An Efficient k-anonymization Algorithm for Privacy Preserving Data Sharing

Md Nurul Huda

Trans-disciplinary Research Integration Center (TRIC)

Noboru Sonehara

National Institute of Informatics (NII), Tokyo

Date: 2012/02/16

k-anonymity

Tools for anonymization

- Generalization
  - Publish more general values
- Suppression
  - Remove tuples, i.e., do not publish outliers

Original Microdata

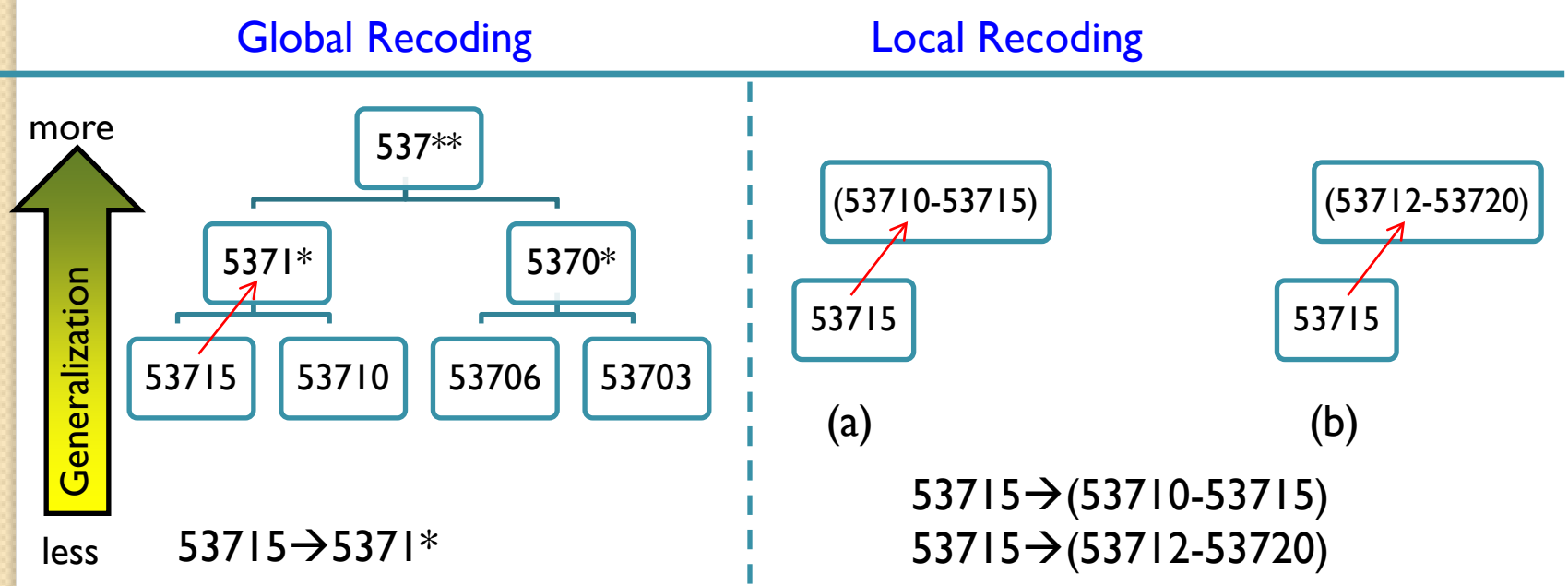
Age	Education	Marital S	Sex	Zip	S. Info
58	College	Divorced	Male	53715	Info_1
42	BA	Married	Male	55410	Info_2
34	College	Married	Male	90210	Info_3
51	HSG	Never	Female	02274	Info_4
46	College	Never	Female	02237	Info_5
		.....	.....	.....	.....

3-anonymous data

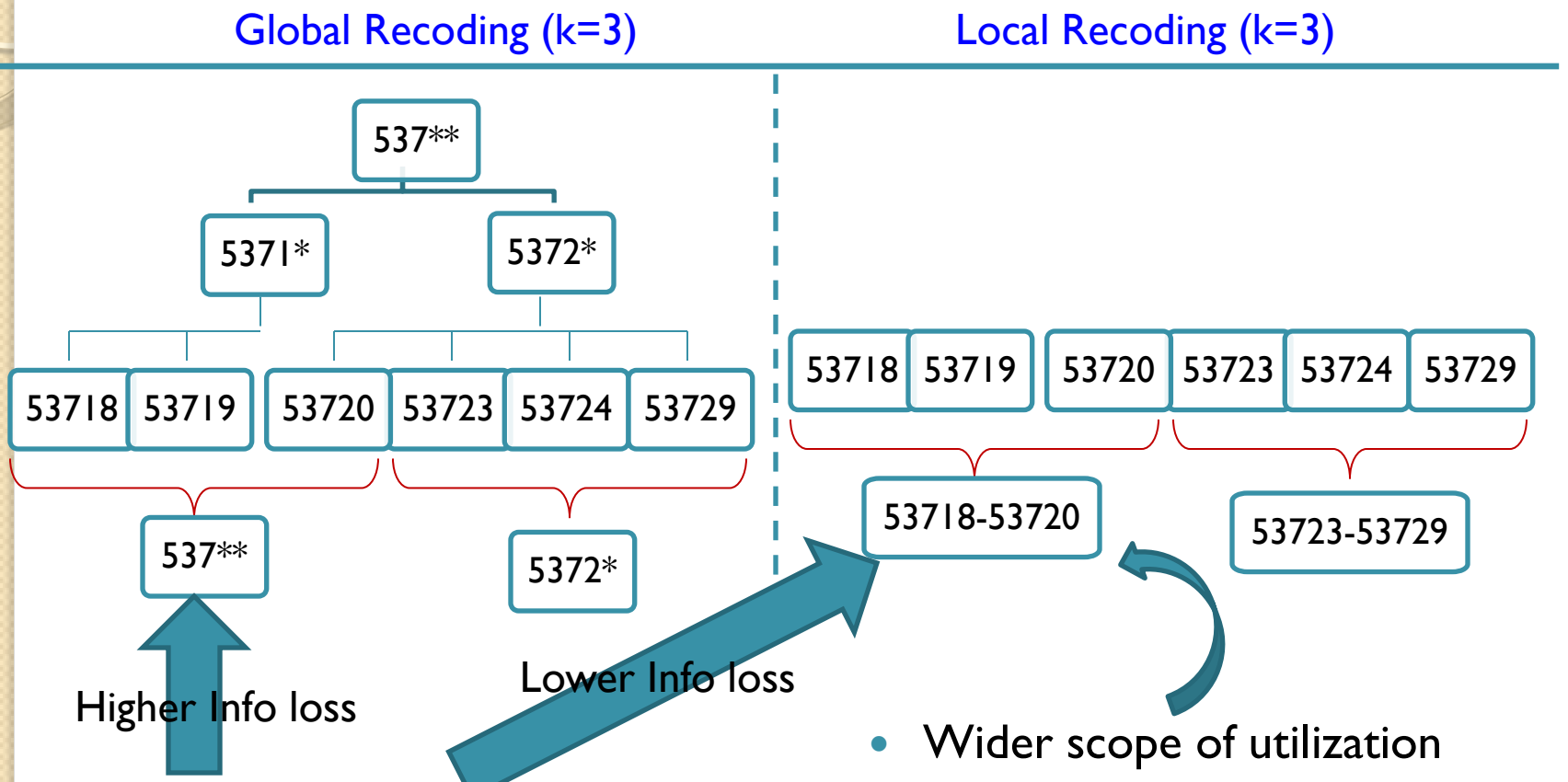
Age	Education	Marital S	Sex	Zip	S. Info
50~	<Bachelor	Married	Male	537**	Info_x, Info_y, Info_z
50~	<Bachelor	Married	Male	537**	
50~	<Bachelor	Married	Male	537**	
40~	<Bachelor	Never	Female	022**	Info_p, Info_q, Info_r
40~	<Bachelor	Never	Female	022**	
40~	<Bachelor	Never	Female	022**	
		.....	.....	.....	.....

Global Recoding vs. Local recoding

- Global Recoding always recodes a specific value to its same general value
- Local Recoding can recode a specific value to different general values



Global Recoding vs. Local recoding



- Information loss is proportional to group sizes.

- Wider scope of utilization

  Ex: Can query "attribute value between 53710 and 53720"

Data Quality (Information Loss) and Privacy Loss

Age	ZipCode	Disease
42-47	430-520	Ulcer
42-47	430-520	Pneumonia
51-55	270-320	Flu
51-55	270-320	Gastritis
62-67	410-550	Dyspepsia
62-67	410-550	Flu

$$\frac{max^{G}_{A_{Num}} - min^{G}_{A_{Num}}}{max_{A_{Num}} - min_{A_{Num}}}$$

$$NCP_{Age}(G_1) = \frac{47-42}{70-0}$$

$$NCP_{Zip}(G_1) = \frac{520-430}{999-000}$$

$$P_i = 1 - \frac{1}{GS_i}$$

LowCost Algorithm

- LowCost(Input T)
- Source=T
- While |Source|>k do {
  - Condition=""
  - Sort Asc Attribute on |Di|
  - For i=1 to |Ai|
    - Find Largest $|G_i|$ so that Cost(G_i) is minimized
    - Condition = Condition + "A_i between $G_{i(min\_value)}$ and $G_{i(max\_value)}$"
    - Source="Select * from Source where" + Condition
  - }
  - T`=T`+ "Select * from Source where" + Condition
  - Source = Source- "Select * from Source where" + Condition
- }
- Output (T`)

Assumptions

- Data are ordered based on group sizes on each column
- Attributes are ordered based on attribute cardinality

Anonymization Example

1. Find the Largest Group with Minimum Cost Attr (A_2)
2. Create Clause: “Sex =1”
3. Update data source for next Attr: (A_3)

……

4. Make Equivalence Class and Move to Anonymous table

$A_2 \rightarrow A_3 \rightarrow A_1$

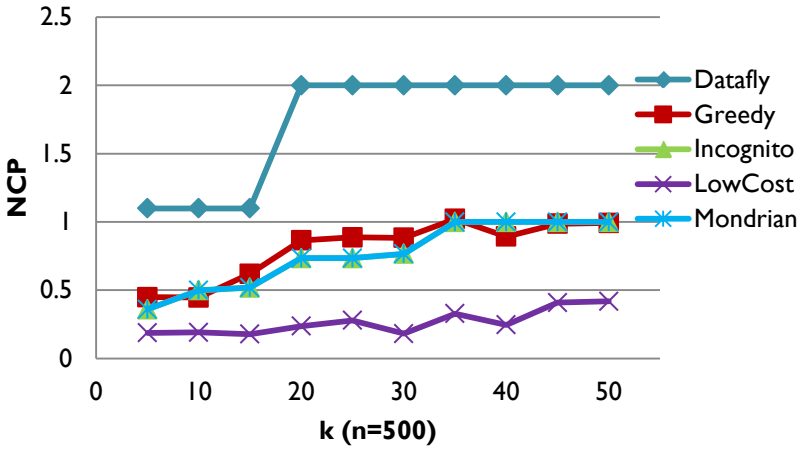
Male: 1
Female: 0

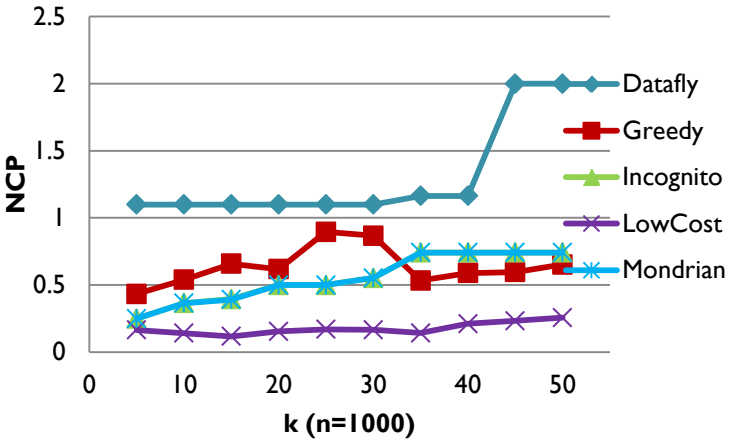
2946	2004/07/13	1	635
1007	2008/01/04	1	636
1324	2008/08/02	1	637
21	2006/03/26	1	637
802	2005/06/13	1	637
1651	2004/06/19	1	637
1665	2004/01/06	1	637
2556	2003/04/16	1	637
2633	2001/01/10	1	637
2608	2000/03/14	1	637
2369	2004/01/30	1	640
1540	2006/01/17	1	641

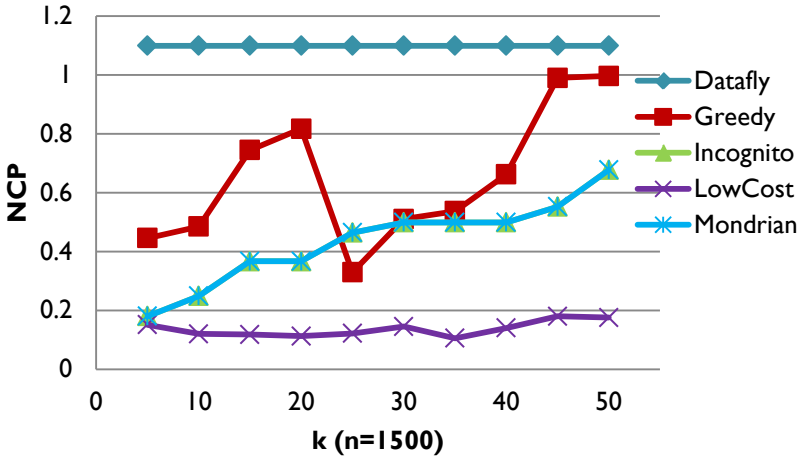
2369	2004/01/30	1	640
1324	2008/08/02	1	637
21	2006/03/26	1	637
802	2005/06/13	1	637
1651	2004/06/19	1	637
1665	2004/01/06	1	637
2556	2003/04/16	1	637
2633	2001/01/10	1	637
2608	2000/03/14	1	637
1007	2008/01/04	1	636
2830	2009/03/17	1	635
2946	2004/07/13	1	635

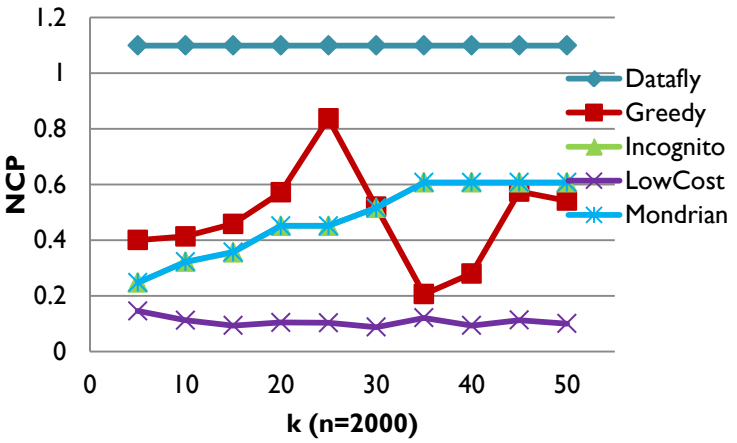
- "Sex = ‘Male’"
- “Sex = ‘Male’ and Zip = 637“
- " Sex = ‘Male’ and Zip = 637 and Birth between #2006/03/26# and #2003/04/16#"

Comparison of NCP Cost (Per record)



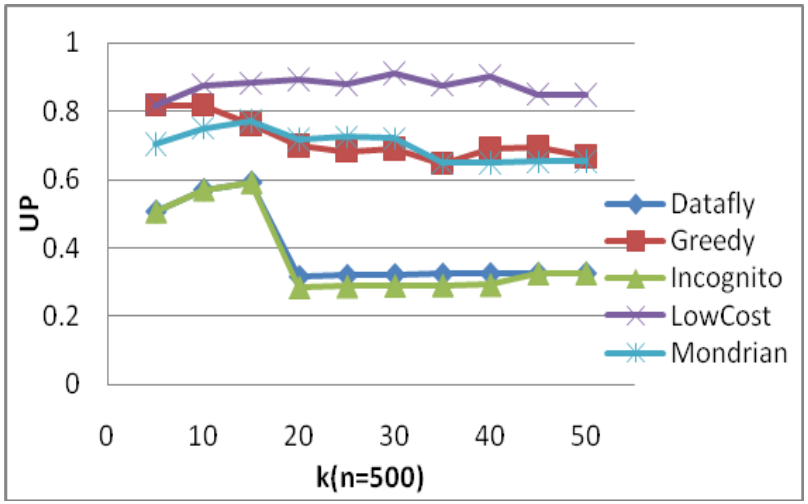


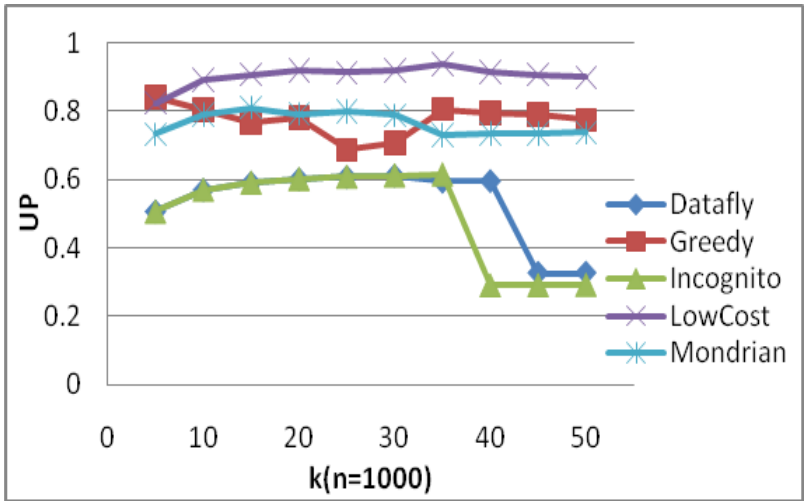




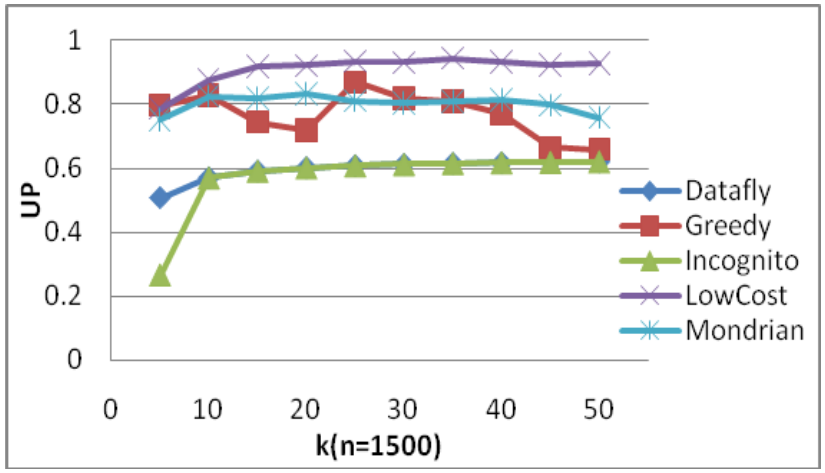
Comparison of Utility-privacy measure (Per Data unit)

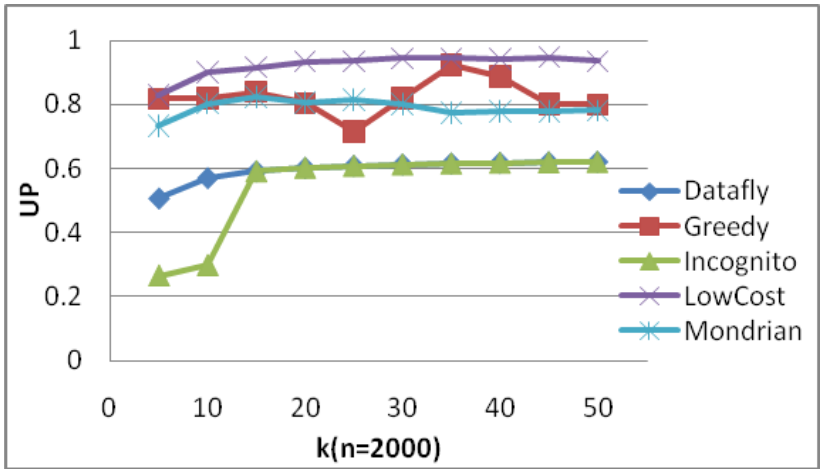
$$UP = \left(1 - \frac{1}{nd} \cdot \sum_{i=1}^{i=n} \sum_{j=1}^{j=d} NCP_{A_{i,j}}^{E_{i,j}}\right) \times \left(1 - \frac{1}{|G|} \sum_{l=1}^{l=|G|} \frac{1}{GS_l}\right)$$

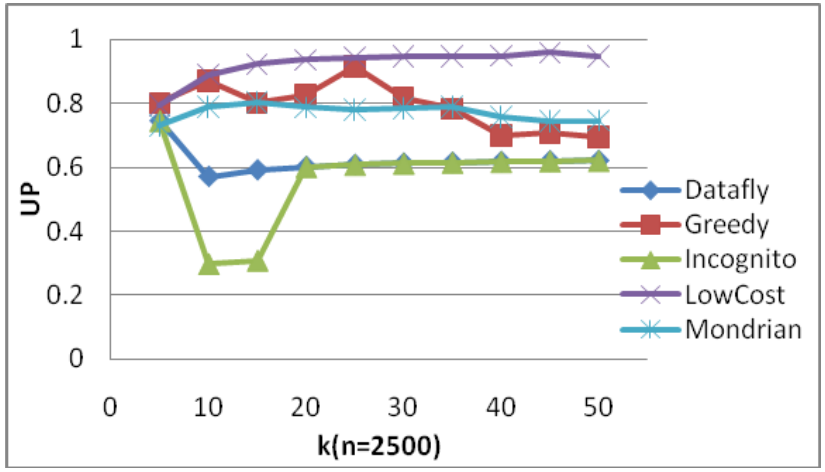


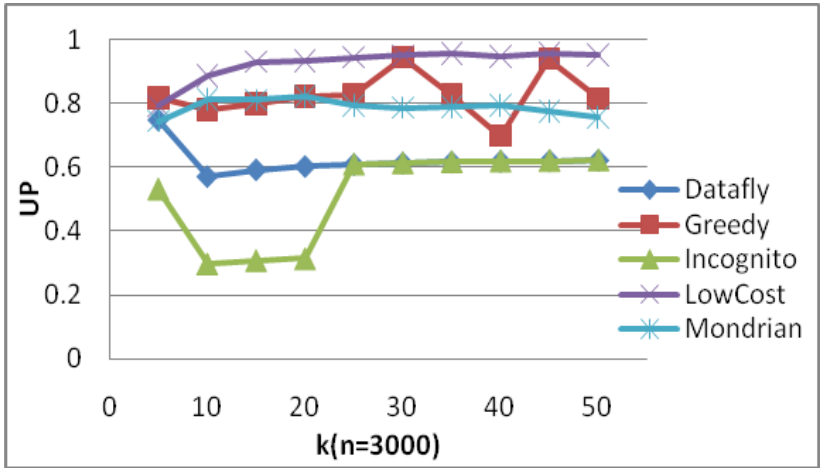


Comparison of Utility-privacy measure (per data unit)









Summary

- An efficient k-anonymization algorithm (MinCost) has been presented
  - Gives high quality for third-party use without compromising privacy level
  - Better performance compared to existing algorithms
- Future Work: make MinCost algorithm satisfy l-diversity

Thank you!