Two sample test for two network real data

Maoyu Zhang[1]

1 Introduction

In recent years, networks have been increasingly applied in many different fields of scientific research, ranging from microscopic networks such as protein-protein interaction networks, gene regulatory networks or brain networks, to macroscopic networks such as social networks, organizational networks, mobility and transport networks.Therefore,we focus on this specific type data, namely networks. A network $G = (V, E)$ is a complex combinatorial object composed of a set V of nodes that can be connected or not, according to the edge set E.

Two-sample tests are widely used in many fields. We hope to study a non-parametric test procedure that uses minimal spanning tree which can represent the similarity between samples. It can be applied to network data. We summary the three tests (Friedman and Rafsky (1979),Chen and Friedman (2017),Chen et al. (2018))based on minimal spanning trees, based on the three tests, we propose two new rank tests called Cross-rank test and Inter-rank test, in this paper, we mainly study the Cross-rank test and introduce the idea of Inter-rank test, finally, we apply the three existed tests and Cross-rank test to our two network datasets: 1.two group brain networks of patients with and without Parkinson's disease; 2.Flight network data at U.S. airports in 2015.

2 Data Information

2.1 Parkinson dataset

We use the function read_csv to read data. "Park_All_1_" is experimental data for patients with Parkinson's disease, and "Park_Control_All" is the data for the control group, that is, those who do not have Parkinson's disease.

[1]Institute of Statistics and Big Data, Renmin University of China, Beijing 100081, China. E-mail: zhangmaoyu@ruc.edu.cn

2.1.1 Data description

Connectome data obtained through functional MRI brain imaging,which enables neuroscientists to measure correlations between activity in brain regions. For example, if stimulation activates both visual and auditory areas of the brain, then these areas can be considered to be related. Connectome data can applicated to health and disease, and can be graphically represented, with nodes defined by regions of interest in the brain (such as hearing, vision, and body movement), and edges that connect nodes are related by metric weights. A group of connections measured on an object may contain hundreds of nodes and tens of thousands of edges.

In this article, we use such a brain connectome data. We've got data from two experiments on whether or not people have Parkinson's disease. Because we have data from a csv file, and each column of the data represents the lower triangle of an adjacency matrix, which is a sample, so we first need to restore the data to the form that each sample is an adjacency matrix.

First, we convert the raw data into the form of adjacency matrix, then convert each adjacency matrix into the form of igraph, and store them in a list, and since the range of the original data is $(-1, 1)$, we normalized it to $(0, 1)$ for the convenience of subsequent work.

2.1.2 Data visualization

First of all, we have a general understanding of the structure of these data through the correlation coefficient graph.We can find that for Parkinson's patients, the correlation between some acupuncture points in the brain is much weaker than people without Parkinson.The first row of figure 1 and figure 2 is the person with and without Parkinson.

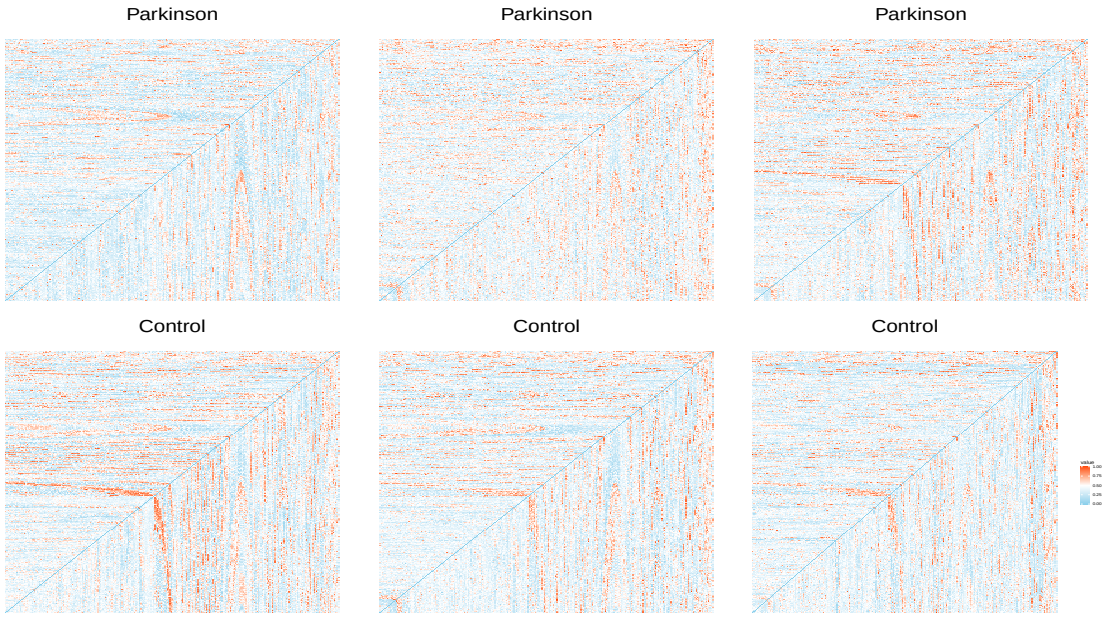


Figure 1: The heatmap of networks for Parkinson's patients and controls.

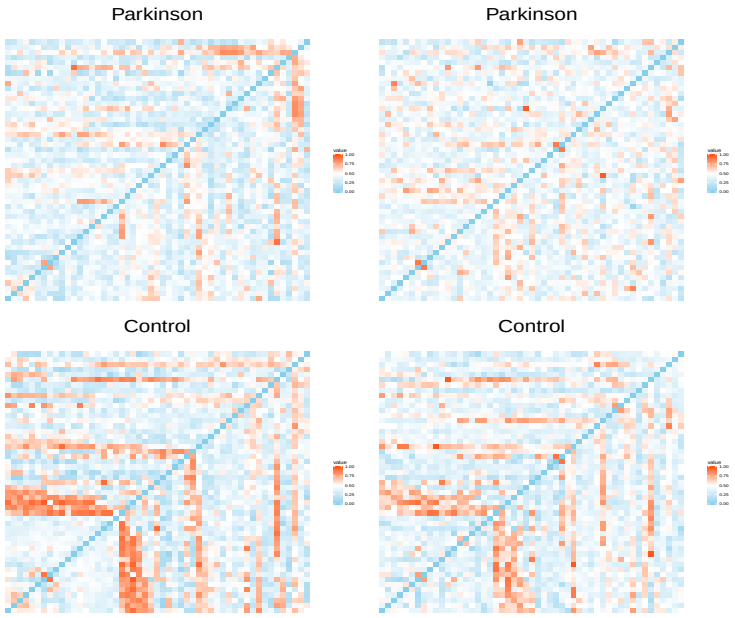


Figure 2: The heatmap of sub-networks for Parkinson's patients and controls.

2.1.3 Data summary

For weighted networks, a useful generalization of degree is the notion of vertex strength, which is obtained simply by summing up the weights of edges incident to a given vertex.

The distribution of strength sometimes called the weighted degree distribution is defined in analogy to the ordinary degree distribution. Therefore, we calculated the strength of network in the two groups to make a simple comparison.

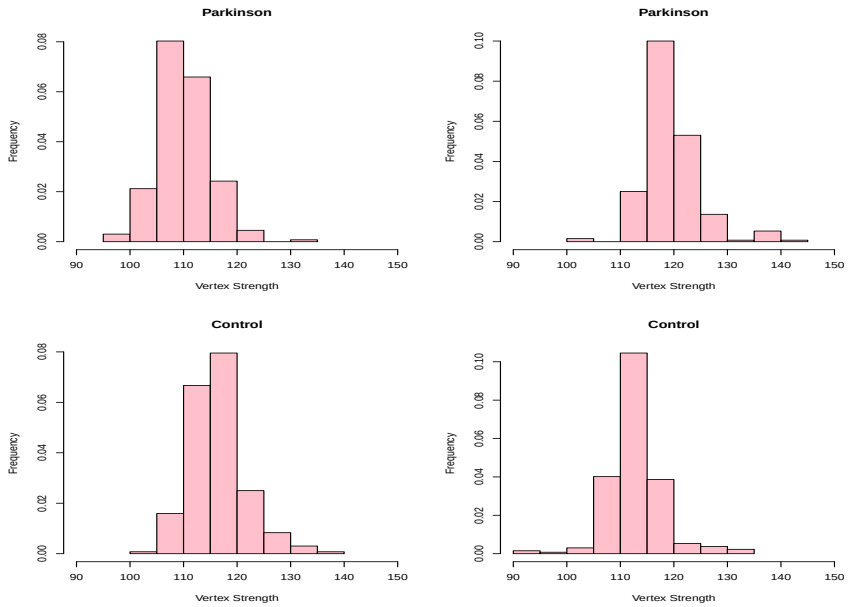


Figure 3: The distribution of virtex strength

Figure 3 shows the similar result as the figure 1 and 2, which indicates that the Parkinson's patients has smaller correlation in some brain region.

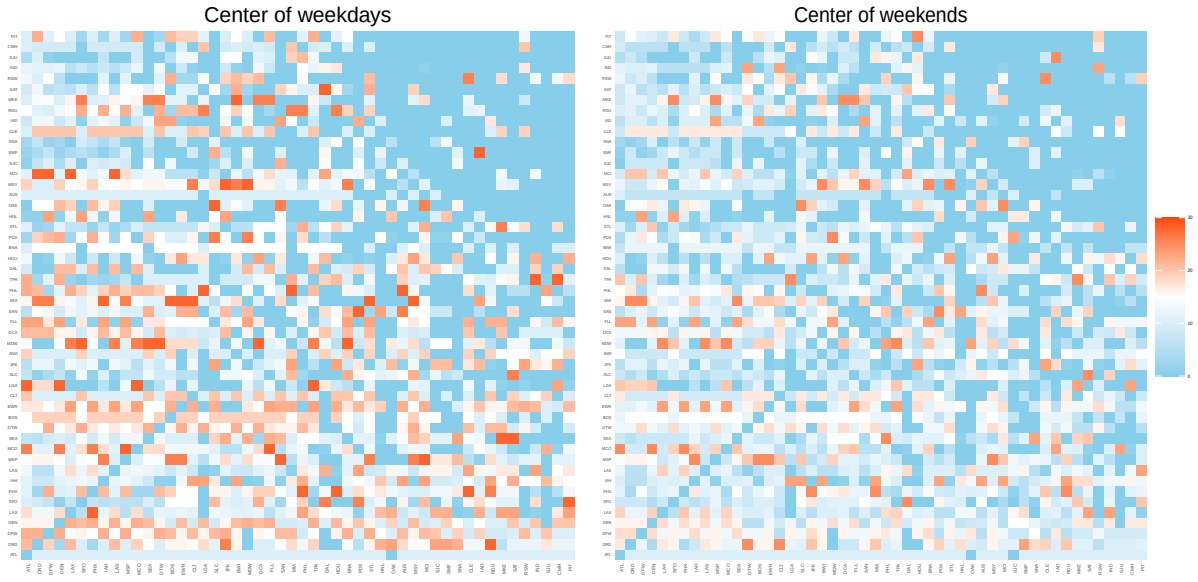
2.2 Flight dataset

We have the real data representing the flight information between 322 airprots in the United States in 2015, there data are reported by the U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics[1]. We only use the 50 busiest airports and their corresponding flights, we donete the 50 airports as the vertices of the network, and the number of flights from one airport to another airport is used as the weight of the network, which is directed, there are a total of 365 days of data, however, the data for october is missing to some extent, we remove the data of october, leaving 334 days, so that we have 334 networks, of which there are 239 days on weekdays and 95 days on weekends. To apply the rank test, we have consider the data on weekdays and on weekends as two groups for the test.

[1] `https://www.kaggle.com/usdot/flight-delays`

2.2.1 Data visualization

The heatmap and the vertex strength distribution of flight data are also shown in figure4 and figure 5. We can clearly see that there will be more flights on weekdays than on weekends.



(a) center of weekdays (b) center of weekends

Figure 4: The heatmap of two networks from two groups.

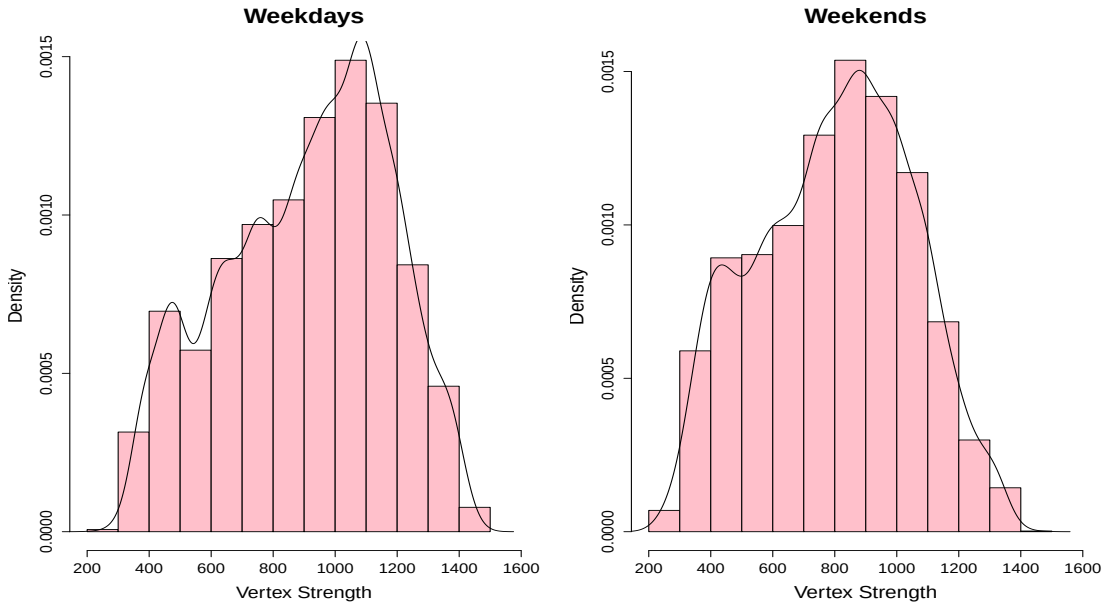
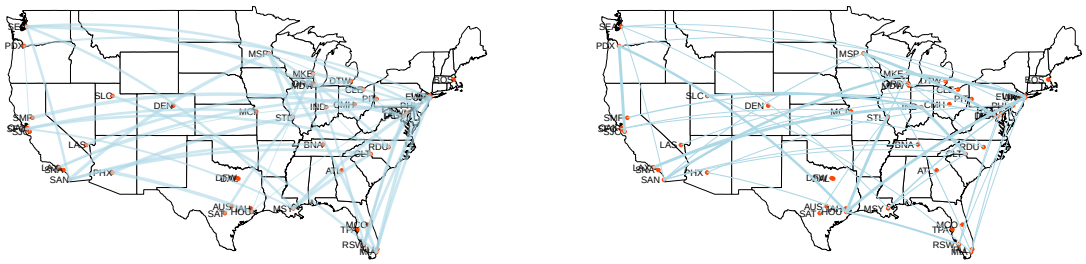


Figure 5: The distribution of virtex strength



(a) center of weekdays (b) center of weekends

Figure 6: The flight map of two networks from two group, the routes shown are displayed by selecting some routes whose absolute value of the difference between two networks is greater than 15.

3 Model

We begin by introducing some concept about minimal spanning tree as follow (Friedman and Rafsky, 1979):

A spanning subgraph of a given graph is a subgraph with node set identical to the node set of the given sraph, a spanning tree of a graph is a spanning subgraph that is a tree. Note that here is a (unique) path between every two nodes in a tree, and thus a spanning tree of a (connected) graph provides a path between every two nodes of the graph.

An edge weighted graph is a graph with a real number assigned to each edge, a minimal spanning tree (MST) of an edge weighted graph is a spanning tree for which the sum of edge weights is a minimum.

k-th MST: we know that MST connects all of the points with minimum total distance. A second MST connects all of the points with minimum total distance subject to the constraint that it be orthogonal to the first MST. A third MST connects the points with minimum total distance subject to the constraint that it be orthogonal to both the first and second MSTs. Generally, the k th MST is a minimal spanning tree orthogonal to the $(k-1)$ th through the first MST. We can use kth MST to conduct our test.

Three existing tests based on MST and two newly proposed tests based on MST rank will be introduced bellow, the null hypothesis of all tests is that the two sets of data come from the same distribution.

3.1 Edge-Count test (1979)

Number the $N-1$ edges of the MST arbitrarily and define $Z_i, 1 \leqslant i \leqslant N-1$, as follows: $Z_i = 1$, if the i th edge links nodes from different samples; $Z_i = 0$, otherwise.

Then

$$R = \sum_{i=1}^{N-1} Z_i + 1 \quad \text{and} \quad E[R] = \sum_{i=1}^{N-1} E\left[Z_i\right] + 1$$

$$W = \frac{R - E[R]}{(\mathrm{Var}[R])^{\frac{1}{2}}}$$

R is the edge-count statistic for this test, which has asymptotic properties, I will not show here. We reject the null hypothesis when R is small.

3.2 General test (2017)

Although many tests have been proposed for generic alternatives, in practice, none of them will be useful for both, to make the test more sensitive to scale alternative, the general test was proposed by Hao chen (2017), which is given by:

$$S = \left(R_1 - \mu_1, R_2 - \mu_2\right) \Sigma^{-1} \begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix}$$

where R_1 is the number of edges connecting observations both from sample $\mathbf{X}$, and R_2 is the number of edges connecting observations both from sample $\mathbf{Y}$, $\mu_1 = \mathbf{E}\left(R_1\right), \mu_2 = \mathbf{E}\left(R_2\right)$, and Σ is the covariance matrix of the vector $(R_1, R_2)'$ under the permutation null distribution. Under the location-alternative, or the scale-alternative for lowdimensional data, we would expect both R_1 and R_2 to be larger than their null expectations, then S would be large.

3.3 Weighted edge-Count test (2018)

When the sample sizes of the two samples are different, the tests above have small power than the balance sample size, Hao chen (2018) et.al solve the problem by applying appropriate weights to different components of the classic test statistic. Let $g_i = 0$ if the observation is from sample otherwise. For an edge $e = (i, j)$, we define:

$$J_e = \begin{cases} 0 & \text{if} \quad g_i \neq g_j \\ 1 & \text{if} \quad g_i = g_j = 0 \\ 2 & \text{if} \quad g_i = g_j = 1 \end{cases}$$

$$R_k = \sum_{e \in G} I_{J_e = k}, k = 0, 1, 2$$

Then R_0 is the number of between-sample edges (which is the test statistic for the edge-count test), R_1 is the number of edges connecting observations both from sample $\mathbf{X}$, and R_2 is the number of edges connecting observations both from sample $\mathbf{Y}$.

Thus the weighted test statistic is defined as:

$$R_w = qR_1 + pR_2, \quad p = \frac{m}{N}, \quad q = 1 - p$$

where m is the sample size of $\mathbf{X}$, n is the sample size of $\mathbf{Y}$, $m + n = N$. When the test statistic is defined in this way, its variance is well controlled no matter how different m and n are.

3.4 Cross-rank test based on MST

From the above three tests, we know that they are all based on the number of edges of the minimum spanning tree to construct statistics, and do not use the information of the distance to the minimum spanning tree. Based on the distance of minimal spanning tree, we propose a new test, Sort the distances of the $n-1$ edges of the minimum spanning tree, and sum the ranks of the between-sample edges to obtain a new statistic. Let d_{ij} is the distance of between $i-$th nodes and $j-$th node of the minimal spanning tree, $Q_{ij} = \text{ranks}(d_{ij})$, which is the rank for the edge connecting $i-$th nodes and $j-$th node, from samllest to highest, giving them a rank from 1 to n, then the new statistic is defined as:

$$W_0 = \sum_{i,j=1}^{n} Q_{ij} I_{\{\text{i,j from the different sample}\}}$$

3.5 Inter-rank test based on MST

Based on the cross-rank test, we propose another new Inter-rank test, first, we define:

$$W_1 = \sum_{i,j=1}^{n} Q_{ij} I_{\{\text{i,j from the sample X}\}}$$

$$W_2 = \sum_{i,j=1}^{n} Q_{ij} I_{\{\text{i,j from the sample Y}\}}$$

the new statistic is defined as:

$$T = \left(W_1 - \mu_1^{'}, W_2 - \mu_2^{'}\right) \Sigma_1^{-1} \begin{pmatrix} W_1 - \mu_1^{'} \\ W_2 - \mu_2^{i} \end{pmatrix},$$

where $\mu_1^{'} = \mathbf{E}(W_1)$, $\mu_2^{'} = \mathbf{E}(W_2)$, Σ_1 is the covariance matrix of the vector $(W_1, W_2)'$ under the permutation null distribution.

4 Simulation

In this section, we compare the size and power for the four tests, we generate samples from standard normal data, the dimention is 20, the sample size of the two distributions is $m = 81$ and $n = 36$, and consider four different types of the two distributions, and then compare their power.

Type 1: The means of the two multivariate Gaussian distributions differ in 1.

Type 2: The variances of two multivariate Gaussian distributions differ in a multiple of 2.

Type 3: The means of two multivariate Gaussian distributions partly differ in 1, which means only part of data are from different means.

Type 4: The variances of two multivariate Gaussian distributions partly differ in a multiple of 2, which means only part of data are from different variances.

Figure $7 - 11$ shows the size and power under different types of data for the four test mentioned above, where k is the number of orthogonal minimal spanning trees.

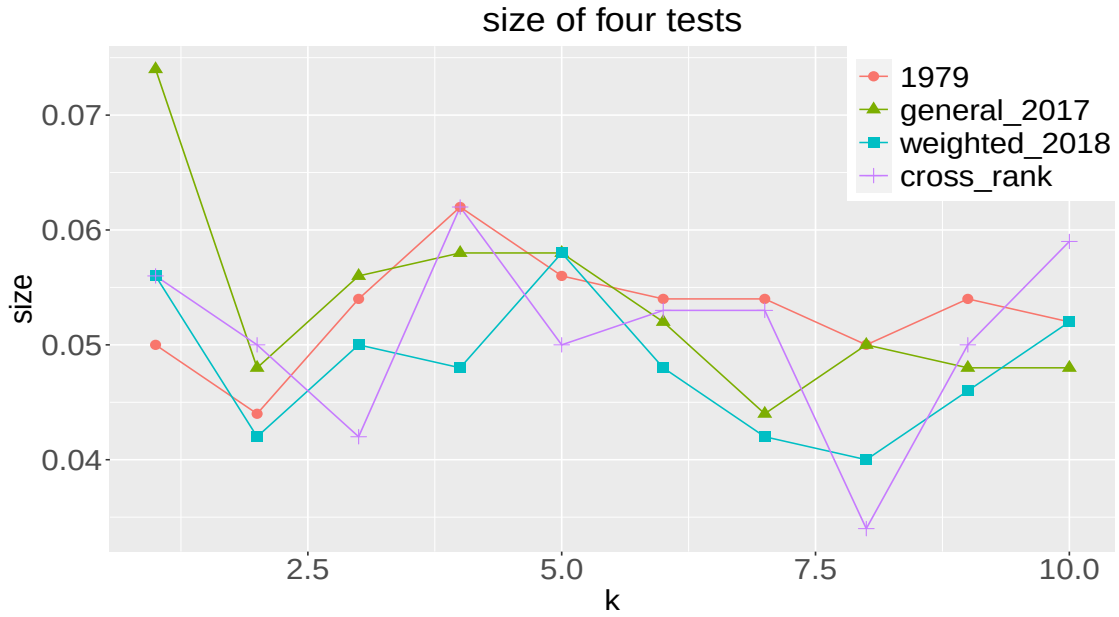


Figure 7: The fraction of trials (out of 500) that the test rejected the null hypothesis at 0.05 significance level of samples from the same distribution.

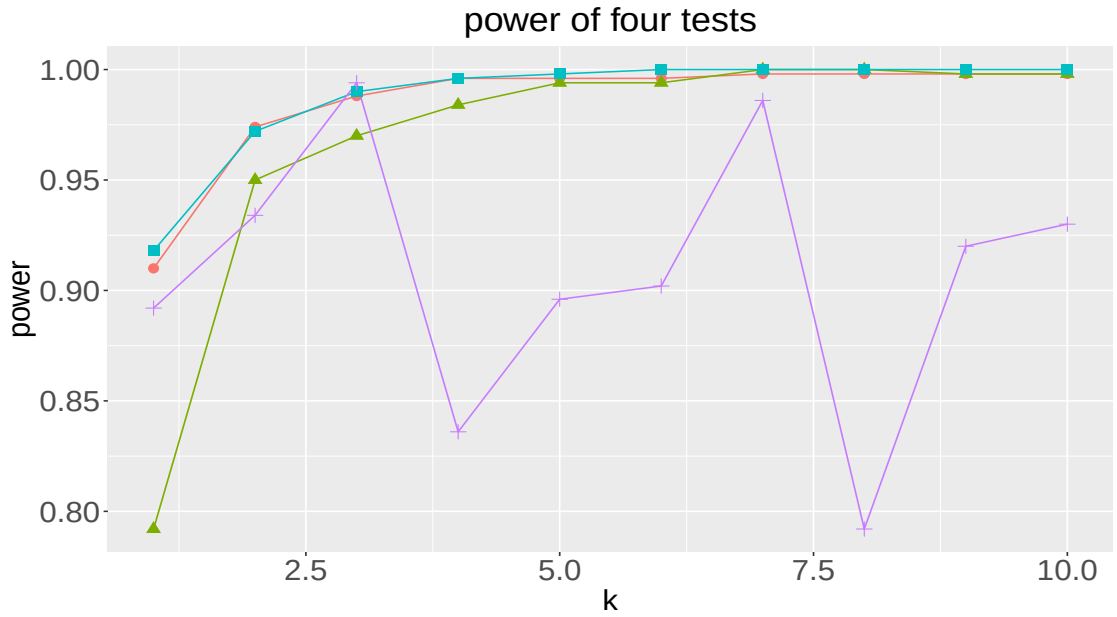


Figure 8: The fraction of trials (out of 500) that the test rejected the null hypothesis at 0.05 significance level of type 1 samples.

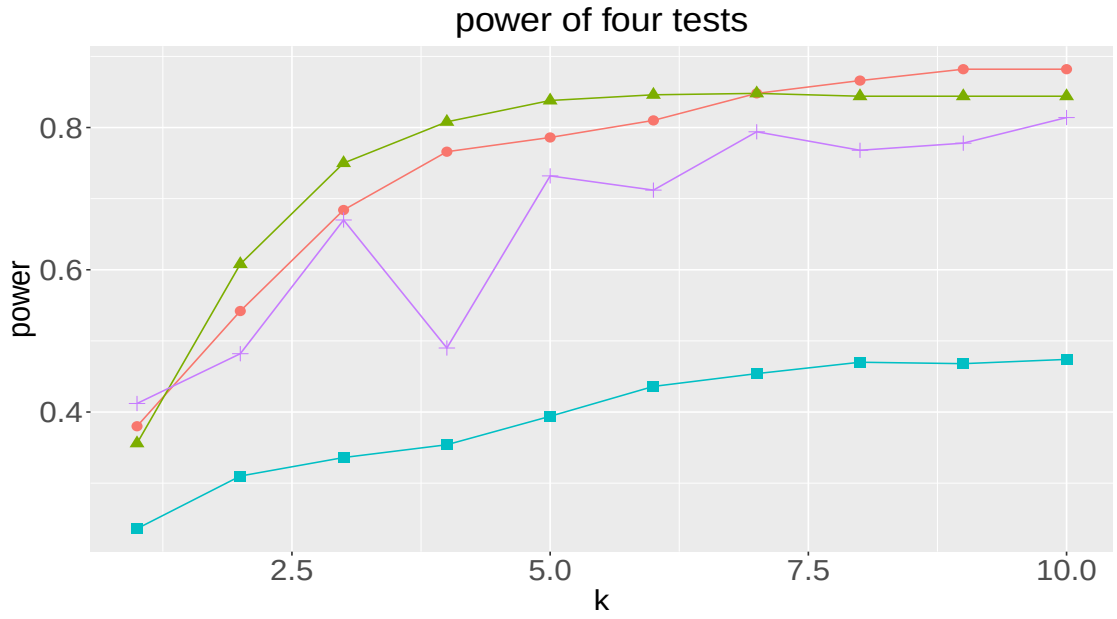


Figure 9: The fraction of trials (out of 500) that the test rejected the null hypothesis at 0.05 significance level of type 2 samples.

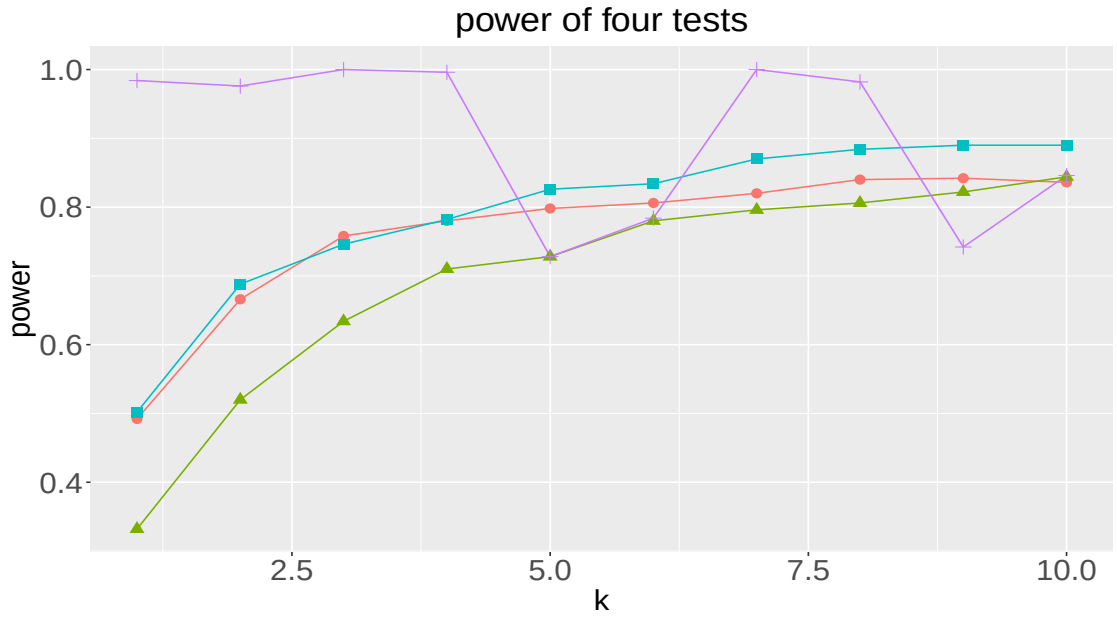


Figure 10: The fraction of trials (out of 500) that the test rejected the null hypothesis at 0.05 significance level of type 3 samples.

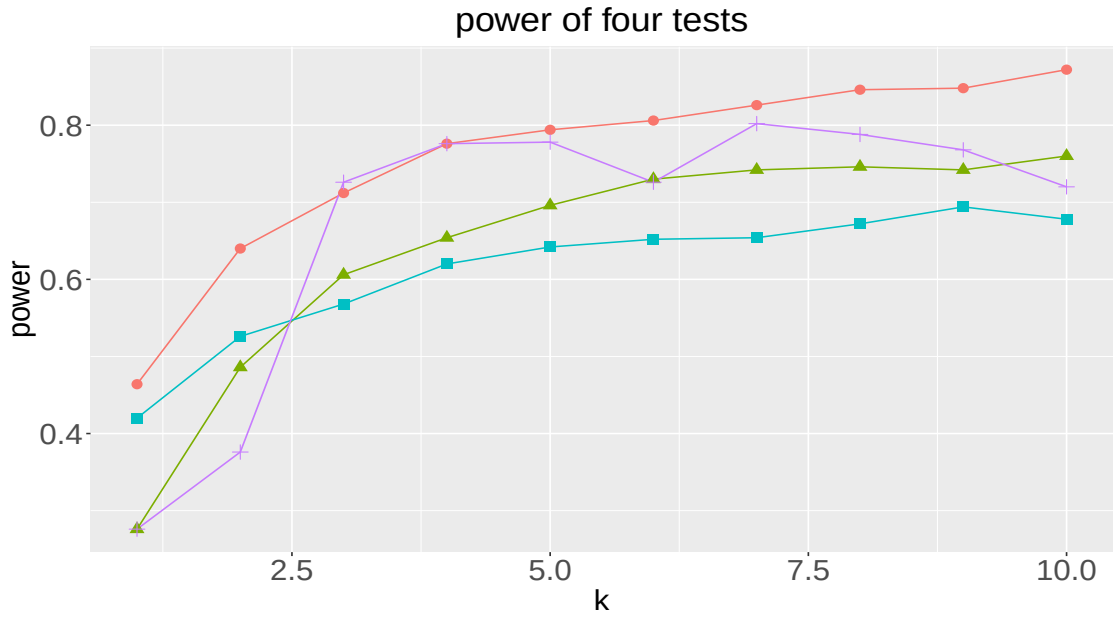


Figure 11: The fraction of trials (out of 500) that the test rejected the null hypothesis at 0.05 significance level of type 4 samples.

We can see from figures 8 and 9 that our method seems to have no advantage when the overall structure of the data changes, but from figures 10 and 11, our method has a slight

advantage in detecting local differences in data, but in general seeing that our method is not particularly stable, it still needs further research and improvement.

5 Real data application

We applied the above four tests to network data, which are the Parkinson data and flight data we mentioned earlier. Figure 12 shows the p-value of Parkinson data, and figure 13 shows the p-value of flight data, almost all test will reject the null hypothesis for both two datasets. Figure 13 show the p-value of different part of parkinson data using cross-rank test, we take out the obvious differences in the previous heat map to test separately, and at the same time check the remaining parts, compare the p-values of the local, all and the rest parts, which indicates that the main differences of the two group data are come from the local part and confirms that the operation of a certain area of the brain of Parkinson's patients is significantly different.

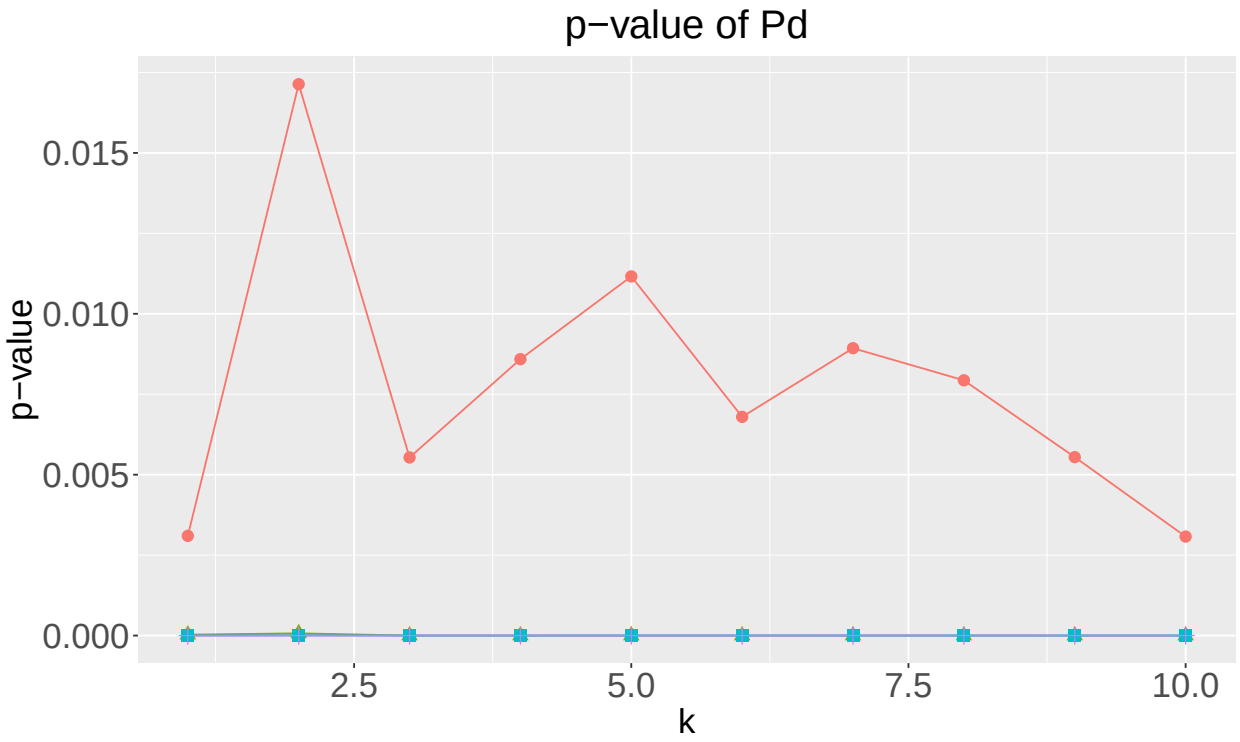


Figure 12

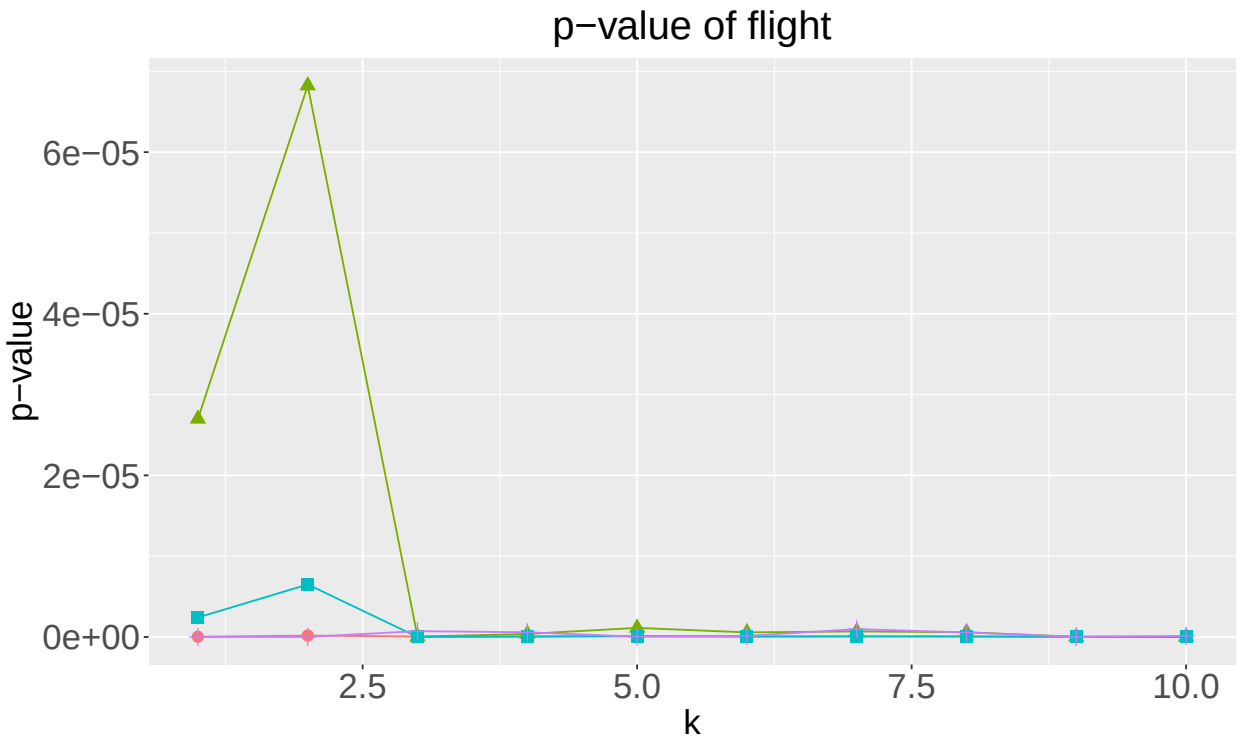


Figure 13

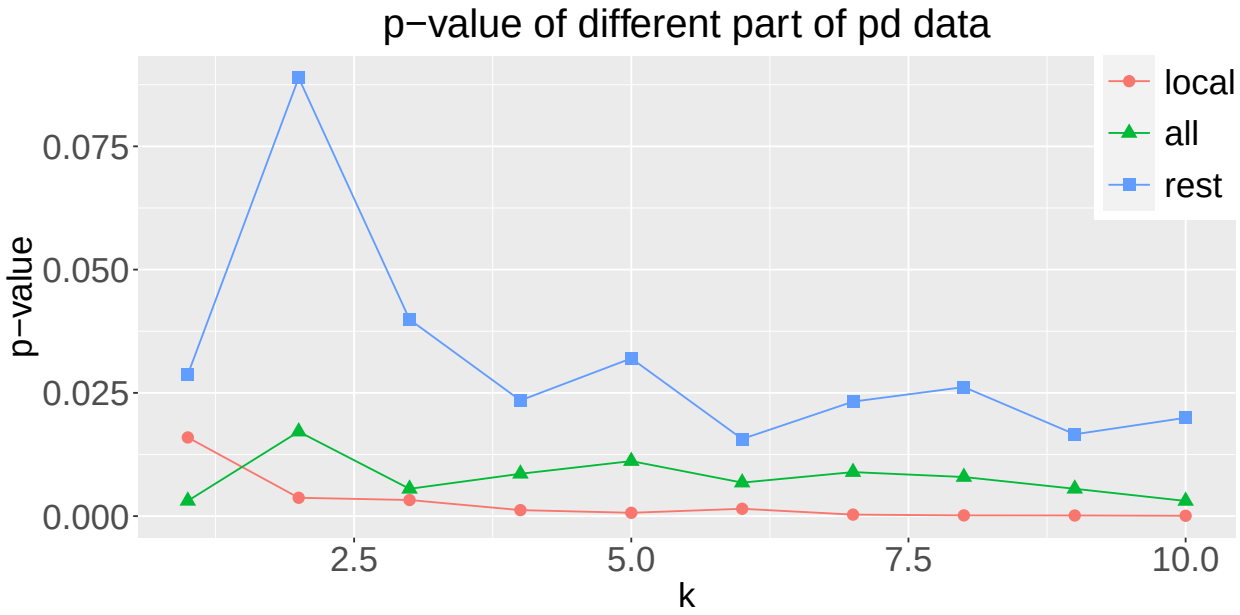


Figure 14

6 Discussion

From the results of the simulation, our method will perform slightly better than other methods in terms of differences in local data characteristics. This may be the place we will focus on next. Our method is generally not particularly good. , So it needs further improvement and thinking.

7 References

References

Chen, H., Chen, X., and Su, Y. (2018), "A weighted edge-count two-sample test for multivariate and object data," *Journal of the American Statistical Association*, 113, 1146–1155.

Chen, H. and Friedman, J. H. (2017), "A new graph-based two-sample test for multivariate and object data," *Journal of the American statistical association*, 112, 397–409.

Friedman, J. H. and Rafsky, L. C. (1979), "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *The Annals of Statistics*, 697–717.