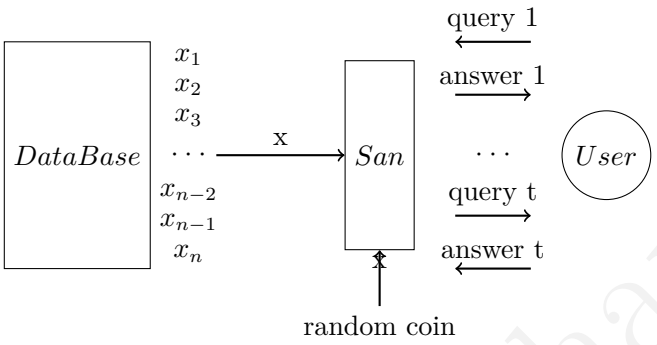
CS3235 Part3 Lec2: Privacy

$_/|/|U_Ch@NgRu!$

December 16, 2021

1 Basic Setting



1.1 San:Sanitization Method

1. input pertubation
   - randomize data before computation/sharing
   - add random noice to database
2. Summary statistic
   - mean
   - variance
   - marginal total
     RTBD
   - regression coefficient
3. output pertubation: summary statistic with noise
4. In this game User decides the query, and the goal is to discover some sensitive information

Mu Changrui

1.1.1 Data Sanitization

Data sanitization involves purposely, permanently deleting, or destroying data from a storage device, to ensure it cannot be recovered.

1.1.2 Data pertubation

Data perturbation is a data security technique that adds 'noise' to databases allowing individual record confidentiality. This technique allows users to ascertain key summary information about the data that is not distorted and does not lead to a security breach.

2 Data anonymisation

Goal

- make a dataset
- the dataset should make the data privacy preserved before sharing
- the dataset should maintain the utility and information of the data as much as possible

2.1 Goal of privacy-preservation

- Membership disclosure: attackers can not say certain person is in the dataset
- Attribute disclosure: attacker cannot tell whether certain person contains certain attributes
- Identify disclosure, attacker cannot tell which records corresponds to a certain given individual

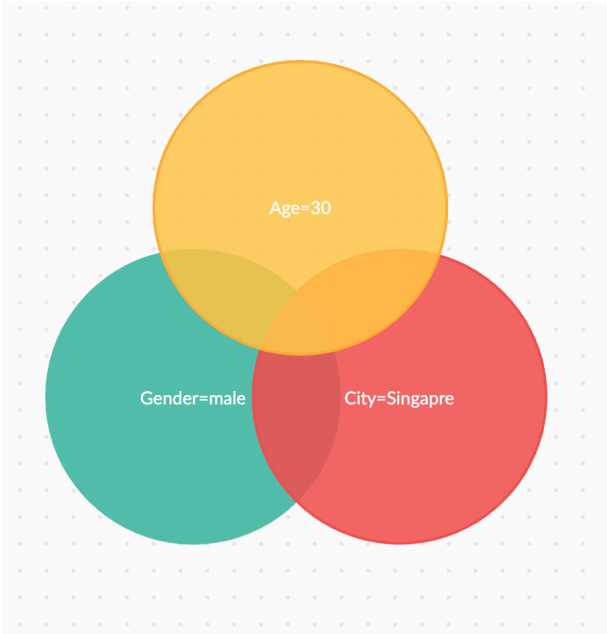
2.2 Personally identifying information(PII)

Any representation of information that permits identity of individuals to whom is implied to be reasonably with direct or indirect means(name, uid etc.)
It is ideal if the PII can be removed.
However in reality, PII has no technical meanning, and in privacy breath, any information can be PII(e.g. If in the class, I am the only male, the sex will be my PII; also in some large database such as ALO dataset and Netflix dataset, there are tremendous number of attributes, with all these attribute's intersection, specific indivisual can be implied)

Mu Changrui

2.3 Latanya Sweeney's attack

Attacker learns sensitive information by joining two datasets on commmon attributes. In this case, the privacy information can be recovered even if the PII is removed.
Here it's possible that one table A contains some personal infroamtion such as HIV test report but all PII in A such as name and ID are removed. There is a table B which contains patients' name, sex, and birthday(perhaps for sending "happy birthday" messages), attackers can join table A and B with common attributes age, sex, birthday etc. to recover the records
It was proved that birthday, ZIP code, gender, uniquely identifies 87% of US population



Mu Xiangrui

3 k-anonymity

A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least k - 1 individuals whose information also appear in the release.
It's generally a NP-hard problem

3.1 Proof on the complex of k-anonymity

RTBD

3.2 Practice of k-anonymity

Even if k-anonymity is NP-hard, there are practically efficient k-anonymization algorithm
Most of the algorithms are based on generalization and suppression

Generalization

Generalization consists in replacing the actual value of the attribute with a less specific, more general value that is faithful to the original; In k-anonymity, each generalized value should have at least k distinct individuals

Example of generalization

ID: 5555555555554444 $\rightarrow$ 555 * * * * * * * * * * * **

Supression

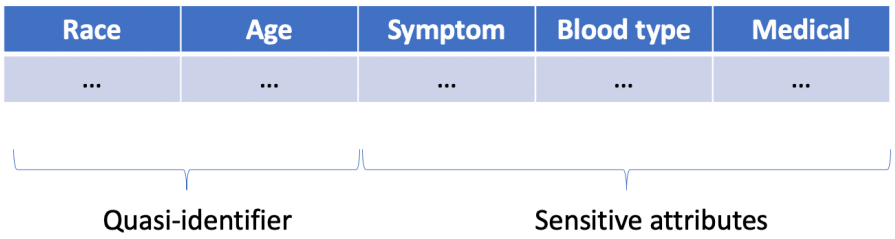
Suppression refers to not releasing a value at all.

3.3 quasi-identifiers and sensitive information

In practice, it seems reasonable to classify the attributes based on the their effects for identifying individuals

- PII: attribute that can identify individuals by itself
- quasi-identifier: identifiers that cannot identify individuals uniquely in most case, but it can still be applied to identify individuals by combining with other quasi-identifier(In short, it cannot identify individual by itself, but has high chance to identify when combining with other quasi-identifier)
- Sensitive attributes: whose information is not willing to be disclosed or linked to individual identity

In practice, PII will be just suppressed, or generalized into quasi-identifier until at least k partition ordered-value domain in interval



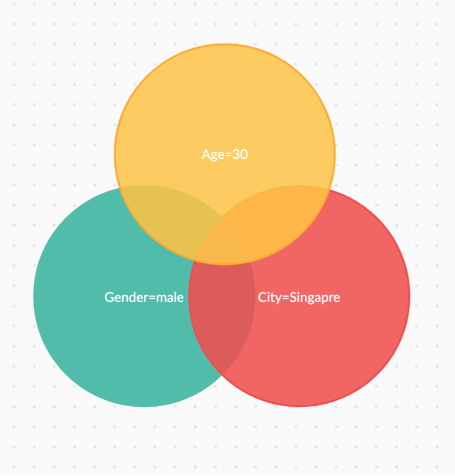
Mu Changrui

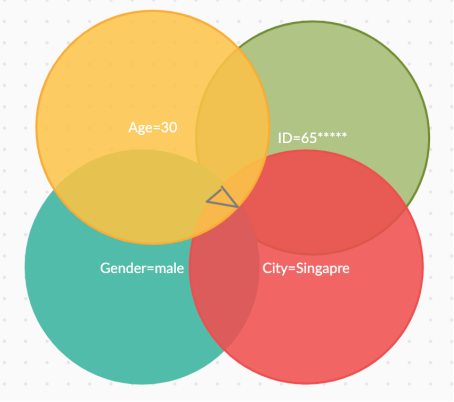
Let's evaluate the privacy goals that the k-anonymity can achieve

- Membership disclosure: with a given person, where we know her/his partial identity information, the k-anonymity cannot achieve membership disclosure, for example, if we know a person's ID is 556023, if there are k records with quasi-ID 55***3, it's likely that this person is a member of the dataset
- Attribute disclosure: k-anonymity cannot achieve attribute; Attribute disclosure can be compromised when membership disclosure is compromise. For example if there is a HIV dataset, with all people in records are HIV patients, after one individual's membership disclosure is compromised, the patient certainly is a HIV patient[1], this also proves that Attribute disclosure $\rightarrow$ Membership disclosure
- Identity disclosure: k-anonymity can achieve identity disclosure because a record in k-anonymous dataset cannot be mapped back to the original dataset[2]; E.g. with quasi-identifier 55***3, it's impossible to know whether it's the record of 556023 or 556133

3.4 Curse of dimentionality

Generalization fundamentally relies on spatial locality, e.g. generalized 65**** needs at least k individuals with ID start form 65
This may not be meet in reality, especially for some large database, such as Netflix, Amazon dataset, the customers' information is actually very parse, and it's hard to find k close neighbor in one dimension;
Also as the dimension increase, the individual indentity is harder to hide It's easy to see that the





posibility of certain individual can be narrowed down as the number of quasi-identifier attributes increase, especially when these attributes have less dependency(in fact there is no clear distinction

between quasi-identity and sensitive information)
In the example of how attribute disclosure can be compromised in k-anonymity, it's easy to see, if all records under the same quasi-identifier has certain/known sensitive characteristics in common, the privacy(membership disclosure and attribute disclosure) will be compromised. Therefore we should search for diversity of sensitive properties within the same quasi-identifier group

4 l diversity/ℓ diversity

Definition: entropy of sensitive attributes within each quasi-identifier group must be at least L
The aim of l-diverse is to enhance the attribute disclosure in k-anonumity

4.1 Drawback of l diversity

l diversity may still help attacker guess attacker "guess" victim's private information
For example, in the whole released data, the rate of cancer is 99%, however, a certain person's membership disclosure is compromised(say he is in the 65****** quasi-identifier group), and the rate of cancer in the 65****** quasi-identifier group is just 40%, attacker can hence guess that this person has no cancer;
Statistics:

- E: the person has cancer
- F: the person in 65****** quasi-identifier group
- $P(E) = 0.99$
- $P(E^c) = 0.99$
- $P(E|F) = 0.4$
- $P(E^c|F) = 0.6$

Hence the membership of the quasi-identifier group can help adversary compromise the attribute disclosure

5 t closeness

Definition:
The distribution of sensitive information in each quasi-identifier group should be close to the distribution in the whole database
Aim: enhance the attribute disclosure in l diverse
This model is ideal, but hard to achieve, because

Mu Changrui

- For large database, there may be many quasi-identifiers, which will make it extremely hard to achive t closeness in any combination of quasi-identifiers
- Also, if we try achive it by adding some noice data, it will take to much extra space
- In reality, there is no clear distinction between sensitive data and quasi-identifier, and hence any attributes is a potential quasi-identifier

6 "Perfect Privacy"

A classic intuition from cryptography may lead us to a definition of "perfect privacy", which is

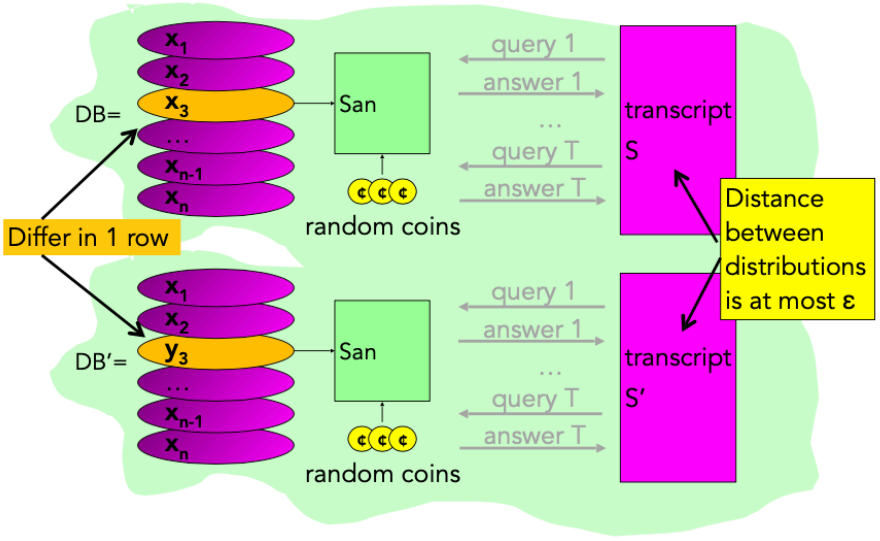
- Pr[Privacy info] = Pr[Privacy info| User has accrss to S]

But this is unachivable, for the following reasons

- The privacy information means any information that may helps indentify the individuals; Once any information related to the individual is released in clear, it can be privacy information. It we let Privacy info and information got from S be independent, the release of the dataset is meaningless
  e.g. if the adversary has prior knowledge that very one in NUS is older than 200. Access to S will definite helps him/her to identify any NUS members
- It's possible that some non-private data helps adversary recover privacy data of victim, i.e. prior information+ non-privacy data $\rightarrow$ privacy information
  e.g. One adversary knows that Bob is 8cm higher than NUS men's average height, if the adversary has access to a anonymous database, which can help him compute the average height of NUS men, then the adversary will able to compromise the privacy of Bob

MuQiangGui

7 Indistinguishability

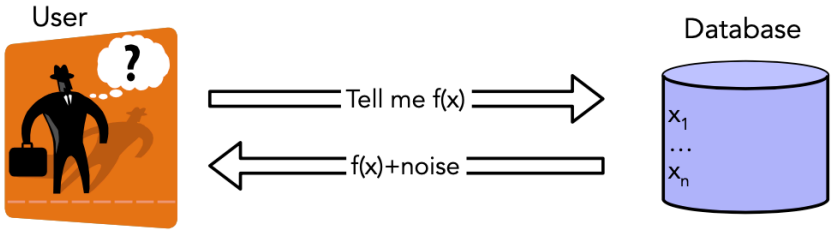


Definition: San is ε-indistinguishable if :

$\forall \mathcal{A}, \forall DB, DB'$, where DB,DB' only differ in 1 row, $\forall$ sets of transcripts S

$\frac{Pr[San(DB)=S]}{Pr[San(DB')=S]} = 1 \pm \varepsilon$

8 Laplacian Mechanism



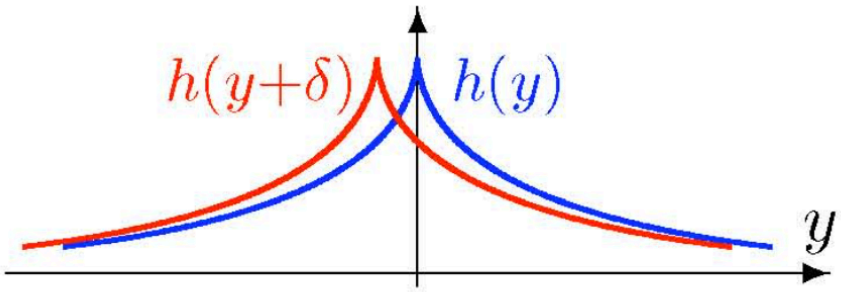
- f(x) is any query, f(x) can be released accurately when f is insensitive to individual entries x1, ... xn
- Global sensitivity: $GS_f = max_{neighborx,x'}|f(x) - f(x')|$

8.1 Laplacian theorem

If $A(x) = f(x) + Lap(\frac{GS_f}{\varepsilon})$, then A is $\varepsilon - indistinguishable$

8.2 Laplace distribution

$Lap(\lambda)$ has density function $h(y) = \frac{\lambda}{2}e^{-\lambda|x|}$



So San gives ε-differential privacy if for all values of DB and Me and all transcripts t:

$$\frac{Pr[San(DB-Me)=t]}{Pr[San(DB+Me)=t]} \leq e^{\varepsilon} \approx 1 \pm \varepsilon$$

Understanding:RTBD

Intuitively,

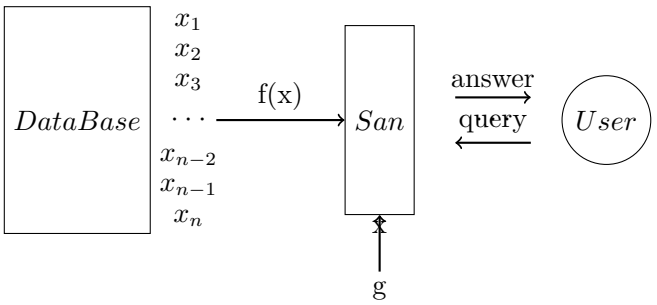
- No perceptible risk is incurred by joining DB
- Anything adversary can do to me, it could do without me (my data)

Mu Changrui

References

[1] Lukas Malina, Jan Hajny, Radek Fujdiak, Jiri Hosek, "On perspective of security and privacy-preserving solutions in the internet of things", 2016, Computer Networks, Volume 102, Pages 83-95, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2016.03.011.

[2] J. Domingo-Ferrer and V. Torra, "A Critique of k-Anonymity and Some of Its Enhancements," 2008 Third International Conference on Availability, Reliability and Security, 2008, pp. 990-993, doi: 10.1109/ARES.2008.97.



Mu Changrui