A Data Application of Graphon Theory

Andreas Alexander Haupt

Born November 2, 1993 in Frankfurt am Main

June 26, 2017

Master's Thesis Mathematics

Advisor: Prof. Ngoc Mai Tran, PhD

Second Advisor: Prof. Joe Neeman, PhD

INSTITUTE FOR APPLIED MATHEMATICS

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT DER

RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT BONN

Contents

Abstract and Acknowledgements

This thesis applies graphon theory [29] to a classification problem for weighted graphs [24]. It connects the theoretical literature on dense graph limits to applied literature on feature-based classification with graph data.

Chapter 2 and Section 4.1 briefly review the literature on graph limits, and introduce binary classification problems. As these sections contain known results, we do not strive for originality. The other sections of this thesis contribute to the study of the statistics of weighted graphs. Our inquiry is led by the following three questions.

Firstly, whether there are weighted generalisations of known sample concentration results for unweighted graph limits. We will formulate our results using a general data-generating model, which we term *decorated graphons*, Definition 3.1. Motivated by shortcomings of the previous study [28] on weighted graph limits, we consider analogues of homomorphism densities and the cut norm for which we show sample concentration. This is the content of Chapter 3 with Theorem 3.7 as central result.

A second question is in how far applied approaches to graph classification can be formulated in the language of random graph models. In Section 4.2, we translate popular graph kernels from the applied literature into the language of random graph models.

The last question is whether one can *prove* bounds relevant for feature-based learning in a random graph model. In Chapter 5, we show two stability estimates bounding the variation of features from above by variations of their data-generating processes. We do this for our adapted notion of homomorphism densities, suppyling a theoretical foundation for the graph kernels presented in Section 4.2 that were originally defined in an ad-hoc manner. We also present and prove a stability estimate that characterises homomorphism densities of cycles, closing a gap in the applied literature. The latter proof combines approximation techniques, KANTOROVICH duality and identities from graph theory. The combination of these tools might be of use for showing stability of a larger class of features.

Acknowledgements I would like to thank my supervisor Ngoc Mai Tran for introducing me to exchangeability and the mathematics of machine learning. I am grateful for her guidance and support throughout the past 14 months. Moreover, I thank my friends and family for their patience and trust.

1 Notation

In thesis we use tools from probability and graph theory. We collect some notation used throughout this thesis here.

Graphs We consider graphs only up to isomorphism: A *node-labelled* weighted finite graph is a tripel (V, E, c), where V the finite node set, $E \subseteq \binom{V}{2}$ is the edge set and $c\colon E \to \mathbb{R}$ is a weight function. A *weighted graph isomorphism* between weighted graphs $G = (V, E, c) \mapsto G' = (V', E', c')$ is a bijective function $\phi\colon V \to V'$ such that $\{i, j\} \in E$ if and only if $\{\phi(i), \phi(j)\} \in E'$ for any $i, j \in V$ and and $c(\{i, j\}) = c'(\{\phi(i), \phi(j)\})$ for any $\{i, j\} \in E$. An equivalence class of node-labelled weighted graphs modulo this isomorphism is called a *weighted graph*. A *complete weighted graph* (V, c) is a weighted graph with $E = \binom{V}{2}$. We view an unweighted graph (V, E) as a complete weighted graph by assigning edges weight one and non-edges weight zero. We denote the set of finite unweighted graphs by $\mathcal{F}$ and the set of unweighted graphs on n nodes by $\mathcal{F}_n$. We view adjacency matrices as equivalence classes up to joint permutations of rows and columns; then graphs and adjacency matrices correspond one-to-one. We use pictograms do denote small graphs: By we denote the complete unweighted graph on 3 nodes (K_3), by the graph on 2 nodes containing one edge (K_2) and by the graph with one node (K_1).

Probability The law of a random variable X on $(\Omega, \mathcal{A}, \mathbb{P})$ will be denoted by $\mathcal{L}(X) = \mathbb{P} \circ X^{-1}$. We denote by Unif_V the uniform distribution on a finite set V and by $\operatorname{Unif}_{[0,1]}$ the uniform distribution on the unit interval. Bern_p is the BERNOULLI distribution with parameter p. For a measurable space $(M, \mathcal{A})$ we denote by $\mathcal{P}(M)$ the set of probability measures on M. Given a measure $\nu \in \mathcal{P}(M)$, its f-fold product is ν^n. If $\nu \in \mathcal{P}(\mathbb{R})$, $\mathbb{E}\nu$ denotes ν's expectation. Denote by δ_x the DIRAC mass at x. The symbol $\perp\!\!\!\perp$ signifies independence of random variables. For $f\colon (M, \mathcal{A}) \to (M', \mathcal{A}')$ measurable and a measure $\nu \in \mathcal{P}(M)$ we denote by $f_*\nu\colon \mathcal{A}' \to \mathbb{R}, f_*\nu[A] = \nu[f^{-1}(A)]$ the pushforward measure of f under ν. Finally, we call $\phi\colon ([0,1], \mathcal{B}([0,1])) \to ([0,1], \mathcal{B}([0,1]))$ measure-preserving if $\phi_* \operatorname{Unif}_{[0,1]} = \operatorname{Unif}_{[0,1]}$.

Miscallena Let $F\colon (M, \|\bullet\|) \to (N, \|\bullet\|')$ be a compact linear operator. Denote by $\Lambda(F)$ the spectrum of F. For sequences (a_n) and (b_n) we use the LANDAU notation

$$(a_n) \in O(b_n) \iff \limsup_{n\to\infty} \frac{a_n}{b_n} < \infty$$

$$
\begin{aligned}
(a_n) \in o(b_n) &\iff \limsup_{n\to\infty} \frac{a_n}{b_n} = 0 \\
(a_n) \in \omega(b_n) &\iff \liminf_{n\to\infty} \frac{a_n}{b_n} > 0.
\end{aligned}
$$

Denote by $\tau_x\colon \mathbb{R} \to \mathbb{R}, \tau_x(y) = x + y$ the translation map. Denote also by $A \triangle B$ the symmetric difference of sets A, B.

2 Graphons

We review the theory of unweighted graph limits based on graphons as introduced in [29]. In Section 2.1 we define random graph models and show which of these can be represented as sampling from exchangeable arrays. Then, we proceed to shows that exchangeable arrays are mixtures of graphons and characterise weak convergence of exchangeable arrays by convergence of homomorphism numbers in Section 2.2. Finally, in Section 2.3, we define a metric structure on the space of graphons.

2.1 Exchangeability and Aldous-Hoover's Representation Result

We start by observing that exchangeable arrays and random graph models are intimately related.

A *random graph model* is a family $(\nu_n)_{n\in\mathbb{N}}$ of probability measures such that each ν_n is a measure on the set $\mathcal{F}_n$ of graphs on n nodes. Let $\pi_n \colon \mathcal{F}_n \to \mathcal{F}_{n-1}$ be the restriction of an $n \times n$ adjacency matrix to its $n-1$ first rows and columns. A random graph model $(\nu_n)_{n\in\mathbb{N}}$ is *projective* if

$$\nu_n \circ \pi_n^{-1} = \nu_{n-1}.$$

Definition. *Let* $\mathbf{X} = (X_{ij})_{i,j\in\mathbb{N}} \colon (\Omega, \mathcal{A}, \mathbb{P}) \to \mathbb{R}^{\mathbb{N}\times\mathbb{N}}$ *be a random variable. We say that* $\mathbf{X}$ *is an* exchangeable array *(or* exchangeable*) if*

$$\mathbf{X} \overset{\mathcal{D}}{=} (X_{\sigma(i)\sigma(j)})_{i,j\in\mathbb{N}},$$

where σ *is a finite permutation of* $\mathbb{N}$*. We call* $\mathbf{X}$ symmetric *if* $(X_{ij})_{i,j\in\mathbb{N}} \overset{a.s.}{=} (X_{ji})_{i,j\in\mathbb{N}}$.

Let $\mathbf{X}$ be a symmetric exchangeable array. Denote by $X_k = (X_{ij})_{1\le i,j\le k}$ the initial k-subarray of $\mathbf{X}$. The measure $\mathbb{G}(k, \mathbf{X}) := \mathcal{L}((X_{ij})_{1\le i,j\le k})$ is the *k-sampling measure*.

Proposition 2.1 ([27, paragraph 11.2.2])**.** *The random graph model* $(\nu_n)_{n\in\mathbb{N}}$ *is projective if and only if there is a binary symmmetric exchangeable array* $\mathbf{X}$ *such that*

$$\nu_n = \mathcal{L}(X_n)$$

for all $n \in \mathbb{N}$.

We need one further definition to be able to present a seminal result characterising the distributions of symmetric exchangeable arrays, in particular, all projective random graph models.

Definition. *A symmetric exchangeable array* $\mathbf{X}$ *is* local *if*

$$(X_{ij})_{i \in I, j \in J} \perp\!\!\!\perp (X_{ij})_{i \in I', j \in J'}$$

for any $I, J, I', J' \subseteq \mathbb{N}$ *such that* $I \cap I', J \cap J' = \emptyset$.

Theorem 2.2 (Aldous-Hoover, [2, 19], [27, Theorem 11.52], [23, Theorem 7.35])**.** *A symmetric exchangeable array* $\mathbf{X}$ *can be represented as follows: There is a random function* $F\colon [0,1]^3 \to \mathbb{R}$*, that is symmetric in its first two arguments such that*

$$(X_{ij})_{i,j \in \mathbb{N}} \overset{\mathcal{D}}{=} (F(U_i, U_j, U_{\{i,j\}}))_{i,j \in \mathbb{N}} \tag{2.1}$$

where $(U_i)_{i \in \mathbb{N}}$ *and* $(U_{\{i,j\}})_{i,j \in \mathbb{N}}$ *are a sequence resp. an array of iid* $\mathrm{Unif}_{[0,1]}$*-variables, which are independent of* F*.*

In addition, the function F *is deterministic if and only if* $\mathbf{X}$ *is local.*

Clearly, for any function $F\colon [0,1]^3 \to [0,1]$ that is symmetric in its first two arguments, the array sampled according to (2.1) is exchangeable and local. Therefore, by means of 2.2, one can parametrise distributions of symmetric local exchangeable arrays by functions $F\colon [0,1]^3 \to [0,1]$ that are symmetric in their first two arguments. In the case of binary arrays $X_{ij} \in \{0,1\}$, we even get an easier parametrisation by symmetric functions $[0,1]^2 \to [0,1]$.

Corollary 2.3. *If* $\mathbf{X}$ *is a local, exchangeable, symmetric, and binary array then there is a symmetric function* $W\colon [0,1]^2 \to [0,1]$ *such that* $U_1, U_2 \ldots \overset{iid}{\sim} \mathrm{Unif}_{[0,1]}$ *and* $X_{ij} \overset{ind}{\sim} \mathrm{Bern}_{W(U_i,U_j)}$*.*

Proof. Let F and $U_i, i \in \mathbb{N}$, $U_{\{j,k\}}, j,k \in \mathbb{N}$ be as in Theorem 2.2. The claim follows by the independence of $(U_i)_{i \in \mathbb{N}}$ and $(U_{\{i,j\}})_{i,j \in \mathbb{N}}$ if we show the following: There is a symmetric function $W\colon [0,1]^2 \to [0,1]$ such that

$$F(U_i, U_j, U_{\{i,j\}}) \overset{\mathcal{D}}{=} 1_{\{(x,y,z) | z \leq W(x,y)\}}(U_i, U_j, U_{\{i,j\}}). \tag{2.2}$$

To prove (2.2) define

$$W(x,y) := \mathbb{E}[F(x,y,U)], U \sim \mathrm{Unif}_{[0,1]}$$

Then

$$\begin{aligned} \mathbb{P}[F(U_i, U_j, U_{ij}) = 1] = \mathbb{E}[F(U_i, U_j, U_{ij})] &= \mathbb{E}[\mathbb{E}[F(U_i, U_j, U_{ij}) | U_i, U_j]] \\ &= \mathbb{E}[W(U_i, U_j)] = \mathbb{P}[U_{\{i,j\}} \leq W(U_i, U_j)]. \end{aligned}$$

□

Let $\mathbf{X}$ be a symmetric, exchangeable, local array and let W be as in Corollary 2.3. Denote $\mathbb{G}(\infty, W) := \mathcal{L}(\mathbf{X})$ and $\mathbb{G}(k, W) := \mathcal{L}(\mathbf{X}_k)$ the *infinite* respectively *k-sampling*. We now restrict to the binary case $X_{ij} \in \{0, 1\}$. The next result characterises classes of functions $W\colon [0,1]^2 \to [0,1]$ that induce the same sampling distributions $\mathbb{G}(\infty, W)$.

Proposition 2.4 ([27, Corollary 10.35 (a)]). *Let $W, W'\colon [0,1]^2 \to [0,1]$ be symmetric functions. Then the following are equivalent:*

(1) $\mathbb{G}(\infty, W) = \mathbb{G}(\infty, W')$*,*

(2) $\mathbb{G}(k, W) = \mathbb{G}(k, W')$ *for any* $k \in \mathbb{N}$*,*

(3) There are measure-preserving maps $\phi, \phi'\colon [0,1] \to [0,1]$ such that $W^\phi \overset{a.s.}{=} W'^{\phi'}$, where $W^\phi\colon [0,1]^2 \to [0,1], (x,y) \mapsto W(\phi(x), \phi(y))$.

This result motivates the following central definition.

Definition 2.5. *A* graphon *W is an equivalence class of functions $W\colon [0,1]^2 \to [0,1]$ modulo measure-preserving transformations, i.e. $W \sim W'$ if and only if there is a measure-preserving function $\phi\colon [0,1] \to [0,1]$ such that $W^\phi(x,y) = W'(x,y)$.*

2.2 Homomorphism Densities

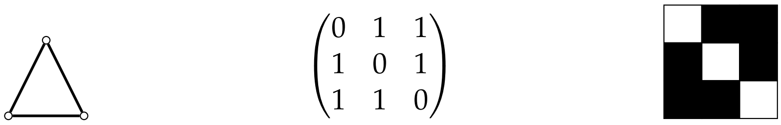
Also subsampling from a finite graph can be interpreted as a grahpon. Let $G = (V, c)$ be a weighted graph. Then interpret the adjacency matrix of G in a chessboard fashion as a graphon W_G: On squares of length and width $\frac{1}{|V|}$ let W_G take the values of each entry in the adjacency matrix. An example of this embedding is given in Figure 2.1. Having this embedding, one can study convergence of the subsampling distributions of graph sequences $(G_n)_{n\in\mathbb{N}}$, $\mathbb{G}(\infty, W_{G_n})$. We would like to characterise weak convergence of $\mathbb{G}(\infty, W_{G_n})$ and for more general graphons in terms of interpretable statistics of graphs.

Definition 2.6. *Let W be a graphon and F be a finite unweighted graph. Define the* induced homomorphism density*, respectively the* homomorphism density *of F in W as*

$$\begin{aligned} t_{ind}(F, W) &:= \mathbb{G}(|V(F)|, W_n)[\{F\}] \\ t(F, W) &:= \mathbb{G}(|V(F)|, W_n)[\{G | E(G) \supseteq E(F)\}]. \end{aligned}$$

$\{G \in V(F_n) | E(G) \supseteq E(F)\}$ is the set of all graphs which contain F as a subgraph.

The following proposition shows that convergence of homomorphism densities characterises weak convergence. Formulas for computing homomorphism densities are given at the end of this section. These will be employed throughout this thesis.



$$\begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$$

(a) Graph (b) Its adjacency matrix (c) Graphon $[0,1]^2 \to [0,1]$

Figure 2.1: A graph, its adjacency matrix and the corresponding graphon.

Proposition 2.7. *Let $(W_n)_{n\in\mathbb{N}}$ be a sequence of graphons and W be a graphon. The following are equivalent:*

(1) $\mathbb{G}(\infty, W_n) \xrightarrow{w} \mathbb{G}(\infty, W)$*,*

(2) $\mathbb{G}(k, W_n) \xrightarrow{w} \mathbb{G}(k, W)$ *for any* $k \in \mathbb{N}$*,*

(3) $t_{ind}(F, W_n) \to t_{ind}(F, W)$ *for any* $F \in \mathcal{F}$*, and*

(4) $t(F, W_n) \to t(F, W)$ *for any* $F \in \mathcal{F}$*.*

Proof. (1) is equivalent to (2) by the definition of weak convergence for infinite index sets. (2) is equivalent to (3) as $(t_{\text{ind}}(F, W))_{F\in\mathcal{F}_k}$ is the probability mass function of the discrete measure $\mathbb{G}(k, W)$ and weak convergence of discrete distributions is equivalent to pointwise convergence of the probability mass function. Finally, also (3) is equivalent to (4): This follows as t_{inj} and t satisfy linear equivalences [27, (7.4), (7.5)]. □

Proposition 2.8. *Let W be a graphon. Then*

$$t(F, W) = \int_{[0,1]^{|V(F)|}} \prod_{\{i,j\}\in E(F)} W(x_i, x_j) \mathrm{d}\,\mathrm{Unif}_{[0,1]}^{|V(F)|}((x_k)_{k\in V(F)}). \tag{2.3}$$

In particular, for a weighted graph $G = (V, c)$, this read

$$t(F, W_G) = \int \prod_{\{i,j\}\in E(F)} c(\{x_i, x_j\}) \mathrm{d}\,\mathrm{Unif}_{V(G)}^{|V(F)|}((x_k)_{k\in V(F)}). \tag{2.4}$$

The following, we will use in an example at the end of this chapter. We observe, that if if G is unweighted and loopless then

$$t(\overset{\circ}{\underset{\circ}{|}}, W_G) = \int c(\{x_i, x_j\}) \mathrm{d}\,\mathrm{Unif}^2_{V(G)}(x_i, x_j) = \sum_{i,j\in V(F)} c(\{i,j\}) = |E(G)|$$

2.3 Cut Distance

In the following, we give the set of graphons metric structure which allows for the introduction of analytic techniques to problems of graph limits.

Definition. *Let W, W' be graphons.*

- *Their* cut distance *is*

$$\delta_\square(W,W') := \inf_\phi \sup_{S,T\subseteq[0,1]} \left| \int_{S\times T} W(x,y) - W'(\phi(x),\phi(y)) \mathrm{d}\,\mathrm{Unif}^2_{[0,1]}(x,y) \right| \tag{2.5}$$

  where the infimum is taken over all measure-preserving transformations $\phi\colon [0,1] \to [0,1]$.

- *The* cut norm *is defined as $\|W\|_\square := \delta_\square(W,0)$, where 0 is the zero graphon.*

We note that the cut norm does *not* induce the cut distance. The name *cut*distance comes from graph *cuts*: If G is a weighted graph, then $\|W_G\|_\square$ is equal to the size of a maximum cut in the graph G. The following three properties along with its explicit definition (2.5) make cut distance a natural notion of similarity of graphs.

Theorem 2.9 ([8, Theorem 3.7])**.** *The topology of weak convergence is metrised by $\delta_\square$.*

Theorem 2.10 ([27, Theorem 11.3])**.** *If the set of graphons is given the topology induced by $\delta_\square$-convergence, then $(\{graphons\}, \delta_\square) \cong \mathrm{cl}_{\delta_\square}(\mathcal{F})$, i.e. the set of graphons is the closure of the set of unweighted graphs with respect to $\delta_\square$.*

Theorem 2.11 ([30, Theorem 5.1])**.** *The metric space $(\{graphons\}, \delta_\square)$ is compact.*

In more general models for random graphs such as [44] or the model we are going to study in the rest of this thesis, an analogue of the cut distance that metrises weak convergence has not been found.

Statements relating the cut distance to sampling by $\mathbb{G}(k,W)$ such as the following hold however in the more general setting we consider.

Theorem 2.12 (Sampling Lemma, [8, Theorem 2.7 (a)])**.** *Let $W_n \sim \mathbb{G}(n,W)$ be sampled from a graphon W. Then there is an exponential tail bound on $\delta_\square(W_n,W)$ around zero. In particular, $W_n \xrightarrow{a.s.} W$ with respect to the metric $\delta_\square$.*

Theorem 2.13 (Counting Lemma, [8, Theorem 2.9])**.** *For any finite unweighted graph F, the function $t(F,\bullet)\colon (\{graphons\},\delta_\square) \to (\mathbb{R}, |\cdot|)$ is* LIPSCHITZ*-continuous.*

We will give quantitative versions of Lemma 2.12 and Lemma 2.13 below. To conclude this section, we illustrate by an example that the structure of the metric space of graphons has relevance even for purely combinatorial questions.

Example (Extremal Graph Theory). *Consider the following problem posed and answered by* TURÀN *[43]: "What is the maximum possible number of edges in an undirected graph G with n vertices that does not contain* $\triangle$ *as a subgraph?" If more complex subgraphs than* $\triangle$ *were forbidden, it would not be clear whether this problem has a solution, i.e. whether the maximum is attained. We can reformulate the problem into a problem of graph limits using homomorphism densities: As we showed below Proposition 2.8, the number of edges is proportional to the density of edges in a graph G,*

$$t(\circ\!\!-\!\!\circ, G) = \frac{|E(G)|}{n^2}.$$

Furthermore, the number of triangles is proportional to the density of triangles

$$t(\triangle, G) = \frac{\#\triangle \textit{ in } G}{n^3}.$$

Hence, TURÀN*'s problem is equivalent to*

$$\max t(\circ\!\!-\!\!\circ, G) \textit{ such that } t(\triangle, G) = 0$$

We know that the the functions $t(\circ\!\!-\!\!\circ, \bullet)$ *and* $t(\triangle, \bullet)$ *are Lipschitz continuous (Theorem 2.13) on the compact space of graphons (Theorem 2.11). Hence, the maximum will be attained. If one can show that the maximum is attained by block models, that is, images of graphs under the embedding given at the end of the last subsection, the problem is solved. This can indeed be done, see [27, Theorem 16.14].*

3 Weighted Graph Limits

We turn our focus to limits of weighted graphs. Their convergence has been studied via homomorphism densities in [28]. After giving the definition of the central object of this study, we review the approach of [28] and argue why it has shortcomings in a statistical setting.

In an effort to address these shortcomings, we study concentration of homomorphism densities. We generalise results known for graphons, in particular Theorem 2.12 and concentration of homomorphism density for the definition put forward.

In the following, we shall make the silent assumption

$$X_{ij} \in [0,1], \forall i,j \in \mathbb{N}. \tag{A}$$

Results for general edge weights on a compact subset of $\mathbb{R}$ can be obtained by an appropriate scaling.

3.1 Definitions

The following is the central object of study in this thesis.

Definition 3.1. *A family of measures*

$$\mathcal{W} = (\mathcal{W}(x,y))_{x,y \in [0,1]} \subseteq \mathcal{P}([0,1])$$

is called a decorated graphon. *Sampling an exchangeable array* $\mathbf{X}$ *from a decorated graphon* $\mathcal{W}$ *is as*

$$U_1, U_2, \ldots \sim \mathrm{Unif}_{[0,1]} \qquad\qquad X_{ij} \overset{iid}{\sim} \mathcal{W}(U_i, U_j).$$

If X_k *is* $\mathbf{X}$*'s initial k-subarray, then we denote* $\mathbb{G}(\infty, \mathcal{W}) := \mathcal{L}(\mathbf{X})$ *and* $\mathbb{G}(k, \mathcal{W}) := \mathcal{L}(X_k)$.

It is worth pointing out that decorated graphons are not defined as equivalence classes, in contrast to Definition 2.5. Therefore, it may happen that $\mathbb{G}(\infty, \mathcal{W}) = \mathbb{G}(\infty, \mathcal{W}')$ even if $\mathcal{W} \neq \mathcal{W}'$. The following propositions however show that all distributions of exchangeable arrays can be viewed as sampled from some decorated graphon.

Proposition. *For any local, exchangeable, symmetric array, there is a family of measures* $(\mathcal{W}(x,y))_{x,y\in[0,1]} \subseteq \mathcal{P}([0,1])$ *such that* $\mathcal{W}(x,y) = \mathcal{W}(y,x)$ *for every* $x, y \in [0,1]$ *such that* $\mathcal{L}(\mathbf{X}) = \mathbb{G}(\infty, \mathcal{W})$.

$$U_1, U_2 \ldots \overset{iid}{\sim} \mathrm{Unif}_{[0,1]} \qquad (X_{ij})_{i,j\in\mathbb{N}} \overset{iid}{\sim} \mathcal{W}(U_i, U_j), i, j \in \mathbb{N}. \tag{3.1}$$

Proof. Let F be the deterministic function from Theorem 2.2. For each $x, y \in [0,1]$, define a random variable

$$W(x,y) = F(x,y,U). \tag{3.2}$$

Define the measure $\mathcal{W}(x,y) := \mathcal{L}(W(x,y))$. The family of measures $(\mathcal{W}(x,y))_{x,y\in[0,1]}$ then satisfies $\mathcal{L}(\mathbf{X}) = \mathbb{G}(\infty, \mathcal{W})$. □

Definition 3.2. *Let* $\mathcal{W} = (\mathcal{W}(x,y))_{x,y\in[0,1]}$ *be a decorated graphon. Its* expectation graphon *is defined as*

$$\mathbb{E}\mathcal{W}\colon \quad [0,1]^2 \to [0,1], (x,y) \mapsto \mathbb{E}[\mathcal{W}(x,y)].$$

As promised, we now discuss the approach of [28]. The authors consider homomorphism densities of unweighted graphs F with continuous, compactly supported functions attached to edges, $c\colon E(F) \to C_c^0([0,1])$ and characterise weak convergence by simultaneous convergence of all homomorphism densities as in Theorem 2.7.

The definition of their homomorphism densities is

$$t(F,W) = \int_{[0,1]^{|V(F)|}} \prod_{\{i,j\}\in E(F)} \left(\int c(\{i,j\}) \mathrm{d}\mathcal{W}(x,y) \right) \mathrm{d}\,\mathrm{Unif}_{[0,1]}^{|V(F)|}, \tag{3.3}$$

that is, they compute for each function $c(\{i,j\})$ associated to an edge $\{i,j\}$ the value of the linear functional $f \mapsto \int f \mathrm{d}\mathcal{W}(x,y)$ associated to $\mathcal{W}(x,y), x, y \in [0,1]$ and integrate with respect to the uniform measure, cf. (2.3). The characterisation uses the one-to-one correspondence of probability measures and positive functionals [34, Theorem 2.14]. Let us point out three shortcomings for statistical analysis in this study.

(1) The authors of [28] do consider sampling from some random object explicitly, which is necessary in a statistical setting. In particular, they do not provide concentration results.

(2) As for each edge separately an infinity of functions has to be considered, their result is inherently infinite even for graphs F with a fixed number of nodes.

(3) The authors give no interpretation of the graphons $\int c(\{i,j\}) \mathrm{d}\mathcal{W}(x,y)$.

Our approach may be rephrased as considering homomorphism densities of type (3.3) in which every edge is assigned the function $f = 1_{[0,1]}$ and we embed graphs as in the example following this discussion. The choice $f = 1_{[0,1]}$ makes the problem finite, addressing (2). In this setting, we are able to generalise concentration results known from graphon theory, formulating our results with reference to sampling from an infinite exchangeable array. This is a remedy for shortcoming (1). Finally, we interpret the graphon $\int c(\{i,j\})\mathrm{d}\mathcal{W}(x,y)$, which is by our choice of $c(\{i,j\})$ independent of $\{i,j\}$ as the expectation graphon, a remedy for shortcoming (3).

Before defining our homomorphism densities of finite-size samples, we give examples of decorated graphons. More can be found in [27, Example 17.1–4].

Example (Embeddings)**. Graphs** *Let $G = (V, c)$ be a weighted graph. Let W_G be the graphon associated to G as defined on page 8. Define $\mathcal{W}_G$ as the decorated graphon that puts* DIRAC *mass on $W_G(x,y)$ at point (x,y); in formulas, $\mathcal{W}_G = (\delta_{W_G(x,y)})_{x,y\in[0,1]}$. The corresponding expectation graphon is W_G.*

Graphons *Let $W\colon [0,1]^2 \to [0,1]$ be a graphon. Then $(\mathrm{Bern}_{W(x,y)})_{x,y\in[0,1]}$ is a decorated graphon with $\mathbb{G}(k,W) = \mathbb{G}(k,(\mathrm{Bern}_{W(x,y)})_{x,y\in[0,1]})$, where the left is sampled from a graphon, the right from a decorated graphon. The corresponding expectation graphon is W.*

Noisy Graphon *Let W be a graphon and let $\nu \in \mathcal{P}(\mathbb{R})$ be a probability measure such that $\operatorname{supp}(\tau_{W(x,y)})_*\nu \subseteq [0,1]$. Denote by $N(W,\nu)$ the decorated graphon.*

$$N(W,\nu)\colon [0,1]^2 \to \mathcal{P}([0,1]), \quad (x,y) \mapsto (\tau_{W(x,y)})_*\nu$$

This decorated graphon adds noise according to ν to a graphon W. Its expectation graphon is W.

It is an interesting open problem whether one can characterise weak convergence as in Theorem 2.9 for an appropriate (quasi-)metric on the space of decorated graphons. Lovàsz claims in [27, p. 324] the existence of such a metric, but "it is awkward to define [it] [. . .] and prove its basic properties".

Formulas (2.3) and (2.5) involve integrals and therefore evaluate a graphon at uncountably many points. Decorated graphons as in Definition 3.1, however, are uninformative about the joint distribution of uncountably many marginals. Therefore, we discretise by sampling an increasing sequence of finite weighted graphs from a decorated graphon and consider limits of homomorphism densities resp. cut distances. As in the following, we will consider subsampling from an array that was sampled from a decorated graphon, we give such arrays a different name.

Definition. *Let $\mathcal{W}$ be a decorated graphon. Let $\mathbf{X} \sim \mathbb{G}(\infty,\mathcal{W})$. Then call X_n a* random n-block model.

We sample the binary array $(Y_{ij})_{1\le i,j\le k}$ from a random n-block model X_n as follows:

$$U_1,\dots,U_k \overset{\text{iid}}{\sim} \mathrm{Unif}_{[n]} \qquad\qquad Y_{ij} \sim \mathrm{Bern}_{X_{ij}}\,. \tag{3.4}$$

The interpretation as a subsampling scheme is as follows: In a first step, sample a random block model from a decorated graphon. In a second step, subsample nodes with replacement from the block model and include edges between the sampled vertices with probability equal to the block model's edge weights.

Definition. *Let X_n be a random n-block model and F be an unweighted graph. Define the* homomorphism density *by*

$$t(F, X_n) := \mathbb{G}(|V(F)|, X_n)[\{H | E(H) \supseteq E(F)\}].$$

Paralleling the case of graphons, one has

$$t(F, X_n) = \int \prod_{\{i,j\} \in E(G)} X_{x_j x_j} \mathrm{d}\, \mathrm{Unif}_{[n]}^{|V(F)|}((x_k)_{k \in V(F)}). \tag{3.5}$$

It will be useful to denote the random variable $t_n(F, \mathcal{W})(\omega) := t_n(F, X_n(\omega))$, for $\mathcal{W}$ a decorated graphon and $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$.

In the next section, we will make use of the following related notion of a density, which closely approximates the homomorphism density.

Definition 3.3. *Let F be an unweighted graph, $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$. Let* $\mathrm{Inj}(V(F), V(G))$ *denote the set of injections $V(F) \hookrightarrow V(G)$. View injections as vectors of disjoint elements $(x_k)_{k \in V(F)}$, $x_k \in V(G)$. Then the* injective homomorphism density *is*

$$t_{inj}(F, G) := \int \prod_{\{i,j\} \in E(F)} X_{x_i x_j} \mathrm{d}\, \mathrm{Unif}_{\mathrm{Inj}(V(F), V(G))}((x_k)_{k \in V(F)}).$$

Proposition ([29, Lemma 2.1.]). *Let F be an unweighted graph and G be a block model. Then*

$$|t_{inj}(F, G) - t(F, G)| \leq \frac{1}{|V(G)|} \binom{V(F)}{2} \tag{3.6}$$

3.2 Concentration around Expectation Graphon

In the case of graphons, homomorphism densities are highly concentrated [27, Corollary 10.4]. In this section, we show that such results also hold in the case of decorated graphons and that there is even concentration of samples in cut distance.

More explicitely, if $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$, $\mathbf{X}' \sim \mathbb{G}(\infty, \mathcal{W}')$, then

$$t_n(F, \mathcal{W}) \xrightarrow[n \to \infty]{\text{a.s.}} t(F, \mathbb{E}\mathcal{W}) \tag{3.7}$$

$$\delta_\square(X_m, X_n') \xrightarrow[m,n \to \infty]{\text{a.s.}} \delta_\square(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}'). \tag{3.8}$$

Equation (3.7) reveals that sampling as in Definition 3.1 will only give information on the expectation graphon in the limit. On the other hand, Proposition 2.7 shows that the ensemble of homomorphism densities characterises the expectation graphon. Hence, using limits of homomorphism densities t_n exactly characterises the expectation graphon. A similar interpretation can be given for the convergence of cut norm (3.8).

We will generalise known result for the theory of graphons as follows: As a first result, in Theorem 3.4, we show that $t_n(F, \mathcal{W})$ converges a.s., using an idea that was outlined for unweighted graph limits in [15, Remark 5.1]. We then identify the limit and prove concentration of $t_n(F, \mathcal{W})$ around the homomorphism densities $t(F, \mathbb{E}\mathcal{W})$ of the expectation graphon in Theorem 3.7. This is the main result of this section and will establish (3.7) in Corollary 3.8. The strategy of proof will be to generalise [8, Lemma 4.4]. Finally, we give a bound on distances $|t_n(F, \mathcal{W}) - t_n(F, \mathcal{W}')|$ in Theorem 3.11.

Theorem 3.4. *Let F be a finite unweighted graph and* $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$*. Consider the descending filtration* $\mathcal{F}_{-n} = \sigma((X_{ij})_{(i,j)\notin[n]\times[n]})$ *and define* $M_n := t_{inj}(F, X_n)$*. Then* $(M_n)_{n\geq|V(F)|}$ *is a reverse martingale.*

The following lemma appears without proof in [27, (5.27)].

Lemma 3.5. *Let F be a finite unweighted graph and* $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$*. Let* $|V(F)| \leq t \leq n$*. Then*

$$t_{inj}(F, (X_{ij})_{1\leq i,j\leq n}) = \frac{1}{\binom{n}{t}} \sum_{S\in\binom{[n]}{t}} \mathbb{E}[t_{inj}(F, (X_{ij})_{i,j\in S})|(X_{ij})_{(i,j)\notin(S\times[n]\triangle[n]\times S)}]. \tag{3.9}$$

Proof. Observe the following: If $(x_A)_{\{A\subseteq[n]||A|=k\}}$ is an array indexed by all k-subsets of a ground set n, then for $S([n])$ denoting the symmetric group acting on $[n]$,

$$\sum_{\substack{i_1,\ldots,i_k\in[n]\\ \text{distinct}}} a_{i_n,\ldots,i_k} = \sum_{\substack{A\subseteq[n]\\ |A|=k}} \sum_{\sigma\in S([n])} x_{\sigma(A)} = \sum_{\substack{A\subseteq[n]\\ |A|=k}} \sum_{\sigma\in\mathrm{Inj}([k],A)} x_{\sigma(A)}, \tag{3.10}$$

where the symmetric group acts on subsets of n by permuting the increasing order of elements.

It suffices to prove (3.9) for $t = |V(F)|$ (the case $t > |V(F)|$ reduces to this case by applying (3.9) with $t = |V(F)|$ to each summand). We may assume without loss of generality that $V(F) = [k]$. Noting that $|\mathrm{Inj}([k], V(G))| = \frac{n!}{(n-k)!}$, we obtain for $\mathbb{E}_n$ the conditional expectation $\mathbb{E}[\bullet|X_n]$ and $\mathbb{E}'_n$ for the conditional expectation $\mathbb{E}\left[\bullet\middle|(X_{ij})_{(i,j)\notin(S\times[n]\triangle[n]\times S)}\right]$

$$\begin{aligned} t_{\mathrm{inj}}(F, X_n) &= \mathbb{E}_n \int \prod_{\{i,j\}\in E(F)} X_{x_i x_j} \mathrm{d}\,\mathrm{Unif}_{\mathrm{Inj}(V(F),V(G))}((x_k)_{k\in V(F)}) \\ &= \mathbb{E}_n \frac{(n-k)!}{n!} \sum_{\substack{i_1,\ldots,i_k\in V(G)\\ \text{distinct}}} \prod_{\{j,k\}\in E(G)} X_{i_j i_k} \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}_n \frac{(n-k)!}{n!} \sum_{\substack{A\subseteq[n]\\|A|=k}} \sum_{\sigma\in\mathrm{Inj}([k],A)} \prod_{\{i,j\}\in E(G)} X_{\sigma(i)\sigma(j)} \\
&= \mathbb{E}'_n \frac{(n-k)!}{n!} \sum_{\substack{A\subseteq[n]\\|A|=k}} \sum_{\sigma\in\mathrm{Inj}([k],A)} \prod_{\{i,j\}\in E(G)} X_{\sigma(i)\sigma(j)} \\
&= \mathbb{E}'_n \frac{(n-k)!k!}{n!} \sum_{\substack{A\subseteq[n]\\|A|=k}} \frac{1}{k!} \sum_{\sigma\in\mathrm{Inj}([k],A)} \prod_{\{i,j\}\in E(G)} X_{\sigma(i)\sigma(j)} \\
&= \mathbb{E}'_n \frac{(n-k)!k!}{n!} \sum_{\substack{A\subseteq[n]\\|A|=k}} \int \prod_{\{i,j\}\in E(F)} X_{x_ix_j} \mathrm{d}\,\mathrm{Unif}_{\mathrm{Inj}(V(F),V(F))}\big((x_k)_{k\in V(F)}\big) \\
&= \mathbb{E}'_n \frac{1}{\binom{n}{k}} \sum_{\substack{A\subseteq[n]\\|A|=k}} t_{\mathrm{inj}}(F,(X_{i,j})_{i,j\in S}).
\end{aligned}
$$

In going from the second to the third line we used (3.10), from the third to fourth we used independence and measurability and to get to the last line, we used the definition of injective homomorphism densities. □

Proof of Theorem 3.4. As $0 \leq M_n \leq 1$ a.s., integrability is trivial.

Fix $n > |V(F)|$. As an instance of exchangeability, we have

$$
\mathbb{E}[t_{\mathrm{inj}}(F,(X_{ij})_{1\leq i,j\leq n-1})|(X_{nj})_{1\leq j\leq n}] = \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{E}[t_{\mathrm{inj}}(F,(X_{ij})_{i,j\in[n]\setminus\{\ell\}})|(X_{\ell j})_{1\leq j\leq n}],
$$

Hence, by Lemma 3.5,

$$
\begin{aligned}
t_{\mathrm{inj}}(F,X_n) &= \frac{1}{n}\sum_{\ell=1}^{n} \mathbb{E}[t_{\mathrm{inj}}(F,(X_{ij})_{i,j\in[n]\setminus\{\ell\}})|(X_{\ell j})_{1\leq j\leq n}] \\
&= \mathbb{E}[t_{\mathrm{inj}}(F,(X_{ij})_{1\leq i,j\leq n-1})|(X_{nj})_{1\leq j\leq n}].
\end{aligned}
$$

Conditioning on $(X_{ij})_{(i,j)\notin[n]\times[n]}$, the claim follows. □

Corollary 3.6. *Let $\mathcal{W}$ be a decorated graphon and $F \in \mathcal{F}$. Then $\lim_{n\to\infty} t_n(F,\mathcal{W})$ exists a.s.*

Proof. Note that for any $n \geq |V(F)|$,

$$
t_{\mathrm{inj}}(F,(X_{ij})_{1\leq i,j\leq n}) = M_n.
$$

Applying the reverse martingale convergence theorem [16, Theorem 5.6.1] to $(M_n)_{n\geq|V(F)|}$, we see that the injective homomorphism densities converge almost surely. Recalling the uniform bound (3.6), $t(F,X_n)$ will converge a.s. to the same limit as $t_{\mathrm{inj}}(F,X_n)$. Hence, it converges a.s. □

The following is the main theorem of this section. It establishes (3.7).

Theorem 3.7. *Let $\mathcal{W} = (\mathcal{W}(x,y))_{x,y\in[0,1]}$ be a decorated graphon and $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$. Let F be an unweighted graph with k nodes. Then*

(1) $\mathbb{E}[t_{inj}(F, X_n)] = t(F, \mathbb{E}\mathcal{W})$, for all $n \geq k$.

(2) With probability at least $1 - 2\exp\left(\frac{n\varepsilon^2}{2k^2}\right)$,

$$|t_n(F, \mathcal{W}) - t(F, \mathbb{E}\mathcal{W})| < \varepsilon. \tag{3.11}$$

In particular, the homomorphism densities are highly concentrated and $t_n(F, \mathcal{W}) \to t(F, \mathbb{E}\mathcal{W})$ a.s.

Proof. (1) Without loss of generality, we may assume that $V(F) = [k]$. Let $\mathbf{X} \sim \mathbb{G}(k, \mathcal{W})$. As a consequence of exchangeability of X_n, it is sufficient in the computation of t_{inj} to consider one injection from $\text{Inj}(V(F), [n])$ instead of the average of all such. Hence, for the identity injection $[k] \hookrightarrow [n]$,

$$\mathbb{E}[t_{\text{inj}}(F, X_n)] = \mathbb{E}\left[\prod_{\{i,j\}\in E(G)} X_{ij}\right].$$

Let $U_1, \ldots, U_n$ be the latent parameters in the sampling of X_n. Then for $\mathtt{W}(U_i, U_j) := \mathbb{E}[X_{ij}|U_i, U_j]$

$$\begin{aligned}\mathbb{E}\left[\prod_{\{i,j\}\in E(G)} X_{ij}\right] &= \mathbb{E}\left[\mathbb{E}\left[\prod_{\{i,j\}\in E(G)} X_{ij}\middle| U_1, \ldots, U_n\right]\right] \\ &= \mathbb{E}\left[\prod_{\{i,j\}\in E(G)} (\mathbb{E}\mathcal{W}(U_i, U_j) + (\mathtt{W}(U_i, U_j) - \mathbb{E}\mathcal{W}(U_i, U_j)))\right]\end{aligned}$$

We multiply out the last product, and use that $(\mathtt{W}(U_i, U_j) - \mathbb{E}\mathcal{W}(U_i, U_j))$ are independent and centered to see that all summands but the one involving only terms from the expectation graphon vanish, i.e.

$$\mathbb{E}\left[\prod_{\{i,j\}\in E(G)} X_{ij}\right] = \mathbb{E}\left[\prod_{\{i,j\}\in E(G)} \mathbb{E}\mathcal{W}(U_i, U_j)\right]$$

(2) Note that the bound in the theorem is trivial for $\varepsilon^2 \leq \ln 2\frac{2k^2}{n} = 4\ln 2\frac{k^2}{2n}$. Hence, in particular, $\varepsilon \leq 4\ln 2\frac{k^2}{2n}$.

Furthermore, $|t(F, X_n) - t(F, \mathbb{E}\mathcal{W})| \leq \frac{1}{n}\binom{k}{2} + |t(F, X_n) - \mathbb{E}[t(F, X_n)]| \leq \frac{k^2}{2n} + |t(F, X_n) - \mathbb{E}[t(F, X_n)]|$ by the first part and (3.6).

Hence

$$\begin{aligned}\mathbb{P}[|t(F,X_n)-t(F,\mathbb{E}\mathcal{W})|\geq\varepsilon] &\leq \mathbb{P}\left[|t(F,X_n)-\mathbb{E}[t(F,X_n)]|\geq\varepsilon+\frac{1}{n}\binom{k}{2}\right]\\ &\leq \mathbb{P}\left[|t(F,X_n)-\mathbb{E}[t(F,X_n)]|\geq\varepsilon\left(1-\frac{1}{4\ln 2}\right)\right].\end{aligned}$$

Set $\varepsilon'=\varepsilon\left(1-\frac{1}{4\ln 2}\right)$.

Let $(X_{ij})_{1\leq i,j\leq n}\sim\mathbb{G}(n,\mathcal{W})$ with latent parameters $U_1,\ldots,U_n$. Define a function depending on n vectors where the i-th vector consists of all values relevant to the i-th column of the array X_n, that is $U_i,X_1,\ldots,X_n$. In formulas,

$$\begin{aligned}f\colon &\bigtimes_{i=1}^{n}[0,1]^{i+1}\to[0,1],\\ &(a_1,\ldots,a_n)=((u_1,x_{11}),(u_2,x_{12},x_{22}),\ldots,(u_n,x_{1n},\ldots,x_{nn}))\\ &\mapsto\mathbb{E}[t(F,(X_{ij})_{1\leq i,j\leq n})|U_1=u_1,\ldots,U_n=u_n,X_{11}=x_{11},\ldots,X_{nn}=x_nn].\end{aligned}$$

We note that the random vectors $(U_i,X_{1i},X_{2i},\ldots,X_{ni})$ are mutually independent for varying i. Claim:

$$|f((a_1,\ldots,a_n)-f((b_1,\ldots,b_n))|\leq\sum_{i=1}^{n}\frac{k}{n}1_{a_i\neq b_i}$$

If this claim is proved, then we have by McDiarmid's inequality [31, (1.2) Lemma],

$$\begin{aligned}&\mathbb{P}[|t(F,X_n)-t(F,\mathbb{E}\mathcal{W})|\geq\varepsilon']\\ &\leq 2\exp\left(-\frac{2\varepsilon'^2}{n\left(\frac{k}{n}\right)^2}\right)\leq 2\exp\left(-\frac{2\varepsilon'^2 n}{k^2}\right)=2\exp\left(-\frac{2n\varepsilon'^2}{k^2}\right),\end{aligned}$$

Which implies the theorem by basic algebra.

Let us now prove the claim: It suffices to consider a,b differing in one coordinate, say n. By (3.5), $t(F,X_n)$ can be written as

$$\int g(x_1,\ldots,x_k)\mathrm{d}\,\mathrm{Unif}^k_{[n]}((x_i)_{i\in[k]})$$

for $g(x_1,\ldots,x_k)=\prod_{\{i,k\}\in E(G)}X_{x_ix_k}$. We observe $0\leq g\leq 1$ (in the case of graphons, one has $g\in\{0,1\}$). It hence suffices to bound the measure where the integrand g depends on a_i by $\frac{k}{n}$. This is the case only if if $x_\ell=i$ at least for one $\ell\in[k]$. But the probability that this happens is upper bounded by,

$$1-\left(1-\frac{1}{n}\right)^k\leq\frac{k}{n},$$

by the Bernoulli inequality.

□

Corollary 3.8. $t_n(F, (X_{ij})_{1\leq i,j\leq n}) \xrightarrow[n\to\infty]{a.s.} t(F, \mathbb{E}\mathcal{W})$

Proof. This follows by the BOREL-CANTELLI lemma [16, Theorem 2.3.1.] from Theorem 3.7. □

We now come to the second limit from (3.8). In the additional material following this section, we show that the rate $O(\log(n)^{-\frac{1}{2}})$ cannot be improved.

Theorem 3.9. *Let $\mathcal{W}$ be a decorated graphon and let $n \in \mathbb{N}$. Then with probability at least $1 - e^{-\frac{n}{2\log n}}$*

$$\delta_\square(\mathbb{G}(n, \mathcal{W}), \mathbb{E}\mathcal{W}) \leq \frac{20}{\sqrt{\log(n)}}$$

The proof strategy from [27, Lemma 10.26] using concentration inequalities remains unchanged. We show where the higher generality of a decorated graphon is important in the proof.

Proof. We only prove a bound on the expectation of the distance. Note that for $n \leq e^{400}$, the bound is trivial as $\delta_\square \leq 1$ by definition. We prove that for $n > e^{400}$,

$$\mathbb{E}[\delta_\square(\mathbb{G}(n, \mathcal{W}), \mathbb{E}\mathcal{W})] \leq \frac{20}{\sqrt{\log n}} - \frac{1}{20} \tag{3.12}$$

From this, the theorem can be proved by the concentration result [27, Theorem 10.3] applied to $n\delta_\square(\bullet, \mathbb{E}\mathcal{W})$ that is also stated in the weighted case.

To prove the claim, first note that for any graphon, in particular $\mathbb{E}\mathcal{W}$, [27, Lemma 10.11] says that

$$\mathbb{E}[\delta_\square(\mathbb{E}\mathcal{W}, H_n)] \leq \frac{18}{\sqrt{\log n}}, \tag{3.13}$$

for $H_n \sim \mathbb{G}(n, (\delta_{\mathbb{E}\mathcal{W}(x,y)})_{x,y\in[0,1]})$. Let also $G_n \sim \mathbb{G}(n, \mathcal{W})$. Let H_n be sampled with latent parameters $(U_i)_{i=1}^n$ and G_n with latent parameters $(U_i')_{i=1}^n$. We show in the additional material to this section, that we can choose any coupling of $(U_i)_{i=1}^n$ and $(U_i')_{i=1}^n$ getting an upper bound on the expected cut distance. We choose the identical coupling $(U_i)_{i=1}^n = (U_i')_{i=1}^n$ and identify nodes $1, \ldots, n$ conditioning on $(U_i)_{i=1}^n$.

Consider the weighted graph $W_n = H_n - G_n$ having the differences of edge weights of H_n and G_n as weights. Denote its adjacency matrix by $(h_{ij})_{1\leq i,j\leq n}$. Recall that the cut norm in graphs equals the normalised maximum cut value in the graph, cf. page 9. In particular, for $H_n[S,T] = \frac{1}{n^2}\sum_{(i,j)\in S\times T} h_{ij}$,

$$\max_{S,T\subseteq[n]} H_n[S,T] = \|G_n - W_n\|_\square.$$

We can write each $H_n[S,T]$ as the sum of $|S||T|$ many random variables

$$\frac{1}{n^2}X_{ij} - \frac{1}{n^2}Y_{ij}$$

$i \in S, j \in T$, where $X_{ij} \sim \mathcal{W}(U_i, U_j)$ and $Y_{ij} = \mathbb{E}\mathcal{W}(U_i, U_j)$. Hence $X_{ij} - Y_{ij}$ are independent centered random variables on $[-1, 1]$. In the case of graphons, these random variables only take two values given U_i and U_j, $-W(U_i, U_j)$ and $1 - W(U_i, U_j)$.

The claim then follows upon applying HOEFFDING's inequality and a union bound over all choices of S, T (one can assume with a multiplicative error of 4 that these are disjoint), see the derivation of [27, (10.9)]. Putting together (3.12) and (3.13), we get

$$\mathbb{E}[\delta_\square(\mathcal{W}, G_n)] \leq \mathbb{E}[\delta_\square(W, H_n)] + \mathbb{E}[\delta_\square(H_n, G_n)] \leq \frac{18}{\sqrt{\log(n)}} + \frac{11}{\sqrt{n}} < \frac{20}{\sqrt{n}} - \frac{1}{20}$$

□

Corollary 3.10. *Let $\mathcal{W}$, $\mathcal{W}'$ be decorated graphon. Let $\mathbf{X} \sim \mathbb{G}(\infty, \mathcal{W})$, $\mathbf{X}' \sim \mathbb{G}(\infty, \mathcal{W}')$. Then $\delta_\square(X_n, X'_m) \xrightarrow[n,m\to\infty]{a.s.} \delta_\square(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}')$.*

Proof. This follows by using a union bound and applying the BOREL-CANTELLI lemma [16, Theorem 2.3.1.] twice. □

Finally, we can give a sampling version of the Counting Lemma Counting Lemma, [27, Lemma 10.23]. In the additional material following this section, we show that its constant cannot be improved.

Theorem 3.11. *Let $\mathcal{W}$, $\mathcal{W}'$ be decorated graphons and let $F \in \mathcal{F}$ be a finite graph with k nodes. Then with probability $1 - 4\exp\left(-\frac{2n\varepsilon^2}{k^2}\right)$, we have*

$$|t_n(F, \mathcal{W}) - t_n(F, \mathcal{W}')| \leq |E(F)|\delta_\square(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}') + \varepsilon$$

Proof. By the triangle inequality

$$\begin{aligned}&|t_n(F, \mathcal{W}) - t_n(F, \mathcal{W}')|\\ &\quad\leq |t_n(F, \mathcal{W}) - t(F, \mathbb{E}\mathcal{W})| + |t(F, \mathbb{E}\mathcal{W}) - t(F, \mathbb{E}\mathcal{W}')| + |t(F, \mathbb{E}\mathcal{W}') - t_n(F, \mathcal{W}')|\end{aligned}$$

By the Counting Lemma [27, Lemma 10.23], we can bound the second term

$$|t(F, \mathbb{E}\mathcal{W}) - t(F, \mathbb{E}\mathcal{W}')| \leq |E(F)|\delta_\square(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}')$$

the first and third term can each be bounded by ε' with probability at least $1 - 2\exp\left(\frac{n\varepsilon'^2}{2k^2}\right)$. Hence, their sum is bounded by $2\varepsilon'$. Now choosing $\varepsilon = 2\varepsilon'$, one arrives at the claim. □

We showed that there is concentration of homomorphism densities and samples not only in the case of unweighted graph limits but also in the more general case of decorated graphons. The *expectation graphon* is the analogue of the graphon in the unweighted theory of graph limits in this more general theory.

3.3 Additional Material

The Counting Lemma is Tight The counting lemma is tight even for graphons as the following example shows: Let W, W' be graphons such that $t(\multimapdotbothvert, W) = 1$ (i.e., W is the all-one graphon) and $t(\multimapdotbothvert, W') = 1 - \varepsilon$. Then $d_\square(W, W') = \varepsilon$. Let $F = E_m$ be the graph of m mutually disjoint edges. Fix m. Then $t(E_m, W) = 1$ as W is the all-one graphon. As the homomorphism density is multiplicative with respect to connected components [27, (5.28)],

$$t(E_m, W) = t(\multimapdotbothvert, W)^m = (1-\varepsilon)^m.$$

Hence,

$$\frac{|t(E_m, W) - t(E_m, W')|}{|E(F)| d_\square(W, W')} = \frac{1-(1-\varepsilon)^m}{m\varepsilon} \xrightarrow{\varepsilon \to 0} 1$$

where one used for the limit DE L'HÔPITAL. A similar calculation shows that also for densities of cycles C_{2m} the limit is bounded from below by $\frac{1}{2}$.

The Sampling Lemma is rate-optimal The rate we prove in Theorem 3.9 is optimal even in the case of graphons, as [25] show: The optimal bound on $\mathbb{E}[d_\square(G, W)]$, $G \sim \mathbb{G}(k, W)$ for a χ-block model W and k nodes is

$$C\sqrt{\frac{\chi}{k \log(\chi)}}. \tag{3.14}$$

The case of an arbitrary block structure corresponds to $\chi = k$. Then (3.14) agrees with the bound in Theorem 3.9.

The Cut Distance as a Coupling Distance First consider the equivalent definition [22, (6.1) and Theorem 6.9] of the cut norm

$$\delta_\square(W_1, W_2) := \inf_{\phi, \psi} \|W_1^\phi - W_2^\psi\|_\square \tag{3.15}$$

where the infimum runs over all couplings $(\psi_1, \psi_2)\colon \Omega \to [0,1] \times [0,1]$, $\psi_1, \psi_2 \sim \mathrm{Unif}_{[0,1]}$. In particular, if $G \sim \mathbb{G}(k, \mathcal{W})$ of $H \sim \mathbb{G}(k, \mathcal{W}')$, then there is a coupling $(\psi_1, \psi_2)\colon \Omega \to [0,1] \times [0,1]$ of the latent variables such that

$$(U_1, U_1'), \ldots, (U_n, U_n') \overset{\mathcal{D}}{=} (\psi_1, \psi_2)$$

$$\begin{aligned} G_{ij} &\overset{\text{ind}}{\sim} \mathcal{W}(U_i, U_j) \\ H_{ij} &\overset{\text{ind}}{\sim} \mathcal{W}(U'_i, U'_j) \end{aligned}$$

Taking the identical coupling (ψ_1, ψ_1), one has by (3.13)

$$\mathbb{E}[\delta_\square(G, H)] \leq \mathbb{E}[\|G - H\|_\square].$$

Under this coupling, sampling is as

$$\begin{aligned} U_1, \ldots, U_n &\sim \text{Unif}_{[0,1]} \\ G_{ij} &\sim \mathcal{W}(U_i, U_j) \\ H_{ij} &\sim \mathcal{W}(U_i, U_j) \end{aligned}$$

Hence, when proving upper bounds on $\delta_\square(G, H)$, one may assume that G and H were sampled with the same latent parameters $(U_i)_{i=1}^k$.

4 Elements of Supervised Learning

We introduce classification problems and present applied approaches to supervised learning with graph data. Supervised learning problems consist of tuple-valued random variables (datum, label), where the datum is in a general space and the label usually is a real number or is categorical. If the label is real, one speaks of *regression problems*; in case of discrete labels one speaks of *classification problems*. In this thesis, we exclusively study classification problems. To be more precise, with the label only taking binary values.

In Section 4.1, we formally define (binary) classification problems and classifier consistency followed by two examples of classifiers. We conclude the section by presenting the kernel trick.

In Section 4.2, we translate features underlying popular graph kernels to the language of random graphs and show that popular benchmark graph kernels use homomorphism densities as features. In Section 4.3, we define WASSERSTEINstability results.

4.1 Binary Classification Problems

Before presenting two famous examples of classifiers, we collect basic definitions.

Classifiers We call a random variable $(X,Y)\colon (\Omega,\mathcal{A},\mathbb{P}) \to (M,d) \times \{0,1\}$ a *binary classification problem*. Denote $\mathcal{L}(X|Y=1)$, $\mathcal{L}(X|Y=1)$ the *alternatives* of the classification problem. A *classifier for data of size n* is a measurable function

$$f\colon (M \times \{0,1\})^n \times M \to \{0,1\}.$$

The *classification error* of a classifier f *trained* on data $(X_1,Y_1),\dots,(X_n,Y_n)$ for (X,Y) is

$$\mathbb{P}[f(((X_1,Y_1),\dots,(X_n,Y_n)),X) \neq Y|((X_1,Y_1),\dots,(X_n,Y_n))]$$

The optimisation problems in *statistical learning theory* is to minimise classification error in the class of classifiers. This will be an optimisation problem conditional on data, hence of finding a function

$$f((x_1,y_1),\dots,(x_n,y_n),\bullet)\colon M \to \{0,1\} \tag{4.1}$$

that minimises classification error.

Bayes Classifier Denote $\eta(x) := \mathbb{E}[Y|X = x]$. Then the *Bayes classifier* is

$$\bar{\eta}(((x_1, y_1), \dots, (x_n, y_n)), x) = 1_{\eta(x) > \frac{1}{2}}.$$

The BAYES classifier is ignorant of data, i.e. it does not depend on the data $((X_1, Y_1), \dots, (X_n, Y_n))$. Nevertheless, it has the lowest expected classification error achievable by a classifier of length n.

Proposition 4.1 ([14, Thm. 2.1]). *Fix any $n \in \mathbb{N}$. Let f be any classifier of length n. Then*

$$\mathbb{P}[\bar{\eta}(X) \neq Y] \leq \mathbb{P}[f(((X_1, Y_1), \dots, (X_n, Y_n)), X) \neq Y]$$

Because of Proposition 4.1, one can define consistency of classification in terms of the BAYES error.

Definition. *Let $(f_n)_{n \in \mathbb{N}}$, $f_n \colon (M \times \{0, 1\})^n \times M \to \{0, 1\}$ be a sequence of classifiers of increasing length. We say that (f_n) is (conditionally)* consistent *if*

$$\mathbb{P}[f(((X_1, Y_1), \dots, (X_n, Y_n)), X) \neq Y|X = x] \xrightarrow[\mathcal{L}(X)\text{-}a.s.]{n \to \infty} \mathbb{P}[\bar{\eta}(X) \neq Y|X = x].$$

Lastly, a *decision boundary* of a classifier f given data $((X_1, Y_1), \dots, (X_n, Y_n))$ is the closed and hence measurable set

$$\partial\{x \in M | f(((X_1(\omega), Y_1(\omega)), \dots, (X_n(\omega), Y_n(\omega))), x) = 1\}$$

Two important classifiers are the support vector machine (SVM) for which M needs HILBERT structure (but we will see how one can apply it on a wider range of spaces) and the nearest neighbor classifier that only assumes metric structure. The shape of the decision boundaries is one main difference of SVM and nearest neighbor classifiers: For SVM, they are linear, for nearest neighbors, they can be very irregular.

Nearest-Neighbor Classifiers The *k-nearest neighbor classifier* is

$$\gamma_k(((x_1, y_1), \dots, (x_n, y_n)), x) = 1_{\sum_{i=1}^{k} y^i - \frac{k}{2}}$$

where (x^i, y^i) is sorted in a way such that $d(x, x^i)$ is increasing in i. It computes hence a majority vote on the k nearest neighbors of a point x among the data points.

It is known that $(\gamma_{k_n})_{n \in \mathbb{N}}$ is consistent for $(k_n)_{n \in \mathbb{N}} \in \omega(\log n) \cap o(n)$ in the following cases:

(1) If $M = \mathbb{R}^n$ [41] and any classification problem.

(2) If the LEBESGUE differentiation theorem holds in the metric measurable space $(M, \mathcal{B}(M), \mathcal{L}(X))$ [13, Thm. 1].

Support-Vector Machines Fix realisations $(x_1, y_1), \dots, (x_n, y_n)$ of (X, Y). For ease of exposition we assume that $A = \{x_i | i \in [n], y_i = 1\}$ and $B = \{x_i | i \in [n], y_i = 0\}$ are linearly separable, i.e. there is a hyperplane $\langle a, x \rangle = b$ separating A and B. For the inseparable case, see [7, Section 7.1.1].

Let hence $\langle a, x \rangle = b$ be a hyperplane separating A and B. By scaling b accordingly, it is without loss to assume $\|a\| = 1$. By basic linear algebra,

$$|\langle a, x \rangle - b| = d(\{y \in F | \langle a, y \rangle = b\}, x),$$

where d is the metric induced by the scalar product of M. Hence, $|\langle a, x \rangle - b|$ gives the distance of point x to the hyperplane $\{\langle a, x \rangle = b\}$. The *support-vector classifier* is

$$f((x_1, y_1), \dots, (x_n, y_n), x) = 1_{\langle a, x \rangle - b > 0}(x)$$

for a hyperplane maximising distance to the closest datum $x \in A \cup B$. In particular, the hyperplane $\{x | \langle a, x \rangle = b\}$ is the decision boundary of this classifier. The minimal distance to a data point is called the *margin* of the support-vector classifier f. The x_i for which $|\langle a, x \rangle - b| = \varepsilon$ are called *support vectors* as they support the hyperplane defined by a and b. This explains the name *support vector machine*.

The Kernel Trick The SVM hyperplane is hence a solution to the following problem

$$\begin{aligned} \underset{a,b,\varepsilon}{\text{maximise}}\ \varepsilon, \quad \text{such that } & \langle a, x_i \rangle - b \geq \varepsilon, \quad \forall i \in [n], y_i = 1 \\ & \langle a, x_i \rangle - b \leq -\varepsilon, \forall i \in [n], y_i = 0 \\ & \langle a, a \rangle = 1. \end{aligned}$$

After substitution $w = \varepsilon a$ and minimisation of the reciprocal this can be written as a linear-quadratic problem

$$\begin{aligned} \underset{w,b}{\text{minimise}}\ \frac{1}{2}\langle w, w \rangle, \quad \text{such that } & \langle w, x_i \rangle - b \geq 1, \quad \forall i \in [n], y_i = 1 \\ & \langle w, x_i \rangle - b \leq -1, \forall i \in [n], y_i = 0. \end{aligned} \tag{4.2}$$

Proposition 4.2. *There is an optimal solution of* (4.2) $w = \sum_{i=1}^n \alpha_i x_i$.

Proof. Note that for the objective of (4.2), $\frac{1}{2}\langle w', w' \rangle \leq \frac{1}{2}\langle w, w \rangle$ if w' is the orthogonal projection of w onto $\operatorname{span}((x_i)_{i=1}^n)$. Furthermore, if w satisfies the constraints of the optimisation problem, then so does its projection onto $\operatorname{span}((x_i)_{i=1}^n)$. Hence there is always an optimal solution in $\operatorname{span}((x_i)_{i=1}^n)$. □

Inserting $w = \sum_{i=1}^n \alpha_i x_i$, the optimisation problem reads

$$\underset{\alpha,b}{\text{minimise}} \sum_{i,j=1}^n \alpha_i \alpha_j \langle x_i, x_j \rangle, \quad \text{such that } \sum_{j=1}^n \alpha_j \langle x_j, x_i \rangle - b \geq 1, \quad \forall i \in [n], y_i = 1 \tag{4.3}$$

$$\sum_{j=1}^{n} \alpha_j \langle x_j, x_i \rangle - b \leq -1, \forall i \in [n], y_i = 0. \tag{4.4}$$

If we substitute into the support-vector classifier, this reads

$$f(x) = 1_{\sum_{i=1}^{n} \alpha_i \langle x_i, x \rangle - b} \tag{4.5}$$

We note that hence the classifier only depends on the GRAM matrix $(K_{ij})_{1 \leq i,j \leq n}$, $K_{ij} = \langle x_i, x_j \rangle$ of the data $x_1, \ldots, x_n$.

Remark 4.3. *In some applications, instead of minimising $\langle w, w \rangle = \|w\|_2$, one maximises $\|w\|_1 = \sum_{i=1}^{n} |w_i|$. This leads to much sparser vectors w, i.e. vectors with much less non-zero entries [46]. This classifier is called 1-norm SVM or* SVM with ℓ^1-penalty*. We will use it in our data application below.*

Having observed that the classifier will only depend on scalar products of data points, we can generalise the support-vector classifier to settings where we merely can embed a space into a HILBERT space and compute scalar products efficiently.

Definition 4.4 (MERCER kernel)**.** *Let $(X, \mathcal{A}, \mu)$ be a probability space. A symmetric function $k \in L^2_{\mu \otimes \mu}(X \times X)$ is called* MERCER kernel *if*

$$\int_{X \times X} \phi(x) k(x,y) \psi(y) \mathrm{d}\mu^2(x,y) \geq 0$$

for all $\psi, \phi \in L^2(\mu)$.

Theorem 4.5 (Mercer's Theorem, [26, Theorem 3.a.1])**.** *Let $(X, \mathcal{A}, \mu)$ be a first countable topological space that is endowed with a complete, locally finite measure μ with $\operatorname{supp} \mu = X$. Let $k \colon X^2 \to \mathbb{R}$ be a MERCER kernel. Then there is a function $\phi \colon X \to H$ for a HILBERT space H such that $k(x,y) = \langle \phi(x), \phi(y) \rangle_H$.*

The vectors $\phi(x_i) \in H$ are called *feature vectors*. If H is Euclidean, then coordinates of $\phi(x_i)$ are called x_i's *features*.

In the next section, we will see how the kernel trick has been used in the classification of graphs.

4.2 Graph Kernels

In this section we translate applied approaches for graph classification into the language of random graphs. In a weighted graph classification problem alternatives take the form

$$k \sim \kappa \in \mathcal{P}(\mathbb{N}) \qquad\qquad X \sim \mathbb{G}(k, \mathcal{W}).$$

In light of the theorems from Chapter 3, we restrict homomorphism densities $t(F,W)$ for a graphon W and W', as homomorphism densities $t_n(F,\mathcal{W})$ are concentrated around the homomorphism densities of the expectation graphon, $t(F,\mathbb{E}\mathcal{W})$. We will analyse discretisation errors in Chapter 5.

To shed light onto which extent popular benchmark classifiers use their graph data and make more transparent which graphons popular graph kernels cannot distinguish, let us define a general class of MERCER kernels on unweighted graphs. We note that this consists of first defining a locally finite measure on the space of graphs and then to define a kernel function. Consider the space $(\mathcal{F},2^{\mathcal{F}},\mu)$ of finite graphs with the discrete topology and a locally-finite measure

$$\mu = \int_{\mathbb{N}} \sum_{F\in\mathcal{F}_n} \delta_F \mathrm{d}\nu,$$

where $\nu \in \mathcal{P}(\mathbb{N})$ has full support on the natural numbers and $\mathcal{F}_n$ is the set of all graphs on exactly n nodes. The measure μ, for varying choices of ν is commonly employed in graph-based classification. It remains to define a general MERCER kernel.

The Complete Graph Kernel Define the *complete graph kernel* using all induced homomorphism densities respectively respectively all homomorphism densities as features,

$$\phi\colon G \mapsto (t_{\text{ind}}(F,G))_{F\in\mathcal{F}} \qquad \phi'\colon G \mapsto (t(F,G))_{F\in\mathcal{F}} \in L^2_\mu(\mathcal{F}), \tag{4.6}$$

cf. [18]. These can distinguish all graphons that can be separated by sampling arrays, as

$$\begin{aligned}\mathbb{G}(\infty,W) = \mathbb{G}(\infty,W') &\overset{2.4}{\Leftrightarrow} t(F,G) = t(F,G')\forall F \in \mathcal{F}\\ &\iff 0\|t(F,G) - t(F,G')\|_{L^2_\mu(\mathcal{F})} = \|\phi(W) - \phi(W')\|_{L^2_\mu(\mathcal{F})},\end{aligned}$$

where 2.4 stands for Proposition 2.4 and the second equality follows from the assumption that ν has full support. For ϕ, one can even show more.

Proposition 4.6. *Weak convergence is metrised by* $d = \|\phi(\bullet) - \phi(\bullet)\|_{L^2_\mu}\colon \mathcal{F}^2 \to \mathbb{R}$.

This shows that the complete graph kernel can distinguish arbitrary decorated graphons up to their expectation graphons.

Proof. Recall the definition of the induced homomorphism densities

$$t_{\text{ind}}(F,G) = \mathbb{G}(k,G)[F]$$

and that simultaneous convergence of $t_{\text{ind}}(F,G)$ for all $F \in \mathcal{F}$ characterises weak convergence, Theorem 2.7. Hence, weak convergence can be metrised by

$$\|\phi(G) - \phi(G')\|_{L^1_\mu(\mathcal{F})} = \|(t_{\text{ind}}(F,G))_{F\in\mathcal{F}} - (t_{\text{ind}}(F,G'))_{F\in\mathcal{F}}\|_{L^1_\mu(\mathcal{F})}$$

$$= \int_{\mathbb{N}} \|\mathbb{G}(k,W) - \mathbb{G}(k,W')\|_1 \mathrm{d}\nu$$
$$= \sum_{i=1}^{\infty} \nu[\{i\}] \sum_{F\in\mathcal{F}_i} |t_{\text{ind}}(F,W) - t_{\text{ind}}(F,W')|$$

where $\|\bullet\|_1$ is the 1-norm for signed measures using that ν has full support. Observe that by definition of ϕ, It remains to shows that $\|\phi(\bullet) - \phi(\bullet)\|_{L^1_\mu(\mathcal{F})}$ and $\|\phi(\bullet) - \phi(\bullet)\|_{L^2_\mu(\mathcal{F})}$ induce the same topology. Fix $\varepsilon > 0$. Let $m \in \mathbb{N}$ such that $\nu(\mathbb{N}_{\geq m}) \leq \frac{\varepsilon}{4}$. Assume that $G_n \to G$ with respect to $\|\phi(\bullet) - \phi(\bullet)\|_{L^1_\mu}$ (The case for L^2_μ is proved analogously). Denoting by $\phi_{\leq n}(G)$ the feature vector that is zero for all graphs with more than n nodes. Because of $\sum_{F\in\mathcal{F}_n} t_{\text{ind}}(F,G) = 1$, we have

$$\|\phi(G_n) - \phi(G)\|_{L^2_\mu(\mathcal{F})} \leq \|\phi_{\leq n}(G_n) - \phi_{\leq n}(G)\|_{L^2_\mu(\mathcal{F})} + 2\frac{\varepsilon}{4}$$

In the first summand only finitely many entries are nonzero. Hence, we can interpret it as a finite-dimensional L^p-space. As all finite-dimensional L^p-spaces are equivalent, there is a k such that $\|\phi_{\leq n}(G_n) - \phi_{\leq n}(G)\|_{L^2_\mu(\mathcal{F})} \leq \frac{\varepsilon}{2}$, for any $n \geq k$, concluding the proof. □

Hence, the complete graph kernel induces even the same topology as weak convergence. Unfortunately, it is likely that it cannot be computed:

Proposition 4.7 ([18, Prop. 2]). *The complete graph kernel is NP-hard to compute .*

It is an open problem whether another graph kernel than the complete graph kernel that metrises weak convergence. If this were the case, it would in particular imply a polynomial-time algorithm for the graph isomorphism problem, which is neither shown to be NP-complete nor to be in P.

State-of-the-Art Graph Kernels We show that three commonly used graph kernels can be written as using homomorphism densities as features.

(1) The authors in [37] propose so-called *graphlet counts* as features. These can be interpreted as using the restriction of the complete graph kernel ϕ' to graphs with at most $k \in \mathbb{N}$ nodes, in [37] for $k = 5$. In this case, the feature vectors are

$$(t_{\text{ind}}(F,G))_{F\in\mathcal{F}_{\leq k}}.$$

(2) The random walk kernel [18, p. 135 center] restricts ϕ' to n-paths P_n. The feature vectors corresponding to the kernel are

$$\phi \colon G \mapsto (t(P_n,G))_{n\in\mathbb{N}}.$$

(3) [33, Prop. 5 and discussion thereafter] restricts ϕ' to trees,

$$\phi\colon G \mapsto (t(T,G))_{T\in\{\text{trees of height} \leq k\}}.$$

The WEISFEILER-LEHMAN graph kernel [38] leverages this idea for node-labelled graphs by strong use of node labels.

Other approaches such as the cyclical pattern kernel [21] use induced densities t_{ind} of graphs. The associated kernel is NP-hard to compute.

This means, that many graph kernels in the literature are associated to graph homomorphism densities. In Section 5, we will give bounds on the quality of homomorphism densities as features using graphon theory. This gives some theoretical foundation to the use of graph kernels and will be based on WASSERSTEIN sufficient conditions studied next.

4.3 Wasserstein Sufficient Conditions

Without an efficient solution to the graph isomorphism problem, all efficiently computable graph kernels will suffer from the inability to distinguish all graphs reliably. Thus, one can only expect sufficient conditions from the use of graph kernels, i.e. differences in feature vector distributions imply differences in data generating processes.

We formulate our stability estimates in terms of WASSERSTEIN distance. Let $(M, d, \mathcal{A})$ be a metric measurable space. The WASSERSTEIN-1 (or KANTOROVICH) distance between probability measures $\nu, \mu \in \mathcal{P}(M)$ (cf. [45, (6.1)]) is

$$\mathcal{W}^1_d(\mu,\nu) := \min_{\gamma\in\Pi(\mu,\nu)} \int d(x,y)\mathrm{d}\gamma(x,y) = \sup_{f\in\mathrm{Lip}_1(F)} \int f\mathrm{d}(\mu-\nu), \tag{4.7}$$

where $\Pi(\mu,\nu)$ is the set of all probability measures γ having first resp. second marginals μ resp. ν and $\mathrm{Lip}_1(M)$ is the set of 1-LIPSCHITZ functions $F \to \mathbb{R}$. The second equality is known as KANTOROVICH *duality* [45, (6.3)].

For a feature map $\phi\colon M \to H$, *Stability estimates* are inequalities of the form

$$\mathcal{W}^1_{d_H}(\phi_*\mathbb{G}(k,W), \phi_*\mathbb{G}(k,W')) \leq c_\phi \delta_\square(W,W') + o_k(1). \tag{4.8}$$

Here, $_*$ denotes the measure push-forward and c_ϕ is a constant only depending on the feature embedding ϕ. They show—up to an $o_k(1)$ additive error—Lipchitz variation of feature vectors with the data generating processes. In Chapter 5, we will derive such estimates.

5 Stable Features for Graph Classification

We now study classification problems with a decorated graphon as underlying data-generating process.

In Section 5.1 we contribute to this study by proving the *stability* of homomorphism densities as features. In Section 5.3 we supply a similar stability estimate for graph spectra. Along the way, we prove convergence of degree sequences and graph spectra in Section 5.2. We give a data application in Section 5.4.

Stability estimates for graph classification are the best we can hope for: As we saw in Chapter 4, computing the complete kernel of all homomorphism densities of a graph is NP-hard. Therefore, one cannot expect bounds on the data generating process in terms of the distribution of features. We show bounds in the other direction, so-called stability estimates, as introduced in Section 4.3.

Data-Generating Process Throughout this section, we assume the following classification problem: Let $\mathcal{W}, \mathcal{W}'$ be decorated graphons. Let $k \in \mathbb{N}$ and let

$$W_1, \ldots, W_n \sim \mathbb{G}(k, \mathcal{W}), \qquad W_1', \ldots, W_n' \sim \mathbb{G}(k, \mathcal{W}').$$

In addition, fix an unweighted graph F.

We assume an equal number of observations from each of the alternatives, which is done for notational convenience only. We will point out where proofs need to be changed to accommodate the general case whenever necessary.

5.1 Stability of Homomorphism Densities

The empirical distributions of homomorphism densities are defined as follows:

$$t := \frac{1}{n} \sum_{i=1}^{n} \delta_{t(F, W_i)} \qquad t' := \frac{1}{n} \sum_{i=1}^{n} \delta_{t(F, W_i')}.$$

The following stability estimate for homomorphism densities is our first main contribution.

Theorem 5.1. *With probability* $1 - 2e^{-cn^{\frac{3}{5}}} - 4e^{-\frac{2}{k^2}n^{\frac{3}{5}}}$,

$$\mathcal{W}^1_{|\bullet|}(t,\bar{t}) \leq |E(F)|\delta_\square(\mathbb{E}\mathcal{W},\mathbb{E}\mathcal{W}') + (1462\sqrt{2}+3)n^{-\frac{1}{5}}$$

Lemma 5.2 ([20, Theorem 1.1])**.** *Let* $X \sim \mu$, $X \in [0,1]$. *Let* $X_1,\dots,X_n \overset{iid}{\sim} \mu$ *and* $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_n}$. *Then for an absolute constant* C

$$\mathbb{E}[\mathcal{W}^1_{|\cdot|}(\mu_n,\mu)] \leq Cn^{-\frac{1}{5}}.$$

In fact, a close inspection of the proof given in [20, Theorem 1.1] shows that the constant can be taken to be $C = 731\sqrt{2}$.

Lemma 5.3 ([17, Theorem 2])**.** *Let* $\mu \in \mathcal{P}(\mathbb{R})$ *such that for* $X \sim \mu$, $\ell = \mathbb{E}[e^{\gamma X^\alpha}] < \infty$ *for some choice of* γ *and* α. *Then one has with probability at least* $1 - e^{-cn\varepsilon^2}$

$$\mathcal{W}^1_{|\bullet|}(\mu_n,\mu) \leq \varepsilon$$

for any $\varepsilon \in [0,1]$ *and* c *only depending on* ℓ, γ *and* α.

Proof of Theorem 5.1. By the triangle inequality,

$$\begin{aligned}\mathcal{W}^1_{|\bullet|}(t,\bar{t}) \leq \mathcal{W}^1_{|\bullet|}(t,\mathcal{L}(t_k(F,\mathcal{W}))) + \mathcal{W}^1_{|\bullet|}(\mathcal{L}(t_k(F,\mathcal{W})),\bar{t}) \\ + \mathcal{W}^1_{|\bullet|}(\mathcal{L}(t_k(F,\mathcal{W}')),\mathcal{L}(t_k(F,\mathcal{W}))).\end{aligned}$$

Combining Lemma 5.2 with Lemma 5.3 and choosing $\varepsilon = n^{-\frac{1}{5}}$, the first two summands are bounded by $(1462\sqrt{2}+2)n^{-\frac{1}{5}}$ with probability at least $1 - 2e^{-cn^{\frac{3}{5}}}$. For the last term take any coupling of $t_k(F,\mathcal{W}')$ and $t_k(F,\mathcal{W})$. Then, by Theorem 3.7, choosing $\varepsilon = n^{-\frac{1}{5}}$, we get with probability at least $1 - 4\exp(-\frac{2n^{\frac{3}{5}}}{k^2})$ that

$$|t_k(F,\mathcal{W}) - t_k(F,\mathcal{W}')| \leq n^{-\frac{1}{5}}.$$

Using a union bound, the theorem follows.

□

As promised, we remark, that if one would like to allow for a different number of observations for the two alternatives, then the concentration bound in combining Lemma 5.2 with Lemma 5.3 has to be applied separately to the first two summands.

5.2 Convergence of Graph Statistics

Convergence of spectra for graphons was established in [9] using a linear identity connecting homomorphism densities to spectra, Lemma 5.5 below. We extend this result in two dimensions. On the one hand, we generalise convergence of spectra to the setting of decorated graphons. On the other hand, we give a convergence result also for the degree sequence. This is the content of the main result, Theorem 5.4. We collect the necessary definitions for the formulation of Theorem 5.4 in the following.

Spectra Let $W\colon [0,1]^2 \to [0,1]$ be a graphon. Define

$$T_W\colon L^2([0,1]) \to L^2([0,1]), \quad T_W f(x) = \int_{[0,1]} W(x,y) f(y) \mathrm{d}\,\mathrm{Unif}_{[0,1]}(y).$$

Let (V,c) be a block model. Then define the linear operators

$$T_c\colon \mathbb{R}^V \cong L^2(\mathrm{Unif}_V) \to L^2(\mathrm{Unif}_V), f \mapsto \left(y \mapsto \int_V c(x,y) f(x) \mathrm{d}\,\mathrm{Unif}_V(x) \right).$$

The operator T_W is a HILBERT-SCHMIDT integral operator and hence compact [3, Section 8.15]. By the RIESZ-SCHAUDER spectral theorem for compact operators [3, Theorem 9.9] the spectrum of W has at most one accumulation point at 0 and all nonzero eigenvalues have finite multiplicity.

Hence, we can enumerate all eigenvalues of T_W order-reservingly with two sequences: Enumerate all non-negative eigenvalues of T_W with multiplicity in a weakly decreasing order to obtain $(\lambda_i^{W+})_{i\in\mathbb{N}} \subseteq \mathbb{R}_{\geq 0}$. Similarly, let $(\lambda_i^{W-})_{i\in\mathbb{N}} \subseteq \mathbb{R}_{\leq 0}$ denote the weakly increasing sequence of all non-positive eigenvalues with multiplicity. A straightforward adaptation of this definition applies to the operators T_{X_n} belonging to a random n-block model X_n, yielding sequences $(\lambda_i^{X_n+})_{i\in\mathbb{N}} \subseteq \mathbb{R}_{\leq 0}$ and $(\lambda_i^{X_n-})_{i\in\mathbb{N}} \subseteq \mathbb{R}_{\geq 0}$.

Graph Transformations Transformations of adjacency matrices and their eigenvalues have important implications for graph structure [40]. We define two such in the following. Let $G = (V,c)$ be a block model and W be a graphon. Denote δ_G the matrix with the column sums of weights (the degrees) on the diagonal. Define the graph LAPLACIAN as

$$l_G := \delta_G - c.$$

Furthermore, define the normalised graph LAPLACIAN as

$$n_G(x,y) := E_{|V(G)|} - \int_{d(x,x)\neq 0} d_G(x,x)^{-1} c(x,y) \mathrm{d}\,\mathrm{Unif}_V(x).,$$

where $E_{|V(G)|}$ is the identity matrix of size $|V(G)|$.

The *degree distribution* is of real-world graphs has been extensively studied [4]. It is defined as

$$d_G = \frac{1}{|V(G)|} \sum_{v \in V(G)} \delta_{\int_{V(G)} c(x,y) \mathrm{d}\,\mathrm{Unif}_{V(G)}(y)} = \left(\int_{V(G)} c(\bullet, y) \mathrm{d}\,\mathrm{Unif}_{V(G)}(y) \right)_* \mathrm{Unif}_{V(G)}$$

for a block model G and

$$D_W = \left(\int_{[0,1]} W(\bullet, y) \mathrm{d}\,\mathrm{Unif}_{[0,1]}(y) \right)_* \mathrm{Unif}_{[0,1]}.$$

for a graphon W.

The main result of this subsection is the following.

Theorem 5.4. *Let $\mathcal{W}$ be a decorated graphon and let $X_n \sim \mathbb{G}(n, \mathcal{W})$. Then*

(1) the eigenvalues of the operator T_W converge to the eigenvalues of the expectation graphon,

$$\lambda_i^{X_n +} \xrightarrow[n \to \infty]{a.s.} \lambda_i^{\mathbb{E}\mathcal{W}+} \qquad\qquad \lambda_i^{X_n -} \xrightarrow[n \to \infty]{a.s.} \lambda_i^{\mathbb{E}\mathcal{W}-}.$$

(2) for $\mathbb{G}(\infty, \mathcal{W})$-a.e. $\omega \in \Omega$, the degree sequences converge weakly,

$$d_{X_n}(\omega) \xrightarrow[n \to \infty]{w} D_{\mathbb{E}\mathcal{W}}.$$

We note that the convergence result for degrees cannot be defined in a pointwise manner, as the degree distribution may be purely continuous.

Lemma 5.5. *For a block model $G = (V, c)$ and a graphon W and C_k the cycle graph on k nodes, one has that $\sum_{i=1}^{\infty} (\lambda_i^{W\pm})^k$ exist and*

$$\sum_{\lambda \in \Lambda(T_c)} \lambda^k = t(C_k, G) \qquad\qquad \sum_{i=1}^{\infty} (\lambda_i^{W+})^k + \sum_{i=1}^{\infty} (\lambda_i^{W-})^k = t(C_k, W). \tag{5.1}$$

If S_k is the star graph on $k+1$ nodes, then

$$t(S_k, G) = \frac{1}{n^{k+1}} \sum_{v \in V(G)} d_G[\{v\}]^k = \int x^k \mathrm{d} d_G(x)$$

The first statement appears in [9, (6.4)], the first equality of the second statement appears in [27, Example 5.10]. The last equality follows by definition of d_G. In the additional material to this section, we will show the following extension of the second statement.

Lemma 5.6. *Let W be a graphon. Then*

$$t(S_k, W) = \int x^k \mathrm{d} D_G(x)$$

Proof of Theorem 5.4. By Corollary 3.8, we have a.s. convergence of any homomorphism density $t_n(F,\mathcal{W}) = t(F,X_n)$. As the set of finite graphs is countable, we get a.s. convergence of

$$(t_n(F,\mathcal{W}))_{F\in\mathcal{F}} \to (t(F,\mathbb{E}\mathcal{W}))_{F\in\mathcal{F}}. \tag{5.2}$$

It suffices to show convergence of spectra for such $\omega \in \Omega$ such that the convergence (5.2) holds. Fix any such ω. Then,

$$\begin{aligned}\sum_{\lambda\in\Lambda(T_{X_n})} \lambda^k(\omega) = t(C_k,X_n)(\omega) \xrightarrow{n\to\infty} t(C_k,\mathbb{E}\mathcal{W}) = \sum_{i=1}^{\infty}(\lambda_i^{W+})^k + \sum_{i=1}^{\infty}(\lambda_i^{W-})^k \\ \int x^k \mathrm{d}d_{X_n}(x)(\omega) = t(S_k,X_n)(\omega) \xrightarrow{n\to\infty} t(S_k,\mathbb{E}\mathcal{W}) = \int x^k \mathrm{d}D_{\mathbb{E}\mathcal{W}}(x)\end{aligned} \tag{5.3}$$

Hence, (5.3) can the be rewritten as

$$\begin{aligned}\sum_{i=1}^{\infty}(\lambda_i^{X_n+})^k(\omega) + \sum_{i=1}^{\infty}(\lambda_i^{X_n-})^k(\omega) \xrightarrow{n\to\infty} \sum_{i=1}^{\infty}(\lambda_i^{\mathbb{E}\mathcal{W}+})^k + \sum_{i=1}^{\infty}(\lambda_i^{\mathbb{E}\mathcal{W}-})^k \\ \int x^k \mathrm{d}d_{X_n}(x)(\omega) \xrightarrow{n\to\infty} \int x^k \mathrm{d}D_{\mathbb{E}\mathcal{W}}(x)\end{aligned}$$

which means we have convergence of all moments for the sequences $(\lambda_i^{X_n\pm})_{i\in\mathbb{N}}$ respectively the measures $d_{X_n}(\omega)$. We would like to conclude convergence of each element in the sequence and weak convergence for the measure. We know that there are subsequential limits $a_i^{\pm}(\omega)$ respectively $d(\omega)$ of the sequences as $((\lambda_i)^{X_n\pm})_{i\in\mathbb{N}} \subseteq [-1,1]$ [27, (7.20)] lies in a compact set and $(d_{X_n})_{n\in\mathbb{N}}$ is tight. Indeed, $d_{X_n}[\{x||x| \geq u\}](\omega) \leq \frac{1}{u^2}\int x^2 d_{X_n}(x)(\omega)$ is a bounded sequence by the assumption of moment convergence. To prove the theorem, it suffices to show that the subsequential limits are equal to the limits in the statement. For the first sequence, this can be concluded from the corollary to the monotone reordering theorem [27, Proposition A.21].

The measure $D_{\mathbb{E}\mathcal{W}}(\omega)$ is compactly supported on $[0,1]$. This implies that its moment generating function has positive radius of convergence, which implies that it is uniquely determined by its moment sequence [6, Theorem 30.1]. Hence also $d(\omega) = D_{\mathbb{E}\mathcal{W}}$, showing the theorem. □

5.3 Stability of Spectra

We turn to proving stability of spectra. By means of Lemma 5.5, we can connect eigenvalues of graphs to homomorphism densities of cycles. These in turn can be connected to cut distance between the expectation graphons through Corollary 3.11.

We will view spectra as point measures: Denote by $\lambda_G = \frac{1}{|V(G)|}\sum_{\lambda\in\Lambda(T_c)}\delta_\lambda \in \mathcal{P}(\mathbb{R})$ the point measure with masses on the eigenvalues of T_c for a weighted graph $G = (V,c)$.

The empirical distribution of eigenvalue point measures then is

$$\bar{\lambda} = \frac{1}{n}\sum_{i=1}^{n} \lambda_{W_i}, \qquad \bar{\lambda}' = \frac{1}{n}\sum_{i=1}^{n} \lambda_{W_i'}. \tag{5.4}$$

Denote in addition the homomorphism densities of length-v cycles by

$$\bar{t}_v := \frac{1}{n}\sum_{i=1}^{n} \delta_{t(C_v, W_i)}, \qquad \bar{t}_v := \frac{1}{n}\sum_{i=1}^{n} \delta_{t(C_v, W_i')}. \tag{5.5}$$

The following is the main result connecting the WASSERSTEIN distance of empirical distributions to the distances of cycle homomorphism densities.

Theorem 5.7. *Let $\bar{\lambda}$, $\bar{\lambda}'$ be as in* (5.4) *and $\bar{t}_v$, $\bar{t}_v'$ be as in* (5.5). *Then*

$$\mathcal{W}^1_{\mathcal{W}^1_{|\bullet|}}(\bar{\lambda}', \bar{\lambda}) \leq \inf_{v \in \mathbb{N}} k^{-1} 2(4e)^v \sum_{i=1}^{v} \mathcal{W}^1_{|\bullet|}(\bar{t}_i, \bar{t}_i') + \frac{3}{\pi v}$$

Corollary 5.8 (Stability of Spectra). *Let $\bar{\lambda}$, $\bar{\lambda}'$ be as in* (5.4). *Then*

$$\mathcal{W}^1_{\mathcal{W}^1_{|\bullet|}}(\bar{\lambda}, \bar{\lambda}') \leq \inf_{v \in \mathbb{N}} \delta_{\square}(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}') 2(4e)^v v^2 k^{-1} + \left(2(4e)^v k^{-1} (1462\sqrt{2} + 3) n^{-\frac{1}{5}} + \frac{3}{\pi v} \right)$$

Lemma 5.9 (Corollary of [1, p. 200]). *Let f be a 1-*LIPSCHITZ *function on $[-1, 1]$. Then there is a polynomial p of degree v such that $\|f - p\|_\infty \leq \frac{3}{\pi n}$.*

Lemma 5.10 ([36, Lemma 4.1]). *Let $\sum_{i=0}^{v} a_i x^v$ be a polynomial on $[-1, 1]$ bounded by M. Then*

$$a_i \leq (4e)^d M$$

Proof of Theorem 5.7. Consider any coupling (λ, λ') of $\bar{\lambda}$ and $\bar{\lambda}'$. Let $(W_i)_{j\ell}$ be the adjacency matrix of graph W_i. Note that $\operatorname{supp} \lambda, \operatorname{supp} \lambda' \subseteq [-1, 1]$, as $0 \leq (W_i)_{j\ell} \leq 1$ for any $i \in [n], j, \ell \in [k]$ and hence

$$\|T_{W_i} f\|_{L^2([0,1])} \leq \|W_i\|_\infty^2 \|f\|_{L^2([0,1])} \leq \|f\|_{L^2([0,1])}$$

where $\|W_i\|_\infty$ is the maximum norm of the matrix $((W_i)_{j\ell})_{j,l \in [k]}$. One has by the definition of the WASSERSTEIN distance $\mathcal{W}^1_{\mathcal{W}^1_{|\bullet|}}$ and KANTOROVICH duality

$$\mathcal{W}^1_{\mathcal{W}^1_{|\bullet|}}(\bar{\lambda}', \bar{\lambda}) \leq \mathbb{E}\left[\mathcal{W}^1_{|\bullet|}(\lambda, \lambda')\right] = \mathbb{E}\left[\sup_{\substack{f\colon [-1,1] \to \mathbb{R} \\ \operatorname{Lip}(f) \leq 1}} \int f(x) \mathrm{d}(\lambda - \lambda')\right] \tag{5.6}$$

Fix any $\omega \in \Omega$. By Lemma 5.9 one can approximate Lipschitz functions by polynomials of bounded degree,

$$\sup_{\substack{f\colon [-1,1] \to \mathbb{R} \\ \operatorname{Lip}(f) \leq 1}} \int f(x) \mathrm{d}(\lambda - \lambda')(\omega) \leq \sup_{\substack{\deg(f) \leq v \\ |f| \leq 2}} \int f(x) \mathrm{d}(\lambda - \lambda')(\omega) + \frac{3}{\pi v}. \tag{5.7}$$

Here, $|f| \leq 2$ can be assumed as f is defined on $[-1, 1]$ and its 1-LIPSCHITZ continuity and because the integral in (5.7) is invariant with respect to the addition of constant functions.

Hence, by Lemma 5.10 and the triangle inequality

$$\sup_{\substack{\deg(f)\leq v \\ |f|\leq 2}} \int f(x)\mathrm{d}(\lambda - \lambda')(\omega) \leq \sum_{i=1}^{v} 2(4e)^{v} \left| \int x^{i}\mathrm{d}(\lambda - \lambda') \right| (\omega)$$

$$= \sum_{i=1}^{v} 2(4e)^{v}k^{-1} \left| \sum_{w\in\lambda} w^{i} - \sum_{w'\in\lambda'} w^{i} \right| (\omega)$$

Inserting this into (5.7) and taking expectations, one gets

$$\mathcal{W}^{1}_{\mathcal{W}^{1}_{|\bullet|}}(\bar{\lambda}, \bar{\lambda}') \leq \frac{3}{\pi v} + \sum_{i=1}^{v} 2(4e)^{v}k^{-1}\mathbb{E}\left[\left| \sum_{w\in\lambda} w^{i} - \sum_{w'\in\lambda'} w^{i} \right| \right]$$

for any coupling (λ, λ') of $\bar{\lambda}$ and $\bar{\lambda}'$. Now consider a coupling (λ, λ') of $\bar{\lambda}$ and $\bar{\lambda}'$ such that $\bar{t}$, $\bar{t}'$ (which are functions of λ, λ' by (5.1)) are optimally coupled. Then by the definition of $\bar{\lambda}$, $\bar{\lambda}'$, $\bar{t}$ and $\bar{t}'$, and Lemma 5.5 one gets that

$$\mathcal{W}^{1}_{\mathcal{W}^{1}_{|\bullet|}}(\bar{t}_{i}, \bar{t}'_{i}) = \mathbb{E}\left[\left| \sum_{w\in\ \lambda} w^{i} - \sum_{w\in\bar{\lambda}'} w'^{i} \right| \right].$$

Hence, by (5.6), one has for any $v \in \mathbb{N}$

$$\mathcal{W}^{1}_{\mathcal{W}^{1}_{|\bullet|}}(\bar{\lambda}', \bar{\lambda}) \leq 2(4e)^{v}k^{-1} \sum_{i=1}^{v} \mathcal{W}^{1}_{|\bullet|}(\bar{t}_{i}, \bar{t}'_{i}) + \frac{3}{\pi v}, \tag{5.8}$$

proving the claim.

□

Remark 5.11. *A statement with slightly worse constants holds for the degree distribution. For the proof, one has to replace cycles by stars in the definition of $\bar{t}_i$ and $\bar{t}'_i$ and use the second part of Lemma 5.5.*

Proof of Corollary 5.8.

$$\begin{aligned}
\mathcal{W}^{1}_{\mathcal{W}^{1}_{|\bullet|}}(\bar{\lambda}, \bar{\lambda}') &\leq \inf_{v\in\mathbb{N}} 2(4e)^{v}k^{-1} \sum_{i=1}^{v} \mathcal{W}^{1}_{|\bullet|}(\bar{t}_{i}, \bar{t}'_{i}) + \frac{3}{\pi v} \\
&\leq \inf_{v\in\mathbb{N}} 2(4e)^{v}k^{-1} \sum_{i=1}^{v} \left(i\delta_{\square}(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}') + (1462\sqrt{2} + 3)n^{-\frac{1}{5}} \right) + \frac{3}{\pi v} \\
&\leq \inf_{v\in\mathbb{N}} \delta_{\square}(\mathbb{E}\mathcal{W}, \mathbb{E}\mathcal{W}')2(4e)^{v}v^{2}k^{-1} + \left(2(4e)^{v}k^{-1}(1462\sqrt{2} + 3)n^{-\frac{1}{5}} + \frac{3}{\pi v} \right)
\end{aligned}$$

where the first bound is from Theorem 5.7 and the second bound is from Theorem 5.1. The third is basic algebra. □

The infimum in Theorem 5.7 signifies a payoff: High powers polynomials corresponding to large cycles are volatile (multiplicative error, variance), but low power polynomials corresponding to small cycles are not very expressive (additive error, bias).

Other stability estimates can be found in the literature on topological data analysis [12] bounding variations of persistence landscapes by variation of underlying data-generating processes.

5.4 Application: Neuroscience

The Dataset and Classification Problem We are given a classification problem on weighted graphs based on diffusion tensor images (DTI) of 56 individuals [35]. There are three different groups of subjects (labels): 17 human controls (HC), 18 persons affected by autoimmune disease *Lupus Erythematodes* (SLE[1]) and 18 SLE patients that in addition have been reported to show neuropsychiatric symptoms.

We consider the classification problems HC versus NPSLE and HC versus SLE as the authors of [24] did, working with the same dataset.

For each subject, there are six different weighted graphs, each consisting of 1164 nodes. Nodes are anatomically localised and correspond to the regions of the finest resolution of the Talairach brain atlas [42]. Four of the six weighted graphs are scalar functions of the DTI tensor field integrated along so-called fibers, i.e. curves connecting the regions that nodes correspond to. We present the data pipeline in Figure 5.1 and give the definitions of the four values in the additional material following this section. The other two weighted graphs give the total length and the number of all fibers connecting two regions. These values in particular measure tissue integrity with respect to whether tractography (fiber tracing) algorithms [5] can detect fibers in the images.

Our Approach First, we reduce dimensions making use of the used brain atlas: We average over regions to get a graph on a coarser level of the Talairach brain atlas hierarchy [42] on 344 nodes. On these graphs, we compute the eigenvalues of the adjacency matrix, the graph Laplacian, the normalised graph Laplacian and the degree matrix for all six values, effectively treating the connectomes as exchangeable random graphs. We restrict to the ten largest and ten smallest eigenvalues of the graph Laplacian of the given graphs and concatenate the features of different measurements to get the final features. We selected the features of the graph Laplacian as it showed different distributions for the three groups for all values, cf. Figure 5.2. We chose smallest and largest eigenvalues as these showed in general stronger variations than intermediate eigenvalues.

[1] Systemic *Lupus Erythematodes* is a chronic inflammatory disease affecting multiple human organs which occurs in 0.01 to 0.1% of the general population [39]. 20 to 70% of those affected by SLE have been reported to show neuropsychiatric symptoms [11].

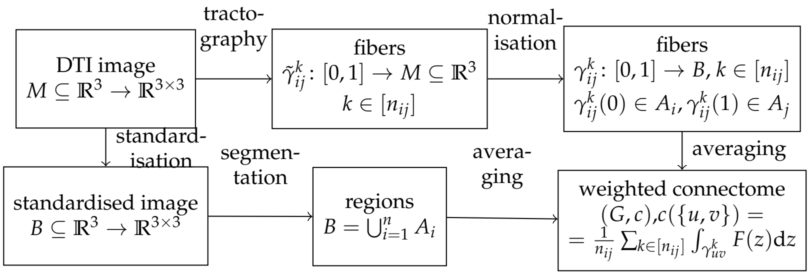

DTI image
$M \subseteq \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$
tractography
fibers
$\tilde{\gamma}_{ij}^k\colon [0,1] \to M \subseteq \mathbb{R}^3$
$k \in [n_{ij}]$
normalisation
fibers
$\gamma_{ij}^k\colon [0,1] \to B, k \in [n_{ij}]$
$\gamma_{ij}^k(0) \in A_i, \gamma_{ij}^k(1) \in A_j$
standardisation
standardised image
$B \subseteq \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$
segmentation
regions
$B = \bigcup_{i=1}^n A_i$
averaging
averaging
weighted connectome
$(G,c), c(\{u,v\}) =$
$= \frac{1}{n_{ij}} \sum_{k\in[n_{ij}]} \int_{\gamma_{uv}^k} F(z)\mathrm{d}z$

Figure 5.1: Preprocessing pipeline for weighted structural connectomes. A brain can be seen as a tensor field $B : M \subseteq \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ of flows. The domain of this tensor field is partitioned into regions $A_1, \dots, A_n$, called brain regions. Fibers are parametrized curves from one region to another. Each scalar function $F : \mathbb{R}^3 \to \mathbb{R}$ converts a brain into a weighted graph on n nodes, where the weight between regions i and j is F averaged or integrated over all fibers between these regions.

	HC vs. NPSLE	HC vs. SLE
[24]	76%	73%
Eigenvalues	78.3%	67.5%

Table 5.1: Accuracies for neuroscience classification tasks compared to [24]. We use concatenation of eigenvalues for the different weighted graphs given and get competitive result comparing to the anatomically localised approach in [24].

In the classification problem HC versus NPSLE we use the 40 fiber tracing-based values "length" (len) and "number of fibers" (num), as we expected that neuropsychiatric symptoms are related to tractography-related features. In the classification problem HC versus SLE we use all 120 features from the six different weighted graphs.

Results We use these features in a support-vector machine with ℓ^1-penalty (cf. Remark 4.3). We summarise the leave-one-out cross validation accuracies in Table 5.1. A permutation test [32] shows that these results are significant (at 10%).

Hence, an easy classifier can get competetive results on this classification task. This can be interpreted as a characterisation of structural connectomes affected by *Lupus Erythematodes*: The average weight of cycles in the expectation graphon of human controls and patients having the disease is altered. As the spectra contain information on cycles of all lengths, these features use both local information, corresponding to very short cycles, and global information, corresponding to very long cycles.

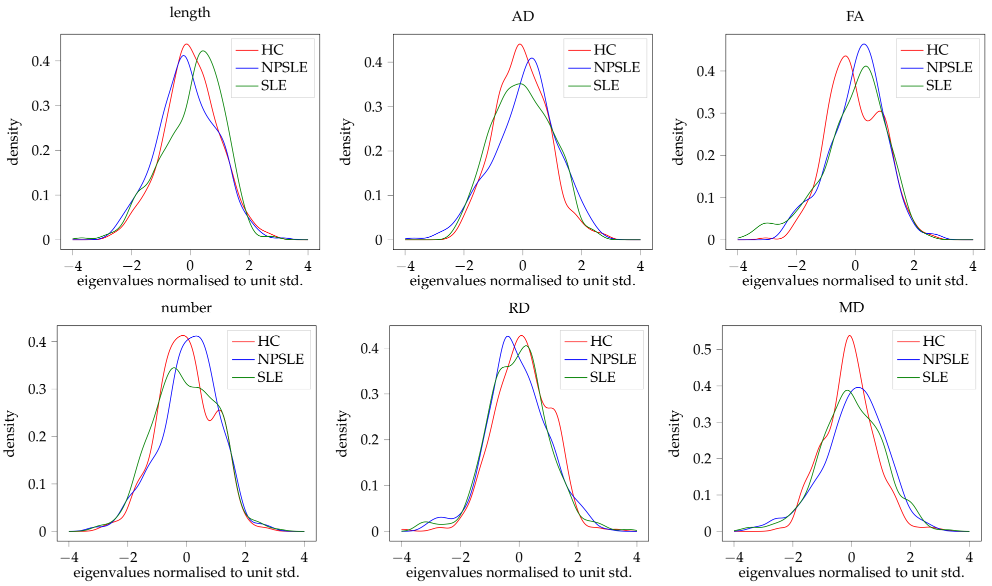


Figure 5.2: Density of first and last ten eigenvalues (normalised to zero mean unit standard deviation) of the graph Laplacian for all six values.

5.5 Additional Material

Omitted Proofs As promised, we prove the two remaining statements.

Proof of Lemma 5.6. Note that by (2.3),

$$\begin{aligned} t(S_k, W) &= \int_{[0,1]^{k+1}} W(x_1, x_2) W(x_1, x_3) \cdots W(x_1, x_{k+1}) \mathrm{d}\,\mathrm{Unif}_{[0,1]}^{k+1} \\ &= \int_{[0,1]} \left(\int_{[0,1]} W(x, y) \mathrm{d}\,\mathrm{Unif}_{[0,1]}(y) \right)^k \mathrm{d}\,\mathrm{Unif}_{[0,1]}(x) \\ &= \int_{[0,1]} x^k \mathrm{d} \left(\int_{[0,1]} W(\bullet, y) \mathrm{d}\,\mathrm{Unif}_{[0,1]}(y) \right)_* \mathrm{Unif}_{[0,1]} \\ &= \int_{[0,1]} x^k \mathrm{d} D_W(x). \end{aligned}$$

□

Explanation of values in Neuroscience Data In a diffusion model of tensor imaging[5], the movement of a water molecule is modelled as a Brownianmotion on a Riemannian manifold with metric d. View the metric as a matrix. Let $\lambda_1 \geq \lambda_2 \geq \lambda_3$ be the eigenvalues of the metric d. Then the radial diffusivity (RD) is defined as $\frac{\lambda_2+\lambda_3}{2}$, the axial diffusivity (AD) as λ_1 and the mean diffusivity (MD) by $\frac{\lambda_1+\lambda_2+\lambda_3}{3}$. Finally, fractional anisotropy (FA) is defined as

$$\sqrt{\frac{\sum_{i=1}^{3}(\lambda_i - \mathrm{MD})^2}{2\sum_{i=1}^{3}\lambda_i^2}}.$$

6 Conclusion and Open Problems

Forming a sound statistical foundation for random graphs is important for the empirical study of complex networks. In an effort to contribute to this endeavour, we studied the classification problem for weighted graphs.

The approach taken was to introduce decorated graphons and generalise known results from the graphon literature to this case. In particular, we provided stability estimates for features in a general weighted random graph model. We hope that the progress made in this thesis will set the groundwork for a better understanding of random networks.

Nevertheless, there remain open problems that we would like to mention:

We do not provide stability estimates for the graph LAPLACIAN. As these features performed well in an applied task, it is worthwhile to characterise their convergence. In particular we believe it is an interesting problem whether the spectrum of the graph LAPLACIAN can be related to homomorphism densities of unweighted graphs by a linear equality.

As we mentioned in Chapter 3, defining a generalisation of the cut metric to decorated graphons is a problem yet to be solved. The challenge is metrising the space of decorated graphons in a way that allows fully characterising the weak convergence of weighted exchangeable random graphs.

The convergence $t_n(F, \mathcal{W}) \to t(F, \mathbb{E}\mathcal{W})$ can be interpreted as an exchangeable instance of a strong law of large numbers. To better understand in how far variants of homomorphism densities scaled in different ways are able to capture moments other than the expectation is a worthwhile task. Ultimately, one would hope for a central limit type result.

Concerning our discussion of the applied literature in Chapter 4, our observation that most features used in classification of graph data are homomorphism densities does not include all benchmarks: The shortest-path histogram kernel [10] uses edge lengths in the metric closure of graphs as features. A convergence result for edge length distributions in a decorated graphon setting would not only constitute an interesting theoretical result, but also lead to a better understanding of such features.

Bibliography

[1] Naum I Achieser. *Theory of approximation*. Courier Corporation, 2013.

[2] David J Aldous. “Representations for partially exchangeable arrays of random variables”. In: *Journal of Multivariate Analysis* 11.4 (1981), pp. 581–598.

[3] Hans Wilhelm Alt. *Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung*. Springer-Verlag, 2012.

[4] Albert-László Barabási and Réka Albert. “Emergence of scaling in random networks”. In: *science* 286.5439 (1999), pp. 509–512.

[5] Peter J Basser et al. “In vivo fiber tractography using DT-MRI data”. In: *Magnetic resonance in medicine* 44.4 (2000), pp. 625–632.

[6] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[7] Christopher M Bishop. “Pattern recognition”. In: *Machine Learning* 128 (2006), pp. 1–58.

[8] Christian Borgs et al. “Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing”. In: *Advances in Mathematics* 219.6 (2008), pp. 1801–1851.

[9] Christian Borgs et al. “Convergent sequences of dense graphs II. Multiway cuts and statistical physics”. In: *Annals of Mathematics* 176.1 (2012), pp. 151–219.

[10] Karsten M Borgwardt and Hans-Peter Kriegel. “Shortest-path kernels on graphs”. In: *Data Mining, Fifth IEEE International Conference on*. IEEE. 2005, 8–pp.

[11] RL Brey et al. “Neuropsychiatric syndromes in lupus Prevalence using standardized definitions”. In: *Neurology* 58.8 (2002), pp. 1214–1220.

[12] Peter Bubenik. “Statistical topological data analysis using persistence landscapes.” In: *Journal of Machine Learning Research* 16.1 (2015), pp. 77–102.

[13] Kamalika Chaudhuri and Sanjoy Dasgupta. “Rates of convergence for nearest neighbor classification”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 3437–3445.

[14] Luc Devroye, László Györfi and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media, 2013.

[15] Persi Diaconis and Svante Janson. “Graph limits and exchangeable random graphs”. In: *arXiv preprint arXiv:0712.2749* (2007).

[16] Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 2010.

[17] N. Fournier and A. Guillin. “On the rate of convergence in Wasserstein distance of the empirical measure”. In: *ArXiv e-prints* (Dec. 2013). arXiv: `1312.2128 [math.PR]`.

[18] Thomas Gärtner, Peter Flach and Stefan Wrobel. “On graph kernels: Hardness results and efficient alternatives”. In: *Learning Theory and Kernel Machines*. Springer, 2003, pp. 129–143.

[19] DN Hoover. *Relations on probability spaces and arrays of random variables*. Tech. rep. Institute for Advanced Study, Princeton, 1979.

[20] Joseph Horowitz and Rajeeva L Karandikar. “Mean rates of convergence of empirical measures in the Wasserstein metric”. In: *Journal of Computational and Applied Mathematics* 55.3 (1994), pp. 261–273.

[21] Tamás Horváth, Thomas Gärtner and Stefan Wrobel. “Cyclic pattern kernels for predictive graph mining”. In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2004, pp. 158–167.

[22] Svante Janson. *Graphons, cut norm and distance, couplings and rearrangements*. Tech. rep.

[23] Olav Kallenberg. *Probabilistic symmetries and invariance principles*. Springer Science & Business Media, 2006.

[24] Mohammad Khatami et al. “BundleMAP: anatomically localized features from dMRI for detection of disease”. In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2015, pp. 52–60.

[25] Olga Klopp and Nicolas Verzelen. “Optimal graphon estimation in cut distance”. In: *arXiv preprint arXiv:1703.05101* (2017).

[26] Hermann König. *Eigenvalue distribution of compact operators*. Vol. 16. Birkhäuser, 2013.

[27] László Lovász. *Large networks and graph limits*. Vol. 60. American Mathematical Soc., 2012.

[28] László Lovász and Balázs Szegedy. “Limits of compact decorated graphs”. In: *arXiv preprint arXiv:1010.5155* (2010).

[29] László Lovász and Balázs Szegedy. “Limits of dense graph sequences”. In: *Journal of Combinatorial Theory, Series B* 96.6 (2006), pp. 933–957.

[30] László Lovász and Balázs Szegedy. “Szemerédi’s lemma for the analyst”. In: *Geometric and Functional Analysis* 17.1 (2007), pp. 252–270.

[31] Colin McDiarmid. “On the method of bounded differences”. In: *Surveys in combinatorics* 141.1 (1989), pp. 148–188.

[32] Markus Ojala and Gemma C Garriga. “Permutation tests for studying classifier performance”. In: *Journal of Machine Learning Research* 11.Jun (2010), pp. 1833–1863.

[33] Jan Ramon and Thomas Gärtner. "Expressivity versus efficiency of graph kernels". In: *Proceedings of the first international workshop on mining graphs, trees and sequences*. 2003, pp. 65–74.

[34] Walter Rudin. *Real and complex analysis*. Tata McGraw-Hill Education, 1987.

[35] Tobias Schmidt-Wilcke et al. "Diminished white matter integrity in patients with systemic lupus erythematosus". In: *NeuroImage: Clinical* 5 (2014), pp. 291–297.

[36] Alexander A Sherstov. "Making polynomials robust to noise". In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM. 2012, pp. 747–758.

[37] Nino Shervashidze et al. "Efficient graphlet kernels for large graph comparison." In: *AISTATS*. Vol. 5. 2009, pp. 488–495.

[38] Nino Shervashidze et al. "Weisfeiler-lehman graph kernels". In: *Journal of Machine Learning Research* 12.Sep (2011), pp. 2539–2561.

[39] Emily C Somers et al. "Population-Based Incidence and Prevalence of Systemic Lupus Erythematosus: The Michigan Lupus Epidemiology and Surveillance Program". In: *Arthritis & Rheumatology* 66.2 (2014), pp. 369–378.

[40] Daniel A Spielman. "Spectral graph theory and its applications". In: *Foundations of Computer Science, 2007. FOCS'07. 48th Annual IEEE Symposium on*. IEEE. 2007, pp. 29–38.

[41] Charles J Stone. "Consistent nonparametric regression". In: *The annals of statistics* (1977), pp. 595–620.

[42] Jean Talairach and Pierre Tournoux. *Co-planar stereotaxic atlas of the human brain. 3-Dimensional proportional system: an approach to cerebral imaging*. Thieme, 1988.

[43] Paul Turán. "On an extremal problem in graph theory". In: *Mat. Fiz. Lapok* 48 (1941), pp. 436–452.

[44] Victor Veitch and Daniel M Roy. "The class of random graphs arising from exchangeable random measures". In: *arXiv preprint arXiv:1512.03099* (2015).

[45] Cédric Villani. *Optimal transport: old and new*. Vol. 338. Springer Science & Business Media, 2008.

[46] Ji Zhu et al. "1-norm support vector machines". In: *Advances in neural information processing systems*. 2004, pp. 49–56.

Index

Nomenclature

$\bar{\lambda}$,$\bar{\lambda}'$	Empirical distribution, spectra, 36
$\bar{t}_v, \bar{t}'_v$	Empirical distributions, cycle homomorphism densities, 36
Bern_p	Bernoulli Distribution, 3
$\delta_\square$	Cut Metric, 9
δ_x	DIRAC Mass, 3
$\mathbb{E}\mathcal{W}$	Expectation Graphon, 12
$\mathbb{E}\nu$	Expectation of measure, 3
⚯	K_2, 3
$\mathbb{G}(\infty, W)$	Array Sampled from Graphon, 7
$\mathbb{G}(\infty, \mathcal{W})$	Array Sampled from Decorated Graphon, 11
$\mathbb{G}(k, \mathbf{X})$	k-sampling, exchangeable array, 5
$\mathbb{G}(k, \mathcal{W})$	k-Sampling, Decorated Graphon, 11
$\mathbb{G}(k, W)$	k-Sampling Graphon, 7
$\perp\!\!\!\perp$	Independence of random variables, 3
⟁	K_3, 3
$\Lambda(F)$	Spectrum of linear operator F, 3
$\mathcal{F}$	set of finite unweighted graphs, 3
$\mathcal{L}(X)$	Law of random variable X, 3
$\mathcal{P}(M)$	Set of probability measures, 3
$\mathcal{W}_d^1$	WASSERSTEIN distance, 29
$\mathcal{W}_G(x, y)$	Weighted graph as graphon, 13
ω	LANDAU small-omega, 3
$\mathrm{Inj}(V(F), V(G))$	Set of injections $V(F) \hookrightarrow V(G)$, 14
$\mathrm{Lip}_1(F)$	Set of Lipschitz functions with constant 1, 29
supp	Support of a measure, 36
ϕ	feature map, 26
$\Pi(\mu, \nu)$	Set of couplings, 29

π_n	Projection of graphs to $n-1$ first rows and columns, 5
τ_x	Translation by x, 4
$\triangle$	symmetric difference, 4
Unif_V	Uniform Distribution, 3
$\circ$	K_1, 3
d_G	Degree Distribution, block model, 34
D_W	Degree Distribution, graphon, 34
E_n	identity matrix, 33
l_G	graph Laplacian, 33
n_G	normalised graph Laplacian, 33
O	LANDAU big-o, 3
o	LANDAU small-o, 3
$S(A)$	Symmetric group acting on $A \subseteq [n]$, 15
t	homomorphism density, 7
$t(F, X_n)$	Block Model Homomorphism Density, 14
t, t'	empirical homomorphism densities, 31
t_{ind}	Induced homomorphism density, 7
t_{inj}	Injective homomorphism density, 14
T_c	Block model as linear operator, 33
$t_n(F, \mathcal{W})$	Finite homomorphism density, 14
T_W	Graphon as linear operator, 33
W^Φ	measure-preservingly transformed graphon, 7
X_k	Initial k-subarray, 5
$_*$	Measure Pushforward, 3