CS 528 Data Privacy and Security
Homework 1
Due Tuesday, 02/26/2019 (11:59 PM)

Name: **CWID:**

Part I (50 points). Design and implement a heuristic algorithm to ensure personalized k-anonymity for the Adult dataset in UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

- To simplify the problem, you only need to consider 4 attributes as quasi-identifiers (QIs) to implement the generalization and/or suppression:

  1. Age: continuous
  2. Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
  3. Marital-Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
  4. Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

  The full dataset description is available at: https://archive.ics.uci.edu/ml/machine-learning-databases/adult/old.adult.names

- Note that the dataset has missing values. If so, it is considered as "Generalized to the top of the hierarchy".

- Each record can specify a different k. In the output, the cardinality of each equivalence class (a group of records with identical QI values) should be no less than all the k values in the group.

- Tasks include:

  1. define a hierarchy for each of the 4 attributes. (**10 points**)
  2. write a program for the heuristic algorithm (which generalizes/suppresses the data while minimizing the utility loss). You can use any programming language you feel comfortable, e.g, Java, Python and C++. (**30 points**)
  3. calculate the distortion and precision of your algorithm. (**10 points**)

Submission includes (input dataset, output dataset, source code files, hierarchies and the distortion/precision results) – all named with the prefix "hw1-I-" (e.g., *hw1-I-PAnonK.java*).

Part II (50 points). Using the same dataset (UCI Machine Learning Adult data) to study differential privacy (centralized).

- *Laplace Mechanism*: query the average age of the records with age greater than 25 (considering each record belonging to an individual adult). Inject Laplacian noise to the query result (average age) to ensure 0.5-differential privacy and 1-differential privacy. Solo a la media parece... el algoritmo es la puta media

  1. in case of $\epsilon = 0.5$, generate 1,000 results for the query over the original dataset, and generate 1,000 results for the query over each of three other datasets (removing a record with the oldest age; removing any record with age 26; removing any record with the youngest age). (**6 points**)
  2. in each of the above 4 groups of 1,000 results, round each number to two decimal places, define a measure and utilize it to validate that each of the last 3 groups of results and the original results are 0.5-*indistinguishable.* (**6 points**) The measure? Is the average, but how could I calculate the probability
  3. repeat all the above for $\epsilon = 1$, utilize the above measure to validate that each of the last 3 groups of results and the original results are 1-*indistinguishable.* (**6 points**)
  4. define another measure and utilize it to justify that the distortion of the 4,000 results for $\epsilon = 1$ is less than that of $\epsilon = 0.5$. (**7 points**)

- *Exponential Mechanism*: query the most frequent "Education" result. Design an exponential mechanism (randomized) to ensure ϵ-differential privacy for the query. Repeat all the procedures for Exponential mechanism ($\epsilon = 0.5$ and $\epsilon = 1$):

  5. in case of $\epsilon = 0.5$, generate 1,000 results for the query over the original dataset, and generate 1,000 results for the query over each of three other datasets (removing a record with the most frequent "Education"; removing any record with the second most frequent "Education"; removing any record with the least frequent "Education"). (**6 points**)
  6. in each of the above 4 groups of 1,000 results, define a measure and utilize it to validate that each of the last 3 groups of results and the original results are 0.5-*indistinguishable.* (**6 points**)
  7. repeat all the above for $\epsilon = 1$, utilize the above measure to validate that each of the last 3 groups of results and the original results are 1-*indistinguishable.* (**6 points**)
  8. define another measure and utilize it to justify that the distortion of the 4,000 results for $\epsilon = 1$ is less than that of $\epsilon = 0.5$. (**7 points**)

For each task (8 in total), submit a source code file (can be only a few lines) and a result file, including the quantitative results and measure (if requested). The files are named with the prefix "hw1-II-" (e.g., *hw1-II-1.java, hw1-II-1.txt*).