67978: A Needle in a Data Haystack
Introduction to Data Science
Homework #2: Similar items, Clustering, Community Detection
Due: **10 Dec, 11:59pm, on Moodle**

Important notes: Coding question should be done in pairs. The other questions should be done **individually**.
Read the instructions **carefully** (that's a good idea in general).

- There will be two submission links (for theoretical and practical parts).
- The theoretical part should be a single zip file (**theoretical_ex2_ID.zip**) containing a single file in pdf format only (no docx or jpg) named **theoretical_ex2_ID.pdf** (ID is your ID). If you are submitting handwritten answers, make sure they are crystal clear.
- Only **one of each pair** needs to submit the practical part code and answers. The other person should only write the id of the one submitting in the pdf.
- The one submitting the practical part (using the second link) should submit a zip file named **practical_ex2_ID1_ID2.zip** (where ID1 your ID number and ID2 is your partner's ID number). The zip should contain a folder named **code** and a pdf for the results, named **practical_ex2_ID1_ID2.pdf**.
- Points may be reduced for submissions which fail to comply.

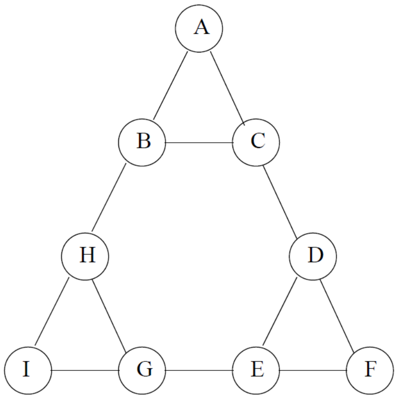
Problem 1 (Finding Similar Items).

(a) What are the first ten 4-shingles of the header (the line that starts with 67978 above)? Use shingle = character, including spaces and punctuation.

(b) Prove or disprove: if the Jaccard similarity of two columns is 0, then minhashing always gives a correct estimate of the Jaccard similarity.

(c) One might expect that we could estimate the Jaccard similarity of columns without using all possible permutations of rows. For example, we could only allow cyclic permutations; start at a randomly chosen row r, which becomes the first in the order, followed by rows $r + 1$, $r + 2$, and so on, down to the last row, and then continuing with the first row, second row, and so on, down to row $r - 1$. There are only n such permutations if there are n rows. However, these permutations are not sufficient to estimate the Jaccard similarity correctly. Give an example of a two-column matrix where averaging over all the cyclic permutations does not give the Jaccard similarity. Compute Jaccard and average similarity.

Problem 2 (Communities).



(a) Use the Girvan-Newman approach to find the number of shortest paths from each of the following nodes that pass through each of the edges. (1) Node A (2) Node B.

(b) Using symmetry, these are all the calculations you need to compute the betweenness of each edge. Do the calculation.

(c) Using your results for (b), pick a threshold and remove the edges with higher betweenness. What is the threshold? What are the communities?

(d) Compute the Laplacian matrix for this graph, find the second-smallest eigenvalue and its eigenvector. What communities does it suggest?

Problem 3 (Clustering Part I (Coding question)).

Synthetic data is information that's artificially manufactured. Synthetic data is created algorithmically, and it is used as a stand-in for test datasets, to validate models and, increasingly, to train machine learning models. It is very useful for testing clustering methods.

Your goal is to generate synthetic data. For each task, generate 300 random points (in $\mathbb{R}^2$. That is, 2D) and plot them. Repeat this **TWICE** (so two plots, 300 points each for each task).

(a) Uniform distribution, $x \in [-1, 1], y \in [0, 5]$

(b) Gaussian with center at [1,1] and std=2.

(c) Three Gaussians with centers at $[i, -i]$ and std=$2i$ $(i = 1, 2, 3)$.

(d) A circle inside a ring (like the CURE slides).

(e) The first letter of both your first names (so the data should look like a noisy version of "AD", for your own letters; if they are the same letter, use last names. :)).

Problem 4 (Clustering Part II (Coding Question)).

You are given ex1.json, the crawling result from HW1 (thanks Alon and Franzisk!). Your goal is to perform clustering. You do not need to implement the clustering algorithms themselves, and you can use packages to parse the json (you might have to fix some small issues with the file, that's ok).

Perform a hierarchical clustering of the one-dimensional set of points defined by the *DolarsPelged* field for items with currency="$" (ignore the rest). Assume clusters are represented by their centroid (average).

(a) At each step merge the clusters with the closest centroids.

(b) At each step merge the two clusters whose resulting cluster has the smallest diameter.

For each clustering above,

(i) Plot the distances throughout the algorithm (x axis: step number, y axis: distance between the clusters merged in that step).

(ii) Let m be the final distance (last merging step). How many clusters would you have for threshold $0.1m$? $0.5m$? (That is, is you stopped merging when the distance was $\geq$ treshold)

Problem 5 (Meta). How long (in hours) did this assignment take? Please answer in the **comments** section of the moodle link, not in the pdf.