Solution Key for Homework 1

1. k-Anonymity (40 Points)

(a) Here are example hierarchies. Many different hierarchies can be defined.

- Hierarchy for age:
  (a) level 0 is defined as the original values
  (b) level 1 is defined as $0-10$, $10-20$, $\ldots 90-100$
  (c) level 2 is defined as $0-20$, $20-40$, $\ldots 80-100$
  (d) level 3 is defined as $1-50$ and $50-100$ (notice that, in the range of $40-60$ in the previous level, values in $40-50$ will be generalized to $1-50$ while values in $50-60$ will be generalized to $50-100$)

  Each value can be generalized up to 4 levels.

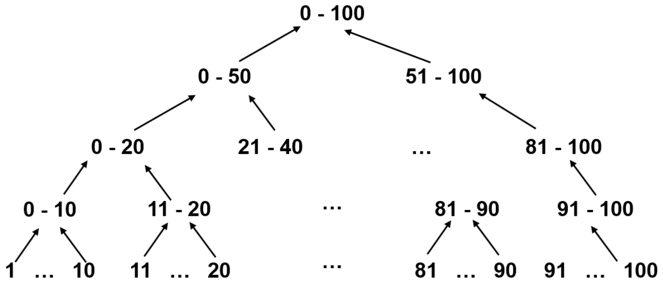

Figure 1: Age Hierarchy

- Hierarchy for education: each value can be generalized up to 3 levels (categorized by general education levels).

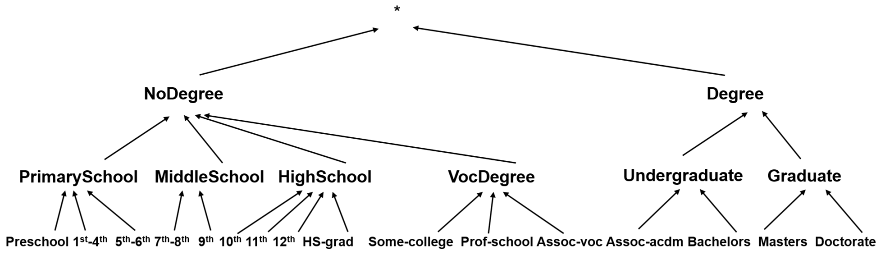

Figure 2: Education Hierarchy

- Hierarchy for marital status: each value can be generalized up to 3 levels (categorized by general marital status).

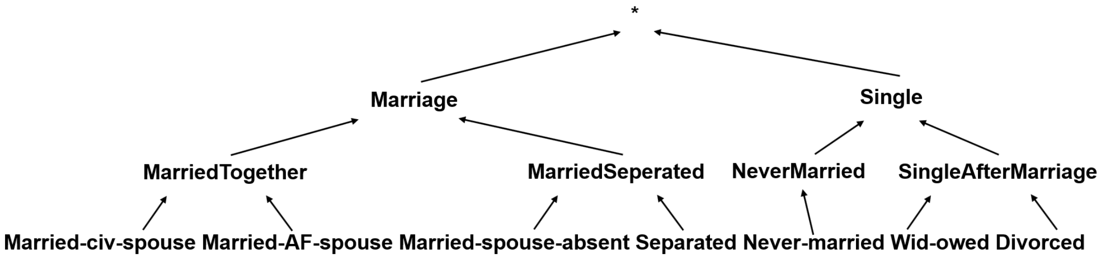

Figure 3: Marital Status Hierarchy

- Hierarchy for race: each value can be generalized up to 2 levels (white and black to "10%+" while other races to "< 10%").

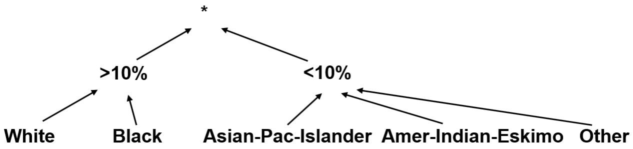

Figure 4: Race Hierarchy

(b) Write a program for the heuristic algorithm (which generalizes/suppresses the data for k-anonymity while minimizing the utility loss).

It has been proven that finding an optimal anonymization solution is an NP-hard problem. Thus, a heuristic algorithm is expected here. Adults with a salary $\leq 50K$ satisfy 10-anonymity while adults with salary $> 50K$ are ok with 5-anonymity. You can design an algorithm that pursues (k_1, k_2)-anonymity for the entire dataset. Or the algorithm can split the dataset to two sub-datasets and satisfy k_1 and k_2 for two sub-datasets, respectively. If mixing adults from two categories into the same equivalence class, $k = 10$ should be satisfied for all the adults in the same equivalence class. This may overly generalize the data. Then, we split the dataset into two partitions, and apply k_1 and k_2, respectively. Here, a representative solution is given in Algorithm 1.

(c) Calculate the distortion and precision of the output $(k_1 = 10, k_2 = 5)$.

The distortion and precision can be different for different algorithms, depending on how hierarchies are constructed. If your hierarchy and algorithm are reasonable, the distortion should not be very large. Otherwise, for instance, if your hierarchy only includes 2 levels for all the attributes, the distortion might be very large.

2. Differential Privacy

(d) calculate the distortion and precision of the output. Note that the distortion and precision results can be different for different algorithms.

Algorithm 1: (k_1, k_2)-Anonymity Algorithm

```
Input  : input dataset $D$
         parameter $(k_1, k_2)$
         quasi-identifier $QI = (A_1, ..., A_4)$
         hierarchies $DGH_{A_i}$ where $i = 1, ..., 4$
Output: dataset $D^*$ satisfying the defined $(k_1, k_2)$-anonymity
split input dataset $D$ into $D_1$ for adults with salary 50K, and $D_2$ for adults with salary 50K
initialize the list SUP=[]
foreach $D_n$, $n \in \{1, 2\}$ do
    freq ← a frequency list containing distinct sequences of values of $D[QI]$, along with the
      count of each sequence
    while exists sequences in freq with count less than $k_n$ and the total count of this
      sequence is more than $k_n$ do
        let $A_j$ be the attribute in freq with the most number of unique values
        freq ← generalized values of $A_j$ in freq
    SUP ← tuples with count less than $k_n$
    suppress tuples in freq with count less than $k_n$
if length of SUP > $k_1$ then
    generalize values for SUP to satisfy $k_1$-anonymity for tuples in SUP
else
    suppress tuples in SUP
Return $D^*$ ← construct table from $freq_n, n \in \{1, 2\}$ and SUP
```