CERIAS
the center for education and research in information assurance and security

T-Closeness: Privacy Beyond *k*-Anonymity and *l*-Diversity

Ninghui Li, Tiancheng Li, Suresh Venkatasubramanian

K-Anonymity

- Definition: Each record is identical to at least k-1 other records.

DOB	Zipcode	Disease
01/10/76	47906	Heart Disease
01/21/76	47905	Hepatitis
02/08/76	47923	Brochitis
02/15/76	47927	Broken Arm
04/03/86	47935	Flu
04/02/86	47931	Hang Nail

DOB	Zipcode	Disease
01/**/76	4790*	Heart Disease
01/**/76	4790*	Hepatitis
02/**/76	4792*	Brochitis
02/**/76	4792*	Broken Arm
04/**/86	4793*	Flu
04/**/86	4793*	Hang Nail

From *k*-Anonymity to *l*-Diversity

- Attacks on *k*-Anonymity
  - **Homogeneity Attack**
    - *Sensitive values lack diversity*

Bob	
Zipcode	***Age***
47678	27

    - *Bob has heart disease!*
  - **Background Knowledge Attack**
    - *Carl doesn't have heart disease*

Carl	
Zipcode	***Age***
47673	36

    - *Carl has cancer!*

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

- *L*-Diversity
  - Each equi-class has at least *l* well-represented sensitive values
  - Instantiations
    - *Distinct l-diversity: Each equi-class has at least l distinct value*
    - *Entropy l-diversity: Entropy(equi-class) ≥ log(l)*
    - *Recursive (c,l)-diversity:* $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$

Limitations of *l*-Diversity (1)

- *L*-diversity may be difficult and unnecessary to achieve
  - A table with two sensitive values: HIV+(1%), HIV-(99%)
  - *l*-diversity is unnecessary to achieve
    - *An equi-class contains only negative records*
  - *l*-diversity is difficult to achieve
    - *There can be at most 1%*10000=100 equi-classes*

Limitations of *l*-Diversity (2)

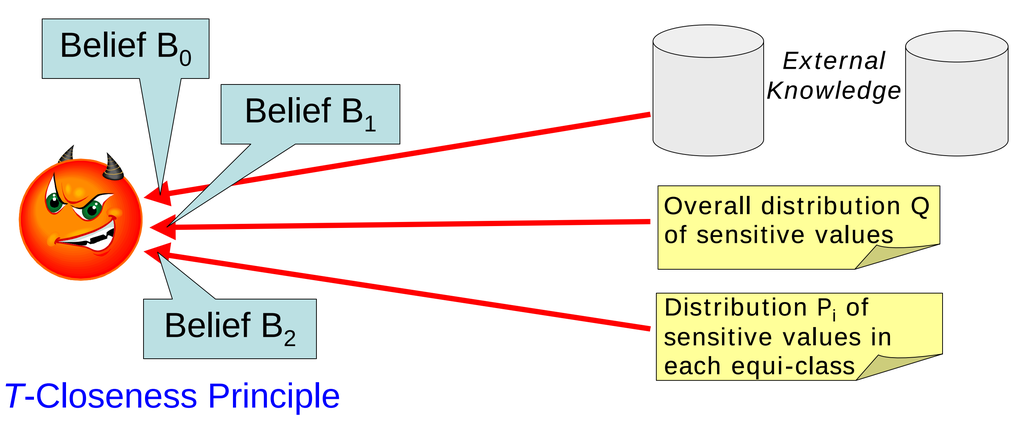
- *l*-diversity is insufficient to prevent attribute disclosure
  - **Skewness Attack**
    - *Exactly the same diversity but very different privacy risks*
      - *Equi-class 1: 49 positive records + 1 negative records*
      - *Equi-class 2: 1 positive records + 49 negative records*
    - *Observation: l-diversity does not consider the overall distribution*
  - **Similarity Attack**
    - *Sensitive values may be semantically close*

Bob	
Zip	***Age***
47678	27

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

    - *Conclusions*
      - *Bob's salary is in [20k,40k]*
      - *Bob has stomach-related disease*
    - *Observation: l-diversity does not consider semantic meanings of sensitive values*

T-Closeness



- *T*-Closeness Principle
  - *Q* should be public information
  - Bound the difference between B_1 and B_2, instead of B_0 and B_1
  - The distance between P_i and *Q* should not exceed *t*
- Earth Mover's Distance
  - Definition: The cost of transforming one distribution to another by moving probabilities
  - Capture the notion of semantic distance
  - Simple formulas for three distances:
    - ordered-distance
    - equal-distance
    - hierarchical-distance