Statistical tools for the analysis of network-valued data: theory, algorithms, and applications

A DISSERTATION PRESENTED
BY
ILENIA LOVATO
TO
THE DEPARTMENT OF MATHEMATICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
MATHEMATICS

UNIVERSITY OF PAVIA
PAVIA, ITALY

DECEMBER 2018

Thesis advisor: Prof. Simone Vantini Ilenia Lovato

Statistical tools for the analysis of network-valued data: theory, algorithms, and applications

Abstract

In the general framework of Object Oriented Data Analysis, the thesis is focused on network-valued data. The atom of the statistical analysis is a network and we have to deal with samples of networks. Network analysis is a widely studied area, but statistical tools for the analysis of *samples of networks* are almost lacking. The thesis presents some statistical tools for null hypothesis testing on network-valued data. After a general introduction on Object Oriented Data Analysis and in particular on network-valued data, a framework for a two-sample test for network-valued data in the context of permutation theory is introduced. The inferential properties of the method are studied as well (theoretically and via simulations) and an illustrative application to a bike sharing data set is presented. A substantial application of the two-sample test on three different data sets of brain networks is presented, together with a comparison of the results obtained via the classical approach that statistically compares samples of brain networks by means of brain summary measures. The data sets analyzed refer to different diseases (autism spectrum disorder and tuberous sclerosis complex in children) and to different acquisition procedures of the data (electroencephalography, functional Magnetic Resonance Imaging and Diffusion Compartment Imaging). The thesis then proceeds stimulated by clinical brain questions. Once it is known that there is a statistical significant difference between two samples, physicians may be interested in finding out which portions of the brain network are responsible for the observed difference. A general multiscale null-hypothesis testing for network-valued data is developed. The proposed method allows to specify a partition of the set of the vertices and to find out in which subnetworks

identified by the partition there is a significant difference between the two samples. The procedure guarantees the control of the probability of at least one false rejection (the so called Family Wise Error Rate) at level α. The inferential properties (the estimation of probability of rejection under the alternative hypothesis and the control of the Family Wise Error Rate) are studied theoretically and via simulations. An application on a brain networks data set is developed to find out in which hemispheres and in which lobes there are significant differences between autistic patients and healthy subjects. Both the (global) two-sample test and the multiscale procedure have been implemented in an R package called `nevada` (NEtwork-VAlued Data Analysis), whose structure and functions are detailed.

Contents

Listing of figures

To my family.

Acknowledgments

I would like to express my sincere gratitude to my supervisor Prof. Simone Vantini for the continuous support of my Ph.D study and research, for his patience, motivation, and enthusiasm. I would like also to thank my co-supervisors Dr. Alessia Pini and Dr. Aymeric Stamm.

A major trend in modern science is the collection of not only larger datasets, but also more complex datasets. While massive data is a term that is currently in very widespread use, data complexity may become the larger scale, and ultimately more important, challenge to the field of statistics.

Marron & Alonso, 2014

1 Introduction

Nowadays, the opportunity of recording data from every phenomenon, in many different conditions and trough different acquisition procedures pushes the collection of more and more elaborate data and open up perspectives in answering complex questions by means of data sets where the single atom is a complex object. The formulation of statistical tools is not going hand in hand with the increasing amount of data sets composed of complex objects. Therefore there is an unmet need for statistical methods able to analyze these type of data sets.

The statistical framework is that of Object Oriented Data Analysis (OODA). OODA is a field of growing interest that emerged from the seminal paper of [72]. It aims at conducting statistical analyses of complex data that cannot be embedded in the standard Euclidean framework [see 44, with discussion], by contrast with more traditional data sets composed of numbers or vectors of numbers that naturally lie in a Euclidean space in which standard statistical methods can be applied. Shapes [18], images [36, 74], manifold-valued data such as directional data [43], trees [72], covariance matrices [19] are examples of so-called object data. Investigating the relationships

between these complex objects requires the development of appropriate statistical tools that can be either generalizations of existing Euclidean methods or novel non-standard approaches [see 60].

In this thesis, we focus on a specific type of object 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. In recent years, networks have indeed become more and more popular in many different areas of scientific investigation, ranging from micro-scale networks such as protein-protein interaction networks, gene regulatory networks or cerebral networks, to macro-scale networks such as social networks, organizational networks, mobility and transport networks [see, for example, 48, chap. 2–5, for possible applications]. The nature of the vertices as well as the role of the edges of the network are application-specific. From the above-cited examples, vertices would be for instance proteins, molecular regulators, regions of the brain, users of a social network, working roles or geographical areas. Edges can be either binary or quantitative with corresponding networks called unweighted and weighted respectively. Binary edges usually encode the presence or absence of a relationship between two vertices. They could be physical interaction of proteins, molecular reactions, structural or functional connections between areas of the brain, friendship on a social network, collaborations between people on a firm or mobility connections between two geographical areas. Differently, quantitative edges measure the strength of the connection between the two vertices, such as the number of structural fibers between two areas of the brain or the amount of vehicles connecting two geographical areas for instance. Moreover, edges can also be directional: for example, in a social network, one might use edges to connect people on the basis of who follows who.

There is a large body of past and current literature on network analysis and its many applications. Yet, a vast majority of that literature has put the attention on the use of a network as an efficient way to represent and analyse data sets in which the interest is on exploring "interactions between entities, whether those entities are individuals in a school [45], species in a food web [33], nodes on a computer network [50], or proteins in metabolic pathways [28]. Network analysis is used to explore the mathematical, statistical and structural properties of a set of items (nodes) and the connections between them (edges; [46])" [4]. Consequently, the scientific effort has then been in the development of tools for constructing, describing and modeling a single network. From the point of view of OODA, these research goals can be framed among the so-called *first generation problems* of OODA in which the effort is spent in the proper construction of object data (S. Marron,

keynote talk at the 6th Nordic-Baltic Biometric Conference, June 19-21, 2017, Copenhagen, DK), which, in the present case, are networks. In this work, we instead focus on the *second generation problems* in OODA which pertains to the statistical analysis of samples of networks. In this setting, networks are considered as the units of the statistical analysis, hence the name of network-valued data. As a result, we have to deal with samples of networks that we model as *i.i.d.* realizations of *network-valued random variables*. The growing amount of available network-valued data urges the need for quantitative statistical tools to face this challenge which, at the moment, has been mostly tackled in a purely heuristic way [64].

Only recently, some proposals have been made in this direction. The first papers on statistical methodologies that investigate network-valued data appeared as a response to neuroimaging problems. Specifically, in [72] and [2] the author developed a Principal Component Analysis analog for a special type of networks coined tree-structured objects. This first OODA for trees is based on the concept of tree lines and the underlying optimization problem is solved in a linear computation time. A dataset of 73 vascular brain trees modeled as acyclic networks with vessels playing the role of edges and bifurcations playing the role of vertices is analysed. More recently, in [49] has been proposed a Principal Component Analysis approach in the space of phylogenetic trees.

When comparing samples of object data, the traditional approach pertains to transforming the individual object data into a multivariate collections of indicators describing the original object data. In the context of network-valued data, this translates into replacing a network by a multivariate vector of graph summary measures such as characteristic path length, clustering coefficient, modularity, global efficiency, betweenness centrality, degree distribution, degree centrality and so on (see [59] for a detailed list of graph summary measures). The comparison between networks is then framed as a classical multivariate data analysis rather than a network-valued data analysis. Despite the high interest coming from the interpretation of these summary measures, their dependence on the network size, the reliance of the resulting inference on the chosen measure and the need for information about the entire structure of networks have encouraged the formulation of new methodologies that do not rely on summary features. The aim of this thesis is to overcome this approach and propose a method that looks at the entire structure of every network in the data set, without the loss of information that comes from summarization. A comparison of these approaches is proposed on simulated data sets and on real data, as detailed in the following. The second chapter of the thesis is devoted to the construction of a framework for a two-sample test for network-valued

data. A substantial application on three different data sets of brain networks is presented in the third chapter together with a comparison of the results obtained via the classical approach that statistically compares samples of brain networks by means of brain summary measures. Clinical brain questions stimulate the fourth chapter. Once it is known that there is a statistical significant difference between two samples, physicians are interested in figure out which portions of the brain network are responsible for the observed difference. A multiscale null-hypothesis testing for network-valued data is developed in the fourth chapter. Both the (global) two-sample test and the multiscale procedure have been implemented in an R [56] package called `nevada` (NEtwork-VAlued Data Analysis), whose structure and functions are detailed in the fifth chapter.

In practice parametric methods reflect a modelling approach and generally require the introduction of a set of stringent assumptions, which are often quite unrealistic, unclear, and difficult to justify. In contrast, nonparametric approaches try to keep assumptions at a lower workable level, avoiding those that are difficult to justify or interpret, and possibly without excessive loss of inferential efficiency.

Pesarin & Salmaso, 2010

2

A two-sample null-hypothesis testing for network-valued data

As mentioned in the first chapter of the thesis, the comparison between samples of networks has been framed as a classical multivariate data analysis rather than a network-valued data analysis. The first attempt to account for the entire network structures when applying null hypothesis significance testing procedures can be found in [63]. The authors define a first statistic based on the Jaccard index to quantify similarity in key vertex locations between groups of networks. Next, they propose a second statistic as the ratio between the means of Kolmogorov-Smirnov statistics to compare the degree distributions of each vertex within and between groups.

In our opinion, the paper by [27] is a cornerstone paper moving in the direction of network-valued data. Motivated by a problem of functional neuroimaging investigation, the authors

[1] See [39].

model the Human brain as a network and derive a sound asymptotic theory for parametric null hypothesis significance testing of network-valued data represented by means of graph Laplacian matrices. In details, they characterize the geometry of the space of graph Laplacian matrices as a manifold with corners, generalize results from [6] to propose a Central Limit Theorem for the Frobenius-based Fréchet mean which allow them to naturally extend classical asymptotic results from textbooks to network-valued data analysis, including k-sample null hypothesis significance testing. They apply the proposed approach to the 1000 Functional Connectomes Project Data Set. Asymptotic theory unfortunately only yields approximate inference, null hypothesis significance testing procedures lack exactness and perform in an unreliable fashion when the sample size is small. Moreover, the proposed procedure requires the computation of the inverse of a covariance matrix which can become very challenging from a numerical point of view when the dimensionality of networks (number of vertices) is large, as stated by the authors themselves.

In a recent paper, [21] the authors introduce a Bayesian framework that can deal with samples of large networks. In details, the authors propose a probabilistic generative model for a network-valued random variable via a flexible Bayesian non-parametric approach. Dimensionality is reduced using a finite mixture model to define the joint distribution of the edges. See also the interesting discussion on [21] recently published on JASA. In [20] the authors further generalize this model for allowing the generative mechanism to change across groups and develop a general Bayesian procedure for inference and testing of group differences in the network structure.

In this chapter, a finite-sample exact and consistent permutation-based two-sample test for making inference on distributions of network-valued data is introduced. The permutation framework has the advantage of not relying on distributional assumptions about the underlying generative models, which comes in handy when these models are complex and/or no simple parametric approximation is available. Moreover, the proposed framework is very flexible: it is indeed possible to choose (i) an appropriate matrix representation for the networks, (ii) a suitable distance between networks and (iii) one or more test statistics for capturing relevant distributional differences. In this thesis, a number of possible representations, distances and statistics is detalied. It is straightforward to add more of them into the framework as well.

This chapter is organized as follows. Section 2.1 presents the statistical framework for network-valued data. It focuses on possible matrix representations of networks for mathematical tractability and proposes a non-exhaustive collection of distances between networks. We discuss

possible interpretation of pairs of representations and distances as well. Next, we introduce the concept of test statistics based on inter-point distances for carrying out null hypothesis significance testing. We review existing test statistics based on inter-point distances and propose two new such statistics which, when used together within the non-parametric combination framework [51, chap. 4], exhibit in three simulation settings the best performances in testing equality of distributions of networks. We then prove exactness and consistency of the permutation-based tests associated to the proposed statistics and the non-parametric combination approach is briefly summarized as well for self-content. Finally, in Section 2.2 and 2.3, we report results from simulation studies and an application to real data pertaining to the bike sharing service in Milan, respectively.

2.1 Statistical framework for network-valued data

2.1.1 Network representations

The first step of our procedure is based on a proper selection of a mathematical representation of each network. Recall that a network $G = (V, E)$ consists of a set V of vertices whose connections are defined within the edge set E. In the literature, three possible matrix representations of networks are mostly used, namely adjacency, Laplacian and modularity matrices, each describing specific aspects of a network.

Adjacency Matrix. The adjacency matrix, often denoted by W, reports at entry w_{ij} the strength of the edge between vertices i and j. Its elements must therefore be non-negative ($w_{ij} \geq 0$). This is the starting point of all matrix representations. If the network is *unweighted*, the strength of the connection boils down to its presence or absence ($w_{ij} = 1$ if $(i, j) \in E$). If the network is *undirected*, the adjacency matrix is symmetric ($w_{ij} = w_{ji}$). If there is no *self-loop* at vertex i (edge connecting vertex i with itself), the corresponding diagonal entry is equal to zero ($w_{ii} = 0$). A network is said *simple* if it is both undirected and without self-loops. In this case, the adjacency matrix W has a null diagonal and is symmetric.

Laplacian Matrix. By definition, the graph Laplacian matrix L can be derived from the adjacency matrix W as $L = D(W) - W$ where $D(W)$ is a diagonal matrix whose diagonal elements

are the degrees d_i of the corresponding vertices:

$$\ell_{ij} = \delta_{ij} d_i - w_{ij}, \text{ with } d_i = \sum_k w_{ik}.$$

This matrix takes its name from the so-called *heat equation* which reads $\partial u/\partial t - a\nabla^2 u$, where ∇^2 is the Laplacian operator. Indeed, the Laplacian matrix is nothing but the discretized version of ∇^2 on the set of vertices [see 48, Chapt. 6]. As a result, similar networks in their Laplacian representation will exhibit configurations of vertices and edges that lead to similar diffusion patterns. Moreover, the Laplacian matrix has some important properties. For example, if there are no self-loops, its eigenvalues are all non-negative, the number of null eigenvalues matches the number of connected components (i.e. subnetworks where any couple of vertices is connected by paths) and the space of simple networks is in bijection with the space of Laplacian matrices.

Modularity matrix. The third matrix representation that we discuss in this work is the modularity matrix B, whose elements are defined as follows:

$$b_{ij} = w_{ij} - \frac{d_i d_j}{2m},$$

where d_i and d_j are the degrees of vertices i and j, respectively, and $m = 1/2 \sum_i d_i$ is the total strength of the edges in the network. We can give a nice interpretation of the modularity matrix in the case of unweighted networks. The element b_{ij} is the difference between the actual weight of edge (i, j) and the expected number of edges between vertices i and j if edges were placed at random. Hence, the presence of non-zero elements in the modularity matrix provides evidence of structure within the network. For this reason, the modularity has been widely used for community detection in networks [47].

The three above-mentioned representations can be straightforwardly adapted to the simpler case of unweighted network by dichotomization. The easiest way consists in assigning 1 to edges with non-zero weight and 0 to the others. A finer dichotimization can be performed via a user-defined threshold above which an edge is assumed to exist.

2.1.2 Distances between networks

Comparing distributions of networks requires a mathematical tool for quantifying how far two networks are from each other. One of the first distances between two networks appeared in the 70's and is defined as the difference between their common number of vertices and the number of vertices in the largest common induced subnetwork [75]. Then, in the 90's, a number of statistics emerged around the concepts of edge rotation or slide [12, 32, 76]. In essence, an edge rotation is defined as a unit operation on a network that pertains to moving a single edge while keeping one of its vertices fixed. Edge slides are a subset of edge rotations in which the moving vertex can only be sent on vertices directly connected to it. Distances between two networks can then be defined as the smallest number of such operations required to transform one network into the other. However, such distances suffer from two major drawbacks: (i) they do not convey an easy interpretation and (ii) their computation is prohibitively time consuming.

In this work we instead take advantage of the matrix representation of a network and consider instead distances that have been recently proposed either on network matrices [14] or on covariance matrices [19], which are not computationally intense and easily interpretable. Let G_1 and G_2 be two networks sharing the same set of vertices V of cardinality N and X and Y be the chosen matrix representation for G_1 and G_2, respectively. We focus on the following distances:

Hamming distance. The Hamming distance between G_1 and G_2 is defined as:

$$\rho_{\mathrm{HA}}(G_1, G_2) = \sum_{i \neq j}^{N} \left| X_{ij} - Y_{ij} \right| ,$$

This distance takes its name after Richard Hamming who needed a way to detect errors in systems [30]. It is easier to grasp its interpretation from unweighted networks. It basically counts "matching errors", i.e. edges that are present in one network but not in the other.

Frobenius distance. The Frobenius distance between G_1 and G_2 is defined as:

$$\rho_{\mathrm{FR}}(G_1, G_2) = \left(\sum_{i \neq j}^{N} \left(X_{ij} - Y_{ij} \right)^2 \right)^{1/2} .$$

This distance is the most frequently used distance in the scientific literature as it is nothing but the

Euclidean distance on the vectorized chosen matrix representation. Interestingly, in the case of unweighted networks represented by the adjacency matrix, it coincides with the Hamming distance.

Spectral distance. The spectral distance between G_1 and G_2 is defined as:

$$\rho_{\mathrm{SP}}(G_1, G_2) = \left(\sum_{i=1}^{N} \left(\Lambda_i^X - \Lambda_i^Y \right)^2 \right)^{1/2},$$

where Λ^X and Λ^Y are vectors storing the (ordered) eigenvalues of X and Y, respectively. This distance only accounts for the eigenvalue structure of a network matrix representation, which captures topological features only, leaving aside the eigenvectors. Under this distance, two networks are considered equal if they differ only by a relabeling of the vertex set. Technically, the spectral distance is defined on the classes of equivalence; otherwise it is a semi-distance since the identity of indiscernibles does not hold in general.

Root-Euclidean distance. It is the Frobenius distance on the squared root of the network matrices:

$$\rho_{\mathrm{RE}}(G_1, G_2) = \rho_{\mathrm{FR}} \left(X^{1/2}, Y^{1/2} \right).$$

This distance can be particularly useful in the case of few large eigenvalues that could have a leverage effect on the comparison which is greatly reduced by the square root transform. This distance is used in the context of matrix-valued data [11, 19, 54], where it has been shown to yield high empirical power in group comparisons. It is defined only for positive definite matrices, which, among the representations proposed in Section 2.1.1, reduces to the Laplacian matrix.

2.1.3 Test statistics based on inter-point distances

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two samples of networks governed by probability distributions $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively. We aim at performing the following two-sample test for equality in distributions:

$$H_0 : \mathbf{F}_1 = \mathbf{F}_2 \quad \text{against} \quad H_1 : \mathbf{F}_1 \neq \mathbf{F}_2. \tag{2.1}$$

Let $G_{11}, \ldots, G_{1n_1} \sim \mathbf{F}_1$ be a sample of n_1 independent and identically distributed network-valued random variables following distribution $\mathbf{F}_1$ and $G_{21}, \ldots, G_{2n_2} \sim \mathbf{F}_2$ be a sample of n_2 independent and identically distributed network-valued random variables following distribution $\mathbf{F}_2$. For

conciseness, let us also introduce the compact notation $\mathbf{G}_k = \{G_{k1}, \ldots, G_{kn_k}\}$ for $k = 1, 2$.

The most frequent approach to the two-sample testing problem pertains to (i) defining a concept of mean element for a given distribution and (ii) using some appropriate distance between the two sample means as statistic for testing equality in distribution. Typically, the sample mean is computed as the element that minimizes its sum of squared distances with each sample unit. It is known as the sample Fréchet mean. This approach however presents a number of drawbacks that are non-trivial to solve. First, the sample Fréchet mean in general metric spaces is not always a consistent estimator of the theoretical Fréchet mean, as stated in 2013 by C. E. Ginestet (arXiv:1204.3183v4) and it could be not unique. Next, object data are often embedded in complex spaces into which there is no closed-form expression of the sample Fréchet mean [54]. It is possible to circumvent this problem either by computing it numerically or by resorting to restricted sample Fréchet means as done by [25] in the context of self-organizing maps. The first solution becomes rapidly prohibitively time-consuming from a computational standpoint. The second solution restricts the search for the minimum to the sample units themselves, which introduces large biases for small sample sizes. Lastly, comparing distributions on the basis of how far their sample means are from each other is too limited since differences in distributions might show up only in higher-order moments.

An alternative approach, that we adopt and promote for general metric spaces, is to define statistics using exclusively distances (denoted by ρ in the following definitions) between the pooled observations (inter-point distances), referred to as *inter-point statistics* or IP-statistics for short in the rest of the manuscript. Most of the state-of-the-art IP-statistics can be classified into two categories.

Characteristic-Based Statistics. These statistics combine inter-point distances in such a way that they can be seen as weighted L^2 distances between characteristic functions of the probability distributions to be compared. They are known in the literature as energy statistics [68] and have been generalized to separable Hilbert spaces [40]. The original energy statistic reads:

$$T_{\mathrm{SR}} := \frac{n_1 n_2}{n_1 + n_2}\left[\frac{2}{n_1 n_2}\sum_{i,j=1}^{n_1,n_2}\rho(G_{1i}, G_{2j}) - \frac{1}{n_1^2}\sum_{i,j=1}^{n_1}\rho(G_{1i}, G_{1j}) - \frac{1}{n_2^2}\sum_{i,j=1}^{n_2}\rho(G_{2i}, G_{2j})\right]. \tag{2.2}$$

Density-Based Statistics. These statistics combine inter-point distances in such a way to compare the density functions of the probability distributions of within-sample and

between-sample inter-point distances, which has been shown to be equivalent to comparing density functions of the two original probability distributions [41]. The easiest statistic along those lines has been proposed by [7] and reads:

$$T_{\mathrm{BG}} := \sum_{k=1}^{2} \left(\binom{n_k}{2}^{-1} \sum_{\substack{i=1 \\ j>i}}^{n_k} \rho(G_{ki}, G_{kj}) - \frac{1}{n_1 n_2} \sum_{i,j=1}^{n_1,n_2} \rho(G_{1i}, G_{2j}) \right)^2 . \tag{2.3}$$

Other statistics that exploit the same result first interpret the matrix of inter-point distances of the pooled sample as the adjacency matrix of a network and then design statistics based on a suitable similarity graph derived from this network. For example, [26] uses the minimum spanning tree while [57] uses the minimum distance non-bipartite pairing tree. [13] nicely reviews statistics based on similarity graphs and proposes a generalized edge-count statistic T_{CF} that is able to identify both mean and variance differences.

Other more complex IP-statistics (not included in this work) exist in the literature [29, 35] but require further modelling assumptions and are not easy to implement.

Inspired by the above-mentioned literature on IP-statistics and motivated by the observation that it might be relevant to detect higher-moment differences between distributions, we hereby introduce two novel IP-statistics, which read:

$$\begin{aligned} T_{\mathrm{IP-Student}} &:= \frac{\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \rho^2(G_{1i}, G_{2j}) - \left(\widehat{\sigma_1^2} + \widehat{\sigma_2^2}\right)}{\frac{\widehat{\sigma_1^2}}{n_1} + \frac{\widehat{\sigma_2^2}}{n_2}} \quad \text{and} \\ T_{\mathrm{IP-Fisher}} &:= \max\left(\frac{\widehat{\sigma_1^2}}{\widehat{\sigma_2^2}}, \frac{\widehat{\sigma_2^2}}{\widehat{\sigma_1^2}} \right), \end{aligned} \tag{2.4}$$

where $\widehat{\sigma_1^2}$ and $\widehat{\sigma_2^2}$ are unbiased estimators of the within-sample variances given by:

$$\widehat{\sigma_1^2} := \frac{1}{n_1(n_1-1)} \sum_{i=1}^{n_1} \sum_{j>i}^{n_1} \rho^2(G_{1i}, G_{1j}) \quad \text{and} \quad \widehat{\sigma_2^2} := \frac{1}{n_2(n_2-1)} \sum_{i=1}^{n_2} \sum_{j>i}^{n_2} \rho^2(G_{2i}, G_{2j}).$$

The first one is a *Student*–like statistic in the sense that it mimics the squared Student-Welch

statistic, which nicely captures mean differences even under unequal variances, and the second one is a *Fisher*–like statistic in that it mimics Fisher variance ratio statistic and is useful in detecting differences in variances. We use a mechanism called Non-Parametric Combination (NPC) that uses both statistics for designing a test that captures both mean and variance differences with high statistical power. The proposed test and the NPC are detailed in the next section.

2.1.4 The permutation framework for hypothesis testing

Given a test statistic, one can design statistical tests in either a parametric or a non-parametric fashion. In the case of network-valued random variables, the generative probabilistic models can be quite complex, making the parametric way almost impractical. Asymptotic results can be achieved as in [27] but suffer from unreliability when sample sizes are small or when network sizes are large. In this section, we instead formalize a non-parametric statistical test using permutation theory [51], which yields exact and consistent inference with minimal distributional assumptions at the cost of increased computational burden.

Permutation Test. Recall that we aim at designing a permutation two-sample test for equality in distributions as specified by Eq. (2.1). Let T be a generic test statistic that grasps – with large positive values – possible differences between $\mathbf{F}_1$ and $\mathbf{F}_2$. Assume that the distributions $\mathbf{F}_1$ and $\mathbf{F}_2$ are continuous. This assumption guarantees that - with probability 1 - independent data observations are all distinct. Let t_{obs} be the value of T obtained from the observed networks. Under the null hypothesis, networks in the two samples are exchangeable. Hence, it is possible to estimate the null distribution of T by randomly permuting the group labels of the observed networks. For each permutation, we obtain a value of the "permuted" test statistic, say t_{perm}. The set of all t_{perm} values is called *permutation distribution* and defines a discrete approximation of the null distribution of the test statistic. The total number m_t of possible permutations is equal to $m_t = (n_1 + n_2)!/n_1!/n_2!$ and if the test is two-sided and $n_1 = n_2$, it is further divided by a factor of two. In any event, the number of possible permutations m_t grows very fast with the sample sizes. For example, when $n_1 = n_2 = 16$, which are not in general considered as large sample sizes, we should enumerate $m_t > 3 \cdot 10^8$ permutations, which, in fact, makes the exhaustive computation of the permutation distribution prohibitively time-consuming. Hence, it is common practice to randomly sample a subset of m permutations with replacement among the m_t possible ones. Given a random set of permutations,

there are different ways of estimating the p-value out of the mechanics of permutations. The most common approach pertains to counting the number of times the value of t_{perm} is equal or exceed the observed value t_{obs} out of the m sampled permutations [51]. This approach, while providing an unbiased estimate of the p-value, fails to provide exact testing procedures in the usual sense of the term because it does not account for the variability introduced by sampling the permutations. In this work, we instead rely on the definition proposed by [53], which takes its roots in randomization tests. We opt for this choice because it always provides an exact test (i.e. $P_{H_0}[p \leq \alpha] = \alpha$) regardless of the sample sizes, the number m of sampled permutations and the value of α [53]. Hence, the choice of m only impacts the power of the test, as expected. This p-value is computed as follows [22, 53]:

$$p(T) = \frac{1}{m_t + 1} \sum_{b_t=0}^{m_t} F\left(b(T); m, \frac{b_t + 1}{m_t + 1}\right) \simeq \frac{b(T) + 1}{m + 1} - \int_0^{0.5/(m_t+1)} F(b(T); m, p_t) \mathrm{d}p_t, \quad (2.5)$$

where F is the cumulative probability function of the binomial distribution and $b(T)$ is the number of t_{perm} greater than t_{obs}. In practice, the exact computation via summation is performed when $m_t < 10,000$. Otherwise, the integral approximation is used. This estimated p-value allows for a fair power comparison in the simulations presented in Section 2.2. In addition to the exactness of the test, it can be shown that a permutation test based on our new test statistics (i.e. $T_{\text{IP}-\text{Student}}$ and $T_{\text{IP}-\text{Fisher}}$) is consistent. The following theorems hold:

Theorem 1. *Let G_1 and G_1' be two network-valued random variables following distribution $\mathbf{F}_1$ and G_2 and G_2' be two network-valued random variables following distribution $\mathbf{F}_2$. If $E[\rho^2(G_1, G_1')] < +\infty$ and $E[\rho^2(G_2, G_2')] < +\infty$, the permutation test based on the IP-Student statistic involving Frobenius, Spectral or Root-Euclidean distance is consistent under the alternative hypothesis of unequal means, namely $P_{H_1}[p(T_{\text{IP}-\text{Student}}) \leq \alpha] \to 1$ as $n_1 + n_2 \to \infty$.*

Proof. In this proof we partially follow [65]. For the law of large numbers, we have that for $n = n_1 + n_2 \to \infty$

$$\frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\rho^2(G_{1i},G_{2j}) \to E[\rho^2(G_1,G_2)],$$

$$\begin{aligned}\frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j>i}^{n_1}\rho^2(G_{1i},G_{1j}) &= \frac{1}{2}\frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}\rho^2(G_{1i},G_{1j})\\ &\to \frac{1}{2}E[\rho^2(G_1,G_1')],\\ \frac{1}{n_2(n_2-1)}\sum_{i=1}^{n_2}\sum_{j>i}^{n_2}\rho^2(G_{2i},G_{2j}) &= \frac{1}{2}\frac{1}{n_2(n_2-1)}\sum_{i=1}^{n_2}\sum_{j\neq i}^{n_2}\rho^2(G_{2i},G_{2j})\\ &\to \frac{1}{2}E[\rho^2(G_2,G_2')].\end{aligned}$$

Therefore, for $n = n_1 + n_2 \to \infty$ the numerator of $T_{\text{IP}-\text{Student}}$ tends to

$$E[\rho^2(G_1,G_2)] - \frac{1}{2}E[\rho^2(G_1,G_1')] - \frac{1}{2}E[\rho^2(G_2,G_2')], \tag{2.6}$$

where G_1, G_1', G_2, G_2' are independent random variables, G_1 and G_1' are independent and identical distributed from $\mathbf{F}_1$ and G_2 and G_2' are independent and identical distributed from $\mathbf{F}_2$. If ρ is one of the distances between Frobenius, Spectral, and Root-Euclidean described in Subsection 2.2, applying [67][Theorem 2], it is possible to prove that the expression in 2.6 is always non–negative and it is equal to zero if and only if $E[G_1] = E[G_2]$. In effect, as mentioned in Section 2, the Frobenius distance is nothing but the Euclidean distance on the vectorized matrix representation. Also Spectral and Root-Euclidean distances can be traced back to an Euclidean distance between vectors. Therefore [67][Theorem 2] can be applied for the three distances mentioned above, yielding the following inequality under the alternative hypothesis H_1 of unequal means:

$$E[\rho^2(G_1,G_2)] - \frac{1}{2}E[\rho^2(G_1,G_1')] - \frac{1}{2}E[\rho^2(G_2,G_2')] > 0.$$

As a result, the numerator of $T_{\text{IP}-\text{Student}}$ tends to a strictly positive constant under H_1 when $n = n_1 + n_2 \to \infty$. The denominator $\widehat{\sigma}_1^2/n_1 + \widehat{\sigma}_2^2/n_2$ tends instead to zero (recall that for hypothesis $E[\rho^2(G_1,G_1')] < +\infty$ and $E[\rho^2(G_2,G_2')] < +\infty$). Eventually, $T_{\text{IP}-\text{Student}} \to +\infty$ when $n = n_1 + n_2 \to \infty$, and hence the permutation test based on $T_{\text{IP}-\text{Student}}$ is consistent for the three

distances mentioned above. □

Moreover, observing that the Hamming distance is a ℓ_1 distance on the vectorized matrix representation, one could think of following the same line of the proof for Frobenius, Spectral, and Root-Euclidean distance. In effect, [66][Theorem 1] guarantees a similar result to that of [67][Theorem 2], but for a general function, instead for a power of the Euclidean distance. This general result is based on the hypothesis that the function must be of strictly negative type. It is well known that ℓ_1 metric space is of negative type but it fulfills the condition of being of strict negative type only in a weaker sense [34] that is not sufficient for our aim. Therefore, the numerator of $T_{\text{IP}-\text{Student}}$ with the Hamming distance could be zero even when $E[G_1] \neq E[G_2]$ and so the consistency is not guaranteed in this case.

Theorem 2. *Let G_1 and G_1' be two network-valued random variables following distribution $\mathbf{F}_1$ and G_2 and G_2' be two network-valued random variables following distribution $\mathbf{F}_2$. If $E[\rho^2(G_1, G_1')] < +\infty$ and $E[\rho^2(G_2, G_2')] < +\infty$, the permutation test based on the IP-Fisher statistic is consistent under the alternative hypothesis of unequal variances, namely $P_{H_1}\left[p(T_{\text{IP}-\text{Fisher}}) \leq a\right] \to 1$ as $n_1 + n_2 \to \infty$.*

Proof. The following limits in probability under H_0 and H_1 hold, respectively:

$$T_{\text{IP}-\text{Fisher}}(n) \to c_0 \quad T_{\text{IP}-\text{Fisher}}(n) \to c_1 \quad \text{as} \quad n = n_1 + n_2 \to \infty,$$

where $c_0 = 1$ and $c_1 > 1$. Therefore, it is immediate to prove by contradiction that there exists $\bar{n}$ such that for all $n \geq \bar{n}$

$$P_{H_1}[T_{\text{IP}-\text{Fisher}}(n) \geq x] \geq P_{H_0}[T_{\text{IP}-\text{Fisher}}(n) \geq x] \quad \text{for all } x$$

and the strict inequality holds for some x. This concludes the proof since the stochastic dominance of $T_{\text{IP}-\text{Fisher}}$ under H_1 on $T_{\text{IP}-\text{Fisher}}$ under H_0 guarantees the consistency of the permutation test [51]. □

Non-Parametric Combination. The IP-statistics proposed in Eq. (2.4) are designed to detect differences in mean – for $T_{\text{IP}-\text{Student}}$ – and variance – for $T_{\text{IP}-\text{Fisher}}$ – independently. In order to make the test sensitive to both mean and variance, we propose to combine the two statistics by means of the Non-Parametric Combination (NPC) methodology [9, 51].

Given a random set of B permutations, we first compute the value t_{obs} and values t_{perm} of the two test statistics and concatenate them into two vectors (one for each statistic) of size $B + 1$, say $\mathbf{T}_{\text{IP}-\text{Student}}$ and $\mathbf{T}_{\text{IP}-\text{Fisher}}$. We then transform these two vectors by ranking them in descending order and dividing their ranks by $B + 1$. This effectively produces two vectors $\pi_{\text{IP}-\text{Student}}$ and $\pi_{\text{IP}-\text{Fisher}}$ of "intermediate p-values", because, for a given permutation, the transformation boils down to a permutation p-value in which the corresponding permuted data is taken as observed data. Next, we combine $\pi_{\text{IP}-\text{Student}}$ and $\pi_{\text{IP}-\text{Fisher}}$ into a single vector $\mathbf{T}_{\text{IP}-\text{StudentFisher}}$ of size $B + 1$, the entries of which are then interpreted as the observed value and permuted values of a new combined statistic $T_{\text{IP}-\text{StudentFisher}}$. There are a number of possible combining functions [51]. One important property is that large combined values should be in favor of the alternative hypothesis. In our framework, we use Tippett's combining function $\psi(x, y) = 1 - \min(x, y)$ [70] which guarantees that the null hypothesis is rejected when at least one of the two independent tests rejects it. The p-value of the combined test is then computed applying Eq. (2.5) using the values in $\mathbf{T}_{\text{IP}-\text{StudentFisher}}$. The non-parametric combination methodology yields consistent tests if the "intermediate" tests based on the individual statistics are marginally unbiased (i.e. $P_{H_1}[p(T) \leq a] \geq P_{H_0}[p(T) \leq a] = a$) and at least one of them is consistent [see 51, chap. 4]. Specifically, we have the following result:

Corollary 1. *The permutation test based on the statistics $T_{\text{IP}-\text{Student}}$ and $T_{\text{IP}-\text{Fisher}}$ combined through the NPC methodology is consistent under the alternative hypothesis of unequal means or variances, namely $P_{H_1}\left[p(T_{\text{IP}-\text{StudentFisher}}) \leq a\right] \to 1$ as $n = n_1 + n_2 \to \infty$.*

Furthermore, the combined test is exact because the "partial" tests based on $T_{\text{IP}-\text{Student}}$ and $T_{\text{IP}-\text{Fisher}}$ are exact [see 51, chap. 4].

2.2 Simulation studies

2.2.1 Impact of different test statistics

The goal of this simulation is to draw a comparison between the proposed IP-statistics (2.4) and the state-of-the-art IP-statistics T_{SR} (2.2), T_{BG} (2.3) and T_{CF} that we compute using a minimal spanning tree of density 5, as suggested by the authors. For this purpose, we generate two samples of networks with 25 vertices. Each network is generated by sampling independent and identically distributed edge weights from a binomial distribution $\mathcal{B}(n, p)$. We simulate three different scenarios

to generate distributions that differ only in their means, only in their variances or in both. The parameters n and p of the binomial distribution are set accordingly. In details, we have:

Scenario 1: Unequal means, equal variances. The two samples are generated using an edge weight distribution with different means such that $\Delta = \mu_1 - \mu_2 = 0{\cdot}000, 0{\cdot}125, 0{\cdot}250, 0{\cdot}375, 0{\cdot}500$ but equal variances $\sigma_1^2 = \sigma_2^2 = 2{\cdot}50$.

Scenario 2: Equal means, unequal variances. The two samples are generated using an edge weight distribution with different variances such that $\Delta = \sigma_2^2/\sigma_1^2 = 1{\cdot}00, 1{\cdot}05, 1{\cdot}10, 1{\cdot}15, 1{\cdot}20$ but equal means $\mu_1 = \mu_2 = 60$.

Scenario 3: Unequal means, unequal variances. The two samples are generated using an edge weight distribution with different means such that $\Delta = \mu_2 - \mu_1 = 0{\cdot}0, 0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4$ and different variances such that $\sigma_2^2/\sigma_1^2 = 1{\cdot}00, 1{\cdot}05, 1{\cdot}10, 1{\cdot}15, 1{\cdot}20$.

The three scenarios are evaluated both under equal sample sizes ($n_1 = n_2 = 20$) and under unequal sample sizes ($n_1 = 30$ and $n_2 = 10$). The balanced sample sizes are typical from many real-life data sets. The unbalanced sample sizes are representative of studies of neurological disorders for instance. For all scenarios and statistics, we use the adjacency matrix representation and the Frobenius distance as done in [13]. The p-value is calculated using Eq. (2.5) and the significance level is set to $\alpha = 0{\cdot}05$. The comparison between statistics is drawn in terms of statistical power, estimated as probability of rejection via Monte-Carlo simulations using a total of 100,000 replicates.

Figure 2.2.1 reports the estimated probability of rejection as the difference between the two samples increases ($\Delta = 0$ yields the nominal level of the test; $\Delta > 0$ yields power estimates). First, we can observe that the effect of unbalanced sample sizes (second row), independently from the statistics and type of differences, almost always generates a slight loss of statistical power. The ranking of the statistics in terms of statistical power is however identical in the balanced and unbalanced cases. The statistics T_{SR} and $T_{\text{IP}-\text{Student}}$ outperforms other statistics for detecting mean-only differences (first column). On the other hand, they feature the worst performances for detecting variance-only differences (second column). The reciprocal holds for the statistics T_{BG} and $T_{\text{IP}-\text{Fisher}}$, which feature the best performances for detecting variance-only differences but are the worst for detecting mean-only differences. Their comparison under both mean and variance differences (third column) is less helpful because it depends on the relative magnitudes of mean and

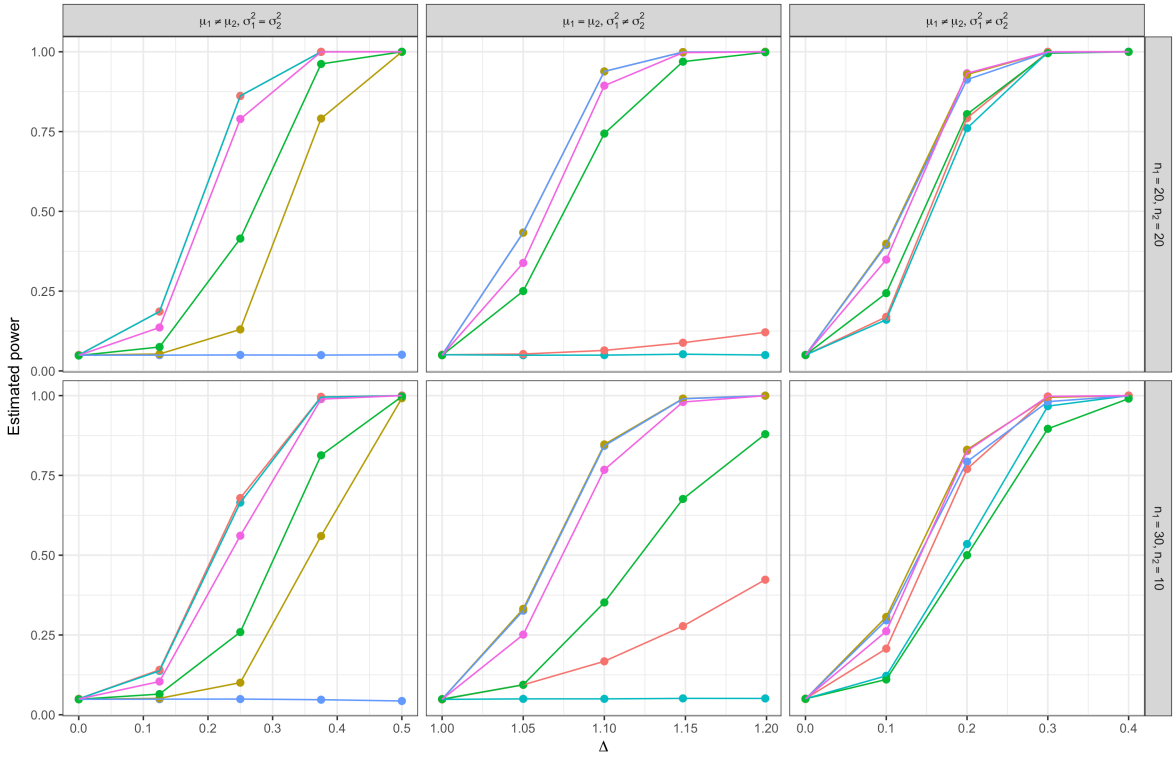


Figure 2.2.1: Power of the test using different test statistics: T_{SR} (2.2) in red, T_{BG} (2.3) in brown, T_{CF} in green, $T_{\text{IP}-\text{Student}}$ (2.4) in light blue, $T_{\text{IP}-\text{Fisher}}$ (2.4) in blue and $T_{\text{IP}-\text{StudentFisher}}$ in pink. The largest standard error is 0·00158.

variance differences. The statistics T_{CF} and $T_{\text{IP}-\text{StudentFisher}}$ lead to statistical powers that are insensitive to the type of differences to be detected. Our combined statistic $T_{\text{IP}-\text{StudentFisher}}$ features however uniformly better performances than T_{CF}. In fact, $T_{\text{IP}-\text{StudentFisher}}$ is the best statistic for detecting simultaneous mean and variance differences and always second-best for detecting mean-only or variance-only differences.

2.2.2 Impact of representations and distances

The goal of this second simulation study is two-fold: (i) to highlight some of the properties of the representations/distances enumerated in this work (Scenarios A, B, C) and (ii) to emphasize that it is critical, when comparing network samples, to focus on the entire network structure and not only on summary indicators (Scenario D). Specifically, we report simulation results pertaining to all

three matrix representations (adjacency, Laplacian and modularity) but, for simplicity, only to two out of the four introduced distances, namely the Frobenius and spectral distances. In effect, simulations showed that, at equal matrix representation, the results with the Hamming and Root-Euclidean distances were similar to those with the Frobenius distance. Similarly to the previous simulation setting, sampled networks are composed of 25 vertices. In all simulations, we assessed the effect of increasing sample size by generating samples S1 and S2 of sizes $n_1 = n_2 = 4$, 8, 12, 16. We designed a total of four scenarios, each with a specific aim, that we hereby describe:

Scenario A. Trivial differences: different edge strengths. The goal is to assess the performances of our test procedures when the probabilistic generative models governing the two samples are different but close. To this end, we defined the two samples using their edge weight distributions. Specifically, we drew the edge weight distribution of S1 from a Poisson distribution with mean $\lambda = 5$ and the edge weight distribution of S1 from a Poisson distribution with mean $\lambda = 6$. This yields an absolute difference of 1 between means and 0·21 between standard deviations of edge weight distributions.

Scenario B. Non-trivial differences: different vertex labelling. The goal is to show that using a relabelling-invariant distance such as the spectral distance to compare network samples coming from distributions that only differ up to a relabelling of the vertices fails to detect differences while other types of distances succeed. To this end, we drew both S1 and S2 from the stochastic block model [31] with different preference matrices. In details, for drawing S1, we used a 3×3 block matrix of edge probabilities with 0·8 in block 1, 0·2 in other blocks and block sizes of 12×12, 12×1, 12×12, 1×12, 1×1, 1×12, 12×12, 12×1 and 12×12, where blocks are enumerated rowwise. For drawing S2, we also used a 3×3 block matrix of edge probabilities with same block sizes but we input the probability of 0·8 to block 9 instead of block 1. These two stochastic block models split the vertices into high- and low-connectivity groups and the two samples differ only from a block swap.

Scenario C. Non-trivial differences: different diffusion patterns. The goal of this scenario is to go deeper into the interpretation of the Laplacian representation. By analogy with the Laplacian operator that plays a central role in the diffusion equation, we hypothesize that the Laplacian representation captures differences in the way a substance can diffuse along the edges of a network. To verify this claim, we drew S1 from the k-regular model [see 8, sec. 2.4] that generates random networks in which all vertices have the same degree and we drew S2 from the G(n,p) Erdös-Renyi model [23] in which every possible edge is created with the same constant probability. In details,

each vertex in networks from S1 is connected to other 8 (out of 24) vertices while we set the probability for drawing an edge between two arbitrary vertices in S2 to $p = 1/3$ such that the edge weight distribution share the same mean in the two samples. The Laplacian structure should be key to capture differences between the two samples because that difference lies in the diffusion patterns induced by the networks.

Scenario D. Matrix representation versus summary indicators. The goal is to demonstrate that using summary indicators (e.g. clustering coefficient) to compare samples of networks, which is the most popular approach [e.g. 1], could yield less powerful test procedures with respect to using the entire network structures. To this end, we propose to generate small-world networks (characterized by a high clustering coefficient) in both samples and add the scale-free property (power-law degree distribution) to networks in S2. We aim at comparing test procedures based on either clustering coefficient (whose high value characterizes small world networks) or whole network representations, respectively. In details, we drew S1 from the Watts & Strogatz model [73] with starting lattice of dimension 1, size of the neighborhood within which the vertices of the lattice will be connected equal to 4 and rewiring probability of 0·15; and we drew S2 from the Barabási-Albert model [3] with quadratic preferential attachment and 4 edges added at each time step.

The simulated scenarios are summarized in Table 2.2.1. Scopes, models and their parameters for the two samples S1 and S2 are summarized. The Bernoulli rate matrices in scenario C are `p1 = matrix(c(0.8, rep(0.2, 3L)), 2L, 2L)` and `p2 = matrix(c(rep(0.2, 3L), 0.8), 2L, 2L)`.

For each scenario, we computed a Monte-Carlo estimate of the probability of rejection of H_0, which can be interpreted as the power of the test. In all simulations, we set the significance level at $a = 0{\cdot}05$ and we performed a total of 100,000 Monte-Carlo runs. For each run, we performed the test with the statistic $T_{\text{IP}-\text{StudentFisher}}$ using $m = 1{,}000$ permutations sampled with replacement and we estimated the p-value according to Eq. (2.5). For a fair comparison, we used the same samples and the same permutations for each combination of representation and distance.

Figure 2.2.2 reports the estimated power, as the sample size increases and for different combinations of matrix representations and distances between networks. The first column of Fig.

Table 2.2.1: Summary table of the simulated scenarios

Scenario	Scope	S_1	S_2
A	Edge strengths	Poisson model: `lambda = 5`	Poisson model: `lambda = 6`
B	Vertex relabelling	Stochastic block model: `pref.matrix = p1` `block.sizes = c(12L, 1L, 12L)`	Stochastic block model: `pref.matrix = p2` `block.sizes = c(12L, 1L, 12L)`
C	Diffusion patterns	k-regular model: `k = 8L`	Erdös-Rényi model: `p = 1/3`
D	Network VS Indicators	Watts & Strogatz model: `dim = 1L` `nei = 4L` `p = 0.15`	Barabási-Albert model: `power = 2L` `m = 4L` `directed = FALSE`

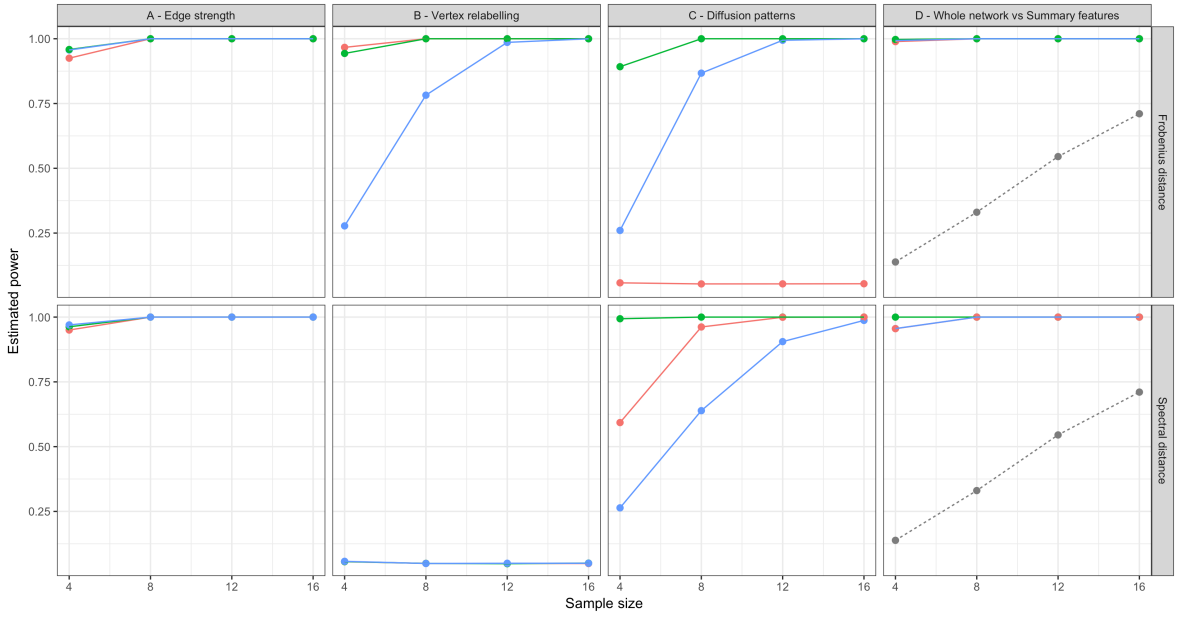


Figure 2.2.2: Power of the test under different representations (adjacency in red, Laplacian in green, modularity in blue), different distances (rows) and different scenarios (columns). The dashed grey curve in Scenario D (last column) represents the statistical power achieved by considering only the clustering coefficient. The largest standard error is 0·00158.

2.2.2 reports estimated probability of rejection for **Scenario A**. It reveals that the power of the test is already close to one for sample sizes as small as $n_1 = n_2 = 4$, despite the fact that the edge weights of the networks in the two samples are drawn from Poisson distributions with close rate parameters. The second column in Fig. 2.2.2 reports results for **Scenario B**. They clearly emphasize that the spectral distance fails to recognize differences for this particular simulated data set, independently from the matrix representations. The spectral distance indeed focuses only on the (ordered) eigenvalues of the matrix representation and therefore it is not sensitive to differences pertaining to vertex relabelling. Fig. 2.2.2 displays the results for **Scenario C** which stress the combined role of representation and distance. First, the test fails to reject the null hypothesis with the Frobenius distance on adjacency matrices for any sample size. This makes sense because the Frobenius distance on the adjacency matrix focuses on differences in edge weight distributions, while samples generated in this scenario differ in the distribution of their nodes. Next, we can see that the power is increasing with the sample size when using the spectral distance on adjacency matrices, reaching values close to 1 from sample sizes as small as 8. This is due to a unique property of the spectrum of adjacency matrix for regular networks that is concentrated on the first eigenvalue equal to k. Finally, tests based on the Laplacian representation succeed in identifying the difference between the two samples, independently from the chosen distance. This is because the feature that discriminates the two samples lies in the fashion a substance can flow through the network, which is exactly what the Laplacian representation captures as shown by the R package `diffusr` [17] that nicely shows that diffusion along the networks is different in the two samples. The fourth column in Fig. 2.2.2 shows that our test is able to distinguish the two samples generated in **Scenario D**. The IP-StudentFisher statistic reaches a statistical power of 1, for sample sizes as small as $n_1 = n_2 = 4$, whereas the same test but based only on the clustering coefficient of the networks goes to 1 with a much lower convergence rate, making it practical only for very large samples. This simulation shows that considering the entire network in the two-sample testing problem allows to achieve a given statistical power with much smaller samples compared to using graph summary measures.

Remark 1. *One may want to use more combinations of representations and distances. This can be done but it is necessary to correct for multiplicity, e.g. by means of Bonferroni-like methods, on the corresponding p-values.*

2.3 Application to bike-sharing data

We chose to demonstrate the usefulness of our approach by applying it to a sharing mobility data set, a case where the test results can be immediately interpreted. Indeed we want to quantitatively answer the question if the sharing mobility shows differences between days of the week. Despite the simplicity of the question, this data presents features which make the parametric approach out of reach: the sample sizes are very small ($n_1 = n_2 = 6$) and the probabilistic generative model of the data is likely to be a mixture distribution accounting for various environmental factors (e.g. precipitation). In the city of Milan a bike sharing service (bikeMi, https://www.bikemi.com) is active since 2008. Milan is divided into 88 neighbourhoods, called Nuclei di Identitá Locale (NILs, http://dati.comune.milano.it/dataset/ds61_infogeo_nil_localizzazione_), and 263 stations are distributed in 39 of these NILs. We are interested in studying the daily bike mobility between the neighbourhoods of the city. Each day is associated to a mobility network which vertices represent neighbourhoods equipped with at least one dock station and edge weights correspond to the number of travels between two neighbourhoods. The data has been collected between January, 25th, 2016 and March, 6th, 2016 where each day starts at 3 a.m. and has been provided by Clear Channel s.r.l.. Since we are interested in the mobility between neighbourhoods, we keep about 300.000 travels of 350.000, excluding travels within the same neighbourhood. In the end, we have a data set of 42 undirected mobility networks (7 days of the week over 6 weeks) to which it is possible to apply all representations and distances presented in the previous sections. Figure 2.3.1 shows a glimpse at the data set by displaying the restricted sample Fréchet means of each day of the week, using the Frobenius distance between Laplacian representations. The colours and the widths of the edges are related to the edge weights: the wider and darker the edge, the larger its weight. We performed pairwise comparisons between days of the week based on samples with sample size $n_1 = n_2 = 6$. The tests have been carried out with the IP-StudentFisher statistic and under all representations and distances discussed in Sections 2.1.1 and 2.1.2. Figure 2.3.2 shows part of the results. In details, the Frobenius distance on adjacency matrix and the spectral distance on Laplacian matrix are considered in the left and right panels, respectively. In the top row, we plotted a multi-dimensional scaling representation of the 42 networks of our data set. Different colours and shapes correspond to different days of the week. The `nevada` package, attached to this work,

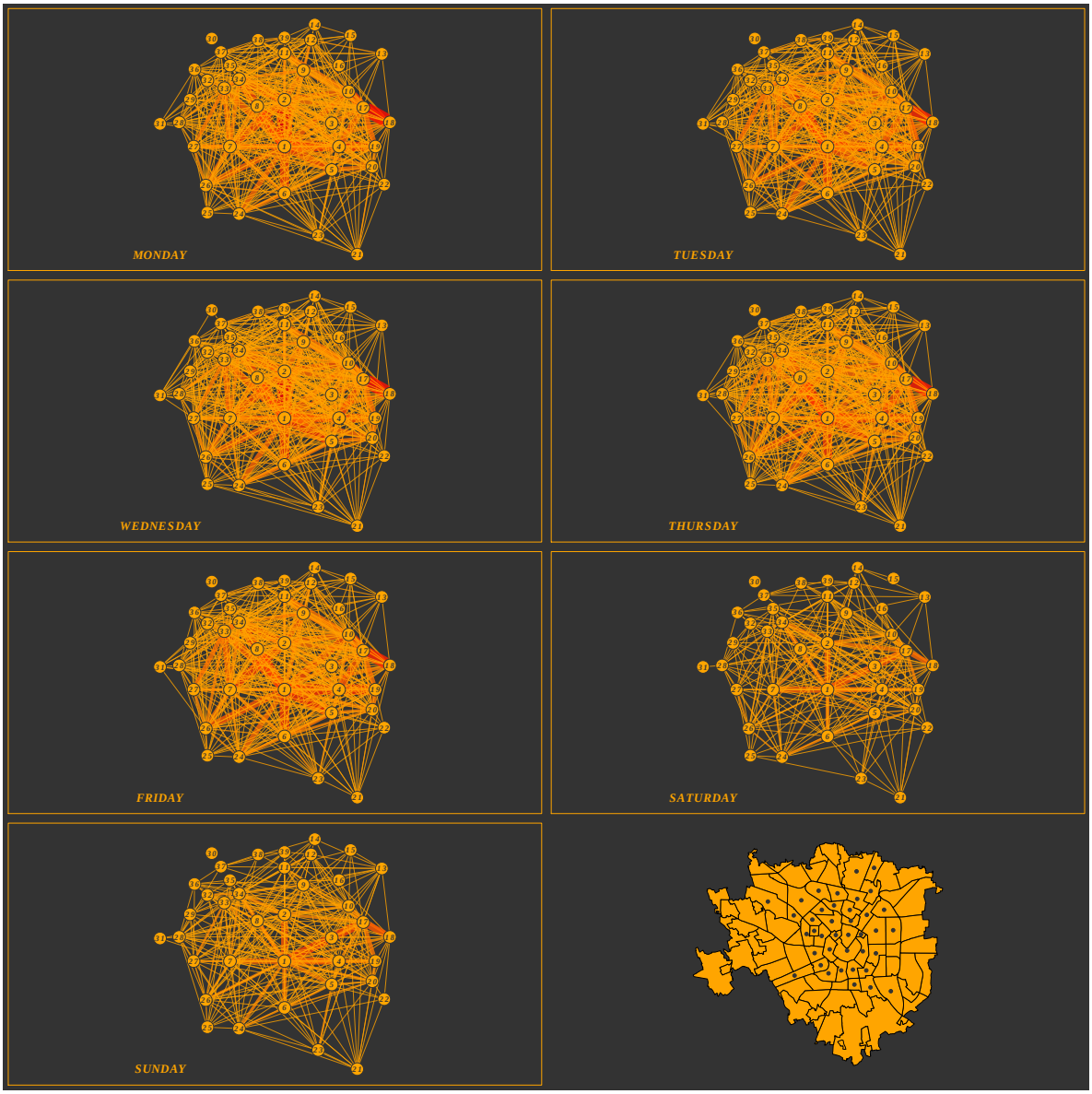


Figure 2.3.1: Restricted sample Fréchet means of each day of the week and, in the last thumbnail (bottom right), the map of the NILs of Milan with a point in the neighbourhoods having at least one dock station.

provides a `plot` function that allows one to visualize multidimensional scaling projections of samples of networks. This is a great supporting tool for picking the best pair of

representation/distance with the scope of highlighting differences between the samples. The second row shows the p-values of each pairwise comparison between different days of the week. The results highlight no significant differences when comparing pairs of week days or Saturday with Sunday. The null hypothesis is instead rejected when comparing week days against weekend days. Results related to the other combinations of representations and distances are similar to those reported in Fig. 2.3.2. These quantitative results are qualitatively visible in both the plots of the entire data set in supplementary material and the multidimensional scaling plots, where there is a separation between working days and non-working days.

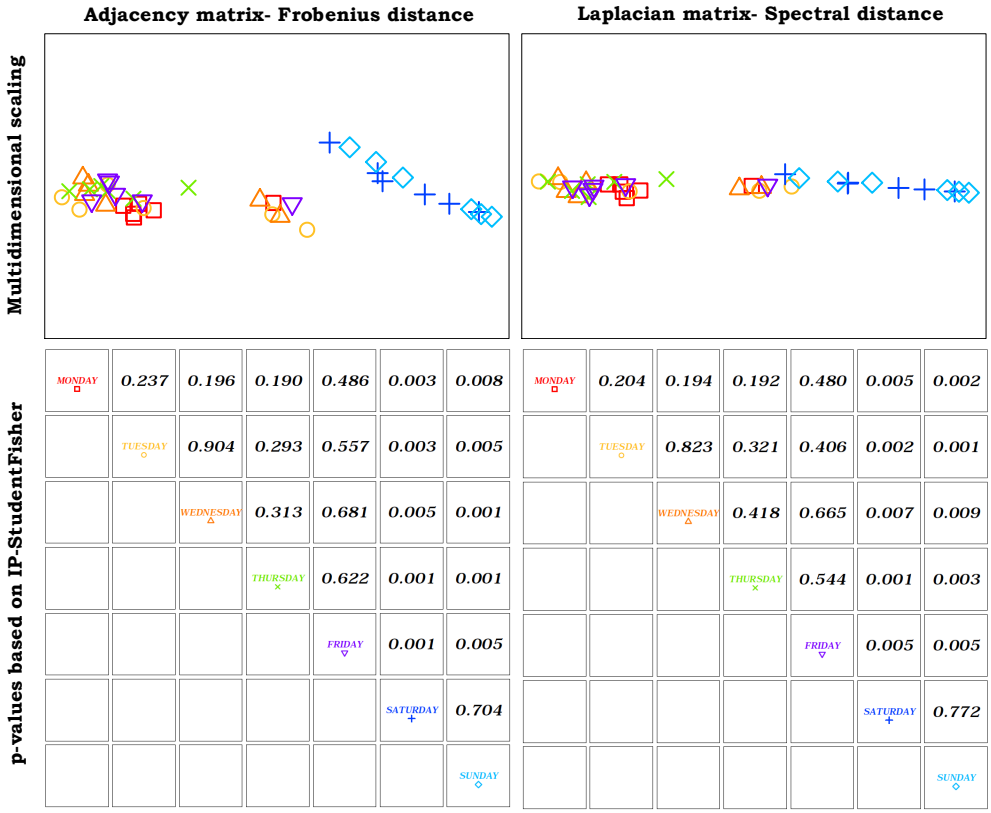


Figure 2.3.2: Results of the application to the bikeMi data set using different matrix representations and distances.

2.4 Discussion

Tackling the two-sample testing problem from the perspective of the permutation framework assumes as null hypothesis that the entire distribution of the two sample is the same (so that, under such an assumption, data in the two samples are exchangeable) while the alternative hypothesis would be that their distribution is different. The choice of the test statistic is then critical because it makes the test sensitive to specific features of the distribution. Therefore, there is no uniformly better statistic for testing equality in distribution but rather many statistic that look at the distribution under different angles. We introduced two statistics which, when combined together through the Non-Parametric Combination methodology, are sensitive to differences in the first two moments of the distributions. A current ongoing work we are pursuing pertains to the definition of statistics sensitive to higher-order moments of the distributions which, when NPC-combined, could make the test sensitive to virtually all moments and thus capture all possible differences.

Starting from standard results on U–statistics, it could be possible to find the asymptotic distributions of $T_{\text{IP}-\text{Student}}$ and $T_{\text{IP}-\text{Fisher}}$. Besides the theoretical interest, the asymptotic distributions might be helpful in reducing the computation time in the case of large sample sizes or large networks. However, permutation tests implemented in our R package `nevada` run a single test within seconds for sample sizes around 20 and networks with 25 nodes.

Furthermore, our proposed method relies only on inter-point distances. This means that all we need is a metric between networks to perform two-sample testing. Hence, we believe that our proposal could be a valid approach not only for network-valued data analysis, but, in a broader context, for Object Oriented Data Analysis, provided that the object data used as sample unit can be embedded into a metric space.

It is not yet established which measures are most appropriate for the analysis of brain networks.

Bullmore & Sporns, 2009

3

An application to brain networks data sets

The structural and functional complexity of a Human brain, captured by time series data or by an image can be explored by means of different techniques. When one is interested in derive some significant evidence from samples of brains, one can treat directly the time series or the images or can construct data set where the statistical unit is a different object. In the latter case, the first step consists in choosing the type of data to extract from the original brain data. It could be a curve, a manifold, a network. Great efforts have been made by the statistical community to establish statistical tools able to deal with samples composed of complex objects. A sound way of represent a brain is by means of a network, where each vertex represents an area of the brain, while an edge can denote either a functional or an anatomical connection.

Among the possible statistical tools that allow to investigate the structure and functions of the Human brain, two-sample test is without any doubt a fundamental extensively used method. The

[1] See [38].

most common approach for two-sample test for samples of brain networks consists in *(i)* selecting a measure that summarises the entire network and *(ii)* statistically comparing the two samples by means of this measure. A large number of brain summary measures have been proposed in the literature to locally and globally characterise a brain: degree, shortest path length, characteristic path length, clustering coefficient, transitivity, efficiency, modularity, etc. Each of these measures captures some aspects of the network (see [58] for a review on their interpretations). The use of a global measure to compare two samples of brain networks entails several drawbacks that we here summarise.

Choice of a summary measure. As mentioned above, a brain network can be characterised by a large number of summary measures. The first issue that one has to tackle when comparing groups of brain networks by means of summary measures pertains to the choice of the measure itself. Indeed, as highlighted in [10], "it is not yet established which measures are most appropriate for the analysis of brain networks". Therefore, what often happens is that a certain number of measures are used and then the choice of the measures to discuss is made a posteriori.

Multiple comparison. The issue of multiple comparisons is directly related to the previous point. In fact, when considering a certain number of brain measures, it might be necessary to correct for multiple comparisons. On the other hand, the correction may require knowing the correlation between the considered measures, making the correction laborious.

Power of the test. Although it cannot be denied the key contribution of network analysis via summary measures on the understanding of the structure and functions of the Human brain, on the other hand these measures have not been specifically designed to maximise the power in group comparisons.

Number of subjects. The considerable problem of the cost of the different acquisition procedures is related to the previous issue. If a test is not designed for maximise the power, a conspicuous number of patients must be involved in the study, leading to very expensive experiments.

In this chapter we will show the potential of the two-sample test introduced in Chapter 2 when applied to data set of brain networks in comparison with the most common approach based on some summary measures. The ultimate goal is an early diagnosis to improve treatments. For this purpose, we first re-generate the four simulated scenarios of Subsection 2.2.2 in the previous chapter with the aim of exploring the estimated power of the test based on five summary measures

and that of the test introduced in the previous chapter; then we consider three data sets previously studied in other papers to re-study them by means of the new test. In detail, each of these data set correspond to a different experimental modality: electroencephalography (EEG), functional MRI (fMRI) and Diffusion Compartment Imaging (DCI). As in the simulation study, for each data set, we propose a comparative analysis involving *(i)* the classical approach of comparing brain networks via summary graph measures and *(ii)* the two-sample test of Chapter 2. In particular, we consider the following graph measures: characteristic path length, global efficiency, clustering coefficient, modularity and small-worldness.

The chapter is organized as follows. In Section 3.1 we briefly review the literature on graph construction and analysis for brain networks. Section 3.2 contains a brief description of the simulated data sets and the results of the simulation study. In Section 3.3 a description of the three real data sets and the results of the inference are reported. Section 3.4 sums up the contributions of the chapter and discuss possible broadening of perspective in treatment and diagnosis.

3.1 Review on construction and analysis of brain connectivity networks

The two main steps in analysis of connectomic data pertain *(i)* building an accurate map of the connectome and *(ii)* analyzing the resulting data. The potential to revolutionize the understanding of the brain organization by means of graph theory is critically dependent upon the validity of the graph representation itself, that involves a non-trivial discretization of the brain into vertices and their interconnecting edges. Moreover, the application of graph theory to brain networks data sets poses several challenges with important implications for how results should be interpreted. In this section we briefly review these two main steps [24].

Construction of a brain connectivity network. The construction of a brain connectivity network involves the definition of vertices and edges able to properly represent brain substructures and their interactions. A first attempt could consist in identifying each neuron with a vertex and a synaptic contact with an edge. This level of resolution for the Human brain is likely unfeasible and there is no clear evidence that it is the most meaningful for understanding structure and function of the Human brain [24]. We here briefly summarize the state-of-the-art pertaining the definition of vertices and edges in Human brain networks [24]:

- *Definition of vertices.* The most common strategies for vertex definition in imaging connectomics are of four types: anatomical, functional, random, and voxel-based. Anatomical and functional parcellations are based on a priori anatomical and functional information, respectively. For example, in a structural brain networks the vertices correspond to anatomically defined regions of histological, MRI or diffusion tensor imaging data; in functional brain networks electroencephalography or multielectrode-array electrodes identify the vertices in the network [10]. Conversely, the random parcellation strategy aims at randomly parcellating the brain into discrete vertices of similar size, while in the voxel-based strategy, each image voxel represents a distinct vertex [24].

- *Definition of edges.* The edges of a brain network are determined by the type of connectivity measured and the method used to quantify it. There are three classes of brain connectivity: structural, functional, and effective. Structural connectivity refers to the anatomical connections between brain regions, derived from diffusion imaging. Functional connectivity pertains to statistical dependencies between spatially distinct neurophysiological recordings, such as coherence measure between two magnetoencephalography sensors. Effective connectivity denotes the causal influence that one neural system exerts over another. See [24] for a detailed review on algorithms, methods, and related challenges and limits of each of these connectivity definitions.

See Figure 3.1.1 for a schematic summary of the construction of structural and functional brain networks.

Analysis of brain connectivity network. Once a proper network representation for the brain has been provided, mathematical tools from graph theory can be used for analyzing brain networks data sets. At this point some issues arise, that we here briefly recall [10, 24]:

- *Multiple comparisons problem.* A difficult multiple comparisons problem is posed by the possible extremely high number of pair-wise interactions between brain regions, that requires the application of methods for controlling the family-wise error rate, such as Bonferroni correction, or other methods that control the false discovery rate, that however perform poorly, especially when the sample sizes are small [24].

- *Graph thresholding.* Spurious or noisy brain connections have to be removed from the data and typically a threshold is applied. As a result, this approach will produce different numbers of edges across different individuals. A common approach to graph thresholding has involved adaptively varying the threshold for each individual to enforce a fixed value of connection density across all participants [24]. However, graph measures depend upon the number of vertices and the connection density and therefore graph measures are often explored over a range of plausible thresholds and connection densities.

- *Reference graphs.* As mentioned, network measures are influenced by the number of vertices, connection density and degree distribution. Therefore, network measures are typically benchmarked against, or normalized to, appropriate null or reference networks that share the same basic properties. The null networks can be typically chosen among random networks, lattice networks, small-world networks. Since each of these networks is appropriate as null model for a particular graph measure, the choice of the null model depends on the network measure, as well as on the connectivity measure used to derive the connectome's edge weights [24].

- *Interpretation of topological measures.* Many graph measures have been developed to study complex systems other than the brain and have been adapted to suit neuroscientific ends. Therefore, their use and interpretation require caution. Moreover, the extent to which each measure provides a meaningful representation of brain function should also be considered. Just to give two examples, path-length based measures are based on assumptions that seem to be unrealistic, and interpretation of variations in clustering coefficient for deriving evidences on local information-processing must account for spatial constraints on connectome architecture [24].

See [24] for a detailed review of these issues and the methods to address them.

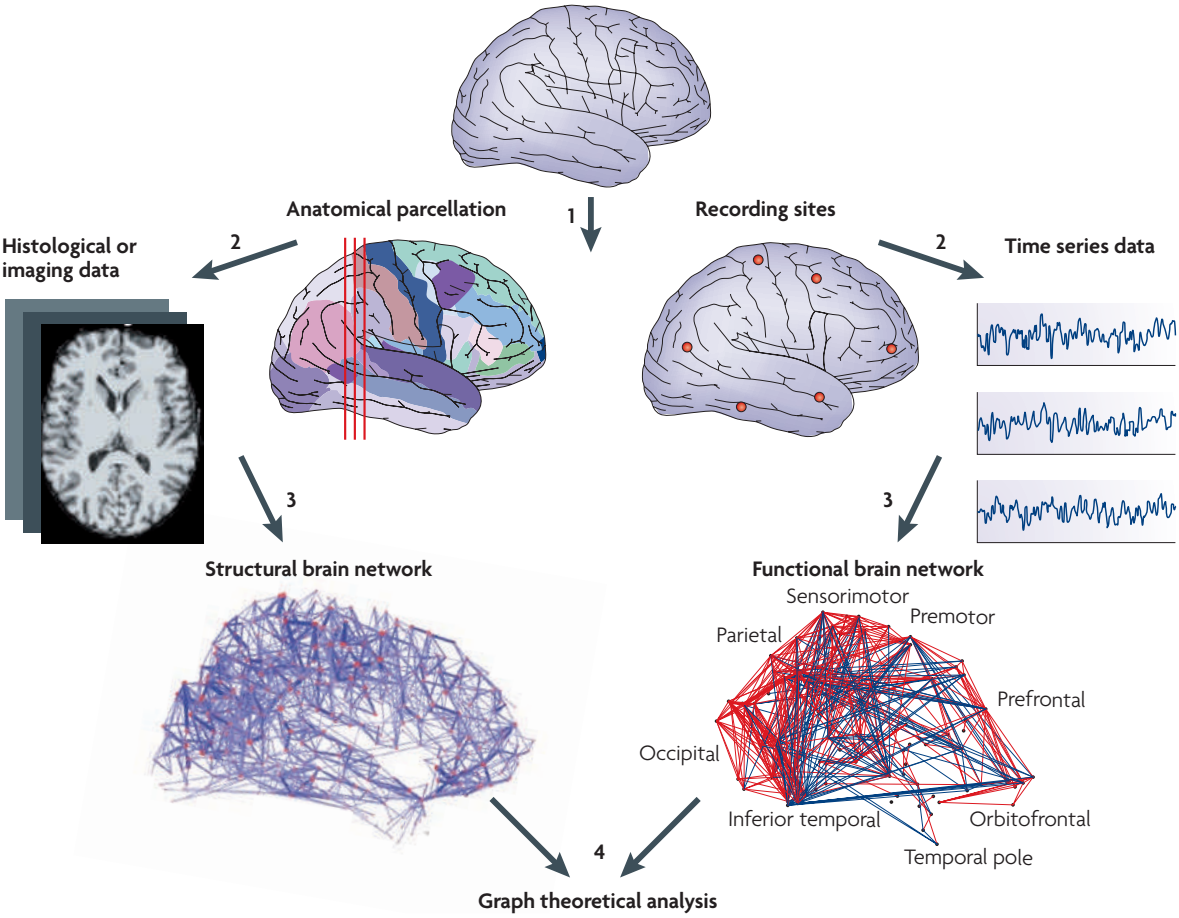


Figure 3.1.1: Construction of structural and functional brain networks. Figure from Bullmore and Sporns, *Complex brain networks: graph theoretical analysis of structural and functional systems*, Nature Review Neuroscience, 2009

3.2 Simulation study

In this section we consider the same simulated scenarios introduced in Subsection 2.2.2 of Chapter 2, that we here briefly describe for self-content. The data sets are generated under the alternative hypothesis and from different generating models. As for the novel test, only one combination of representation and distance is here considered, while the classical approach is explored via five summary measures.

3.2.1 Description of the generated scenarios

Scenario 1. We defined the two samples drawing the edge weight distribution of the first sample from a Poisson distribution with parameter $\lambda = 5$ and the edge weight distribution of the second sample from a Poisson distribution with parameter $\lambda = 6$.

Scenario 2. We drew both the first and the second sample from the stochastic block model [31] with different preference matrices. In details, for drawing the first sample, we used a 3×3 block matrix of edge probabilities with 0.8 in block 1, 0.2 in other blocks while for drawing the second sample, we also used a 3×3 block matrix of edge probabilities with same block sizes but we input the probability of 0.8 to block 9 instead of block 1. These two stochastic block models split the vertices into high- and low-connectivity groups and the two samples differ only from a block swap.

Scenario 3. We drew the first sample from the k-regular model [see 8, sec. 2.4] that generates random networks in which all vertices have the same degree ($k = 8$) and we drew the second sample from the Erdös-Renyi model [23] in which every possible edge is created with the same constant probability ($p = 1/3$).

Scenario 4. We generated small-world networks in both samples and added the scale-free property to networks in the second sample. In details, we drew networks in the first sample from the Watts & Strogatz model [73] with starting lattice of dimension 1, size of the neighborhood within which the vertices of the lattice will be connected equal to 4 and rewiring probability of 0.15; and we drew networks in the second sample from the Barabási-Albert model [3] with quadratic preferential attachment and 4 edges added at each time step.

In each case sampled networks are composed of 25 vertices and four increasing sample sizes are taken into account: $n_1 = n_2 = 4, 8, 12, 16$. In all simulations, we set the significance level at $\alpha = 0.05$ and we estimated the power drawing 1000 simulated data sets.

3.2.2 Results of the simulation study

Figure 3.2.1 reports the estimated power, as the sample size increases, of the test based on the methodology proposed in Chapter 2 (using Laplacian representation and Frobenius distance) and that of the test based on some summary measures. In all the scenarios, the performance of the novel test is always better than that of the classical approach.

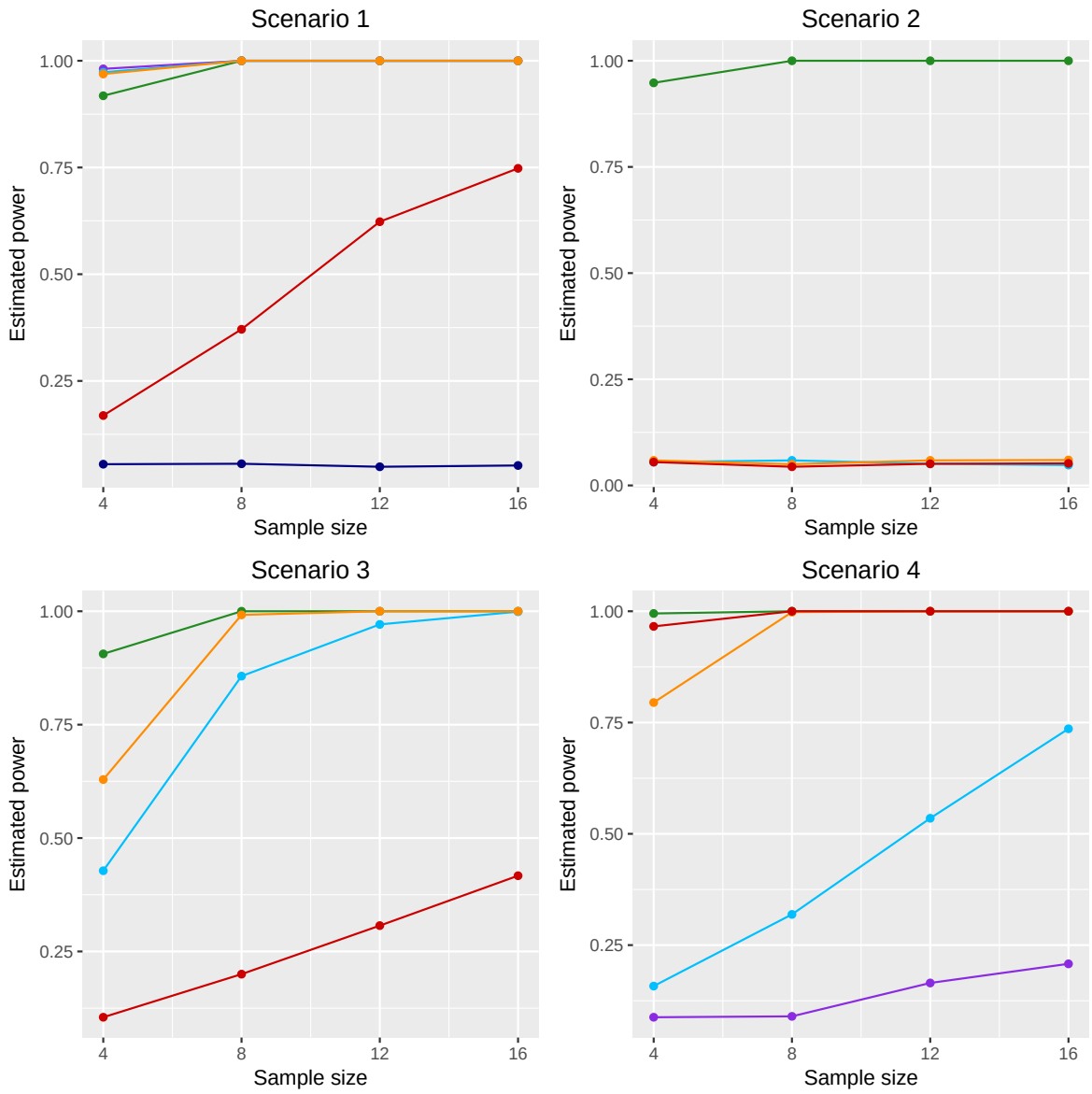


Figure 3.2.1: Power of the test under the novel test (green), characteristic path length (purple), clustering coefficient (light blue), efficiency (orange), modularity (red) and small-worldness (dark blue).

In detail, the top-left block (**Scenario 1**) of Fig. 3.2.1 reveals that the power of the novel test is

already close to one for sample sizes as small as $n_1 = n_2 = 4$. The test based on characteristic path length, clustering coefficient and efficiency has a similar performance as the novel test, while the comparison based on modularity has a low-increasing power. The test that takes into account the small-worldness has power that it is essentially at the nominal level, so it fails in finding differences. The novel test performs very well also in the simulated data sets of **Scenario 2** (top-right block of Fig. 3.2.1) while the test based on transitivity, efficiency and modularity are essentially at the nominal level for all the sample sizes considered. For the networks in this particular data sets, characteristic path length is not meaningfully computed (because there could be disconnected networks) and therefore also small-worldness (whose computing involves the characteristic path length) does not exist. Also in the case of **Scenario 3** (bottom-left block in Fig. 3.2.1), characteristic path length and small-wordlness are not available. The novel test has high power already for sample size equal to 4. The test based on transitivity and that based on efficiency have similar performances: the power is pretty low for sample size $n_1 = n_2 = 4$, while it increases as the sample size increases. The test based on modularity has instead a very low increasing power. In **Scenario 4** (lbottom-right block in Fig. 3.2.1) the novel test and the test based on modularity reveals high power for all the sample sizes. The second-best is the comparison via efficiency while the test based on characteristic path length and transitivity has a low increasing power. For this simulated data sets small-worldness is not available.

In addition to the fact that the novel test performs at least as or always better than the classical approach, this simulation study shows that there is no a summary measure that is better than the others in all the scenarios, remarking the issue of choosing which summary measure to use.

3.3 Population study

In this section we describe the three data sets analysed and the results of the inference.

3.3.1 Description of the data sets

The data sets we analyse in this chapter have been provided by the Computational Radiology Laboratory, Boston Children's Hospital, Harvard Medical School (Boston, MA, USA). The first data set concerns electroencephalographic connectivity. In [52] the authors studied brain functional networks of electroencephalographic connectivity by means of graph theory in order to

investigate syndromic and non-syndromic autism. In detail, the data set is composed of brain networks with a total of 19 vertices from patients with Tuberous Sclerosis Complex (TSC for short) ($n = 29$), patients with Tuberous Sclerosis Complex and Autism Spectrum Disorder (ASD+TSC for short) ($n = 14$), patients with non-syndromic Autism Spectrum Disorder (ASD for short) ($n = 16$) and controls ($n = 13$). See [52] for all the details pertaining the selection of patients, EEG recording and the connectivity measure (i.e. coherence) used to construct the edges in the networks. Three frequency bands have been considered: theta band, lower alpha band and upper alpha band (see [52] for the motivations of this choice). We consider here two group comparisons: ASD vs ASD+TSC and ASD vs controls. The second data set is derived from functional Magnetic Resonance Imaging and it is composed of 31 patients affected by Tuberous Sclerosis Complex and 28 controls. This data set has been previously studied in [69]. Each brain network has 116 vertices and the correlation coefficient as weights. We finally consider a DCI data set that has been analized as part of a TSC Autism Center for Excellence Research Network (TACERN) study [61]. A total of 31 patients affected by Tuberous Sclerosis Complex were included with age between one year and three years. 29 patients come with a score predictive of autism (for the others this score is unknown). The brain has been divided into 134 regions and the edges have been defined by means of different connectivity measures. We here consider Compartmental Fractional Anisotropy (cFA for short), the volume occupied by the streamlines between two regions divided by the total volume of the streamlines (volume ratio for short) and the variance of the fraction of free water in a streamline where the values along the streamlines are weighted with the streamline volume (isoF for short). We consider two different kinds of group comparison. The first one is based on the age of the patients and we compare one-year-old patients with three-years-old patients. The second one aims at comparing patients with respect to the score predictive of autism. In detail, we looked for differences between patients with a low risk (score less or equal to 10) and a high risk (score greater than 10) of autism.

3.3.2 Results of the population study

EEG connectivity data

ASD vs ASD+TSC. Table 3.3.1 contains the p-values obtained comparing ASD patients and ASD+TSC patients in terms of graph measures. No significant differences are revealed by these

results (except in the case of the characteristic path length in the theta band).

	Theta band	Lower alpha band	Upper alpha band
Charact. path length	**0.046**	0.612	0.134
Global efficiency	0.098	0.754	0.168
Clustering coeff.	0.096	0.534	0.096
Modularity	0.678	0.093	0.347
Small-worldness	0.633	0.619	0.057

Table 3.3.1: p-values for the comparison ASD vs TSC+ASD based on five summary graph measures.

Table 3.3.2 contains the results of the new test in the case of theta, lower alpha and upper alpha band, where the significant p-values are in bold (the chosen level is $\alpha = 0.05$). Two different matrix representations (adjacency and laplacian) and three different distances (hamming, frobenius and spectral) are considered. The two-sample test has been conducted looking for differences in mean (loc), variance (scale) and in mean and variance (l+s).

Contrary to the comparison by means of summary measures, except in the case of the laplacian representation in the lower alpha band, the results show significant differences between the two samples under adjacency and laplacian representations and hamming and frobenius distance. On the other hand, regardless of the matrix representation, the tests based on spectral distance lead to non significant p-values.

ASD vs controls. Results of the group comparison based on the summary measures are reported in Table 3.3.3.

Table 3.3.4 contains the results of the test in the case of theta band, lower alpha band and upper alpha band. The same combinations of representations and distances and moment distributions of the other cases are considered.

For the theta and lower alpha band, the p-values are significant under hamming and frobenius distance, regardless of the matrix representation while under spectral distance they are not significant. On the other hand, in the case of upper alpha band the p-values are all significant, even

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.001**	0.326	**0.002**	**0.001**	0.457	**0.002**	0.071	0.331	0.150
Laplacian	**0.002**	0.549	**0.002**	**0.005**	0.851	**0.012**	**0.049**	0.272	0.104

(a) *Theta band*

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.013**	0.704	**0.008**	**0.004**	0.803	**0.004**	0.361	0.790	0.593
Laplacian	0.052	0.664	0.123	0.086	0.706	0.170	0.401	0.896	0.635

(b) *Lower alpha band*

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.001**	0.386	**0.002**	**0.001**	0.454	**0.004**	0.114	0.758	0.207
Laplacian	**0.010**	0.344	**0.019**	**0.024**	0.383	**0.037**	0.119	0.662	0.169

(c) *Upper alpha band*

Table 3.3.2: p-values of the comparison between ASD and ASD+TSC.

	Theta band	Lower alpha band	Upper alpha band
Charact. path length	0.172	0.186	**0.020**
Global efficiency	0.183	0.076	**0.017**
Clustering coeff.	0.173	0.153	**0.009**
Modularity	0.079	**0.004**	0.0554
Small-worldness	0.833	0.969	0.627

Table 3.3.3: p-values for the comparison ASD vs controls based on five summary graph measures.

under spectral distance. The latter may be related to the fact that the upper alpha band is the only case where the three topological measures are able to detect differences in the two groups (see

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.001**	0.265	**0.002**	**0.001**	0.194	**0.002**	0.073	0.628	0.179
Laplacian	**0.001**	0.357	**0.004**	**0.004**	0.405	**0.020**	0.081	0.572	0.179

(a) *Theta band*

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.001**	0.528	**0.004**	**0.001**	0.775	**0.002**	0.064	0.101	0.151
Laplacian	**0.012**	0.258	**0.024**	**0.013**	0.211	**0.026**	0.080	0.073	0.160

(b) *Lower alpha band*

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.001**	0.375	**0.002**	**0.001**	0.647	**0.002**	**0.003**	0.076	**0.005**
Laplacian	**0.002**	0.112	**0.004**	**0.001**	0.089	**0.002**	**0.004**	**0.046**	**0.008**

(c) *Upper alpha band*

Table 3.3.4: p-values of the comparison between ASD and controls.

Table 3.3.3).

fMRI data

As in the previous case, we report the results of the two approaches compared in this chapter. Table 3.3.5 contains the results of the comparison between a sample of TSC patients and a sample of controls. The novel test [39] highlights differences between the two samples under all the combinations of representations and distances.

On the other hand, the test based on the comparison of summary measures (see Table 3.3.6) fails in recognise differences in three over five measures: only global efficiency and clustering coefficient

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.009**	0.213	**0.026**	**0.009**	0.076	**0.018**	**0.010**	0.247	**0.028**
Laplacian	**0.021**	0.706	**0.032**	**0.030**	0.824	**0.044**	**0.016**	0.352	**0.043**

Table 3.3.5: p-values for the comparison between TSC patients and controls for the fMRI data set.

Charact. path length	0.109
Global efficiency	**0.028**
Clustering coeff.	**0.012**
Modularity	0.747
Small-worldness	0.396

Table 3.3.6: p-values for the comparison TSC vs controls based on five summary graph measures for the fMRI data set.

succeed in finding differences.

DCI DATA

1 year old vs 3 years old. In this data set the edges of the brain networks have been defined by means of different connectivity measures. We here compare 1-year-old patients with 3-years-old patients in the case of cFA, volume ratio and isoF. Brain networks constructed by means of the latter connectivity measure have been compared also in terms of the predictive score of autism. The results of the comparison between 1-year-old and 3-years-old patients are displayed in Tables 3.3.7 – 3.3.12. The sample size is 9 for both groups. In all the cases it is possible to observe that the novel test is able to detect differences better than how it is able to do a test based on the summary measures. In detail, in this particular case, the test based on the summary measures suffers of the lack of the characteristic path length due to the great number of zeros in the adjacency matrices. Therefore, the small-worldness is not available in the two samples for all the patients as well. These two facts make impossible a comparison of the two samples based on characteristic path length and small-worldness. Among the other summary measures, only global efficiency in the case of cFA ($p = 0.022$) and clustering coefficient in the case of volume ratio ($p = 0.018$) succeed in finding

differences between the two samples.

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.014**	0.969	**0.030**	**0.025**	0.992	0.064	0.097	0.112	0.188
Laplacian	**0.024**	0.704	**0.049**	**0.032**	0.436	0.054	0.058	0.401	0.123

Table 3.3.7: p-values for the comparison between 1 years patients and 3 years patients for the DCI data set with cFA.

Global efficiency	**0.022**
Clustering coeff.	0.056
Modularity	0.311

Table 3.3.8: p-values for the comparison between 1 years patients and 3 years patients based on five summary graph measures for the DCI data set with cFA.

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.019**	0.966	**0.040**	**0.043**	0.584	0.063	0.117	0.168	0.232
Laplacian	**0.013**	0.987	**0.032**	**0.018**	0.739	**0.030**	**0.040**	0.881	0.066

Table 3.3.9: p-values for the comparison between 1 years patients and 3 years patients for the DCI data set with volume ratio.

Global efficiency	0.117
Clustering coeff.	**0.018**
Modularity	0.622

Table 3.3.10: p-values for the comparison between 1 years patients and 3 years patients based on five summary graph measures for the DCI data set with volume ratio.

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.010**	0.283	**0.034**	**0.017**	0.706	**0.024**	0.076	0.415	0.151
Laplacian	**0.001**	0.670	**0.008**	**0.009**	0.411	**0.014**	**0.007**	0.089	**0.010**

Table 3.3.11: p-values for the comparison between 1 years patients and 3 years patients for the DCI data set with isoF.

Global efficiency	0.475
Clustering coeff.	0.257
Modularity	0.161

Table 3.3.12: p-values for the comparison between 1 years patients and 3 years patients based on five summary graph measures for the DCI data set with isoF.

High risk vs low risk. Also in the case of the comparison between low and high risk of autism patients the novel test (see Table 3.3.13) detects differences where the summary measures don't (see Table 3.3.14). In detail, we consider for this comparison only the isoF connectivity measure. The total number of high risk patients is 9, while the total number of low risk patients is 20. Obviously, as in the 1-year-old and 3-years-old comparison, characteristic path length and small-worldness cannot be used for testing differences. Only clustering coefficient recognised a difference ($p = 0.048$).

	Hamming			Frobenius			Spectral		
	loc	scale	l+s	loc	scale	l+s	loc	scale	l+s
Adjacency	**0.012**	0.126	**0.035**	**0.013**	0.605	**0.042**	**0.048**	0.569	0.100
Laplacian	**0.005**	0.327	**0.012**	**0.004**	0.991	**0.010**	**0.001**	0.669	**0.010**

Table 3.3.13: p-values for the comparison between low risk patients and high risk patients for the DCI data set with isoF.

Global efficiency	0.299
Clustering coeff.	**0.048**
Modularity	0.103

Table 3.3.14: p-values for the comparison between low risk patients and high risk patients based on five summary graph measures for the DCI data set with isoF.

3.4 Discussion

In this chapter we compared two different ways of conducting two-sample tests on brain networks. We consider the classical way of summarise an entire brain network with a brain measure and then test the vectors of this measure [10, 58]; on the other hand, we took into account the novel two-sample test for network-valued data of the previous chapter. We consider brain networks derived from different acquisition procedures (i.e. EEG, fMRI, DCI) and with different number of vertices (19, 116, 134). Both the sample sizes and the criteria that define the samples are different (i.e. the type of disease, the presence/absence of a disease, the age, the risk of autism). In all these different cases the two-sample test for network-valued data performs better than the standard method of comparing brain networks by means of a univariate test involving summary measures. More over, the test allows to overcome the drawbacks explained in the introduction: the issue of choosing which summary measure consider, multiple comparisons correction, the lack of power and the number of subjects involved in the studies. As a result, this new method opens up to possible new discoveries in the field of neuroimaging and to possible improvements in treatment and diagnosis. The results contained in this chapter are now on the attention of some neurologists at the Boston Children's Hospital.

We are able to obtain not only a global p-value, like in traditional tests, but also a p-value for each of the defined aspects or domains. In this way, if we find a significant departure from H_o, we can investigate the nature of this departure in detail.

Brombin & Salmaso, 2009

4

Multiscale null-hypothesis testing for network-valued data

Once it is known that there is a significant difference between two samples of brain networks, physicians may be interested in finding out which portions of the brain are responsible for the observed (global) difference. In many fields where the atom of the statistical analysis is a complex data, a novel trend towards "local" inference is growing. The interest is in looking for which features of the data show statistical significant differences between the two samples. Besides the methodological challenge, computational issues may arise, due to the possibly high number of tests to be performed. In the context of high dimensional data, one of the best known method is that of False Discovery Rate [5], that aims at finding which components of a vector are statistically different between the two samples. In the Functional Data Analysis (FDA)

[1] See [37].

framework great attention has been put on this issue. The interest is in finding which parts of the domain are statistically different between two sample. The most recent proposals are those by [55] and [71]. See [55] for a review of this topic in FDA. In the context of OODA, [9] proposed a generalization of the nonparametric combination methodology that allows to obtain a p-value for each of the defined aspects or domains in shape analysis. With respect to network-valued data, the interest is in finding which edges, vertices or subnetworks are different between the two samples. Both [27] and [20] faced the problem of finding which edges are responsible for a global difference in the two samples. In detail, [27] took advantage of the linear decomposition of the test statistics they introduced to represent the individual contribution of each edge, in a sound asymptotic framework. [20] incorporated in their analyses the multiple local tests where the differences are explored in terms of a bayesian non-parametric approach on each edge, while controlling for multiple comparisons. Because of the modelling formulation of their procedure, the method in [20] is suitable to be applied only to unweighted networks.

In view of the complex structure of the data we have to deal with, we propose a fully non-parametric approach to locally compare samples of networks. Our framework allows to test both weighted and unweighted networks defining a partition of the vertices on which testing differences between samples. In detail, such a partition (that may be inspired by application requests) defines *intra*– and *inter*–subnetworks of interest that consist of connectivity connections inside each element of the partition and between elements of the partition. The major contributions of this chapter are the formulation of a framework that allows to identify subnetworks that exhibit statistical significant differences between the two samples while guaranteeing a finite-sample strong control of the family wise error rate on the subnetworks defined by the partition of the vertices, in a fully non-parametric approach that allows to deal with data generated by complex models.

The chapter is organized as follows. In Section 4.1 the test of hypothesis is detailed. The mathematical framework, the related theoretical properties and two different procedures (i.e. Complete multiscale testing procedure and Adaptive multiscale testing procedure) for local inference for network-valued data are reported in Section 4.2. In Section 4.3 two simulation studies are reported. Finally, we study the EEG data set in Section 4.4.

4.1 The test of hypothesis

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two random samples of networks of cardinality n_1 and n_2, respectively and suppose that every network $G = (V, E)$ in the samples has the same set V of N vertices. We are interested in finding out which subnetworks show statistical significant differences between the two samples. For this reason, the first step consists in identifying subnetworks of interest on which a significance null hypothesis testing will be conducted and whose characteristics are detailed in the following. Subnetworks are exclusively identified by means of the vertices. In detail, the key ingredient is a partition of the set V of the vertices, i.e. a collection $\mathcal{V} = \{V_i\}_{i \in I}$ of subsets of V such that

1. $V = \bigcup_{i \in I} V_i$
2. $V_i \cap V_j = \varnothing \quad \forall i \neq j$

Depending on the origin of the data set and on the research questions, the partition of the set V of vertices may be indicated by an expert. Both the richness of the partition and the possible coarseness/nicety of the partition's elements can be arbitrarily chosen.

Remark 2. *The possibility of choosing a partition of the vertices allows to consider also the two extreme (and opposite) cases. If a vertex splitting is not suggested and therefore all the vertices of the set V belong to the same element of the partition (i.e. the set V itself and so $\mathcal{V} \equiv V$) this local test is reduced to the global test of the second chapter. If each vertex of the network belongs to a different element of the partition (and, as a result, $|I| = N$) the* intra*–subnetworks are not defined and the test of hypothesis is reduced to testing* inter*–differences on every single edge, that is the most common approach.*

Once a partition is selected, two different classes of subnetworks can be identifyed, that we call *intra*–subnetworks and *inter*–subnetworks and we here describe:

***intra*–subnetworks.** An *intra*–subnetwork $G^{intra}_{V_i} = (V_i, E^{intra}_{V_i})$ has the set V_i as set of vertices and the edges in E that have both endpoints in V_i as the set of edges.

***inter*–subnetworks.** An *inter*–subnetwork $G^{inter}_{V_i \cup V_j} = (V_i \cup V_j, E^{intra}_{V_i \cup V_j})$ has the set $V_i \cup V_j$ as set of vertices and the edges in E that have an endpoint in V_i and the other endpoint in V_j as the set of edges.

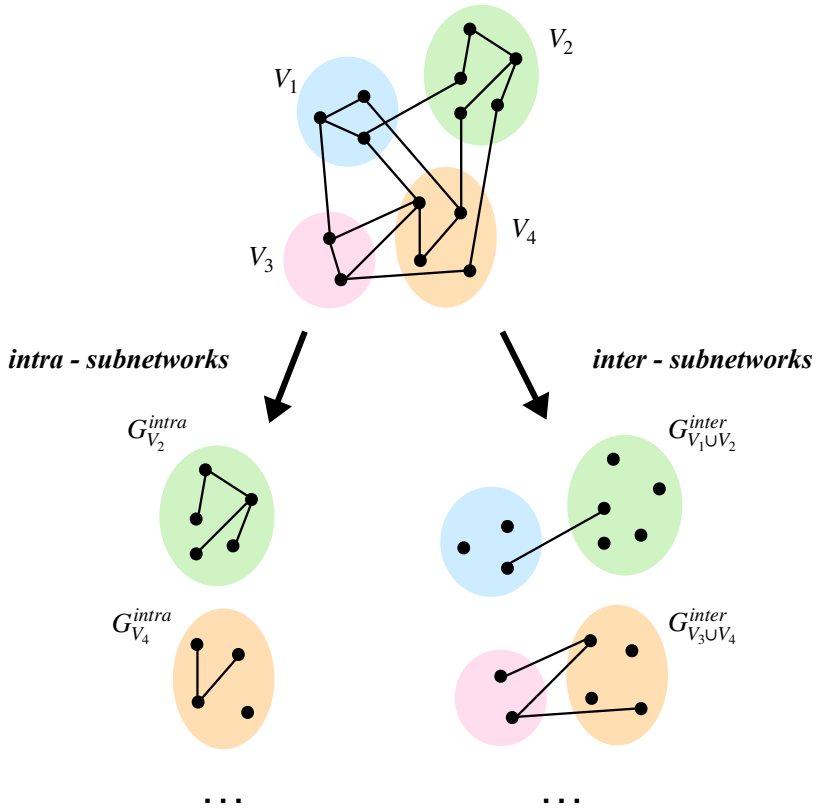


Figure 4.1.1: An example of partition of the vertices of a network in four elements and the corresponding $G^{intra}_{V_i}$ for $i = 2, 4$ on the left and $G^{inter}_{V_i \cup V_j}$ for $(i, j) = (1, 2)$ and $(3, 4)$ on the right.

Figure 4.1.1 illustrates an example of a partition of the set of vertices of a network in four subsets. Two *intra*–subnetworks $G^{intra}_{V_2}$ and $G^{intra}_{V_4}$ (over the four possible ones) and two *inter*–subnetworks $G^{inter}_{V_1 \cup V_2}$ and $G^{inter}_{V_3 \cup V_4}$ (over the six possible ones) are reported to clarify the definitions.

The fundamental objects on which we define our test of hypotheses are precisely these subnetworks just defined. Indeed our aim is to conduct a null hypothesis significant testing to figure out *(i)* if the global difference between the two samples is due to some subnetworks rather than others and *(ii)* if the (possible) "local" differences between the two samples lie in the *intra*–subnetworks or in the *inter*–subnetworks (and therefore to find out if an *intra*–difference or

an *inter*–difference, or both, is present, respectively).
Coherently to these concepts, we introduce the following two families of tests of hypotheses:

***intra* test of hypothesis.** In relation to the *intra*–differences, the subnetworks involved in the test are the *intra*–subnetworks. Let $\mathbf{F}_1^i$ and $\mathbf{F}_2^i$ be the distributions that govern, in the two populations respectively, the *intra*–subnetworks defined by the element V_i of the partition $\mathcal{V}$. For all $i = 1, \cdots, |I|$ we want to test the following:

$$H_0^{intra,i} : \mathbf{F}_1^{G_{V_i}} = \mathbf{F}_2^{G_{V_i}} \quad \text{against} \quad H_1^{intra,i} : \mathbf{F}_1^{G_{V_i}} \neq \mathbf{F}_2^{G_{V_i}}. \tag{4.1}$$

Therefore, the number of hypotheses we are interested in is $|I|$.

***inter* test of hypothesis.** Switching to the "*inter*–differences", we want to test hypotheses on the the *inter*–subnetworks. Let $\mathbf{F}_1^{ij}$ and $\mathbf{F}_2^{ij}$ be the distributions that govern, in the two populations respectively, the *inter*–subnetworks defined by the union of the elements V_i and V_j of the partition $\mathcal{V}$. For all $i \neq j = 1, \cdots, |I|$ we want to test the following:

$$H_0^{inter,ij} : \mathbf{F}_1^{G_{V_i \cup V_j}} = \mathbf{F}_2^{G_{V_i \cup V_j}} \quad \text{against} \quad H_1^{inter,ij} : \mathbf{F}_1^{G_{V_i \cup V_j}} \neq \mathbf{F}_2^{G_{V_i \cup V_j}}. \tag{4.2}$$

The number of hypotheses that are tested is $\frac{|I| \cdot |I-1|}{2}$.

Testing this collection of hypotheses on network-valued data entails the facing of two main difficulties. First, the number of hypotheses that are simultaneously tested could be very high and therefore, a testing procedure able to control the probability of a Type I error is needed. The challenge is both methodological and computational (see Section 4.2). Second, the statistical unit is a complex object and, in particular, each single test is a test on network-valued data. This requires the application of inferential tools for this kind of object data.

4.2 Methods

4.2.1 Mathematical framework

Besides the control on each single tested hypothesis, in this work, as in many high dimensional application, we want to guarantee the control of the family wise error rate at a global level α, that is $\mathbb{P}[\text{at least one I type error}] < \alpha$. Different methods assuring such a control on a family of tests have

been proposed in the literature. One of the most flexible is the Closed Testing Procedure. This well-known methodology has been introduced for the first time by [42]. With this method a null hypothesis H_o is rejected if all the possible intersection hypotheses that involve H_o are rejected by an α level test. This procedure guarantees that the probability of making no type I error is at least $1 - \alpha$ [42]. The Closed Testing Procedure is therefore composed of an hierarchy of auxiliary hypotheses (i.e. all the possible intersections of the null hypotheses to be tested) with the hypotheses one is interested in at the bottom of this hierarchy. In our specific application, the latter are hypotheses on the *intra*–subnetworks (see test (4.1)) and on the *inter*–subnetworks (see test (4.2)). At the bottom of the hierarchy there are therefore these $|I| + \binom{|I|}{2}$ hypotheses. At this point, we need to clarify how the intermediate hypotheses in the hierarchy are defined in the case of network-valued data. For this purpose, we first generalize the *intra*–subnetworks and the *inter*–subnetworks (previously defined for the elements of the partition) to subnetworks with a different set of vertices. To define these new extended families of subnetworks we make use of the σ–algebra generated by the partition of the vertices. Recall that a σ-algebra over a non empty set $\mathcal{V}$ is a family of subsets of $\mathcal{V}$ that includes $\mathcal{V}$, is closed under complement and is closed under countable unions. If $K \subseteq \mathcal{P}(V)$ (with $\mathcal{P}(V)$ the power set) is a non-empty family of sets over V (i.e. a subset of $\mathcal{P}(V)$) the σ-algebra $\sigma(K)$ generated by K is the smallest σ-algebra that contains K, that is the intersection of all the σ-algebras that contain K. If $\{V_i\}_{i\in I}$ is a finite or countable partition of V, $\sigma(\{V_i\}_{i\in I}) = \{\bigcup_{j\in J} V_j, J \subseteq I\}$. Consider the σ-algebra generated by the partition $\mathcal{V} = \{V_i\}_{i\in I}$ of V and fix an element $A \in \sigma(\mathcal{V})$. Hence A is the union of a certain number of subsets V_i of the partition $\mathcal{V}$. Let us define a concept of dimensionality for such an element A. We say that $A \in \sigma(\mathcal{V})$ *has dim* d if d is the number of subsets V_i of $\mathcal{V}$ whose union constitutes A and we introduce the following notation: $dim(A) = d$.
We can now introduce three different classes of networks that will be used to define the auxiliary hypotheses. What is common between these three networks is the set of the vertices: given an element $A \in \sigma(\mathcal{V})$, all the vertices in A constitute the set of vertices in all the three classes of networks and we refer to it as V_A; what changes is the set of the edges. We define the following:

- for all $A \in \sigma(\mathcal{V})$, $G_A^{total} = (A, E_A^{total})$ is the network with edges the edges in E that have the endpoints in any of the subsets in the partition $\mathcal{V}$ contained in A (i.e., in terminology of graph theory, the subgraph induced in G by A [see 16, Chapt. 1]). See the example in the top right of Fig. 4.2.1.

- for all $A \in \sigma(\mathcal{V})$, $G_A^{intra} = (A, E_A^{intra})$ is the network with edges the edges in E that have both endpoints in the same subset in the partition $\mathcal{V}$ contained in A. A network of this type contains only the edges that do not exit each single subset V_i in A. See the example in the bottom left of Fig. 4.2.1.

- for all $A \in \sigma(\mathcal{V})$, $G_A^{inter} = (A, E_A^{inter})$ is the network with edges the edges in E that have the endpoints in two different subsets in the partition $\mathcal{V}$ contained in A. This type of network contains only the edges that connect vertices belonging to two different subsets $V_i \neq V_j$ in A. See the example in the bottom rigth of Fig. 4.2.1.

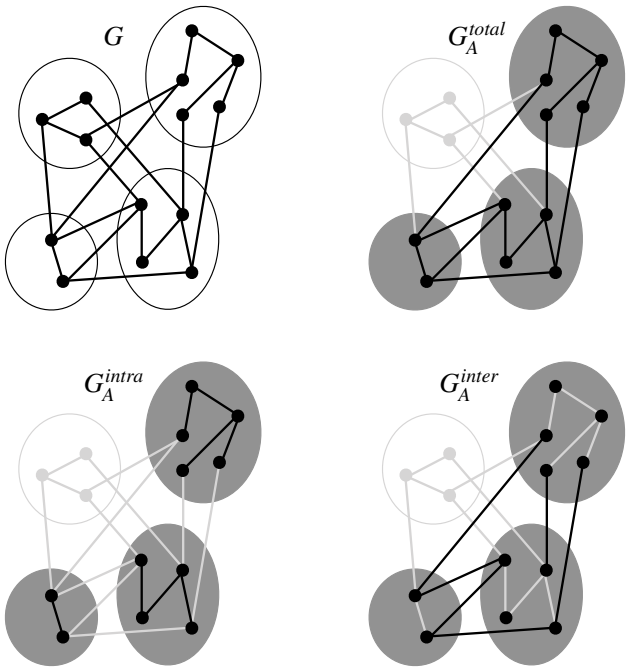


Figure 4.2.1: An example of partition of the vertices of a network in four elements (first block). In the other three blocks an element A with $dimens(A) = 3$ is highlighted in grey and the corresponding G_A^{total}, G_A^{intra} and G_A^{inter} are reported. In each case, the vertices and the edges that actually constitute the subnetworks are in black.

Remark 3. *If A is an element of the partition, that is $A \equiv V_i$ for some $i = 1, \cdots, |I|$, G_A^{inter} is not defined and $G_A^{intra} \equiv G_A^{total}$.*

Before defining the auxiliary tests of hypothesis, we need to introduce the following families of subnetworks:

$$\mathcal{G}^{total} := \left\{ G_A^{total} : A \in \sigma(\mathcal{V}) \right\} \tag{4.3}$$

$$\mathcal{G}^{intra} := \left\{ G_A^{intra} : A \in \sigma(\mathcal{V}) \right\} \tag{4.4}$$

$$\mathcal{G}^{inter} := \left\{ G_A^{inter} : A \in \sigma(\mathcal{V}), dim(A) > 1 \right\}. \tag{4.5}$$

The procedure we introduce in this chapter tests the null hypothesis against the alternative on all the subnetworks in the previous families. We now define three auxiliary tests.
In the following, $\mathbf{F}_1^A$ and $\mathbf{F}_2^A$ are the distribution functions that generate in the two samples the subnetworks in the family under scrutiny (i.e. $\mathcal{G}^{total}, \mathcal{G}^{intra}, \mathcal{G}^{inter}$).
The first auxiliary test doesn't distinguish between the *intra*–differences and the *inter*–differences. It tests the hypothesis on elements of the family $\mathcal{G}^{total}$ to see where the differences between the two samples are. The test is the following:

$$H_0^{total,A} : \mathbf{F}_1^{G_A^{total}} = \mathbf{F}_2^{G_A^{total}} \quad \text{against} \quad H_1^{total,A} : \mathbf{F}_1^{G_A^{total}} \neq \mathbf{F}_2^{G_A^{total}}, \tag{4.6}$$

where A is an element of σ–algebra $\sigma(\mathcal{V})$. The corresponding p-value is p_A^{total}.
The second auxiliary test is focus on possible *intra*–differences and therefore is performed on the elements of the family $\mathcal{G}^{intra}$:

$$H_0^{intra,A} : \mathbf{F}_1^{G_A^{intra}} = \mathbf{F}_2^{G_A^{intra}} \quad \text{against} \quad H_1^{intra,A} : \mathbf{F}_1^{G_A^{intra}} \neq \mathbf{F}_2^{G_A^{intra}}, \tag{4.7}$$

where A is an element of σ–algebra $\sigma(\mathcal{V})$. The corresponding p-value is p_A^{intra}. This test is analogs of test (4.1) for a general element of the σ–algebra $\sigma(\mathcal{V})$ instead for the elements of the partition $\mathcal{V}$.
The third "auxiliary" test puts the attention on possible *inter*–differences and therefore is performed on the elements of the family $\mathcal{G}^{inter}$:

$$H_0^{inter,A} : \mathbf{F}_1^{G_A^{inter}} = \mathbf{F}_2^{G_A^{inter}} \quad \text{against} \quad H_1^{inter,A} : \mathbf{F}_1^{G_A^{inter}} \neq \mathbf{F}_2^{G_A^{inter}}, \tag{4.8}$$

where A is an element of σ–algebra $\sigma(\mathcal{V})$ such that $dimens(A) > 1$. The corresponding p-value is p_A^{inter}. This test is analogs of test (4.2) for a general element of the σ–algebra $\sigma(\mathcal{V})$ instead for the

elements of the partition $\mathcal{V}$.

Remark 4. *In some applications it may happen that some vertices belong to more than one element of the subdivision of the vertices. In our approach, this is not a drawback. Indeed it is possible to redefine the given subdivision in order to have a partition where the overlapping vertices are elements of the partition itself and then construct the σ-algebra as usual. In detail, the σ-algebra generated by a family $\mathcal{B}$ of sets when $\mathcal{B}$ is not necessarily a partition can be defined. Let be Ω a non empty set and define $\mathcal{B} := \{B_i, 1 \leq i \leq k < \infty\} \subset \mathcal{P}(\Omega)$. Then, the σ-algebra $\sigma(\mathcal{B})$ generated by $\mathcal{B}$ is given by*

$$\sigma(\mathcal{B}) = \left\{ E : E = \bigcup_{\delta \in J} B_\delta, J \subset \{1, 2, \cdots, k\} \right\}$$

where for each $\delta = (\delta_1, \delta_2, \cdots, \delta_k)$, $\delta_i \in \{0, 1\}$, B_δ is defined as $B_\delta = \bigcap_{i=1}^{k} B_i(\delta_i)$, where $B_i(0) = B_i^c$ and $B_i(1) = B_i$, $i \geq 1$.

4.2.2 Complete multiscale testing procedure

Given the general Closed Testing Procedure and the auxiliary hypotheses (4.6), (4.7) and (4.8) (Subsection 4.2.1), we now describe what we here call *Complete multiscale testing procedure*. Recall that we have two samples $\mathcal{G}_1$ and $\mathcal{G}_2$ of networks of cardinality n_1 and n_2, respectively and that each network in $\mathcal{G}_1$ and $\mathcal{G}_2$ has the same set of N vertices. Define $m := 2^{|I|}$, the cardinality of the σ–algebra $\sigma(\{V_i\}_{i \in I}) = \sigma(\mathcal{V})$ generated by the partition $\{V_i\}_{i \in I} = \mathcal{V}$.
The Complete multiscale testing procedure we propose tests the null hypothesis of no differences in the generating distributions against the alternative, on gradually smaller subnetworks defined on the elements of the σ–algebra $\sigma(\mathcal{V})$. The description of the procedure is the following:
$\forall i = |I|, |I| - 1, \cdots, 2, 1$ and $\forall A \in \sigma(\mathcal{V})$ such that $dim(A) = i$ perform the tests (4.6), (4.7) and (4.8).
This procedure allows to define an adjusted p-value for the *intra*–subnetworks and for the *inter*–subnetworks defined according to the partition of the set of vertices.
The adjusted p-value for the *intra*–subnetworks (test (4.1)) is defined as

$$p_{V_i} := \max_{A: V_i \in A} p_A^{intra}, p_A^{total}$$

whereas the adjusted p-value for the *inter*–subnetworks (test (4.2)) is defined as

$$p_{V_iV_j} := \max_{A:V_i,V_j\in A} p_A^{inter}, p_A^{total}$$

The Complete multiscale testing procedure is summarized in Algorithm 1.

Algorithm 1 Complete multiscale testing procedure

1: **procedure**
2: *loop*:
3: $p^A_{total} \leftarrow$ local *total* two-sample test on G^{total}_A
4: $p^A_{intra} \leftarrow$ local *intra* two-sample test G^{intra}_A
5: $p^A_{inter} \leftarrow$ local *inter* two-sample test G^{inter}_A
6: **goto** *loop*
7: computing of the adjusted p-values

Lemma 1. *The procedure* Complete multiscale testing procedure *guarantees the strong control of the family wise error rate on the following set (see* (4.3), (4.4) *and* (4.5)*):*

$$\mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter},$$

that is, if $G \in \mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter}$ *is the larger subnetwork where* H_o *is true,*

$$\mathbb{P}[\exists G^{intra}_{V_i} \subseteq G : p_{V_i} \leq \alpha \quad or \quad \exists G^{inter}_{V_i\cup V_j} \subseteq G : p_{V_i\cup V_j} \leq \alpha] \leq \alpha,$$

where with the notation $G_1 \subseteq G_2$ *we mean the* G_1 *is a subnetwork of* G_2*.*

Proof. Let A be the set of vertices of G and denote with H^G_o the hypothesis among all the $H^{total,A}_o$, $H^{intra,A}_o$ and $H^{inter,A}_o$ that refers to G and p_G the corresponding p-value. Thanks to the structure of the Multiscale testing procedure, an hypothesis $H^{intra,i}_o$ ($H^{inter,i,j}_o$) is rejected only if all the corresponding $H^{total,A}_o$ and $H^{intra,A}_o$ ($H^{total,A}_o$ and $H^{inter,A}_o$) are rejected too. H^G_o corresponds to one of these hypotheses. Each single hypothesis is tested at level α and in particular this is true for H^G_o. Therefore, $\mathbb{P}[p_{V_i} \leq \alpha] \leq \mathbb{P}[p_G \leq \alpha] = \alpha$ ($\mathbb{P}[p_{V_i\cup V_j} \leq \alpha] \leq \mathbb{P}[p_G \leq \alpha] = \alpha$). □

Remark 5. *Lemma 1 guarantees the control of the family wise error rate on the larger subnetwork G in $\mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter}$ where the null hypothesis is true. It is clear from the proof that this type of control is guaranteed on all the subnetworks G in $\mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter}$ where the null hypothesis is true.*

Remark 6. *As stated in Lemma 1, the control of the family wise error rate is guaranteed on subnetworks belonging to the families $\mathcal{G}^{total}$, $\mathcal{G}^{intra}$ and $\mathcal{G}^{inter}$. These families are induced by the partition of the vertices suggested by the user. On the other hand, the control of the family wise error rate is not guaranteed on subnetworks that are not included in the previous families. The type of family wise error rate control is coherent with the initial choice of the partition.*

4.2.3 Adaptive multiscale testing procedure

In the previous subsection we briefly described the well-known Closed testing procedure and we introduced tools to adapt this method to perform multiscale null hypothesis testing on network-valued data. Both in the Closed testing procedure and in the Complete multiscale testing procedure, the number of hypotheses to be tested might be very high, leading to remarkable computational costs. In addition to the possible high computational cost, we want to highlight another possible limit. If an α level has been fixed prior to the beginning of the closed testing procedure and if one is looking only for those null hypotheses that will be rejected by the procedure, in many cases it can happen that lot of auxiliary hypotheses are tested uselessly. In fact, if the procedure starts from the top of the hierarchy of the auxiliary hypotheses, at every step if an hypothesis is not rejected, it is not necessary to test the hypotheses included in this one. Consider for example the set of hypothesis in Figure 4.2.2. If the hypothesis are tested starting from the top of the hierarchy and if an α level of, for example, 0.05 has been fixed, once one figures out that the hypothesis $H_0^{2,3}$ is not rejected, it is not necessary to test H_0^2 and H_0^3 because for sure they will not be rejected.

We therefore propose a general alternative to the classical Closed Testing Procedure, i.e. what we here call *Adaptive Closed Testing Procedure*. The peculiarity of this method consists, at each step, in selecting the hypotheses that are suitable to be tested. Similarly to the classical Closed Testing Procedure, an hierarchy of hypothesis is introduced, with the global hypothesis at the top. Given an α level, if a significant difference is found at the global level, the hypothesis of the subsequent level of the hierarchy are tested. Then the hypothesis of the third level that are contained in non-rejected

hypothesis are not tested in the successive step and so on. An adjust p-value for each hypothesis ω_β is therefore defined as the maximum of the set of p-values pertaining to hypothesis containing ω_β.

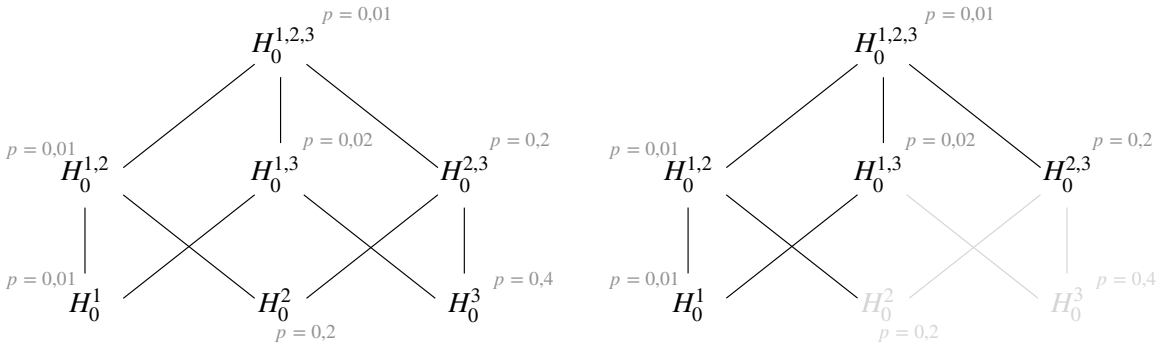


Figure 4.2.2: An example of closed testing procedure (first block) and of adaptive closed testing procedure (second block). The hypothesis that are subject to be tested are in black.

As in the previous subsection, we generalize this procedure to the case of network-valued data. The method we introduce here and that we call *Adaptive multiscale testing procedure* tests the null hypothesis against the alternative not on all the gradually smaller subnetworks, but selecting which of them are suitable to be tested. The aim is to avoid testing hypothesis on subnetworks on which the test on the upper level has returned a non significant p-value. Precisely, if the test on a portion of the network is not significant, then the hypothesis is not tested in the possible subnetworks of that portion. This criteria is applied also to the three families of auxiliary tests defined in (4.6), (4.7) and (4.8). In particular, if the global test returns a non significant p-value, the method doesn't go on. If the *global*–p-value is significant, the global–*intra* test and the global–*inter* test are performed. If there is not a global–*intra*–difference (global–*inter*–difference), all the subsequent *intra* tests (*inter* tests) are not performed. This concepts are formalized and clarified in the following paragraph.

In the description of the Adaptive multiscale testing procedure we make use of the auxiliary tests 4.6, 4.7 and 4.8 introduced in Section 4.2. The difference with the Complete multiscale testing procedure when referring to the auxiliary tests consists in the subnetworks involved in the tests. After fixing an α level, as first step a global test on $G \equiv G_V^{total}$ is required to check if the two samples

are different. If that is the case, the method proceeds looking for the nature of the difference: it might be due to *intra*–differences or *inter*–differences or both. Therefore, a global test on G_V^{intra} and on G_V^{inter} is perfomed, i.e. the tests 4.7 and 4.8 are performed for the element in the σ–algebra with *dimension* equal to $|I|$. The procedure now goes on only for the type of connection (i.e. *intra*–connections or *inter*–connections) where there is a difference. In detail, if the p-value resulting from testing (4.7) is significant (i.e. that referred to a global *intra*–difference), tests (4.6) and (4.7) are performed for elements A with $dimens(A) = |I| - 1$. At this point, before proceeding to the subnetworks referring to A such that $dimens(A) = |I| - 2$, it is necessary to evaluate on which subnetworks it is suitable to conduct the test. Subnetworks are constructed based on the elements of the σ–algebra $\sigma(\mathcal{V})$, so let's focus on these elements to describe how to choose the subnetworks to test. At the end of the first step, $|I|$ tests have been performed, each based on an element A of *dimension* $|I| - 1$ belonging to the σ–algebra $\sigma(\mathcal{V})$. Those elements of *dimension* $|I| - 2$ that are subsets of elements of *dimension* $|I| - 1$ for which the test on the corresponding subnetworks is not significant, are not considered for the test in the next step. The subnetworks that are identified as suitable to be tested are involved in tests (4.6) and (4.7). The procedure goes on in this way for all the subsequent steps, until the hypothesis pertaining to the elements of the σ–algebra of *dimension* 1 (i.e. the test (4.1)) are tested. The procedure for the *inter*–differences is analogue to that for the *intra*–differences. Similarly, if p-value resulting from testing (4.8) on the element of *dimension* $|I|$ (i.e. the global test on G_V^{inter}) is significant, the same steps just described, but for the auxiliary hypotheses (4.6) and (4.8), are carried out until the hypothesis pertaining to the elements of the σ–algebra of *dimension* 2 (i.e. the tests (4.2)) are tested.
The adjusted p-values for the *intra*– and *inter*–subnetworks are defined as in the case of the Multiscale testing procedure where the subnetworks involved are only those on which the tests have been performed:

$$\begin{aligned} \tilde{p}_{V_i} &:= \max_{A:V_i \in A} p_A^{intra}, p_A^{total} \\ \tilde{p}_{V_i V_j} &:= \max_{A:V_i, V_j \in A} p_A^{inter}, p_A^{total} \end{aligned} \tag{4.9}$$

for the G_A^{total}, G_A^{intra} and G_A^{inter} on which the test has been performed.

Algorithm 2 summarizes the Adaptive multiscale testing procedure. The word *suitable* used in the algorithm stands for those elements A of the σ–algebra generated by the partition $\mathcal{V}$ that at each step of the Adaptive multiscale testing procedure are identified as subsets to consider to construct

(and then test on) the corresponding *intra*– and/or *inter*–subnetworks.

Algorithm 2 Adaptive multiscale testing procedure

procedure
  input level α
  $p \leftarrow$ global two-sample test
  if $p > \alpha$ **then return** false
  $p_{intra} \leftarrow$ global *intra* two-sample test
  $p_{inter} \leftarrow$ global *inter* two-sample test
  if $p_{intra} < \alpha$ **then**
    loop:
    $p^A_{total} \leftarrow$ local *total* two-sample test on G^{total}_A for suitable A
    $p^A_{intra} \leftarrow$ local *intra* two-sample test G^{intra}_A for suitable A
    goto *loop*
  if $p_{inter} < \alpha$ **then**
    loop:
    $p^A_{total} \leftarrow$ local *total* two-sample test on G^{total}_A for suitable A
    $p^A_{inter} \leftarrow$ local *inter* two-sample test G^{inter}_A for suitable A
    goto *loop*
    computing of the adjusted p-values

Remark 7. *At the same fixed α level, the Complete multiscale testing procedure and the Adaptive multiscale testing procedure identify the same sets of significant and non significant subnetworks. If a subnetwork is statistically different between the two samples, the p-value found with the Complete multiscale testing procedure and that with the Adaptive multiscale testing procedure are exactly the same ($\tilde{p}_{V_i} = p_{V_i}$ or $\tilde{p}_{V_iV_j} = p_{V_iV_j}$); if a subnetwork is not statistically different, the p-value found with the Complete multiscale testing procedure is larger or equal than the results obtained from the Adaptive multiscale testing procedure ($\tilde{p}_{V_i} \leq p_{V_i}$ or $\tilde{p}_{V_iV_j} \leq p_{V_iV_j}$).*

Lemma 2. *The Adaptive multiscale testing procedure guarantees the control of the family wise error rate on the following set:*

$$\mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter},$$

that is, if $G \in \mathcal{G}^{total} \cup \mathcal{G}^{intra} \cup \mathcal{G}^{inter}$ is the larger subnetwork where H_0 is true,

$$\mathbb{P}[\exists G^{intra}_{V_i} \subseteq G : \tilde{p}_{V_i} \leq \alpha \quad or \quad \exists G^{inter}_{V_i \cup V_j} \subseteq G : \tilde{p}_{V_i \cup V_j} \leq \alpha] \leq \alpha,$$

where with the notation $G_1 \subseteq G_2$ we mean the G_1 is a subnetwork of G_2.

Proof. From Remark 7, it follows that, fixed an α level, $\tilde{p}_{V_i} \leq \alpha$ if and only if $p_{V_i} < \alpha$ and $\tilde{p}_{V_i V_j} \leq \alpha$ if and only if $p_{V_i V_j} < \alpha$ The proof now follows from that of Lemma 1. □

Remark 8. *As stated in the description of the Adaptive multiscale testing procedure, at the beginning it is necessary to fix an α level and each step of the procedure depends upon the chosen level. Once the Adaptive multiscale testing procedure has been performed at level α, if one is interested in finding significant differences at a level $\alpha_<$ smaller than that fixed, it is possible to simply look at the results of the α level Adaptive multiscale testing procedure. In fact, the $\alpha_<$ significant p-values found with the procedure at level α are the same ones that would be obtained if the Adaptive multiscale testing procedure is carried on at level $\alpha_<$. As for a level $\alpha_>$ larger than that fixed, this is not longer true. In fact, there's no guarantee that all the $\alpha_>$ significant differences resulting from the procedure at level α would be found also with the procedure under the level $\alpha_>$. With the Adaptive multiscale testing procedure at level $\alpha_>$, p-values that are between α and $\alpha_>$ could indeed grow over the level $\alpha_>$.*

We finally briefly compare the computational costs of the complete multiscale testing procedure and the adaptive multiscale testing procedure in terms of number of performed tests. We carry on this comparison analytically in an abstract situation, i.e. under the hypotheses that $n_1, n_2 \to +\infty$ and $\alpha \to 0^+$ and we express the costs in terms of: *(i)* number of (intra or inter) subnetworks where there is a difference between the two samples and *(2)* cardinality of the partition of the vertices. We consider separately the costs deriving from testing intra differences and from testing inter differences.

Cost of intra test of hypothesis. Let us suppose that the partition $\mathcal{V}$ of the vertex set V is composed of a total of m elements and that the number of significant intra–differences is k (i.e., k intra subnetworks exhibit a significant difference between the two samples). Varying the dimension of the elements of the σ-algebra generated by $\mathcal{V}$, we now count how many tests are not performed at each step of the adaptive multiscale testing procedure. For

$i = 1, \cdots, m-k$, if we are considering the step of the adaptive multiscale testing procedure where elements of the σ-algebra of dimension equal to i are tested, the number of tests that are not performed is $\binom{m-k}{i}$, that is the number of way of choosing i subsets V_h among the set of non significant $m-k$ subsets V_g. For dimension $i = m-k+1, \cdots, m$, the adaptive multiscale testing procedure is instead based on all the possible intra test of hypothesis. Therefore, the total number of tests that are not performed is:

$$\sum_{i=1}^{m-k} \binom{m-k}{i} = 2^{m-k} - \binom{m-k}{0} = 2^{m-k} - 1.$$

Hence, the total cost C_{intra} of the branch of the adaptive multiscale testing procedure involving intra test of hypothesis is

$$C_{\text{intra}} = 2^m - (2^{m-k} - 1).$$

Cost of inter test of hypothesis. In the case of inter test of hypothesis is not possible to count exactly how many tests are (or are not) performed at each step of the procedure because the cost might depend on the position of the inter–differences. We hence provide an estimation of the supremum of the cost. In effect, some particular positions of inter difference can lead to a lower computational cost. Let m be again the number of elements of the partition $\mathcal{V}$ and l the number of inter subnetworks that exhibit a difference between the two samples (that can be up to $m(m-1)/2$). For $i = 1, \cdots, m$, if we are considering the step of the adaptive multiscale testing procedure where elements of the σ-algebra of dimension equal to i are tested, the cost of each step is at most

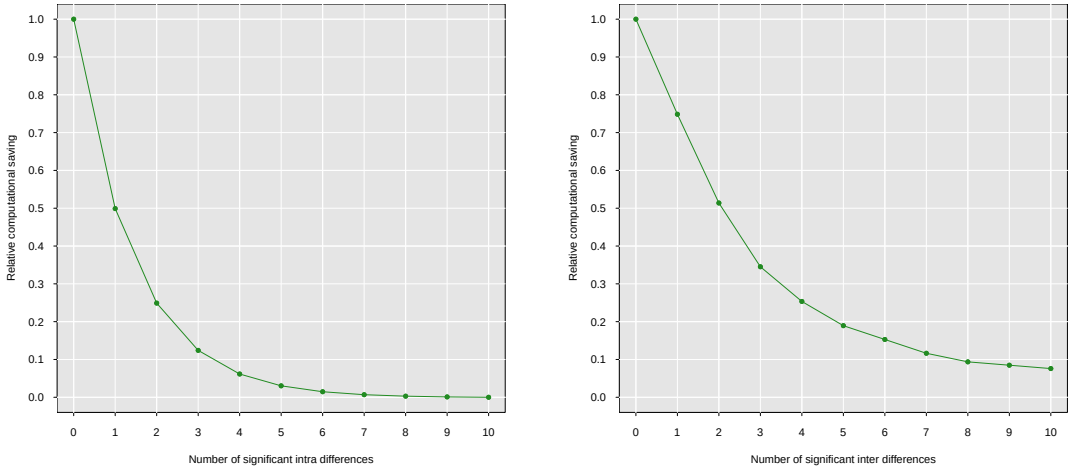
$$\binom{m-2}{i-2} l,$$

since for a given dimension i, and for a given inter–difference, the total number of subnetworks to be tested corresponds to the number of way of choosing $i-2$ subsets V_h among $m-2$ subsets V_g. The supremum of the total cost C_{inter} of the adaptive multiscale

testing procedure is therefore given by:

$$C_{\text{inter}} \leq \sum_{i=1}^{m} \binom{m-2}{i-2} l = l \sum_{i=1}^{m} \binom{m-2}{i-2}.$$

Figure 4.2.3 shows the relative computational savings computed as the difference between the cost of the complete procedure and that of the adaptive procedure, over the former. The fewer the differences that are present, the greater the computational saving. These relative savings do not depend on the number of elements of the partition of the vertices. If there are no differences between the two samples, the saving is total. If there is only one difference between the two samples, the adaptive multiscale testing procedure guarantees a saving of 50% in the case of intra differences and a saving of at least 75% in the case of inter differences. If the number of differences is two, savings of 25% and of at least 50% are guaranteed for intra and inter differences, respectively.



(a) *Relative computational savings in the case of intra differences.*

(b) *Relative computational savings in the case of inter differences.*

Figure 4.2.3: Relative computational savings.

4.3 Simulation studies

The aim of this section is to explore the potential of our methodology on simulated data sets where different levels and kinds of differences are present. In detail, the goal of this section is twofold. First of all we focus on the identification of local differences, showing that our procedure is able to detect in which areas of the network there are difference between the two populations, while controlling the family wise error rate; we also show that a naive approach that does not control for multiple testing fail in controlling the family wise error rate. Second, we generate a particular data set in order to highlight how important it is to consider the entire network approach also in the case of local inference instead of other summary objects that somehow summarize the entire structure of the network and we show how this latter approach loses power. Inspired by atlases commonly used in the clinical practise, we generate samples of networks with 68 vertices. We use a partition with four elements and the sample sizes are $n_1 = n_2 = 10$. We used a total of 1000 replicates and an α level equal to 0.05.

4.3.1 Identification of local differences

Simulated scenarios. In this first simulation study we simulate four different scenarios, all characterized by the presence of specific subnetworks where there is a differences in the edge strength distribution; what distinguishes the scenarios is which subnetworks are different between the two populations. To generate the samples, we rely on the stochastic block model [31], that is a useful model that allows to choose the probability of existence of an edge inside and between pre-specified areas of the network. The parameters of this model are the partition of the vertex set into disjoint subsets $C_1, \cdots, C_m$ and an edge probability matrix with dimension $m \times m$ whose element ij is the probability of existence of an edge between vertices belonging to area C_i and to area C_j. Therefore, the stochastic block model allows to specify the probabilities both of edges connecting vertices inside the pre-specified blocks (if $i = j$) and of edges connecting vertices belonging to different blocks (if $i \neq j$). Selecting $C_1, \cdots, C_m = V_1, \cdots, V_m$, through this model we are able to generate samples that have *intra*–differences and/or *inter*–differences in different locations. Table 4.3.1 (where we indicate the areas with RoI, i.e. Region of Interest) describes the four scenarios by means of a 4×4 matrix where a cross ($\times$) in entry ij (ii) represents an *inter*–difference between the two populations in the connections between area i and area j

(*intra*–difference in area *i*); on the other hand, a checkmark (✓) stands for no differences in the corresponding areas. As a result, in correspondence of a cross the null hypothesis is false, while it is true in correspondence of a checkmark. In detail, the two populations have both *intra*– and *inter*–differences. We start with a very trivial case where the differences are present in all the (four) *intra*–subnetworks and in all the (six) *inter*–subnetworks defined by the partition (see Table 4.3.1a). In the second and third scenario we explore separately *intra*– and *inter*–differences. The second scenario (see Table 4.3.1b) aims at testing our Adaptive multiscale testing procedure on a simulated data set where the differences in the two samples are located only in the *intra*–subnetworks; the third scenario (see Table 4.3.1c) is instead focused on differences only in the *inter*–subnetworks. Finally, we explore a more realistic scenario (see Table 4.3.1d) where the differences between the two populations are located in some of the *intra*– and *inter*–subnetworks. The four elements of the partition of the vertices contains 17 vertices each. In Appendix all the edge probability matrices are reported for each scenario.

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	×	×	×	×
RoI 2		×	×	×
RoI 3			×	×
RoI 4				×

(a) *First scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	×	✓	✓	✓
RoI 2		×	✓	✓
RoI 3			×	✓
RoI 4				×

(b) *Second scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	✓	×	×	×
RoI 2		✓	×	×
RoI 3			✓	×
RoI 4				✓

(c) *Third scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	✓	✓	×	✓
RoI 2		✓	×	✓
RoI 3			×	×
RoI 4				✓

(d) *Fourth scenario.*

Table 4.3.1: Explaining tables for the generated scenarios of the first simulation study.

Estimation of probability of rejection. Table 4.3.2 reports the estimated probability of rejection

of the test on the four simulated scenarios just described. As in the explaining Table 4.3.1, for each simulation we report a 4×4 table with the estimated probability of rejection for each tested hypothesis. The entry *ij* is referred to the comparison of the subnetwork identified by the areas *i* and *j*. If $i = j$, the result regards a *intra*–subnetwork, while if $i \neq j$, the result is inherent to a *inter*–subnetwork. The results show that in all the generated scenarios the Adaptive multiscale testing procedure is sensitive to the violation of the null hypothesis, regardless of the type of difference (*intra*– and *inter*–difference) that is present between the two populations. We explore the same scenarios also by means of a naive approach that simply tests all the null hypotheses separately without applying any strategy to correct for multiple comparisons. In this case a total of ten null hypotheses is tested (four *intra* hypotheses and six *inter* hypotheses). As in Table 4.3.2, Table 4.3.3 reports the estimated power for each tested hypothesis: the *intra*– and *inter*–differences are correctly detected and under the null hypothesis the power of the test is at the nominal level. What dramatically changes between the Adaptive multiscale testing procedure and the naive approach that simply tests all the null hypotheses without any correction procedure, is the Family Wise Error Rate (FWER), that is the probability of at least one false rejection. Table 4.3.4 compares the estimated FWER committed with the Adaptive multiscale testing procedure and the naive approach in Scenarios 2, 3 and 4. The results show that our procedure correctly controls the FWER at level α (0.012, 0.013 and 0.005 in Scenarios 2, 3 and 4, respectively) while the naive approach that does not control for the multiplicity fails in controlling the FWER, leading, as expected, to a probability of at least one false rejection of 0.255, 0.180 and 0.260 in each scenario, respectively.

4.3.2 The validity of the network approach

Simulated scenarios. When locally comparing two samples of networks, one may think of construct a partition-driven "aggregated network" starting from the original one. One way of doing it consists in constructing a new network where the vertices of an element of the partition in the original network boils down to a single vertex. Coherently, in the "aggregated network", the weight of an edge connecting vertex *i* and vertex *j* might be defined as the mean (or the sum) of the weights of the edges connecting the vertices belonging to area *i* and *j* of the original network. Once having

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	1.000	1.000	1.000	1.000
RoI 2		1.000	1.000	1.000
RoI 3			1.000	1.000
RoI 4				1.000

(a) *First scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	1.000	0.004	0.005	0.002
RoI 2		1.000	0.000	0.001
RoI 3			1.000	0.001
RoI 4				1.000

(b) *Second scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	0.002	1.000	1.000	1.000
RoI 2		0.002	1.000	1.000
RoI 3			0.006	1.000
RoI 4				0.003

(c) *Third scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	0.000	0.000	1.000	0.002
RoI 2		0.000	1.000	0.002
RoI 3			1.000	1.000
RoI 4				0.001

(d) *Fourth scenario.*

Table 4.3.2: Estimation of the probability of rejection with the Adaptive multiscale testing procedure on the generated data sets of the first simulation.

these two samples of "aggregated networks" one may think of testing local group differences using the maximal partition in the Adaptive multiscale testing procedure to see if there are difference in the connections between the given areas of the network. In this simulation we want to show that an approach based on "aggregated networks" fails to capture the actual complexity of the network. For this purpose, we generate two samples of networks with 68 vertices and chose a partition of four elements (with 10, 20, 17 and 21 vertices each). The weights of the networks are given by a Poisson distribution with parameter equal to 8 and a difference between the two samples is then introduced in 24 edges (over the $(10 \times 20)/2$ total possible edges) connecting the first two areas identified by the partition (so it is an *inter*–difference, see Table 4.3.5). In the first sample, the weight of 12 edges is modified in a Poisson distribution with parameter equal to 5 and other 12 edges are modified according to Poisson distribution with parameter equal to 11. In the second sample, exactly the same edges that have been modified in the first sample are modified too. Those edges that in the first sample were modified in a $Pois(5)$ in the second sample are turned into a $Pois(11)$ and vice versa. Appendix reports the details on the edges that have been modified.

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	1.000	1.000	1.000	1.000
RoI 2		1.000	1.000	1.000
RoI 3			1.000	1.000
RoI 4				1.000

(a) *First scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	1.000	0.042	0.048	0.065
RoI 2		1.000	0.049	0.054
RoI 3			1.000	0.049
RoI 4				1.000

(b) *Second scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	0.045	1.000	1.000	1.000
RoI 2		0.045	1.000	1.000
RoI 3			0.042	1.000
RoI 4				0.058

(c) *Third scenario.*

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	0.058	0.039	1.000	0.051
RoI 2		0.050	1.000	0.055
RoI 3			1.000	1.000
RoI 4				0.044

(d) *Fourth scenario.*

Table 4.3.3: Estimation of the probability of rejection with the naive approach on the generated data sets of the first simulation.

	Scenario 2	Scenario 3	Scenario 4
Adaptive multiscale procedure	0.012	0.013	0.005
Naive approach	0.255	0.180	0.260

Table 4.3.4: Family wise error rate on scenarios 2, 3 and 4 of the first simulation study.

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	✓	×	✓	✓
RoI 2		✓	✓	✓
RoI 3			✓	✓
RoI 4				✓

Table 4.3.5: Explaining tables for the generated scenario of the second simulation study.

In order to highlight the validity of our approach, we compare this two samples with the Adaptive multiscale testing procedure in two different ways. The first one makes use of the Adaptive multiscale testing procedure on the original network; the second one consider samples of "aggregated networks" instead of the original networks. In detail, for each element in the two samples, an "aggregated network" with four vertices is constructed. Each vertex corresponds to one of the four elements of the partition and the weight of an edge between a couple of vertices is defined as the arithmetic mean of the weights of edges connecting the vertices belonging to the two areas of the original network. Eventually, for each replicate, from two samples of networks with 68 vertices we derive two samples of networks with 4 vertices.

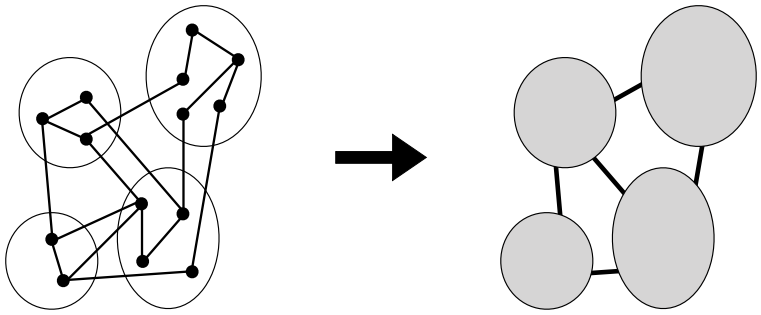


Figure 4.3.1: An illustrative example of an "aggregated network".

Estimation of probability of rejection. The results reported in Table 4.3.6 point out how important it is to consider the entire network instead of reducing the complexity of the data to a simpler object. From Table 4.3.6a it is possible to observe that the Adaptive multiscale testing procedure is able to detect where there is a difference between the two samples. In the case of the samples composed of "aggregated networks", the power of the global test (that is the first step of the Adaptive multiscale testing procedure) is approximately at the nominal level (0.042). Table 4.3.6b contains the estimate of the probability of rejection of the Adaptive multiscale testing procedure in this second case. It is clear that testing samples of "aggregated networks" instead of the original networks leads to a method that is not able to find out differences between the two samples. As a result, this example clearly emphasizes that an approach that takes into account of the entire

structure of the network finds differences while an approach based on an "aggregated network" do not. Moreover, boiling down a subset of vertices (i.e. an element V_i of the partition) to a single vertex loses information on the *intra*–subnetworks $G_{V_i}^{intra}$ (so that's why in Table 4.3.6b there is not the diagonal).

	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1	0	1	0.001	0.004
RoI 2		0.001	0.004	0.005
RoI 3			0	0.002
RoI 4				0

(a) *With the entire networks.*

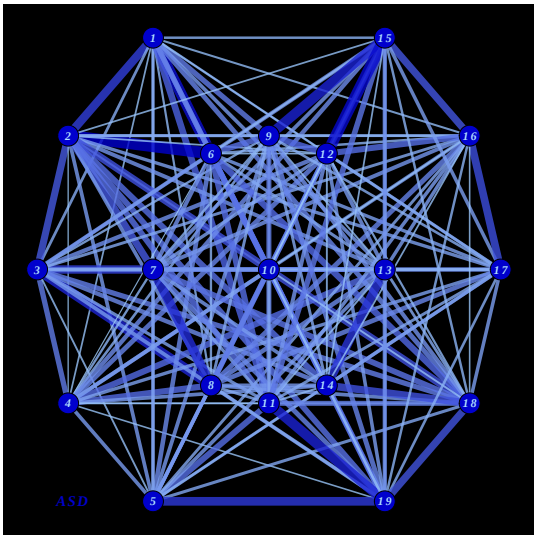
	RoI 1	RoI 2	RoI 3	RoI 4
RoI 1		0.001	0.008	0.004
RoI 2			0	0
RoI 3				0.002
RoI 4				

(b) *With the "aggregated networks".*

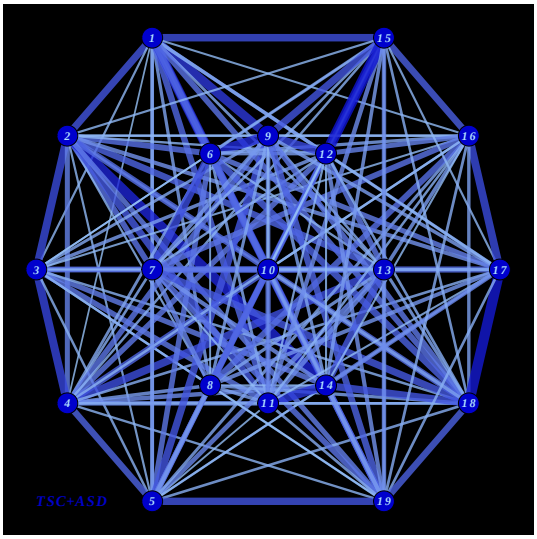
Table 4.3.6: Estimation of probability of rejection for the generated data sets of the second part.

4.4 Analysis of autistic subjects data set

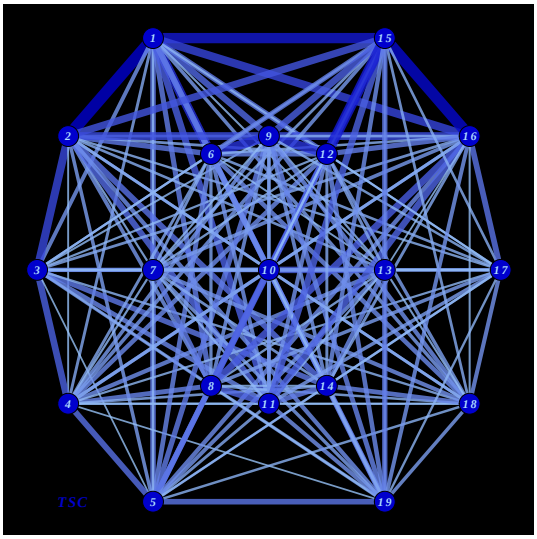
Description of the data set. We consider here one the three data sets studied in the previous chapter, that we briefly describe for self-content. We apply our methodology to brain functional networks of electroencephalographic (EEG) connectivity previously studied in [52]. The data has been collected from patients with Tuberous Sclerosis Complex (TSC) ($n = 29$), patients with Tuberous Sclerosis Complex and Autism Spectrum Disorder (ASD) ($n = 14$), patients with non-syndromic Autism Spectrum Disorder ($n = 16$) and controls ($n = 13$). The TSC is a multisystem, autosomal dominant disorder affecting children and adults and it results from mutations in one of two genes, TSC1 or TSC2 [15]. Approximately 40% of patients affected by TSC develop ASD [62]. See [52] for the details on the process of identification and diagnosis of the patients included in the study. Each network has 19 vertices identified by the electrode locations from the international 10-20 system of electrode placement while the edges are given by the coherence measure (see [52] for the detail on this measure and the validity of this approach). The



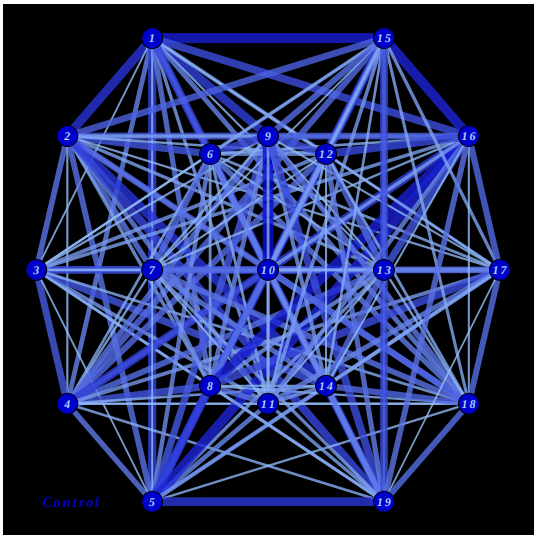
(a) *Autism Spectrum Disorder patient*



(b) *Tuberous Sclerosis Complex and Autism Spectrum Disorder patient*



(c) *Tuberous Sclerosis Complex patient*

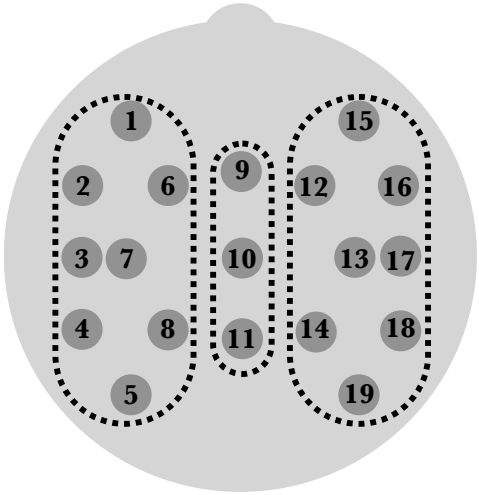


(d) *Control*

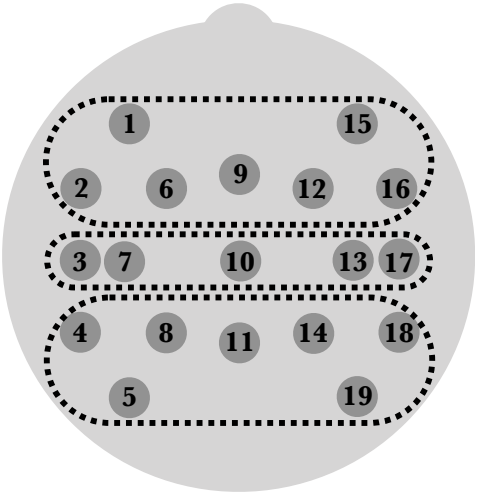
Figure 4.4.1: Representation of brain networks of a patient with non-syndromic autism, a patient affected by tuberous sclerosis complex and autism, a patient with tuberous sclerosis complex and a control.

networks are therefore weighted with weight between 0 and 1.

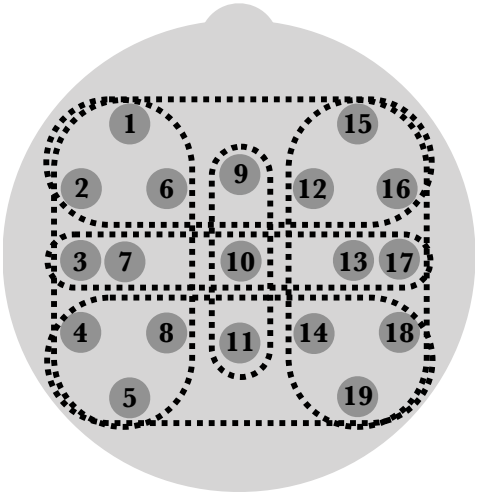
The choice of the partitions. This specific application allows to highlight a strength of our approach. While in some case it is possible to choose the position of the vertices through an application-driven approach, in the case of EEG connectivity data, the electrode locations, given by an international system of electrode positions, have not a physiological meaning and therefore the same holds for the vertices in the networks. It is possible to overcome this limitation and give a meaning at each vertices adopting our approach. Indeed, our procedure for local inference allows to identify some regions of interest and conduct the test on these regions, even if the original networks have positions of vertices without a strong meaning. In this application we initially consider two different partitions of the set of vertices. The first one, indicated with $\mathcal{P}_{LR}$, is that into right and left hemisphere and it practically translates into a partition of three elements, as represented in Figure 4.4.2a. The second partition, indicated with $\mathcal{P}_{FT}$, is composed of three elements: frontal lobe, intermediate area and temporal lobe, as represented in Figure 4.4.2b. In the third case we consider a splitting of the vertices given by the union of the two previous partitions $\mathcal{P}_{LR}$ and $\mathcal{P}_{FT}$. This is not an actual partition but, as stated in Remark 4, it is however possible to construct the σ–algebra generated by it. Practically, it is as we were considering a partition $\mathcal{P}_{LR-FT}$ of nine elements, represented by Figure 4.4.2c. In detail, the partition is composed of these elements: $\{1, 2, 6\}, \{9\}, \{12, 15, 16\}, \{3, 7\}, \{10\}, \{13, 17\}, \{4, 5, 8\}, \{11\}, \{14, 18, 19\}$.



(a) *Partition $\mathcal{P}_{LR}$ of the brain in left hemisphere, central area and right hemisphere*



(b) *Partition $\mathcal{P}_{FT}$ of the brain in frontal lobe, intermediate area and temporal lobe*



(c) *Partition $\mathcal{P}_{LR-FT}$ with nine elements.*

Figure 4.4.2: The three partitions of the vertices considered in the application on EEG data set.

Results. We consider the Adaptive multiscale testing procedure with a level equal to 0.1 and a number B of permutations equal to 10000. Tables 4.4.1 and 4.4.2 reports the results of the local inference for the comparison between patients with non-syndromic ASD and patients with ASD and TSC and between patients with non-syndromic ASD and controls. We present the results using adjacency matrix representation and frobenius distance and theta band. Figures 4.4.3 and 4.4.4 represent the locations of significant *intra*–differences (light blue areas) and the locations of *inter*–differences (dark blue arrows). As for the comparison between autistic patients and patients affected both by autism and by tuberous sclerosis complex (see Table 4.4.1), there is a correspondence between the results in the case of the partitions with three elements and that with nine elements. In detail, the macro significant differences found in the two samples where the first two partitions are considered (see Tables 4.4.1a and 4.4.1b) are found also when the vertices are split in nine subgroups (see Table 4.4.1c). As for the first partition (Figure 4.4.3a), there are *intra*–differences in left hemisphere, central area and right hemisphere, while the *inter*–differences are between left and right hemisphere, between left hemisphere and central area and between central area and right hemisphere. As for the second partition (Figure 4.4.3b), the *intra*–differences are located in frontal and temporal lobe, while the *inter*–differences are between frontal lobe and intermediate area and intermediate area and temporal lobe. With the third partition (Figure 4.4.3c), a more detailed representation of the locations of the differences is provided. The *intra*–differences are located in frontal-left area, frontal-right area and temporal-left area. The *inter*–differences are the following: between frontal-left area and frontal-right area, between frontal-central area and intermediate-right area, between frontal-left area and intermediate-right area, between intermediate-right area and temporal-right area, between frontal-right area and intermediate-central area, between intermediate-central area and central-temporal area, between central-temporal area and temporal-right area, between temporal-left area and central-temporal area, between intermediate-left area and central-temporal area and finally between intermediate-left area and temporal-left area.

As for the comparison between autistic patients and controls (see Table 4.4.2), the macro significant differences found in the two samples where the first two partitions are considered (see Tables 4.4.2a and 4.4.2b) are found also when the vertices are split in nine subgroups (see Table 4.4.2c), except in one case. As for the first partition (Figure 4.4.4a), the significant *intra*–differences and *inter*–differences found are exactly the same found in the previous group

comparison. As for the second partition (Figure 4.4.4b), the *intra*–differences are located in frontal lobe, intermediate area and in temporal lobe, while the *inter*–differences are between frontal and temporal lobe and between intermediate area and temporal lobe. The partition with nine elements gives more information (Figure 4.4.4c), except in the case of the the *intra*–difference in the central area for which it is not possible to establish where the differences are (among the connections frontal-intermediate, the connections intermediate-temporal, the connections frontal-temporal). In detail, the *intra*–differences are in frontal-left area, frontal-right area, temporal-left area and temporal-right area. The *inter*–differences are the following: between frontal-left area and frontal-right area, between frontal-left area and temporal-left area, between frontal-central area and frontal-right area, between frontal-right area and temporal-right area, between frontal-right area and temporal-left area, between intermediate-left area and intermediate-central area, between intermediate-left area and temporal-left area, between intermediate-right area and temporal-left area, between intermediate-left area and temporal-right area, between temporal-left area and temporal-central area, between temporal-central area and temporal-right area and finally between temporal-left area and temporal-right area. The results contained in this section, together with the results of the previous chapter, are now on the attention of some neurologists at the Boston Children's Hospital.

	Left	Central	Right
Left	$<$ **0.001**	**0.002**	**0.002**
Central		**0.016**	**0.002**
Right			**0.004**

(a) *Left and right hemisphere.* $p_{intra} =$ ***0.0002*** *and* $p_{inter} =$ ***0.0006***

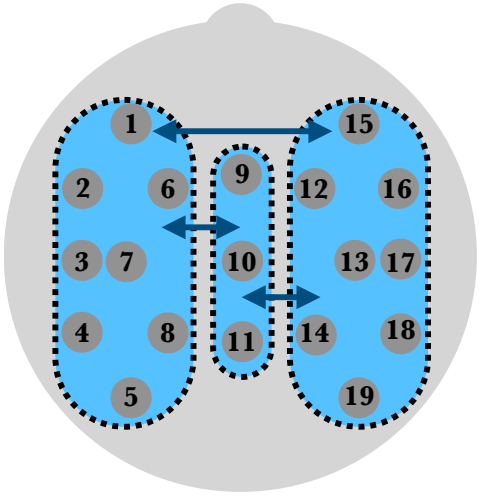
	Front.	Interm.	Temp.
Front.	**0.002**	$<$ **0.001**	$\geq$ 0.231
Interm.		$\geq$ 0.169	$<$ **0.001**
Temp.			**0.002**

(b) *Frontal and temporal lobe.* $p_{intra} =$ ***0.0002*** *and* $p_{inter} =$ ***0.0004***

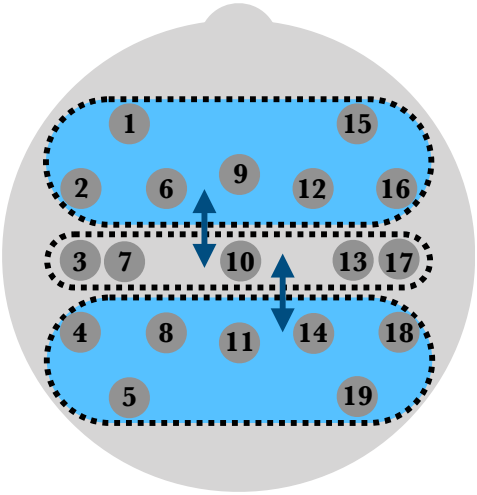
	Front. left	Front. centr.	Front. right	Interm. left	Interm. centr.	Interm. right	Temp. left	Temp. centr.	Temp. right
Front. left	**0.031**	$\geq$0.269	**0.023**	$\geq$0.181	$\geq$0.269	**0.093**	$\geq$0.197	$\geq$0.193	$\geq$0.269
Front. centr.			$\geq$0.296	$\geq$0.181	$\geq$0.287	**0.093**	$\geq$0.296	$\geq$0.208	$\geq$0.287
Front. right			**0.094**	$\geq$0.112	**0.067**	$\geq$0.152	$\geq$0.206	$\geq$0.208	$\geq$0.112
Interm. left				$\geq$0.309	$\geq$0.169	$\geq$0.169	**0.050**	**0.095**	$\geq$0.181
Interm. centr.						$\geq$0.169	$\geq$0.140	**0.022**	$\geq$0.287
Interm. right						$\geq$0.510	$\geq$0.152	$\geq$0.510	**0.060**
Temp. left							**0.037**	**0.072**	$\geq$0.187
Temp. centr.									**0.060**
Temp. right									$\geq$0.309

(c) *Right and left hemisphere, frontal and temporal lobe.* $p_{intra} =$ ***0.0004*** *and* $p_{inter} =$ ***0.0004***

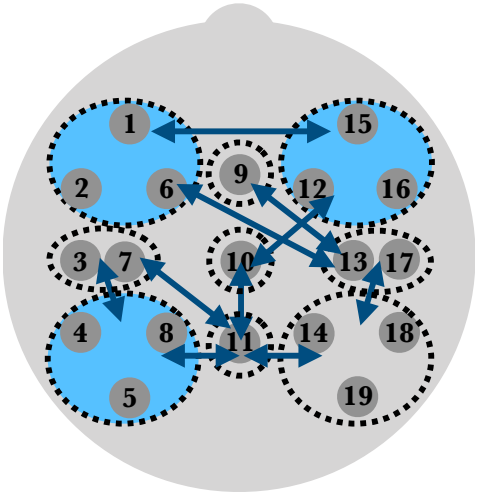
Table 4.4.1: Results of the Adaptive multiscale testing procedure for the comparison between autistic patients and patients with autism and tuberous sclerosis complex.



(a) *Differences found with the partition in left hemisphere, central area and right hemisphere.*



(b) *Differences found with the partition in frontal lobe, intermediate area and temporal lobe.*



(c) *Differences found with the partition of nine elements.*

Figure 4.4.3: Locations of the differences between autistic patients and patients with autism and tuberous sclerosis complex. Light blue areas refers to those subnetworks where there is an *intra*–difference between the two samples, while dark blue arrows indicate the presence of an *inter*–difference.

	Left	Central	Right
Left	$<$ **0.001**	**0.003**	**0.001**
Central		**0.026**	**0.002**
Right			$<$ **0.001**

(a) *Left and right hemisphere.* $p_{intra} =$ ***0.0002*** *and* $p_{inter} =$ ***0.0010***

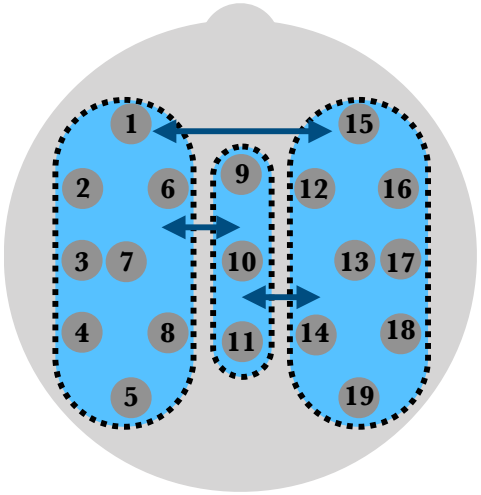
	Front.	Interm.	Temp.
Front.	$<$ **0.001**	$\geq$ 0.146	**0.021**
Interm.		**0.010**	**0.003**
Temp.			$<$ **0.001**

(b) *Frontal and temporal lobe.* $p_{intra} =$ ***0.0002*** *and* $p_{inter} =$ ***0.0034***

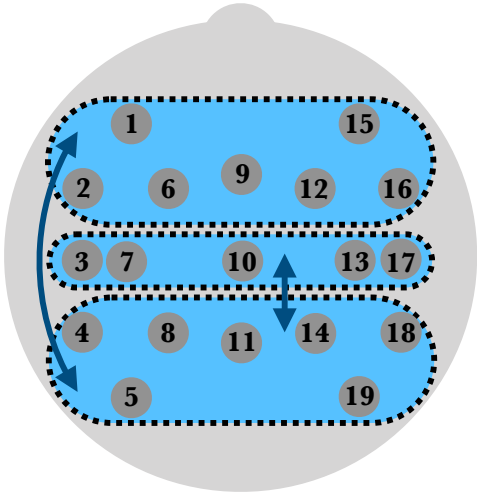
	Front. left	Front. centr.	Front. right	Interm. left	Interm. centr.	Interm. right	Temp. left	Temp. centr.	Temp. right
Front. left	**0.003**	$\geq$0.172	**0.005**	$\geq$0.163	$\geq$0.172	$\geq$0.172	**0.033**	$\geq$0.110	$\geq$0.163
Front. centr.			**0.038**	$\geq$0.163	$\geq$0.172	$\geq$0.172	$\geq$0.151	$\geq$0.110	$\geq$0.163
Front. right			**0.001**	$\geq$0.209	$\geq$0.171	$\geq$0.209	**0.032**	$\geq$0.209	**0.095**
Interm. left				$\geq$0.155	**0.077**	$\geq$0.209	**0.006**	$\geq$0.209	$\geq$0.163
Interm. centr.						$\geq$0.172	$\geq$0.639	$\geq$0.171	$\geq$0.151
Interm. right						$\geq$0.187	**0.032**	$\geq$0.219	**0.085**
Temp. left							**0.002**	**0.008**	**0.029**
Temp. centr.									**0.038**
Temp. right									**0.006**

(c) *Right and left hemisphere, frontal and temporal lobe.* $p_{intra} =$ ***0.0002*** *and* $p_{inter} =$ ***0.0006***

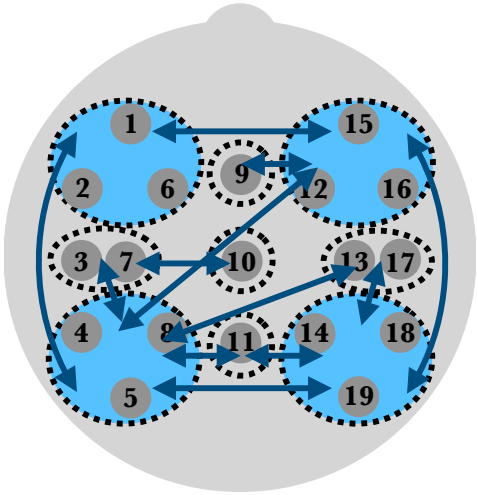
Table 4.4.2: Results of the local inference for the comparison between autistic patients and controls.



(a) *Differences found with the partition in left hemisphere, central area and right hemisphere.*



(b) *Differences found with the partition in frontal lobe, intermediate area and temporal lobe.*



(c) *Differences found with the partition of nine elements.*

Figure 4.4.4: Locations of the differences between autistic patients and controls. Light blue areas refers to those subnetworks where there is an *intra*–difference between the two samples, while dark blue arrows indicate the presence of an *inter*–difference.

4.5 Discussion

The aim of this chapter is to formulate a procedure that, once it is known that there is a significant difference between two samples, allows to find out which subnetworks are responsible for the global observed difference. This aim opens up to theoretical and computational challenges. A possible high number of null-hypothesis can be stated and therefore it is necessary to control the error due to multiple comparisons; meanwhile an adaptive procedure that reduces the computational costs has been introduced. The proposed procedure is very flexible and allows to include the expertise of the sector specialist in the definition of the partition of the vertices.

Appendix

In this section we detail how to replicate the generated scenarios studied in Section 4.3.

Details on the generated scenarios of the first simulation study

We report the edges probability matrices used to generate the simulated scenarios in the first simulation study (see Section 4.3.1).

Scenario 1

$$p_1 = \begin{pmatrix} 0.4 & 0.1 & 0.4 & 0.1 \\ 0.1 & 0.1 & 0.1 & 0.4 \\ 0.4 & 0.1 & 0.1 & 0.4 \\ 0.1 & 0.4 & 0.4 & 0.1 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0.1 & 0.4 & 0.1 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.1 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.4 \end{pmatrix}$$

Scenario 2

$$p_1 = \begin{pmatrix} 0.1 & 0.4 & 0.4 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.1 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0.4 & 0.4 & 0.4 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.4 \end{pmatrix}$$

Scenario 3

$$p_1 = \begin{pmatrix} 0.1 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.4 & 0.4 & 0.1 \\ 0.1 & 0.4 & 0.1 & 0.4 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0.1 & 0.1 & 0.4 & 0.4 \\ 0.1 & 0.1 & 0.1 & 0.1 \\ 0.4 & 0.1 & 0.4 & 0.4 \\ 0.4 & 0.1 & 0.4 & 0.4 \end{pmatrix}$$

Scenario 4

$$p_1 = \begin{pmatrix} 0.1 & 0.4 & 0.1 & 0.4 \\ 0.4 & 0.4 & 0.1 & 0.4 \\ 0.1 & 0.1 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0.1 & 0.4 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.4 & 0.4 \\ 0.4 & 0.4 & 0.1 & 0.1 \\ 0.4 & 0.4 & 0.1 & 0.4 \end{pmatrix}$$

Details on the generated scenario of the second simulation study

We here specify which edges connecting vertices in the first two elements of the partition are modified from $Pois(8)$ to $Pois(5)$ and $Pois(11)$ to introduce a difference between the two samples:
First sample

- from $Pois(8)$ to $Pois(5)$: 1-11, 1-13, 2-12, 2-17, 2-18, 3-14, 3-19, 4-12, 4-16, 5-11, 5-15, 5-1
- from $Pois(8)$ to $Pois(11)$: 6-20, 7-20, 7-23, 7-27, 8-24, 8-26, 8-28, 9-21, 9-22, 9-29, 10-24, 10-25.

Second sample

- from $Pois(8)$ to $Pois(5)$: 6-20, 7-20, 7-23, 7-27, 8-24, 8-26, 8-28, 9-21, 9-22, 9-29, 10-24, 10-25
- from $Pois(8)$ to $Pois(11)$: 1-11, 1-13, 2-12, 2-17, 2-18, 3-14, 3-19, 4-12, 4-16, 5-11, 5-15, 5-1.

5

The R package `nevada`

The package `nevada` (NEtwork-VAlued Data Analysis) is an R package for the statistical analysis of network-valued data sets. It has been developed during the Ph.D. program and it is itself a contribution of the thesis. It is available on GitHub (https://github.com/astamm/nevada) and can be simply installed from it. The package provides a set of matrix representations for networks so that network-valued data can be transformed into matrix-valued data. Subsequently, a number of distances between matrices is provided as well to quantify how far two networks are from each other and a number of distance-based statistics is proposed for testing equality in distribution between samples of networks using exact permutation testing procedures. The implementation is largely made in C++ and the matrix of inter– and intra–sample distances is pre-computed, which alleviates the computational burden often associated with permutation tests. A multiscale null-hypothesis procedure is also implemented and the choice of matrix representation, distance between networks and test statistics is possible as well. In details:

- the `repr_*()` functions return the chosen matrix representation of the input network;

- the `dist_*()` functions return the chosen distance between two networks;
- the `stat_*()` functions return the value of the chosen test statistic;
- the `test_twosample()` function returns the p-value of a permutation test in which the null hypothesis is that the two samples come from the same distribution of networks;
- the `network_localtest2p()` function returns *intra*– and *inter*– p-values of a multiscale null-hypothesis testing of no differences in the distribution of networks based on a partition of the set of vertices suggested by the user.

In the following each function is detailed.

5.1 Network Representation Functions

Description

This is a collection of functions that convert a network stored as an `igraph` object into a desired matrix representation among adjacency matrix, graph laplacian or modularity matrix.

Usage

```
repr_adjacency(network, validate = TRUE)
repr_laplacian(network, validate = TRUE)
repr_modularity(network, validate = TRUE)
```

Arguments

`network`	An `igraph` object.
`validate`	A boolean specifying whether the function should check the class of its input (default: `TRUE`).

Value

A numeric square matrix giving the desired network representation recorded in the object's class.

Examples

```
g <- igraph::sample_smallworld(1, 25, 3, 0.05)
repr_adjacency(g)
repr_laplacian(g)
repr_modularity(g)
```

5.2 Distances Between Networks

Description

This is a collection of functions computing the distance between two networks.

Usage

```
dist_hamming(x, y, representation = "adjacency")
dist_frobenius(x, y, representation = "laplacian")
dist_spectral(x, y, representation = "laplacian")
dist_root_euclidean(x, y, representation = "laplacian")
```

Arguments

x	An igraph object or a matrix representing an underlying network.
y	An igraph object or a matrix representing an underlying network. Should have the same number of vertices as x.
representation	A string specifying the desired type of representation, among: "adjacency" [default], "laplacian" and "modularity".

Value

A scalar measuring the distance between the two input networks.

Examples

```
g1 <- igraph::sample_gnp(20, 0.1)
g2 <- igraph::sample_gnp(20, 0.2)
dist_hamming(g1, g2, "adjacency")
dist_frobenius(g1, g2, "adjacency")
dist_spectral(g1, g2, "laplacian")
dist_root_euclidean(g1, g2, "laplacian")
```

5.3 Test Statistics for Network Populations

Description

This is a collection of functions that provide statistics for testing equality in distribution between samples of networks.

Usage

```
stat_lot(d, indices)
stat_sot(d, indices)
stat_biswas(d, indices)
stat_energy(d, indices, alpha = 1)
stat_edge_count(d, indices, type = "generalized")
```

Arguments

d	Either a matrix of dimension $(n_1 + n_2)(n_1 + n_2)$ containing the distances between all the elements of the two samples put together (for distance-based statistics) or a list of edge properties of a similarity graph for the graph-based edge count statistics.
indices	A vector of dimension n1 containing the indices of the elements of the first sample.
alpha	An integer specifying to which power elevating the Euclidean distance of the energy-based statistic (default: 1L).
type	A string specifying the version of the edge count test statistic to be used. Choices are "original", "generalized" [default] or "weighted".

Value

A scalar giving the value of the desired test statistic.

Examples

```
n1 <- 30L
n2 <- 10L
x <- nvd("smallworld", n1)
y <- nvd("pa", n2)
d <- dist_nvd(x, y, representation = "laplacian", distance =
"frobenius")
stat_lot(d, 1:n1)
stat_sot(d, 1:n1)
stat_biswas(d, 1:n1)
stat_energy(d, 1:n1)
```

```
r <- repr_nvd(x, y, representation = "laplacian")
e <- edge_count_global_variables(d, n1, k = 5L)
stat_edge_count(e, 1:n1, type = "original")
stat_edge_count(e, 1:n1, type = "generalized")
stat_edge_count(e, 1:n1, type = "weighted")
```

5.4 Comparison of Network Distributions

Description

This function carries out an hypothesis test where the null hypothesis is that the two populations of networks share the same underlying probabilistic distribution against the alternative hypothesis that the two populations come from different distributions. The test is performed in a non-parametric fashion using a permutational framework in which several statistics can be used, together with several choices of network matrix representations and distances between networks.

Usage

```
test_twosample(x, y, representation = "adjacency", distance =
"frobenius", statistic = "lot", B = 1000L, alpha = 0.05, test =
"exact", k = 5L)
```

Arguments

`x`	An `nvd` object listing networks in sample 1.
`y`	An `nvd` object listing networks in sample 2.
`representation`	A string specifying the desired type of representation, among: "`adjacency`" [default], "`laplacian`" and "`modularity`".
`distance`	A string specifying the chosen distance for calculating the test statistic, among: "`hamming`", "`frobenius`" [default], "`spectral`" and "`root-euclidean`".
`statistic`	A string specifying the chosen test statistic(s), among: "`lot`" [default], "`sot`", "`biswas`", "`energy`", "`original`", "`generalized`", "`weighted`" or a combination from c("`lot`", "`sot`", "`biswas`", "`energy`").

`B` The number of permutation or the tolerance (default: `1000L`). If this number is lower than 1, it is intended as a tolerance. Otherwise, it is intended as the number of required permutations.

`alpha` The significance level (default: "`0.05`).

`test` A character string specifying if performing an exact test through the use of Phipson-Smyth estimate of the p-value or an approximate test through a Monte-Carlo estimate of the p-value (default: "`exact`").

`k` An integer specifying the density of the minimum spanning tree used for the edge count statistics (default: `5L`).

Value

A list with three components: the value of the statistic for the original two samples, the p-value of the resulting permutation test and a numeric vector storing the values of the permuted statistics.

Examples

```
n <- 10L
x <- nvd("smallworld", n)
y <- nvd("pa", n)
test1 <- test_twosample(x, y, "modularity")
x <- nvd("smallworld", n)
y <- nvd("smallworld", n)
test2 <- test_twosample(x, y, "modularity")
```

5.5 Multiscale null hypothesis testing for networks

Description

This function carries out a local hypothesis test via a multiscale testing procedure, based on a partition of the vertices suggested by the user. Each single null hypothesis is that the two populations of networks share the same underlying probabilistic distribution against the alternative hypothesis that the two populations come from different distributions. Several statistics can be used, together with several choices of network matrix representations and distances between networks.

Usage

```
  network_localtest2p <- function(x, y, location, representation =
"adjacency", distance = "frobenius", statistic = c("sot", "lot"), alpha
= 0.05, B = 1000)
```

Arguments

x	An nvd object listing networks in sample 1.
y	An nvd object listing networks in sample 2.
location	A vector where entry *i* specifies in which element of the partition vertex *i* belongs.
representation	A string specifying the desired type of representation, among: "adjacency" [default], "laplacian" and "modularity".
distance	A string specifying the chosen distance for calculating the test statistic, among: "hamming", "frobenius" [default], "spectral" and "root-euclidean".
statistic	A string specifying the chosen test statistic(s), among: "lot", "sot", "biswas", "energy", "original", "generalized", "weighted" or a combination from c("lot", "sot", "biswas", "energy") (default: c("lot", "sot") .
B	The number of permutation or the tolerance (default: 1000L). If this number is lower than 1, it is intended as a tolerance. Otherwise, it is intended as the number of required permutations.
alpha	The significance level (default: "0.05).

Value

A list with four components: the value of the *intra* global p-value, the results of the inference for the *intra*–subnetworks, the value of the *inter* global p-value and the results of the inference for the *inter*–subnetworks.

Examples

```
  p1 <- matrix(data = c(0.1, 0.4, 0.1, 0.4, 0.4, 0.4, 0.1, 0.4, 0.1,
0.1, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4), nrow = 4, ncol = 4, byrow = TRUE)
```

```
  p2 <- matrix(data = c(0.1, 0.4, 0.4, 0.4,0.4, 0.4, 0.4, 0.4, 0.4,
0.4, 0.1, 0.1, 0.4, 0.4, 0.1, 0.4), nrow = 4, ncol = 4, byrow = TRUE)
  block1 <- NULL
  block2 <- NULL
  for (i in 1:s) {
  block1[[i]] <- as_adjacency_matrix(sample_sbm(68, p1,
c(17,17,17,17)), type = "both", sparse = FALSE)
  block2[[i]] <- as_adjacency_matrix(sample_sbm(68, p2,
c(17,17,17,17)), type = "both", sparse = FALSE)
  }
  network_localtest2p(block1, block2, c(rep(1,17), rep(2, 17), rep(3,
17), rep(4,17)))
```

6

Conclusion

The main topic of this thesis is null-hypothesis testing for network-valued data. It starts with the most simple null-hypothesis problem, i.e. two-sample test. Due to the complexity of the data, the problem is tackled from the perspective of the permutation framework. The choice of the test statistic is then critical because it makes the test sensitive to specific features of the distribution. Therefore, there is no uniformly better statistic for testing equality in distribution but rather many statistic that look at the distribution under different angles. We introduced two statistics which, when combined together through the Non-Parametric Combination methodology, are sensitive to differences in the first two moments of the distributions. Furthermore, our proposed method relies only on inter-point distances. This means that all we need is a metric between networks to perform two-sample testing. Hence, we believe that our proposal could be a valid approach not only for network-valued data analysis, but, in a broader context, for Object Oriented Data Analysis, provided that the object data used as sample unit can be embedded into a metric space.

Staying within the context of network-valued data, we explore the potential of the two-sample

test introduced when it is applied to brain networks data sets. We compared two different ways of conducting two-sample tests on brain networks: we consider the classical way of summarise an entire brain network with a brain measure and then test the vectors of this measure and we took into account the two-sample test for network-valued data. We consider brain networks derived from different acquisition procedures (i.e. EEG, fMRI, DCI) and with different number of vertices (19, 116, 134). Both the sample sizes and the criteria that define the samples are different (i.e. the type of disease, the presence/absence of a disease, the age, the risk of autism). In all these different cases the two-sample test for network-valued data performs better than the standard method of comparing brain networks by means of a univariate test involving summary measures. This new method opens up to possible new discoveries in the field of neuroimaging and to possible improvements in treatment and diagnosis.

The natural continuation of this work is the formulation of a procedure that, once it is known that there is a significant difference between two samples, allows to find out which subnetworks are responsible for the global observed difference. This aim opens up to theoretical and computational challenges. A possible high number of null-hypothesis can be stated and therefore it is necessary to control the error due to multiple comparisons; meanwhile an adaptive procedure that reduces the computational costs has been introduced. The proposed procedure is very flexible and allows to include the expertise of the sector specialist in the definition of the partition of the vertices.

Appendix

6.1 Visualization of the entire bikeMi data set

Figures 6.1.1–6.1.7, represent all the networks in the data set considered in the Section 2.3 of the second chapter. The relative positions of the vertices are coherent with the real positions of the neighbourhoods of Milan.

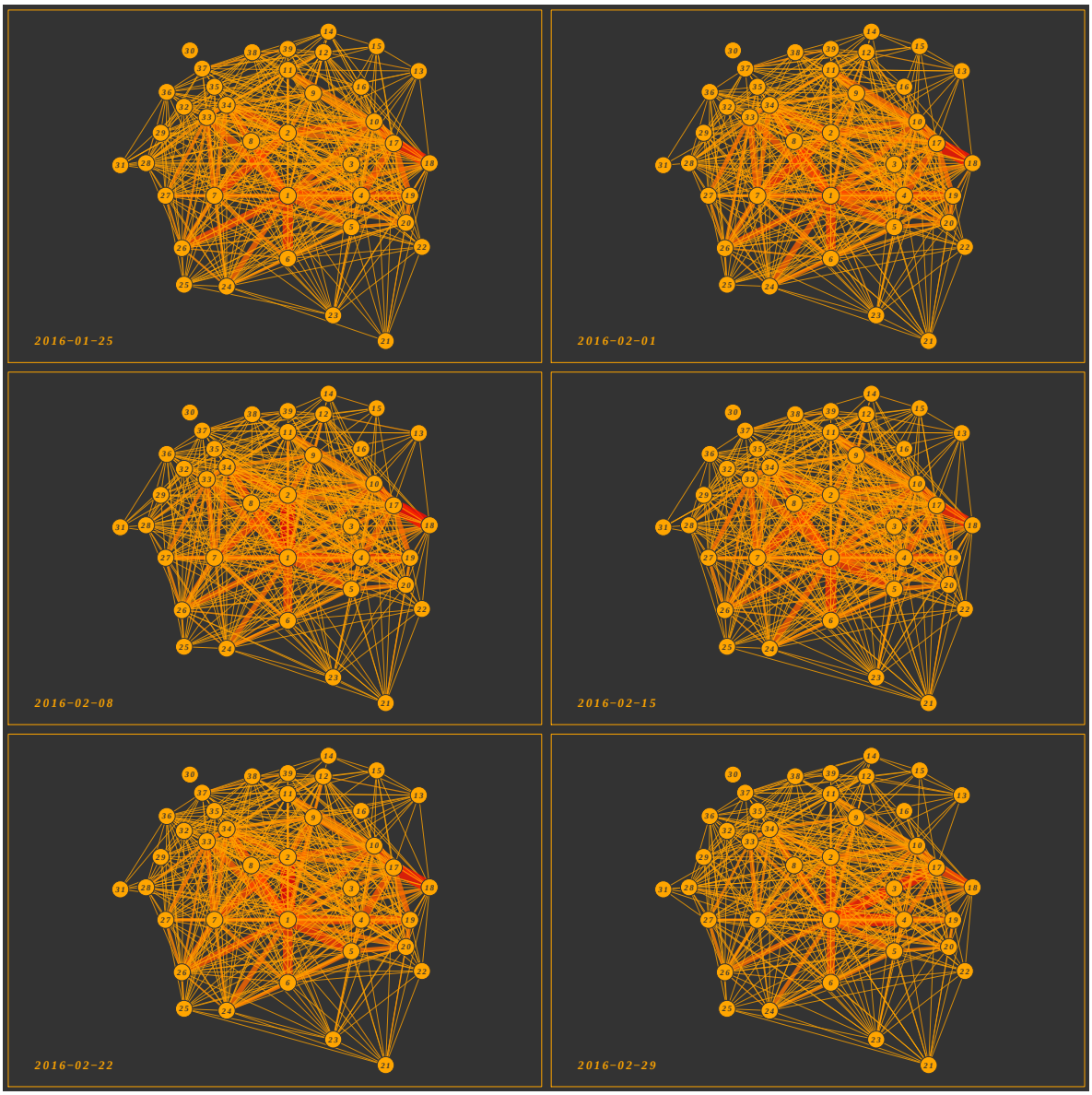


Figure 6.1.1: Networks representing the trips between NILs of Milan on Monday.

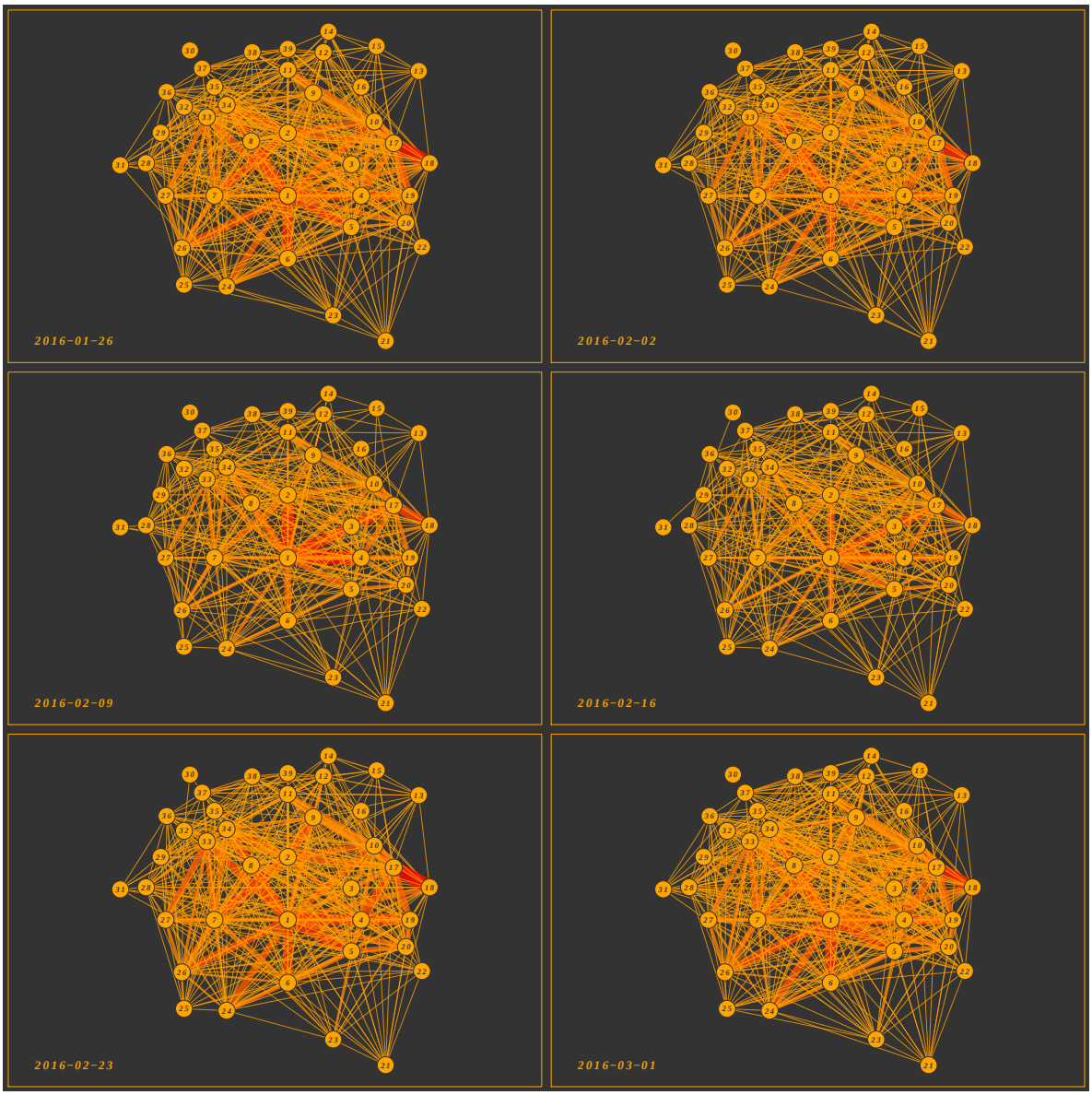


Figure 6.1.2: Networks representing the trips between NILs of Milan on Tuesday.

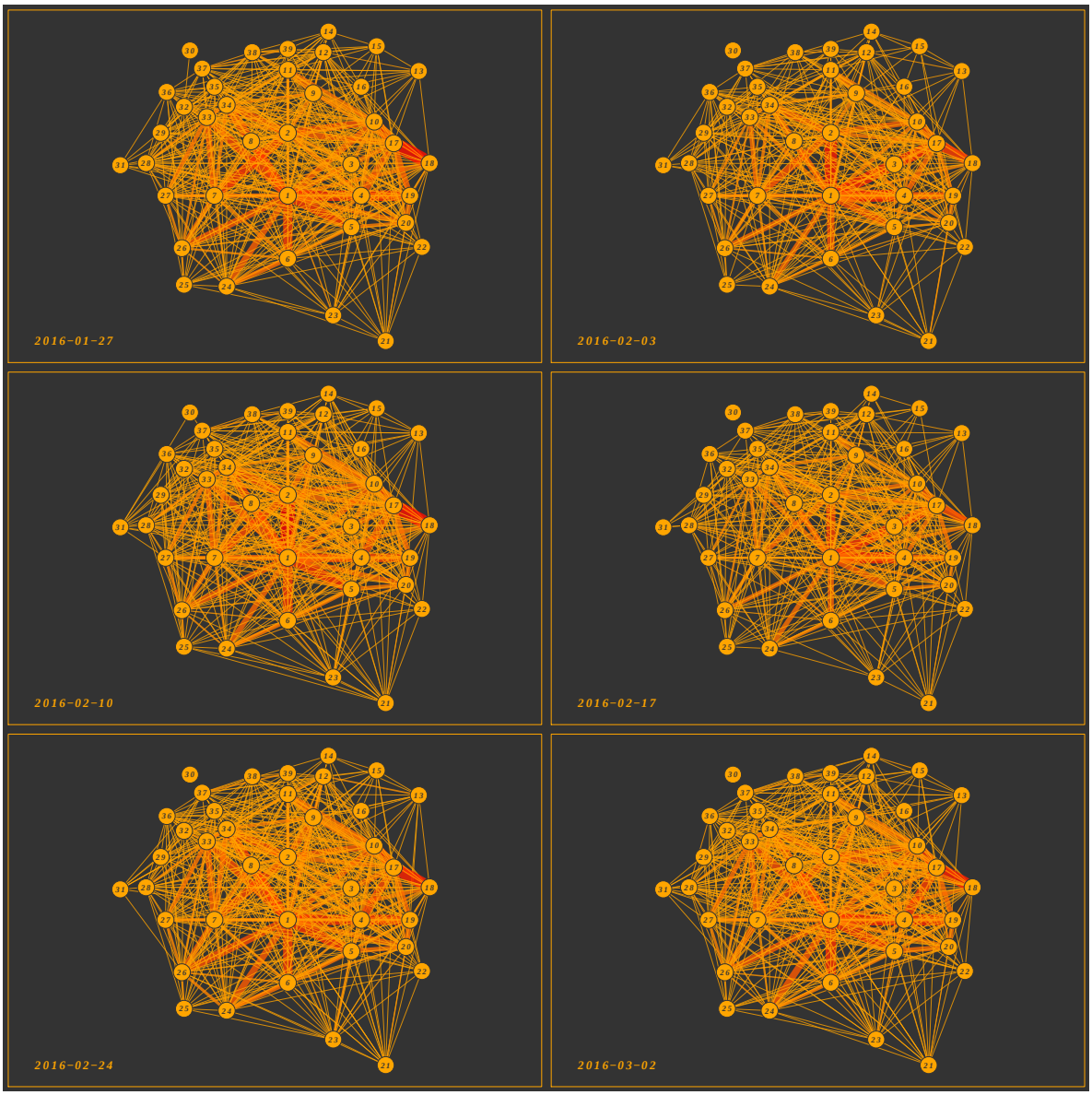


Figure 6.1.3: Networks representing the trips between NILs of Milan on Wednesday.

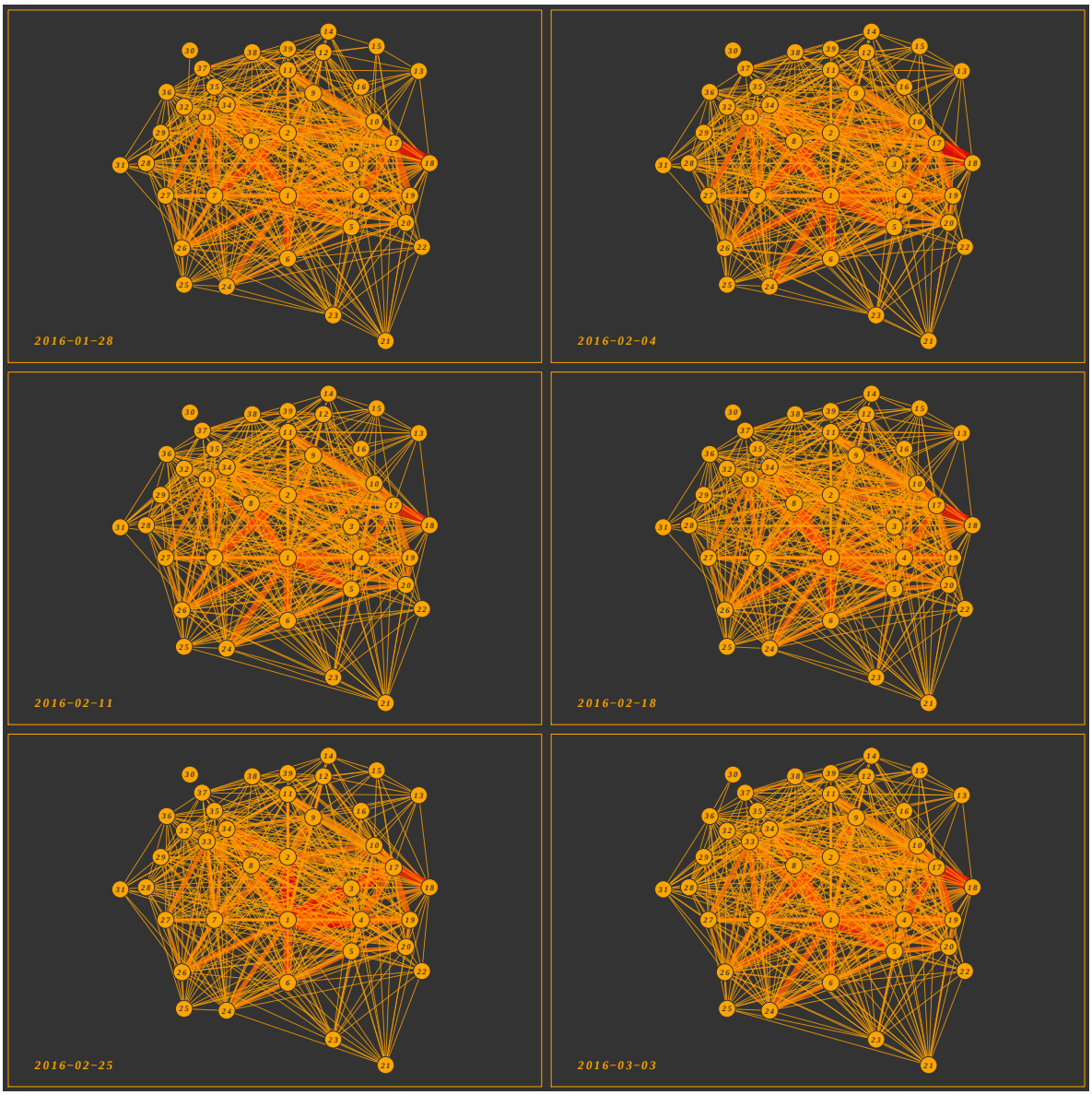


Figure 6.1.4: Networks representing the trips between NILs of Milan on Thursday.

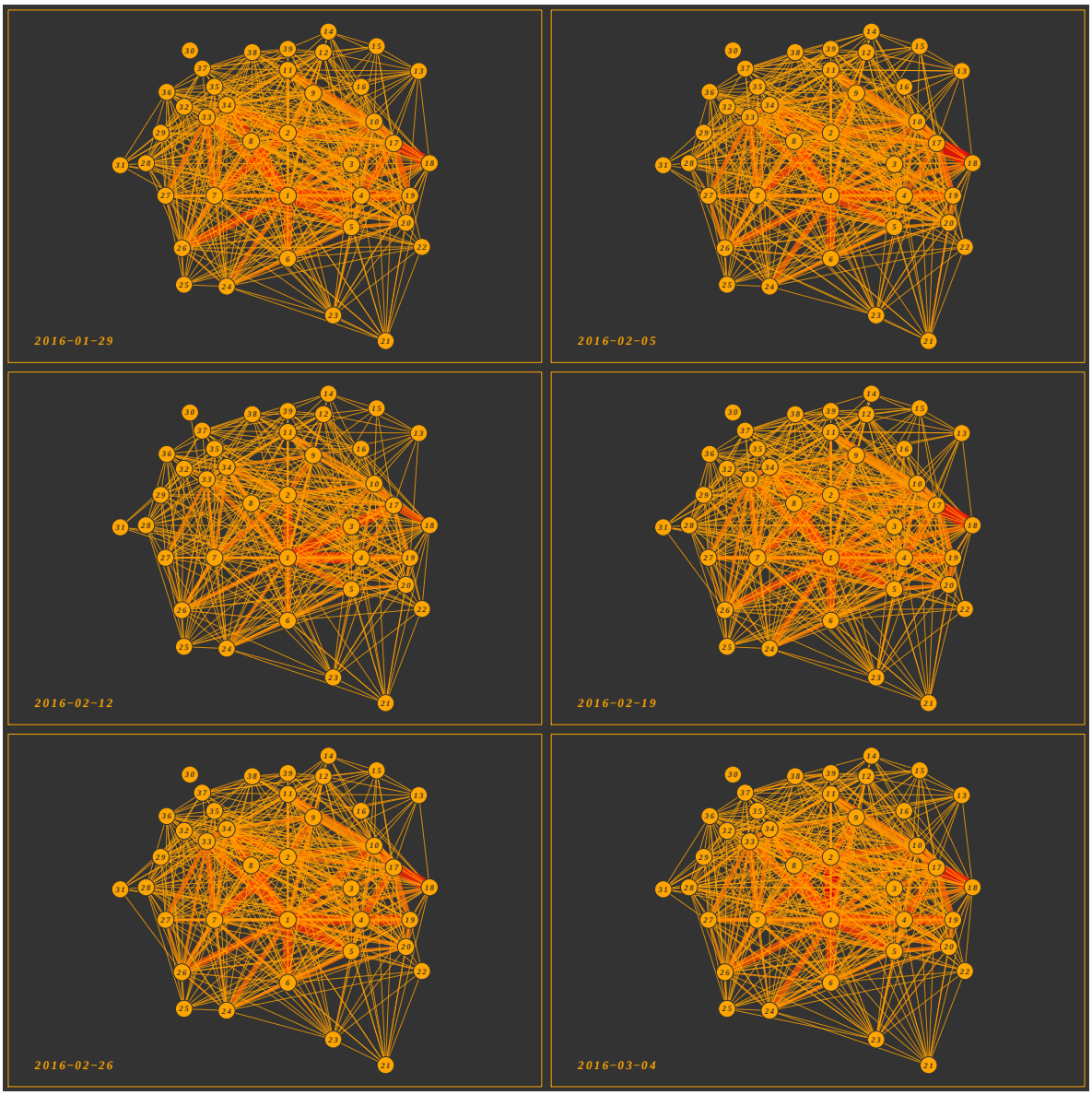


Figure 6.1.5: Networks representing the trips between NILs of Milan on Friday.

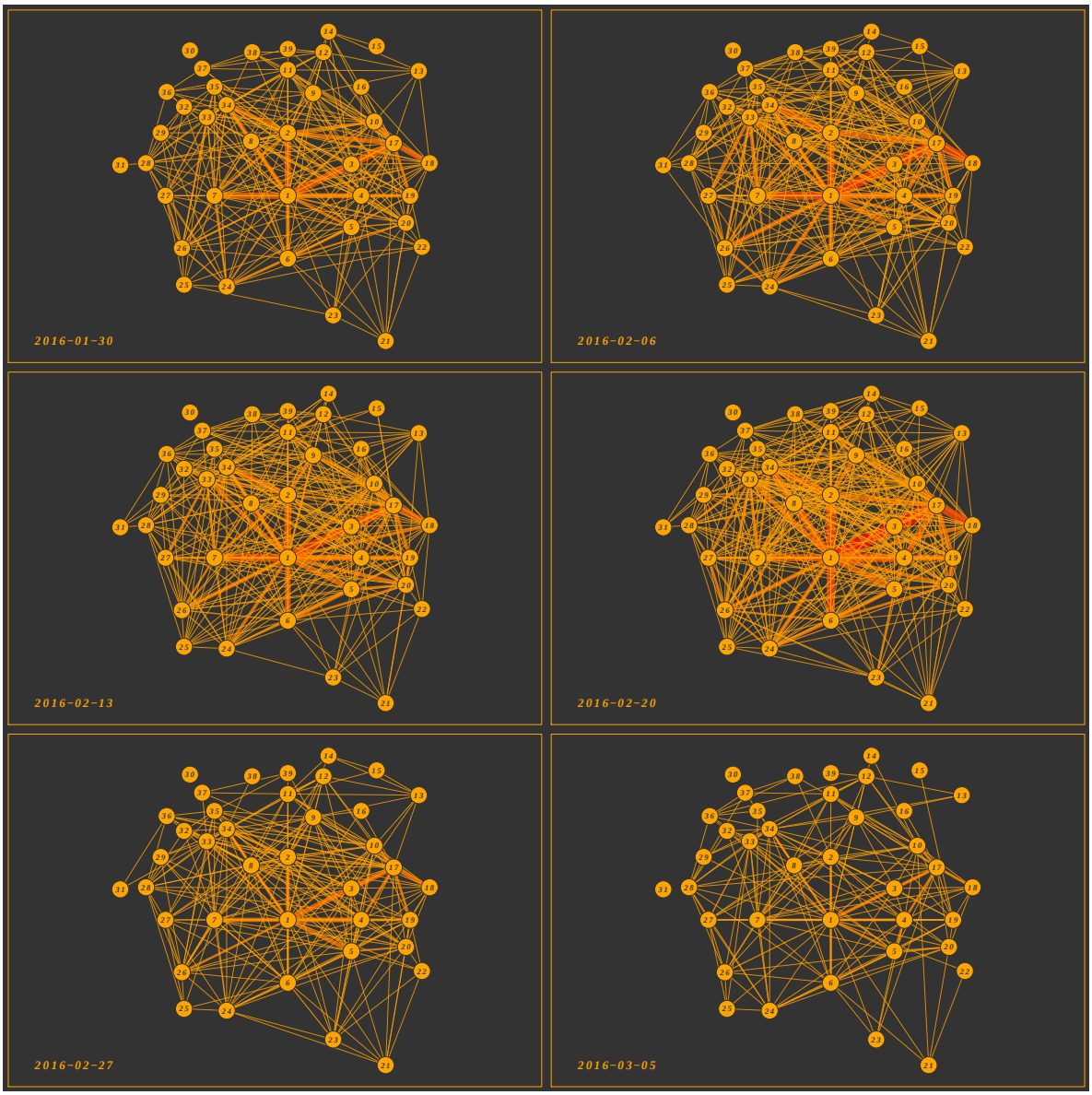


Figure 6.1.6: Networks representing the trips between NILs of Milan on Saturday.

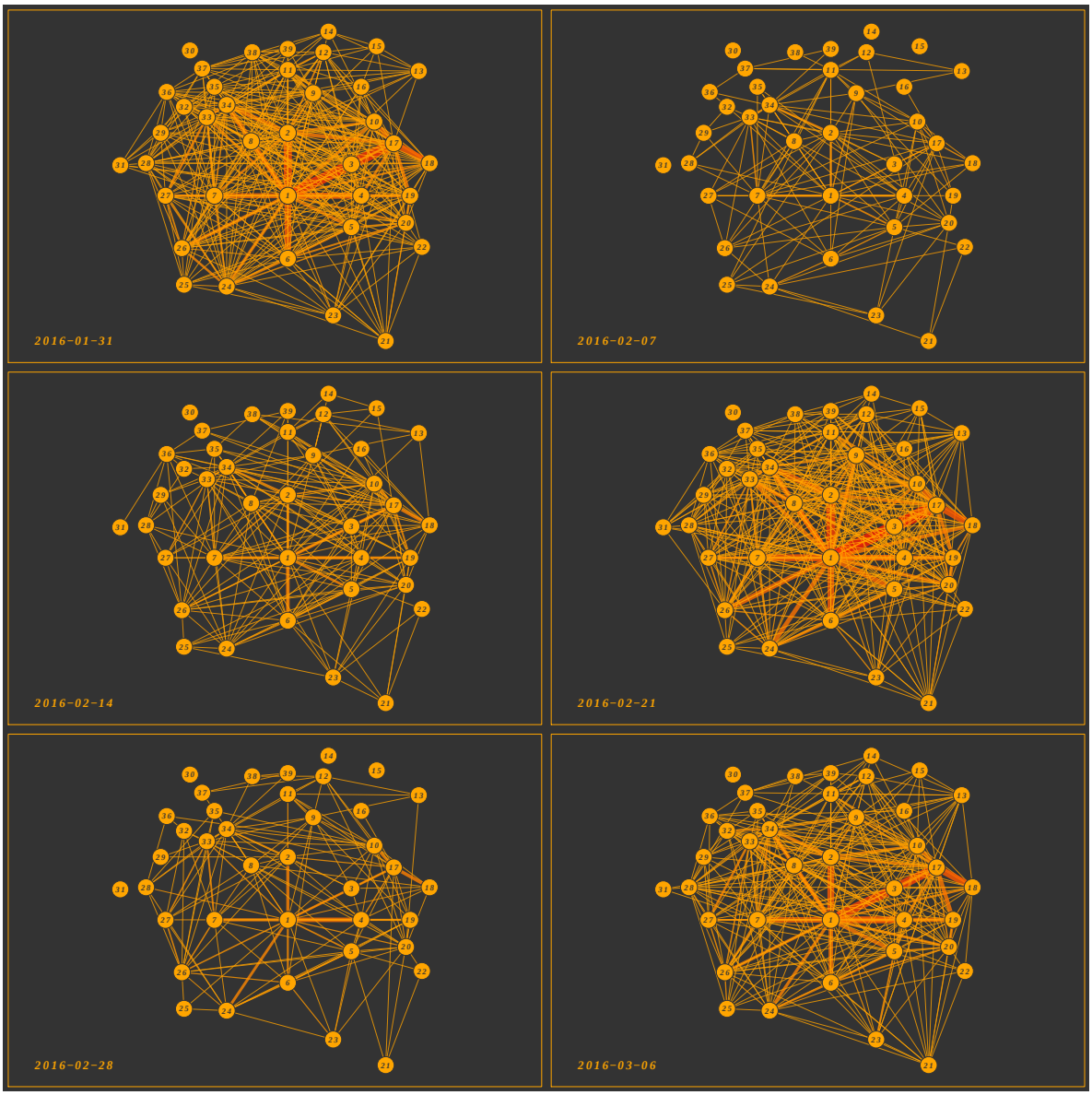


Figure 6.1.7: Networks representing the trips between NILs of Milan on Sunday.

References

[1] M. E. Airoldi, X. Baib, and K. M. Carley. Network sampling and classification: an investigation of network model representations. *Decis Support Syst.*, 51(3):506–518, 2011.

[2] B. Aydin, G. Pataki, Wang H., E. Bullitt, and J. S. Marron. A principal component analysis for trees. *Ann. Appl. Stat.*, 3:1597–1615, 2009.

[3] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286: 509–512, 1999.

[4] A. Barberán, Bates S. T., E. O. Casamayor, and N. Fierer. Using network analysis to explore co-occurrence patterns in soil microbial communities. *ISME J.*, 6(2):343–351, 2012.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[6] R. Bhattacharya and L. Lin. Omnibus central limit theorems for fréchet means and nonparametric inference on non-euclidean spaces. *Proc. Amer. Math. Soc.*, 145:413–428, 2017.

[7] M. Biswas and A. K. Ghosh. A nonparametric two-sample test applicable to high dimensional data. *J. Mult. Anal.*, 123:160–171, 2014.

[8] B. Bollobas. *Random Graphs, Second Edition*. Cambridge University Press, Cambridge, 2001.

[9] C. Brombin and L. Salmaso. Multi-aspect permutation tests in shape analysis with small sample size. *Computational Statistics and Data Analysis*, 53:3921–3931, 2009.

[10] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci 2009*, 10:186 – 198, 2009.

[11] A. Cabassi, D. Pigoli, P. Secchi, and P. A. Carter. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electron. J. Statist.*, 11(2):3815–3840, 2017. doi: 10.1214/17-EJS1347.

[12] G. Chartrand, F. Saba, and H. B. Zou. Edge rotations and distance between graphs. *Casopis pro pestovani matematiky*, 110(1):87–91, 1985.

[13] H. Chen and J. H. Friedman. A new graph-based two-sample test for multivariate and object data. *J. Am. Statist. Ass.*, 112:397–409, 2017.

[14] F. Comellas and J. Diaz-Lopez. Spectral reconstruction of complex networks. *Physica A*, 387: 6436––6442, 2008.

[15] P. B. Crino, K. L. Nathanson, and E. P. Henske. The tuberous sclerosis complex. *N Engl J Med*, 356, 2006.

[16] R. Diestel. *Graph Theory: 5th edition*. Springer Graduate Texts in Mathematics. 2017. ISBN 9783961340057.

[17] S. Dirmeier. *diffusr*. R Foundation for Statistical Computing, 2017.

[18] I. L. Dryden and K. V. Mardia. *Statistical Analysis of Shape*. Wiley, New York, 1998.

[19] I. L. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with application to diffusion tensor imaging. *Ann. Appl. Stat.*, 3(3):1102–1123, 2009.

[20] D. Durante and D. B. Dunson. Bayesian inference and testing of group differences in brain networks. *Bayesian Anal.*, 13(1):29–58, 2018.

[21] D. Durante, D. B. Dunson, and J. T. Vogelstein. Nonparametric bayes modeling of populations of networks. *J. Am. Statist. Ass.*, 112(520):1516–1530, 2017. doi: 10.1080/01621459.2016.1219260.

[22] M. Dwass. Modified Randomization Tests for Nonparametric Hypotheses. *Ann. Math. Statist.*, 28(1):181–187, 1957. ISSN 0003-4851. doi: 10.1214/aoms/1177707045.

[23] P. Erdös and A. Rényi. On random graphs. i. *Publicationes Mathematicae*, 6:290–297, 1959.

[24] A. Fornito, A. Zalesky, and M. Breakspear. Graph analysis of the human connectome: Promise, progress, and pitfalls. *NeuroImage*, 80:426–444, 2013.

[25] A. P. Fournel, E. Reynaud, M. J. Brammer, A. Simmons, and C. E. Ginestet. Group analysis of self-organizing maps based on functional MRI using restricted Fréchet means. *NeuroImage*, 76:373–385, 2013.

[26] J. H. Friedman and L. C. Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Ann. Statist.*, 7(4):697–717, 1979.

[27] C. E. Ginestet, J. Li, P. Balanchandran, S. Rosenberg, and E. D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *Ann. Appl. Stat.*, 11(2):725–750, 2017.

[28] R. Guimerá and L. A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028):895–900, 2005.

[29] P. Hall and N. Tajvidi. Permutation test for equality of distribution in high dimensional settings. *Biometrika*, 89(2):359–374, 2002.

[30] R. W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29(2):147–160, 1950. doi: 10.1002/j.1538-7305.1950.tb00463.x.

[31] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5:109–137, 1983.

[32] E. B. Jarret. Edge rotation and edge slide distance graphs. *Computers Math. Applic.*, 34(11): 81–87, 1997.

[33] A. E. Krause, K. A. Frank, D. M. Mason, Ulanowicz R. E., and W. W. Taylor. Compartments revealed in food-web structure. *Nature*, 426(6964):282–285, 2003.

[34] H. Li and A. Wenston. Strict p-negative type of a metric space. *Positivity*, 14(3):529–545, 2010.

[35] Z. Liu and R. Modarres. A triangle test for equality of distribution functions in high dimensions. *Journal of Nonparametric Statistics*, 23(3):605–615, 2011.

[36] N. Locantore, J. S. Marron, D. G. Simpson, N. Tripoli, J. T. Zhang, K. L. Choen, G. Boente, R. Fraiman, B. Brumback, and C. Croux. Robust principal component analysis for functional data. *Test*, 8:1–73, 1999.

[37] I. Lovato, A. Pini, A. Stamm, and S. Vantini. Multiscale null hypothesis testing for network-valued data: analysis of brain functional networks of autistic subjects. in preparation.

[38] I. Lovato, A. Stamm, A. Pini, S. Vantini, B. Scherrer, S. K. Warfield, and M. Taquet. A new statistical test to detect differences in samples of brain networks. in preparation.

[39] I. Lovato, A. Pini, A. Stamm, and S. Vantini. Model-free two-sample test for network-valued data. submitted.

[40] R. Lyons. Distance covariance in metric spaces. *Ann. Probab.*, 41(5):3284–3305, 2013.

[41] J.-F. Maa, D. K. Pearl, and R Bartoszynsk. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. *Ann. Statist.*, 24(3):1069–1074, 1996.

[42] R. Marcus, E. Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[43] K. V. Mardia. *Statistics of Directional Data*. Academic Press London, UK, 1972.

[44] J. S. Marron and A. M. Alonso. Overview of object oriented data analysis. *Biom. J.*, 56: 732–753, 2014.

[45] J. Moody. Race, school integration, and friendship segregation in america. *Am J Sociol.*, 107: 679–716, 2001.

[46] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45(2): 167–256, 2003.

[47] M. E. J. Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23):8577–8582, 2006.

[48] M. E. J. Newman. *Networks: an introduction*. Oxford University Press, UK, 2010.

[49] T. W. Nye, X. Tang, G. Weyenberg, and R. Yoshida. Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. *Biometrika*, 104(4):901–922, 2017.

[50] R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. *Phys Rev Lett.*, 86(14):3200–3203, 2001.

[51] F. Pesarin and L. Salmaso. *Permutation Tests for Complex Data*. Wiley, 2010.

[52] J. M. Peters, M. Taquet, C. Vega, S. S. Jeste, I. S. Fernández, J. Tan, C. A. Nelson, M. Sahin, and S. K. Warfield. Brain functional networks in syndromic and non-syndromic autism: a graph theoretical study of EEG connectivity. *BMC Medicine*, 54(11), 2013.

[53] B. Phipson and G. K. Smyth. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Stat Appl Genet Mol Biol.*, 9(1):1–12, 2010. ISSN 1544-6115. doi: 10.2202/1544-6115.1585.

[54] D. Pigoli, J. A. D. Aston, I. L. Dryden, and P. Secchi. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 2014.

[55] A. Pini and S. Vantini. Interval-wise testing for functional data. *Journal of Nonparametric Statistics*, DOI:10.1080/10485252.2017.1306627, 2017.

[56] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[57] P. M. Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Statist. Soc. B*, 67:515–530, 2005.

[58] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059 – 1069, 2010. Computational Models of the Brain.

[59] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

[60] L. M. Sangalli, P. Secchi, and S. Vantini. Object oriented data analysis: a few methodological challenges. *Biom. J.*, 56:774–777, 2014.

[61] B. Scherrer, K. Kapur, A. K. Prohl, M. Taquet, J. M. Peters, X. Tomas-Fernandez, P. E. Davis, M. Bebin, D. A. Krueger, H. Northrup, J. Wu, M. Sahin, and S. K. Warfield. The connectivity fingerprint of the fusiform gyrus encodes the risk of developing autism in infants with tuberous sclerosis complex. submitted.

[62] S. J. Shafali, M. Sahin, P. Bolton, G. B. Ploubidis, and A. Humphrey. Characterization of autism in young children with tuberous sclerosis complex. *Journal of Child Neurology*, 23(5): 520–525, 2008.

[63] S. L. Simpson, R. G. Lyday, S. Hayasaka, A. P. Marsh, and P. J. Laurienti. A permutation testing framework to compare groups of brain networks. *Front Comput Neurosci.*, 7:171, 2013.

[64] S. L. Simpson, F. D. Bowman, and P. J. Laurienti. Analyzing complex functional brain networks: fusing statistics and network science to understand the brain. *Stat Surv*, 7:1–36, 2014.

[65] G. J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5 (16.10), 2004.

[66] G. J. Székely and M. L. Rizzo. A new test for multivariate normality. *J. Mult. Anal.*, 93:58–80, 2005.

[67] G. J. Székely and M. L. Rizzo. Hierarchical clustering via joint between-within distances: extending ward's minimum variance method. *Journal of Classification*, 22:151–183, 2005.

[68] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *J. Statist. Plann. Inference*, 143:1249–1272, 2013.

[69] M. Taquet, R. Gong, and S. K. Warfield. Contradictory conclusions in analysis of brain functional networks: the role of image registration. *"Organization for Human Brain Mapping (OHBM)", Seattle, USA*, 2013.

[70] L. H. C. Tippett. *The Methods of Statistics*. Williams & Norgate, London, 1931.

[71] O. Vsevolozhskaya, M. Greenwood, and D. Holodov. Pairwise comparison of treatment levels in functional analysis of variance with application to erythrocyte hemolysis. *The Annals of Applied Statistics*, 8:905–925, 2014.

[72] H. Wang and J. S. Marron. Object oriented data analysis: sets of trees. *Ann. Statist.*, 35: 1849–1873, 2007.

[73] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 6 (6684):440–442, 1998.

[74] S. Wei, C. Lee, L. Wichers, and J. S. Marron. Direction-projection-permutation for high-dimensional hypothesis tests. *Journal of Computational and Graphical Statistics*, 25(2): 549–569, 2016. doi: 10.1080/10618600.2015.1027773.

[75] B. Zelinka. On a certain distance between isomorphism classes of graphs. *Casopis pro pestovani matematiky*, 100(4):371–373, 1975.

[76] B. Zelinka. Edge shift distance between trees. *Archivum Mathematicum*, 28(1-2):5–9, 1992.