Heat diffusion distance processes : a statistically founded method to analyze graph data sets

Etienne Lasalle   ETIENNE.LASALLE@UNIVERSITE-PARIS-SACLAY.FR
Université Paris-Saclay, CNRS,
Laboratoire de mathématiques d'Orsay,
91405, Orsay, France.

Editor:

Abstract

We propose two multiscale comparisons of graphs using heat diffusion, allowing to compare graphs without node correspondence or even with different sizes. These multiscale comparisons lead to the definition of Lipschitz-continuous empirical processes indexed by a real parameter. The statistical properties of empirical means of such processes are studied in the general case. Under mild assumptions, we prove a functional central limit theorem, as well as a Gaussian approximation with a rate depending only on the sample size. Once applied to our processes, these results allow to analyze data sets of pairs of graphs. We design consistent confidence bands around empirical means and consistent two-sample tests, using bootstrap methods. Their performances are evaluated by simulations on synthetic data sets.

Keywords: graphs, two-sample tests, heat diffusion, empirical processes, topological data analysis.

1. Introduction

Considering the current growth of available data and the modeling power of networks, methods to analyze graph-structured data have gained interest over the last few decades. Particular attention has been devoted to designing notions of distance between graphs. The design of these notions is highly constrained by the working framework. In particular, different types of information can be used depending on whether the graphs are directed or undirected, weighted or unweighted, have the same size or not. Another key factor to define distances is whether a node correspondence (NC) is known or not. In the case of a known NC, we can consider the graphs to be defined on the same vertex set and comparisons can be made at the edge scale. In this context, people have applied various metrics to compare adjacency matrices, Laplacian matrices, heat kernels (Hammond et al., 2013) and other matrices whose entries represent quantities associated to pairs of nodes (Koutra et al., 2013). On the other hand, when no NC is available, graphs are often compared at a mesoscopic or macroscopic level using structural summaries. People have used global statistics on graphs like degree distributions, network diameters, or clustering coefficients (Pržulj et al., 2004). Another well studied approach is to consider graphlets (Pržulj et al., 2004), *i.e.*, small given subgraphs that are counted in graphs. Then, various methods have been developed to compare the graphlet counts (Pržulj, 2007; Yaveroğlu et al., 2014; Ali et al., 2014; Faisal

et al., 2017). Some work has also been pursued to exploit the structural information carried by spectra of operators (Wilson and Zhu, 2008; Gera et al., 2018).

Another powerful way to encode structural information about graphs is to use diffusion processes, like heat diffusion. When working with weighted graphs, one can interpret weights as the thermal conductivity of edges, meaning that heat diffuses faster along edges with higher weights. Note that unweighted graphs can always be seen as weighted graphs with weights in $\{0, 1\}$. Given initial conditions, the way heat diffuses can be used to characterize and compare graphs (Coifman and Lafon, 2006; Coifman and Hirn, 2014; Hammond et al., 2013; Tsitsulin et al., 2018). This approach is appealing as it allows to analyze graphs at different scales by looking at different diffusion times t. For small values of t, the diffusion only concerns a small neighborhood of the initially heated nodes, while for larger values it involves larger and possibly more complex structures, taking into account topological properties of the graph. Thus, the choice of relevant and informative diffusion times is essential.

For more references on comparisons of graphs, we refer the reader to Soundarajan et al. (2014), Emmert-Streib et al. (2016), Tantardini et al. (2019) and references therein.

1.1 Our contributions

While a lot of the above notions of distances are often supported by experimental results and applications to learning or data mining tasks, they usually lack statistical foundations. In this context, we provide new tools to analyze and compare graphs or even data sets of graphs, that benefit from statistical guarantees. Our methods take advantage of the desirable multiscale property of heat diffusion. Moreover, one of our methods can deal with graphs without known NC or even graphs of different sizes, by using topological descriptors from topological data analysis (TDA). To circumvent the difficulty of choosing a suitable diffusion time, we opt to take into account the whole diffusion process. As a result, we define two real-valued processes, indexed by all the diffusion times in $[0, T]$ for some $T > 0$, representing comparisons of heat distributions.

The first process, called Heat Kernel Distance (HKD) process is defined by comparing heat kernels with the Frobenius norm. In this case, an NC between graphs needs to be known for the entry-wise comparison of the heat kernels to be meaningful.

The second process, called Heat Persistence Distance (HPD), is defined using tools from TDA and can deal with graphs of different sizes. To do so, each graph is equipped with a real-valued function, the Heat Kernel Signature (HKS) (Sun et al., 2009; Hu et al., 2014), defined on the vertex set. Then, graphs are converted into topological descriptors called persistence diagrams. They are multisets of points in $\mathbb{R}^2$, encoding how topological features, like connected components and loops, evolve along with the families of sublevel and superlevel subgraphs. The diagrams are then compared with the so-called Bottleneck distance. Using persistence diagrams allows to switch from node-based representations of graphs to comparable topological summaries, hence requiring no assumption on graph sizes and NC.

To statistically study the HKD and HPD processes, we prove general results on Lipschitz-continuous real-valued empirical processes indexed by a real parameter. Namely, we show that they verify a functional Central Limit Theorem and admit Gaussian approximations

with rates depending only on the sample sizes. These results ensure the asymptotic validity of bootstrap methods to design confidence bands around empirical mean processes, as well as consistent two-sample tests. They are applied to the HKD and HPD processes and could be applied to any other Lipschitz-continuous processes indexed by a real parameter under mild assumptions. We illustrate these results on simulated data sets of pairs of graphs, drawn from various models: Erdős-Rényi model, stochastic block model, and random geometric graph models.

The results from this article allow to statistically analyze data sets of pairs of graphs. While it might seem more natural to study data sets of graphs, there are various situations where graphs naturally come in pairs, and where the relevant information is actually the structural changes inside the pairs. For example, when monitoring brain diseases (Rocca et al., 2016; Farahani et al., 2019), patients' brains can be observed at different time points. In this case, the changes of the brains' connectivity are of prime interest to assess the evolution of the disease (Faivre et al., 2016; Castellazzi et al., 2018). Moreover, even when graphs don't come in pairs, there exist several ways to turn a data set of graphs into a data set of pairs of graphs. For example, one can consider all the possible pairs of graphs from the data set or split the data set into two groups and construct pairs that contain a graph of each group. However, designing relevant approaches to combine graphs into pairs is probably a task-dependent problem, which is beyond the scope of this article.

1.2 Organisation

The rest of the paper is organized as follows. Section 2 introduces the graph framework and heat diffusion on graphs. We define the HKD and HPD processes, and present their statistical properties. These results being actually not restricted to the HKD and HPD processes, we generalize the study of such processes. In Section 3, we introduce a framework for general continuous real-valued empirical processes indexed by a real parameter. We prove their statistical properties and present some consequences on bootstrap methods. Finally, we illustrate the construction of confidence bands and two-sample tests in Section 4 using several generative models of random graphs. All Python codes are freely available at `https://github.com/elasalle/HeatDistanceProcess`.

2. Study of graph data using heat distance processes

In this section, we introduce notions of distance between graphs based on heat diffusion. We define the resulting HKD and HPD processes and study their statistical properties.

2.1 Background and definitions

Before introducing the key notions of this work, we present general definitions and notations that will be used in the rest of the paper. We start by introducing notations relative to graph theory before presenting the theory of extended persistence adapted to graphs.

2.1.1 Graphs

For $0 \leq w_{\min} \leq w_{\max}$, we denote by $\mathcal{G}_n(w_{\min}, w_{\max})$ the set of undirected weighted graphs of size n, without self-loop and whose weights are in $\{0\} \cup [w_{\min}, w_{\max}]$. The special case of

unweighted graphs correspond to $w_{\min} = w_{\max} = 1$. For clarity in the notation, we remove the $w_{\min}$ and $w_{\max}$, whenever there is no ambiguity. We also consider $\mathcal{G}^n$ (with n as an exponent) the set of graphs of size at most n, i.e., $\mathcal{G}^n = \cup_{1 \leq i \leq n} \mathcal{G}_i$. For a graph G in $\mathcal{G}_n$, we denote by $W(G)$ its weight matrix (or adjacency matrix), i.e., the $n \times n$ symmetric matrix whose (i, j)-coefficient is the weight $w_{i,j}$ of edge $\{i, j\}$. $D(G)$ denotes the diagonal matrix whose entry $D(G)_{i,i}$ is the degree of node i defined by $\sum_j w_{i,j}(G)$. The combinatorial Laplacian $L(G)$ is defined by $D(G) - W(G)$. Taking non-negative weights ensures that $L(G)$ is a real symmetric positive-semidefinite matrix. From now on, we forget the dependence in G in the notation, whenever there is no ambiguity. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of L and let $(\phi_1, \ldots, \phi_n)$ be a family of orthonormal eigenvectors. We denote by Λ the diagonal matrix containing the eigenvalues on the diagonal and ϕ the matrix whose columns are the ϕ_i's so that L admits the following decomposition

$$L = \phi \Lambda \phi^T = \sum_{k=1}^{n} \lambda_k \phi_k \phi_k^T. \tag{1}$$

Note that $\lambda_1 = 0$ and that ϕ_1 can always be chosen to be the vector whose entries are equal to $1/\sqrt{n}$. In the following, this choice will always be made.

2.1.2 Persistence on graphs

We present here the basics of ordinary and extended persistence. We refer the reader to Cohen-Steiner et al. (2009), Edelsbrunner and Harer (2010) and Oudot (2015) for a complete description of these theories.

Persistence theory allows to study the topology of topological spaces in a multiscale manner. Usually, given a topological space X and a continuous real-valued function $f : X \to \mathbb{R}$, one considers the family of sublevel sets $X_\alpha := \{x \in X, \ f(x) \leq \alpha\}$, for α varying from $-\infty$ to $+\infty$. Ordinary persistence records the levels at which topological features (connected components, loops, cavities or higher dimensional holes...) appear and disappear. For each feature, its birth and death levels are stored as the coordinates (b, d) of a point in $\mathbb{R}^2$. The multiset of these points is called a persistence diagram. This framework can be applied to graphs.

Let $G = (V, E)$ be a graph with vertex set V and edge set E, and f be a real-valued function on V. Consider the family of sublevel subgraphs $(G_\alpha)_{\alpha \in \mathbb{R}}$, where $G_\alpha = (V_\alpha, E_\alpha)$ with $V_\alpha = \{v \in V, \ f(v) \leq \alpha\}$ and $E_\alpha = \{\{v, v'\} \in E, \ v, v' \in V_\alpha\}$. Across the family of sublevel graphs, as α increases, we can record birth and death levels of connected components, and birth levels of loops. As a connected component dies when it gets connected to an older connected component, remark that the connected components of G will never die. Similarly, as a loop dies when it gets "filled-in" by a 2-dimensional object, loops appearing in G_α (an object of maximal dimension 1) will never die. To prevent topological features from having no death levels (or infinite death levels), the theory of extended persistence suggests also considering the family of superlevel subgraphs. Define $G^\alpha = (V^\alpha, E^\alpha)$ similarly to G_α with $V^\alpha = \{v \in V, \ f(v) \geq \alpha\}$. A death level is now assigned to connected components of G and loops as the level at which they appear in the family of superlevel subgraphs when α decreases. Additionally, we record the birth and death of connected components in the family of superlevel subgraphs. Hence, extended persistence is able to detect four types

of topological features and extract their birth and death levels (b, d) corresponding to four types of points :

- Ord_0 : birth and death of a connected component in (G_α).

- Rel_1 : birth and death of a connected component in (G^α).

- Ext_0 : birth and death of a connected component of G when using (G_α) and (G^α).

- Ext_1 : birth and death of a loop when using both (G_α) and (G^α).

These four types of topological features can be seen as downward branches, upward branches, connected components, and loops, respectively; with the orientation being taken with respect to f. For a graph G and a function f on its vertices, we will denote by $Ord_0(G, f)$, $Rel_1(G, f)$, $Ext_0(G, f)$ and $Ext_1(G, f)$ the persistence diagrams containing the corresponding points. In the following, $Dg(G, f)$ will generically denote any of these four diagrams. We refer the reader to Carrière et al. (2020, Section 2.1) for a more precise and illustrative presentation of extended persistence diagrams on graphs.

The space of diagrams can be equipped with the Bottleneck distance d_B. We recall its definition. Let μ and ν be two diagrams, *i.e.*, two multisets of points in $\mathbb{R}^2$, and let $\Delta := \{(a, a), a \in \mathbb{R}\}$ be the diagonal. Denote by $\Pi(\mu, \nu)$ the set of bijections from $\mu \cup \Delta$ to $\nu \cup \Delta$. d_B is defined by

$$d_B(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \sup_{x \in \mu \cup \Delta} \|x - \pi(x)\|_\infty.$$

We state a stability result for extended persistence diagrams computed on graphs. It is a consequence of a more general stability result for persistence diagrams.

Theorem 1 (Chazal et al. 2016; Cohen-Steiner et al. 2009) *For all graph $G = (V, E)$, for all $f, f' : V \to \mathbb{R}$, and for all diagram construction Dg among Ord_0, Rel_1, Ext_0 and Ext_1,*

$$d_B(Dg(G, f), Dg(G, f')) \le \|f - f'\|_\infty,$$

with $\|f\|_\infty = \max\{|f(v)|, \ v \in V\}$.

2.2 Heat distance processes

Let G be a graph in $\mathcal{G}_n$ and let L be its Laplacian. For $t \ge 0$, let $u_t \in \mathbb{R}^n$ be the vector whose i-th coefficient represents the amount of heat of node i at time t. Then, u_t follows the heat equation :

$$\forall t \ge 0, \quad \frac{d}{dt} u_t = -L u_t$$

The solution is given by $u_t = e^{-tL} u_0$. The matrix e^{-tL} is called the *heat kernel* and describes how heat diffuses in the graph. The i-th column of e^{-tL} contains the amount of heat of each node at time t, when a single unit of heat was placed at node i at time $t = 0$. From (1), the heat kernel decomposes in

$$e^{-tL} = \phi e^{-t\Lambda} \phi^T = \sum_{k=1}^{n} e^{-t\lambda_k} \phi_k \phi_k^T.$$

2.2.1 HEAT KERNEL DISTANCE

Assume that we know the NC for graphs in $\mathcal{G}_n$ and that we number the nodes such that the identity mapping gives the correspondences. Hence, comparing adjacency matrices, Laplacians, or heat kernels entry-wise becomes meaningful. Here we compare graphs through their heat kernels. For two graphs G and G', define their Heat Kernel Distance (HKD) at time t by

$$D_t((G,G')) = \|e^{-tL} - e^{-tL'}\|_F, \tag{2}$$

where L and L' are the laplacians of G and G' respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. This notion of distance was introduced by Hammond et al. (2013). To turn the HKD into a parameter-free notion of distance, Hammond et al. (2013) define the *Graph Diffusion Distance* as $\max_t D_t((G,G'))$. This has the drawback of comparing different pairs of graphs at different times. Instead, our approach consists in using the whole function $t \to D_t((G,G'))$. More precisely, considering a probability distribution P on $\mathcal{G}_n \times \mathcal{G}_n$ and a random pair of graphs $(G,G') \sim P$, we are interested in the stochastic process $\{D_t((G,G')), \ t \in [0,T]\}$ for some $T > 0$. That is, the process obtained by evaluating the functions of the family $\mathcal{F}_{HKD} := \{D_t, \ t \in [0,T]\}$ on a random pair of graphs (G,G'). This framework corresponds to the general framework of *empirical processes*. The properties of the process associated with $\mathcal{F}_{HKD}$ are studied in Section 2.3, while a more general study of such empirical processes is carried out in Section 3.

2.2.2 HEAT PERSISTENCE DISTANCE

In practice, the NC between graphs is not always known. Additionally one may be interested in comparing graphs of different sizes. In these cases, HKD can not be computed. To circumvent these issues and following ideas from Carrière et al. (2020), we define the Heat Persistence Distance (HPD) by using extended persistence diagrams computed with the Heat Kernel Signature (HKS). These persistence diagrams can be compared with the Bottleneck distance d_B without any assumption on graph sizes and node identification.

The HKS was first introduced by Sun et al. (2009) for the study of shapes. Here we restrict ourselves to the definition of the HKS on graphs of Hu et al. (2014). For a graph G of size n with vertex set $V = \{1, \dots, n\}$, the HKS at time t is the function $h_t(G) : V \to \mathbb{R}$ such that

$$h_t(G)(i) = \sum_{k=1}^{n} e^{-t\lambda_k} \phi_k(i)^2, \quad 1 \le i \le s.$$

Intuitively, the image of $h_t(G)$ corresponds to the diagonal of the heat kernel e^{-tL}. Hence, $h_t(G)(i)$ represents the remaining amount of heat at node i after a diffusion time t, when a single unit of heat was placed at node i at time $t = 0$. For each value of t, the HKS provides a function on the vertices of a graph, that we use to compute an extended persistence diagram. The HPD at time t between two graphs G, G' in $\mathcal{G}^n$ is defined by

$$H_t((G,G')) = \max_{Dg} d_B(Dg(G, h_t(G)), Dg(G', h_t(G'))), \tag{3}$$

where the maximum is taken over the four diagram constructions Ord_0, Rel_1, Ext_0 and Ext_1. In our simulations, persistence diagrams are computed by following the approach of Carrière et al. (2020) and using the Gudhi library (Maria et al., 2014).

Similarly to $\mathcal{F}_{HKD}$, we define the family $\mathcal{F}_{HPD} := \{H_t, \ t \in [0, T]\}$ in order to study the induced stochastic process : $\{H_t((G, G')), \ t \in [0, T]\}$, for some random pair of graphs $(G, G') \in \mathcal{G}^n \times \mathcal{G}^n$.

2.3 Results

The goal of this section is to show that the HKD and HPD processes admit a functional version of the central limit theorem, as well as a Gaussian approximation (see Section 3.1 for precise definitions of these notions). As we will show in Section 3, this is a special case of a more general result on uniformly bounded Lipschitz-continuous processes. For the sake of completeness of this section, we choose to state the results on the distance processes here, before exposing and proving the general results in Section 3.

Let us start by proving a lemma on the extremal laplacian eigenvalues of graphs in $\mathcal{G}_n$. We define $\Lambda_{\min}$ and $\Lambda_{\max}$ as, respectively, the minimal and maximal positive laplacian eigenvalues of graphs in $\mathcal{G}_n$:

$$\Lambda_{\min} := \inf\{\lambda > 0, \ \text{s.t.} \ \lambda \text{ is an eigenvalue of } L(G), G \in \mathcal{G}_n\}$$
$$\Lambda_{\max} := \sup\{\lambda > 0, \ \text{s.t.} \ \lambda \text{ is an eigenvalue of } L(G), G \in \mathcal{G}_n\}.$$

Lemma 2 $\Lambda_{\min}$ *and* $\Lambda_{\max}$ *satisfy the following bounds :*

$$\Lambda_{\min} \geq \frac{8 w_{\min}}{n^2}, \tag{4}$$

$$\Lambda_{\max} \leq n w_{\max}. \tag{5}$$

Note that (4) will not be used in the rest of the paper. Still, we choose to present it as we believe it could be of future use.

Proof Take $G \in \mathcal{G}_n$, we want to prove that λ_{min}, the smallest positive eigenvalue of $L(G)$, verifies $\lambda_{min} \geq 8 w_{\min} n^{-2}$. To do so, we apply a Cheeger-type inequality. First assume that G is connected, hence $\lambda_{min} = \lambda_2(G)$. Let V be the set of vertices of G. Following Fiedler (1995, Section 3) we define the *average minimal cut* of G by

$$\gamma(G) = \min_{\emptyset \neq U \subsetneq V} \sum_{i \in U, \ j \in V \setminus U} \frac{w_{i,j}}{|U| \, (n - |U|)}.$$

As G is connected, there exists at least one edge with a weight greater than $w_{\min}$ joining U and $V \setminus U$. Hence,

$$\sum_{i \in U, \ j \in V \setminus U} w_{i,j} \geq w_{\min}.$$

Moreover, for all U, $|U| \, (n - |U|) \leq n^2/4$. This yields to $\gamma(G) \geq 4 w_{\min}/n^2$. From Fiedler (1995, Theorem 2), we have that $\lambda_2(G) \geq 2\gamma(G)$, hence $\lambda_2(G) \geq 8 w_{\min} n^{-2}$. If now G is not connected, one can check that there exists G_{sub} a connected subgraph of G of size n_{sub}, such that $\lambda_{min} = \lambda_2(G_{sub})$. This gives

$$\lambda_{min} = \lambda_2(G_{sub}) \geq \frac{8 w_{\min}}{n_{sub}^2} \geq \frac{8 w_{\min}}{n^2}.$$

7

This finishes the proof of the first bound.

We now prove the second bound. The largest eigenvalue of L(G) is denoted by λ_n. Letting x be an eigenvector associated to λ_n verifying $\|x\|_2 = 1$, we have

$$\lambda_n = x^T L(G)x = \sum_{i<j} w_{i,j}(x_i - x_j)^2 \leq w_{\max} \sum_{i<j} (x_i - x_j)^2 = w_{\max}\, x^T L(G_{comp})x$$

where G_{comp} is the complete undirected graph, hence

$$L(G_{comp}) = \begin{pmatrix} n-1 & -1 & \cdots & -1 \\ -1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \cdots & -1 & n-1 \end{pmatrix}.$$

One can check that $x^T L(G_{comp})x \leq \lambda_n(L(G_{comp})) = n$, so $\lambda_n \leq n w_{\max}$. $\blacksquare$

2.3.1 HEAT KERNEL DISTANCE PROCESS

Let P be a probability distribution on $\mathcal{G}_n \times \mathcal{G}_n$. Let T be a positive real number and recall that $\mathcal{F}_{HKD} := \{D_t,\ t \in [0,T]\}$, with D_t defined in (2). We will study the centered and rescaled empirical process $\{G_N D_t,\ t \in I\} = \{\sqrt{N}(P_N - P)D_t,\ t \in I\}$, where P_N is the empirical measure $N^{-1} \sum_i \delta_{(G_i, G'_i)}$ associated to a N-sample $((G_1, G'_1), \ldots, (G_N, G'_N))$ drawn under P. We denote by $\langle \cdot, \cdot \rangle$ the standard scalar product in $\mathbb{R}^n$.

Let us prove that the HKD process is uniformly bounded and Lipschitz-continuous. Note that these properties will exactly be the conditions required in Section 3 to prove the statistical results on general empirical processes.

Proposition 3 *For all G, G' in $\mathcal{G}_n$, denote by $(\lambda_k)_{1 \leq k \leq n}$ and $(\phi_k)_{1 \leq k \leq n}$ (resp. $(\lambda'_l)_{1 \leq l \leq n}$ and $(\phi'_l)_{1 \leq l \leq n}$) the eigenvalues and orthonormal eigenvectors of $L(G)$ (resp. $L(G')$). Remember that we always choose ϕ_1 and ϕ'_1 equal to the vector with all entries equal to $1/\sqrt{n}$. Then, the application $t \to D_t((G, G'))$ verifies :*

1. *For all $t \in [0,T]$, $D_t((G, G'))$ can be written in terms of the eigen-elements of $L(G)$ and $L(G')$:*

$$D_t((G, G')) = \left(\sum_{k,l=2}^{n} (e^{-t\lambda_k} - e^{-t\lambda'_l})^2 \langle \phi_k, \phi'_l \rangle^2 \right)^{1/2}. \tag{6}$$

2. *For all $t \in [0,T]$, $D_t((G, G')) \leq \sqrt{n}$.*

3. *$t \mapsto D_t((G, G'))$ is $(n^{3/2} w_{\max})$-Lipschitz continuous on $[0,T]$.*

Proof Let G, G' be in $\mathcal{G}_n$. We start by proving (6). This is done through the following computation :

$$\|e^{-tL} - e^{-tL'}\|_F^2$$

$$= \|e^{-tL}\|_F^2 + \|e^{-tL'}\|_F^2 - 2Tr\left(e^{-tL}e^{-tL'}\right)$$

$$= \sum_{k=1}^n e^{-2t\lambda_k}\|\phi_k\phi_k^T\|_F^2 + \sum_{l=1}^n e^{-2t\lambda_l'}\|\phi_l'\phi_l'^T\|_F^2 - 2\sum_{k,l=1}^n e^{-t\lambda_k}e^{-t\lambda_l'}Tr\left(\phi_k\phi_k^T\phi_l'\phi_l'^T\right)$$

One can prove that $Tr\left(\phi_k\phi_k^T\phi_l'\phi_l'^T\right) = \langle\phi_k, \phi_l'\rangle^2$.

Similarly $\|\phi_k\phi_k^T\|_F^2 = \|\phi_k\|_2^2 = \sum_{l=1}^n \langle\phi_k, \phi_l'\rangle^2$ and $\|\phi_l'\phi_l'^T\|_F^2 = \|\phi_l'\|_2^2 = \sum_{k=1}^n \langle\phi_k, \phi_l'\rangle^2$. So

$$\|e^{-tL} - e^{-tL'}\|_F^2 = \sum_{k,l=1}^n (e^{-t\lambda_k} - e^{-t\lambda_l'})^2\langle\phi_k, \phi_l'\rangle^2.$$

The sums can start at $k = 2$ and $l = 2$ thanks to the facts that $\lambda_1 = \lambda_1' = 0$ and $\phi_1 = \phi_1'$ combined with the orthogonality of the eigenvectors families. This finishes the proof of (6).

Bounding all terms $(e^{-t\lambda_k} - e^{-t\lambda_l'})^2$ by 1 in (6), and using the orthonormality of the eigenvectors families yields to $D_t((G, G')) \leq \sqrt{n}$.

We now prove the Lipschitz result. One can check that $t \to D_t((G, G'))$ is $\mathcal{C}^1$ on $[0, T]$. The case $G = G'$ is easily dealt with. Assume now that $G \neq G'$. For all $t \in (0, T]$,

$$\left|\frac{d}{dt}D_t((G, G'))\right| = \left|\frac{-\sum_{k,l=2}^n (e^{-t\lambda_k} - e^{-t\lambda_l'})(\lambda_k e^{-t\lambda_k} - \lambda_l' e^{-t\lambda_l'})\langle\phi_k, \phi_l'\rangle^2}{\left(\sum_{k,l=2}^n (e^{-t\lambda_k} - e^{-t\lambda_l'})^2\langle\phi_k, \phi_l'\rangle^2\right)^{1/2}}\right|$$

$$\leq \sum_{k,l=2}^n \frac{\left|e^{-t\lambda_k} - e^{-t\lambda_l'}\right|\left|\lambda_k e^{-t\lambda_k} - \lambda_l' e^{-t\lambda_l'}\right||\langle\phi_k, \phi_l'\rangle|^2}{\left(\sum_{i,j=2}^n (e^{-t\lambda_i} - e^{-t\lambda_j'})^2\langle\phi_i, \phi_j'\rangle^2\right)^{1/2}}$$

$$\leq \sqrt{\sum_{k,l=2}^n \left|\lambda_k e^{-t\lambda_k} - \lambda_l' e^{-t\lambda_l'}\right|^2 \langle\phi_k, \phi_l'\rangle^2}\sqrt{\frac{\sum_{k,l=2}^n (e^{-t\lambda_k} - e^{-t\lambda_l'})^2\langle\phi_k, \phi_l'\rangle^2}{\sum_{i,j=2}^n (e^{-t\lambda_i} - e^{-t\lambda_j'})^2\langle\phi_i, \phi_j'\rangle^2}}$$

$$= \sqrt{\sum_{k,l=2}^n \left|\lambda_k e^{-t\lambda_k} - \lambda_l' e^{-t\lambda_l'}\right|^2 \langle\phi_k, \phi_l'\rangle^2},$$

where the last inequality comes from the Cauchy-Schwarz inequality.

According to the mean value theorem, for all k and l

$$\left|\lambda_k e^{-t\lambda_k} - \lambda_l' e^{-t\lambda_l'}\right| \leq |\lambda_k - \lambda_l'| \leq \Lambda_{\max}.$$

9

Hence, for all $t \in (0, T]$,

$$\left| \frac{d}{dt} D_t((G, G')) \right| \leq \Lambda_{\max} \sqrt{\sum_{k,l=2}^{n} \langle \phi_k, \phi'_l \rangle^2} \leq n w_{\max} \sqrt{n} = n^{3/2} w_{\max}.$$

This finishes the proof.   ∎

We now state the statistical results concerning the HKD process.

Theorem 4 *For all probability distribution P on $\mathcal{G}_n \times \mathcal{G}_n$, the family $\mathcal{F}_{HKD}$ is P-Donsker. That is, the process $\{\sqrt{N}(P_N - P)D_t, \ t \in [0, T]\}$ converges weakly to $\mathbb{G}$ in $\mathcal{C}([0, T])$, where $\mathbb{G} = \{\mathbb{G}_t, \ t \in [0, T]\}$ is a zero mean Gaussian process with covariance function $\kappa(t, s) = P(D_t D_s) - PD_t.PD_s$.*

This theorem can be seen as a functional central limit theorem for the HKD process. As Section 3.3 will show, it validates the construction of consistent confidence bands and consistent two-sample tests. To strengthen this result, we provide information about the speed of convergence happening in Theorem 4.

Theorem 5 *For all probability distribution P on $\mathcal{G}_n \times \mathcal{G}_n$, the process $\{G_N D_t, \ t \in I\}$ admits a Gaussian approximation with rate $r_N = N^{-1/7} \log N^{9/14}$.*

Note that this rate is independent of the graph size n.

These two theorems are direct applications of Theorem 11 and Theorem 12 from Section 3, combined with Proposition 3.

2.3.2 HEAT PERSISTENCE DISTANCE PROCESS

In this section we work in $\mathcal{G}^n$. We denote by $n(G)$ the size of a graph G. Let P be a probability distribution on $\mathcal{G}^n \times \mathcal{G}^n$. Let T be a positive real number and recall that $\mathcal{F}_{HPD} := \{H_t, \ t \in [0, T]\}$, with H_t defined in (3). Again, we study the empirical process $\{G_N H_t, \ t \in I\} = \{\sqrt{N}(P_N - P)H_t, \ t \in I\}$, where P_N is the empirical measure $N^{-1} \sum_i \delta_{(G_i, G'_i)}$ associated to a N-sample $((G_1, G'_1), \ldots, (G_N, G'_N))$ drawn under P.

Before stating the statistical results on HPD processes, we prove that they are uniformly bounded and Lipschitz-continuous.

Proposition 6 *For all G, G' in $\mathcal{G}^n$, the application $t \to H_t((G, G'))$ verifies :*

1. *For all $t \in [0, T]$, $0 \leq H_t((G, G')) \leq 1$.*

2. *$t \mapsto H_t((G, G'))$ is $(2nw_{\max})$-Lipschitz-continuous on $[0, T]$.*

Proof Recall that for all $G \in \mathcal{G}^n$, and for a vertex i,

$$h_t(G)(i) = \sum_{k=1}^{n(G)} e^{-t\lambda_k} \phi_k(i)^2,$$

where $(\lambda_k)_{1 \leq k \leq n(G)}$ and $(\phi_k)_{1 \leq k \leq n(G)}$ are the eigenvalues and orthonormal eigenvectors of $L(G)$. Hence $0 \leq h_t(G)(i) \leq 1$ for all i, meaning that all points in the diagram $Dg(G, h_t(G))$ are contained in $[0,1]^2$. So from the definition of the Bottleneck distance, $0 \leq H_t((G, G')) \leq 1$, for all G, G' in $\mathcal{G}^n$ and for all $t \in [0, T]$.

Let us now compute the first derivative of $h_t(G)(i)$:

$$\frac{d}{dt} h_t(G)(i) = -\sum_{k=1}^{n(G)} \lambda_k e^{-t\lambda_k} \phi_k(i)^2.$$

Its absolute value is upper-bounded by $\lambda_{n(G)}$, the largest eigenvalue of $L(G)$. From Lemma 2, we have $\lambda_{n(G)} \leq n(G) w_{\max}$. Hence, $t \to h_t(G)$ is $(nw_{\max})$-Lipschitz continuous on $[0, T]$. To conclude, we come back to the definition of the HPD in terms of distance between persistence diagrams. Applying the triangular inequality to the Bottleneck distance gives for all $G, G' \in \mathcal{G}^n$, for all $t, t' \in [0, T]$, and for all diagram construction Dg,

$$\left| d_B(Dg(G, h_t(G)), Dg(G', h_t(G'))) - d_B(Dg(G, h_{t'}(G)), Dg(G', h_{t'}(G'))) \right|$$
$$\leq d_B(Dg(G, h_t(G)), Dg(G, h_{t'}(G))) + d_B(Dg(G', h_t(G')), Dg(G', h_{t'}(G'))).$$

Similarly, the same inequality but with maxima over the diagram constructions holds :

$$\left| H_t((G, G')) - H_{t'}((G, G')) \right|$$
$$\leq \max d_B(Dg(G, h_t(G)), Dg(G, h_{t'}(G))) + \max d_B(Dg(G', h_t(G')), Dg(G', h_{t'}(G'))).$$

Applying Theorem 1 and using the Lispchitz continuity of the HKS yields

$$\left| H_t((G, G')) - H_{t'}((G, G')) \right|$$
$$\leq \|h_t(G) - h_{t'}(G)\|_\infty + \|h_t(G') - h_{t'}(G')\|_\infty$$
$$\leq 2nw_{\max} |t - t'| .$$

$\blacksquare$

We now state the statistical results concerning the HPD process.

Theorem 7 *For all probability distribution P on $\mathcal{G}^n \times \mathcal{G}^n$, the family $\mathcal{F}_{HPD}$ is P-Donsker. Thus, the process $\{\sqrt{N}(P_N - P)H_t,\ t \in [0, T]\}$ converges weakly to $\mathbb{G}$ in $\mathcal{C}([0, T])$, where $\mathbb{G} = \{\mathbb{G}_t,\ t \in [0, T]\}$ is a zero mean Gaussian process with covariance function $\kappa(t, s) = P(H_t H_s) - PH_t . PH_s$.*

Theorem 8 *For all probability distribution P on $\mathcal{G}^n \times \mathcal{G}^n$, the process $\{G_N H_t,\ t \in I\}$ admits a Gaussian approximation with rate $r_N = N^{-1/7} \log N^{9/14}$.*

As previously, these results are a direct consequence of Theorem 11 and Theorem 12, as well as Proposition 6.

11

3. General continuous empirical processes

In this section, we generalize the previous results on empirical processes. We properly introduce the general framework for continuous empirical processes, then show that uniform boundedness and Lipschitz-continuity implies a functional central limit theorem, as well as a Gaussian approximation. Finally, we derive consequences on the construction of confidence bands and two-sample tests.

3.1 Background and definitions

Let I be a compact interval of $\mathbb{R}$ and $\mathcal{C}(I)$ the space of continuous real-valued functions on I endowed with the metric induced by the uniform norm : $\|h\|_\infty = \sup_{t \in I} |h(t)|$. Consider a measurable space $(\mathbb{X}, \mathcal{X})$. For all measure Q on $(\mathbb{X}, \mathcal{X})$ and all measurable function $g : \mathbb{X} \to \mathbb{R}$, we denote the integral of g with respect to Q by $Qg := \int_{\mathbb{X}} g(x) dQ(x)$. Consider a probability measure P on $(\mathbb{X}, \mathcal{X})$ and $\mathcal{F} := \{f_t, \ t \in I\}$, a family of measurable real-valued functions on $\mathbb{X}$ indexed by I. For all $x \in \mathbb{X}$, define $f(x)$ as the function $t \to f_t(x)$, and assume that $f(x) \in \mathcal{C}(I)$. Therefore, given a random variable X with distribution P, one can equivalently see $\{f_t(X), \ t \in I\}$ either as a random process or as $f(X)$ a random variable in $\mathcal{C}(I)$.

Given an i.i.d sample $X_1, \ldots, X_N$ drawn under P, we are interested in the statistical properties of the mean function $N^{-1} \sum_i f(X_i)$ and its centered and scaled version $N^{-1/2} \left(\sum_i f(X_i) - Pf \right)$. Equivalently, one can study the empirical processes $\{P_N f_t, \ t \in I\}$ and $\{G_N f_t, \ t \in I\}$, where $P_N = N^{-1} \sum_i \delta_{X_i}$, and $G_N = \sqrt{N}(P_N - P)$. In the following, we see random processes and random functions as the same objects.

When studying the statistical properties of $\mathcal{F}$, one might be interested in a functional version of the central limit theorem. This corresponds to the concept of Donsker families.

Definition 9 (P-Donsker) *For P a probability measure on $(\mathbb{X}, \mathcal{X})$, the family $\mathcal{F}$ is called P-Donsker if the process $\{G_N f_t, \ t \in I\}$ converges in distribution to the centered Gaussian process $\{\mathbb{G}_t, \ t \in I\}$ with covariance function κ defined as $\kappa_{s,t} := Pf_t f_s - Pf_t Pf_s$ for all $t, s \in I$..*

Here convergence in distribution means weak convergence in the space $\mathcal{C}(I)$. That is, for all continuous bounded function $h : \mathcal{C}(I) \to \mathbb{R}$, $\lim_{N \to \infty} \mathbb{E}[h(G_N f)] = \mathbb{E}[h(\mathbb{G})]$, where the expectation on the left-hand side is taken over the distribution of the sample $X_1, \ldots, X_N$, and the one on the right-hand side is taken over the distribution of the Gaussian process $\mathbb{G}$.

Going further into the statistical analysis of $\mathcal{F}$, one might want to assess the speed at which $\{G_N f_t, \ t \in I\}$ converges in distribution to $\{\mathbb{G}_t, \ t \in I\}$. This can be done by proving Gaussian approximation results.

Definition 10 (Gaussian Approximation) *Let $(r_N)_{N \geq 1}$ be a vanishing sequence of positive real numbers. We say that the process $\{G_N f_t, \ t \in I\}$ admits a Gaussian approximation with rate r_N, if for all $\lambda > 1$, there exists a constant C such that for all $N \geq 1$ one can construct on the same probability space both the sample $X_1, \ldots, X_N$ and a version $\mathbb{G}^{(N)}$ of the Gaussian process $\mathbb{G}$ verifying*

$$\mathbb{P}\left(\left\| G_N f - \mathbb{G}^{(N)} \right\|_\infty > C.r_N \right) \leq N^{-\lambda}.$$

Note that if $\{G_N f_t, \ t \in I\}$ admits a Gaussian approximation with rate r_N, applying the Borel-Cantelli Lemma would yield to $\left\|G_N f - \mathbb{G}^{(N)}\right\|_\infty = O(r_N)$ *almost surely.*

ASSUMPTIONS

A few assumptions on $\mathcal{F}$ will be needed in the following. We present the main ones here.

(L) - There exits $k > 0$ such that for all $x \in \mathbb{X}$ the function $t \to f_t(x)$ is k-Lipschitz continuous on I, meaning that for all $t, s \in I$

$$|f_t(x) - f_s(x)| \leq k \, |t - s| \, .$$

(B) - $\mathcal{F}$ is uniformly bounded. That is, there exists a constant $M > 0$ such that for all $x \in \mathbb{X}$ and for all $t \in I$,

$$|f_t(x)| \leq M.$$

Note that the assumption **(B)** implies the existence of the covariance function κ. It also implies that all moments of $f_t(X)$ are finite.

3.2 Donsker theorem and Gaussian approximation

In this section, we prove that assumptions **(L)** and **(B)** are sufficient to obtain a Donsker theorem and a Gaussian approximation result. This will allow us to derive, in the next section, more practical consequences of these results. Namely, we will prove the validity of bootstrap methods to construct consistent confidence bands and consistent two-sample tests. Note that these results are very general and assume very little about the family $\mathcal{F}$.

Theorem 11 *Assume that $\mathcal{F}$ verifies assumptions **(L)** and **(B)** .*
Then $\mathcal{F}$ is P-Donsker.

Proof We need to prove the weak convergence of $\{G_N f_t, \ t \in I\}$ to the centered gaussian process $\mathbb{G}$. To do so, remark that assumption **(B)** ensures that second moments of $f(X)$ are finite. This allows applying the multidimensional version of the central limit theorem to all finite-dimensional marginals of the random function $G_N f$. Hence, these finite-dimensional marginals converge in distribution to those of $\mathbb{G}$. To conclude to the weak convergence of the whole process, the sequence of random functions $\{G_N f, \ N \geq 1\}$ needs to be tight. According to Billingsley (2013, Theorem 12.3), tightness of $\{G_N f, \ N \geq 1\}$ is implied by the following two conditions :

1. There exists $t_0 \in I$, such that $\{G_N f_{t_0}, \ N \geq 1\}$ is tight.

2. There exist $\gamma \geq 0$, $\alpha > 1$ and a non-decreasing function $\psi : I \to \mathbb{R}$ such that $\forall t, s \in I$, $\forall N \geq 1$,
$$\mathbb{E}\left[|G_N f_t - G_N f_s|^\gamma\right] \leq |\psi(t) - \psi(s)|^\alpha \, .$$

Let us start by proving point 1. Let t_0 be any point in I. Recall the definition of tightness : $\{G_N f_{t_0}, \ N \geq 1\}$ is tight if for all $\eta > 0$, there exists $\alpha > 0$ such that for all $N \geq 1$

$$P\left(|G_N f_{t_0}| \leq \alpha\right) > 1 - \eta. \tag{7}$$

If $Var(f_{t_0}(X)) = 0$, $\{G_N f_{t_0}, \ N \geq 1\}$ is tight, since for all N, $G_N f_{t_0} = 0$ P-a.s.. Otherwise, fix $\eta > 0$. As the left hand side in (7) is non-decreasing with respect to α, we may as well show that there exists α such that for all N, $P\left(-\alpha < G_N f_{t_0} \leq \alpha\right) > 1 - \eta$. Following from **(B)** , $f_{t_0}(X)$ admits a third moment. Combined with the positive variance, we can apply the Berry-Essen Theorem : if F_N and ϕ denote the cumulative distribution functions of, respectively, $G_N f_{t_0}$ and a centered Gaussian variable of variance $Var(f_{t_0}(X))$, then there exists a constant C such that for all $\alpha \in \mathbb{R}$ and all $N \geq 1$

$$|F_N(\alpha) - \phi(\alpha)| \leq \frac{C}{\sqrt{N}}.$$

Now we take N_η such that for all $N > N_\eta$, $C/\sqrt{N} < \eta/4$, and we choose α_η such that $\phi(\alpha_\eta) > 1 - \eta/4$. Then, for all $N > N_\eta$

$$
\begin{aligned}
&P\left(-\alpha_\eta < G_N f_{t_0} \leq \alpha_\eta\right) \\
=&F_N(\alpha_\eta) - F_N(-\alpha_\eta) \\
=&\phi(\alpha_\eta) - \phi(-\alpha_\eta) + (F_N(\alpha_\eta) - \phi(\alpha_\eta)) - (F_N(-\alpha_\eta) - \phi(-\alpha_\eta)) \\
\geq&\phi(\alpha_\eta) - \phi(-\alpha_\eta) - 2\frac{C}{\sqrt{N}} \\
=&2\phi(\alpha_\eta) - 1 - 2\frac{C}{\sqrt{N}} \\
>&2\left(1 - \frac{\eta}{4}\right) - 1 - 2\frac{\eta}{4} = 1 - \eta
\end{aligned}
$$

One can easily choose $\alpha > \alpha_\eta$ to extend the inequality $P\left(-\alpha_\eta < G_N f_{t_0} \leq \alpha_\eta\right) > 1 - \eta$ to all $N \geq 1$. This finishes the proof of point 1.

The proof of point 2, is a consequence of assumption **(L)** . For all $t, s \in I$, one has the following inequality

$$
\begin{aligned}
&\mathbb{E}\left[|G_N f_t - G_N f_s|^2\right] \\
=&\mathbb{E}\left[|(f_t(X) - Pf_t) - (f_s(X) - Pf_s)|^2\right] \\
\leq&(2k|t - s|)^2 \\
=&(2kt - 2ks)^2.
\end{aligned}
$$

This proves point 2. with $\gamma = 2$, $\alpha = 2$ and $\psi(u) = 2ku$, and finishes the proof of Theorem 11. ∎

Theorem 12 *Assume that $\mathcal{F}$ verifies assumptions **(L)** and **(B)** .*
Then $\{G_N f_t, \ t \in I\}$ admits a Gaussian approximation with rate $r_N = N^{-1/7} \log N^{9/14}$.

The proof of Theorem 12 is based on a result from Berthet and Mason (2006). They derive rates for the Gaussian approximation of more general processes. They do so by approaching the process by finite-dimensional marginals and applying a multidimensional Gaussian approximation result. Controlling the covering number of the family (or equivalently its metric entropy) allows them to derive a good trade-off between a good approximation by the marginals and keeping the dimension small enough to obtain good rates in the multidimensional Gaussian approximation. For the sake of completeness, let us recall their result.

Let $\mathcal{M}$ be the set of all measurable real-valued functions on $(\mathbb{X}, \mathcal{X})$. The authors work with a centered and scaled empirical process $\{G_N \tilde{f}, \ \tilde{f} \in \tilde{\mathcal{F}}\}$ indexed by a general family $\tilde{\mathcal{F}} \subset \mathcal{M}$, not necessarily indexed by I. Their result shows that, under mild assumptions, this process can be approached by a centered Gaussian process $\tilde{\mathbb{G}}$ indexed by $\tilde{\mathcal{F}}$ with covariance $\mathbb{E}\left[\tilde{\mathbb{G}}(\tilde{f})\tilde{\mathbb{G}}(\tilde{g})\right] = P\tilde{f}\tilde{g} - P\tilde{f}P\tilde{g}$, for $\tilde{f}, \tilde{g} \in \tilde{\mathcal{F}}$. Let us define the covering number of $\tilde{\mathcal{F}}$. First, assume that there exists $\tilde{F} \in \mathcal{M}$ such that for all $\tilde{f} \in \tilde{\mathcal{F}}$ and all $x \in \mathbb{X}$, $|\tilde{f}(x)| \leq \tilde{F}(x)$. We say that $\tilde{F}$ is an *envelope* of $\tilde{\mathcal{F}}$. For all probability measure Q on $(\mathbb{X}, \mathcal{X})$, we consider the semi-metric $d_Q(g, h)^2 = \int (g - h)^2 dQ$, for $g, h \in \mathcal{M}$. Under d_Q, we define the ball of radius $\delta > 0$ centered in $h \in \mathcal{M}$ by $B_Q(h, \delta) := \{g \in \mathcal{M}, \ d_Q(g, h) < \delta\}$. We also define $\tilde{F}_Q^2 := \int \tilde{F}^2 dQ$. For $\delta > 0$, let $N(\tilde{\mathcal{F}}, d_Q, \delta)$ be the size of the smallest finite subset $K \subset \mathcal{M}$ verifying that the union of balls $B_Q(h, \delta)$ for h in K covers $\tilde{\mathcal{F}}$. Finally, we set the covering number of $\tilde{\mathcal{F}}$ to be $N(\tilde{\mathcal{F}}, \delta) := \sup_Q N(\tilde{\mathcal{F}}, d_Q, \delta.\tilde{F}_Q)$, where the supremum is taken over all Q such that $0 < \tilde{F}_Q < \infty$.

Before stating the result of Berthet and Mason (2006), consider these two basic assumptions.

(F.i) For some $M > 0$ and for all $\tilde{f} \in \tilde{\mathcal{F}}$, $\sup_{x \in \mathbb{X}} |\tilde{f}(x)| \leq M/2$.

(F.ii) The class $\tilde{\mathcal{F}}$ is point-wise measurable, i.e., there exists a countable subclass $\tilde{\mathcal{F}}_\infty$ of $\tilde{\mathcal{F}}$ such that we can find for any function $\tilde{f} \in \tilde{\mathcal{F}}$ a sequence of functions $\{\tilde{f}_m\}$ in $\tilde{\mathcal{F}}_\infty$ for which $\lim_{m \to \infty} \tilde{f}_m(x) = \tilde{f}(x)$ for all $x \in \mathbb{X}$.

Proposition 13 *(Berthet and Mason, 2006, Proposition 1.) Assume that $\tilde{\mathcal{F}}$ verifies (F.i) and (F.ii). Take $\tilde{F} := M/2$ as the envelope of $\tilde{\mathcal{F}}$. Moreover, assume that there exist positive constants c_0 and v_0 such that $N(\tilde{\mathcal{F}}, \delta) \leq c_0 \delta^{-v_0}$. Then, for each $\lambda > 1$ there is a constant $\rho(\lambda)$ such that for each N, one can construct $X_1, \ldots, X_N$ and $\tilde{\mathbb{G}}^{(N)}$ such that*

$$\mathbb{P}\left(\sup_{\tilde{f} \in \tilde{\mathcal{F}}} \left|G_N \tilde{f} - \tilde{\mathbb{G}}^{(N)}(\tilde{f})\right| > \rho(\lambda) N^{-\frac{1}{2+5v_0}} \log N^{\frac{4+5v_0}{4+10v_0}}\right) \leq N^{-\lambda}.$$

Proof (Theorem 12) The proof consists in applying Proposition 13 to $\mathcal{F}$ with $v_0 = 1$. Clearly, assumption **(B)** implies (F.i). Since the paths $t \to f_t(x)$ are continuous for all $x \in \mathbb{X}$, taking $\mathcal{F}_\infty = \{f_t, \ t \in I \cap \mathbb{Q}\}$ gives (F.ii), where $\mathbb{Q}$ is the set of rational numbers.

Let us prove the upper-bound on the covering number. Let L be the length of I. From **(L)** , we know that there exists a constant k, such that $t \to f_t(x)$ is k-lipschitz, for all x. Considering a regular grid on I , $\min I = t_0 < t_1 < \cdots < t_q = \max I$. For all $t \in I$ there

exists a integer j such that for all $x \in \mathbb{X}$, $\left| f_t(x) - f_{t_j}(x) \right| \leq k \left| t - t_j \right| \leq kL/q$. Hence, for all probability measure Q, $d_Q(f_t, f_{t_j}) \leq kL/q$, meaning that $N(\mathcal{F}, d_Q, kT/q) \leq q + 1$. As this last inequality stands for all Q, we can find a constant $c_0 > 0$ such that $N(\mathcal{F}, \delta) \leq c_0 \delta^{-1}$, for all $\delta > 0$. This finishes the proof. ∎

3.3 Statistical consequences

The previous section ensures the asymptotic validity of bootstrap methods for the construction of confidence bands and two-sample tests.

CONFIDENCE BAND

Let $c_\alpha := \inf\{u, \ \mathbb{P}\left(\|\mathbb{G}\|_\infty > u \right) \leq \alpha\}$ be the upper α-quantile of the maximum of the Gaussian process. As a consequence of Theorem 11, we have

$$\lim_{N \to \infty} \mathbb{P}\left(\forall t, P f_t \in \left[P_N f_t - \frac{c_\alpha}{\sqrt{N}}, P_N f_t + \frac{c_\alpha}{\sqrt{N}} \right] \right) \geq 1 - \alpha.$$

Unfortunately, as the distribution of $\mathbb{G}$ is unknown, c_α can not be directly computed. Instead, consider a bootstrap sample $\hat{X}_1, \ldots, \hat{X}_N$ drawn under P_N and let $\hat{P}_N$ be its empirical probability measure. Consider the process $\{\hat{G}_N f_t, t \in I\}$ where $\hat{G}_N$ is the measure $\sqrt{N}(\hat{P}_N - P_N)$. Theorem 2.6 in Kosorok (2008) ensures that in the case where $\mathcal{F}$ is P-Donsker, $\{\hat{G}_N f_t, t \in I\}$ converges weakly to $\mathbb{G}$, given the data. Let $\hat{c}_\alpha$ be the upper α-quantile of $\|\hat{G}_N f\|_\infty$ given the data. We can use $\hat{c}_\alpha$ to design a consistent confidence band around $P_N f$:

$$\lim_{N \to \infty} \mathbb{P}\left(\forall t, P f_t \in \left[P_N f_t - \frac{\hat{c}_\alpha}{\sqrt{N}}, P_N f_t + \frac{\hat{c}_\alpha}{\sqrt{N}} \right] \right) \geq 1 - \alpha.$$

Note that $\hat{c}_\alpha$ can be estimated with Monte-Carlo simulations, by drawing as many bootstrap samples as we want.

TWO SAMPLE TESTS

Consider the following setup. Let P and Q be two probability distributions on $\mathbb{X}$, and assume we are given two independent iid samples $(X_1, \ldots, X_M)$ and $(Y_1, \ldots Y_N)$, drawn under P and Q, respectively. We denote by P_M and Q_N the empirical measures. We would like to test the null hypothesis $H_0 : P = Q$ against the alternatives $H_1 : P \neq Q$, by using the family $\mathcal{F}$, assuming that it is Donsker with respect to both P and Q. We follow the approach described in Section 3.7 of Van Der Vaart and Wellner (1996), and consider the following test statistic :

$$D_{M,N} := \sqrt{\frac{MN}{M + N}} \, \|P_M f - Q_N f\|_\infty.$$

The strategy is to define a data-dependent threshold $\hat{c}_{M,N}(\alpha)$ and reject the null hypothesis whenever $D_{M,N} > \hat{c}_{M,N}(\alpha)$. Consider the pooled data $(Z_1, \ldots Z_{M+N}) = (X_1, \ldots, X_M, Y_1, \ldots, Y_N)$,

and its empirical measure H_{N+M}. Let $(\hat{Z}_1, \ldots \hat{Z}_{M+N})$ be a bootstrap sample drawn from H_{M+N} and consider the bootstrap empirical measures

$$\hat{P}_M = \frac{1}{M} \sum_{i=1}^{M} \delta_{\hat{Z}_i} \qquad \text{and} \qquad \hat{Q}_N = \frac{1}{N} \sum_{i=1}^{N} \delta_{\hat{Z}_{M+i}}.$$

We can define

$$\hat{D}_{M,N} := \sqrt{\frac{MN}{M+N}} \left\| \hat{P}_M f - \hat{Q}_N f \right\|_\infty,$$

as well as

$$\hat{c}_{M,N}(\alpha) = \inf \left\{ t, \ \mathbb{P}\left(\left\| \hat{D}_{M,N} f \right\|_\infty > t \ \Big| \ Z_1, \ldots, Z_{N+M} \right) \leq \alpha \right\},$$

for $\alpha \in (0,1)$. Note that $\hat{c}_{M,N}(\alpha)$ can be estimated with Monte-Carlo simulations. Using $\hat{c}_{M,N}(\alpha)$ as the threshold to accept or reject H_0 leads to a consistent test.

Theorem 14 (Van Der Vaart and Wellner 1996, Section 3.7.2) *Assume that $\mathcal{F}$ is Donsker with respect to both P and Q and that $\|Pf\|_\infty$ and $\|Qf\|_\infty$ are finite. Furthermore assume that $M/(M+N) \to \lambda \in (0,1)$. Then the test that rejects H_0 whenever $D_{M,N} > \hat{c}_{M,N}(\alpha)$ is consistent, in the sense that the asymptotic level is α and under any alternative verifying $\|Pf - Qf\|_\infty > 0$, $\mathbb{P}(D_{M,N} > \hat{c}_{M,N}(\alpha)) \to 1$.*

4. Experiments

We illustrate the construction of confidence bands and two-sample tests on synthetic data sets of pairs of graphs. For that, we consider different random graph models and combine them to create independent pairs of graphs.

4.1 Random graph models

In this section, we present the models generating the random graphs, namely the Erdős-Rényi model (Erdos et al., 1960), the stochastic block model (Holland et al., 1983) and the geometric model (Penrose et al., 2003). For each, we specify the parameters used for the simulations.

Erdős-Rényi model (ER)

This model generates random graphs where each edge appears with probability p, independently from all the others. It requires two parameters: n the graph size and p the edge probability. Because of the independence and their homogeneity, ER graphs are considered to have no structure.

In our simulations, we take $n = 50$ and $p = 0.5$. Weights may be added by assigning a uniform weight between 0 and 2 to each existing edge, independently from all the others.

Stochastic block model (SBM)

This model is a generalization of the ER model that introduces a block structure. The n nodes are clustered in K groups $C_1, \ldots, C_K$, of respective sizes $n_1, \ldots, n_K$. Edges appear

independently from the others, but with a probability depending on the groups : edge $\{i, j\}$ appears with probability $p_{k,l}$ when $i \in C_k$ and $j \in C_l$.

In our simulations, we take $K = 2$, $n_1 = n_2 = 25$, and $p_{1,1} = p_{2,2} = 0.75$, $p_{1,2} = p_{2,1} = 0.25$. So graphs are composed of two dense clusters, with few edges between them. Similarly to the ER model, we may add random weights following the uniform distribution between 0 and 2.

GEOMETRIC MODEL (GM)

Given a compact domain U of $\mathbb{R}^d$ for some d, a graph is generated from the GM by drawing n points uniformly on U and creating an edge between two points if their Euclidean distance is smaller than a given threshold. Here we choose a slight variation of this model by considering a number $p \in [0, 1]$ and creating the edges corresponding to the $\lfloor p\binom{n}{2} \rfloor$ pairs of points with the smallest euclidean distances.

In our simulations, the compact domain U is either A_ε the annulus in $\mathbb{R}^2$ with outer radius 1 and inner radius $\varepsilon > 0$ or A_0 the unit disk. We either fix $n = 50$ or for each graph, n is drawn from a Poisson distribution of parameter 50. We take $p = 0.5$. To obtain weighted graphs, we may assign the weight e^{-2d} to an edge, where d is the distance between the two points forming the edge.

We combine these models to generate pairs of independent graphs on which we can compute HKD and HPD processes. We consider the pairs of independent ER graphs (ER-ER) and the pairs containing one ER graph and one SBM graph (ER-SBM). For these distributions, the groups' composition is known and nodes are treated independently among groups. Thus, we can consider that we know an NC between graphs. As a result, we can compute both HKD and HPD processes. Similarly, we consider pairs of independent geometric random graphs: Disk-Disk and Disk-Annulus. However, as nodes in these models correspond to random points, there is no default NC. Hence, only HPD processes are computed.

4.2 Simulation results

In this section, we present the simulation results concerning the construction of confidence bands and the performances of the two-sample tests.

4.2.1 CONFIDENCE BANDS

In this section, we compute confidence bands under the different models of pairs of graphs defined above. For each model, we draw a sample $(G_1, G'_1), \ldots, (G_N, G'_N)$ with $N = 100$. We compute the mean process, that is $t \to N^{-1} \sum_i D_t((G_i, G'_i))$ or $t \to N^{-1} \sum_i H_t((G_i, G'_i))$ and compute a confidence band of level 99% around this empirical mean using the bootstrap method presented in Section 3.3. Computations are done with 1000 bootstrap samples. Results are shown in Figure 1, 2 and 3, where solid lines represent empirical means and transparent areas represent confidence bands.

Remark that confidence bands around HKD processes (Figure 1) seem to be narrower than those around HPD processes (Figure 2). Therefore, users should rather use HKD processes whenever NC's are available. Nonetheless, the versatility of HPD processes does not totally reduce their efficiency. As Figure 3 indicates, HPD empirical means seem to be able
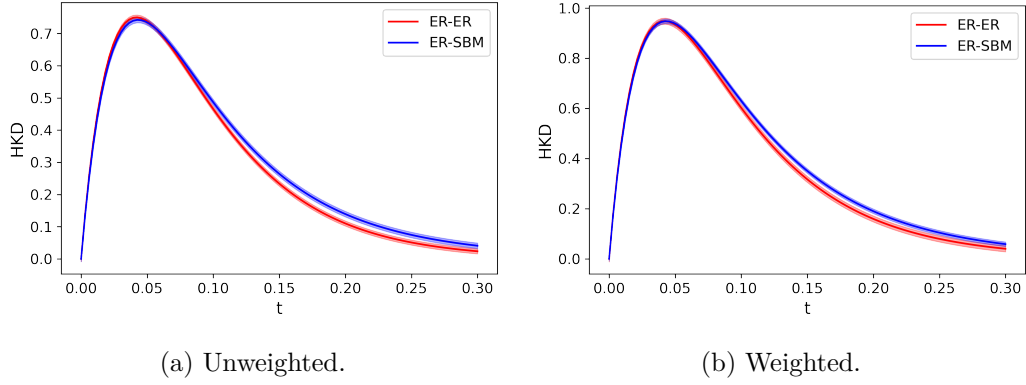
(a) Unweighted.

(b) Weighted.

Figure 1: Confidence band around the mean HKD processes with ER-ER (red) and ER-SBM (blue) distributions.
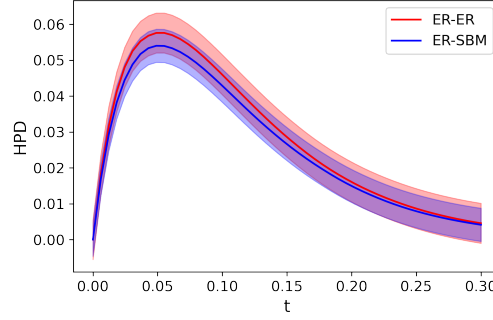

Figure 2: Confidence band around the mean HPD processes with unweighted ER-ER (red) and ER-SBM (blue) distributions.

to discriminate between the different distributions. These observations will be confirmed in the next section, where the performances of the two-sample tests are investigated.
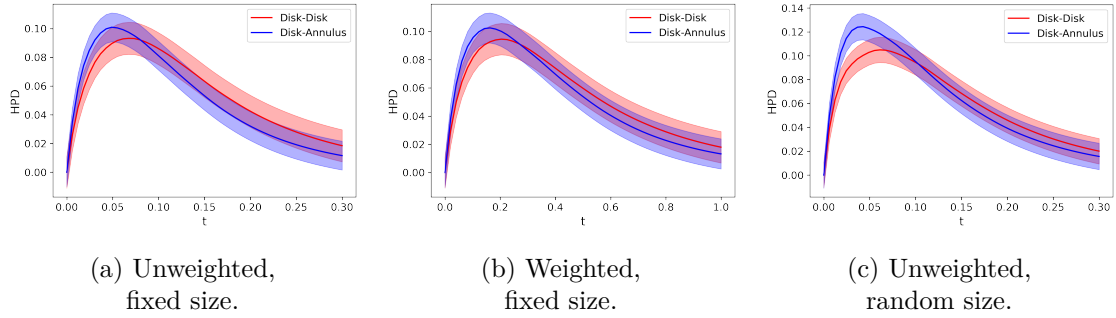

(a) Unweighted,
fixed size.

(b) Weighted,
fixed size.

(c) Unweighted,
random size.

Figure 3: Confidence band around the mean HPD processes with Disk-Disk (red) and Disk-Annulus (blue) distributions.

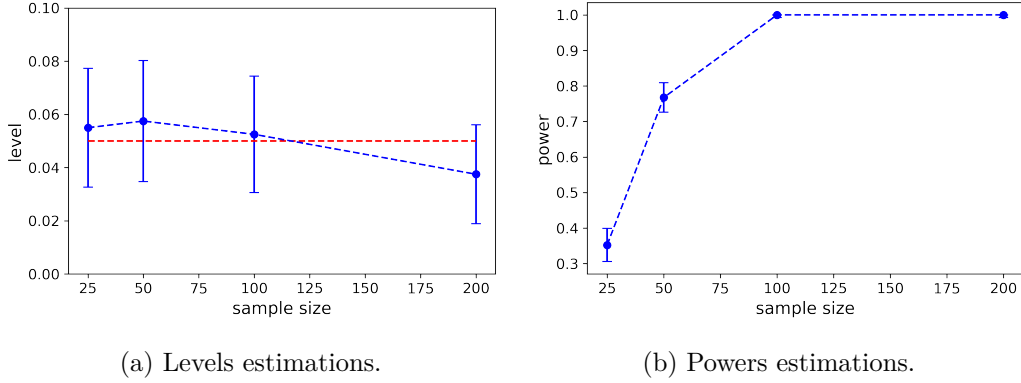(a) Levels estimations. (b) Powers estimations.

Figure 4: Performances of the two-sample test using HKD processes. The level estimations are made with weighted ER-ER distributions, while the power estimations are made with weighted ER-ER and ER-SBM distributions. The samples size varies in [25,50,100,200]. Estimations are done by repeating 400 independent tests. Vertical lines represent 95%-confidence intervals of the estimations.
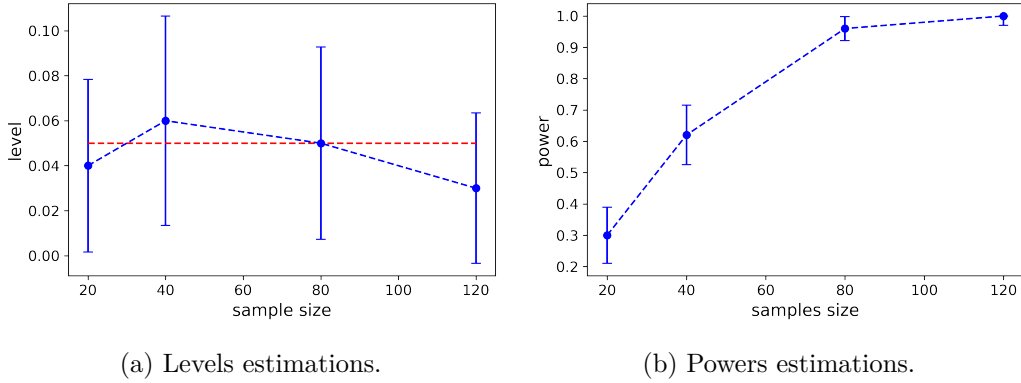

(a) Levels estimations. (b) Powers estimations.

Figure 5: Performances of the two-sample test using HPD processes. The level estimations are made with Disk-Disk distributions, while the power estimations are made with Disk-Disk and Disk-Annulus distributions. Graphs are unweighted and with fixed size. The samples size varies in [20,40,80,120]. Estimations are done by repeating 100 independent tests. Vertical lines represent 95%-confidence intervals of the estimations.

4.2.2 TWO-SAMPLE TESTS

Levels and Powers. Simulations are run to evaluate the performances of the two-sample tests using HKD and HPD processes, see Figure 4 and Figure 5. In all tests, the desired level is set to 0.05, and computations are done with 1000 bootstrap samples. Figures 4a and 5a illustrate that the asymptotic levels of the tests correspond to the set level. On the other hand, Figures 4b and 5b illustrate that the powers tend to 1 when sample sizes increase, indicating that the tests manage to distinguish between the different distributions.
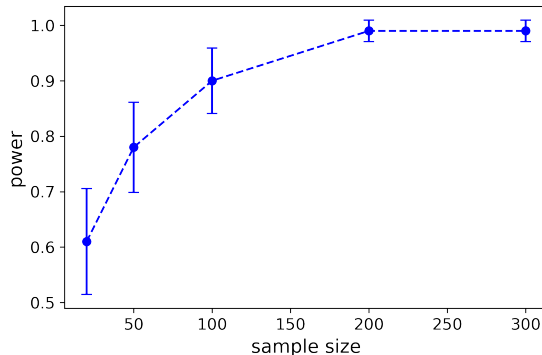
Figure 6: Evolution of the power of the HKD two-sample test when $|p_0(N) - p_1(N)| = C \log N/\sqrt{N}$. The samples size N varies in [20,50,100,200,300]. Estimations are done by repeating 100 independent tests. Vertical lines represent 95%-confidence power intervals of the estimations.

Comparison with the Neyman-Pearson Test. The Neyman-Pearson test (Neyman and Pearson, 1933) is an optimal testing procedure, in the sense that it is the test with the highest power for a given level. But being based on likelihood ratios, it rarely is computable. Its performances are determined by the total variation distance (TV) between the two distributions. If we consider the case of distributions that depend on the sample size, the speed at which their TV tends to 0 determines if the Neyman-Pearson test will asymptotically be able to distinguish between the two distributions. For ER models, independence of the edges allows to theoretically compute bounds of the TV and determine the phase transition. Let us take a real number p such that $0 < p < 1$. And for each sample size $N \geq 1$ consider the parameters $p_0(N)$ and $p_1(N)$, such that both converge to p. Following Lehmann and Romano (2006, Section 13.1.), we can show that as long as $|p_0(N) - p_1(N)| \gg N^{-1/2}$, the Neyman-Pearson test asymptotically distinguishes between the distributions of independent pairs of $ER(n, p_0(N))$ and independent pairs of $ER(n, p_1(N))$ when using N-samples. Figure 6 shows that for large enough sample sizes, our two-sample test based on HKD processes distinguishes between ER models up to $|p_0(N) - p_1(N)| = C \log N/\sqrt{N}$, with $C = 0.01$. The tests are computed with a set level of 0.05 and with 1000 bootstrap samples.

5. Conclusion

We proposed two multiscale comparisons of graphs using heat diffusion processes, namely the HKD and HPD. The first one requires the assumption of equal graph sizes and a known NC, while the second one is free of these assumptions. The multiscale approach solves the problem of choosing an informative diffusion time. We proposed to use these processes to analyze data sets of pairs of graphs and were able to design consistent confidence bands and two-sample tests. The methods are supported by theoretical results: the HKD and HPD families are Donsker, meaning that the processes verify a functional central limit theorem. Moreover, the processes admit Gaussian approximations with rates that are independent of the graph sizes. These results are very general and can be applied to other processes

under mild assumptions. Essentially, the processes are required to be uniformly bounded and Lipschitz-continuous. Moreover, the performances of our methods were evaluated by simulations on synthetic data sets. We showed that the two-sample tests were able to distinguish between Erdős-Rényi and SBM graphs, as well as between geometric graphs sampled on different domains. On Erdős-Rényi models with parameters depending on the sample size, the tests were still distinguishing between the different distributions, even when working close to the phase transition of Neyman-Pearson tests.

As future work, we would like to apply these methods to real-world data sets and extend them to be able to perform learning tasks, e.g., clustering, classification, or change point detection. Extensions to be able to deal with data sets of graphs by opposition to pairs of graphs should also be developed to broaden the application spectrum. On the theoretical side, studying the interplay between graph sizes and sample sizes could be a first step toward non-asymptotic methods to analyze data sets of graphs. Nonetheless, we believe that the introduction of the HKD and HPD processes has the potential to bring innovative and statistically founded ways to analyze data sets of graphs. Moreover, the theoretical results presented in this work being very general, our methods could be extended to other fields.

Acknowledgments

---

1. Inria Saclay.
2. Université Paris-Saclay.

References

Waqar Ali, Tiago Rito, Gesine Reinert, Fengzhu Sun, and Charlotte M Deane. Alignment-free protein interaction network comparison. *Bioinformatics*, 30(17):i430–i437, 2014.

Philippe Berthet and David M Mason. Revisiting two strong approximation results of dudley and philipp. In *High dimensional probability*, pages 155–172. Institute of Mathematical Statistics, 2006.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: a neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pages 2786–2796. PMLR, 2020.

Gloria Castellazzi, Laetitia Debernard, Tracy R Melzer, John C Dalrymple-Alford, Egidio D'Angelo, David H Miller, Claudia AM Gandini Wheeler-Kingshott, and Deborah F Mason. Functional connectivity alterations reveal complex mechanisms based on clinical and radiological status in mild relapsing remitting multiple sclerosis. *Frontiers in neurology*, 9:690, 2018.

Frédéric Chazal, Vin De Silva, Marc Glisse, and Steve Oudot. *The structure and stability of persistence modules*. Springer, 2016.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Extending persistence using poincaré and lefschetz duality. *Foundations of Computational Mathematics*, 9(1):79–103, 2009.

Ronald R Coifman and Matthew J Hirn. Diffusion maps for changing data. *Applied and computational harmonic analysis*, 36(1):79–107, 2014.

Ronald R Coifman and Stéphane Lafon. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Herbert Edelsbrunner and John Harer. *Computational topology: an introduction*. American Mathematical Soc., 2010.

Frank Emmert-Streib, Matthias Dehmer, and Yongtang Shi. Fifty years of graph matching, network alignment and network comparison. *Information sciences*, 346:180–197, 2016.

Paul Erdos, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

Fazle E Faisal, Khalique Newaz, Julie L Chaney, Jun Li, Scott J Emrich, Patricia L Clark, and Tijana Milenković. Grafene: Graphlet-based alignment-free network approach integrates 3d structural and sequence (residue order) data to improve protein structural comparison. *Scientific reports*, 7(1):1–15, 2017.

Anthony Faivre, Emmanuelle Robinet, Maxime Guye, Celia Rousseau, Adil Maarouf, Arnaud Le Troter, Wafaa Zaaraoui, Audrey Rico, Lydie Crespy, Elisabeth Soulier, et al. Depletion of brain functional connectivity enhancement leads to disability progression in multiple sclerosis: a longitudinal resting-state fmri study. *Multiple Sclerosis Journal*, 22 (13):1695–1708, 2016.

Farzad V Farahani, Waldemar Karwowski, and Nichole R Lighthall. Application of graph theory for identifying connectivity patterns in human brain networks: a systematic review. *frontiers in Neuroscience*, 13:585, 2019.

Miroslav Fiedler. An estimate for the nonstochastic eigenvalues of doubly stochastic matrices. *Linear algebra and its applications*, 214:133–143, 1995.

Ralucca Gera, Lázaro Alonso, Brian Crawford, Jeffrey House, JA Mendez-Bermudez, Thomas Knuth, and Ryan Miller. Identifying network structure similarity using spectral graph theory. *Applied network science*, 3(1):1–15, 2018.

David K Hammond, Yaniv Gur, and Chris R Johnson. Graph diffusion distance: A difference measure for weighted graphs based on the graph laplacian exponential kernel. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 419–422. IEEE, 2013.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Nan Hu, Raif M Rustamov, and Leonidas Guibas. Stable and informative spectral signatures for graph matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2305–2312, 2014.

Michael R Kosorok. Introduction to empirical processes. *Introduction to Empirical Processes and Semiparametric Inference*, 2008.

Danai Koutra, Joshua T Vogelstein, and Christos Faloutsos. Deltacon: A principled massive-graph similarity function. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 162–170. SIAM, 2013.

Erich L Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer Science & Business Media, 2006.

Clément Maria, Jean-Daniel Boissonnat, Marc Glisse, and Mariette Yvinec. The gudhi library: Simplicial complexes and persistent homology. In *International congress on mathematical software*, pages 167–174. Springer, 2014.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

Steve Y Oudot. *Persistence theory: from quiver representations to data analysis*, volume 209. American Mathematical Society Providence, 2015.

Mathew Penrose et al. *Random geometric graphs*, volume 5. Oxford university press, 2003.

Nataša Pržulj. Biological network comparison using graphlet degree distribution. *Bioinformatics*, 23(2):e177–e183, 2007.

Natasa Pržulj, Derek G Corneil, and Igor Jurisica. Modeling interactome: scale-free or geometric? *Bioinformatics*, 20(18):3508–3515, 2004.

Maria A Rocca, Paola Valsasina, Alessandro Meani, Andrea Falini, Giancarlo Comi, and Massimo Filippi. Impaired functional integration in multiple sclerosis: a graph theory study. *Brain Structure and Function*, 221(1):115–131, 2016.

Sucheta Soundarajan, Tina Eliassi-Rad, and Brian Gallagher. A guide to selecting a network similarity method. In *Proceedings of the 2014 Siam international conference on data mining*, pages 1037–1045. SIAM, 2014.

Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Computer graphics forum*, volume 28, pages 1383–1392. Wiley Online Library, 2009.

Mattia Tantardini, Francesca Ieva, Lucia Tajoli, and Carlo Piccardi. Comparing methods for comparing networks. *Scientific reports*, 9(1):1–19, 2019.

Anton Tsitsulin, Davide Mottin, Panagiotis Karras, Alexander Bronstein, and Emmanuel Müller. Netlsd: hearing the shape of a graph. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2347–2356, 2018.

Aad W Van Der Vaart and Jon A Wellner. Weak convergence. In *Weak convergence and empirical processes*. Springer, 1996.

Richard C Wilson and Ping Zhu. A study of graph spectra for comparing graphs and trees. *Pattern Recognition*, 41(9):2833–2841, 2008.

Ömer Nebil Yaveroğlu, Noël Malod-Dognin, Darren Davis, Zoran Levnajic, Vuk Janjic, Rasa Karapandza, Aleksandar Stojmirovic, and Nataša Pržulj. Revealing the hidden language of complex networks. *Scientific reports*, 4(1):1–9, 2014.