Thesis Problem Statement

Maximiliano Martino

March 15, 2025

Problem Definition

Our objective is to determine whether two or more brain networks, represented as graphs, belong to the same population or not, with the additional aim of identifying specific connections (nodes or edges) responsible for any detected differences. Each brain is modeled as an undirected graph $G = (V, E)$, where V represents fixed brain regions (nodes), and E represents the connections (edges) between them, which may vary among individuals.

Let $\mathcal{G}_A = \{G_1^A, G_2^A, \ldots, G_{N_A}^A\}$ be a collection of graphs corresponding to a control population (normal brains), $\mathcal{G}_B = \{G_1^B, G_2^B, \ldots, G_{N_B}^B\}$ a population with a disease like schizophrenia, and $\mathcal{G}_C = \{G_1^C, G_2^C, \ldots, G_{N_C}^C\}$ a population with Alzheimer's. All three populations have the same number of nodes $|V|$, but their edges E differ due to variations in brain connectivity.

For each population, we define a central or representative graph, denoted M_A for the control population, M_B for the schizophrenia group, and M_C for the Alzheimer's group. The representative graph M_A can be obtained by averaging the adjacency matrices of all graphs in $\mathcal{G}_A$, and similarly for M_B and M_C.

To assess if a given graph G_i^A belongs to its respective population, we define the distance $d(G_i^A, M_A)$ as a measure of how far this graph is from the central graph of its group. Similarly, we can calculate the distance between each graph of a population and the central graph of another population, such as $d(G_i^A, M_B)$, $d(G_i^A, M_C)$, $d(G_i^B, M_A)$, and so on.

Hypothesis Test

The hypothesis test, previously developed by other researchers, assesses whether the distributions of distances $d(G, M_A)$, $d(G, M_B)$, and $d(G, M_C)$ are significantly different across populations, allowing us to infer if the graphs originate from the same population or not. This test is based on the following hypotheses:

- **Null Hypothesis H_0:** Brain networks from all populations (control, schizophrenia, and Alzheimer's) come from the same distribution, meaning

there are no significant differences between the three groups. Mathematically, this is expressed as:

$$H_0 : \mathcal{M}_1 = \mathcal{M}_2 = \cdots = \mathcal{M}_m$$

- **Alternative Hypothesis H_A:** At least one of the brain populations differs from the others:

$$H_A : \text{At least one } \mathcal{M}_i \neq \mathcal{M}_j$$

The proposed test statistic for the null hypothesis is given by:

$$T = \frac{\sqrt{m}}{a} \sum_{i=1}^{m} \sqrt{n_i} \left(\frac{n_i}{n_i - 1} \overline{d_{G_i}(M_i)} - \frac{n}{n-1} \overline{d_G(M_i)} \right)$$

where the variables involved are:

- m: The number of populations (e.g., control, schizophrenia, and Alzheimer's).
- n_i: The number of graphs (brains) in population i.
- n: The total number of graphs across all populations.
- $\overline{d_{G_i}(M_i)}$: The average distance of graphs in population i to their corresponding central graph M_i.
- $\overline{d_G(M_i)}$: The total average distance of all graphs to the central graph of population M_i.
- a: A normalization constant chosen based on the specific form of the test statistic.

The test statistic T measures the overall deviation of distances $d(G_i, M_i)$ from their expected values under the null hypothesis. A highly negative T suggests that the graphs differ significantly, leading us to reject the null hypothesis in favor of the alternative hypothesis, indicating that brain networks from the compared populations do not originate from the same distribution.

Objective

Building on this detection test, our work focuses on *identifying specific nodes or connections* responsible for differences between networks across populations. Our hypothesis is that critical nodes exist whose removal or modification would reduce the observed distance $d(G_i, M)$, bringing the networks closer together.

Our approach involves two potential algorithms:

1. **Node Removal and Re-test:** Iteratively remove nodes and recalculate distances to identify critical nodes contributing to the difference.

2. **Edge Removal and Re-test:** Remove edges between nodes to find significant structural differences.

These approaches should work well when there are large differences between groups. It is important to explore under what probability laws or graph sizes this method might not be effective. For instance, we need to investigate cases where dependencies exist between certain groups of edges ("if edge k appears, then edge l always appears").

We will develop an algorithm, implemented in `R`, to efficiently identify critical nodes and/or edges responsible for population-level differences in brain networks.

Exploratory Conditions

1. **Non-informative Edges:** Given the potentially large number of edges, we might consider eliminating edges that never appear. That is, we retain only positive edges.

2. **Separability of T:** We will examine whether the statistic can be rewritten recursively. If T is separable, i.e., if T of 3 edges can be decomposed into T of 2 of those edges plus a contribution from the third, this could significantly reduce computational complexity. Further analysis is required to confirm if this is applicable.

Edge Addition Approach

Idea: Suppose we have only two groups; to check if they differ, it is natural to base this on the frequency of the edge between node k and node l in graphs from group A, and then compare it to occurrences in graphs from group B:

1. **Calculation of p-values:** For each edge, we calculate a p-value using an ANOVA test for proportions. This approach focuses on adding edges instead of removing them.

2. **Edge Addition Based on p-values:** The approach begins by calculating T considering only the edge with the lowest p-value. This edge is added to the set of edges to be discovered. Then the next edge with the second-lowest p-value is added, and so on, until the increase in T is minimal. We will investigate how to define the criterion for a change in slope.

Challenges with Subtle Differences

- **False Positives:** When group differences are subtle, large groups of edges may introduce noise. Some edges might show low p-values by chance and could be incorrectly selected by the algorithm. This issue must be addressed to ensure robustness in detecting critical edges.

Input and Output

Input: Adjacency matrices for each graph and their group labels, or an array for each group containing the matrices.
Output: An adjacency matrix with 1 in the edges identified by the algorithm as responsible for differences between populations.