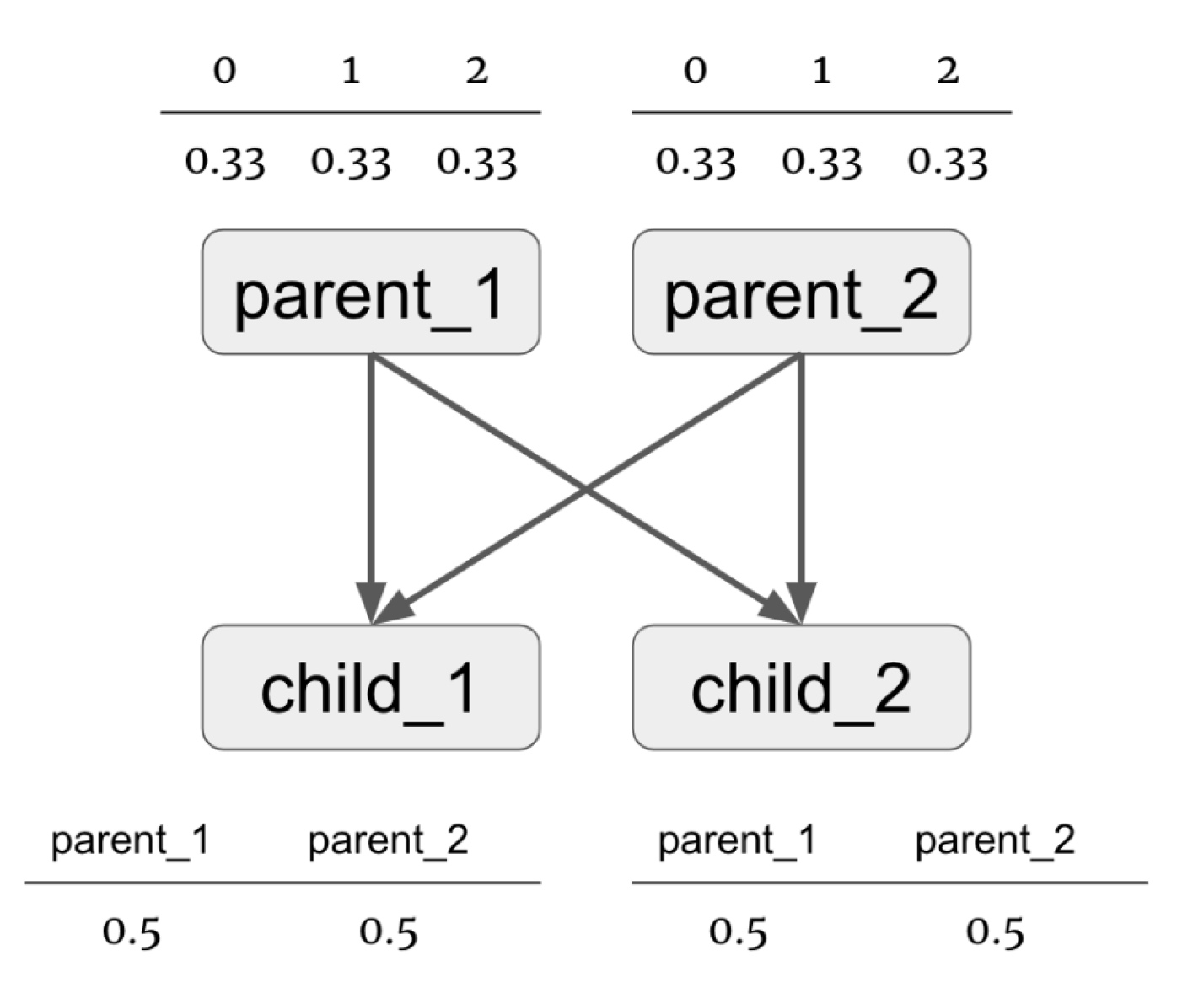
In this problem, you will take on the role of a data owner, who owns two sensitive datasets, called **hw\_compas** and **hw\_fake**, and is preparing to release differentially private synthetic versions of these datasets.

The first dataset, **hw\_compas** is a subset of the dataset released by ProPublica as part of their [COMPAS investigation](https://propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing). The **hw\_compas** dataset has attributes age, sex, score, and race, with the following domains of values: age is an integer between 18 and 96, sex is one of ‘Male’ or ‘Female’, score is an integer between -1 and 10, race is one of 'Other', 'Caucasian', 'African-American', 'Hispanic', 'Asian', 'Native American'.

The second dataset, **hw\_fake**, is a synthetically generated dataset. We call this dataset “fake” rather than “synthetic” because you will be using it as *input* to a privacy-preserving data generator. We will use the term “synthetic” to refer to privacy-preserving datasets that are produced as *output* of a data generator.

We generated the **hw\_fake** dataset by sampling from the following Bayesian network:



In this Bayesian network, **parent\_1**, **parent\_2**, **child\_1**, and **child\_2** are random variables. Each of these variables takes on one of three values {0, 1, 2}.

* Variables **parent\_1** and **parent\_2** take on each of the possible values with an equal probability. Values are assigned to these random variables independently.
* Variables **child\_1** and **child\_2** take on the value of one of their parents. Which parent’s value the child takes on is chosen with an equal probability.

To start, use the [Data Synthesizer library](https://github.com/DataResponsibly/DataSynthesizer) to generate 4 synthetic datasets for each sensitive dataset **hw\_compas** and **hw\_fake** (8 synthetic datasets in total), each of size N=10,000, using the following settings:

* A: random mode
* B: independent attribute mode with **epsilon = 0.1**.
* C: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=1**
* D: correlated attribute mode with **epsilon = 0.1**, with Bayesian network degree **k=2**

For guidance, you can use the [HW2\_Template](https://drive.google.com/file/d/1xXm9bWKSYnBxR5VEJyyb4TN0qyZ-mFuC/view?usp=sharing) here. Please make sure to duplicate this file rather than put your code directly here.

**(a)** Execute the following queries on synthetic datasets and compare their results to those on the corresponding real datasets:

* **Q1** (**hw\_compas** only): Execute basic statistical queries over synthetic datasets.

The **hw\_compas** has numerical attributes **age** and **score.** Calculate **Median, Mean, Min, Max** of **age** and **score** for the synthetic datasets generated with settings A, B, C, and D (described above). Compare to the ground truth values, as computed over **hw\_compas**. Present results in a **table**. Discuss the accuracy of the different methods in your report. Which methods are accurate and which are less accurate? If there are substantial differences in accuracy between methods - explain these differences.

* **Q2 (hw\_compas** only): Compare how well random mode (A) and independent attribute mode (B) replicate the original distribution.

Plot the distributions of values of **age** and **sex** attributes in **hw\_compas** and in synthetic datasets generated under settings A and B. Compare the **histograms** visually and explain the results in your report.

Next, compute cumulative measures that quantify the difference between the probability distributions over age and sex in **hw\_compas** vs. in privacy-preserving synthetic data. To do so, use the Two-sample Kolmogorov-Smirnov test (KS test) for the numerical attribute and Kullback-Leibler divergence (KL-divergence) for the categorical attribute, using provided functions **ks\_test** and **kl\_test**. Discuss the relative difference in performance under A and B in your report.

For Two-sample Kolmogorov-Smirnov test and Kullback-Leibler divergence, you might find functions such as ‘*entropy*’ and *‘ks\_2samp*’ from *scipy.stats* useful.

* **Q3** (**hw\_fake** only): Compare the accuracy of correlated attribute mode with k=1 (C) and with k=2 (D).

Display the pairwise mutual information matrix by heatmaps, showing mutual information between all pairs of attributes, in **hw\_fake** and in two synthetic datasets (generated under C and D). Discuss your observations, noting how well / how badly mutual information is preserved in synthetic data.

To compute mutual information, you can use functions from <https://github.com/DataResponsibly/DataSynthesizer/blob/master/DataSynthesizer/lib/utils.py>

For heatmaps, we suggest considering functions (*heatmap*) provided in the seaborn library (see example: <https://seaborn.pydata.org/examples/many_pairwise_correlations.html>) and remember to set up *vmax* and *vmin* when plotting.

**(b) (hw\_compas** only**)**: Study the variability in accuracy of answers to Q1 under part (a) for A, B, and C for attribute **age**.

To do this, fix epsilon = 0.1, generate 10 synthetic databases (by specifying different seeds) for each setting A, B, and C. Calculate the **mean** and **median** of age in each of the generated datasets. Then, for each setting, plot the 10 **median** values and the 10 **mean** values using a **box-and-whiskers** plot. Compare these metrics to the ground truth median and mean from the real data. Carefully explain your observations: which mode gives more accurate results and why? In which cases do we see more or less variability?

Specifically, you should generate 30 datasets in total: 10 under setting A, 10 under setting B, 10 under setting C. For the box-and-whiskers plots, we expect to see two subplots: one for each of the **median** and **mean** withthe three parameter settings (A, B and C) along the X-axis and age on the Y-axis. You should include these plots in your report.

**(c) (hw\_compas** only**)**: Study how well statistical properties of the data are preserved as a function of the privacy budget, epsilon. To see robust results, execute your experiment with 10 different synthetic datasets (with different seeds) for each value of epsilon, for each data generation setting (B, C, and D). Compute the following metrics, visualize results as appropriate with box-and-whiskers plots, and discuss your findings in the report.

* For each setting (B, C, and D), vary epsilon from 0.02 to 0.1 in increments of 0.02. Specifically, the epsilons are [0.02, 0.04, 0.06, 0.08, 1]. In total, you should generate 150 synthetic datasets (3\*10\*5) and calculate the KL-divergence for race in each dataset. Create three box-and-whiskers plots, one for each setting (B, C, D). Each plot should have epsilon on the X-axis and KL-divergence on the Y-axis. Discuss your findings in the report and include your plots.