Anonymizing Multidimensional Data: k-Anonymity

UNIVERSITÀ DEGLI STUDI DI GENOVA | Dibris

Davide Caputo
davide.caputo@dibris.unige.it

Multidimensional Data: EI, QI and SD

EI		QI				SD			
ID	Name	Gender	Age	Address	Zip	Basic	HRA	Med	All
12345	John	M	25	1, 4th St.	560001	10,000	5,000	1000	6,000
56789	Harry	M	36	358, A dr.	560068	20,000	10,000	1000	12,000
52131	Hari	M	21	3, Stone Ct	560055	12,000	6,000	1000	7,200
85438	Mary	F	28	51, Elm st.	560003	16,000	8,000	1000	9,600
91281	Srini	M	40	9, Ode Rd	560001	14,000	7,000	1000	8,400
11253	Chan	M	35	3, 9th Ave	560051	8,000	4,000	1000	4,800

Anonymization Techniques: Requirements

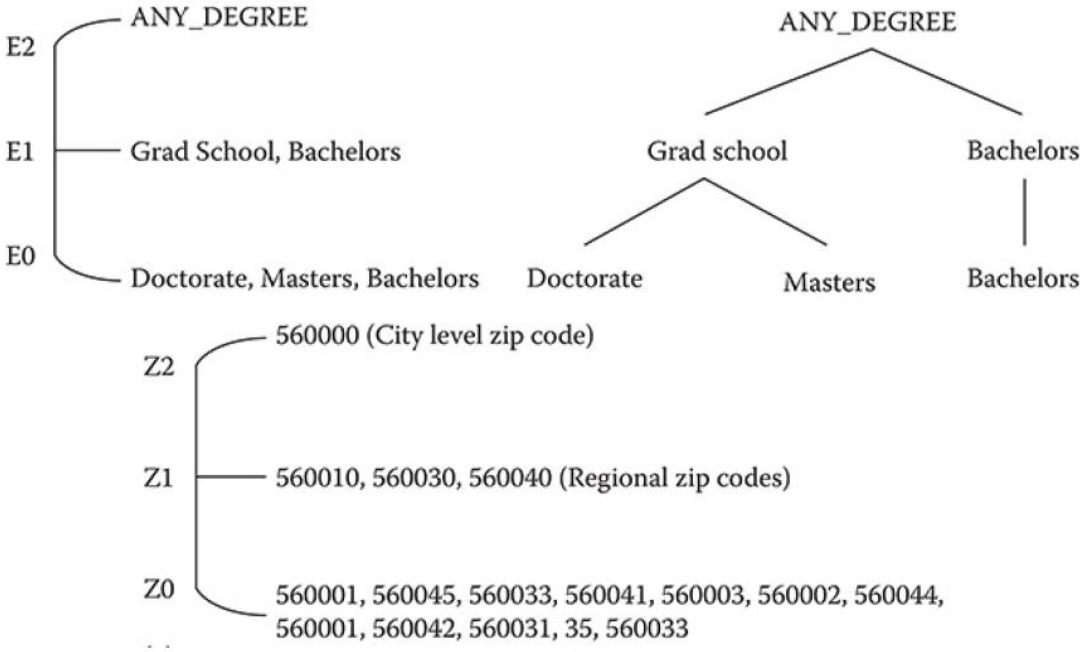
1. Prevent the **Record Linkage**
2. Preserve the **utility of the transformed data**
3. Ensure the **protection of outlier record**
4. Preserve **the correlation/association between QI and SD**

Group-based Anonymization: K-Anonymity

- **k-Anonymization** is a technique for preserving individual identification by transforming the record set that **each record of a table identical to at least *k - 1* other records**

- k-Anonymization is granted by **generalizing** and **suppressing** the value of attribute

Data Generalization



Example 4-anonymous salary table

ID	Gender	Day	Month	Year	Address	City	Zip Code	Education	Years of Experience	Salary
1	Any Sex	—	—	1973	—	BLR	560000	Any_Degree	20	35,000
2	Any Sex	—	—	1975	—	BLR	560040	Any_Degree	17	28,000
3	Any Sex	—	—	1977	—	BLR	560030	Any_Degree	18	15,000
4	Any Sex	—	—	1974	—	BLR	560040	Any_Degree	20	38,000
5	Any Sex	—	—	1985	—	BLR	560000	Any_Degree	12	10,000
6	Any Sex	—	—	1980	—	BLR	560000	Any_Degree	10	9,000
7	Any Sex	—	—	1977	—	BLR	560040	Any_Degree	15	18,000
8	Any Sex	—	—	1978	—	BLR	560000	Any_Degree	18	22,000
9	Any Sex	—	—	1980	—	BLR	560040	Any_Degree	20	15,000
10	Any Sex	—	—	1982	—	BLR	560030	Any_Degree	15	32,000
11	Any Sex	—	—	1980	—	BLR	560030	Any_Degree	12	14,000
12	Any Sex	—	—	1979	—	BLR	560030	Any_Degree	14	16,000

k-Anonymity: Datafly Algorithm

Input: Private Table **PT**; quasi-identifier QI = $(A_1, \ldots, A_n)$,
k constraint; hierarchies DGH_{Ai}, where i=1,…,n.
Output: MGT, a generalization of PT[QI] with respect to k
Assumes: |PT |≥ k
Method:

1. freq ← a frequency list contains distinct sequences of values of PT[QI], along with the number of occurrences of each sequence.
2. **while there exists** sequences in freq occurring less than k times that account for more than k tuples **do**
   2.1. **let** A_j be attribute in freq having the most number of distinct values
   2.2. freq ← generalize the values of A_j in freq
3. freq ← suppress sequences in freq occurring less than k times.
4. freq ← enforce k requirement on suppressed tuples in freq.
5. **Return** MGT ← construct table from freq

k-Anonymity: Datafly Algorithm (1)

Race	BirthDate	Gender	ZIP	#occurs	
black	9/20/65	male	02141	1	t1
black	2/14/65	male	02141	1	t2
black	10/23/65	female	02138	1	t3
black	8/24/65	female	02138	1	t4
black	11/7/64	female	02138	1	t5
black	12/1/64	female	02138	1	t6
white	10/23/64	male	02138	1	t7
white	3/15/65	female	02139	1	t8
white	8/13/64	male	02139	1	t9
white	5/5/64	male	02139	1	t10
white	2/13/67	male	02138	1	t11
white	3/21/67	male	02138	1	t12
2	12	2	3		

A

Race	BirthDate	Gender	ZIP	#occurs	
black	1965	male	02141	2	t1,t2
black	1965	female	02138	2	t3, t4
black	1964	female	02138	2	t5, t6
white	1964	male	02138	1	t7
white	1965	female	02139	1	t8
white	1964	male	02139	2	t9, t10
white	1967	male	02138	2	t11, t12
2	3	2	3		

B

k=2, QI = {Race, BirthDate, Gender, Zip}

k-Anonymity: Datafly Algorithm (1)

Race	BirthDate	Gender	ZIP	#occurs	
black	9/20/65	male	02141	1	t1
black	2/14/65	male	02141	1	t2
black	10/23/65	female	02138	1	t3
black	8/24/65	female	02138	1	t4
black	11/7/64	female	02138	1	t5
black	12/1/64	female	02138	1	t6
white	10/23/64	male	02138	1	t7
white	3/15/65	female	02139	1	t8
white	8/13/64	male	02139	1	t9
white	5/5/64	male	02139	1	t10
white	2/13/67	male	02138	1	t11
white	3/21/67	male	02138	1	t12
2	12	2	3		

A

WILL BE SUPPRESSED!!
(STEP 2)

Race	BirthDate	Gender	ZIP	#occurs	
black	1965	male	02141	2	t1,t2
black	1965	female	02138	2	t3, t4
white	1964	male	02138	1	t7
white	1965	female	02139	1	t8
white	1967	male	02138	2	t11, t12
2	3	2	3		

B

k=2, QI = {Race, BirthDate, Gender, Zip}

k-Anonymity: Datafly Algorithm (2)

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

References

- L. Sweeney, “Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, ”International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,vol. 10,no. 05,pp. 571–588,2002. [Online]. Available:http://www.worldscientific.com/doi/abs/10.1142/S0218488502000165X

Clone and Complete the Code

https://github.com/Dado1513/python-datafly-to-complete

```
while True:
    ## Complete with the implementation of datafly algorithm
    do_stuff()
```