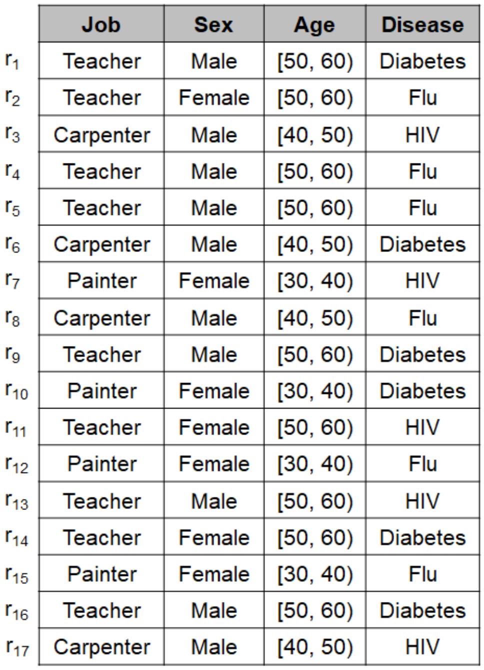
COMP 430/530: Data Privacy and Security - Fall 2020
Homework Assignment # 2

General rules that apply to all questions:

- You must show all your work and calculations. Do not only write the final answers.
- Correct answers without justification will receive little or no credit.
- Assume base 10 logarithms wherever necessary.
- For ℓ-diversity: Values of ℓ may be floats in entropy ℓ-diversity. In the remaining ℓ-diversity definitions, values of ℓ are integers.

Question 1. [19 pts] Consider the following table with QI = {Job, Sex, Age} and SA = Disease. The record IDs r_1, ..., r_{17} are included only so that you can easily refer to them in your answers; they are not actually part of the table.



	Job	Sex	Age	Disease
r_1	Teacher	Male	[50, 60)	Diabetes
r_2	Teacher	Female	[50, 60)	Flu
r_3	Carpenter	Male	[40, 50)	HIV
r_4	Teacher	Male	[50, 60)	Flu
r_5	Teacher	Male	[50, 60)	Flu
r_6	Carpenter	Male	[40, 50)	Diabetes
r_7	Painter	Female	[30, 40)	HIV
r_8	Carpenter	Male	[40, 50)	Flu
r_9	Teacher	Male	[50, 60)	Diabetes
r_{10}	Painter	Female	[30, 40)	Diabetes
r_{11}	Teacher	Female	[50, 60)	HIV
r_{12}	Painter	Female	[30, 40)	Flu
r_{13}	Teacher	Male	[50, 60)	HIV
r_{14}	Teacher	Female	[50, 60)	Diabetes
r_{15}	Painter	Female	[30, 40)	Flu
r_{16}	Teacher	Male	[50, 60)	Diabetes
r_{17}	Carpenter	Male	[40, 50)	HIV

Answer the following questions using this table.

(a) [2 pts] How many equivalence classes (ECs) does the table contain? Write which records belong to which EC.

(b) [2 pts] For each individual EC, the EC is k-anonymous for what largest value of k?

(c) [2 pts] For each individual EC, the EC is distinct ℓ-diverse for what largest value of ℓ?

(d) [2 pts] For each individual EC, the EC is entropy ℓ-diverse for what largest value of ℓ?

(e) [2 pts] For each individual EC, the EC is recursive $(1, \ell)$-diverse for what largest value of ℓ?

(f) [2 pts] For each individual EC, the EC is recursive $(c, 2)$-diverse for what smallest value of c? Treat c as a float.

(g) [2 pts] For each individual EC, the EC is t-close for what smallest value of t, using variational distance as the distance metric?

(h) [2 pts] For each individual EC, the EC is t-close for what smallest value of t, using Kullback-Leibler divergence as the distance metric?

(i) [1 pt] The table, as a whole, is k-anonymous for what value of k?

(j) [1 pt] The table, as a whole, is entropy ℓ-diverse for what value of ℓ?

(k) [1 pt] The table, as a whole, is t-close for what value of t, using variational distance as the distance metric?

Probabilistic ℓ-diversity: We learned in the lectures that there are multiple formulations of ℓ-diversity, such as distinct ℓ-diversity, entropy ℓ-diversity, and recursive (c, ℓ)-diversity. Probabilistic ℓ-diversity is also one possible formulation. It is defined as follows.

Let $p_{(q^*,s)}$ be defined the same way as in entropy ℓ-diversity. An equivalence class (EC) q^* satisfies probabilistic ℓ-diversity if for all s, it holds that: $p_{(q^*,s)} \leq \frac{1}{\ell}$. A table satisfies probabilistic ℓ-diversity if all ECs satisfy probabilistic ℓ-diversity.

Question 2. [12 pts] True or False?

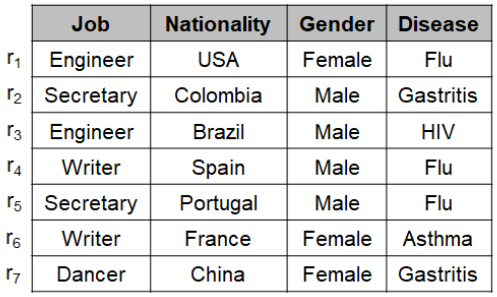
For each of the items below, decide whether they are True or False. If they are true, provide a formal proof (or argue as formally and clearly as you can) that they are true. If they are false, provide a counter-example. Correct answers without justification will get little or no credit.

(a) [4 pts] If an EC satisfies probabilistic ℓ-diversity for some value of ℓ, then it also satisfies distinct ℓ-diversity for the same value of ℓ.

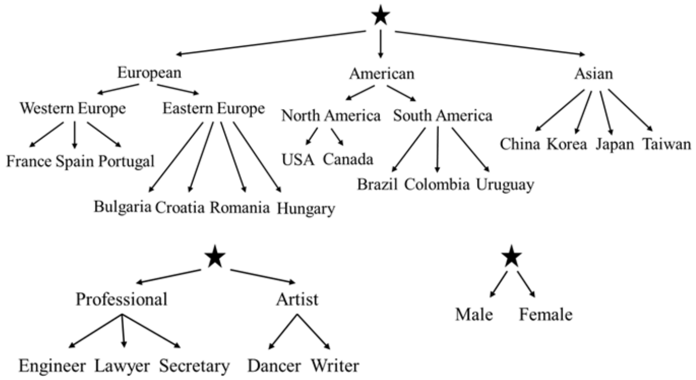
(b) [4 pts] If an EC satisfies distinct ℓ-diversity for some value of ℓ, then it also satisfies probabilistic ℓ-diversity for the same value of ℓ.

(c) [4 pts] Generalization is monotonic with respect to probabilistic ℓ-diversity.

Question 3. [15 pts] Consider the following table with QI = {Job, Nationality, Gender} and SA = Disease. Also consider the DGHs that are given next to the table.



	Job	Nationality	Gender	Disease
r1	Engineer	USA	Female	Flu
r2	Secretary	Colombia	Male	Gastritis
r3	Engineer	Brazil	Male	HIV
r4	Writer	Spain	Male	Flu
r5	Secretary	Portugal	Male	Flu
r6	Writer	France	Female	Asthma
r7	Dancer	China	Female	Gastritis



(a) [5 pts] Using the given DGHs, perform the minimum amount of generalization needed to achieve 2-anonymity such that the first EC contains $\{r_1, r_2, r_3\}$, the second EC contains $\{r_4, r_5\}$, and the third EC contains $\{r_6, r_7\}$.

(b) [5 pts] Calculate the utility loss of the anonymization in part a using the Distortion Metric.

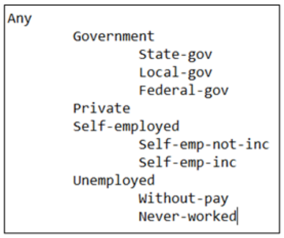
(c) [5 pts] Calculate the utility loss of the anonymization in part a using the Loss Metric (LM). Assume $w_{\text{job}} = w_{\text{nationality}} = 0.4$ and $w_{\text{gender}} = 0.2$.

Question 4. [36 pts] In this question, you will implement a k-anonymization algorithm from scratch, in multiple steps. You can use any programming language you choose (Python is recommended). Submit your source code and a readme text file showing how we can run your code.

Make your code modular according to the steps below. We can give partial credit if we can test each of the below steps independently from the others; but if you do not follow these steps or your code is not modular, we cannot give partial credit.

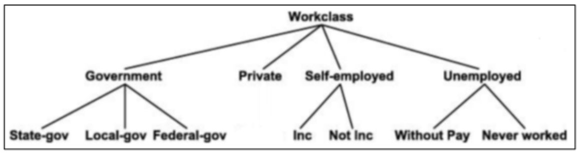
Step 1: Reading inputs [3 pts]. Write two functions *read_data(filename)* and *read_DGHs(directory)*. *read_data* should take the filename as input (file "adult.csv" is provided to you), read and store this dataset in the program's memory. You can assume that the file is always provided in csv (comma separated values) format and the first row of the file contains attribute names.

read_DGHs should take the folder/directory address as input (the folder "DGHs" is provided to you). Each file in this folder is a txt file containing the DGH of one QI attribute (example: "education.txt", "gender.txt"). Your *read_DGHs* function should read and store the DGH of each attribute in the program's memory. Here is an example of how a file represents a visual DGH:



```
Any
	Government
		State-gov
		Local-gov
		Federal-gov
	Private
	Self-employed
		Self-emp-not-inc
		Self-emp-inc
	Unemployed
		Without-pay
		Never-worked
```

DGH in text file



Visual representation of a DGH

Step 2: Data pre-processing [0 pts]. This step is data-dependent, i.e., it will vary from one dataset to another. For the adult.csv dataset that is provided to you, you need to apply the following pre-processing rules:

- Make sure that the dataset is clean. Remove missing values. (Similar to Homework 1)
- Treat the following attributes as QIs = {education, gender, marital-status, native-country, occupation, relationship, workclass}.
- Treat the following attributes as SA = income.
- Remove all other attributes from the dataset.

Step 3: Cost measurement [4+4=8 pts]. Implement two functions to calculate the cost of an anonymized dataset: *cost_MD* and *cost_LM*.

cost_MD(actual_dataset, anonymized_dataset, DGHs): This function takes as input the actual and anonymized datasets, and calculates the cost of anonymization using the Distortion Metric.

cost_LM(actual_dataset, anonymized_dataset, DGHs): This function takes as input the actual and anonymized datasets, and calculates the cost of anonymization using the Loss Metric. You can assume the weights of all attributes are identical, i.e., for N attributes, $w_1 = w_2 = ... = w_N = \frac{1}{N}$.

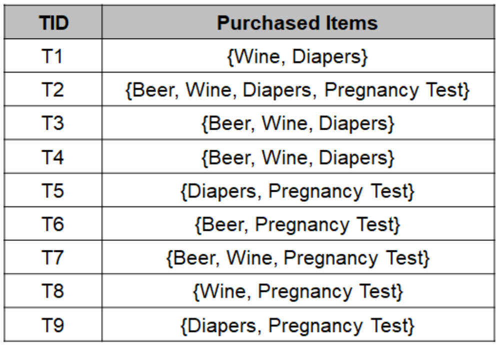
Step 4: Top-down anonymization [15 pts]. Implement a top-down k-anonymization algorithm similar to what we covered in class: *anonymize(actual_dataset, DGHs, k)*.

Assume all records are maximally generalized at the beginning. At each node, split recursively into its children. The split criteria (which attribute specialization to perform) should be the specialization that will yield the highest LM cost improvement. That is: Let D denote the dataset prior to specialization and D_s denote the dataset after specialization s is performed. Then, the specialization you select should be the s^* that maximizes: $s^* = \arg\max_s LM(D) - LM(D_s)$. For tie-breaking, choose specialization s^* alphabetically among those that are tied. If the specialization leads to $\leq k$ records in one or more ECs, then the split should not be performed. When the algorithm terminates, the resulting k-anonymous dataset should be returned.

Step 5: Empirical analysis [10 pts]. Run your top-down anonymization algorithm on the Adult dataset with the given DGHs, for different values of k: $k = 2, 10, 25, 50, 75, 100, 200$. How do MD cost and LM cost values change for varying k? Illustrate with graph(s) and/or table(s). Discuss your observations. Are the results expected or not?

Important: Your program should be generic. It should work for any csv file and any set of DGHs. We may test it with other csv files and DGHs. Do not hardcode data, DGHs, or specializations. Submit your source code and a readme text file showing how we can run your code.

Question 5. [14 pts] Consider the following set-valued dataset and answer parts (a)-(e). Explain and justify your answers. Show all necessary calculations. Correct answers without justification or calculation will get no credit.

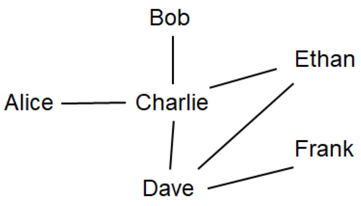


TID	Purchased Items
T1	{Wine, Diapers}
T2	{Beer, Wine, Diapers, Pregnancy Test}
T3	{Beer, Wine, Diapers}
T4	{Beer, Wine, Diapers}
T5	{Diapers, Pregnancy Test}
T6	{Beer, Pregnancy Test}
T7	{Beer, Wine, Pregnancy Test}
T8	{Wine, Pregnancy Test}
T9	{Diapers, Pregnancy Test}

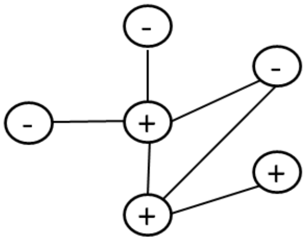
(a) [2 pts] Does the dataset satisfy 2-anonymity?
(b) [3 pts] Does the dataset satisfy 2^2-anonymity?
(c) [3 pts] Does the dataset satisfy 2^3-anonymity?
(d) [3 pts] Does the dataset satisfy 3^2-anonymity?
(e) [3 pts] (This part is not tied to the above dataset.) True or False: Any set-valued dataset that is k-anonymous is also k^m-anonymous for the same value of k and for all $m \geq 1$.

Question 6. [9 pts] MegaCorp has 6 employees: Alice, Bob, Charlie, Dave, Ethan, Frank. Since each employee is aware of the risks of COVID, they keep track of who they have physically interacted with. The employees only know their own physical interactions with others; they do not know whether two other employees have been in physical contact. For example, Alice knows that she has interacted with Charlie but does not know who else Charlie has interacted with.

MegaCorp has actually built a physical contact tracing app which allows MegaCorp to track ALL employees' physical interactions. Here is the full physical interaction graph of MegaCorp:



MegaCorp decides to pseudo-anonymize this graph as follows. They remove all employee names, but they add whether each employee tested COVID positive or not (+ means they tested positive, - means they tested negative). The following is the resulting graph, which MegaCorp shares publicly:



(a) [6 pts] Describe an attack that Charlie and Dave can accomplish together, using just the second graph (public graph) and their knowledge of their own physical interactions. By the end of the attack, Charlie and Dave should learn whether Alice, Bob, Ethan and Frank have tested COVID positive or negative. Note: Charlie and Dave are allowed to communicate among themselves when performing the attack, but they are not allowed to communicate with other employees.

(b) [3 pts] Make the public graph 2-anonymous using a single edge addition.

Best of luck!
ps: I know there are 105 points total. Good for you! ☺