 - Other parameters chosen for worst-case impact on privacy breaches.

Can we still find frequent itemsets?

Privacy Breach level = 50%.

Soccer:

$$s_{\min} = 0.2\%$$

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	266	254	12	31
2	217	195	22	45
3	48	43	5	26

Mailorder:

$$s_{\min} = 0.2\%$$

Itemset Size	True Itemsets	True Positives	False Drops	False Positives
1	65	65	0	0
2	228	212	16	28
3	22	18	4	5

-
-
-

Talk Outline

- Classification
- Association Rules
- Open Problems



Privacy Breaches

- We know how to control privacy breaches for boolean data (associations) – what about quantitative data?
- Example: 80% of the people whose randomized value of age is in $[80,90]$ and whose randomized value of income is [...] have their true age in $[70,80]$.
- Challenge: How do you limit privacy breaches without prior knowledge of data distributions?



-
-
-

Clustering

- Classification: Partitioned the data by class & then reconstructed attributes.
 - Assumption: Attributes independent given class attribute.
- Clustering: Don't know the class label.
 - Assumption: Attributes independent.
 - Latter assumption is much worse!
- Can we reconstruct a set of attributes together?
 - Amount of data needed increases exponentially with number of attributes.



Summary

- Can have our cake and mine it too!
 - Randomization is an interesting approach for building data mining models while preserving user privacy.
- Algorithms for privacy-preserving classification and association rules.
- Lots of interesting open problems.



