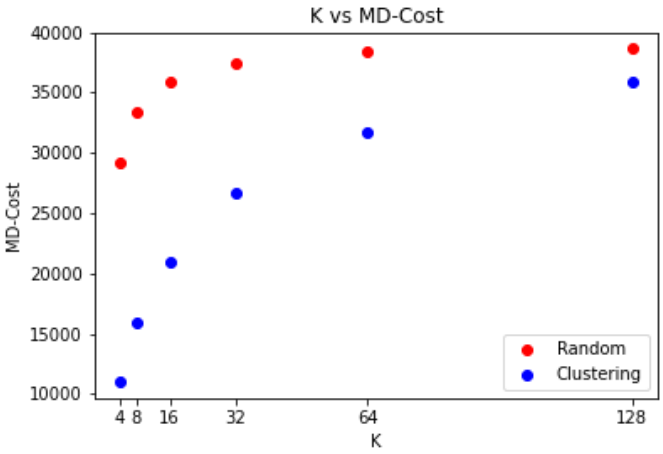
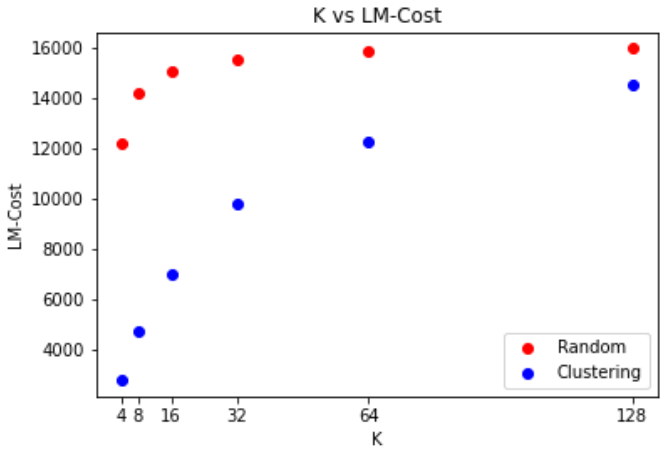
Ertan Can Güner
60610

Comp430 HW1 Report
K-Anonymization Algorithms

I have implemented 2 different anonymization algorithms: random and clustering. Algorithms ran 3 times for each K value and averages are taken to form the graphs. Random anonymization is the fastest algorithm to run but it makes the larges loss in utility in comparison to other algorithms. Clustering would result in least utility loss if a better closeness heuristic can be used other than naïve LM cost solution I have implemented. Slowest algorithm would be the bottom-up approach because in worst case it would've checked all possible generalization methods to find the one that satisfies k-anonymity. Results fit my expectation in regards the time costs because during the implementation I could calculate the complexity of the algorithms, but I didn't expect the similar figures for "k vs MD cost" and "k vs LM cost". I couldn't learn much from the assignment other that the main concepts because I had time management problems on my part and there are a lot of unanswered questions on my mind because of the lack of resources on the internet and in the lectures. For example, generalization method I used uses a naïve approach: after parsing the DGHs into a tree structure I find the least common ancestor of two child nodes which results in a lot the field values in the data set to become the root node. This results in too much utility loss but if there are enough data in the data set then this can be overcome but still there is too much utility loss.





	K					
	4	8	16	32	64	128
Random	1.2 sec	1.2 sec	1.3 sec	1.4 sec	1.6 sec	2 sec
Clustering	17 min 40 sec	20 min 50 sec	22 min 40 sec	23 min 30 sec	24 min 30 sec	25 min