Model-free Two-sample Test for Network-Valued Data

Ilenia Lovato[☆]

Department of Mathematics, University of Pavia
C.so Strada Nuova 65, 27100 Pavia, Italy

Alessia Pini

Department of Statistical Sciences, Università Cattolica del Sacro Cuore
Largo A. Gemelli 1, 20123, Milano, Italy

Aymeric Stamm

Laboratoire de Mathmatiques Jean Leray
UMR CNRS 6629, Nantes, France

Simone Vantini

MOX - Department of Mathematics, Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy

Abstract

In the framework of Object Oriented Data Analysis, a permutation approach to the two-sample testing problem for network-valued data is proposed. In details, the present framework proceeds in four steps: (i) matrix representation of the networks, (ii) computation of the matrix of pairwise (inter-point) distances, (iii) computation of test statistics based on inter-point distances and (iv) embedding of the test statistics within a permutation test. The proposed testing procedures are proven to be exact for every finite sample size and consistent. Two new test statistics based on inter-point distances (i.e., IP-Student and IP-Fisher) are defined and a method to combine them to get a further inferential tool (i.e., IP-StudentFisher) is introduced. Simulated data shows that tests with our statistic exhibit a statistical power that is either the best or second-best but very close to the best on a variety of possible

[☆]Corresponding author
Email address: ilenia.lovato01@universitadipavia.it (Ilenia Lovato[☆])

alternatives hypotheses and other statistics. A second simulation study that aims at better understanding which features are captured by specific combinations of matrix representations and distances is presented. Finally, a case study on mobility networks in the city of Milan is carried out. The proposed framework is fully implemented in the `R` package `nevada` (NEtwork-VAlued Data Analysis).

Keywords: Network-valued data, Null-hypothesis testing, Object-oriented data analysis, Permutation test, Shared mobility

---

1. Introduction

Object Oriented Data Analysis (OODA) is a field of growing interest that emerged from the seminal paper of Wang and Marron (2007). It aims at conducting statistical analyses of complex data that cannot be embedded in the standard Euclidean framework (see Marron and Alonso, 2014, 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 (Dryden and Mardia, 1998), images (Locantore et al., 1999; Wei et al., 2016), manifold-valued data such as directional data (Mardia, 1972), trees (Wang and Marron, 2007), covariance matrices and operators (Dryden et al., 2009; Pigoli et al., 2014), density functions (Menafoglio and Secchi, 2017) 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 Sangalli et al., 2014).

In this work, we focus on a specific type of object data, namely networks. 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, Newman, 2010, 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 cor-

responding 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 (Moody, 2001), species in a food web (Krause et al., 2003), nodes on a computer network (Pastor-Satorras and Vespignani, 2001), or proteins in metabolic pathways (Guimerá and Nunes Amaral, 2005). Network analysis is used to explore the mathematical, statistical and structural properties of a set of items (nodes) and the connections between them (edges; Newman (2003))" (Barberán et al., 2012). 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 (Simpson et al., 2014).

Only recently, some proposals have been made in this direction. The first papers on statistical methodologies that investigate network-valued data ap-

peared as a response to neuroimaging problems. Specifically, Wang and Marron (2007) and Aydin et al. (2009) 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, Nye et al. (2017) 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 characterizing 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 Rubinov and Sporns (2010) 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 first attempt to account for the entire network structures when applying null hypothesis significance testing procedures can be found in Simpson et al. (2013). 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 Ginestet et al. (2017) is a cornerstone paper moving in the direction of network-valued data. Motivated by a problem of functional neuroimaging investigation, the authors 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 Bhattacharya and Lin (2017) 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, Durante et al. (2017) 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 Durante et al. (2017) recently published on JASA. Durante and Dunson (2018) 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.

Recently, Chen and Friedman (2017) propose a new graph-based two-sample test for multivariate and object data that is able to detect difference both in location and in scale and that can be applied if a similarity measure between observations can be defined. This paper has its root in the paper of Friedman and Rafsky (1979), where the authors proposed a non-parametric two-sample test based on a minimum spanning tree constructed using the pairwise distances among the pooled observations. The test is then based on a count statistic on the number of edges that connect observations from different samples. Other similarity graphs can be used in this framework: minimum spanning tree with higher density (Friedman and Rafsky, 1979), $k-$nearest neighbour graphs (Schilling, 1986; Henze, 1988), minimum distance non-bipartite pairing tree (Rosenbaum, 2005). See Chen and Friedman (2017) for a complete and exhaustive literature review. From a practical point of view, none of these test is sensitive to differences both in location and in scale. To overcome this limit, Chen and Friedman (2017) propose a new test statistic that works better than other tests for both location and

scale alternatives and for location-scale alternative, while the power of the test is still dependent from the chosen similarity graph and its density.

In this work, we propose a finite-sample exact and consistent permutation-based two-sample test for making inference on distributions of network-valued data. 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 paper, we detail a number of possible representations, distances and statistics. It is straightforward to add more of them into the framework as well.

The paper is organized as follows. Section 2 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 (Pesarin and Salmaso, 2010, chap. 4), exhibit the best performances in testing equality of distributions of networks. The two novel test statistics we introduce target specific moments of the distribution – thus are easily interpretable – and, when jointly used through non-parametric combination, make the test more sensitive to global differences in distributions. 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 3 and 4, we report results from simulation studies and an application to real data pertaining to the bike sharing service in Milan, respectively. Our statistical framework for inference on network-valued data has been implemented in the R (R Core Team, 2016) package `nevada`, available on GitHub (`https://github.com/astamm/nevada`).

2. Statistical framework for network-valued data

2.1. 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:

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

where δ_{ij} is the Kronecker symbol ($\delta_{ij} = 1$ if $i = j$ or 0 otherwise). This matrix takes its name from the so-called *heat equation* which reads $\partial u/\partial t - \alpha \nabla^2 u$, where ∇^2 is the Laplacian operator. Indeed, the Laplacian matrix is nothing but the discretized version of ∇^2 on the set of vertices (see Newman, 2010, 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 eigenvalues which are zero 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 paper 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 (Newman, 2006).

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.

2.2. 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 (Zelinka, 1975). Then, in the 90's, a number of statistics emerged around the concepts of edge rotation or slide (Chartrand et al., 1985; Zelinka, 1992; Jarret, 1997). In essence, an edge rotation pertains to replacing one of the two end-vertices of an edge by another vertex, keeping the other end-vertex fixed. An edge slide is a particular type of edge rotation: the moving end-vertex of the edge can be replaced only by adjacent vertices. 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 paper we instead take advantage of the matrix representation of a network and consider instead distances that have been recently proposed

either on network matrices (Comellas and Diaz-Lopez, 2008) or on covariance matrices (Dryden et al., 2009), 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:

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

This distance takes its name after Richard Hamming who needed a way to detect errors in systems (Hamming, 1950). 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:

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

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 (Dryden et al., 2009; Pigoli et al., 2014; Cabassi et al., 2017), 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, reduces to the Laplacian matrix.

2.3. Test statistics based on inter-point distances

Let $\mathcal{G}_1$ and $\mathcal{G}_2$ be two samples of networks with the same set of vertices V 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). Next, object data are often embedded in complex spaces into which there is no closed-form expression of the sample Fréchet mean (Pigoli et al., 2014). It is possible to circumvent this problem either by computing it numerically or by resorting to restricted sample Fréchet means as done by Fournel et al. (2013) 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 (Székely and Rizzo, 2013) and have been generalized to separable Hilbert spaces (Lyons, 2013). The test based on the energy distance statistic is a special case of the kernel-based two-sample tests proposed in Gretton et al. (2012). The latter relies on a test statistic called maximum mean discrepancy which measures the distance between two probability distributions embedded in a topological (possibly infinite-dimensional) space. 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 (Maa et al., 1996). The easiest statistic along those lines has been proposed by Biswas and Ghosh (2014) 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, Friedman and Rafsky (1979) uses the minimum spanning tree while Rosenbaum (2005) uses the minimum distance non-bipartite pairing tree. Chen and Friedman (2017) 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 (Hall and Tajvidi, 2002; Liu and Modarres, 2011) 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:

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

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

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

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.

2.4. 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 Ginestet et al. (2017) 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 (Pesarin and Salmaso, 2010), 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 (Pesarin and Salmaso, 2010). 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 Phipson and Smyth (2010), 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 (Phipson and Smyth, 2010). Hence, the choice of m only impacts the power of the test, as expected. This p-value is computed as follows (Dwass, 1957; Phipson and Smyth, 2010):

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

where F is the cumulative probability function of the binomial distribution, $b(T)$ is the number of t_{perm} greater than t_{obs}, m is the number of randomly sampled permutations, m_t is the total number of possible permutations and b_t is the index of summation. 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 3. 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. It is useful to recall that, if $(\mathcal{X}, \rho)$ is a metric space and P is a probability measure, the Fréchet mean of P is defined as

$$\arg\min_{x\in\mathcal{X}} \int_{\mathcal{X}} \rho^2(x, y)P(dy).$$

If $Y = \{Y_1, \cdots, Y_n\}$ is a sample drawn from P, the sample Fréchet mean is

$$\arg\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^{n} \rho^2(x, Y_i),$$

and the minimum value of the summation is the sample Fréchet variance. See Jain (2016) for an exhaustive generalisation of the concept of sample Fréchet mean and its properties to graph spaces. The following theorems hold:

Theorem 2.1. *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 Fréchet means (with respect to distance* ρ*), namely* $P_{H_1}\left[p(T_{\text{IP}-\text{Student}}) \leq \alpha\right] \to 1$ *as* $n_1 + n_2 \to \infty$.

Theorem 2.2. *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 Fréchet variances (with respect to distance* ρ*), namely* $P_{H_1}\left[p(T_{\text{IP}-\text{Fisher}}) \leq \alpha\right] \to 1$ *as* $n_1 + n_2 \to \infty$.

Proofs of Theorem 2.1 and Theorem 2.2 are reported in Appendix A.

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 (Brombin and Salmaso, 2009; Pesarin and Salmaso, 2010).

Given a set of B randomly chosen 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 replacing the component values by their rank (sorting the values in decreasing order) divided by $B + 1$. This boils down to computing, for each component of the concatenated vectors, the permutational p-value where the corresponding permuted data is considered as the observed one. This effectively produces two vectors $\boldsymbol{\pi}_{\text{IP}-\text{Student}}$ and $\boldsymbol{\pi}_{\text{IP}-\text{Fisher}}$ of "intermediate p-values", which are hence on the same scale and thus comparable. Next, we combine $\boldsymbol{\pi}_{\text{IP}-\text{Student}}$ and $\boldsymbol{\pi}_{\text{IP}-\text{Fisher}}$ 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 (Pesarin and Salmaso, 2010). 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 $\psi(x, y) = 1 - \min(x, y)$ (Tippett, 1931) 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. $P_{H_1}[p(T) \leq \alpha] \geq P_{H_0}[p(T) \leq \alpha] = \alpha$) and at least one of them is consistent (see Pesarin and Salmaso, 2010, chap. 4). Specifically, we have the following result:

Corollary 2.1. *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* $P_{H_1}\left[p(T_{\text{IP}-\text{StudentFisher}}) \leq \alpha\right] \to 1$ *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 Pesarin and Salmaso, 2010, chap. 4).

3. Simulation studies

3.1. 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 (see Table *Simulation 1* in Appendix B for details on the specific parameters that have been used to that effect). 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 Chen and Friedman (2017). 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.

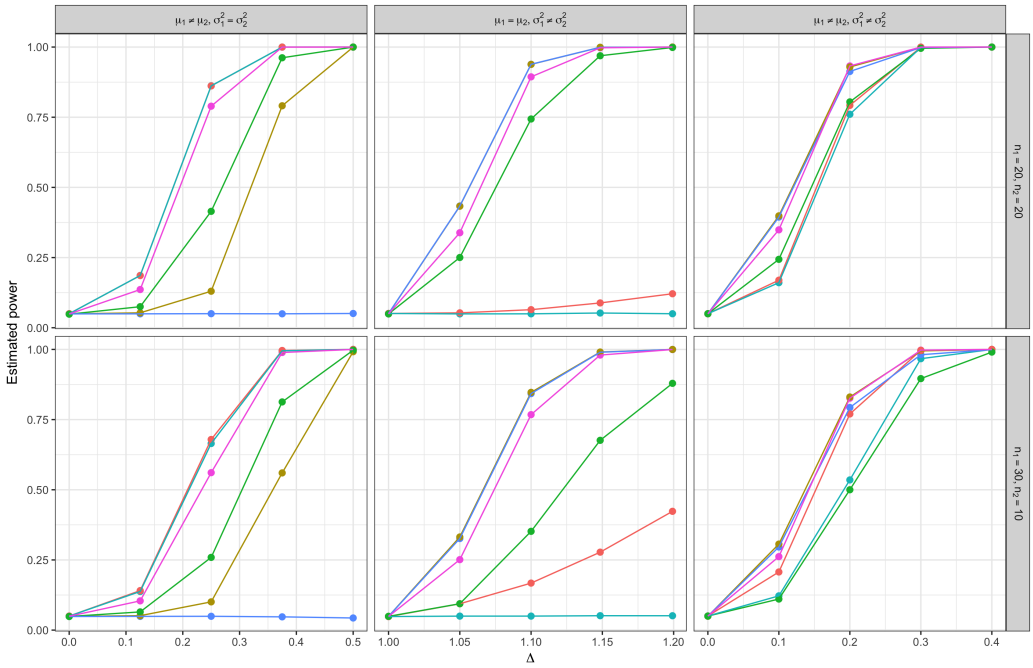

Figure 1: Power of the test using different test statistics: T_{SR} (2.2) in red, T_{BG} (2.3) in brown, T_{CF} in green, $T_{\mathrm{IP-Student}}$ (2.4) in light blue, $T_{\mathrm{IP-Fisher}}$ (2.4) in blue and $T_{\mathrm{IP-StudentFisher}}$ in pink. The largest Monte Carlo standard error is 0.00158. If two curves coincide, the dots of one curve and the line of the other curve are shown.

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

3.2. 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. The edge weights are i.i.d. sampled.

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 (Holland et al., 1983) 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 Bollobas, 2001, sec. 2.4) that generates random networks in which all vertices have the same degree and we drew S2 from the $G(n, p)$ Erdős-Renyi model (Erdös and Rényi, 1959) 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. Airoldi et al., 2011), 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 (Watts and Strogatz, 1998) 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 (Barabási and Albert, 1999) with quadratic preferential attachment and 4 edges added at each time step.

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 $\alpha = 0{\cdot}05$ and we performed a total of 100,000 Monte-Carlo runs. For each run, we performed the test with the statistic $T_{\mathrm{IP-StudentFisher}}$ 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. Modeling details of the generated scenarios are summarized in Appendix B.

Figure 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. 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 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 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

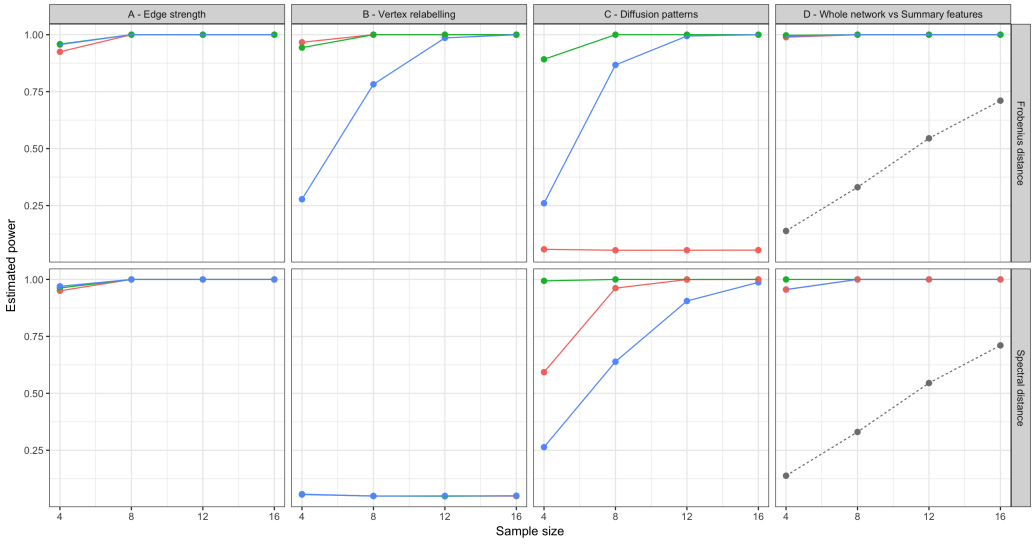

Figure 2: Power of the test under different representations (adjacency in red, Laplacian in green, modularity in blue), different distances (rows) and different scenarios (columns). The test is conducted under the combined test statistic $T_{\mathrm{IP-StudentFisher}}$. The dashed grey curve in Scenario D (last column) represents the statistical power achieved by considering only the clustering coefficient. The largest Monte Carlo standard error is 0.00158.

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` (Dirmeier, 2017) that nicely shows that diffusion along the networks is different in the two samples. The fourth column in Fig. 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 3.1. *Scenario A shows the performance of our test when the edge weights are i.i.d. sampled from a discrete distribution. We performed a similar simulation study (not reported in the paper) drawing the edge weights i.i.d. from the Exponential distribution to see how the test perform on a continuous edge weight distribution. The results are similar to those of Scenario A.*

Remark 3.2. *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.*

4. 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.. 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 3 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

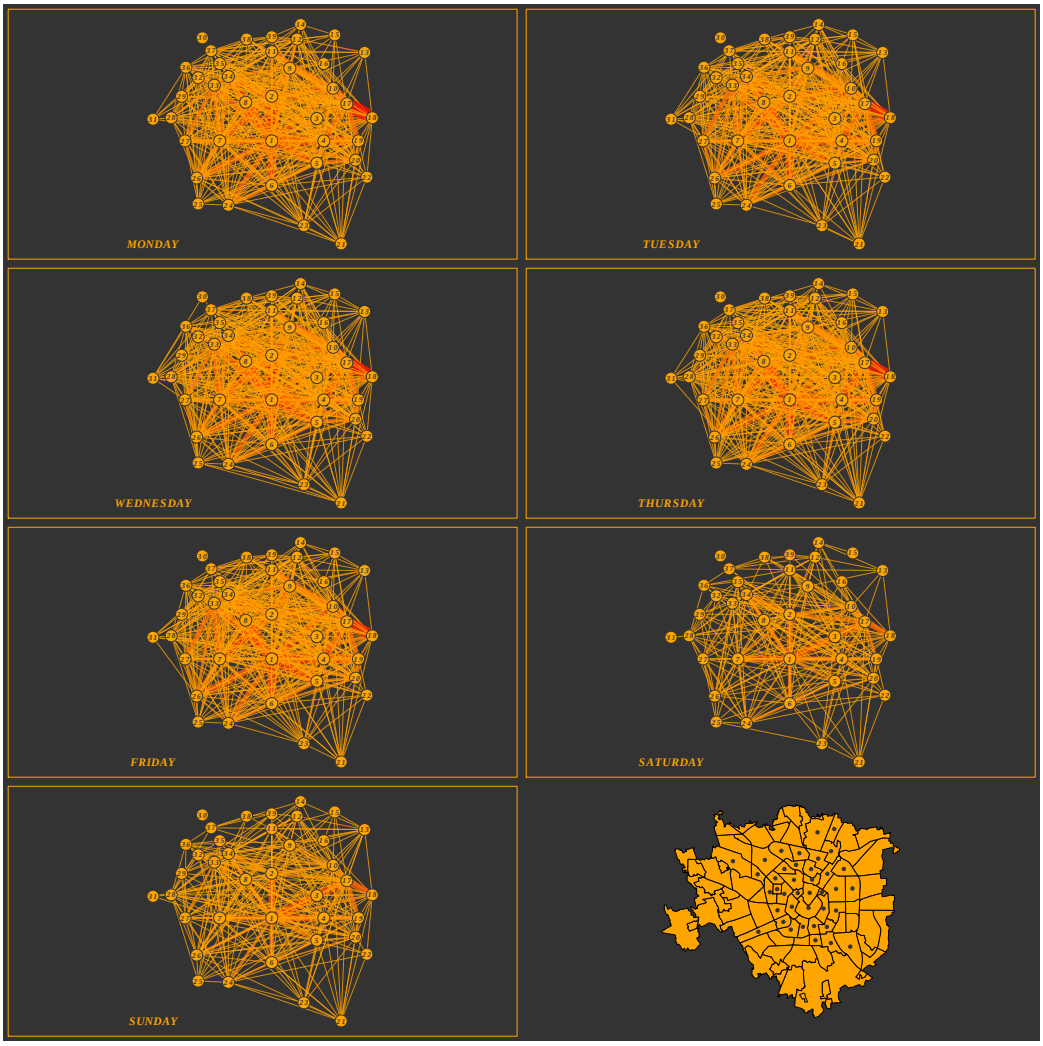

Figure 3: Restricted sample Fréchet means of each day of the week and, in the last thumbnail (bottom right), the map of the NILs of Milan with a point in the neighbourhoods having at least one dock station.

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 and 2.2. Figure 4 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, 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. 4. 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.

5. Discussion

Flexibility in choosing representation, distance and test statistics is a strength of our proposed comprehensive framework. In effect, different choices of the pair representation/distance allow to analyze the data set under different angles, focusing therefore on different types of differences. For example, the difference between two samples of networks may be in the intensity of the weights of the networks – in which case using the adjacency matrix representation is preferable – or in the way a substance can diffuse along the nodes of the networks – in which case the Laplacian matrix representation shall be used. For the choice of distances, similar arguments can be made. For example, while the Frobenius distance is the natural Euclidean distance for matrices, we might want to be insensitive to differences that are only due to a relabelling of the nodes – in which case the spectral distance should be considered. In definitive, the practical user who knows the type of difference (s)he expects can target specific representations and distances. If, on

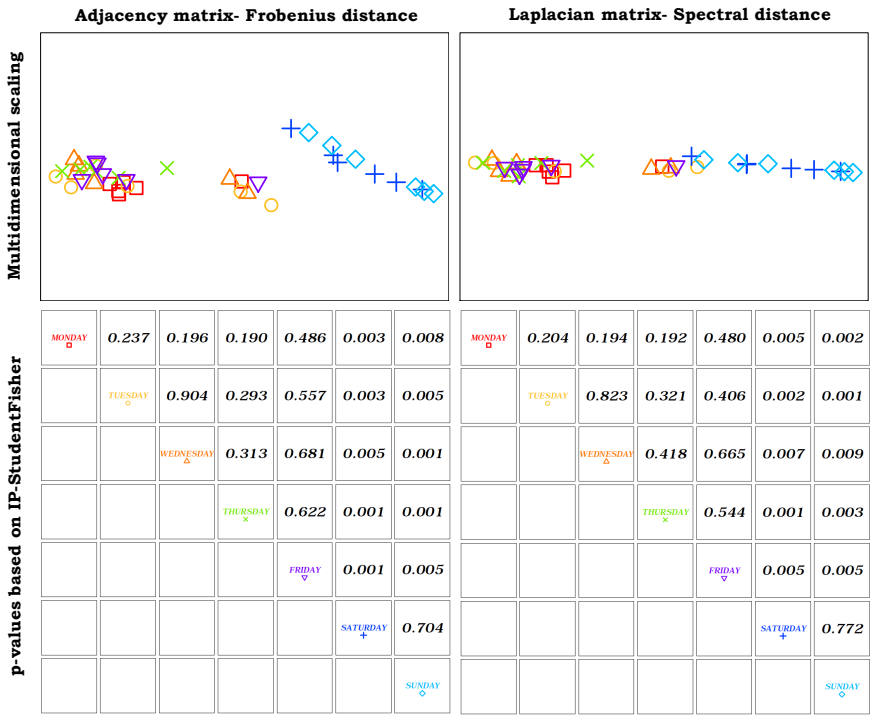

Figure 4: Results of the application to the bikeMi data set using different matrix representations and distances.

the contrary, the practical user is conducting a purely exploratory investigation, we strongly believe it is still a strength of the framework to allow testing for differences under different angles to make new discoveries about the phenomenon under investigation from the data set at hand. Depending on the results that are obtained using the different representations and distances, the user can obtain a clear idea of what type of differences (if any) are present.

Tackling the two-sample testing problem from the perspective of the permutation framework assumes as null hypothesis that the entire distribution of the two samples 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 statistics that look at the distribution under different angles. We advocate in favor of the permutation non-parametric combination approach which allows to integrate multiple test statistics in the estimation of the p-value. Along these lines, we define and propose a set of novel test statistics based on inter-point distances only that individually target each moment of the distribution. In the `nevada` package available freely on GitHub, we thus set this approach as the default for the test statistic(s). For the sake of completeness and because many other statistics exist and might be used, we also give in the package flexibility to the practical user to use other statistics taken from the literature or even to implement her own preferred statistic(s).

Starting from standard results on U–statistics, it could be possible to find the asymptotic distributions of $T_{\mathrm{IP-Student}}$ and $T_{\mathrm{IP-Fisher}}$. 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 the 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.

Appendix A. Proof of the theorems

Theorem 1. In this proof we partially follow Székely and Rizzo (2004). For the law of large numbers, we have that for $n = n_1 + n_2 \to \infty$

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

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

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

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

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

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 Székely and Rizzo (2005a)[Theorem 2], it is possible to prove that the expression in A.1 is always non–negative and it is equal to zero if and only if $E[G_1] = E[G_2]$ (where the expectation is defined in the Fréchet sense). In effect, the representation of a network by means of a matrix allows to trace it back to the vectorization of the representation matrix. Hence, all considered distances can be seen as Euclidean distances between properly defined vectors. Indeed: the Frobenius distance is nothing but the Euclidean distance on the vectorized matrix representation; The Spectral distance is the Euclidean distance between the vectors of the eigenvalues of the representation matrix; and the Root-Euclidean distances is an Euclidean distance between the vectorization of the square roots of the matrix representations. Therefore Székely and Rizzo (2005a)[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, Székely and Rizzo (2005b)[Theorem 1] guarantees a similar result to that of Székely and Rizzo (2005a)[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 (Li and Wenston, 2010) 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. 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 (Pesarin and Salmaso, 2010).

□

Appendix B. Modeling details of the generated scenarios

Table *Simulation 1* on page 29 reports all the parameters used to generate the simulated scenarios in Subsection 3.1.

Simulation 1: Parameters of the independent Binomial distributions and corresponding location and scale alternatives.

	n_1	p_1	n_2	p_2	μ_1	μ_2	σ_1^2	σ_2^2	$\|\mu_1 - \mu_2\|$	σ_2^2 / σ_1^2
LOCATION-ONLY	10	0.50000	10	0.50000	5.00000	5.00000	2.50000	2.50000	0.00000	1.00000
	10	0.50625	10	0.49375	5.06250	4.93750	2.50000	2.50000	0.12500	1.00000
	10	0.51250	10	0.48750	5.12500	4.87500	2.50000	2.50000	0.25000	1.00000
	10	0.51875	10	0.48125	5.18750	4.81250	2.50000	2.50000	0.37500	1.00000
	10	0.52500	10	0.47500	5.25000	4.75000	2.50000	2.50000	0.50000	1.00000
SCALE-ONLY	300	0.20000	300	0.20000	60.00000	60.00000	48.00000	48.00000	0.00000	1.00000
	300	0.20000	375	0.16000	60.00000	60.00000	48.00000	50.40000	0.00000	1.05000
	300	0.20000	500	0.12000	60.00000	60.00000	48.00000	52.80000	0.00000	1.10000
	300	0.20000	750	0.08000	60.00000	60.00000	48.00000	55.20000	0.00000	1.15000
	300	0.20000	1500	0.04000	60.00000	60.00000	48.00000	57.60000	0.00000	1.20000
LOCATION & SCALE	20	0.10000	20	0.10000	2.00000	2.00000	1.80000	1.80000	0.00000	1.00000
	20	0.10000	21	0.10000	2.00000	2.10000	1.80000	1.89000	0.10000	1.05000
	20	0.10000	22	0.10000	2.00000	2.20000	1.80000	1.98000	0.20000	1.10000
	20	0.10000	23	0.10000	2.00000	2.30000	1.80000	2.07000	0.30000	1.15000
	20	0.10000	24	0.10000	2.00000	2.40000	1.80000	2.16000	0.40000	1.20000

Table B.1 on page 30 reports all the parameters used to generate the simulated scenarios in Subsection 3.2. Scopes, models and their parameters for the two samples S1 and S2 are summarized. All networks are made of 25 vertices. 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)`.

Acknowledgements

BikeMi data have been kindly provided by Clear Channel s.r.l. Alessia Pini was also supported by MOX - Department of Mathematics, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano, Italy. Alessia Pini also acknowledges partial support from COST Action CA 15109 and UCSC (Research grant track D1). Aymeric Stamm was also supported by MOX - Department of Mathematics, Politecnico di Milano, P.za Leonardo da Vinci 32, 20133 Milano, Italy, by Computational Radiology Laboratory - Boston Children's Hospital - Harvard Medical School, 360 Longwood Ave, 02115 Boston, MA, USA and by Center for Analysis, Decisions, and Society, Human

Table B.1: Summary table of the simulated scenarios

Scenario	Scope	S1	S2
A	Edge strengths	Poisson model:	
		`lambda = 5`	`lambda = 6`
B	Vertex relabelling	Stochastic block model:	
		`pref.matrix = p1`	`pref.matrix = p2`
		`block.sizes = c(12L, 1L, 12L)`	
C	Diffusion patterns	k-regular model:	Erdős-Rényi model:
		`k = 8L`	`p = 1/3`
D	Network VS Indicators	Watts & Strogatz model:	Barabási-Albert model:
		`dim = 1L`	`power = 2L`
		`nei = 4L`	`m = 4L`
		`p = 0.15`	`directed = FALSE`

Technopole, Palazzo Italia, Via Cristina Belgioioso, 20157 Milano, Italy. We sincerely thank two anonymous reviewers for their highly constructive and insightful comments.

Supplementary material

The supplementary material contains plots of the entire bikeMi data set.

References

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

Aydin, B., Pataki, G., H., W., Bullitt, E., Marron, J.S., 2009. A principal component analysis for trees. Ann. Appl. Stat. 3, 1597–1615.

Barabási, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.

Barberán, A., T., B.S., Casamayor, E.O., Fierer, N., 2012. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. 6, 343–351.

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

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

Bollobas, B., 2001. Random Graphs, Second Edition. Cambridge University Press, Cambridge.

Brombin, C., Salmaso, L., 2009. Multi-aspect permutation tests in shape analysis with small sample size. Comput. Stat. Data Anal. 53, 3921–3931.

Cabassi, A., Pigoli, D., Secchi, P., Carter, P.A., 2017. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. Electron. J. Statist. 11, 3815–3840. URL: https://doi.org/10.1214/17-EJS1347, doi:10.1214/17-EJS1347.

Chartrand, G., Saba, F., Zou, H.B., 1985. Edge rotations and distance between graphs. Casopis pro pestovani matematiky 110, 87–91.

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

Comellas, F., Diaz-Lopez, J., 2008. Spectral reconstruction of complex networks. Physica A 387, 6436–6442.

Dirmeier, S., 2017. diffusr. R Foundation for Statistical Computing.

Dryden, I.L., Koloydenko, A., Zhou, D., 2009. Non-euclidean statistics for covariance matrices, with application to diffusion tensor imaging. Ann. Appl. Stat. 3, 1102–1123.

Dryden, I.L., Mardia, K.V., 1998. Statistical Analysis of Shape. Wiley, New York.

Durante, D., Dunson, D.B., 2018. Bayesian inference and testing of group differences in brain networks. Bayesian Anal. 13, 29–58.

Durante, D., Dunson, D.B., Vogelstein, J.T., 2017. Nonparametric bayes modeling of populations of networks. J. Am. Statist. Ass. 112, 1516–1530. URL: https://doi.org/10.1080/

01621459.2016.1219260, doi:10.1080/01621459.2016.1219260, arXiv:https://doi.org/10.1080/01621459.2016.1219260.

Dwass, M., 1957. Modified Randomization Tests for Nonparametric Hypotheses. Ann. Math. Statist. 28, 181–187. doi:10.1214/aoms/1177707045.

Erdös, P., Rényi, A., 1959. On random graphs. i. Publicationes Mathematicae 6, 290–297.

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

Friedman, J.H., Rafsky, L.C., 1979. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. Ann. Statist. 7, 697–717.

Ginestet, C.E., Li, J., Balanchandran, P., Rosenberg, S., Kolaczyk, E.D., 2017. Hypothesis testing for network data in functional neuroimaging. Ann. Appl. Stat. 11, 725–750.

Gretton, A., Borgwardt, K., Rasch, M., Scholkopf, B., Smola, A., 2012. A kernel two-sample test. Journal of Machine Learning Research 16, 723–773.

Guimerá, R., Nunes Amaral, L.A., 2005. Functional cartography of complex metabolic networks. Nature 433, 895–900.

Hall, P., Tajvidi, N., 2002. Permutation test for equality of distribution in high dimensional settings. Biometrika 89, 359–374.

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

Henze, N., 1988. A multivariate two-sample test based on the number of nearest neighbor type coincidences. Ann. Statist. 16, 772–783. URL: https://doi.org/10.1214/aos/1176350835, doi:10.1214/aos/1176350835.

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

Jain, B.J., 2016. Statistical graph space analysis. Pattern Recogn. 60, 802–812. URL: https://doi.org/10.1016/j.patcog.2016.06.023, doi:10.1016/j.patcog.2016.06.023.

Jarret, E.B., 1997. Edge rotation and edge slide distance graphs. Computers Math. Applic. 34, 81–87.

Krause, A.E., Frank, K.A., Mason, D.M., E., U.R., Taylor, W.W., 2003. Compartments revealed in food-web structure. Nature 426, 282–285.

Li, H., Wenston, A., 2010. Strict p-negative type of a metric space. Positivity 14, 529–545.

Liu, Z., Modarres, R., 2011. A triangle test for equality of distribution functions in high dimensions. Journal of Nonparametric Statistics 23, 605–615.

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

Lyons, R., 2013. Distance covariance in metric spaces. Ann. Probab. 41, 3284–3305.

Maa, J.F., Pearl, D.K., Bartoszynsk, R., 1996. Reducing multidimensional two-sample data to one-dimensional interpoint comparisons. Ann. Statist. 24, 1069–1074.

Mardia, K.V., 1972. Statistics of Directional Data. Academic Press London, UK.

Marron, J.S., Alonso, A.M., 2014. Overview of object oriented data analysis. Biom. J. 56, 732–753.

Menafoglio, A., Secchi, P., 2017. Statistical analysis of complex and spatially dependent data: a review of object oriented spatial statistics. European journal of operational research 258, 401–410.

Moody, J., 2001. Race, school integration, and friendship segregation in america. Am J Sociol. 107, 679–716.

Newman, M.E.J., 2003. The structure and function of complex networks. SIAM Rev. 45, 167–256.

Newman, M.E.J., 2006. Modularity and community structure in networks. Proc Natl Acad Sci U S A 103, 8577–8582.

Newman, M.E.J., 2010. Networks: an introduction. Oxford University Press, UK.

Nye, T.W., Tang, X., Weyenberg, G., Yoshida, R., 2017. Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees. Biometrika 104, 901–922.

Pastor-Satorras, R., Vespignani, A., 2001. Epidemic spreading in scale-free networks. Phys Rev Lett. 86, 3200–3203.

Pesarin, F., Salmaso, L., 2010. Permutation Tests for Complex Data. Wiley.

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

Pigoli, D., Aston, J.A.D., Dryden, I.L., Secchi, P., 2014. Distances and inference for covariance operators. Biometrika 101, 409–422.

R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

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

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: Uses and interpretations. NeuroImage 52, 1059–1069.

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

Schilling, M.F., 1986. Multivariate two-sample tests based on nearest neighbors. Journal of the American Statistical Association 81, 799–806.

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

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

Székely, G.J., Rizzo, M.L., 2004. Testing for equal distributions in high dimension. InterStat 5.

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

Székely, G.J., Rizzo, M.L., 2005b. A new test for multivariate normality. J. Mult. Anal. 93, 58–80.

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

Tippett, L.H.C., 1931. The Methods of Statistics. Williams & Norgate, London.

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

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 6, 440–442.

Wei, S., Lee, C., Wichers, L., Marron, J.S., 2016. Direction-projection-permutation for high-dimensional hypothesis tests. Journal of Computational and Graphical Statistics 25, 549–569. URL: https://doi.org/10.1080/10618600.2015.1027773, doi:10.1080/10618600.2015.1027773, arXiv:https://doi.org/10.1080/10618600.2015.1027773.

Zelinka, B., 1975. On a certain distance between isomorphism classes of graphs. Casopis pro pestovani matematiky 100, 371–373.

Zelinka, B., 1992. Edge shift distance between trees. Archivum Mathematicum 28, 5–9.