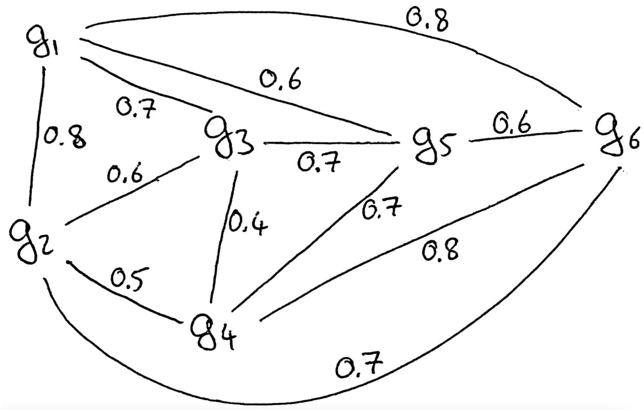
BCB 570 Homework Network Inference **120 points** **Due Wednesday, Mar 2 at 11:59pm**

Bagging: Bootstrap Aggregating

1. (5 pts) Describe the bagging process in your own words.
2. (5 pts) Use R or Python to generate a sample of 100 instances of a normally distributed random variable with a mean of 2 and a variance of 8. Estimate the mean and variance of the data from the entire sample of 100.
3. (10 pts) Write a program that subsamples the data (with replacement) into 20 groups of 10 and estimates the mean and variance for each subsample. Plot histograms of the sample mean and variance for each subsample. Use the histograms to estimate the sample parameters.

Data Processing Inequality (DPI)

1. (10 pts) Describe how the DPI is used to simplify mutual information networks and discuss possible issues. Hint: The ARACNE paper may prove helpful.
2. (5 pts) The graph below shows six genes ($g_1, \ldots, g_6$) and all pairwise mutual information values (edges) above a threshold of 0.38. Draw the simplified network after applying the DPI.



Network inference challenge

1. (15 pts) Describe how the five methods WGCNA, TIGRESS, GENIE3, ARACNE, CLR find associations between genes.
2. Pick two of the methods we discussed in class (WGCNA, TIGRESS, GENIE3, ARACNE/CLR), or for extra credit pick a new hybrid method from the recent literature (2014 or later). Note: You cannot pick both ARACNE AND CLR as they are too similar. Use the two methods you picked to infer the genetic regulatory network using
   (a) (30 pts) the five in silico multifactorial datasets from the DREAM4 Challenge (provided in Canvas). Use the provided ground truth data to generate precision recall curves as well as ROC curves averaged across the five datasets to measure the performance of your method.
   (b) (10 pts) the E.Coli blind data set (provided in Canvas). Here, you may only pick one method. Pick whichever method you believe will perform best based on what you learned in (a). We will assess the performance using the CURRENT version of RegulonDB.
   Details: Submit a ranked list of regulatory link predictions ordered according to the confidence you assign to the predictions, from the most reliable (first row) to the least reliable (last row) prediction. Use a 3 tab-separated column format as in the example below:

```
A \tab B \tab XYZ
```

where A and B are two different genes (no self-interactions). Links are directed: the gene in the first column regulates the gene in the second column. (If both A regulates B and B regulates A, then both lines should be included.) XYZ is a score between 0 and 1 that indicates the confidence level you assign to the prediction. (E.g., XYZ = 1 if gene A is deemed to regulate gene B with highest confidence and XYZ = 0 if A is deemed not to directly regulate B). All pairs omitted from the list will be considered to appear randomly ordered at the end of the list. Save the file as text, and name it: YourName.txt

(c) Extra credit: Download the current version of the RegulonDB TF-gene interactions and assess the performance of your prediction.

3. (10 pts) Design an algorithm with a weighted voting scheme that uses the "wisdom of crowds" that we can implement. See Marbach's 2012 Nature Methods paper for details on the "wisdom of crowds". We will use this algorithm after the main assignment has been turned in to create an ensemble estimate of the different blind data set results to see if the community prediction performs better than our individual predictions.

Network inference using single-cell data

1. (0 pts) Read Murali's 2020 Nature Methods paper and the 2021 Network Inference review (pdf posted on Canvas).

2. (10 pts) Summarize the new approach taken by the 2020 Nature Methods paper to evaluate network inference methods? Why was a new approach needed?

3. (10 pts) What are, in your opinion, two of the biggest obstacles in network inference today? Name at least two and discuss ideas how to overcome these issues.
   Note: This is an open-ended question. I am aware that you will (likely) not solve this problems as part of this assignment ;)