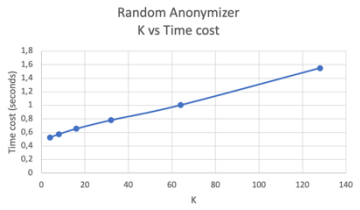
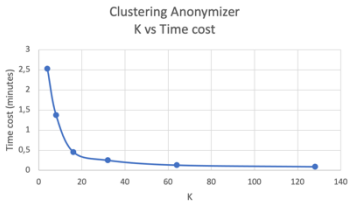
Comp430 HW#1
Burak Yıldırım 72849

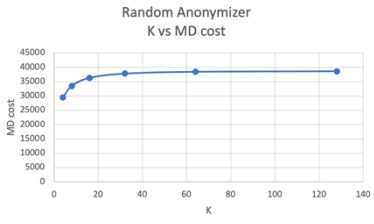
Random Anonymizer:
For K=4, LM cost = 1559.9186480186522, MD cost = 29514, time cost = 0:00:00.527636
For K=8, LM cost = 1787.3353535353444, MD cost = 33506, time cost = 0:00:00.574737
For K=16, LM cost = 1898.8717948717683, MD cost = 36306, time cost = 0:00:00.654780
For K=32, LM cost = 1956.4102564102323, MD cost = 37810, time cost = 0:00:00.781818
For K=64, LM cost = 1986.4615384615279, MD cost = 38450, time cost = 0:00:01.004709
For K=128, LM cost = 2000.0, MD cost = 38642, time cost = 0:00:01.550596

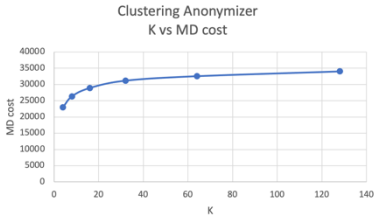
Clustering Anonymizer:
For K=4, LM cost = 1133.3459984459907, MD cost = 23022, time cost = 0:02:53.355317
For K=8, LM cost = 1323.783216783209, MD cost = 26362, time cost = 0:01:38.182128
For K=16, LM cost = 1446.412742812729, MD cost = 28930, time cost = 0:00:46.207690
For K=32, LM cost = 1532.9926961926883, MD cost = 31186, time cost = 0:00:25.277160
For K=64, LM cost = 1610.275058275048, MD cost = 32562, time cost = 0:00:13.671979
For K=128, LM cost = 1701.9474747474671, MD cost = 34034, time cost = 0:00:09.370776



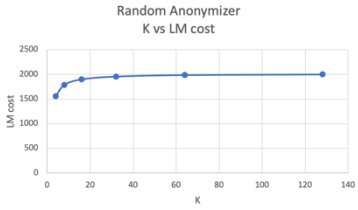


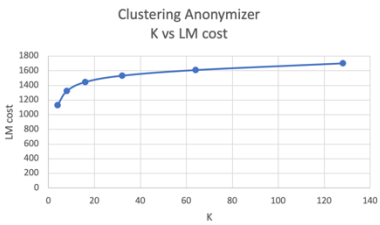
In the random anonymizer, when k increases time cost also increases because generalizing more records takes more time. In the clustering anonymizer, when k increases time cost decreases because we mark k records in each iteration and since we don't have to look same records again and again it gets faster in each iteration.





In the random and clustering anonymizers, when k increases MD cost also increase because more generalization will be required to satisfy higher k-anonymity. In the random anonymizer, MD costs tend to be higher because it's totally random whereas in the clustering anonymizer we are using a dist metric and selecting ECs by keeping the dist metric as minimum reduced the MD cost.





LM considers branching factor and depth of attributes. In the random and clustering anonymizers, when k increases LM cost also increase because more generalization will be required to satisfy higher k-anonymity. In the random anonymizer, LM costs tend to be higher because it's totally random whereas in the clustering anonymizer we are using a dist metric and selecting ECs by keeping the dist metric as minimum reduced the LM cost.

I explained the trade-offs next to the charts above. Random anonymizer is the fastest due to it doesn't select ECs with a certain rule. Clustering anonymizer has the lowest utility loss in terms of both LM and MD costs since it tries to make LM cost minimum which leads to reducing the total table cost. I would prefer random anonymizer to produce a faster result but more utility loss. I would prefer clustering anonymizer to produce lower utility loss and better results but slower in time. The results fit my expectations because I was also expecting the clustering anonymizer perform better but slower. In addition, I expected the LM cost equal to 2000 at max because in the adult dataset, there are 2000 records and 8 fields. If all of them becomes any lm cost will be 1 for each field and 8*1=8 for each record. Assuming weights are identical, weight of an attribute is 1/8. Weight * lm cost of a record = 1 and for 2000 records it can be 2000 at max which we can observe in the charts it holds. From this assignment, I learned that trying to minimize the LM cost while selecting the ECs leads to lower utility loss. Our focus should be achieving k-anonymity with a lowest possible utility loss.