**BIODS 253 Final Project Design Document – Brandon Bergsneider**

**Overview**

I aim to implement a software pipeline for performing network analysis and unsupervised clustering on patient-reported outcome (PRO) data. These methods can be used to better understand and model symptom co-occurrence and co-severity patterns in PRO data, informing precision-health based symptom care.

**Background**

Last year, I worked as a research fellow at the National Institutes of Health (NIH) Neuro-Oncology Branch, and my research focused on using computational network analysis and unsupervised clustering methods to model symptom co-occurrence and co-severity patterns in cancer patients and survivors. The goal of this research is to better understand heterogeneity in cancer patients’ symptom experiences and inform more personalized and effective symptom care strategies. We published [our analysis](https://doi.org/10.1093/noajnl/vdac188) on a cohort of 1128 primary brain tumor patients, and we are currently collaborating with a group at UCSF to conduct the same analysis on PRO data from over 3000 cancer patients.

So far, all the code I have written for these analyses are in relatively disorganized .Rmd files that are not easily reproducible (the original scripts can be viewed [here](https://drive.google.com/drive/folders/1r5XDUa_xDLxOCfzYX6Pe9Dq7ROywNc48?usp=sharing)). The goal of my BIODS 253 final project is to organize my analysis pipeline into a GitHub repository that can be easily understood and reproduced. Many aspects of the project can be improved with better software engineering techniques including better file organization, version control, use of functions for repetitive code, code annotation, and testing functions. I hope to be able to share this GitHub repository with my mentors and other fellows at the NIH who will continue my work. Others will make use of this code in the near future, as there are other patient-reported outcomes (PROs) datasets that our group has recently gained access to and plans to conduct network analysis on. Additionally, a long-term goal of our group is to share this methodology through a user-friendly website that would enable non-computationalists to conduct network analysis on PRO data, and this GitHub repository is a first step towards making the code reproducible enough to implement into a web-based platform.

**Current Goals:**

I aim to create an R package that allows users to easily implement the following analyses given a .csv file of patient-reported symptom data:

1. Identify symptoms that are redundant (aka variables that represent the same underlying construct) and consolidate them into latent variables using the Unique Variable Analysis method described in [Christensen, Garrido, & Golino (PsyArXiv)](https://psyarxiv.com/4kra2/)
2. Construct a Gaussian Graphical Model (GGM) network for the entire cohort and calculate network centrality measures using the method described in [Epskamp, Borsboom, & Fried (2018)](https://link.springer.com/article/10.3758/s13428-017-0862-1)
3. (Time permitting) Conduct second-order unsupervised clustering to stratify patients into subgroups based on their symptom co-severity or co-occurrence patterns, using a modification of the method described in [Henry et al. ( 2018)](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0191981)

**Non-Goals:**

For this course, I am not aiming to implement all parts of the network analysis and unsupervised clustering pipeline. Rather, my goal is to develop an R package and GitHub repository that contains the core aspects of the analysis and will serve as a foundation for me to expand upon and fully flesh out in the coming months. Given short time frame of this project and the limits on my persona schedule as a first-year medical student, I do not think it is reasonable to try to implement the entire analysis pipeline for the project, but I will at least attempt to create a foundation that I can build upon.

**Future Goals:**

In the future, I aim to fully implement the entire analysis pipeline into a package that others can use on their own data. This package will be used by others in the NIH Neuro-Oncology branch to analyze new datasets that we recently gained access to. Additionally, our group aims to implement these methods into a user-friendly website that non-computationalists can use to conduct network analysis and unsupervised clustering on patient-reported symptom data, and I hope that this package and GitHub repository will serve as the foundation for such a web platform.

**Detailed Design:**

I will create an R package that implements various functions on a .csv file of patient-reported symptom data. The .csv file will need to be formatted in the following way:

* Each row should represent a patient.
* The first column must be titled ‘ID’ and it must contain unique ID numbers for each patient.
* The remaining columns should represent symptom severity measurements. The title of each column should be the name of the symptom, and the column should contain numerical ratings of symptom severity.

I will implement the following functions:

* plot\_severity: Takes the symptom .csv file and plots a bar chart of the severity of each symptom in the dataset. This is helpful in determining which symptoms are at the extremes of the distribution and may be redundant.
* plot\_occurrence: Takes the symptom .csv file and plots a bar chart of the percent occurrence of each symptom in the dataset. This is also helpful in determining which symptoms may be redundant.
* perform\_uva: Takes the .csv file and performs Unique Variable Analysis (UVA) on the data, allowing users to see which variables are redundant and manually choose which variables to consolidate. This function will plot a bar chart of the UVA results and output a new .csv file with redundant variables consolidated using the Maximum Likelihood with Robust standard errors estimator.
* construct\_network: Takes the .csv file, constructs a Gaussian Graphical Model based on the patient symptom data, and discovers symptom clusters. This function will output a plot of the symptom network and plots of the network centrality and bridge centrality measures. Additionally, it will provide users the option to output .csv files containing edge weights and centrality measure calculations.

Backend functions will include:

* check\_csv\_file: A function that checks whether the .csv file is formatted properly (described above). If not, it will return an error message. This will be called by all of the functions that take in a .csv file.

To test these functions, I will be using a publicly available dataset on PTSD symptoms (unfortunately, I am not allowed to share the cancer patient symptom data that my group is working on due to IRB restrictions). The data can be found at this link: <https://datashare.nida.nih.gov/study/nida-ctn-0015>. To access the data, click on “DTN-0015 Data Files,” fill out the required data sharing agreement, and navigate to the file titled “qs.csv.” This data is in long format and contains a column with subject id’s, a column with the names of the administered items, and a third column containing item responses. Instructions for reformatting the data into the format described above will be provided in the ReadMe file. Additionally, I will include the reformatted data into the R package available on Github.

**Work Estimates:**

I expect to spend 20-30 hours on this project in total, with the following rough time estimates:

* 2 hrs: Write design document.
* 2 hrs: Set up Github repository and R package files. Download and reformat the data.
* 10-20 hrs: Write required functions and .R files.
* 2 hrs: Write tests for the R package.
* 2 hrs: Write ReadMe file.