Berke Can Rizai - 69282 | | Costs below are averaged over two runs.

Randomized Anonymizer:

K	Time Cost	MD Cost	LM Cost
4	30s	29350	1535
8	34s	33340	1775
16	37.5s	36050	1886
32	38.5s	37640	1950
64	40s	38500	1990
128	41s	38650	2000

Clustering Anonymizer:

K	Time Cost	MD Cost	LM Cost
4	45s	28670	1580
8	44s	30052	1695
16	34.2s	30758	1749
32	28.8s	30770	1760
64	25.7s	30785	1760
128	24s	30785	1760

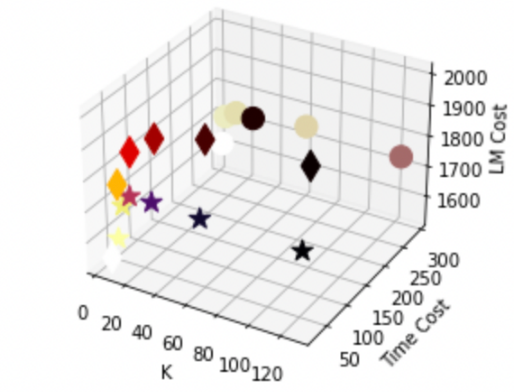
Bottom-up Anonymizer:

K	Time Cost	MD Cost	LM Cost
4	4:52 mins	32642	1610
8	4:47 mins	34640	1714
16	4:45 mins	30758	1739
32	4:29 mins	30775	1750
64	4:44 mins	34642	1739
128	4:35 mins	34642	1739

If we compare the LM Cost, MD Cost and Time cost for K values, we see that random anonymizer had positive correlation between more generalization & higher utility loss and the time taken for computation however, there is no clear line in the bottom-up anonymizer and time is more unpredictable while in the clustering, time and LM are negatively correlated.

Another observation is that in the bottom-up anonymizer, k=64 and k=128 gave the same solution, this might be due to fact that after completing the whole level in the branching tree, I first check the all features generalized version in the next level because otherwise it would take hours since there are many nodes possible even though I stored them in set so that I don't revisit them again. This lets the algorithm iterate over first layers faster. In my implementation, levels are decreasing as we generalize (as opposed to increasing in the lecture) and if we have (a, b, c) at current node, if (a-1, b-1, c-1) does not satisfy K anonymity we don't need to try (a, b, c-1), (a, b-1, c) and (a-1, b, c).

We can see that as K grows, bottom-up approach is better in terms of utility loss however that comes at great cost of time even in the dataset of 2000 rows. For small K, clustering seems to be the most viable option in terms of MD, LM and time combined. And it doesn't grow much in time with bigger Ks.



Here on the left, we can see K, Time vs LM graph. Star is clustering, diamonds are random, and circles are bottom-up strategy. We can see that LM cost of randomized is highest for same amount of K and for the time cost, bottom up is the worst by a wide margin. Stars (clustering) seem to perform better than the competitors if we can trade some LM cost with time complexity. Colors are representing the time compared to itself (other runs of same method). For small K, if we can accept some utility loss, randomized method might be feasible.