



Privacy-Preserving Data Management and Mining

Yücel SAYGIN

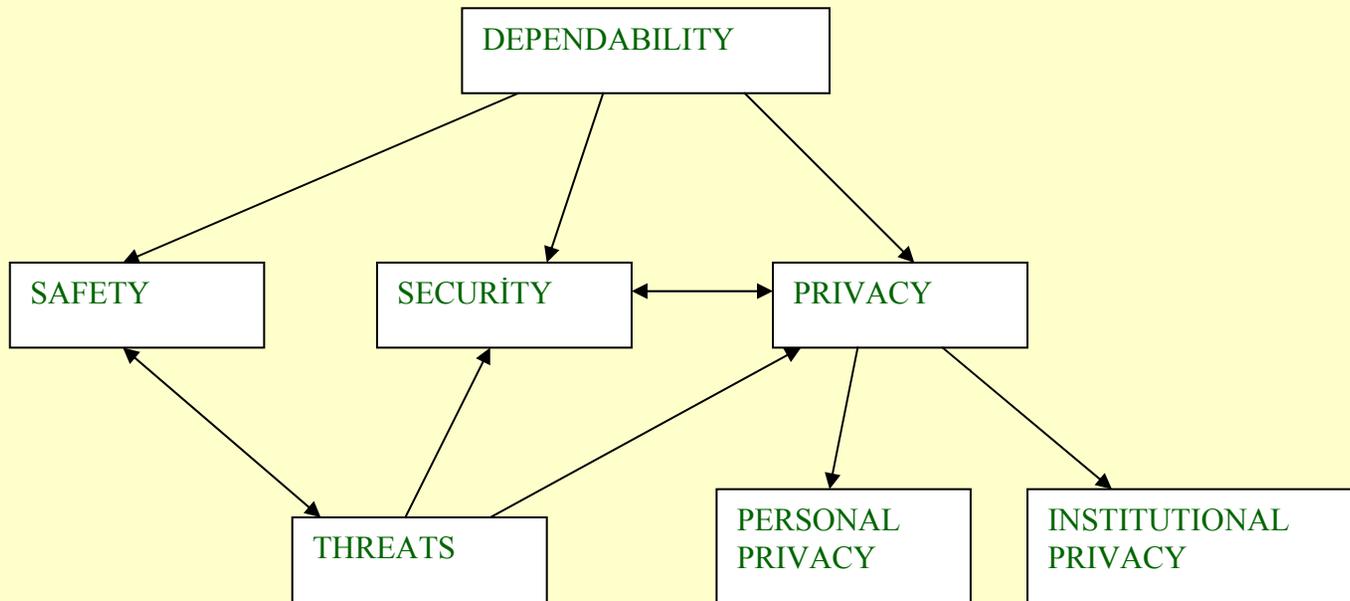
ysaygin@sabanciuniv.edu

<http://people.sabanciuniv.edu/~ysaygin/>

WHAT IS PRIVACY

- One of the fundamental rights of people
 - The right to be left alone
 - The right of people to have control over their private space
 - Data Privacy : Giving the data owners the right to say what can be done with their data
-

The Big Picture



PRIVACY

- 9/11 as a turning point in the US
 - Fight against terrorism!!!!
-

Why privacy is a really big issue these days?

CAPPS II (Computer Assisted Passenger Prescreening System) collects flight reservation information as well as commercial information about passengers. This data, in turn, can be utilized by government security agencies. Although CAPPS represents US national data collection efforts, it also has an effect on other countries.

Why privacy is a really big issue these days?

The following sign at the KLM ticket desk in Amsterdam International Airport demonstrates the point:

Please note that KLM Royal Dutch Airlines and other airlines are required by new security laws in the US and several other countries to give security customs and immigration authorities access to passenger data. Accordingly any information we hold about you and your travel arrangements may be disclosed to the concerning authorities of these countries in your itinerary

Why privacy is a really big issue these days?

Some of the largest airline companies in US, including American, United and Northwest, turned over millions of passenger records to the FBI

SSchwartz J. & Micheline M. (2004). Airlines Gave F.B.I. Millions of Records on Travelers After 9/11 *NY Times*, May 1.

Why privacy is a really big issue these days?

- Technology is really integrated with our personal life
 - With new technology : Networking, WEB
 - New devices: Mobile Phones, RFID tags, Computers, digital cameras
 - Which means that data about us, and about what we are doing can be collected easily and at a fraction of the cost 10 years ago.
 - Navigation patterns in WEB
 - Location information (wireless phones, RFID tags)
 - Transactions (e-commerce, POS...)
 - Your emails (now scanned by gmail to display ads) (was a big discussion in the CFP conference at Berkeley, 2004)
-

Why privacy is a really big issue these days?

Total Information Awareness (TIA) project in US, which aims to build a centralized database that will store the credit card transactions, emails, web site visits, flight details of Americans was not funded by the Congress due to privacy concerns.

Why privacy is a really big issue these days?

Now there is actually the means to create the cyber-big-brother!

- Video-cameras,
 - Various sensors
 - RFID tags and readers
 - Web navigation, emails
 - Electronic transactions
-

So what's Data Mining got to do with it?

- Databases and Data Mining deal with the data about people
 - Data mining was seen as a threat by public since:
 - The term data mining was used a lot by the officials
 - Data mining is seen by the public as a magic tool to discover secrets
 - Data mining projects were not funded recently due to privacy concerns
 - This fact pushed data mining researchers towards what is called
 - “Privacy Preserving Data Mining”
 - And “Privacy Preserving Data Management” in general
-

Data mining and Privacy Issues Gained Momentum in US

- Pentagon has released a study that recommends the government to pursue specific technologies as potential safeguards against the misuse of data-mining systems similar to those now being considered by the government to track civilian activities electronically in the United States and abroad

Perhaps the strongest protection against abuse of information systems is Strong Audit mechanisms... we need to watch the watchers

Markoff J. (2002). Study Seeks Technology Safeguards for Privacy. *NY Times*, 19 December.

- This shows us that even the most aggressive data collectors in the US are aware of the fact that the data mining tools could be misused and we need a mechanism to protect the confidentiality and privacy of people.
-

Another example

- Micro-data publishing at the ports
 - May reveal business secrets
-

Another example

- In August 2006, AOL releases user search data that includes
 - 20 million web queries of
 - 650.000 AOL users
 - i.e., all searches of its users over 3 months period
 - AOL realized its mistake and removed the data, but it was already there and downloaded by many.
 - Data did not contain the user identifiers
 - But it is a fact that users search about themselves on the web, and about their friends.
-

A severe privacy violation

- An anonymous user with id 4417749 searched for the following
 - “numb fingers”
 - “60 single men”
 - “dog that urinates on everything”
 - “landscapers in lilburn, Ga”
 - Several people with name “Arnold”
 - Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends’ medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.
-

Privacy violation by search engines

- Consider google, yahoo, altavista storing your searches
 - Over years
 - Which can be cross-linked to private information
 - Is the big brother becoming a reality in cyberspace?
-

Is data privacy something new?

- Maybe termed differently but it has been studied in the past
 - Statistical databases, statistical disclosure control ...
 - The inference problem
 - Late 90s: k-anonymity for data privacy
-

Is data privacy something new?

- Micro-data : actual data that is collected (not statistical summaries)
 - Micro-data Publishing by
 - census bureau demographic information
 - Internal revenue services
 - Health information
 - Identifiers should be removed before publishing
 - But is that enough?
-

Linkage attacks!

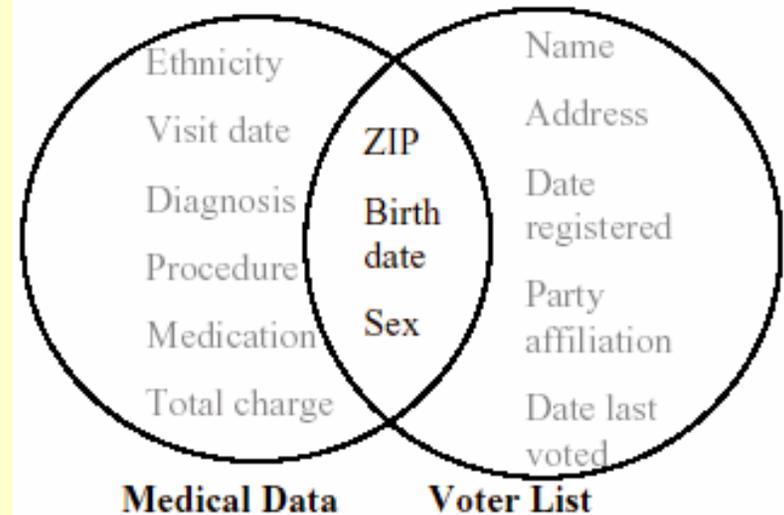
- Multiple information sources about individuals
 - Ex: (from Samarati, TKDE, 2001) In US, there are estimated 5 billion privately owned records about
 - Finances
 - Interests
 - Demographics
 - In some states you can pay 25 USD and buy license and license plate files that contain
 - Name, SSN, bdate, address, telephone, family status, employment, salary status
-

Linkage attacks!

- Healthcare data is one of the most sensitive
 - And prone to attacks
 - Demographics and medical records even de-identified may be linked to other databases
-

An Example: Re-identifying “anonymous” data (Sweeney ’01)

- 37 US states mandate collection of information
- She purchased the voter registration list for Cambridge Massachusetts
 - 54,805 people
- 69% unique on postal code and birth date
- 87% US-wide with all three



Samarati, TKDE 2001

Medical Data Released as Anonymous

SSN	Name	Race	DateOfBirth	Sex	ZIP	Marital Status	HealthProblem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath



Voter List

Name	Address	City	ZIP	DOB	Sex	Party
.....
.....
Sue J. Carlson	900 Market St.	San Francisco	94142	9/15/61	female	democrat
.....



Samarati, TKDE 2001

- Possible techniques for identity protection:
 - Scrambling
 - Swapping
 - Adding noise,
 - While maintaining the statistical properties
- Many applications require true data

Samarati, TKDE 2001

- Generalization and suppression are better approaches in terms of data quality
- Quasi-identifiers for possible linkage attacks
- K-anonymity as a privacy measure
- Try to achieve minimal generalization for a given k

Samarati, TKDE 2001

- K-anonymity: (def from Samarati, TKDE 2001)
 - Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k-individuals

Medical Data Released as Anonymous

SSN	Name	Race	DateOfBirth	Sex	ZIP	Marital Status	HealthProblem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party
.....
.....
Sue J. Carlson	900 Market St.	San Francisco	94142	9/15/61	female	democrat

Samarati, TKDE 2001

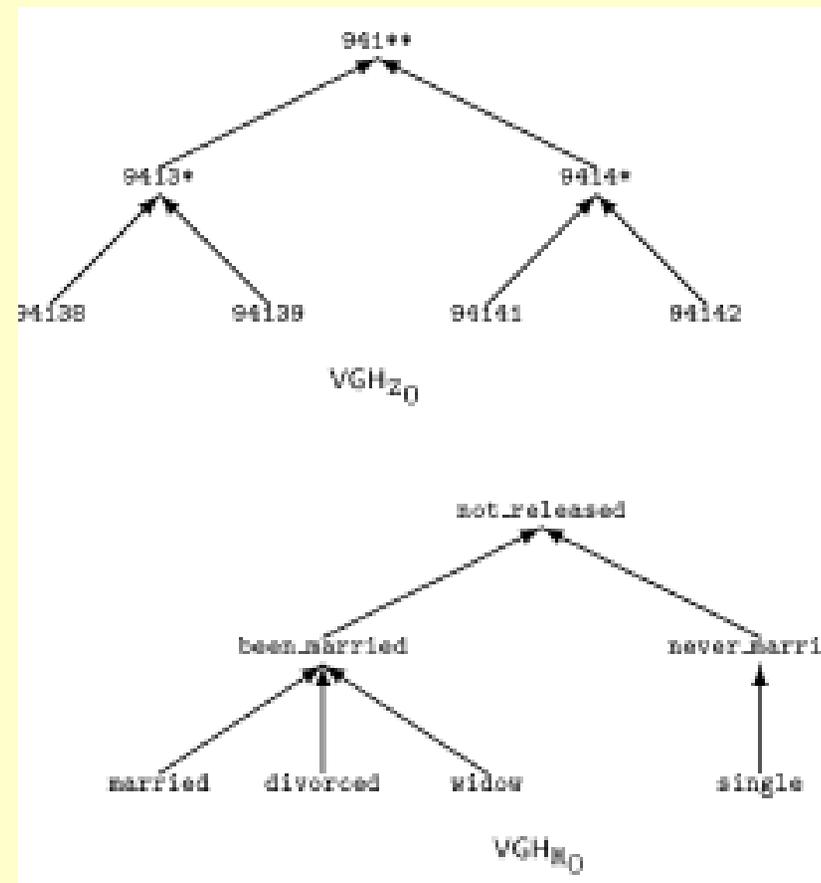
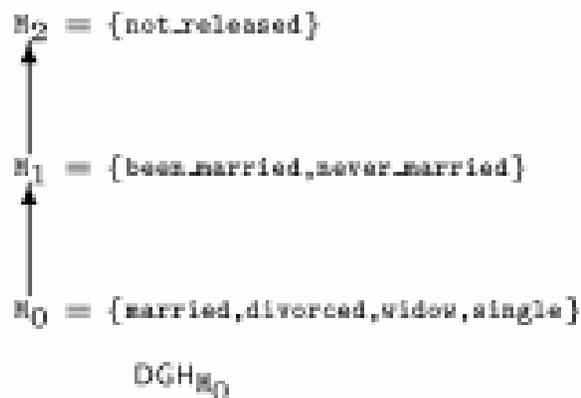
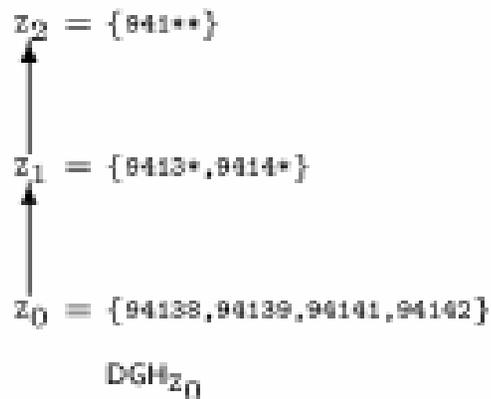
- K-anonymity: (def from Samarati, TKDE 2001)
 - Let $T(A_1, \dots, A_n)$ be a table and QI be a quasi identifier associated with T . T is said to satisfy k -anonymity wrt QI , iff each sequence of values in $T[QI]$ appears at least with k occurrences in $T[QI]$

Medical Data Released as Anonymous

SSN	Name	Race	DateOfBirth	Sex	ZIP	Marital Status	HealthProblem
		asian	09/27/64	female	94139	divorced	hypertension
		asian	09/30/64	female	94139	divorced	obesity
		asian	04/18/64	male	94139	married	chest pain
		asian	04/15/64	male	94139	married	obesity
		black	03/13/63	male	94138	married	hypertension
		black	03/18/63	male	94138	married	shortness of breath
		black	09/13/64	female	94141	married	shortness of breath
		black	09/07/64	female	94141	married	obesity
		white	05/14/61	male	94138	single	chest pain
		white	05/08/61	male	94138	single	obesity
		white	09/15/61	female	94142	widow	shortness of breath

Samarati, TKDE 2001

- K-anonymity: (def from Samarati, TKDE 2001)
 - Domain Generalization Hierarchies (DGH)
 - Value Generalization Hierarchies (VGH)



Samarati, TKDE 2001

- Private Table (PT) and Generalized Tables (GT)
- GT[0,1] satisfies k-anonymity for k=1,2
- GT[1,0] for k = 1,2,3

Race:R ₀	ZIP:Z ₀
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R ₁	ZIP:Z ₀
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142
person	94138
person	94139
person	94141
person	94142

GT_[1,0]

Race:R ₁	ZIP:Z ₁
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*
person	9413*
person	9413*
person	9414*
person	9414*

GT_[1,1]

Race:R ₀	ZIP:Z ₁
asian	9413*
asian	9413*
asian	9414*
asian	9414*
black	9413*
black	9413*
black	9414*
black	9414*
white	9413*
white	9413*
white	9414*
white	9414*

GT_[0,1]

Samarati, TKDE 2001

- Private Table and Generalized Tables

Race:R ₀	ZIP:Z ₀
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

PT

Race:R ₀	ZIP:Z ₂
asian	941**
black	941**
white	941**

GT_[0,2]

Race:R ₁	ZIP:Z ₂
person	941**

GT_[1,2]

Samarati, TKDE 2001

- How about assigning the most general value for each attribute?
- What is a good generalization?

Samarati, TKDE 2001

- Distance vector $GT[0,2]$, $GT[1,2]$

Race: R_0	ZIP: Z_0
asian	94138
asian	94139
asian	94141
asian	94142
black	94138
black	94139
black	94141
black	94142
white	94138
white	94139
white	94141
white	94142

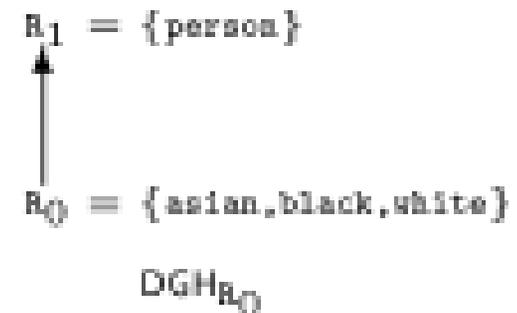
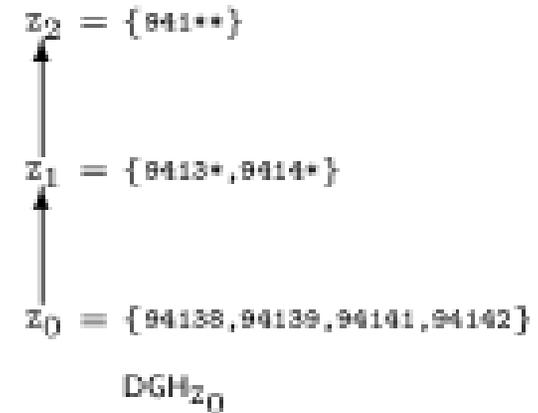
PT

Race: R_0	ZIP: Z_2
asian	941**
black	941**
white	941**

$GT_{[0,2]}$

Race: R_1	ZIP: Z_2
person	941**

$GT_{[1,2]}$



Samarati, TKDE 2001

- K-minimal generalization: T_j is the k-minimal generalization of T_i iff
 - T_j satisfies k-anonymity
 - There are no other generalized tables that satisfy k-anonymity with smaller distance vectors than T_j .

Samarati, TKDE 2001

Race	DOB	Sex	ZIP	MaritalStatus
asian	09/27/64	female	94139	divorced
asian	09/30/64	female	94139	divorced
asian	04/18/64	male	94139	married
asian	04/15/64	male	94139	married
black	03/13/63	male	94138	married
black	03/18/63	male	94138	married
black	09/13/64	female	94141	married
black	09/07/64	female	94141	married
white	05/14/61	male	94138	single
white	05/08/61	male	94138	single
white	09/15/61	female	94142	widow

Race	DOB	Sex	ZIP	MaritalStatus
asian	64	not_released	941**	not_released
asian	64	not_released	941**	not_released
asian	64	not_released	941**	not_released
asian	64	not_released	941**	not_released
black	63	not_released	941**	not_released
black	63	not_released	941**	not_released
black	64	not_released	941**	not_released
black	64	not_released	941**	not_released
white	61	not_released	941**	not_released
white	61	not_released	941**	not_released
white	61	not_released	941**	not_released

Race	DOB	Sex	ZIP	MaritalStatus
person	[60-64]	female	9413*	been_married
person	[60-64]	female	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	male	9413*	been_married
person	[60-64]	female	9414*	been_married
person	[60-64]	female	9414*	been_married
person	[60-64]	male	9413*	never_married
person	[60-64]	male	9413*	never_married
person	[60-64]	female	9414*	been_married

Samarati, TKDE 2001

- Suppression:

Race:R ₀	ZIP:Z ₀
asian	94138
asian	94138
asian	94142
asian	94142
black	94138
black	94141
black	94142
white	94138

PT

Race:R ₁	ZIP:Z ₀
person	94138
person	94138
person	94142
person	94142
person	94138
person	94141
person	94142
person	94138

GT_[1,0]

Race:R ₀	ZIP:Z ₀
asian	94138
asian	94138
asian	94142
asian	94142

GT_[0,0]

Race:R ₁	ZIP:Z ₀
person	94138
person	94138
person	94142
person	94142
person	94138
person	94142
person	94138

GT_[1,0]

Privacy and Data Mining

- There are two aspects of data mining when we look at it from a privacy perspective
 - Protecting the privacy of people against the misuse of data via data mining tools
 - Being able to mine the data without actually seeing it
-

Privacy and Data Mining

- The CODMINE Project funded by EU FP5 Future and Emerging Technologies Program.
 - GeoPKDD Project funded by EU FP6
 - KdUbiq Project funded by FP6
-

How can we protect the sensitive knowledge against data mining?

- Types of sensitive knowledge that could be extracted via data mining techniques are
 - Patterns (Association rules, sequences)
 - Clusters that describe the data
 - Classification models for prediction
-

Motivating example

- Rules + background knowledge.
 - Ex: Sniffing Prozac users.
-

Association Rules and k-anonymity (Atzori et. AI. VLDB Journal, 2006)

- Queries vs association rule mining
 - Data anonymity to pattern anonymity

 - A pattern p with support s is k-anonymous if $s > k$
 - A collection of patterns is k-anonymous iff
 - Each pattern is k-anonymous
 - And, no other patterns that are not k-anonymous can be inferred from the patterns
-

Association Rules and k-anonymity

- **Rule1:** Age=27 and ZIP=45254 and religion=Christian => NativeCountry=USA
(sup = 758, conf = 99.8%)
- **Rule2:** Age=27 and ZIP=45254 => NativeCountry=USA

Support (27, 45254, Christian, USA) = 758

Confidence (rule1) = support (rule1)/support(27,45254,Christian)

From above, we can infer the support of (27,45254,Christian) and support of (27,45254,not Christian)

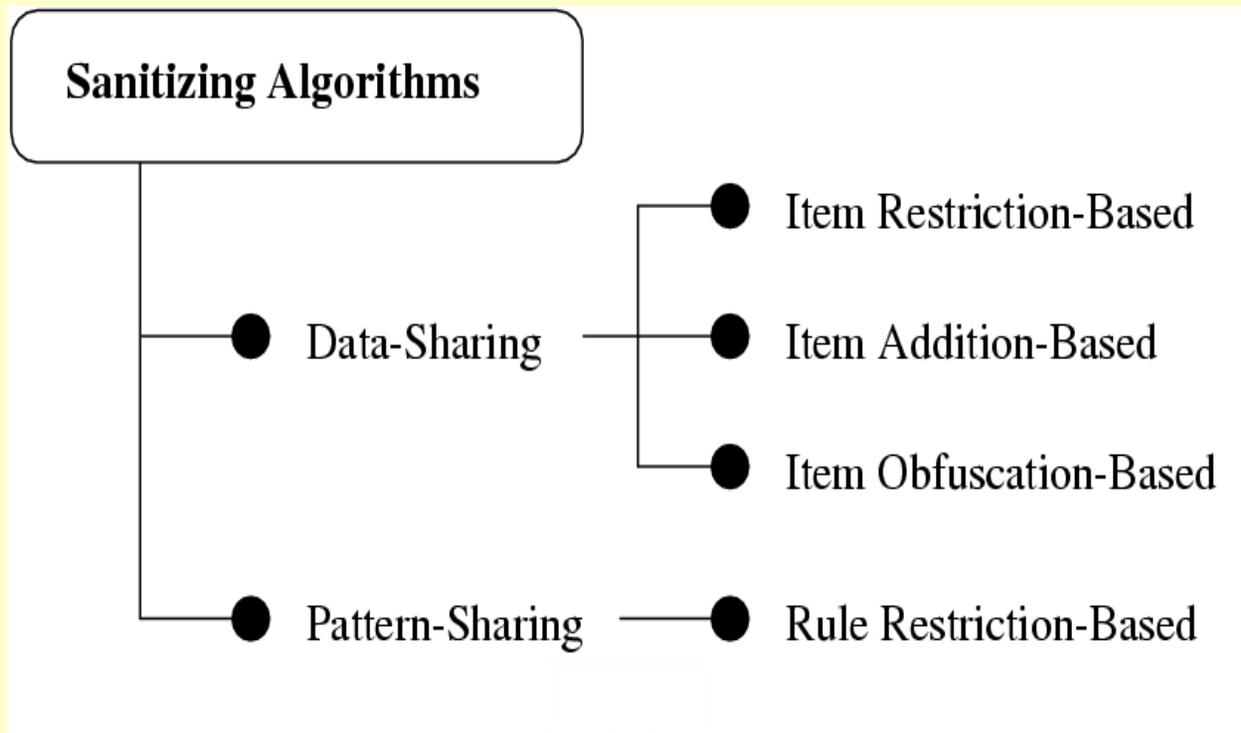
Association Rules and k-anonymity

- **Rule1:** Age=27 and ZIP=45254 and religion=Christian => NativeCountry=USA
(sup = 758, conf = 99.8%)
 - **Rule2:** Age=27 and ZIP=45254 => NativeCountry=USA
 - From the rules above, we can infer the following rule:
 - Age=27 and ZIP=45254 and not (NativeCountry=USA) => Religion=Christian
(sup=1, confidence=100%)
 - Age, post code, and nativeCountry are public and quasi-identifiers in this case.
-

Association Rules and k-anonymity

- Possible solutions:
 - Anonymize the database (anonymize the transactions that are about few people)
 - Anonymize the rules (anonymize the rules about few people)
-

Taxonomy of sanitization algorithms (Oliveira, Zaiane, Saygin PAKDD 2004)

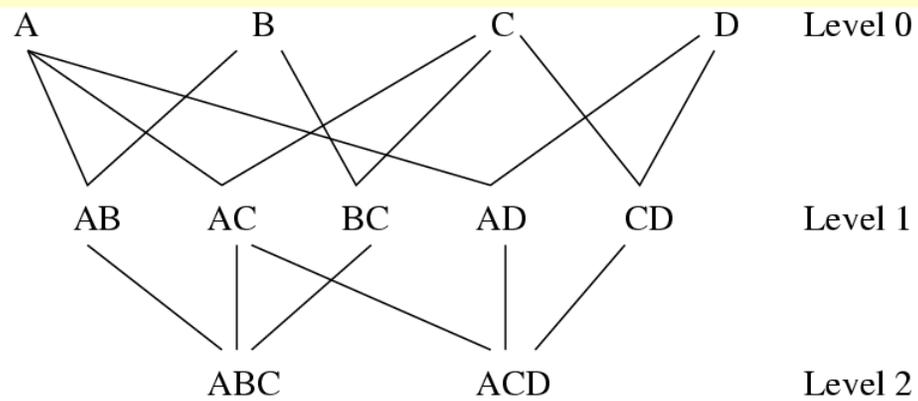


Itemset Lattice

- Sensitive rules

TID	List of Items
T1	A B C D
T2	A B C
T3	A C D
T4	A B C
T5	A B

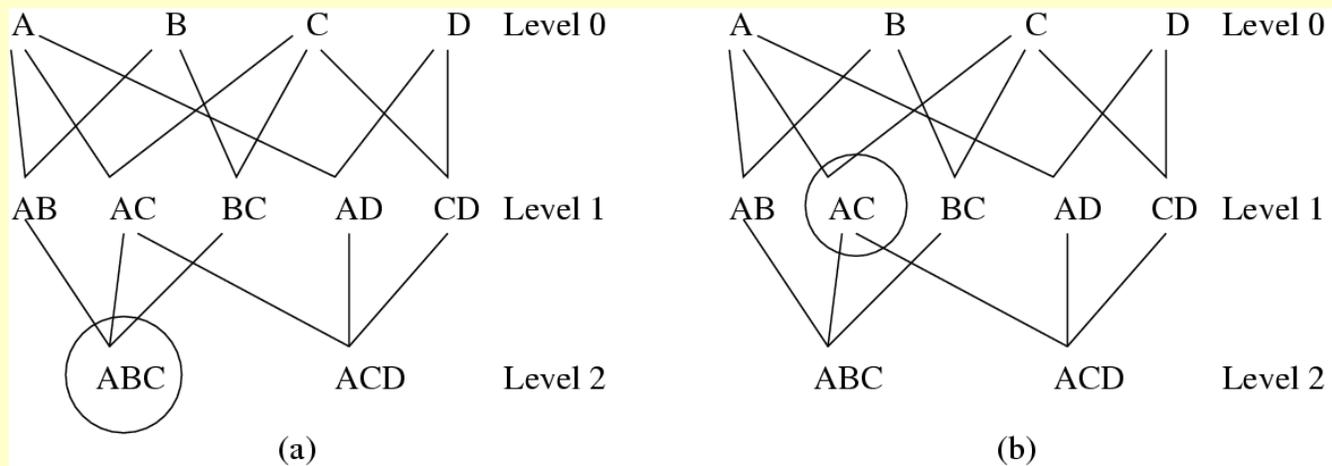
(a)



(b)

Itemset Lattice

- Possible attacks



Association Rule Hiding

- Large amounts of customer transaction data is collected in supermarket chains to find association rules in customer buying patterns
 - lots of research conducted on finding association rules efficiently and tools were developed.
 - Association rule hiding algorithms are deterministic with given support and confidence thresholds
 - Therefore association rules are a good starting point.
-

Association Rule Hiding

- Rules: “**Body => Head**”
 - Ex1: “Diaper => Beer”
 - Ex2: “Internetworking with TCP/IP” => ” Interconnections: bridges, routers,…”
 - parameters: (support, confidence)
 - Minimum Support, and Confidence Thresholds are used to prune the non-significant rules
-

Trans ID	Items Bought
1	A, B, D
2	B, E, F
3	A, C, D
4	A, B
5	A, B, D

Min. support 50%
Min. confidence 70%

Frequent Itemset	Support
{A}	%80
{B}	%80
{D}	%60
{A,B}	%60
{A,D}	%60

Rule	Support
$A \Rightarrow B$	60%
$A \Rightarrow D$	60%

Algorithms for Rule Hiding

- What we try to achieve is:
 - Let D be the source database
 - Let R be the set of significant association rules that are mined from D with certain thresholds
 - Let r_i be a sensitive rule in R
 - Transform D into D' so that all rules in R can still be mined from D' except r_i
 - It was proven that optimal hiding of association rules with minimal side effects is NP-Hard
-

Heuristic Methods

- We developed heuristics to deal with the problem.
 - Different techniques are implemented based on:
 - Modifying the database by inserting false data or by removing some data.
 - Inserting unknown values to fuzzify the rules
-

Basic Approach for Rule Hiding

- Reduce the support of confidential rules
 - Reduce the confidence of rules
 - This way prevent tools to discover these rules
 - The challenge is the data quality
 - metric for data quality is the number of rules that can still be mined and the number of rules that appear as a side effect
 - heuristic algorithms to minimize the newly appearing rules, and to minimize the accidentally hidden rules.
-

Basics of the heuristic algorithms

- If we want to remove an item from a transaction to reduce the support or the confidence
 - Which item should we start from
 - Which transaction should we choose to hide the selected item
 - We can either
 - Select an item and a transaction in round robin fashion, I.e., select the next item from the next transaction that supports that item, and move to another item and another transaction.
 - Select the item that will probably have the minimal impact on the other rules
-

Basics of rule hiding

- $\text{conf}(X \Rightarrow Y) = \text{sup}(XY) / \text{sup}(X)$
 - Decreasing the confidence of a rule can be done by:
 - Increasing the support of X in transactions not supporting Y
 - Decreasing the support of Y in transactions supporting both X and Y
 - Decreasing support of rule can be done by
 - Decreasing the support of the corresponding large itemset XY
-

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Min. support 20%
Min. confidence 80%

Rule	Confidence
$AB \Rightarrow C$	100%
$BC \Rightarrow A$	100%

Hiding $AB \rightarrow C$ by increasing support of AB

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A, B
5	B

Rule	Confidence
$AB \Rightarrow C$	66%
$BC \Rightarrow A$	100%

Hiding $AB \rightarrow C$ by decreasing support of ABC

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Trans ID	Items Bought
1	A, C
2	A, B
3	A, C
4	A, B
5	B

Rule	Confidence
$AB \Rightarrow C$	0%
$BC \Rightarrow A$	0%

Hiding $AB \rightarrow C$ by decreasing the support of C

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Trans ID	Items Bought
1	A, B
2	A, B, C
3	A, C
4	A, B
5	B

Rule	Confidence
$AB \Rightarrow C$	50%
$BC \Rightarrow A$	100%

Rule Hiding by Fuzzification

- In some applications where publishing wrong data is not acceptable, then unknown values may be inserted to blur the rules.
 - When unknown values are inserted, support and confidence values would fall into a range instead of a fixed value.
 - Similar heuristics for rule hiding can be employed to minimize the side effects
-

TID	A	B	C	D
1	1	1	0	1
2	0	1	0	0
3	1	0	1	1
4	1	1	0	0
5	1	1	0	1

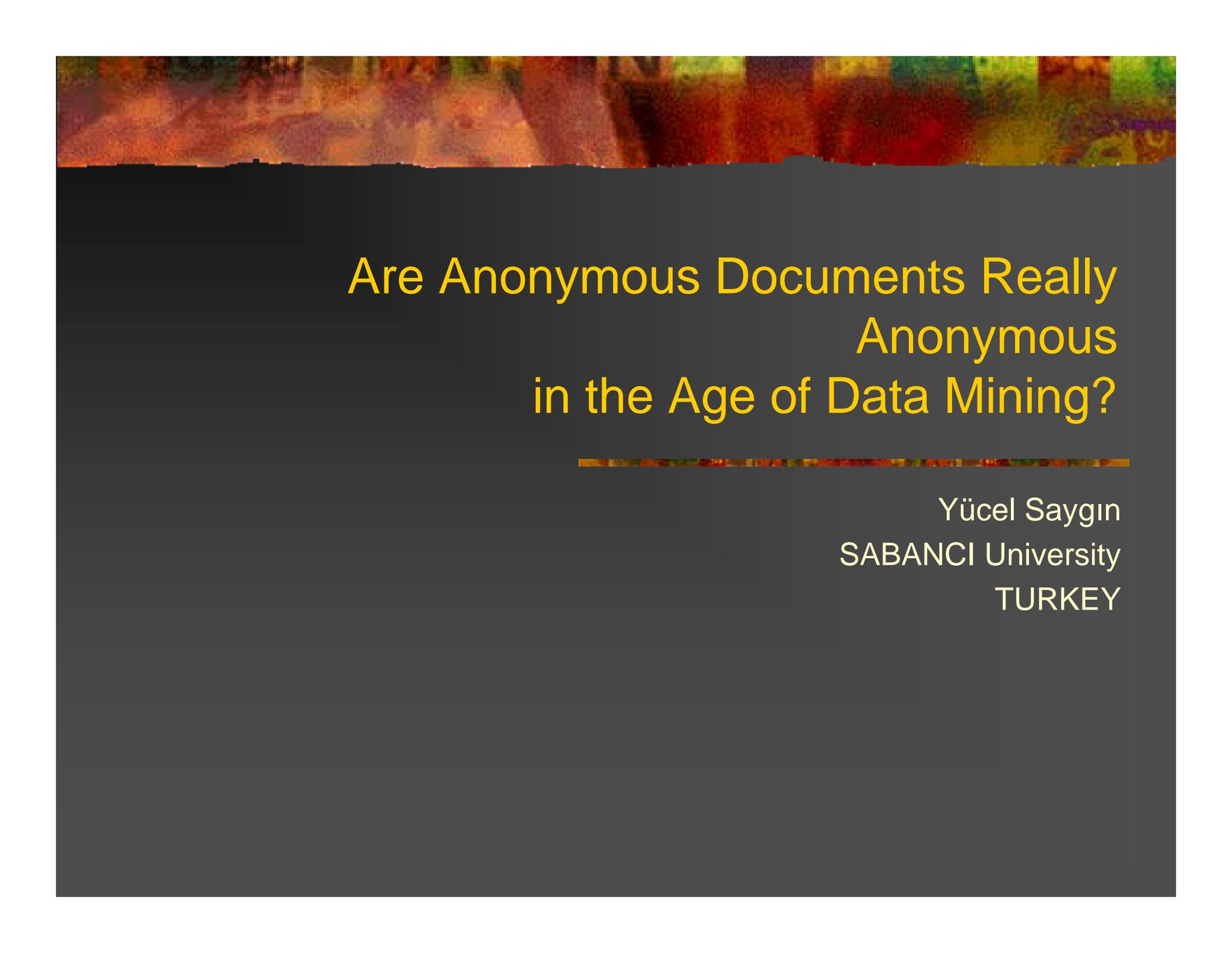
Support and confidence
Becomes a range of values

TID	A	B	C	D
1	?	1	0	1
2	0	1	0	0
3	1	0	1	?
4	1	?	0	0
5	1	?	0	1

Exercise: Try to calculate
The support and confidence
Ranges from the database
With ? symbols

Classification model as a threat to privacy

- The fact that each author uses a characteristic frequency distribution over words and phrases helps us
 - Feature representation used:
 - T: total number of tokens
 - V: total number of types
 - C: total number of characters
 - Classify the document by a learning algorithm and then try to perturb the classification
-



Are Anonymous Documents Really Anonymous in the Age of Data Mining?

Yücel Saygın
SABANCI University
TURKEY

Motivation

- Document collection efforts
 - Internet: emails, newsgroups
 - Critical applications such as health care
 - Documents may contain sensitive private information
 - What if documents need to be disclosed to third parties?
-

Sanitization and Anonymization is needed before disclosure!



Task of sanitization and anonymization is to remove personally identifying information from text and make sure that the text itself cannot be identified with an individual (whenever necessary)

Private or personally identifying information in documents

- Explicit private information can be removed by hand
 - For example, instead of writing the name and surname, just write the initials.
 - Ex: “G.O. (52 yo) was sexually harassed in her office last week. G.O. who is a professor in the law department of Trabzon university in Turkey told the police that...”
 - We need better mechanisms than that!
-

Private or personally identifying information in documents

- Explicit private information
 - names of persons, phone numbers, addresses etc
 - Relationships between entities (people, events, occurrence in a document etc.)
- Explicit private information can be anything that could be identified with a person.
 - Ex1 : for a report the author could be private info.
 - Ex2 : “Dr. John Marshall lives in 12th Broadway, Manhattan, NY”
address could be confidential.
 - Ex3: “John Marshall met with Edward Jacobs in 1997”.
The fact that two people met could be confidential.

Private information in documents

- Implicit Information
 - Statistical information, data mining results



Private information in documents

- Implicit Information
 - Statistical information, data mining results
 - Ex1: call center statistics may tell that 50% of the customers are not happy with the service provided by a company
 - Ex2: The author of a document, category of a document
-

Private information in documents

- Combination of explicit and implicit info
 - Name of a person appears in a document with a certain category
 - In order to fully anonymize a document, we need to remove all personally identifying information
-

Sanitization and Anonymization needed before disclosure!



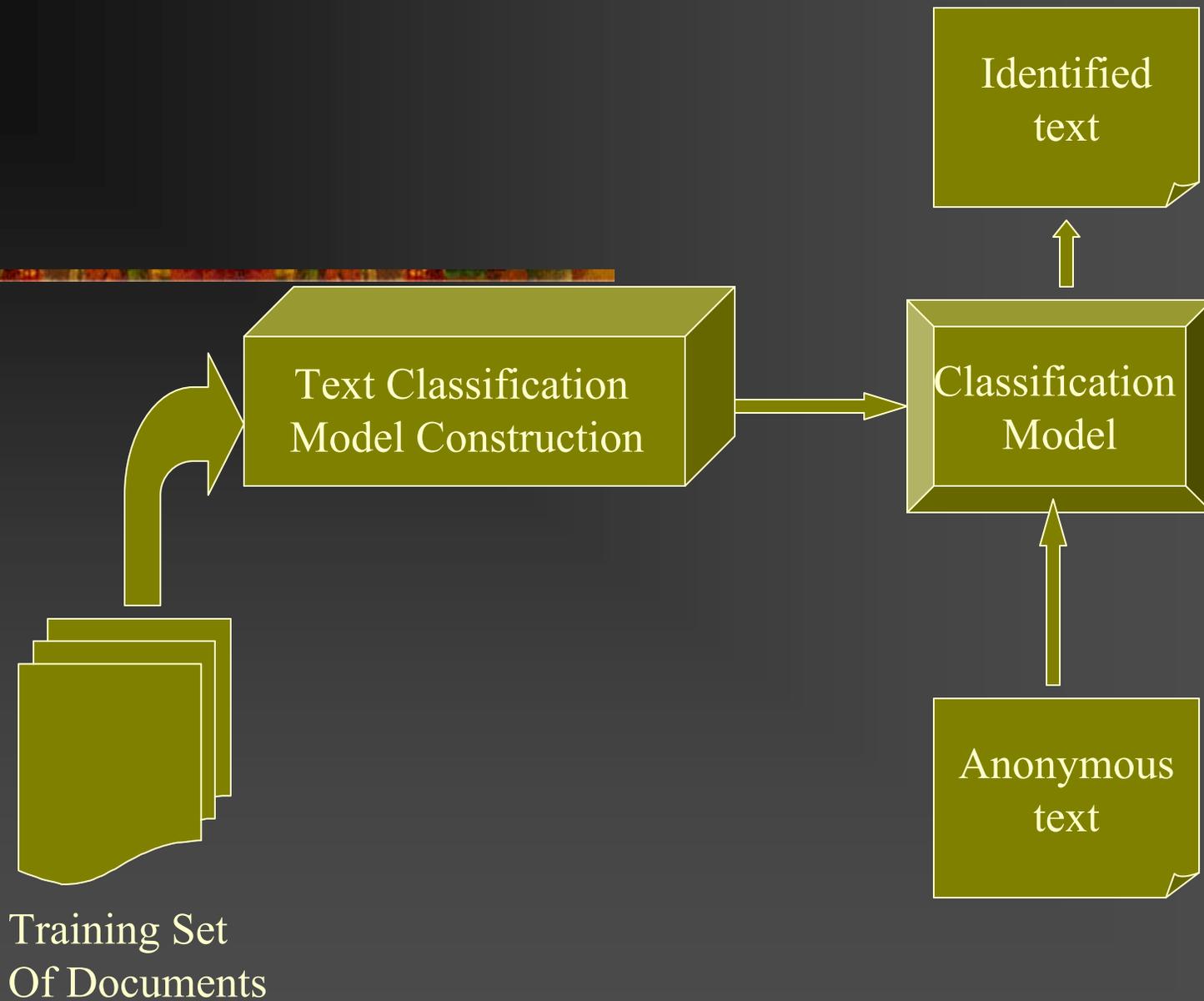
Task of sanitization and anonymization needs to consider data mining techniques as well!

Possible threats to privacy in docs

- Direct threats: against explicit information that could be obtained with browsers or querying tools
 - Web Crawlers are doing a terrific job searching through hypertext and collect everything they can.
 - What if someone accidentally stores his/her emails in the public_html directory?
 - Search the internet to find the phone number of a person
 - Lots of email messages, homepages etc are indexed
 - Just google a person name using your browser
-

Implicit information and indirect threats

- Indirect threats : against implicit information, could be obtained by data mining tools
 - Ex: the authorship information for a document
 1. If you know the possible authors, you can collect a set of documents written by these authors
 2. Build a classification model out of these documents to profile the set of authors
 3. Use the model to predict the author of the anonymous text
 - Even if a document is anonymized, the authorship information could still be obtained
-



Implicit information and indirect threats

- Ex: Conferences are doing anonymous peer reviewing, i.e., both the reviewers and authors of papers are anonymous
 - In order to find the identify the reviewer
 - Get the list of the program committee (available at the conference web site)
 - Obtain a list of papers written by the committee (reviewers are among PC members) from the internet
 - Use this list to train a classification model
 - Feed the (bad) review you got to the classification model to get the reviewer identity
-

Implicit information and indirect threats

- For example
 - we selected a Turkish news paper and trained a SVM using features such as the specific words used, punctuation marks, average word length etc.
 - Using the SVM we were able to predict the author of the articles in the newspaper with more than 90% accuracy
-

How do we defeat direct threats?

- Better to automate the process instead of people doing it (more privacy, and less expensive)
 - Detecting private (sensitive) explicit information
 - Removing private information
 - Completely deleting it (not a good approach for text, watch for data quality too)
 - Generalize the private information
 - Ex: Replace the actual name of a person with tags like <person>
 - Ex: Generalizing an address by removing the street number, and the zip code
-

How do we avoid direct threats?

- Detecting private (sensitive) explicit information
 - By Named Entity Extraction (NEA) from NLP
 - Named Entity : Atomic elements of information in text. Such as: person names, phone numbers
 - Named entities are specified by grammars which are then converted to finite state machines to extract them from text
 - NEA is still not very accurate but can be tuned for specific applications
-

How do we avoid direct threats?

- Relationships between named entities
 - Ex: A person being a friend with another person
 - Much harder to extract
 - Need to first extract named entities, identify the relationship and go from there
 - Relationships among NEs can be identified using association rule mining
 - Exercise: Develop a process for association rule mining from text documents
 - Identify the type of data
 - Identify the association type
 - Preprocessing
-

How do we avoid indirect threats?

- Data mining to detect indirect threats
 - Ex: text classification to identify author of a text
 - Updating the text to disturb the accuracy of classifiers
 - Identify the features that play an important role in classification and try to make them homogeneous in the training set
 - Change words by synonyms, or rephrase the sentences
-

How do we avoid indirect threats?

- Difficulties:
 - Different classification methods, hard to find an anonymization method for all
 - Quality of text should not be degraded
-

Privacy in Documents

- Data mining tools can actually be used to reach private implicit information
 - Anonymized documents may not really anonymous wrt text mining tools
 - NLP, data mining should work hand in hand to overcome the threats for full anonymization
-

Motivation, cont.

- Removing confidential data values from the data set is not enough to suppress it
 - Using certain data mining techniques it is still possible to infer the removed data values
 - e.g. Classification
-

Motivating Example : Healthcare Research Application

Cell Size Uniformity	Cell Shape Uniformity	Gender	Clump Thickness	Diagnosis
1	1	F	1	Benign
1	1	F	2	Benign
2	2	M	1	Malignant
...				



Motivating Example, cont.

- Suppose that the following data set will be disclosed
 - But the confidential data value must be suppressed
 - The confidential data values might be replaced with a “?” denoting unknown
 - Not enough ☹️

Cell Size Uniformity	Cell Shape Uniformity	Gender	Clump Thickness	Diagnosis
2	2	M	1	?
1	1	F	2	Benign
1	1	F	1	Benign
2	1	M	1	Malignant
1	2	F	2	Malignant
2	2	M	1	Benign
2	2	F	1	Malignant

Mining the Data Without Seeing it

- Privacy Preserving Mining of Association Rules in Distributed Environments
 - Reference: *Murat Kantarcioglu, Chris Clifton, "Privacy-Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data," IEEE Transactions on Knowledge and Data Engineering, vol. 16, no. 9, pp. 1026-1037, Sept., 2004.*
-

Data Distribution (Horizontal)

Data

TID	A	B	C	D
1	1	1	0	1
2	0	1	0	0
3	1	0	1	1
4	1	1	0	0
5	1	1	0	1

Site1

TID	A	B	C	D
1	1	1	0	1
2	0	1	0	0

Site2

TID	A	B	C	D
3	1	0	1	1
4	1	1	0	0
5	1	1	0	1

Data Distribution (Vertical)

Data

TID	A	B	C	D
1	1	1	0	1
2	0	1	0	0
3	1	0	1	1
4	1	1	0	0
5	1	1	0	1

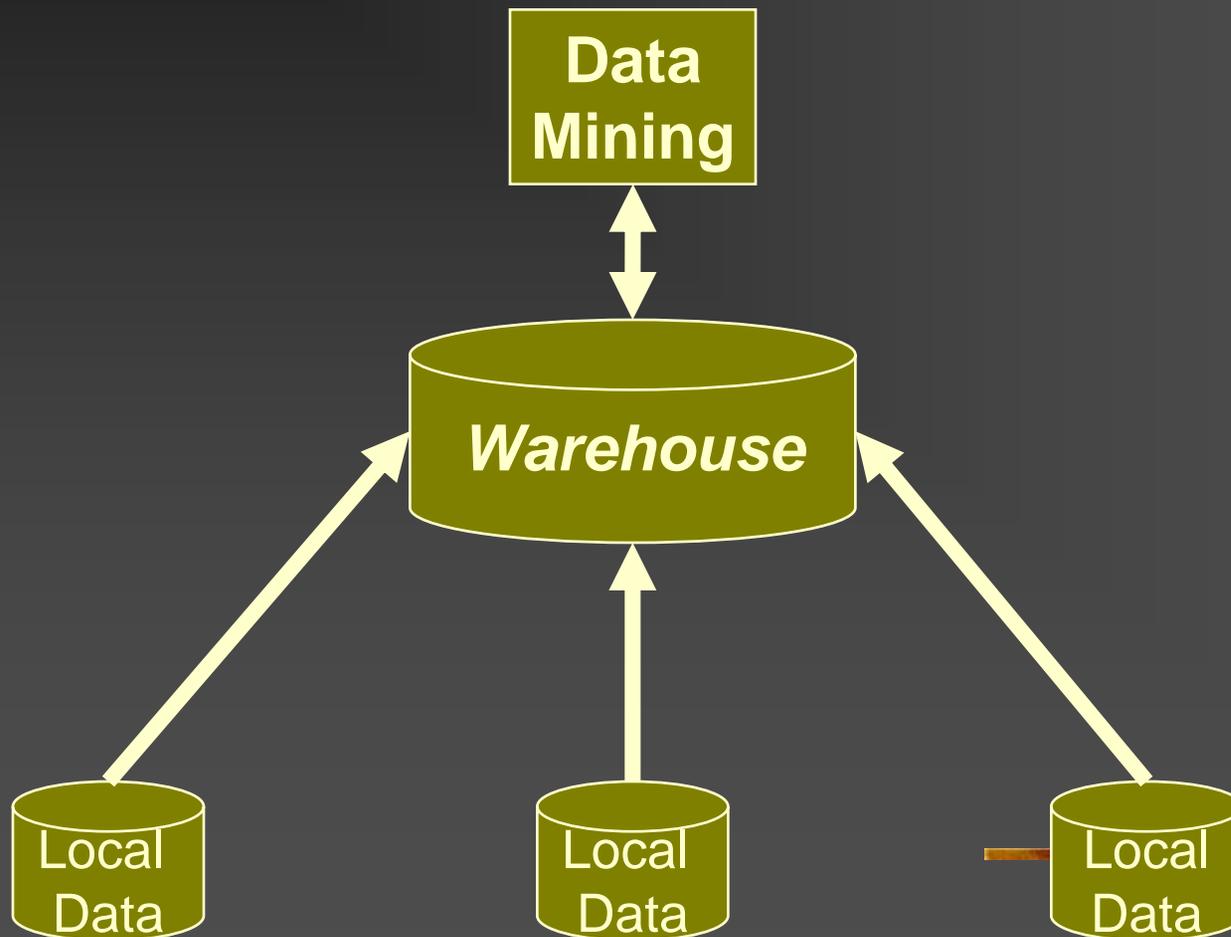
Site1

TID	A	B
1	1	1
2	0	1
3	1	0
4	1	1
5	1	1

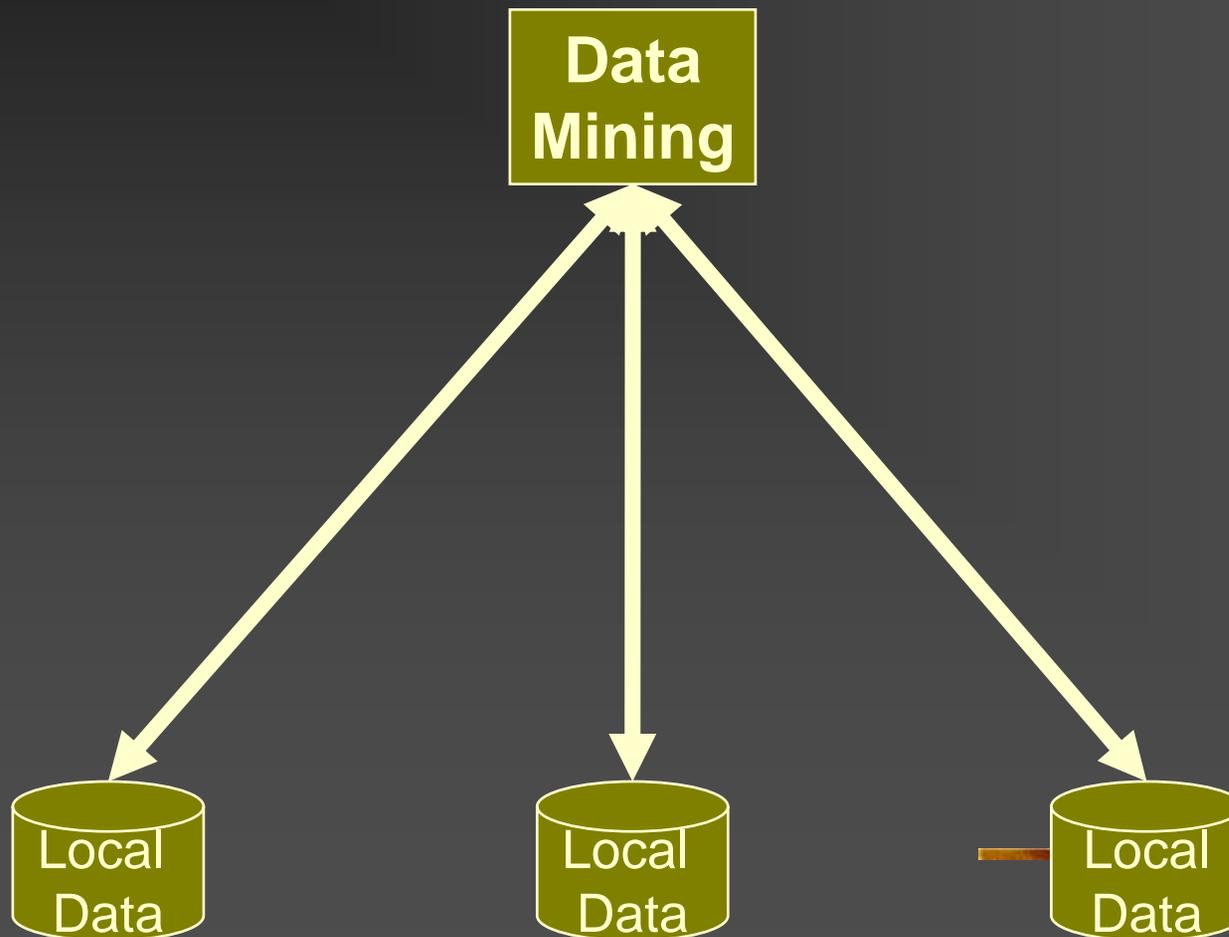
Site2

TID	C	D
1	0	1
2	0	0
3	1	1
4	0	0
5	0	1

Centralized Data Mining



Distributed Mining



Distributed rule mining

- We know how rules are mined from centralized databases
- The distributed scenario is similar
- Consider that we have only two sites S1 and S2, which have databases D1 (with 3 transactions) and D2 (with 5 transactions)

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Distributed rule mining

- We would like to mine the databases as if they are parts of a single centralized database of 8 transactions
- In order to do this, we need to calculate the local supports
 - For example the local support of A in D1 is 100%
 - The local support of the itemset {A,B,C} in D1 is 66%, and the local support of {A,B,C} in D2 is 40%.

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C
4	A
5	B

Distributed rule mining

- Assume that the minimum support threshold is 50% then {A,B,C} is frequent in D1, but it is not frequent in D2.
- However when we assume that the databases are combined then the support of {A,B,C} in D1 U D2 is 50%
- which means that an itemset could be locally frequent in one database, but not frequent in another database. And it can be frequent globally
- In order for an itemset to be frequent globally, it should be frequent in at least one database

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C

Trans ID	Items Bought
4	A, B, C
5	A, B, C
6	A, C
7	A
8	B

Distributed rule mining

- The algorithm is based on apriori which prunes the rules by looking at the support
- Apriori also uses the fact that an itemset is frequent only if all its subsets are frequent
- Therefore only frequent itemsets should be used to generate larger frequent itemsets

Trans ID	Items Bought
1	A, B, C
2	A, B, C
3	A, C

Trans ID	Items Bought
4	A, B, C
5	A, B, C
6	A, C
7	A
8	B

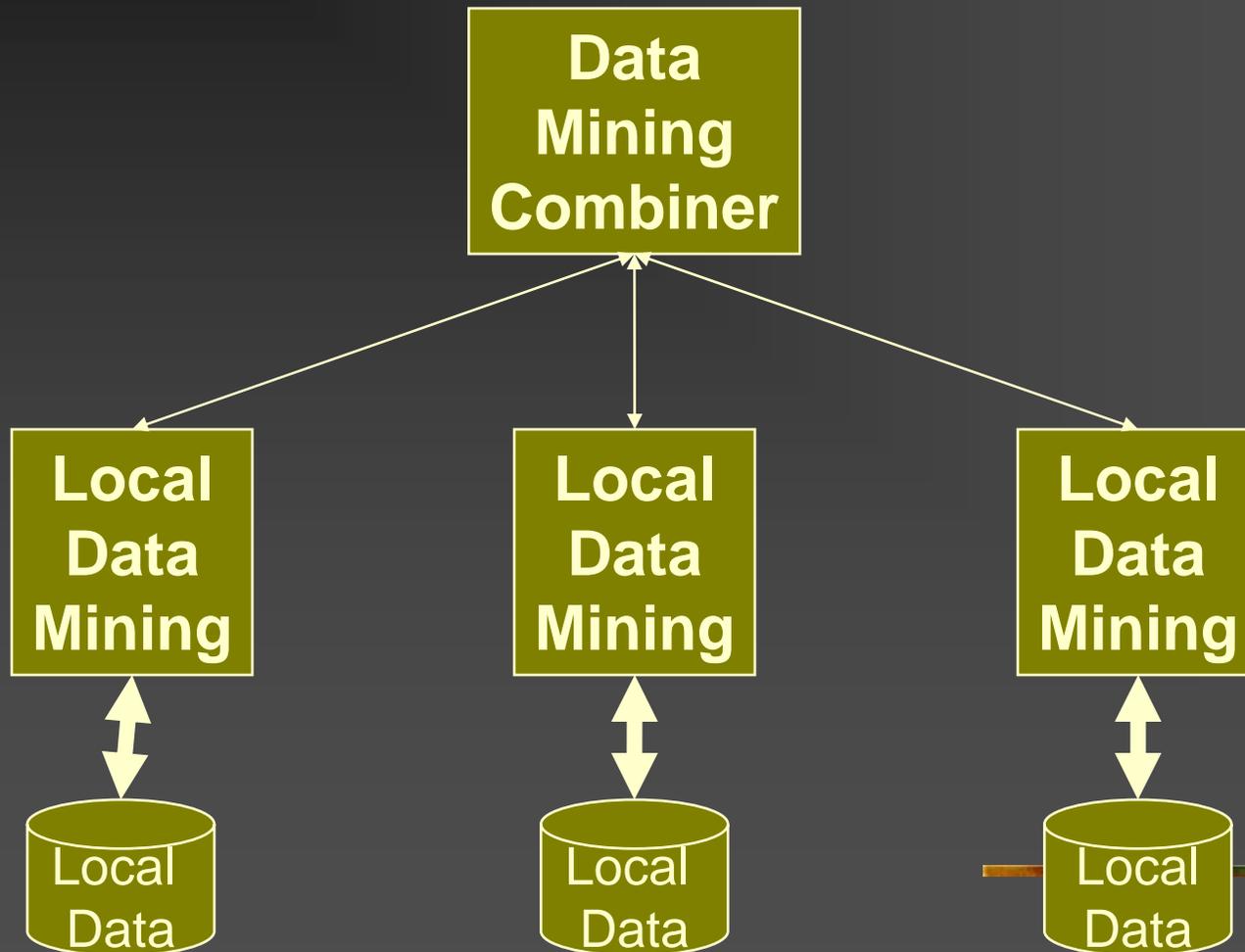
Distributed Mining Overview

- We will look at horizontally partitioned case
 - Lemma: if a rule has support $> s\%$ globally, it must have support $> s\%$ in at least one of the sites.
 - Basic algorithm
 1. First set the global support threshold (s) which is also used as the local support
 2. Each site sends its locally frequent rules to a coordinator site together with the total number of transactions they have
 3. Coordinator distributes the collected rules to all the sites
 4. Sites count the support of the received rules
 5. Sites send the supports of the rules back to the coordinator
 6. Coordinator calculates the global supports and broadcasts the rules back to the sites
-

Distributed Data Mining and Privacy

- Distributed Data Mining does not require the disclosure of data
 - Only rules are disclosed
 - But rules themselves may disclose private information about the sites
 - Ex: 2% of the patients who had a heart surgery died of heart attack in Hospital A
 - Therefore, companies would like to receive the global rules, but they do not want to disclose the supports and confidence of the individual rules.
 - So we need a method to mine the rule without sites seeing the data or the rules of each other
-

Private Distributed Mining:



The assumptions

- There are more than two parties (two party case needs special treatment)
 - Parties are non-colluding
 - Parties follow the protocol for privacy preserving distributed mining
 - Parties can be curious
-

Basic Tools

- One way hash functions
 - Ex: Secure comparison of two numbers:
 - Person1 has : N
 - Person 2 has : M
 - They want to check if $M=N$
 - But they do not want to disclose their numbers, or the fact that $M>N$, or $N>M$
 - The only information that could be revealed is equality
 - How can we do that?
-

Basic Tools

- Can we generalize this and do secure set intersection?
 - Two sites S, and R have a single column table
 - They want to learn which values are in the intersection set
 - But they do not want to disclose the their sets to each other
 - What are possible attacks in case of a small domain?
-

Basic Tools

- Encryption: $E(A)$
 - An encryption is commutative if the following two equations hold for any given feasible encryption keys K_1, K_2, \dots, K_n , any M , and any permutations of i, j
 - $E_{K_{i1}}(\dots E_{K_{in}}(M)) = E_{K_{j1}}(\dots E_{K_{jn}}(M))$
 - For different M_1 , and M_2 the probability of collusion is very low
 - For the case of two keys, $E_{k1}(E_{k2}(A)) = E_{k2}(E_{k1}(A))$
 - RSA is a famous commutative encryption technique
 - Ex: Secure comparison of two numbers (A , and B) with commutative encryption. Assume that A , and B have their own encryption keys, say K_1 , and K_2 . And we go from there
-

Distributed PP Association Rule Mining

- For distributed association rule mining, each site needs to distribute its locally frequent itemsets to the rest of the sites
 - Instead of circulating the actual itemsets, the encrypted versions are circulated
 - Example:
 - S1 contains A, S2 contains B, S3 contains A. Each of them have their own keys, K1, K2, K3.
 - At the end of step 1, each all sites will have items encrypted by all sites.
 - The encrypted items are then passed to a common site to eliminate the duplicates and to start decryption. This way they will not know who has sent which item.
 - Decryption can now start and after everybody finished decrypting, then they will have the actual items.
-

Distributed PP Association Rule Mining

- Now we need to see if the global support of an item is larger than the threshold.
- We do not want to reveal the supports, since support of an item is assumed to be confidential.
- A secure multi-party computation technique is utilized for this
 - Assume that there are three sites, and each of them has {A,B,C} and freq in S1 is 5 (out of 100 transactions), in S2 is 6 (out of 300), and in S3 20 (out of 300), and minimum support is 5%.
 - S1 selects a random number, say 17
 - S1 adds the difference $5 - 5\% \times 100$ to 17 and sends the result (17) to S2
 - S2 adds $6 - 5\% \times 200$ to 17 and sends the result (13) to S3.
 - S3 adds $20 - 5\% \times 300$ to 13 and sends the result (18) back to S1
 - $18 >$ the chosen random number (17), so {A,B,C} is globally frequent.

Privacy Preserving Clustering

- Ali Inan, Yucel Saygin, Erkay Savas, Ayca Azgin Hintoglu, Albert Levi. *Privacy Preserving Clustering on Horizontally Partitioned Data*. Proceedings of the IEEE International Workshop on Privacy Data Management , 2006.
-

MOTIVATION

Patient 1

Patient 2



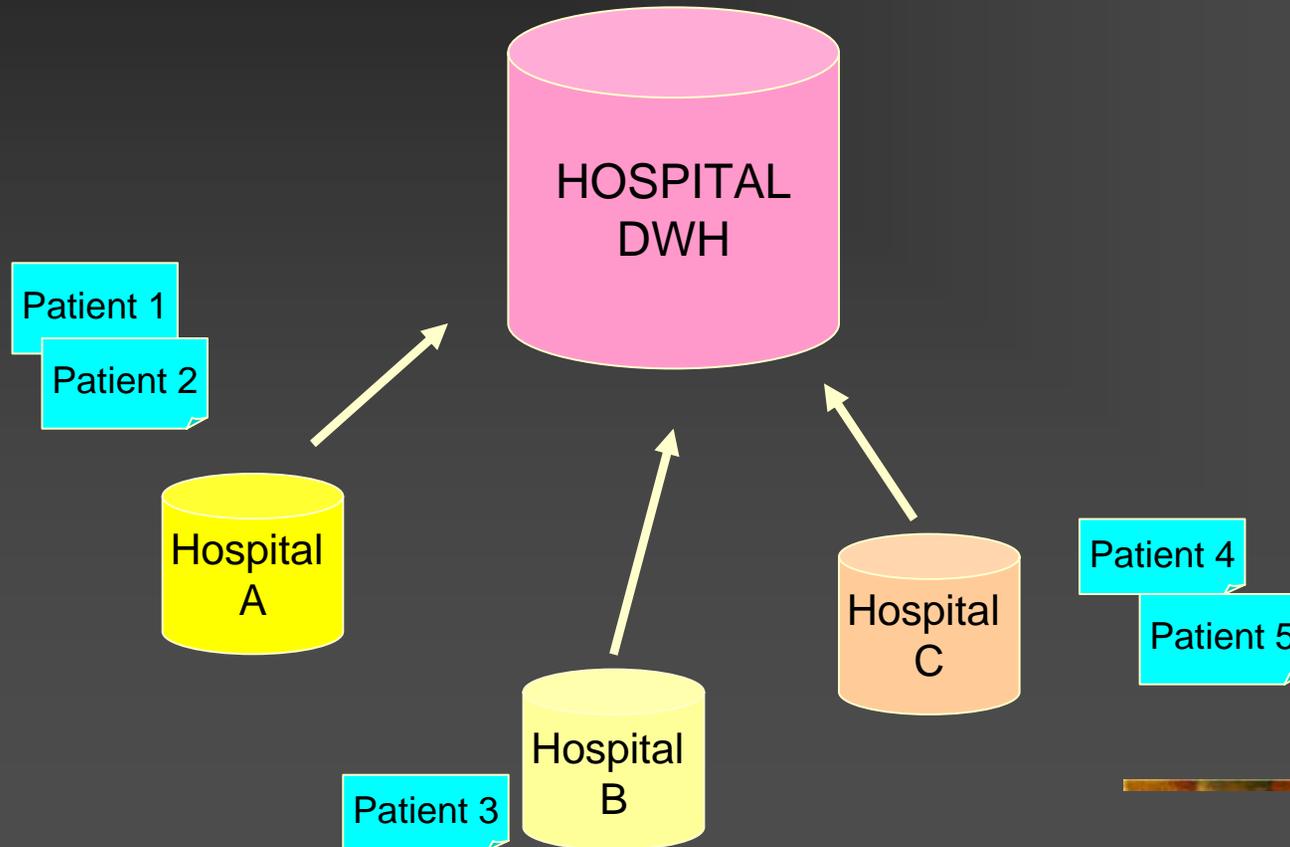
Patient 3



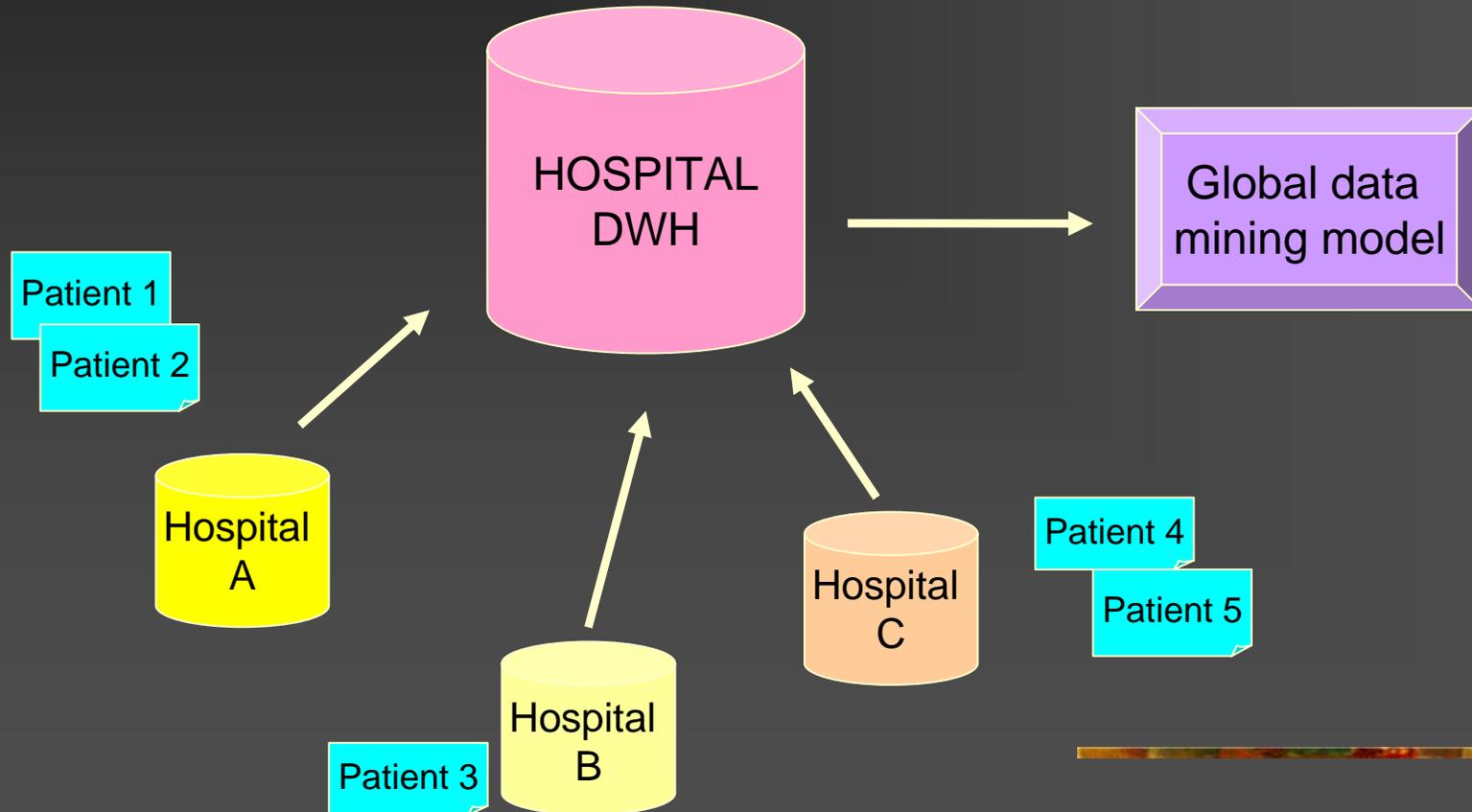
Patient 4

Patient 5

MOTIVATION



MOTIVATION



MOTIVATION

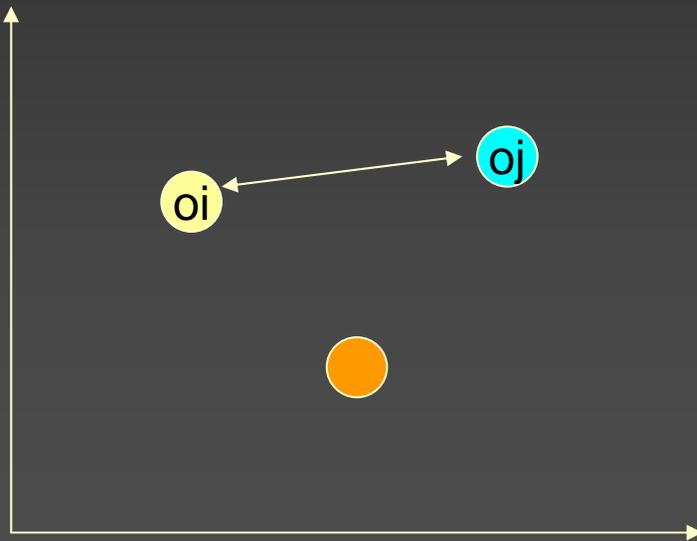
- Constructing a warehouse of patients may not be possible due to privacy concerns
 - In some applications companies may not want to share their customer data.
 - Solution is :
 - PRIVACY PRESERVING DATA MINING
 - ON HORIZONTALLY PARTITIONED DATA
 - WE ARE PARTICULARLY INTERESTED IN CLUSTERING
-

Related Work

- Sanitization techniques:
 - Trade-off accuracy versus privacy
 - Secure Multi-Party Computation (SMC) techniques:
 - High computation/communication costs
 - No accuracy loss
 - K-means clustering protocol by Kruger et al.
 - We propose an SMC based clustering protocol
-

Clustering

- Grouping similar objects together in multidimensional space
- Requires a dissimilarity matrix as input, which stores the distance between objects



Contributions

- Secure comparison protocols for numeric, categorical and **alphanumeric** attributes
 - DNA clustering
- SMC protocol for constructing the dissimilarity matrix of horizontally partitioned data
- Proof of security of the protocols and communication cost analysis
- Proof-of-concept implementation

Data Matrix

- Object-by-variable structure
- Rows are scattered among data holders in case of horizontally partitioned data
- $m \times n$ matrix: m objects on n attributes

$$D = \begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

Dissimilarity Matrix

- Object-by-object structure
- $\text{dist}(i, j)$ is the comparison function, returning the distance
- $d[i][j]$ is the distance between objects i and j
- $m \times m$ matrix: pair-wise comparison of m objects

$$d = \begin{bmatrix} 0 & & & & & \\ \text{dist}(2,1) & 0 & & & & \\ \text{dist}(3,1) & \text{dist}(3,2) & 0 & & & \\ \vdots & \vdots & \vdots & 0 & & \\ \text{dist}(m,1) & \text{dist}(m,2) & \dots & \dots & 0 & \end{bmatrix}$$

Example: Secure Comparison

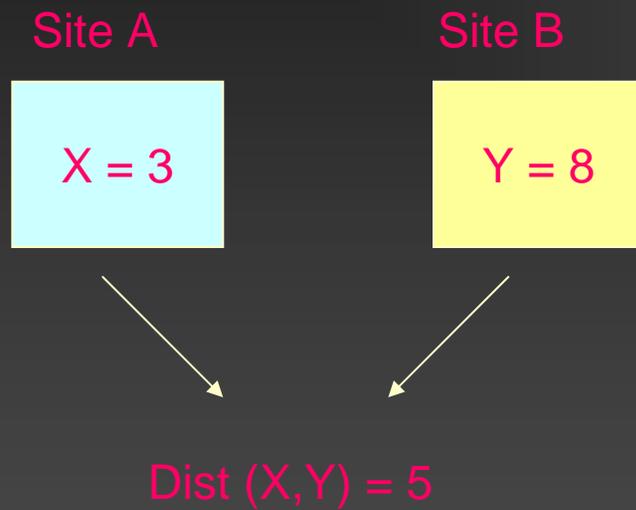
Site A

$X = 3$

Site B

$Y = 8$

Example: Secure Comparison



Example: Secure Comparison

Site A

$X = 3$
 $R1 = 7$

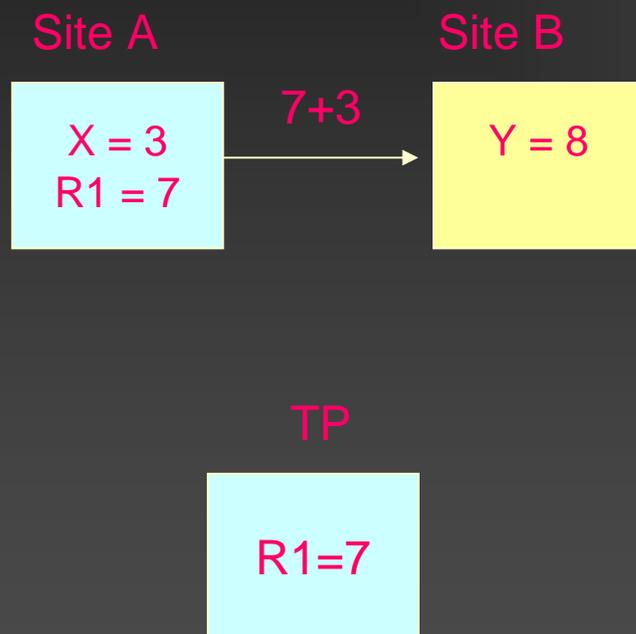
Site B

$Y = 8$

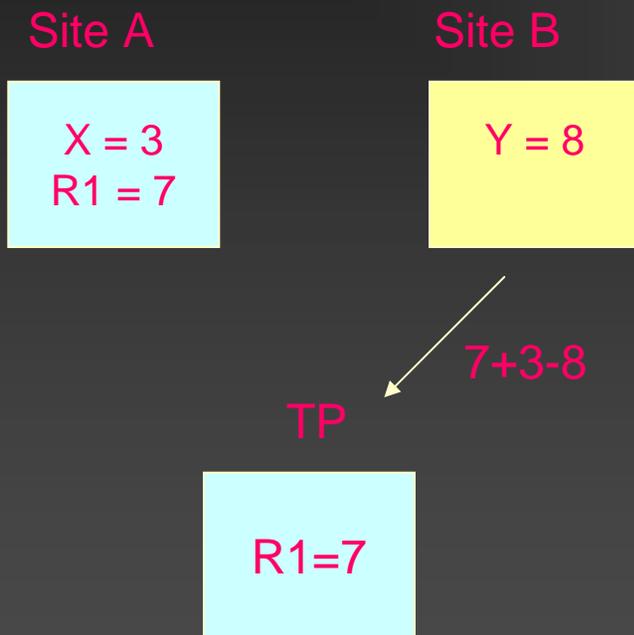
TP

$R1 = 7$

Example: Secure Comparison



Example: Secure Comparison



Example: Secure Comparison

Site A

$X = 3$
 $R1 = 7$

Site B

$Y = 8$

TP

$7+3-8-7$
 $= -5$

Example: Secure Comparison

Site A

$X = 3$
 $R1 = 7$

Site B

$Y = 8$

TP

$7+3-8-7$
 $= -5$

Problem: TP learns that $Y > X$

ex: sales of company A is greater
than the sales of company B

Example: Secure Comparison

Site A

X = 3
R1 = 7
R2 = 5

Site B

Y = 8
R2 = 5

TP

$7 + 3 - 8 - 7$
 $= -5$

- Use another random number R2 that can be shared Between A and B
- If R2 is odd then site A negates its value otherwise B negates its value

Example: Secure Comparison

Site A

X = 3
R1 = 7
R2 = 5

Site B

Y = 8
R2 = 5

TP

$7 + 3 - 8 - 7$
 $= -5$

Another problem: if A and TP use the same random number for all the distance calculations, then site B can do a statistical attack to recover the value X

USE a pseudo-random number generator that generates sequence of random numbers one for each distance calculation.

MAIN IDEA

- Build the dissimilarity matrix from the objects distributed to multiple sites using the secure difference calculation
 - Dissimilarity (Distance) matrix will just contain the object IDs and the distances, therefore will not reveal private information.
 - We are also going to show how categorical and alphanumeric attributes are handled.
-

Problem Definition

- There are $k \geq 2$ data holder parties and a third party (TP)
- Every party is semi-trusted
 - Parties follow the protocol as they are supposed to do
 - But they may store any data revealed to them in order to infer private information
 - Also called honest-but-curious behavior
- Involved parties are non-colluding
 - Parties can not share private information among themselves

Problem Definition (cont'd)

- Third party
 - Has access to comparison functions
 - Shares a secret key with every data holder
 - Governs the communication btw. data holders
 - Builds the dissimilarity matrix, d
 - Clusters according to d
 - Publishes results
- Data holder parties
 - Has access to comparison functions
 - Share a secret key with every data holder (and TP)

Distance Metrics

- Numeric Data:

$$\text{distance}(x, y) = |x - y|$$

- Categorical Data:

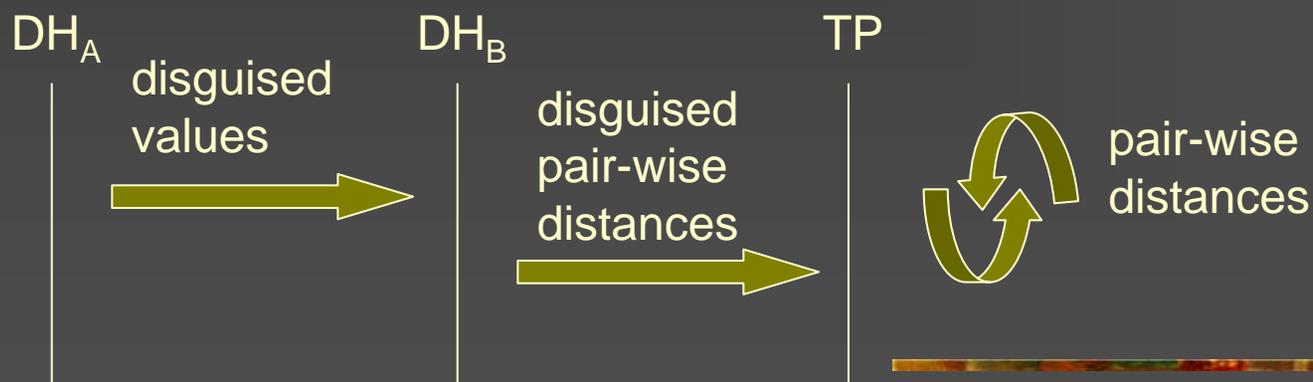
$$\text{distance}(a, b) = \begin{cases} 0, & \text{if } a = b; \\ 1 & \text{otherwise.} \end{cases}$$

- Alphanumeric Data:

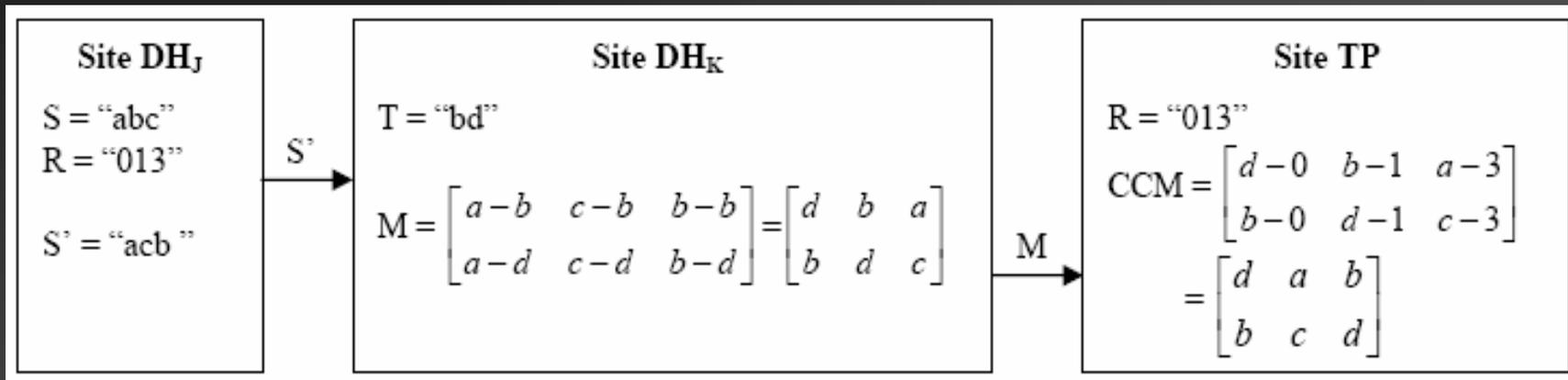
- Any form of edit distance

Privacy Preserving Comparison Protocols

- Involved parties: data holders, DH_A and DH_B , and TP
- Pseudo-random number generators
 - Disguise private data
 - R_{AB} : shared among DH_A and DH_B (only for numeric data)
 - R_{AT} : shared among DH_A and TP



Alphanumeric Data Comparison



- S : DH_j 's private input
- T : DH_k 's private input
- R : random vector shared btw. DH_j and TP
- $CCM[i][j] = 'a'$ implies $S[i] = T[j]$
 - Adequate to compute any form of Edit distance for S and T

Dissimilarity Matrix Construction

- TP also needs local dissimilarity matrices of data holders
 - Entry $d[i][j]$ of dissimilarity matrix d
 - If i and j are held by different sites
 - TP computes $\text{dist}(i,j)$ through the protocol btw. these sites
 - Else if i and j are held by the same site
 - $\text{dist}(i,j)$ is in this sites local dissimilarity matrix
-

Privacy at site TP

- TP knows the distance between each data object
- But has no information on any data value
- TP can not infer any private data
 - Proof: given the distance d between two data points, one can generate infinitely many pairs of data points that are d distant from each other
 - Corollary: data holders can simply build their local dissimilarity matrices and share it with TP so as to decrease the number of private comparisons

Publishing the Results

- Each data holder party can assign different weights to attributes in our protocol
- TP keeps the dissimilarity matrix secret
 - Since data holders have a portion of private data, they can use dissimilarity matrix to infer other private information
- TP publishes clustering results and also parameters for computing the quality of the clustering such as average square distance btw. cluster members

Complexity Analysis

- Numeric data

	Data Holder J	Data Holder K
Data Size	n	m
Local Diss. Matrix	$O(n^2)$	$O(m^2)$
Comparison Protocol	$O(n)$	$O(m*n)$
Total	$O(n^2)$	$O(m^2+m*n)$

- Alphanumeric data

	Data Holder J	Data Holder K
Data Size	n	m
Max. String Length	p	q
Local Diss. Matrix	$O(n^2)$	$O(m^2)$
Comparison Protocol	$O(n*p)$	$O(m*n*p*q)$
Total	$O(n^2+n*p)$	$O(m^2+m*n*p*q)$

Possible attacks

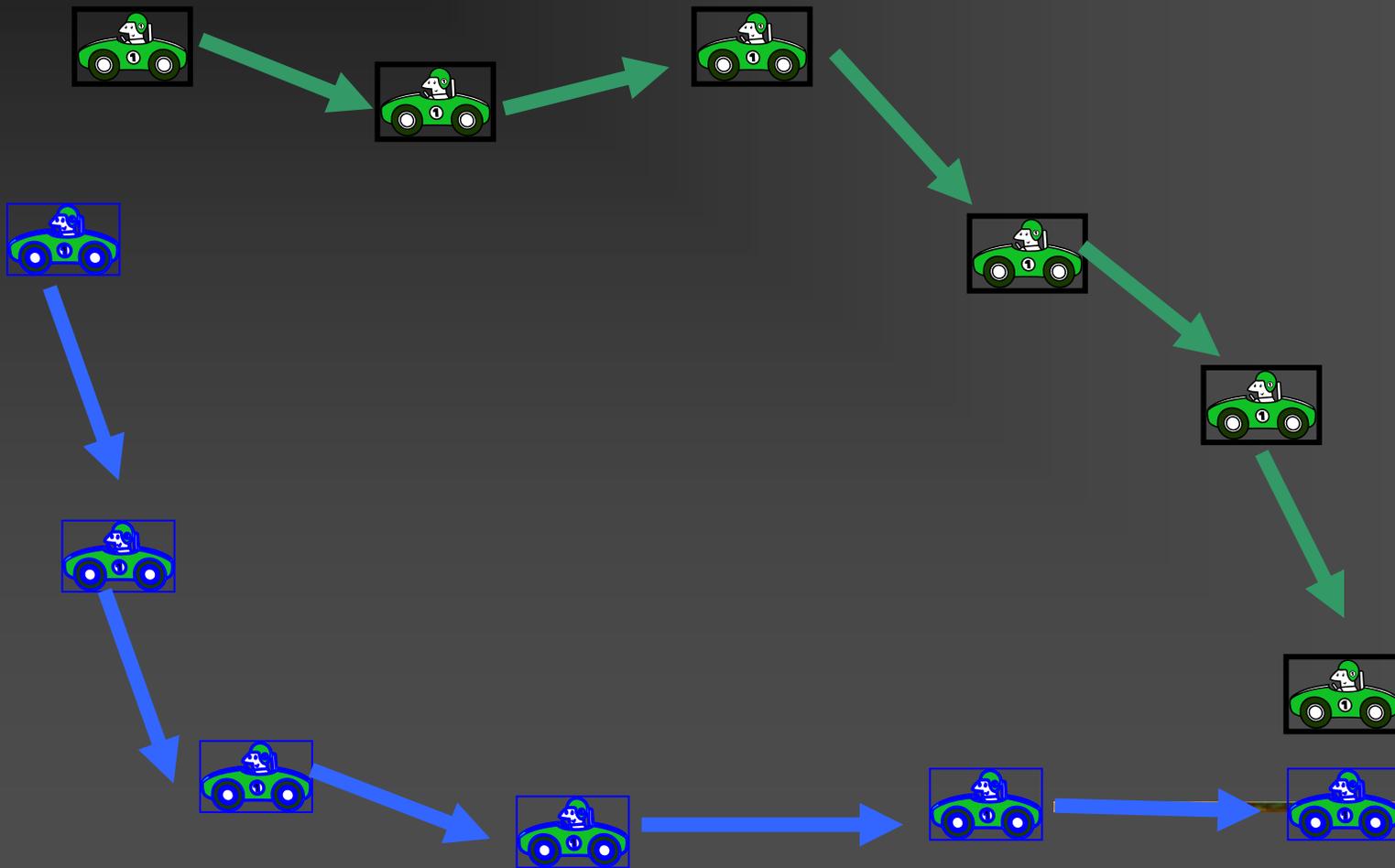
- The dissimilarity matrix is prone to attacks when there is domain knowledge available about the attributes
-

Spatio-temporal Data

- PDAs, mobile phones
- Location based services
- Vehicle tracking
- Various Sensors



Saptio-temporal Trajectories



ST Trajectories

- Still an ongoing research topic
 - What is a trajectory?
 - How do we extract trajectories
 - How do we represent trajectories?
 - How do we store/index trajectories?
 - How do we query trajectories?
 - How do we mine in trajectories?
 - What are the privacy/security issues with trajectories?
 - These questions will be addressed under GeoPKDD
-

Why do we collect ST-Trajectories

- **Discovering**
 - Frequent trajectories
 - Clusters of trajectories
 - **For**
 - Traffic management/planning
 - **BUT** we do not want others to know our location information
 - **Because**
 - We do not want to receive ads,
 - Safety reasons
 - Some people just want to be left alone
 - **So, privacy is a main concern**
-

Privacy vs Utility

- Balance between privacy and utility



Why do we collect ST-Trajectories

- Balance between privacy and utility



Similarity of ST Trajectories

- Euclidean is the simplest one
 - Assumes that the trajectories are the same length and they are sampled at the same time points.
 - Returns the sum of real penalties in terms of the distance of the corresponding points in space.
 - Edit Distance with Real Penalty
 - Cost of replacement is the distance between the replaced points
 - Cost of insertion/deletion is done with reference to a common object.
 - Dynamic Time Warping: Similar to edit distance, uses stutters (instead of deletion or insertion, repeats an element in the trajectory)
-

Similarity of ST Trajectories

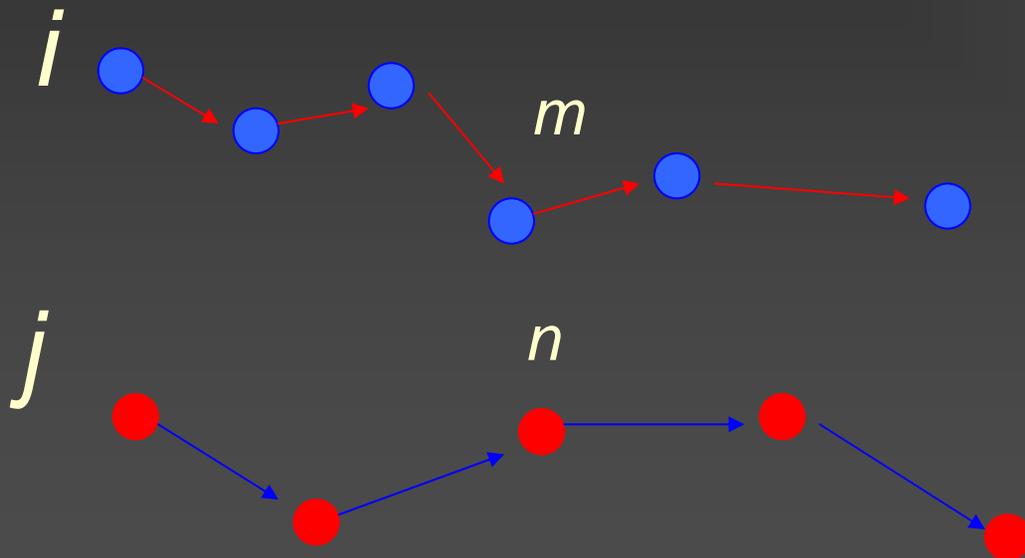
- Longest Common Subsequence
 - Longest common subsequence is the distance between two ST trajectories
 - Edit Distance on Real Sequence
 - Number of insertion/deletion operations to convert one trajectory to the other is the distance
 - Both Edit distance and LCSS require pair-wise comparison of all the points in trajectories
 - Two points match if they are close in space by less than a threshold value
-

Trajectory Matrix

- Denoted by D , and represents trajectories
- $D[i][j]$ is the j^{th} observation of trajectory i .
- Rows are scattered among data holders in case of horizontally partitioned data
- $m \times n$ matrix: m trajectories on maximum n observations

Dissimilarity Matrix

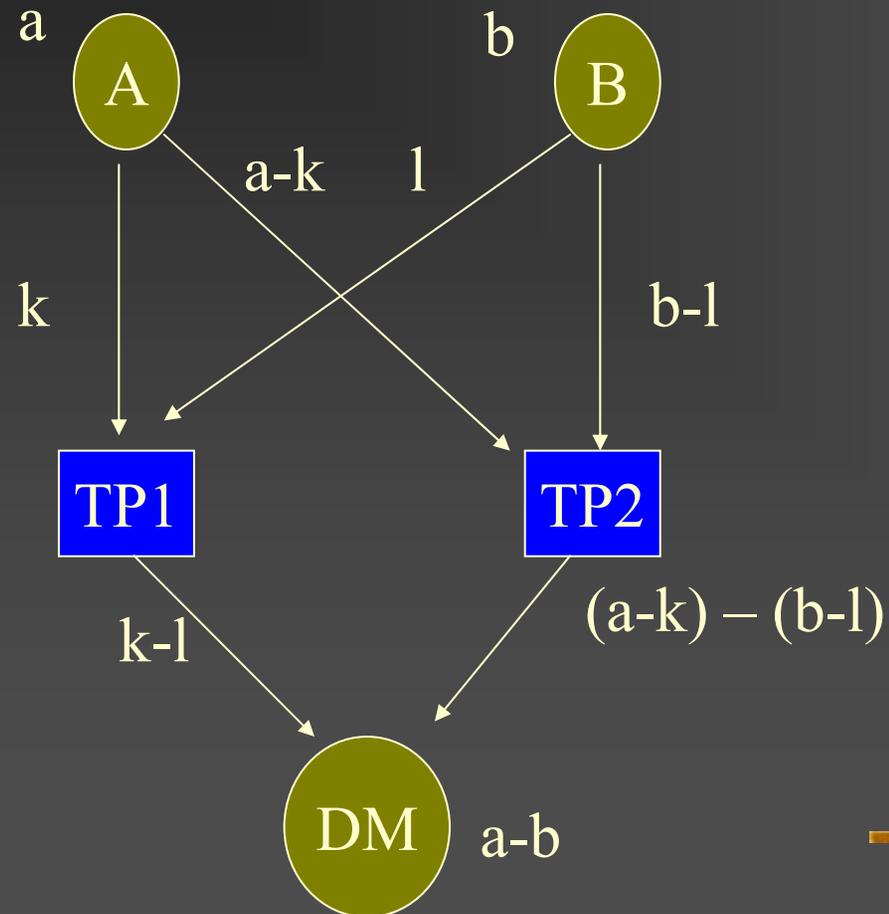
- Object-by-object structure
- $D[i][j][m][n]$ is the distance between the m^{th} observation of trajectory i and n^{th} observation in trajectory j



Homomorphic Secret Sharing

- Secret sharing among m players so that at least t players need to collaborate to compute the secret
 - A secret s will be distributed to m players as s_1, s_2, \dots, s_m
-

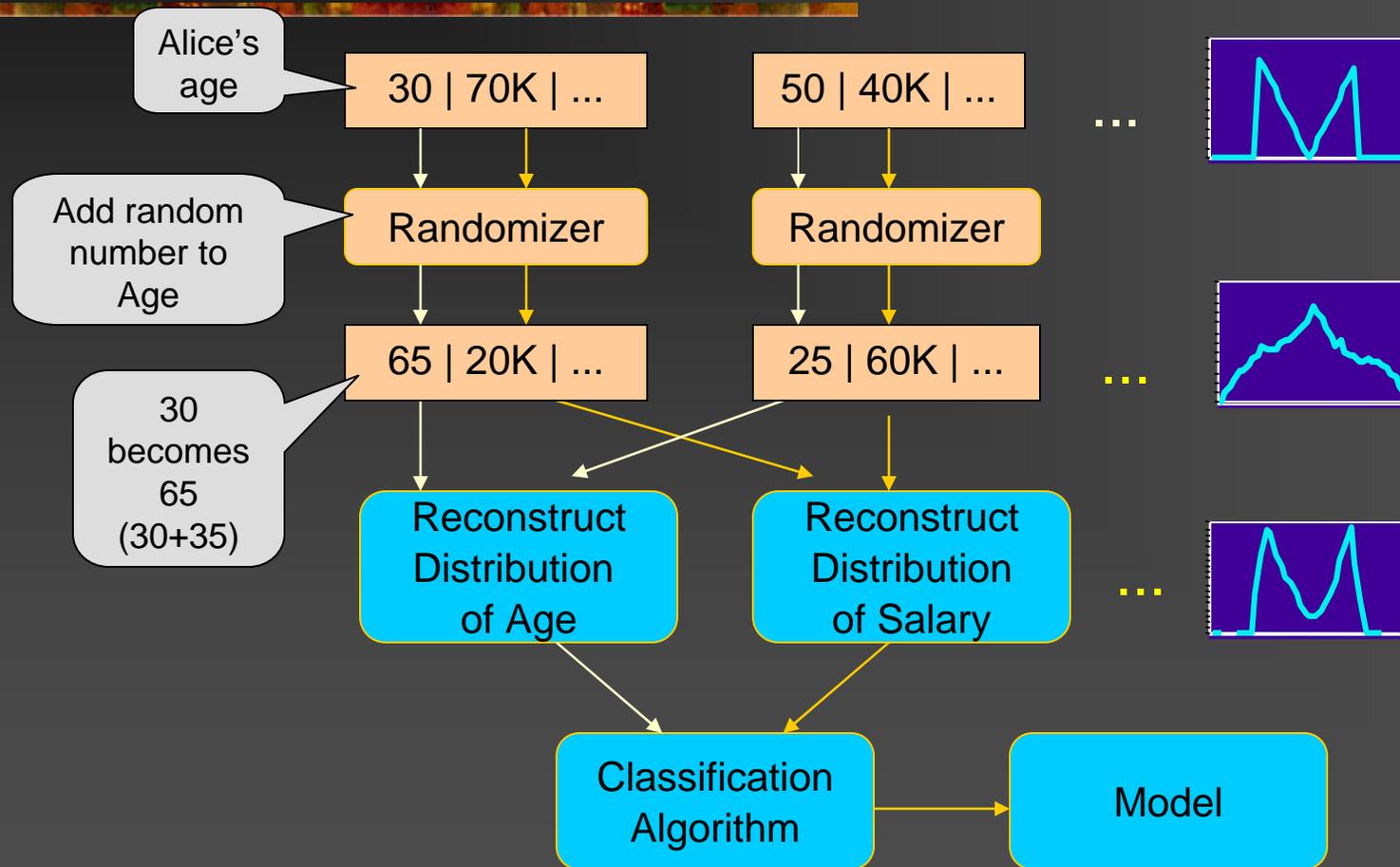
Homomorphic Secret Sharing (Less trust on TPs)



Classification on perturbed data

- Reference: Rakesh Agrawal and Ramakrishnan Srikant. “Privacy-Preserving Data Mining”. SIGMOD, 2000, Dallas, TX.
 - They developed a technique for constructing a classification model on perturbed data.
 - The data is assumed to be stored in a centralized database
 - And it is outsourced to a third party for mining, therefore the confidential values need to be handled
 - The following slides are based on the slides by the authors of the paper above
-

Randomization Approach Overview



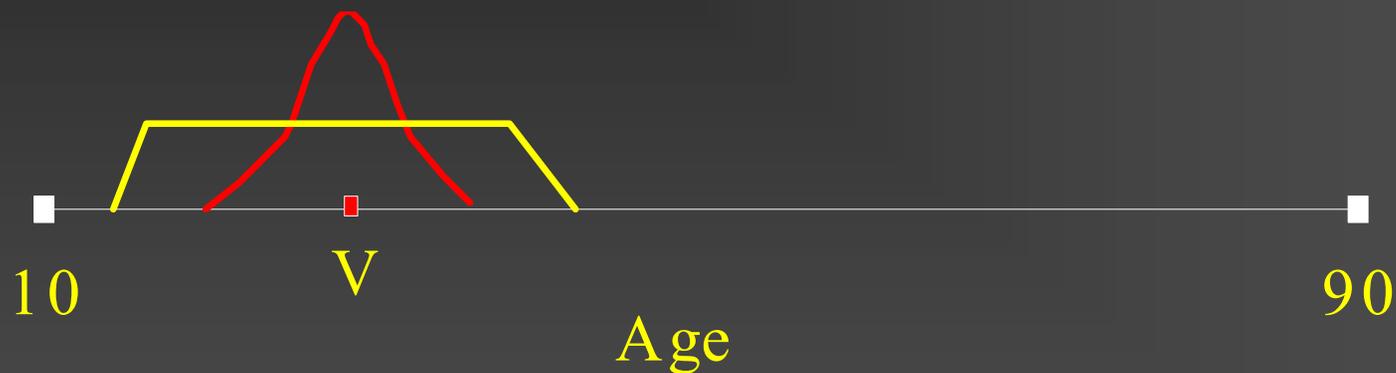
Reconstruction Problem

- 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

Intuition (Reconstruct single point)

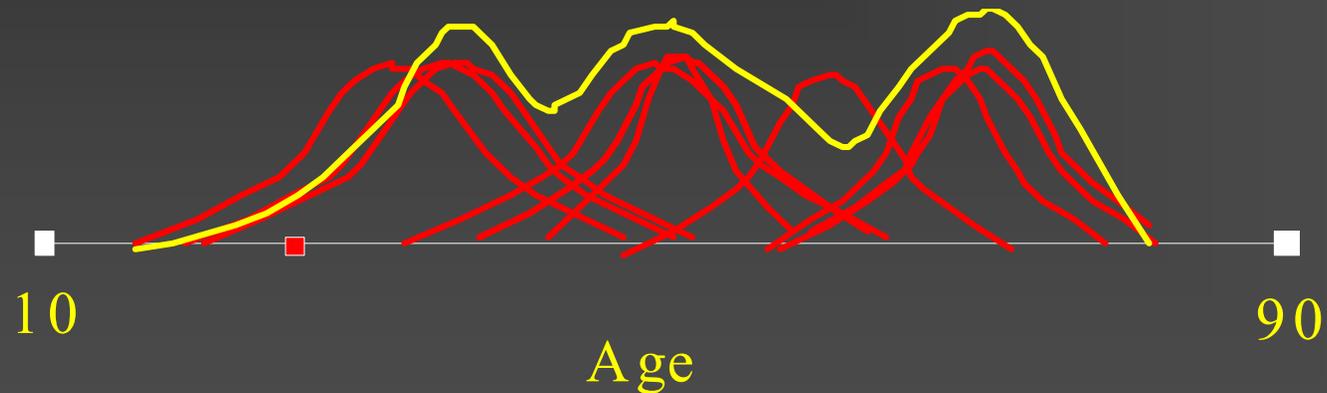
- Use Bayes' rule for density functions



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

Reconstructing the Distribution

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



Reconstruction: Bootstrapping

$f_X^0 :=$ Uniform distribution

$j := 0$ // Iteration number

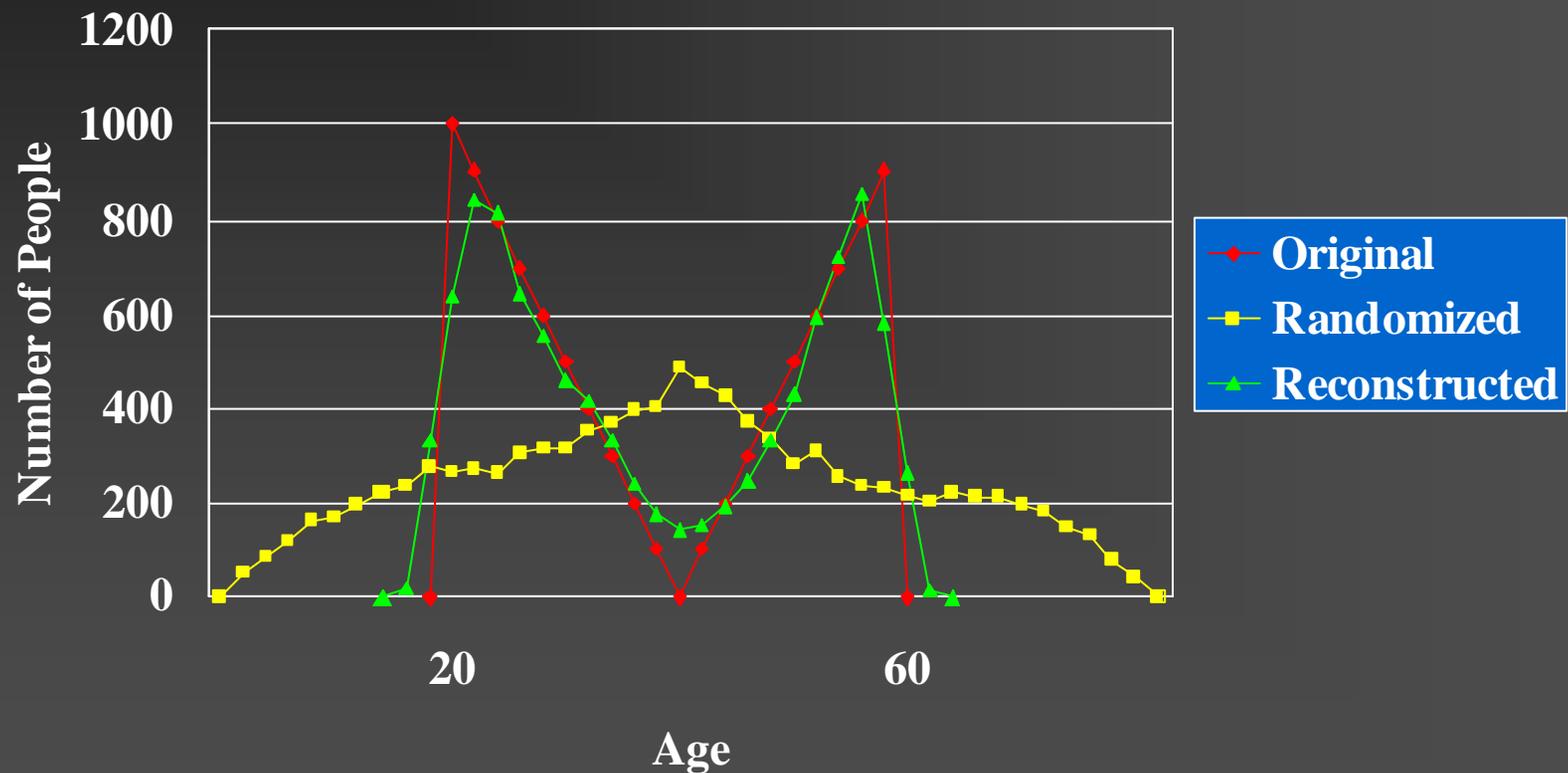
repeat

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

$j := j+1$

until (stopping criterion met)

Shown to work in experiments on large data sets.



Algorithms

- “Global” Algorithm
 - Reconstruct for each attribute once at the beginning
 - “By Class” Algorithm
 - For each attribute, first split by class, then reconstruct separately for each class.
 - See SIGMOD 2000 paper for details.
-

Experimental Methodology

- Compare accuracy against
 - **Original**: unperturbed data without randomization.
 - **Randomized**: perturbed data but without making any corrections for randomization.
 - Test data not randomized.
 - Synthetic data benchmark.
 - Training set of 100,000 records, split equally between the two classes.
-

Quantifying Privacy

- 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 “ amount of privacy.”
 - Example: (Interval Width : 54) / (Range of Age: 100) \approx 54% randomization level @ 90% confidence
-

Privacy Metric

If, from perturbed data, the original value x can be estimated to lie between $[x_1, x_2]$ with $c\%$ confidence, then the privacy at $c\%$ confidence level is related to $x_2 - x_1$

	Confidence		
	50%	95%	99.9%
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	$1.34 \times \sigma$	$3.92 \times \sigma$	$6.8 \times \sigma$

Example

- Salary 20K - 150K
- 95% Confidence
- 50% Privacy in Uniform
- $2\alpha = 0.5 * 130K / 0.95$
= 68K

• Issues

- For very high privacy, discretization will lead to a poor model
- Gaussian provides more privacy at higher confidence levels