Privacy Preserving Data Mining

Danushka Bollegala

COMP 527

UNIVERSITY OF LIVERPOOL

Privacy Issues

- Data mining attempts to find (mine) interesting patterns from large datasets
- However, some of those patterns might reveal information that users would not like to be disclosed
  - diseases/medical records of patients
  - credit worthiness
  - past conviction records
  - exam marks?
- It is like that an organization X and Y possess mutually useful data, and would like to use each other's data for data mining, but do not want to share the actual data.
- Can data mining and privacy co-exist?
  - Privacy Preserving Data Mining (PPDM)

Privacy Preserving Data Mining (PPDM)

- PPDM is a sub-field in DM that studies methods that can be used to perform various data mining tasks (e.g. decision tree learning, k-means clustering etc.) at the same time preserving the privacy of the users.
- Two main approaches exist
  - Anonymization (perturb with noise, abstract)
    - Add some noise to the data points so that it is not possible to uniquely determine a user
    - We would like to add the least amount of noise such that it is easier to perform data mining tasks on the anonymized data.
  - Encryption (perform DM tasks on encrypted data)
    - Each party that possess some data encrypts their data using their private keys.
    - We will perform DM operations directly on the encrypted data
    - Secure but is time consuming (encryption/decryption)

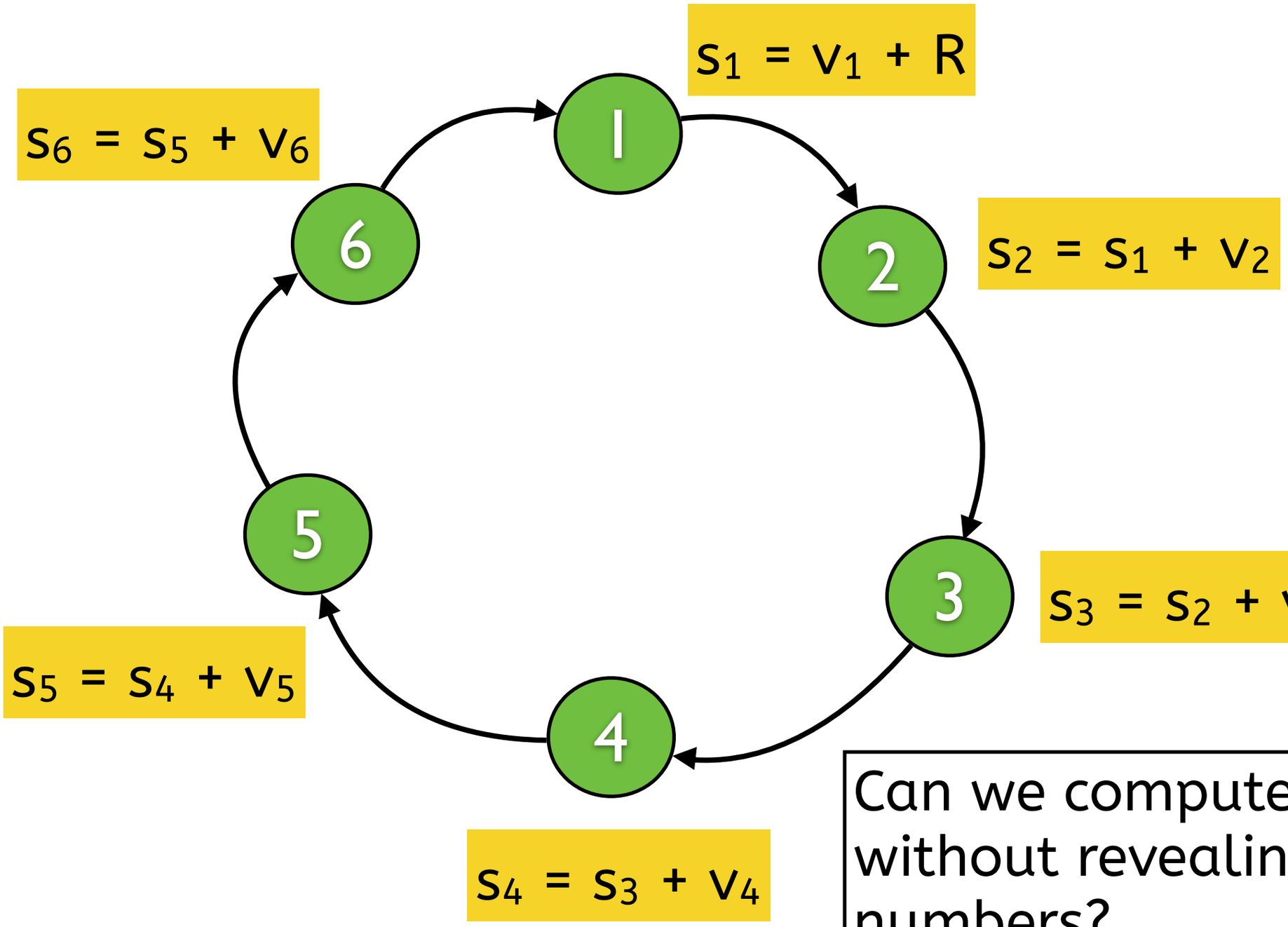
Dinning Cryptographers

- A group of cryptographers go out for dinner. After the dinner the waiter brings the bill and says it has already been paid for. However, the cryptographers are uneasy about this and would like to know whether any one of them paid the bill or was it an outsider. The cryptographer who paid the bill (if some one from this group did so) would like to keep this fact a secret.
- Conditions
  - waiter is honest and cannot be bribed.
- How can we find out whether some one in this group paid the bill or was it an outsider (NSA paid the bill!)?

Two Millionaires Problem

- Two millionaires want to know who has more money. But they do not want to disclose their wealth.
- cf. Yao’s millionaires’ problem

Secure Distributed Sum



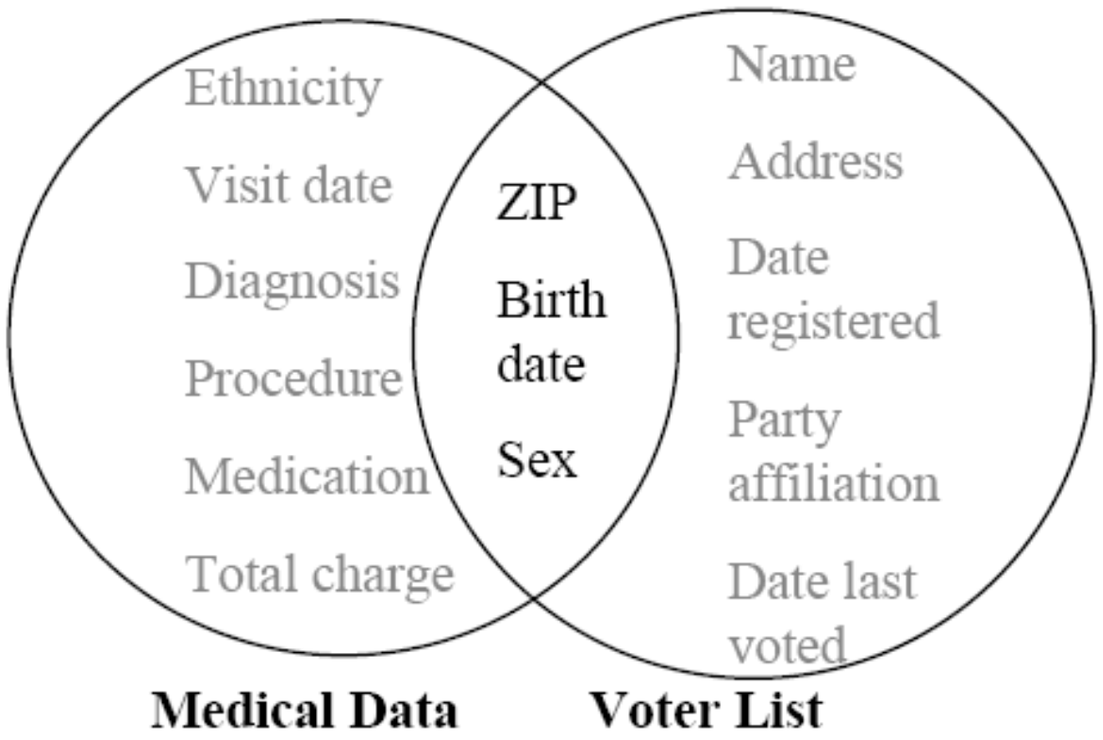
Can we compute the sum without revealing the individual numbers?

$v_1 + v_2 + v_3 + v_4 + v_5 + v_6$

Link Attack

- Although we might have anonymized two databases, by combining (linking) the two we might be able to identify the users.
- Can identify the medical record of the MA mayor by linking the voting database and medical record database.
- 87% of the people can be identified by combining zipcode, sex, and date of birth according to 1990 US census.

6 people have the date
of birth same as the mayor's.
3 males
1 matches the zip code!



Anonymization

- Explicit identifiers
  - Can uniquely identify a person.
  - Needs to be deleted or anonymized
- Quasi identifiers (QI)
  - By combining with external resources can be used to identify a person

name	date of birth	sex	zip	disease
Mark Taylor	21/1/70	M	53715	influenza
Ann Silvia	10/1/81	F	55410	AIDS
Lindsay Smith	1/10/44	F	90210	tooth ache
Michael Jordon	21/2/84	M	10285	bronchitis
Steve Jobs	19/4/72	M	11567	cancer

k-anonymity

- Proposed by Sweeney and Samarati [2001, 2002]
- By modifying the quasi identifiers anonymize an individual with (k-1) others in the database.
- In a k-anonymized database, we will have at least k different individuals with the same combination of values for the quasi identifiers.
- The probability of uniquely identifying an individual via a link attack reduces to 1/k.
- Techniques for implementing k-anonymity
  - Generalization
  - Suppression

Example

original data

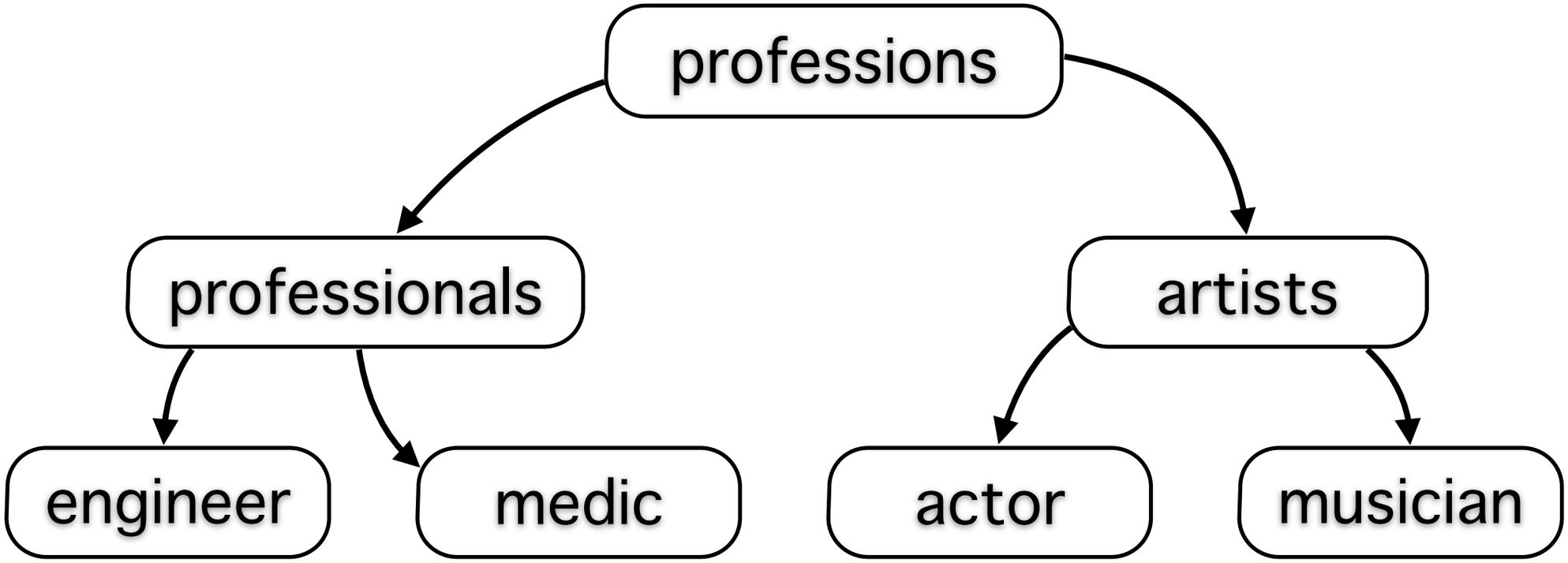
date of birth	sex	zipcode
21/1/79	Male	53715
10/1/79	Female	55410
1/10/44	Female	90210
21/2/83	Male	2274
19/4/82	Male	2237

k-anonymized data

	date of	sex	zipcode
group 1	*/1/79	Human	5****
	*/1/79	Human	5****
suppress	1/10/44	Female	90210
group 2	*/*/8*	Male	22**
	//8*	Male	22**

Generalization/Abstraction

- Given a hierarchy (ontology) of concepts/attributes, various generalization methods exist.
- We must select the generalization method with the least compromise with the anonymization.



Minimal distortion metric (MD)

- The number of data points (entries/records/rows) lost due to anonymization.
  - e.g. If we anonymize 5 males/females as human then MD = 5.
- Anonymization has a trade-off between the distortion and the level of privacy achieved.

$$IGPL(s) = \frac{IG(s)}{PL(s)+1}$$

Information Gain (IG)

Privacy Loss (PL)

Issues of k-anonymity

- After anonymization we can still make some inferences about a particular individual because there is no noise in the data.
- When the dimensionality of the data increases, the probability of uniquely identifying an individual increases for a fixed k value.

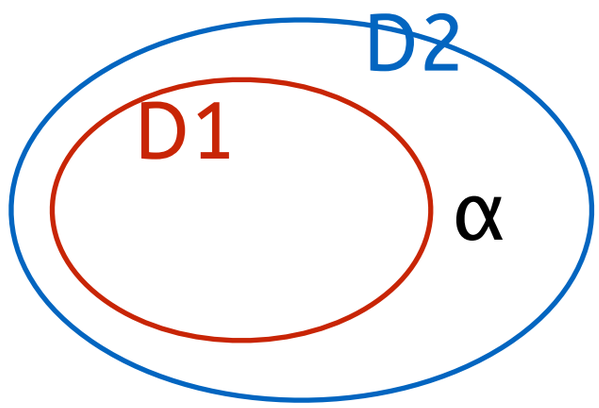
Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

If John is from Keral and is 19 years olf then we know that he has cancer, heart diseases, or viral infection.

slide credit: Wikipedia

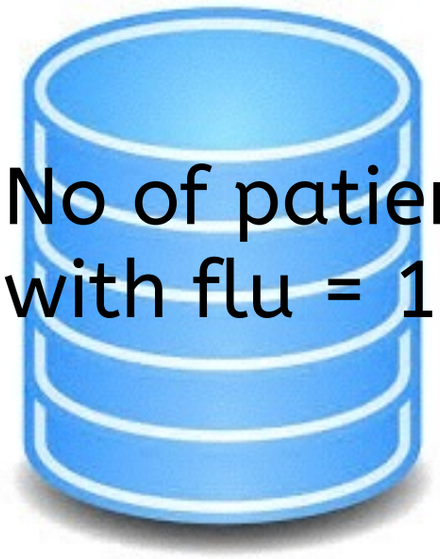
Differential Privacy

- Let us assume that we have two databases D1 and D2 that are differing in only one record.
- External users must not be able to identify this record by issuing any queries to D1 and D2.
- We must answer to the queries issues to D1 and D2 such that the answer contains sufficient noise to avoid revealing the differing record.

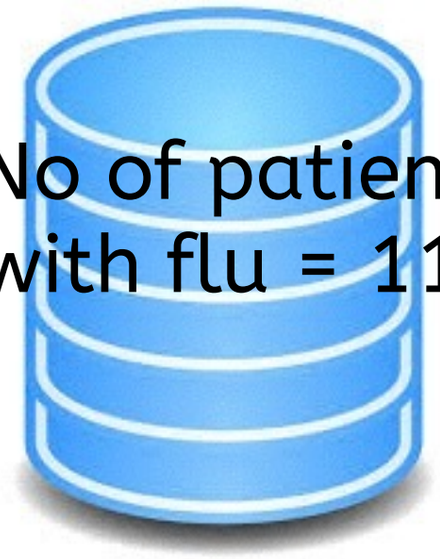


Example

Database of a hospital H before and after John has been admitted



No of patients with flu = 10



No of patients with flu = 11

f(getFluPatients) = 10 + 1

f(getFluPatients) = 11 - 2

We answer the query (asked via a function f) using some noise shown in red so that an attacker will not be able to find out that John has flu.

Differential Privacy

$$t = f(X) + Y \qquad p(t - f(D1)) \leq e^{\epsilon} p(t - f(D2))$$

$$\log \frac{p(t - f(D1))}{p(t - f(D2))} \leq \epsilon \qquad \text{Laplace}(\lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|t|}{\lambda}\right)$$

$$\Rightarrow \frac{p(t - f(D1))}{p(t - f(D2))} = \frac{\exp\left(-\left|t - f(D1)\right| / \lambda\right)}{\exp\left(-\left|t - f(D2)\right| / \lambda\right)} = \exp\left(\frac{\left|t - f(D1)\right| - \left|t - f(D2)\right|}{\lambda}\right)$$

$$\leq \exp\left(\frac{\left|f(D1) - f(D2)\right|}{\lambda}\right) = \exp\left(\varepsilon \left|f(D1) - f(D2)\right|\right) = \exp\left(\varepsilon \Delta p\right)$$

$$\Delta f = \max_{D1, D2} \left\| f\left(D1\right) - f\left(D2\right) \right\|_1 \qquad \varepsilon = \lambda^{-1}$$

$$\Rightarrow p\left(Y\right) = p\left(t - f\left(X\right)\right) = Laplace\left(\frac{\Delta f}{\varepsilon}\right) \quad \Rightarrow \quad Y = t - f\left(X\right) \sim Laplace\left(\frac{\Delta f}{\varepsilon}\right)$$

Privacy using Encryption

- A and B would like to perform data mining using both their databases. But they do not want to share their raw data.
- Solution
  - Encrypt the data using their public keys and perform statistical operation on the encrypted data.
- An important property of the **homomorphic encryption**
  - If we denote the encryption of a message x using a public key pk as Epk(x), then the following holds
    - Epk(x + y) = Epk(x) x Epk(y)
  - In RSA for example $Epk(x) = x^e \bmod m$, where m is the public key and e is an exponent.

Applications

- k-means clustering over a distributed database.
- Each party has a subset of (non-overlapping) attributes. We would like to cluster the data using k-means but do not want the parties to share their attributes.
  - vertical partitioning of a database
- Clustering data points based on each partition might lead to incorrect results.

Vertical partitioning of a database

age	sex	height	weight	profession	location

Issues of cryptography-based PPDM

- Slow in practice
  - Multiple encryption-decryption steps required, which can be slow for large databases.
    - k-means for a database with 1000 records taking as much as one hour!
- Tend to be complicated algorithms
  - see Vaidya+Clifton KDD'03 for the details of the vertical partitioned k-means algorithm