Homework 1

Swathi Dinakaran (sdinaka)

Q1. (a) Using Definition 4 of the paper, complete the function compute weights() and compute the weights of each movie. Tabulate the weights obtained for each movie. This should be a table with 15 movie-ids and their corresponding weights. [10]

movie : weight

03124 : 0.08158933699959232

06315 : 0.08160505491876055

07242 : 0.08164844369534954

16944 : 0.08161489477641136

17113 : 0.0816603191038822

10935 : 0.08164646622070376

11977 : 0.08166626360766739

03276 : 0.08168611145531343

14199 : 0.0816326379485879

08191 : 0.08166824611782734

06004 : 0.0876552503936246

01292 : 0.08768793722221335

15267 : 0.08772896619363886

03768 : 0.08765932965065752

02137 : 0.08763895214961775

(b) How many users are present in the database? Using Definition 7 in the paper, complete the function score() and compute the scores of the auxiliary information with respect to every user's ratings in the database. What is the highest score? What is the second highest score? [15+5]

# users in db: 44651

Highest score: 0.07857135454521895

Second highest: 0.07133917881063469

c) What is the user-id of the user with the highest score? Write out the ratings of this user from the database, and verify if they are similar to the ratings in the auxiliary information

user-id of the user with the highest score :1664010

Comparison:

```
movie |    user_rating | aux_rating | difference

14199 :          4          4.5              0.5

17113 :          4          4.2              0.20000000000000018

06315 :          4          4.0              0.0

01292 :          3          3.3              0.2999999999999998

11977 :          4          4.2              0.20000000000000018

15267 :          4          4.2              0.20000000000000018

08191 :          4          3.8              0.20000000000000018

16944 :          4          4.2              0.20000000000000018

07242 :          4          3.9              0.10000000000000009

06004 :          4          3.9              0.10000000000000009

03768 :          4          3.5              0.5

03124 :          4          3.5              0.5
```

(d) Assume the eccentricity metric is γM, where $M=\sum w(i)\ |supp(aux)|\ i \in supp(aux)$ is the scaled sum of weights of attributes in aux. Say, γ = 0.1 then what is the value of the eccentricity threshold? What is the difference between the highest and second highest score? Is it greater than the eccentricity metric? [4+4+2]

value of the eccentricity threshold : 0.008365139005235106

difference between the highest and second highest score : 0.007232175734584262

Threshold is greater than difference between max and second max value

Q2. For the following questions consider table below-

(a) What are the quasi-identifiers and sensitive attributes in this table? [5]

Quasi-identifiers are such attributes that are not individually unique but Attributes whose values when taken together can potentially identify an individual.. They are such attributes that donot contain sensitive information about the entity in database. In the given table, **(ZIP code, Age)** are the set of quasi identifiers, since they are not individually unique and also do not contain any sensitive information.

Sensitive attributes are attributes that can contain sensitive information about the user, which may reveal information about the user that can be used for malicious purposes. These attributes need utmost protection from attackers. In the given table, **Salary and Disease** are sensitive attributes, since Salary of a person can be maliciously used by the attacker. And disease information can be misused by insurance companies. Overall misuse of such attributes place the privacy of user at stake.

(b) Compose a 3-anonymous, 3-diverse table. Show intermediate steps for deriving the final solution and also draw the generalization lattice (i.e., using incognito algorithm) for your solution. [15+10]

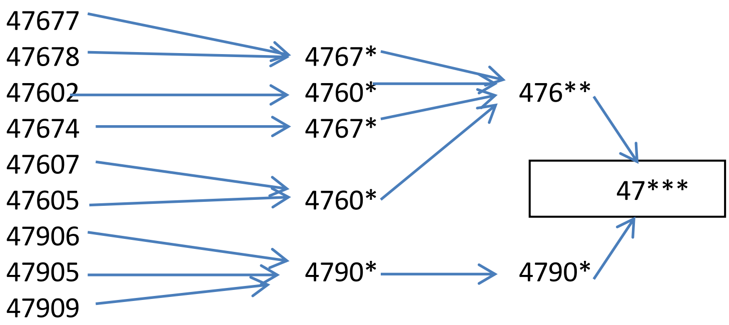
i) classify attributes

Primary key	Quasi Identifiers		Sensitive attributes	
ID	ZIP code	Age	Salary	Disease
1	47677	29	3K	Gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	Stomach cancer
4	47905	43	6K	gastritis
5	47605	30	7K	flu
6	47906	47	8K	bronchitis
7	47674	36	9K	bronchitis
8	47607	32	10K	pneumonia
9	47909	52	11K	Stomach cancer

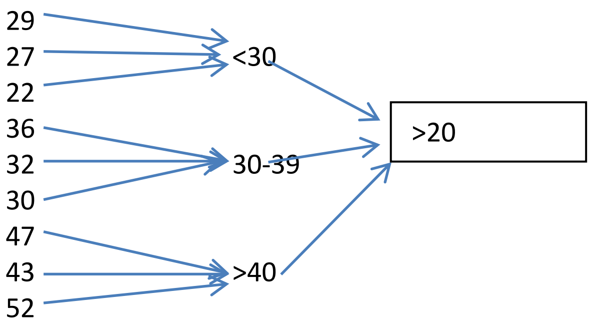
ii) k anonymity: A dataset is k-anonymous if every combination of identity-revealing characteristics occurs in at least k different rows of the data set. Here the equivalence class should contain atleast 3 different rows of data set

iii) Generalisation: A technique used to replace more specific value with generic and semantically similar values. Here generalisation can be done on quasi identifiers, Age and Zip. It could be divided into 3 categories. Age: 20-29 , 30-39 ,>40 and ZIP: 476**, 479**.

Zip Values:



Age:



iv) Generalised 3-anonymous table:

1	476**	2*	3K	Gastric ulcer
3	476**	2*	5K	Stomach cancer
2	476**	2*	4K	gastritis
7	476**	3*	9K	bronchitis
8	476**	3*	10K	pneumonia
5	476**	3*	7K	flu
6	4790*	>40	8K	bronchitis
4	4790*	>40	6K	gastritis
9	4790*	>40	11K	Stomach cancer

Where k=3

v)3 diverse: k-Anonymity does not provide privacy if: Sensitive values in an equivalence class lack diversity and if the attacker has background knowledge

Therefore, each equivalence class has at least l well-represented sensitive values. Here each equivalence class has 3 distinct diseases evenly distributed. Hence the table is 3 diverse.

vi) Generalisation lattice:

Hierarchy:

Z3={47***}

↑

Z2={476*,4790*}

↑

Z1={4767*, 4760*, 4790*}

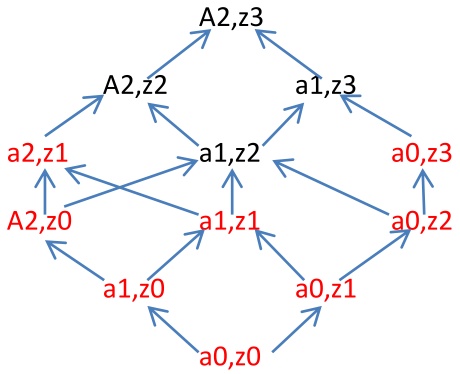
↑

Z0={47677, 47602 , 47678 ,47905 ,47605 ,47906 ,47674 ,47607 ,47909} a2{>20}

a1{<29 , (30-39) , >40}

a0{29, 22, 27, 43, 30, 47, 36, 32, 52}

Lattice:



(c) Compute the t-closeness of your solution with respect to salary. Does your solution resolve the ‘similarity attack’ for the sensitive attribute ‘disease’? If not, compute an alternative solution. [10+10]

T closeness:

The idea of t-closeness is that the distribution of sensitive data in every group is not too far from the distribution in the full population

L-diversity does not consider semantics of sensitive values! T-closeness considers overall distribution Q of sensitive value.

Q(Salary)={ 3K,4K,5K,6K,7K,8K,9K,10K,11K }

P1(Eq 1 Salary)=(3K, 5K, 4K)

P2(Eq 2 Salary)=(9K, 10K, 7K)

P3(Eq 3 Salary)=(8K, 6K, 11K)

*D[P1,Q] : transform P1 to Q –

Move 1/9 probability for each of the following pairs

Cost= amount of dirt moved * the distance moved

- 3k ->6k, 3k ->7k cost: 1/9*(3+4)/8 =1/9*7/8
- 4k->8k,4k->9k cost: 1/9*(4+5)/8 =1/9*9/8
- 5k->10k,5k->11k cost: 1/9*(5+6)/8 =1/9*11/8

– Total cost: 1/9*27/8=0.375

* D[P2,Q] : transform P2 to Q –

Move 1/9 probability for each of the following pairs

Cost= amount of dirt moved * the distance moved

- 9k ->8k, 9k ->5k cost: 1/9*(1+4)/8 =1/9*5/8
- 10k->11k, 10k->8k cost: 1/9*(1+2)/8 =1/9*3/8
- 7k->3k, 7k->4k cost: 1/9*(4+3)/8 =1/9*7/8

– Total cost: 1/9*15/8=0.208

* D[P3,Q] : transform P3 to Q –

Move 1/9 probability for each of the following pairs

Cost= amount of dirt moved * the distance moved

- 8k ->7k, 8k ->5k cost: 1/9*(1+3)/8 =1/9*4/8
- 6k->4k, 6k->3k cost: 1/9*(2+3)/8 =1/9*5/8
- 11k->10k, 11k->9k cost: 1/9*(1+2)/8 =1/9*3/8

– Total cost: 1/9*12/8=0.167

Q=(0.375+0.208+0.167)/3= 0.25

* D[P1,Q]> Q => 0.375> 0.25

* D[P2,Q] < Q => 0.208< 0.25

* D[P3,Q] < Q => 0.167< 0.25

Out of EQ1,EQ2 and EQ3, EQ3(0.167) is the one with most optimal t-closeness.

Therefore the table has 0.167 t-closeness w.r.t salary.

ii)

The table is not immune to similarity attack. If we know a person has a low salary (3k-5k) then we know that he has a stomach related disease.

This is because l-diversity takes into account the diversity of sensitive values in the group, but does not take into account the semantical closeness of the values.

To make the table immune to similarity attack, make each equivalence class satisfy t-closeness .

1	476**	20-39	3K	Gastric ulcer
3	476**	20-39	5K	Stomach cancer
2	476**	20-39	4K	gastritis
7	476**	20-39	9K	bronchitis
8	476**	20-39	10K	pneumonia
5	476**	20-39	7K	flu
6	4790*	>40	8K	bronchitis
4	4790*	>40	6K	gastritis
9	4790*	>40	11K	Stomach cancer

The above table has 2 equivalence classes that are not vulnerable to similarity attack..The above table satisfies 3 anonymity , 3 diverse and similarity attack can't be performed

Q(Salary)={ 3K,4K,5K,6K,7K,8K,9K,10K,11K }

P1(Eq 1 Salary)=(3K, 5K, 4K, 9K, 10K, 7K)

P2(Eq 2 Salary)=(8K, 11K, 6K)

*D[P1,Q] : transform P1 to Q –

- 3k ->8k,cost: 3/9*(5)/8 =1/3*5/8=5/24
- 5k ->8k, cost: 3/9*(3)/8 =1/3*3/8=3/24
- 4k->11k, cost: 3/9*(7)/8 =1/3*7/8=7/24
- 9->11k, cost: 3/9*(2)/8 =1/3*2/8=2/24
- 10k->6k, cost: 3/9*(4)/8 =1/3*4/8=4/24
- 7k->6k, cost: 3/9*(1)/8 =1/3*1/8=1/24
  -total cost: 22/24=11/12

* D[P2,Q] : transform P2 to Q –

- 8k ->3k, cost: 3/9*(5)/8 =1/3*5/8=5/24
- 8k ->5k, cost: 3/9*(3)/8 =1/3*3/8=3/24
- 11k->4k, cost: 3/9*(7)/8 =1/3*7/8=7/24
- 11k->9k, cost: 3/9*(2)/8 =1/3*2/8=2/24
- 6k->10k, cost: 3/9*(4)/8 =1/3*4/8=4/24
- 6k->7k, cost: 3/9*(1)/8 =1/3*1/8=1/24
  -total cost: 22/24=11/12

D[P1,Q]= D[P2,Q]=avg=Threshold.