Homework 7

For each of the following questions, assess whether the proposed approach is well suited to the problem. If it is, explain why it is better than some alternative. If it is not, propose a more reasonable approach. **You should answer all 3 of the questions**. You don’t need to write more than a short paragraph for each question. Your 3 answers together must be less than one page, single spaced.

Question 1:

Ammon has single cell transcriptome data from 100s of neurons, and he would like to use them to infer relationships between neuron types that he suspects are developmentally related. He produces a graph where each node has the expression level of various transcription factors known to be involved in development, and the edge weights defined by the absolute differences in expression, weighted by the distance between the brain regions of the cells. He wants to describe the relationship between any two cells as the most likely path between them. His idea is to produce a Minimal Spanning Tree (MST) of the entire graph, and then describe the relationship between any two cells as the connecting path across the MST. Assuming that his original graph is well constructed, is MST a good approach for defining these pair-wise relationships?

Question 2:

Yongjun is interested in how disrupting circadian rhythm affects language production. He has a series of recordings from subjects relating the plots of well know movies following varying amounts of sleep deprivation, and he wants to use graph theory approaches to analyze word choice. He wants to design his feature space such that he can characterize subject speech by measuring path length and average degree of chosen nodes (his thought is they will indicate some sense of the consistency of word choice within and across subjects). His idea is to define a large feature space of language such that each node is a single recording of a word (subject 1 “cat” would be a different node from subject 2, “cat”) and the edge weights are defined by acoustic similarity (defined by the cross correlation of the recordings). Given his question and proposed analysis, is this a good approach for defining the feature space and edge weights?

Question 3:

In order to assess claims of a lack of diversity\* at Penn, Junhyong has been given access to facebook’s data on faculty at Penn, specifically anonymous profile information of identified interests\*\*, without friend’s lists (i.e., what’s listed here: <https://www.facebook.com/ads/preferences/edit/>). This provides a high dimensional space (where each point is an individual and each dimension is the presence/absence of some interest) He wants to use tSNE to perform a dimensional reduction on these data so that he can visualize Penn faculty and also perform a clustering algorithm to assess the diversity of background found at Penn. Given his dataset and his question, is tSNE a good approach here?

\* - We're not talking about racial/gender diversity only. You might wonder whether Penn has diverse socio-economic background, or diverse politics, or diversity of philosophy, or diversity of interests, etc. These sorts of questions are harder to assess based on standard census metrics.

\*\* - If you're unfamiliar with facebook's data, it's pretty comprehensive. Mine includes assumed interests such as Indie Rock, Video Games, FactCheck.org, Reason, and identified categories such as "Birthday in February, Away from hometown, US politics (moderate), Frequent Visitors: Pizza Places, Parks". It would be a pretty cool data set to get access to.