CDS DS 210 — Programming Data Science

Name: Emmy Blumenthal | Final Project Proposal | **BU ID:** U87312711
Due Date: Nov 11, 2022 | **Email:** emmyb320@bu.edu

Clustering, partitioning, and archetypal representation of embryonic mouse brain cells

The structure and differentiation of different cell types is a central question in biology and biophysics that has a history of combining epistemological questions (e.g., what is a cell type?) and quantitative methods. In this project, I will employ graph clustering methods to try to approach some of these questions from a basic level. I will investigate the 'megascale cell-cell similarity network' from the Stanford network analysis project (`https://snap.stanford.edu/biodata/datasets/10023/10023-CC-Neuron.html`). Specifically, I will attempt to use a variety of graph clustering algorithms to find subgraphs/cliques of the similarity network which represent cells that are more similar to each other than other clusters of cells; an example algorithm I will implement is the highly connected subgraphs (HCS) clustering algorithm (`https://en.wikipedia.org/wiki/HCS_clustering_algorithm`). The relative success of the algorithm at finding independent clusters may tell me something about cell types. If clusters are sufficiently independent and there are many clusters, this tells us that cell types may be well-differentiated. On the other hand, if there are few clusters and/or it is difficult to find independent cliques/clusters, I might be able to conclude that cell types are difficult to differentiate using this method.

If I am able to find sufficiently differentiated clusters, a next possible step would be implementing vertex centrality measures which may tell us which cells can serve as archetypes/representatives for their given cell type. These centrality measures would be implemented within the individual cell sub-graphs. Because individual cell's gene expressions often fluctuate significantly, it is likely that finding archetypal example cells may be difficult/infeasible. However, understanding the way in which various centrality methods may succeed or fail to identify only a few archetypes for each subgraph may provide some insights into how cell types could be defined and their structures of differentiation. Possible centrality measures that I may employ include closeness and eigenvector centrality.

To accomplish this project, I will attempt to adhere to the following schedule:

- Week of Nov. 18: Download data and ensure that I can read data in the necessary formats (e.g., adjacency matrix, sparse representation)
- Week of Nov. 25: Give a first pass at implementing the HCS clustering algorithm; identify questions and errors that will need to be tackled after Thanksgiving break.
- Week of Dec. 2: Polish HCS clustering algorithm and seek support for persistent issues. Assess feasibility of implementing sub-graph centrality measures.
- Week of Dec. 9: If implementing sub-graph centrality measures is feasible given implemented HCS algorithm, implement two centrality measures and assess results relative to each other. If implementing sub-graph centrality measures is unsuccessful, attempt to implement a different sub-graph/clique clustering algorithm besides HCS.
- Week of Dec. 15: Finalize report on results; clean repository and source code.