Data Privacy and Security Assignment 1 Report

Ismayil Ismayilov

	Random			Clustering			Top-down		
k	Time	MD	LM	Time	MD	LM	Time	MD	LM
5	0.59	701304	31598	5287	157587	5862	51	438934	10146
10	0.56	781512	35261	2529	255312	10093	48	470127	11686
20	0.51	831012	37065	1269	355068	14860	44	495724	12831
40	0.52	862492	38231	638	447932	19395	47	542582	17424
80	0.52	873332	38859	339	537652	23875	48	572887	18974
160	0.52	875092	38988	182	627892	28125	41	607981	21126
320	0.49	875092	38988	95	724052	32163	36	650511	23643

Table 1: Experiment results

Discussion

Random

The fastest among the anonymizers taking an average of 0.5 seconds for any k. Not surprisingly, it is also the worst performing one achieving 701403 and 31598 for the minimum MD and LM cost respective (for $k = 5$). Anonymization gets worse as k increases with $k = 160$ and $k = 320$ getting the worst possible MD and LM costs.

Clustering

The best performing anonymizer achieving 157587 and 5862 for the MD and LM costs respectively. Also the most compute intensive one. Executing for $k = 5$ takes about 1.5 hours. Overall, the execution time is far higher than that for the other anonymizers.

Top-Down

For all values of k, on average, takes less than a minute to execute. The utility loss results are not as good for those clustering but not as bad as those for random; utility losses are approximately halfway between results for random and clustering.

Summary

Overall, clustering gives the lowest utility loss at the cost of much higher execution time. Random is the fastest but also the worst performing one; since the clusters are picked randomly they are very dissimilar leading to excessive generalization. Top-down gives adequate utility loss and acceptable execution time. If the goal is to get the best anonymization it would be better to use Clustering; if time is a factor but adequate anonymization is required, Top-Down would be a good choice; Random should probably not be used at all.