

Privacy-Preserving Publishing Data with Outliers

Wendy Hui Wang, Ruilin Liu
Department of Computer Science, Stevens Institute of Technology

Introduction

Identifiers		Quasi-Identifiers		Sensitive
Id	Name	Age	Gender	Salary
1	Alice	20	F	20K
2	Bob	20	M	25K
3	Justin	20	M	120K
4	Carol	30	F	30K
5	Allan	30	F	50K
6	Bill	30	M	2 Billion
7	Ben	40	M	100K
8	Susan	40	F	110K
9	David	40	M	130K

Quasi-Identifiers			Sensitive
Age	Gender	Salary	
20	*	20K	
20	*	25K	
20	*	120K	
30	*	30K	
30	*	50K	
30	*	2 Billion	
40	*	100K	
40	*	110K	
40	*	130K	

← Bad anonymization

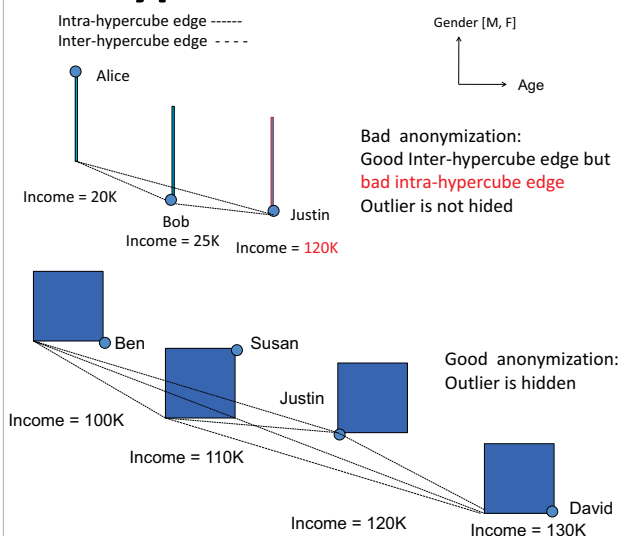
Quasi-Identifiers			Sensitive
Age	Gender	Salary	
[20, 40]	*	20K	
[20, 40]	*	50K	
[20, 40]	*	100K	
[20, 40]	*	110K	
[20, 40]	*	25K	
[20, 40]	*	30K	
[20, 40]	*	120K	
[20, 40]	*	130K	

→ Good anonymization

Challenge:

How to hide outliers in the published dataset with minimized information loss?

QI-Hypercube



Privacy model

Plain k-anonymity

In each anonymization group:

- At least k distinct sensitive values
- Attacker cannot infer the existence of any outlier from the anonymized data

QI-hypercube Group

- c -dimension hypercube (c : numbers of QIs)
- **Node** consists of boundaries of QIs
- **Edge** connects two nodes in a group
 - intra-hypercube edge (two nodes have at least one pair of QI-values that is the same in the same hypercube)
 - inter-hypercube edge (two nodes have all pairs of QI-values that are the same in different hypercubes)

Plain Group

- For all hypercubes in the group, at least one intra-hypercube edge is good.
- At least one inter-hypercube edge is good.

Algorithm

Step 1: Removal of global outliers

Remove the outliers that cannot be included into any plain, k -anonymous group.

Step 2: Expansion-based grouping

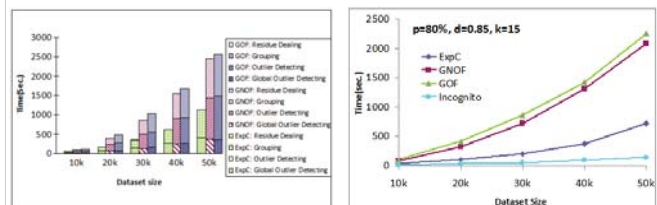
1. Pick k tuples to construct a group.
2. Examine the plainness of the group. If it does not satisfy plain k -anonymity, add more tuples to expand the hypercube, until it reaches the plain k -anonymity.

Step 3: Processing of residue tuples

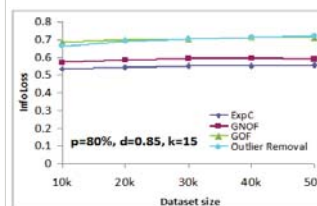
1. Pick the anonymization group that produces the minimal information loss by including residue tuples as the seed group.
2. Merge the seed groups until it reaches plain k -anonymity.

Experiment

Performance



Information Loss



*ExpC: our algorithm, GOF, GNDF: different group algorithms which need to locate outliers first