Data Privacy

J- Component

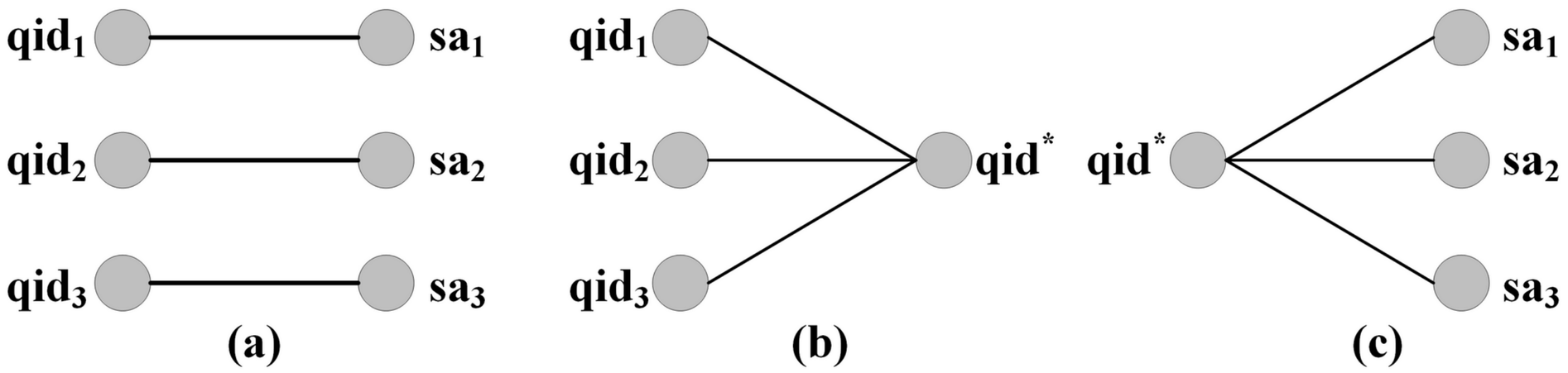
TEAM

Saurabh Singh - 19BCI0184
Maurya Goyal - 19BCI0191
Ayush Gupta - 19BCI0222

Topic

New multidimensional recoding model and a greedy algorithm for k-anonymization

What is k-anonymity?



Assuming the record is in this format: [QID, SA]. The basic idea of k-anonymity is safety in group , which means that we are safe if we are in a group of people whose QIDs are the same.

Nobody can infer our sensitive information (SA) from this group using QID, as shown in Fig. 1 (k=3 in 1(b) and 1(c)). If each of these group has at least k people, then this dataset satisfy k-anonymity.

- The raw datasets usually don't satisfy k-anonymity.
- To transform raw datasets into anonymised datasets, we require the assistance of an anonymization algorithm.
- Our base paper talks about that and is based on generalization.
- Generalization is a kind of transformation, which finds a result QID* that covers all QIDs (QID1~QID3)

Literature Review:

G. Ghinita, P. Karras, P. Kalnis, N. Mamoulis. Fast data anonymization with low information loss. Proceedings of the 33rd international conference on Very large data bases, VLDB Endowment, 2007, 758-769

K. LeFevre, D. J. DeWitt, R. Ramakrishnan. Multidimensional K-Anonymity ICDE '06: Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, , 2

Literature Review:

Y. He, J. F. Naughton, Anonymization of set-valued data via top-down, local generalization. Proceedings of VLDB, 2009, 2, 934-945

J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, A. W.-C. Fu. Utility-based anonymization using local recoding. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, 785-790

Idea from research papers:

Using multidimensional partitioning, a k-anonymization is generated in two steps. In the first step, multidimensional regions are defined that cover the domain space, and in the second step, recoding functions are constructed using summary statistics from each region.

Algorithm Proposed:

Anonymize($partition$)

if (no allowable multidimensional cut for $partition$)

return $\phi : partition \rightarrow summary$

else

$dim \leftarrow$ choose_dimension()

$fs \leftarrow$ frequency_set($partition$, dim)

$splitVal \leftarrow$ find_median(fs)

$lhs \leftarrow \{t \in partition : t.dim \leq splitVal\}$

$rhs \leftarrow \{t \in partition : t.dim > splitVal\}$

return Anonymize(rhs) $\cup$ Anonymize(lhs)

Algorithm with k-d tree explaination:

We'll take k=2 for example:

Quasi-identifiers

Zipcode

Age

Age	Sex	Zipcode	Disease
25	Male	53711	Flu
25	Female	53712	Hepatitis
26	Male	53711	Brochitis
27	Male	53710	Broken Arm
27	Female	53712	AIDS
28	Male	53711	Hang Nail

Patient Data

Motive:

1. **Partition the raw dataset into k-groups using kd-tree. k-groups means that each group contains at least k records.**
2. **Generalization each k-group such that each group has the same QID*.**

Workflow

Data

Partition

Generalization

Output

Why using k-d tree?

- It is fast, straight-forward and sufficient.
- KD-trees are a specific data structure for efficiently representing our data.
- KD-trees helps organize and partition the data points based on specific conditions.

Implementation

Thankyou