Network Learning with Semi-relaxed Gromov-Wasserstein: Graphon Estimation is a Barycenter Problem

Charles Dufour
EPFL (SDS)
Lausanne, Switzerland

Ulysse Naepels
EPFL (SMAT)
Lausanne, Switzerland

Leonardo V. Santoro
EPFL (SMAT)
Lausanne, Switzerland

Abstract

Estimating the generative mechanism of large-scale networks is a fundamental challenge in statistical machine learning. It requires the identification of the latent connectivity structure, which is in general an NP-hard combinatorial problem due to the absence of canonical node labels. We address this challenge by allowing for probabilistic couplings, thereby relaxing the assignment problem. Our estimation framework can be formulated as a semi-relaxed Gromov–Wasserstein objective and provides a low-dimensional representation of the generative structure. We solve this via a block-coordinate conditional gradient algorithm. Despite the relaxation, the resulting solution is typically deterministic: in fact, we show that the optimality gap between the relaxed solution and the deterministic assignment vanishes at rate $O(1/n)$, where n is the number of nodes. This allows for tractable recovery of the underlying model and enables rigorous statistical analysis: we establish consistency and minimax-optimal convergence rates for both stochastic block models and Hölder-smooth graphons. Our implementation scales efficiently with n, as demonstrated on both synthetic and real-world datasets.

1 Introduction

A central challenge in the analysis of network data is the recovery of latent structure. In statistics and machine learning, this includes tasks such as identifying communities, summarizing connectivity patterns, and inferring underlying generative mechanisms from large networks of structured data. Such problems are relevant across a wide range of applications, including the analysis of social networks [1], biological interaction networks [2, 3] such as protein-protein interactions [4, 5], web recommender systems [6], or infrastructure and communication systems [7]. In these settings, networks are often large, noisy, and unlabeled, which makes inference particularly challenging.

A fundamental difficulty common to network inference problems is the absence of a canonical node labeling. This permutation invariance introduces both conceptual and technical challenges: since model parameters are inherently identifiable only up to relabeling of the nodes, the problem is linked to a challenging latent variable inference problem (i.e. a non-parametric regression with unknown design points [8]). Any meaningful notion of estimation error must therefore compare networks *up to permutations*, rather than at the level of individual node indices. This renders standard loss functions for adjacency matrices – such as Frobenius or ℓ^2 norms – ill-suited to this setting. As a result, defining a suitable notion of distance and, in particular, what it means for an estimator to be close to the truth, becomes nontrivial. These issues already arise in simple parametric models, such as stochastic block models (SBMs, [9]) – where nodes are assigned to latent blocks that govern their connectivity patterns – and become even more pronounced in nonparametric settings, such as graphons [10]. From a computational perspective, inference in these models typically requires estimating both the latent structural parameters and the assignment of nodes to communities, a high-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

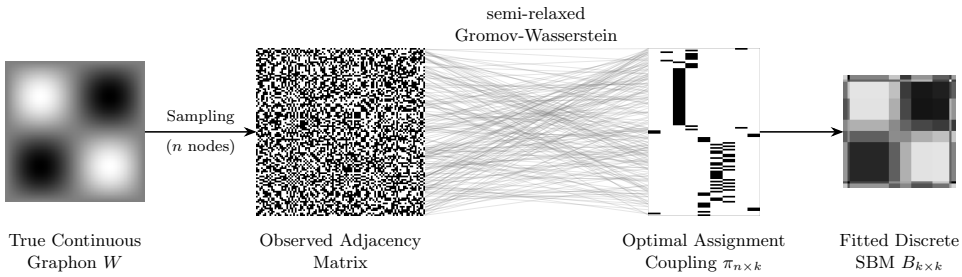


Figure 1: A finite adjacency matrix is sampled from a continuous latent graphon W. By minimizing the semi-relaxed Gromov-Wasserstein distance, we identify an optimal assignment coupling $\pi_{n\times k}$—observed here to be deterministic—which maps nodes to latent blocks. This results in a fitted discrete SBM $B_{k\times k}$ that serves as a low-dimensional approximation of W.

dimensional combinatorial problem. Determining the assignment map is itself NP-hard, rendering standard estimation approaches computationally intractable.

A natural and popular approach to the lack of canonical ordering in network data is offered by optimal transport, and in particular by the Gromov–Wasserstein (GW) metric [11]. The GW distance compares graphs by optimally aligning their pairwise relations, treating networks as metric measure spaces and assessing similarity by matching their internal geometry rather than their node identities. This framework has inspired a rich line of recent work, with applications to graph comparison, alignment, clustering, and barycenter computation [12–15]. While GW distances provide a natural, permutation-invariant notion of similarity between networks, their role in recovering generative parameters has only been recently algorithmically explored [16, 17] but not theoretically clarified. In particular, it remains unclear whether optimizing GW-based objectives can be interpreted as a statistically meaningful estimation procedure for latent network models or how the resulting transport plans relate to underlying node assignments and block structure. This gap motivates the present work, in which we show that relaxing hard node assignments to probabilistic couplings leads to a principled estimation framework that can be formulated as a semi-relaxed Gromov–Wasserstein barycenter problem, and that admits sharp statistical guarantees. Figure 1 sketches the overall pipeline: from an observed adjacency matrix, we recover a low-dimensional block model that summarises the latent generative mechanism. The srGW objective [18] and the conditional-gradient scheme [19] are established tools, and we use them as such. Existing srGW work targets matching, clustering, or dictionary learning [18, 20–22]. We instead cast SBM and graphon estimation as a srGW barycenter problem and prove consistency of the procedure.

Contributions In this work, we address the problem of estimating latent network structure from a single observed adjacency matrix, framing it as a semi-relaxed Gromov-Wasserstein (srGW) barycenter problem. Our theoretical analysis is developed for simple (binary undirected) graphs, though the method can be extended to handle (multiple) decorated graphs (i.e., with edge features), as shown in our real-data applications.

We define our estimate for the latent generating mechanism as the minimizer of a srGW objective between the observed adjacency matrix ($n \times n$) and a latent block model ($k \times k$), which discretizes the underlying truth; this can be interpreted as a form of nonparametric dimensionality reduction for network data. This formulation naturally yields an optimal transport plan π that defines a probabilistic alignment mapping observed nodes to latent blocks, while simultaneously inferring block weights. Because the srGW framework operates intrinsically over the space of probabilistic couplings, it bypasses the combinatorial intractability of deterministic node assignments, ensuring its computational tractability. A central challenge with probabilistic relaxations is ensuring the resulting soft assignments carry meaningful discrete interpretations. We provide a rigorous theoretical bridge for this and show that the optimal solutions of the semi-relaxed problem concentrate near deterministic assignments. Specifically, we bound the optimality gap between the relaxed solution and its corresponding discrete partition, showing it vanishes at a rate of $O(1/n)$. This result provides a formal justification for employing probabilistic relaxations as a tool to recover discrete assignments: the excess cost of discretizing the optimal coupling decays as the observed network grows, allowing

us to obtain *near-optimal* deterministic solutions (NP-hard to find) by projecting *optimal* probabilistic ones (efficiently computable).

These insights are realized through a scalable block-coordinate conditional gradient algorithm that recovers latent block structures and weights efficiently. Our implementation is designed for modern computational workflows: it is GPU-accelerated, scales to networks with thousands of nodes in seconds, and can handle multiple observed networks and graphs with edge attributes.

Our main contributions can be summarized as follows:

1. We cast SBM and graphon estimation as a semi-relaxed Gromov–Wasserstein barycenter problem, introducing a probabilistic relaxation of node assignments.
2. We prove that relaxed solutions concentrate near deterministic assignments, establishing a finite-sample bound on the optimality gap (Proposition 3.3). We show that our estimator is statistically consistent and achieves minimax-optimal rates for both SBMs and smooth graphons (Theorem 3.5), in both the dense and the sparse [23] regimes (Corollary 3.6).
3. We demonstrate a practical, GPU-accelerated block-coordinate solver (Algorithm 1, BCM). Our implementation scales efficiently with network size and demonstrates strong performance on both synthetic and real-world networks.

Overall, a key strength of the proposed framework – beyond its minimax-optimal guarantees in both the dense and sparse regimes – is its flexibility: it applies without modification to bounded edge-decorated graphs in the sense of [24]; we use this in our real-data examples (Section 4)

Related Work Inference of latent network structures has a rich history, particularly through stochastic block models and graphons, see for example [25, 23, 26–28]. The maximum likelihood estimator [29] is statistically optimal but comes with NP-hard computation. Spectral methods [30, 31] offer efficiency but are sensitive to degree heterogeneity. Our work sidesteps the direct labeling problem and leverages the notion of semi-relaxed Gromov-Wasserstein barycenters. The GW distance [11, 12] provides a metric between graphs that avoids explicit node-to-node correspondence. It has been used for graph alignment [13], graphon estimation [16], and extended to attributed networks [14, 32, 15]. The semi-relaxed GW divergence, introduced by Vincent-Cuaz et al. [18] relaxes the constraint on one marginal, and was successfully applied to graph clustering, dictionary learning [20] and to non-linear dimensionality reduction problems [21, 22]. In a related setting, Murray and Pickarski [33] showed that optimal semi-relaxed Gromov–Wasserstein plans are typically deterministic for inner-product–based costs. We generalize their proof strategy to network inference, where costs are induced by adjacency matrices, and establish finite-sample bounds on the optimality gap between relaxed (probabilistic) and constrained (deterministic) assignments; controlling this gap allows us to extend the approach of [8] and prove minimax-optimal rates for stochastic block models and smooth graphons. This distinguishes our work from Gromov–Wasserstein barycenter methods for graphs such as [16], which target graph *averaging*, require the barycenter size to be fixed in advance, and rely on entropic regularization that promotes diffuse transport plans – none of which target the recovery of a discrete latent generating mechanism. In contrast, our conditional-gradient solver, by adapting the barycenter size and yielding near-deterministic couplings, is precisely what allows for statistical guarantees on the latent block structure.

Structure of the paper. In Section 2, we introduce network models, including stochastic block models and graphons, and define the semi-relaxed GW objective. Section 3 presents our main theoretical results, including the relaxation strategy via srGW barycenters (Section 3.1); bounds on the optimality gap between relaxed and deterministic assignments (Section 3.2); finite-sample convergence rates for the proposed estimator (Section 3.3); implementation aspects by a conditional gradient algorithm (Section 3.4). Section 4 provides a concise illustration of the method's empirical behavior on simulated and real-world networks; additional experiments, comparisons and discussions are deferred to the appendix. Section A contains additional background material, B the proofs, and E extensive numerical study and comparison to existing methods.

2 Setting and background

Given a metric space Ω, we denote by $\Omega^{n\times n}$ the space of all graphs on n nodes represented by their adjacency matrices with edges in Ω. In the following we consider undirected graphs with no

self-loops so that any adjacency matrix $E \in \Omega^{n\times n}$ is symmetric with zeros on its diagonal, and the edges are binary indicators ($\Omega = \{0, 1\}$), but our method extends to broader settings: see Section 4.

To compare network structures while accounting for their inherent non-identifiability, we leverage the *Gromov-Wasserstein distance* (see Section A.2), which focuses on the relational structure *within each space*, rather than across. Formally, given two graphs on n, m nodes represented by their adjacency matrices $E \in \Omega^{n\times n}$, $F \in \Omega^{m\times m}$ respectively, their GW distance can be written as

$$\mathrm{GW}_2^2(E, F) = \inf_{\pi} \sum_{i,i'=1}^{n} \sum_{j,j'=1}^{m} |E_{i,i'} - F_{j,j'}|^2 \, \pi_{i,j} \, \pi_{i',j'},$$

where π is taken over the set of couplings whose left and right marginals are the discrete uniform distributions on $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. The GW distance thus provides a mean to compare graphs in a way that is invariant to permutations of the nodes and works for different numbers of nodes, making it well-suited for our estimation problem.

SBMs and graphons. A widely used generative model for networks is the *stochastic block model* [9]. In the SBM, each node i is assigned to one of k latent blocks or communities, encoded by a membership function $z : [n] \to [k]$. In this framework, the probability that an edge between nodes i and j is present depends only on their block memberships:

$$\mathbb{P}(E_{ij} = 1 \mid z(i), z(j)) = B_{z(i),z(j)},$$

where $B \in [0, 1]^{k\times k}$ is a symmetric matrix specifying the connection probabilities between blocks. The SBM can be viewed as a special case of a graphon [10], where the graphon W is piecewise constant over a partition of $[0, 1]^2$ into $k \times k$ blocks. That is, $W(x, y) = B_{a,b}$, for $x \in I_a, y \in I_b$, with $\{I_1, \ldots, I_k\}$ a partition of $[0, 1]$ determining the blocks. Under a more general graphon model generating mechanism, each node i in an observed graph $E \in \Omega^{n\times n}$ is associated with an unobserved latent feature $\xi_i \in [0, 1]$, drawn i.i.d., and edges are formed independently given the latent variables:

$$\mathbb{P}(E_{ij} = 1 \mid \xi_i, \xi_j) = W(\xi_i, \xi_j),$$

where $W : [0, 1]^2 \to [0, 1]$ is the underlying (unknown) graphon. This framework provides a nonparametric limit for dense graphs and captures a wide range of network structures beyond classical stochastic block models [34, 35].

3 Main results

We consider the problem of estimating the generating mechanism of a network from a single observed undirected graph. From an observed adjacency matrix E on n nodes, generated from an unknown graphon W, we estimate a block-connectivity matrix $\mathrm{B}(\hat{z}) \in [0, 1]^{k\times k}$ and weights $\nu(\hat{z}) \in \Delta^{k-1}$ parameterising a low-dimensional approximation of W. It is well-known that any smooth graphon can be approximated by a SBM on k nodes arbitrarily well, for k sufficiently large [25]. Therefore, a common strategy is to fit a block model on at most $k = k(n) \ll n$ blocks to the observed graph E, for k possibly diverging with n. Fitting an SBM with k blocks to an observed network on n nodes amounts to partitioning the nodes into the corresponding blocks, and then estimating the connectivity probabilities across blocks. A suitable partition can be found by optimizing over all assignment maps from n nodes to k blocks, which is an intrinsically difficult combinatorial problem. To tackle the issue, we propose a relaxation of the problem: instead of assigning each node to a single block, we map it to a probability distribution over the blocks. The advantage of the relaxation is that it yields a practical implementation method which remains tractable even for large graphs. Notably, we prove that the gap between the relaxed and combinatorial problems can be bounded for finite samples and that in the large network limit such a convex relaxation comes at virtually no cost. We will then show that this procedure yields an asymptotically consistent estimator: see Figure 1.

3.1 From hard to soft assignment

Hard assignment. Finding the generating mechanism of a SBM corresponds to estimating simultaneously two quantities: the block membership map $z : [n] \to [k]$ and the connection probability

matrix $B \in [0,1]^{k\times k}$ [8]. This can be naturally stated as a least squares problem:

$$\underset{\substack{z\in\mathcal{Z}_{n,k}\\ B\in[0,1]^{k\times k}}}{\arg\min}\ \frac{1}{n^2}\sum_{i,j=1}^{n}\|E_{i,j}-B_{z(i),z(j)}\|_2^2, \qquad \text{with} \quad \mathcal{Z}_{n,k}=\{z:[n]\to[k]\}. \tag{1}$$

For a fixed assignment map $z \in \mathcal{Z}_{n,k}$, the value of B minimizing (1) is given by:

$$\mathrm{B}(z)_{a,b} = \frac{\sum_{i,j=1}^{n} E_{i,j}\,\mathbb{1}_{\{z(i)=a,z(j)=b\}}}{\sum_{i,j=1}^{n}\mathbb{1}_{\{z(i)=a,z(j)=b\}}}.$$

Therefore, optimizing (1) over both memberships and block-matrices can be reduced to optimizing solely over memberships $\arg\min_{z\in\mathcal{Z}_{n,k}} \mathcal{L}(E,\mathrm{B}(z))$. This approach theoretically achieves an estimation error which vanishes like $O_p(n^{-2\alpha/(\alpha+1)}+\log n/n)$ under standard smoothness assumptions [8]. However, solving such an optimization problem is NP-hard in general [36, 27]: it requires searching over *all* possible assignments of n nodes to k blocks, a discrete set of size k^n which renders exhaustive search computationally intractable already for moderate n.

Probabilistic alignment: a tractable alternative. To overcome the intractability of searching over all deterministic labelings, we propose to relax the problem by considering *probabilistic labelings*. Instead of assigning each node i to a single block $z(i) \in [k]$, we associate each node with a probability distribution over the blocks. Formally, we define a probabilistic labeling as a matrix $\pi \in \mathbb{R}^{n\times k}$, where $\pi_{i,j}$ represents the probability that node i belongs to the j-th block. That is, π is in:

$$\Pi_{[n],k} = \left\{\pi\in\mathbb{R}_+^{n\times k} : \sum_{a=1}^{k}\pi_{i,a}=\frac{1}{n}, \quad \forall i\in[n]\right\}.$$

This can be interpreted as the set of couplings over $[n]\times[k]$ with left marginal fixed and equal to the uniform measure over $[n]$, while the right marginal is left free; this is the semi-relaxed GW formulation of Vincent-Cuaz et al. [18]. The relaxed optimization problem now reads:

$$\begin{aligned}(\hat{\pi},\mathrm{B}(\hat{z}))&\in \underset{\pi\in\Pi_{[n],k},B\in[0,1]^{k\times k}}{\arg\min}\ \mathcal{L}_E(\pi,B), \text{ where}\\ \mathcal{L}_E(\pi,B)&:=\sum_{i,j=1}^{n}\sum_{a,b=1}^{k}\|E_{i,j}-B_{a,b}\|_2^2\pi_{i,a}\pi_{j,b}.\end{aligned} \tag{2}$$

Observe that, as for the combinatorial problem, when π is fixed there is a unique probability matrix $\mathrm{B}(\pi)$ minimizing the objective (2) given by:

$$\mathrm{B}(\pi)_{a,b}=\sum_{i,j=1}^{n}\frac{E_{i,j}\pi_{i,a}\pi_{j,b}}{\nu(a)\nu(b)}, \qquad \text{with } \nu(a)=\sum_{i=1}^{n}\pi_{i,a}, \tag{3}$$

where ν encodes the relative block size associated to π. This allows us to rewrite Equation (2) as:

$$\hat{\pi}\in\underset{\pi\in\Pi_{[n],k}}{\arg\min}\ \mathcal{L}_E(\pi), \qquad \text{with} \quad \mathcal{L}_E(\pi)=\mathcal{L}_E(\pi,\mathrm{B}(\pi)). \tag{4}$$

This relaxation transforms the original hard combinatorial problem (1) into a continuous, easier one: the search space is now a convex set, which is significantly more tractable than the discrete set of deterministic labelings. However, probabilistic couplings are less easily interpretable, rendering their statistical analysis more challenging. Furthermore, the optimization landscape remains complex: in effect, we are burdened with the task of solving a GW objective, with cost function depending on the coupling, producing an intricate structure with no inherent simple optimization procedure.

In the optimization problem (4), we seek a barycenter structure that can be interpreted as a form of dimensionality reduction in the form of a lower-complexity block model that preserves the main relational structure of the graphs. The grouping of nodes into blocks enforces information sharing across nodes and allows for estimation, even if clustering is not our primary goal[1].

[1] see Gao et al. [8, section 3.6] for a discussion between clustering and parameter estimation

3.2 Optimal soft assignments are nearly deterministic

As observed by Murray and Pickarski [33], since the right marginal of the coupling is free, optimal π tend to concentrate on deterministic assignments: each node is typically mapped to a single block. We first study the optimal coupling for a fixed block matrix, then combine with the optimal B from (3) to obtain rates for the estimator.

Let $B \in [0,1]^{k\times k}$ be a fixed block matrix and consider the problem of finding the optimal coupling $\pi \in \Pi_{[n],k}$ for fixed block structure B:

$$\underset{\pi\in\Pi_{[n],k}}{\arg\min}\ \mathcal{L}_{E,B}, \qquad \text{where} \quad \mathcal{L}_{E,B} = \mathcal{L}_E(\pi, B) \tag{5}$$

Before targeting (4), a simpler question is to understand the nature of the optimal solutions of (5). We extend the marginal minimization result appearing in Murray and Pickarski [33, Thm. 10] and show that optimal solutions to (5) are typically not diffuse.

Lemma 3.1. *Fix $B \in [0,1]^{k\times k}$ and $E \in \Omega^{n\times n}$. If $\tilde{\pi}$ is a local-minima of $\mathcal{L}_{E,B}$. Then:*

$$\tilde{\pi}_{i,a} > 0 \Longrightarrow a \in \underset{\circ\in[k]}{\arg\min} \sum_{j=1}^{n}\sum_{b=1}^{k} \|E_{i,j} - B_{\circ,b}\|_2^2\, \tilde{\pi}_{j,b}, \qquad \textit{for all} \quad (i,a) \in [n]\times[k].$$

In words: at any local minimum, the blocks supporting positive mass for node i are those minimising the marginal cost given the assignments of all other nodes. Two consequences follow: (i) when this argmin is unique, $\tilde{\pi}$ *must* be deterministic; (ii) when ties occur, we can rebalance the mass at node i onto a single tied block without violating the first-order condition. Carefully handling all such ties yields a deterministic labeling whose excess cost is $O(1/n)$.

Definition 3.2 (Rounding map)**.** For $\pi \in \Pi_{[n],k}$, define the *rounding map* $\mathrm{z} : \Pi_{[n],k} \to \mathcal{Z}_{n,k}$ by

$$\mathrm{z}(\pi) \ = \ \left(\min\left\{ a \in \underset{a'\in[k]}{\arg\max}\ \pi_{i,a'} \right\},\ i \in [n] \right), \tag{6}$$

assigning each node to its heaviest block and breaking ties by smallest index.

Proposition 3.3. *Let $B \in [0,1]^{k\times k}$, and let $\tilde{\pi}$ be a local minimum for* (5)*. Then, there exists a deterministic labeling $\tilde{z} = \mathrm{z}(\tilde{\pi})$ such that the optimality gap satisfies:*

$$0 \le \mathcal{L}_{E,B}(\tilde{z}) - \mathcal{L}_{E,B}(\tilde{\pi}) \le n^{-1}\|B\|_\infty^2.$$

The bound is independent of k: the gap remains $O(1/n)$ whether k is fixed or grows with n, which is what enables the graphon rate where $k \asymp n^{1/(\alpha\wedge 1+1)}$.

Remark 3.4 (From local minima to Frank-Wolfe stationarity). The Frank-Wolfe (FW) gap at a coupling $\tilde{\pi} \in \Pi_{[n],k}$ vanishes exactly when $\tilde{\pi}$ itself minimises the linear functional $\pi \mapsto \langle \nabla\mathcal{L}_{E,B}(\tilde{\pi}),\, \pi\rangle$ over $\Pi_{[n],k}$. By the product structure of $\Pi_{[n],k}$, this minimisation decouples across nodes and assigns each node to a block minimising $\mathcal{J}_{\tilde{\pi}}(i,\cdot)$, so $\mathrm{gap}(\tilde{\pi}) = 0$ is exactly the per-node argmin condition $\tilde{\pi}_{i,a} > 0 \Rightarrow a \in \arg\min_\circ \mathcal{J}_{\tilde{\pi}}(i,\circ)$ of Lemma 3.1. Inspection of the proofs (Section B) shows that Lemma 3.1 and Proposition 3.3 use only this condition; the conclusions therefore apply to any FW-stationary coupling, including the output of the conditional-gradient solver of Section 3.4 at termination, and – via Theorem 3.5 and Corollary 3.6 – to the resulting estimate $(\mathrm{B}(\hat{z}), \nu(\hat{z}))$. Whether Algorithm 1 further converges to a global minimum remains open [18].

3.3 Network learning with srGW barycenters

To approximate W from E we fit a block model (B,ν) with adaptive k to E. Rather than solving the NP-hard (1) directly, we solve the relaxed (4) and exploit Proposition 3.3 to extract a near-optimal deterministic assignment via the following three-step procedure:

1. **Solve the relaxed problem:** $\hat{\pi} \in \arg\min\{\mathcal{L}_E(\pi) :\ \pi \in \Pi_{[n],k}\}$ (via Algorithm 1).
2. **Extract a deterministic assignment:** $\hat{z} = \mathrm{z}(\hat{\pi})$ by Eq. (6).
3. **Compute the block estimate:** $(\hat{B}, \hat{\nu}) = \big(\mathrm{B}(\hat{z}),\ \nu(\hat{z})\big)$ via Eq. (3).

The procedure described above enjoys the following optimal rates of convergence.

Theorem 3.5. *Let $E \in \Omega^{n\times n}$. **(i) SBM.** Suppose that E is generated from a stochastic block model with parameters (B^*, ν^*) and true block assignment z^*. Let $(\mathrm{B}(\hat{z}), \nu(\hat{z}))$ be as above. For any $C' > 0$ there exists universal $C > 0$ such that with probability at least $1 - \exp(-C' n \log k)$:*

$$\frac{1}{n^2}\sum_{i,j}\left\|B^*_{z^*(i),z^*(j)} - \mathrm{B}(\hat{z})_{\hat{z}(i),\hat{z}(j)}\right\|^2 \le C\left(\frac{k^2}{n^2} + \frac{\log k}{n}\right).$$

(ii) Graphon. *Let W be an α-Hölder-smooth graphon and E be generated from W, with $\theta^* = \mathbb{E}[E]$. Setting $k = \lceil n^{1/(\alpha\wedge 1+1)}\rceil$, for any $C' > 0$ there exists universal $C > 0$ such that with probability at least $1 - \exp(-C' n \log n)$,*

$$\frac{1}{n^2}\sum_{i,j}\left\|\theta^*_{ij} - \mathrm{B}(\hat{z})_{\hat{z}(i),\hat{z}(j)}\right\|^2 \le C\left(n^{-\frac{2\alpha}{\alpha+1}} + \frac{\log n}{n}\right).$$

Both rates are minimax-optimal [8, 23]. The rates are usually attained by the (NP-hard) least-squares estimator over discrete assignments, whereas Theorem 3.5 obtains them from the minimiser of a different, continuously relaxed objective – the srGW barycenter loss in (4). The clustering rate $\log k/n$ in (i) accounts for the unknown node structure; in (ii) the rate $n^{-2\alpha/(1+\alpha)}$ is the standard nonparametric rate for α-Hölder functions, and $\log n/n$ dominates when $\alpha \ge 1$. Unsurprisingly, since the optimality gap in Proposition 3.3 depends on $\|B\|^2_\infty$, we can refine the rates under sparsity.

Corollary 3.6 (Sparse regime)**.** *Assume the true parameter scales as $\theta^*_{ij} = \rho_n W^*(\xi_i, \xi_j)$ with $W^* : [0,1]^2 \to [0,1]$ and $\rho_n \in (0,1]$ dictates the sparsity level. Assuming that all blocks contain a non-vanishing fraction of the nodes, $(\mathrm{B}(\hat{z}), \nu(\hat{z}))$ match the minimax rates established by Klopp et al. [23] up to an additive gap of order $\|\mathrm{B}(\hat{\pi})\|^2_\infty n^{-1}$.*

Comparison with prior work. Our argument adapts the marginal-minimisation template of Murray and Pickarski [33]. Two refinements are needed: (a) Lemma 3.1 and Proposition 3.3 hold for *local* minima, since the conditional-gradient solver only certifies stationary points [19]; (b) a direct adaptation yields an $O(k/n)$ gap, insufficient when $k \asymp n^{1/(\alpha\wedge 1+1)}$, so Proposition 3.3 sharpens it to $O(\|B\|^2_\infty/n)$ independent of k via a refined ties argument. In the sparse case, the additive gap $\|\mathrm{B}(\hat{\pi})\|^2_\infty$ could be further controlled, but we leave this direction to future work. We also discuss in Section E.1 the relation with structured GW-barycenter methods [16]: while these methods address graph averaging via entropically regularised barycenters, our goal is recovery of the latent graphon structure through near-deterministic low-dimensional couplings, making the two approaches related but not directly comparable. See Section E.1 for details.

3.4 Block-coordinate descent algorithm

We now describe an efficient implementation for solving Eq. (4). Since the objective jointly optimizes over the block matrix and the assignment map, we adopt an alternating scheme that iteratively updates each component [18]. For a fixed labeling, the block matrix B admits a closed-form solution (3), so no iterative optimization is required. The algorithm therefore alternates between a Gromov-Wasserstein transport update for π, and a closed-form barycenter update for B. We give pseudo-code in Algorithm 1. The inner step relies on the Frank-Wolfe algorithm, which converges to stationary points at rate $O(1/\sqrt{T})$ in the primal-dual gap for non-convex objectives [19] and is sketched in Algorithm 2. The dominant cost is the conditional-gradient inner loop, which we parallelise on GPU through the `POT` library [37]; the closed-form $\tilde{B}$-update is $O(nk)$ and runs on either device.

Initialisation and choice of k. Since (4) is non-convex, $\pi^{(0)}$ matters: Section E compares three strategies (k-means on the spectral embedding, spectral clustering, product/uniform), with k-means most stable on graphons and product fastest on SBMs (Figures 7 to 9). We use k-means by default, since it is most stable across graphon settings. The graphon rate uses oracle $k_{\mathrm{opt}} = \lceil n^{1/(\alpha\wedge 1+1)}\rceil$ depending on the unknown smoothness α. We use the following self-pruning scheme: pick a conservative upper bound $k_{\max}$, run Algorithm 1 with $k = k_{\max}$, and report the effective block count $\widehat{k} = |\{a : \widehat{\nu}(a) > 0\}|$ from the pruned right marginal. The free right marginal in (4) acts as implicit

Algorithm 1 Block-Coordinate Minimization (BCM) for srGWB

Input: Graph $E \in \{0,1\}^{n\times n}$; number of blocks k
Initialize $\pi^{(0)} \in \Pi_{[n],k}$
for $t = 0, 1, 2, \ldots, T-1$ **do**
  $B_{a,b}^{(t+1)} = \frac{\sum_{i,j=1}^{n} E_{i,j}\, \pi_{i,a}^{(t)}\, \pi_{j,b}^{(t)}}{\sum_{i,j=1}^{n} \pi_{i,a}^{(t)}\, \pi_{j,b}^{(t)}}$, **(Update B)**
  $\pi^{(t+1)} = \underset{\pi\in\Pi_{[n],k}}{\operatorname{argmin}} \sum_{i,j=1}^{n} \sum_{a,b=1}^{k} \left(E_{i,j} - B_{a,b}^{(t+1)}\right)^2 \pi_{i,a}\, \pi_{j,b}$, **(Optimize π:** Algorithm 2 **)**
end for; **Output:** π^T

complexity regularisation: blocks that do not improve the srGW objective receive zero mass and drop out, so $\widehat{k} \le k_{\max}$ adapts to the data without an explicit penalty. Across our 10 synthetic settings, $k_{\max} = \lceil \sqrt{n} \rceil$ (graphons) and $k_{\max} = 20$ (SBMs) are sufficient: $\widehat{k}$ tracks the effective block count, and the GW estimation error is essentially flat in $k_{\max}$ once $k_{\max} \ge 2k_{\text{opt}}$ (Figures 7 and 8).

Across all 10 synthetic settings, the BCM solver returns near deterministic couplings at convergence (the fraction of nodes carrying mass on more than one block is almost zero in every run; Section E). Lemma 3.1 and Proposition 3.3 guarantee that any local minimum lies within $O(1/n)$ of a deterministic assignment in cost. Still, these results do not fully explain why gradient-based methods typically converge to deterministic solutions – which can arguably be seen as a form of implicit regularization, with gradient descent converging to more structured solutions.

4 Data analysis

Synthetic experiments use binary graphs; real-data examples use weighted networks fitted via the expected adjacency matrix [38] (the method extends naturally to decorated graphons [24]).

Synthetic experiments. We consider 6 graphons and 4 SBMs, running 20 Monte Carlo replicates per (model, initialisation, n). Figure 2 shows the decaying error rate for $n \in \{100, \ldots, 10'000\}$ on a periodic graphon – well beyond the $n \lesssim 10^3$ scale at which graphon-estimation methods are typically benchmarked [28, 39, 26, 16] – matching the theoretical rate of Section 3.3; analyses for the remaining models are in Section E.1. Code is available at this repository [40].

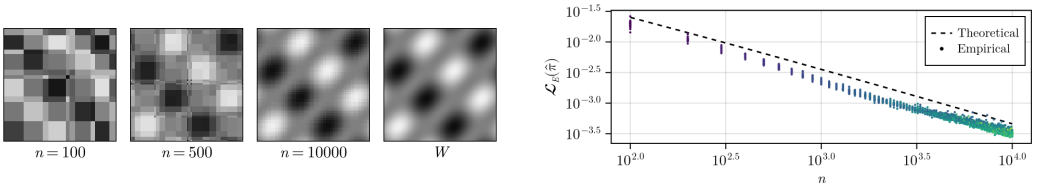


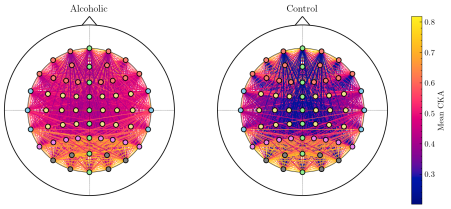
Figure 2: Left: estimator for increasing network size in the 'G4' example. Right: empirical loss, matching theoretical rates. Obtained from 20 Monte Carlo runs with clustering-based initialization. Similar and additional plots for the other considered models are found in Section E.1.

Table 1: GW distance to truth ($\times 10^{-3}$) at $n = 2000$, mean$_{\pm\text{std}}$ over 20 runs. **Bold**: best per setting; *italic*: second best. **Ours** uses the best-performing initialization per setting, and perform very well across all settings. Additional details can be found in Section E, including timing results.

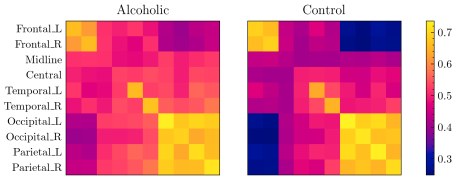
	SBMs				Graphons					
Method	S1	S2	S3	S4	G1	G2	G3	G4	G5	G6
LG	$88.4_{\pm 2.9}$	$88.6_{\pm 2.6}$	$71.9_{\pm 17}$	$40.7_{\pm 14}$	$2.8_{\pm 0.66}$	$2.2_{\pm 0.39}$	$1.3_{\pm 0.12}$	$42.6_{\pm 1.1}$	$42.8_{\pm 1.2}$	$55.7_{\pm 9.8}$
NBDsmooth	$0.82_{\pm 0.007}$	$0.82_{\pm 0.007}$	$0.82_{\pm 0.004}$	$0.81_{\pm 0.010}$	$0.70_{\pm 0.012}$	$1.0_{\pm 0.009}$	$1.2_{\pm 0.011}$	$\mathit{0.88}_{\pm 0.007}$	$\mathit{0.91}_{\pm 0.010}$	$0.93_{\pm 0.015}$
SAS	$81.3_{\pm 11}$	$82.2_{\pm 9.8}$	$0.97_{\pm 0.19}$	$5.8_{\pm 3.5}$	$\mathit{0.46}_{\pm 0.011}$	$\mathit{0.62}_{\pm 0.009}$	$1.0_{\pm 0.019}$	$42.0_{\pm 1.1}$	$41.9_{\pm 1.1}$	$54.3_{\pm 11}$
USVT	$\mathit{0.32}_{\pm 0.007}$	$\mathit{0.32}_{\pm 0.007}$	$\mathit{0.32}_{\pm 0.004}$	$\mathit{0.45}_{\pm 0.010}$	$\mathbf{0.18}_{\pm 0.005}$	$1.7_{\pm 0.07}$	$\mathit{0.73}_{\pm 0.016}$	$1.3_{\pm 0.018}$	$2.7_{\pm 0.59}$	$\mathbf{0.34}_{\pm 0.008}$
Ours	$\mathbf{0.013}_{\pm 0.0003}$	$\mathbf{0.31}_{\pm 0.0003}$	$\mathbf{0.31}_{\pm 0.0004}$	$\mathbf{0.24}_{\pm 0.001}$	$0.51_{\pm 0.04}$	$\mathbf{0.61}_{\pm 0.05}$	$\mathbf{0.63}_{\pm 0.08}$	$\mathbf{0.73}_{\pm 0.03}$	$\mathbf{0.76}_{\pm 0.03}$	$\mathit{0.60}_{\pm 0.009}$

Baseline comparison. Table 1 benchmarks our estimator at $n = 2000$ against the standard SBM/graphon baselines with mature implementations that target the same recovery loss and scale to $n = 2000$: Largest-Gap (LG, [41]), Neighbourhood Smoothing (NBDsmooth, [39]), Sort-and-Smooth (SAS, [28]), and USVT [26]. Our method delivers the strongest overall performance: it attains the smallest GW distance on the SBM settings (often by one to two orders of magnitude over every competitor), while ranking competitively on the smooth graphons and being computationally efficient (see Fig. 10). Crucially, ours is the only estimator that performs uniformly well across both graphons *and* block models, while being interpretable (in opposition to USVT): each baseline collapses on at least one regime. The histogram estimator [25]– which subsumes spectral-clustering with local refinement – and the variational MLE `sparseBM` [42] did not terminate within our computational budget; restricted comparisons for $n \leq 300$ are in Figure 11, and our method is order of magnitude faster. The real-data examples lack ground truth, so we evaluate qualitatively against domain priors (Section E).

EEG dataset. We apply our method to recordings from 64 electrodes on 20 subjects (10 alcoholic, 10 control) sampled at 256 Hz over 1 second under three stimulus paradigms [43]. Edge weights are Centered Kernel Alignment (CKA) between replicated electrode time series, with a linear kernel [44]; we apply srGWB with $k = 40$ blocks. Figure 3b displays the fitted co-activation structures and their mapping to brain regions; Figure 3a shows the alcoholic-vs-control difference, with increased co-activation in alcoholic subjects between frontal and parietal/occipital regions, a pattern commonly interpreted as a compensatory mechanism [45, 46]. See Section E for further analysis.



(a) Inferred latent connectivity in alcoholic (left) and control (right).



(b) Mapped co-activation structures in alcoholic (left) and control (right).

Figure 3: EEG dataset: inferred block connectivity (left) and co-activation structures (right) for the first stimulus condition.

OpenFlight Airport Data. We analyse the OpenFlight airport network [47], with edges encoding the existence of direct flights, partitioning 3257 airports into 23 clusters. Figure 4 shows the inferred block-level connectivity; the full geographically-laid-out clustering is in Section E and reveals strong spatial coherence and major hubs.

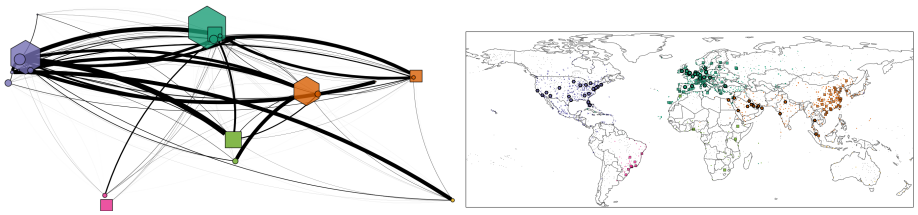


Figure 4: OpenFlight network: inferred block-level connectivity. Geographic map in Section E.

5 Discussion

We introduced a tractable framework for latent network inference via semi-relaxed Gromov–Wasserstein barycenters. Relaxing hard assignments to probabilistic couplings turns SBM/graphon estimation into a continuous optimization that avoids NP-hard combinatorial search, with a regime-agnostic optimality gap of order $\|B\|_\infty^2/n$. This yields minimax-optimal rates for SBMs and Hölder graphons in both dense [8] and sparse [23] regimes, via a scalable conditional gradient algorithm. Scope and natural extensions are discussed in Section 3.4. An important and

promising extension would be to incorporate in the method and its analysis a penalty for the number of blocks selected by the algorithm. Such an extension would offer a novel framework for adaptively selecting the preferred complexity by which to summarize network data. However, the addition of a sparsity-promoting regularization to the semi-relaxed Gromov-Wasserstein problem (4) significantly modifies the optimization landscape and requires a separate, careful analysis. Finally, while the numerical examples are mainly illustrative, we see the proposed framework as an important step toward more general nonparametric methods for network data, for example, for sieve bootstrap or data augmentation in networks with edge and node attributes.

References

[1] Amanda L Traud, Peter J Mucha, and Mason A Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

[2] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[3] Monica Agrawal, Marinka Zitnik, and Jure Leskovec. Large-scale analysis of disease pathways in the human interactome. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 23, page 111, 2018.

[4] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[5] Ziqi Gao, Chenran Jiang, Jiawen Zhang, Xiaosen Jiang, Lanqing Li, Peilin Zhao, Huanming Yang, Yong Huang, and Jia Li. Hierarchical graph learning for protein–protein interaction. *Nature Communications*, 14(1):1093, 2023.

[6] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 974–983, 2018.

[7] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[8] Chao Gao, Yu Lu, and Harrison H. Zhou. Rate-Optimal Graphon Estimation. *The Annals of Statistics*, 43(6):2624–2652, 2015. ISSN 0090-5364. doi: 10.1214/15-AOS1354.

[9] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. doi: 10.1016/0378-8733(83)90021-7.

[10] László Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012. ISBN 0-8218-9085-9.

[11] Facundo Mémoli. Gromov–Wasserstein Distances and the Metric Approach to Object Matching. *Foundations of Computational Mathematics*, 11(4):417–487, August 2011. ISSN 1615-3383. doi: 10.1007/s10208-011-9093-5.

[12] Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 2664–2672. PMLR, June 2016.

[13] Hongteng Xu, Dixin Luo, Hongyuan Zha, and Lawrence Carin Duke. Gromov-Wasserstein Learning for Graph Matching and Node Embedding. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6932–6941. PMLR, May 2019.

[14] Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Fused Gromov-Wasserstein Distance for Structured Objects. *Algorithms*, 13(9):212, August 2020. ISSN 1999-4893. doi: 10.3390/a13090212.

[15] Martin Bauer, Facundo Mémoli, Tom Needham, and Mao Nishino. The Z-Gromov-Wasserstein Distance, January 2025.

[16] Hongteng Xu, Dixin Luo, Lawrence Carin, and Hongyuan Zha. Learning Graphons via Structured Gromov-Wasserstein Barycenters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10505–10513, May 2021. ISSN 2374-3468. doi: 10.1609/aaai.v35i12.17257.

[17] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. G-Mixup: Graph Data Augmentation for Graph Classification. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8230–8248. PMLR, June 2022.

[18] Cédric Vincent-Cuaz, Rémi Flamary, Marco Corneli, Titouan Vayer, and Nicolas Courty. Semi-relaxed Gromov Wasserstein divergence with applications on graphs. In *ICLR 2022 - 10th International Conference on Learning Representations*, pages 1–28, Virtual, France, April 2022.

[19] Simon Lacoste-Julien. Convergence Rate of Frank-Wolfe for Non-Convex Objectives, July 2016.

[20] Cédric Vincent-Cuaz, Titouan Vayer, Rémi Flamary, Marco Corneli, and Nicolas Courty. Online graph dictionary learning. In *International conference on machine learning*, pages 10564–10574. PMLR, 2021.

[21] Ranthony A Clark, Tom Needham, and Thomas Weighill. Generalized dimension reduction using semi-relaxed Gromov-Wasserstein distance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16082–16090, 2025. ISBN 2374-3468.

[22] Hugues Van Assel, Cédric Vincent-Cuaz, Nicolas Courty, Rémi Flamary, Pascal Frossard, and Titouan Vayer. Distributional reduction: Unifying dimensionality reduction and clustering with gromov-wasserstein. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

[23] Olga Klopp, Alexandre B. Tsybakov, and Nicolas Verzelen. Oracle inequalities for network models and sparse graphon estimation. *The Annals of Statistics*, 45(1):316–354, February 2017. ISSN 0090-5364, 2168-8966. doi: 10.1214/16-AOS1454.

[24] Charles Dufour and Sofia C. Olhede. Graphon estimation beyond binary edges: inference for decorated graphs with applications to multiplex and weighted networks, 2026. URL `http://arxiv.org/abs/2408.12339`.

[25] Sofia C. Olhede and Patrick J. Wolfe. Network histograms and universality of blockmodel approximation. *Proceedings of the National Academy of Sciences*, 111(41):14722–14727, October 2014. doi: 10.1073/pnas.1400374111.

[26] Sourav Chatterjee. Matrix estimation by Universal Singular Value Thresholding. *The Annals of Statistics*, 43(1):177–214, February 2015. ISSN 0090-5364, 2168-8966. doi: 10.1214/14-AOS1272.

[27] Arash A. Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *The Annals of Statistics*, 46(1):149–179, February 2018. ISSN 0090-5364, 2168-8966. doi: 10.1214/17-AOS1545.

[28] Stanley Chan and Edoardo Airoldi. A Consistent Histogram Estimator for Exchangeable Graph Models. In *Proceedings of the 31st International Conference on Machine Learning*, pages 208–216. PMLR, January 2014.

[29] Alain Celisse, Jean-Jacques Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6: 1847–1899, 2012. doi: 10.1214/12-EJS729.

[30] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011. doi: 10.1214/11-AOS887.

[31] Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015. doi: 10.1214/14-AOS1274.

[32] Vayer Titouan, Nicolas Courty, Romain Tavenard, Chapel Laetitia, and Rémi Flamary. Optimal Transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6275–6284. PMLR, May 2019.

[33] Ryan Murray and Adam Pickarski. On probabilistic embeddings in optimal dimension reduction. *Journal of Machine Learning Research*, 26(234):1–39, 2025.

[34] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and Newman–Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50): 21068–21073, 2009. doi: 10.1073/pnas.0907096106.

[35] Peter Orbanz and Daniel M. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2): 437–461, February 2015. ISSN 1939-3539. doi: 10.1109/TPAMI.2014.2334607.

[36] Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2016. doi: 10.1109/TIT.2015.2490670.

[37] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. ISSN 1533-7928.

[38] Etienne Donier-Meroz, Arnak S. Dalalyan, Francis Kramarz, Philippe Choné, and Xavier D'Haultfoeuille. Graphon Estimation in bipartite graphs with observable edge labels and unobservable node labels, September 2023.

[39] Yuan Zhang, Elizaveta Levina, and Ji Zhu. Estimating network edge probabilities by neighbourhood smoothing. *Biometrika*, 104(4):771–783, December 2017. ISSN 0006-3444. doi: 10.1093/biomet/asx042.

[40] Charles Dufour, Ulysse Naepels, and Leonardo Santoro. dufourc1/srgw_graphon_submission: v0.1, 2026. URL `https://doi.org/10.5281/zenodo.23687985`.

[41] Antoine Channarond, Jean-Jacques Daudin, and Stéphane Robin. Classification and estimation in the Stochastic Blockmodel based on the empirical degrees. *Electronic Journal of Statistics*, 6 (none):2574–2601, January 2012. ISSN 1935-7524, 1935-7524. doi: 10.1214/12-EJS753.

[42] Gabriel Frisch, Jean-Benoist Leger, and Yves Grandvalet. SparseBM: A Python Module for Handling Sparse Graphs with Block Models. 2022.

[43] Henri Begleiter. EEG Database, 1995.

[44] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, May 2019.

[45] Sandra Chanraud, Anne-Lise Pitel, Eva M Müller-Oehring, Adolf Pfefferbaum, and Edith V Sullivan. Remapping the brain to compensate for impairment in recovering alcoholics. *Cerebral cortex*, 23(1):97–104, 2013.

[46] Rui Cao, Zheng Wu, Haifang Li, Jie Xiang, and Junjie Chen. Disturbed connectivity of eeg functional networks in alcoholism: A graph-theoretic analysis. *Bio-Medical Materials and Engineering*, 24(6):2927–2936, 2014. doi: 10.3233/BME-141112.

[47] OpenFlights. Openflights airports database. https://openflights.org/data.html, 2017. Accessed: 2026-01-14.

[48] Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, August 2012. ISSN 0090-5364, 2168-8966. doi: 10.1214/12-AOS1036.

[49] Svante Janson. Graphons, cut norm and distance, couplings and rearrangements, June 2011.

[50] László Lovász and Balázs Szegedy. Limits of compact decorated graphs, October 2010.

[51] Romain Abraham, Jean-François Delmas, and Julien Weibel. Probability-graphons: Limits of large dense weighted graphs, December 2023.

[52] Charles Dufour and Sofia C. Olhede. Decorated graphons for temporal network estimation, 2026. URL `http://arxiv.org/abs/2607.23145`.

[53] Martin Jaggi. Revisiting Frank-Wolfe: Projection-Free Sparse Convex Optimization. In *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435. PMLR, February 2013.

A Background

A.1 Generative models for networks

Generative models for networks provide a probabilistic framework for describing graph-valued data. In this section we briefly review classical models relevant to our setting: namely, stochastic block models, graphons, and their extensions. We refer the readers to Orbanz and Roy [35] for details.

Stochastic block models (SBMs). The stochastic block model assumes that each node belongs to one of a finite number of latent groups, and that the probability of an edge between two nodes depends only on their group memberships [9]. Formally, let $K \in \mathbb{N}$ denote the number of blocks, let $z_i \in \{1, \dots, K\}$ be the latent block assignment of node i, and let $B \in [0,1]^{K \times K}$ be a symmetric matrix of connection probabilities. Conditional on the assignments $(z_i)_{i=1}^n$, edges are generated independently according to

$$\mathbb{P}(E_{ij} = 1 \mid z_i, z_j) = B_{z_i, z_j}, \qquad 1 \le i < j \le n.$$

SBMs capture community structures and provide a tractable model for clustering and inference. They admit many extensions, including degree-corrected SBMs, mixed-membership SBMs, and dynamic or hierarchical variants (see e.g., [48]). Importantly, observe that SBMs correspond to graphons that are piecewise constant with respect to a finite partition of $[0,1]$. As such, they can be viewed as finite-dimensional approximations to more general graphon models.

Graphons. Graphons provide a nonparametric framework for modeling large graphs and serve as limit objects for sequences of graphs with growing size [10]. Formally, a graphon is defined as a measurable function $W : [0,1]^2 \to [0,1]$. Given W, a random graph with n nodes can be generated by sampling latent variables $u_1, \dots, u_n \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]$ and then, conditionally on these variables, drawing edges independently as

$$\mathbb{P}(E_{ij} = 1 \mid u_i, u_j) = W(u_i, u_j), \qquad 1 \le i < j \le n.$$

Graphons are identifiable only up to measure-preserving transformations of the unit interval. That is, if $\phi : [0,1] \to [0,1]$ is measure-preserving, then W and $W^\phi(u,v) = W(\phi(u), \phi(v))$ define the same distribution over graphs [49]. This inherent non-identifiability motivates the use of comparison tools, such as the cut distance or Gromov–Wasserstein distance, that are invariant under such transformations. A useful way to interpret graphons is as infinite-dimensional analogues of SBMs, where the discrete block assignments are replaced by continuous latent positions. Many statistical properties of graphons depend on regularity assumptions.

A graphon W is said to be (α, L)-Hölder continuous for some $\alpha \in (0,1]$ and $L > 0$ if

$$|W(x,y) - W(x',y')| \le L\big(|x - x'| + |y - y'|\big)^\alpha \quad \text{for all } (x,y),(x',y') \in [0,1]^2.$$

Continuity and Hölder smoothness captures the idea that nodes with close latent positions should have somewhat close connectivity patterns. This assumption enables approximation of smooth graphons by block models with a controlled number of blocks and underlies convergence rates for graphon estimation [23]. In our context, smoothness assumptions help relate finite-sample graph observations to their underlying graphon representations.

Decorated graphs. Many real-world networks carry additional information beyond binary connectivity, such as node features, and edge attributes or weights. Decorated and probability graphons extend classical graphons by incorporating this supplementary information [50, 51]. Formally, we consider a node feature space $\mathcal{X}$ and an edge space Ω and, for each node i and j, sample $x_i \in \mathcal{X}$ and $E_{ij} \in \Omega$. The distributions of the node features and edges are still independent conditionally on the latent variables; for example, a decorated graphon model may specify a joint distribution

$$(u_i, x_i) \sim \mu, \qquad E_{ij} \mid u_i, u_j \sim W(u_i, u_j),$$

where the node attributes (x_i) are informative about the latent positions (u_i) and $W(u_i, u_j)$ is a probability measure supported on Ω. It is possible to define a distance for general networks encompassing all features previously described [15], and our results extend naturally to the case where we do not consider node features and Ω is compact and bounded as described in [24].

A.2 Gromov–Wasserstein distance

Gromov–Wasserstein distance. Define a metric measure space as a triple (X, d_X, μ_X), where (X, d_X) is a metric space and μ_X is a probability measure on X. Given two metric measure spaces (X, d_X, μ_X) and (Y, d_Y, μ_Y), the squared 2-Gromov–Wasserstein distance is defined as

$$\mathrm{GW}_2^2(\mu_X, \mu_Y) = \inf_{\pi \in \Pi(\mu_X, \mu_Y)} \iint_{X^2 \times Y^2} \left| d_X(x, x') - d_Y(y, y') \right|^2 \mathrm{d}\pi(x, y)\, \mathrm{d}\pi(x', y'),$$

where $\Pi(\mu_X, \mu_Y)$ denotes the set of couplings (joint probability measures) with marginals μ_X and μ_Y.

Unlike classical optimal transport, the GW distance does not require the spaces X and Y to be embedded in a common ambient space. Instead, it aligns points so as to best preserve pairwise distances, making it invariant to isometries and, more generally, to measure-preserving re-labelings. We refer the reader to Mémoli [11] for details.

GW distance between graphs. In the context of graphs, nodes are typically equipped with the uniform probability measure, and the metric is induced by the adjacency matrix or edge weights. Let $E \in \Omega^{n \times n}$ and $F \in \Omega^{m \times m}$ denote the adjacency matrices of two (possibly weighted) graphs. The GW distance between them can be written as

$$\mathrm{GW}_2^2(E, F) = \min_{\pi \in \Pi([n], [m])} \sum_{i,i'=1}^{n} \sum_{j,j'=1}^{m} \left| E_{i,i'} - F_{j,j'} \right|^2 \pi_{i,j}\, \pi_{i',j'},$$

where $\Pi([n], [m])$ denotes the set of couplings whose left and right marginals are the discrete uniform distributions on $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$, respectively. This formulation makes explicit that GW compares graphs up to node permutations, and relates GW to the Mean Squared Error (MSE) used in Gao et al. [8]. See Peyré et al. [12], Xu et al. [13], Vayer et al. [14], Bauer et al. [15] for details.

Semi-relaxed Gromov–Wasserstein distance. In some cases, it may be desirable to relax one of the marginal constraints. This leads to the *semi-relaxed Gromov–Wasserstein* distance [18]. Given again metric measure spaces (X, d_X, μ_X) and (Y, d_Y, μ_Y), the semi-relaxed GW problem is given by relaxing the marginal constraint on Y:

$$\mathrm{srGW}_2^2(\mu_X, \mu_Y) = \inf_{\substack{\pi \geq 0 \\ (\pi)_X = \mu_X}} \iint_{X^2 \times Y^2} \left| d_X(x, x') - d_Y(y, y') \right|^2 \mathrm{d}\pi(x, y)\, \mathrm{d}\pi(x', y'),$$

where $(\pi)_X$ denotes the marginal of π on X. In the discrete graph setting, this translates as: $E \in \Omega^{n \times n}$ and $F \in \Omega^{m \times m}$, the semi-relaxed GW distance takes the form

$$\mathrm{srGW}_2^2(E, F) = \min_{\substack{\pi \in \mathbb{R}_+^{n \times m} \\ \sum_{j=1}^m \pi_{i,j} = \frac{1}{n}}} \sum_{i,i'=1}^{n} \sum_{j,j'=1}^{m} \left| E_{i,i'} - F_{j,j'} \right|^2 \pi_{i,j}\, \pi_{i',j'}.$$

The semi-relaxed formulation retains invariance to relabeling of the source graph while allowing greater flexibility in matching to the target. This relaxation is particularly useful when comparing finite graphs to graphons, or when the target space represents a reference structure rather than a fixed empirical distribution.

Semi-relaxed Gromov–Wasserstein barycenters. Let a collection of metric measure spaces $(X_\ell, d_\ell, \mu_\ell)_{\ell=1}^L$ and weights $(\lambda_\ell)_{\ell=1}^L$, $L \geq 1$, such that $\lambda_\ell \geq 0$ and $\sum_{\ell=1}^L \lambda_\ell = 1$. A srGW barycenter is defined as a space (Y, d_Y, μ_Y) minimizing the weighted sum of semi-relaxed GW discrepancies

$$\sum_{\ell=1}^{L} \lambda_\ell \, \mathrm{srGW}_2^2(\mu_\ell, \mu_Y).$$

In contrast to classical GW barycenters, the semi-relaxed formulation fixes the marginal on each space while allowing flexibility in the barycenter measure, which is particularly advantageous when aggregating graphs of varying sizes or when estimating a population-level structure [18]. In the graph or graphon setting, srGW barycenters can be interpreted as representative structures that best preserve the relational patterns of the input graphs up to measure-preserving transformations. This makes srGW barycenters well suited for tasks such as averaging networks, template learning, and estimating graphons.

B Proofs

Proof of Lemma 3.1. We adapt the ideas introduced in [33] to our context. Let $\tilde{\pi}$ be a local-minima of $\mathcal{L}_{E,B}$ Denote

$$\mathcal{J}_{\tilde{\pi}} : (i, a) \in [n] \times [k] \mapsto \sum_{j=1}^{n} \sum_{b=1}^{k} \|E_{i,j} - B_{a,b}\|_2^2 \, \tilde{\pi}_{j,b}.$$

Fix $i \in [n]$ and $a \in [k]$ and assume that $\tilde{\pi}_{i,a} > 0$. By contradiction, suppose $a \notin \arg\min_{\circ \in [k]} \mathcal{J}_{\tilde{\pi}}(i, \circ)$. Let $a' \in \arg\min_{\circ \in [k]} \mathcal{J}_{\tilde{\pi}}(i, \circ)$ and define the perturbation $\psi = \frac{\epsilon}{n} \tilde{\pi}_{i,a} \, (\delta_{i,a'} - \delta_{i,a})$ for some $\epsilon > 0$. Then, observe that:

$$\begin{aligned}\mathcal{L}_{E,B}(\tilde{\pi} + \psi) - \mathcal{L}_{E,B}(\tilde{\pi}) = 2 \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \pi_{j,d} \psi_{j',d'} \\ + \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \psi_{j,d} \psi_{j',d'}.\end{aligned}$$

By definition of ψ, the first term reduces to $\frac{\epsilon}{n}(\mathcal{J}_{\tilde{\pi}}(i, a') - \mathcal{J}_{\tilde{\pi}}(i, a)) < 0$. The second term is quadratic in ϵ/n: we can find $\epsilon > 0$ small enough such that $\mathcal{L}_{E,B}(\tilde{\pi} + \psi) - \mathcal{L}_{E,B}(\tilde{\pi}) < 0$, which contradicts the minimality of $\tilde{\pi}$. This concludes the proof. □

Proof of Proposition 3.3. Suppose a (local) optimal embedding π maps a single node i in E to multiple blocks in B. Hence, there exist $2 \leq k_i \leq k$ blocks $\omega(i) := \{a_{k_1}, \ldots, a_{k_i}\} \in [k]$ such that $\pi_{i,a} > 0$ for all $a \in \omega(i)$. By Lemma 3.1, necessarily it must be that for any $a \in \omega(i)$

$$a \in \arg\min_{d \in [k]} \sum_{j=1}^{n} \sum_{b'=1}^{k} \|E_{i,j} - B_{d,b'}\|_2^2 \, \pi_{j,b'}. \tag{7}$$

Without loss of generality, assume that $\omega(i) = \{1, \ldots, k_i\}$ and that $\pi_{i,1} = \max_{a \in \omega(i)} \pi_{i,a}$. We can construct a new coupling π' from π by transporting all the mass in $\{(i, b)\}_{b=2}^{k_i}$ to $(i, 1)$, without violating our first-order condition:

$$\pi' = \pi + \underbrace{\sum_{b=2}^{k_i} \pi_{i,b} \, (\delta_{i,1} - \delta_{i,b})}_{:= \, \psi},$$

with ψ being a signed measure with zero total mass. We emphasize that π' is still a valid probabilistic mapping:

1. **(Nonnegativity):** For $(j,d) \notin \{(i,a),(i,b)\}$ we have $\pi'_{j,d} = \pi_{j,d} \geq 0$. For $(j,d) = (i,a)$, $\pi'_{i,a} = \pi_{i,a} + \sum_{b=2}^{k_i} \pi_{i,b} \geq 0$. For $(j,d) = (i,b)$, $\pi'_{i,b} = \pi_{i,b} - \pi_{i,b} = 0$. Hence, $\pi'_{j,d} \geq 0$ for all j, d.

2. **(Left marginal)** By assumption π has left marginal uniform, i.e. $\sum_{d=1}^{k} \pi_{j,d} = 1/n$ for every j. For any fixed $j \neq i$ we have $\psi_{j,d} = 0$ for all d, hence $\sum_d \pi'_{j,d} = \sum_d \pi_{j,d} = 1/n$. For $j = i$, $\sum_{d=1}^{k} \pi'_{i,d} = \sum_d \pi_{i,d} + \sum_d \psi_{i,d} = \frac{1}{n} + 0 = \frac{1}{n}$, since ψ has zero total mass. Therefore, the left marginal of π' is still uniform on $[n]$.

3. **(Total mass)** Since the left marginal is preserved for every j, the total mass $\sum_{j,d} \pi'_{j,d}$ remains $\sum_j (1/n) = 1$, so π' is normalized as a coupling.

Since $\mathcal{L}_{E,\tilde{B}}$ is quadratic in the coupling, we can split it into what is essentially a Taylor decomposition in the first and second-order variations of $\mathcal{L}_{E,\tilde{B}}$. Explicitly, one can see that:

$$\begin{aligned}
\mathcal{L}_{E,\tilde{B}}(\pi') &= \mathcal{L}_{E,\tilde{B}}(\pi + \psi) \\
&= \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, (\pi+\psi)_{j,d} \, (\pi+\psi)_{j',d'} \\
&= \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \pi_{j,d}\pi_{j',d'} \\
&\qquad + 2 \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \pi_{j,d}\psi_{j',d'} + \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \psi_{j,d}\psi_{j',d'} \\
&= \mathcal{L}_{E,\tilde{B}}(\pi) + 2 \underbrace{\sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \pi_{j,d}\psi_{j',d'}}_{\text{(D1)}} + \underbrace{\sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \psi_{j,d}\psi_{j',d'}}_{\text{(D2)}}.
\end{aligned}$$

We first show that (D1) $= 0$. Since $\psi = \sum_{b=2}^{k_i} \pi_{i,b} \, (\delta_{i,1} - \delta_{i,b})$ the sum collapses to:

$$\begin{aligned}
\text{(D1)} &= \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \pi_{j,d}\psi_{j',d'} \\
&= \sum_{b=2}^{k_i} \pi_{i,b} \left(\sum_{j=1}^{n} \sum_{d=1}^{k} \|E_{j,i} - B_{d,1}\|_2^2 \, \pi_{j,d} - \sum_{j=1}^{n} \sum_{d=1}^{k} \|E_{j,i} - B_{d,b}\|_2^2 \, \pi_{j,d} \right)
\end{aligned}$$

and the last displayed equation is null by (7).

Next, we upper bound (D2).

$$\begin{aligned}
\text{(D2)} &= \sum_{j,j'=1}^{n} \sum_{d,d'=1}^{k} \|E_{j,j'} - B_{d,d'}\|_2^2 \, \psi_{j,d}\psi_{j',d'} \\
&= \sum_{b=2}^{k_i} \pi_{i,b}^2 \left(\|E_{i,i} - B_{b,b}\|_2^2 + \|E_{i,i} - B_{1,1}\|_2^2 \right) - 2 \sum_{2 \leq b < b' \leq k_i} \pi_{i,b}\pi_{i,b'} \|E_{i,i} - B_{b,b'}\|_2^2 \\
&= \sum_{b=2}^{k_i} \pi_{i,b}^2 \left(\|B_{b,b}\|_2^2 + \|B_{1,1}\|_2^2 \right) - 2 \sum_{2 \leq b < b' \leq k_i} \pi_{i,b}\pi_{i,b'} \|B_{b,b'}\|_2^2 \\
&\leq \min_{b=2,\ldots,k_i} \left\{ \pi_{i,b}^2 \right\} \sum_{b=1}^{k_i} \left(\|B_{b,b}\|_2^2 + \|B_{1,1}\|_2^2 \right) \\
&\leq \min_{b=2,\ldots,k_i} \left\{ \pi_{i,b} \right\}^2 k_i 2\|B\|_\infty^2 \leq 2 \left(\frac{1}{nk_i} \right)^2 k_i \|B\|_\infty^2 = \frac{2\|B\|_\infty^2}{n^2 k_i} \leq \frac{\|B\|_\infty^2}{n^2},
\end{aligned}$$

where we used that there are no self-loops in E, $k_i \geq 2$, and that $\sum_{d=1}^{k_i} \pi_{i,d} = \frac{1}{n}$.

Repeating this operation for all $i \in [n]$ manifests a deterministic mapping with increased cost bounded by $\|B\|_\infty^2/n$, hence proving the result.

□

Lemma B.1. *Let $\hat{\pi}$ be the optimal probabilistic estimator and $\mathrm{z}(\hat{\pi})$ be the rounded deterministic labeling estimator, according to Eq.* (6)*. Consider the solution over deterministic couplings $\bar{z} \in \arg\min_{z \in \mathcal{Z}_{n,k}} \mathcal{L}\left(E, \mathrm{B}\left(z\right)\right)$. Then, we have:*

$$\mathcal{L}_E\left(\bar{z}\right) \leq \mathcal{L}_E\left(\mathrm{z}(\hat{\pi})\right) \leq \mathcal{L}_E\left(\bar{z}\right) + O(1/n)$$

Proof of Lemma B.1. From the definition of $\bar{z}$, and since $\mathrm{z}(\hat{\pi})$ is a feasible point for the optimization problem defining $\bar{z}$, we have

$$\mathcal{L}_E\left(\bar{z}\right) \leq \mathcal{L}_E\left(\mathrm{z}(\hat{\pi})\right).$$

Now by definition of $\hat{\pi}$, we have

$$\mathcal{L}_E\left(\hat{\pi}\right) \leq \mathcal{L}_E\left(\bar{z}\right).$$

Finally, using the definition of $\mathrm{z}(\hat{\pi})$ and the optimality gap in Proposition 3.3, we have

$$\mathcal{L}_E\left(\mathrm{z}(\hat{\pi})\right) \leq \mathcal{L}_{E,\mathrm{B}(\hat{\pi})}\left(\mathrm{z}(\hat{\pi})\right) \leq \mathcal{L}_E\left(\hat{\pi}\right) + D_n \leq \mathcal{L}_E\left(\bar{z}\right) + D_n.$$

where $D_n = \|\,\mathrm{B}\left(\hat{\pi}\right)\|_\infty^2/n$. □

Proof of Theorem 3.5, SBM. We define

$$\hat{\theta}_{ij} = \mathrm{B}\left(\hat{z}\right)_{\hat{z}(i),\hat{z}(j)}, \text{ and } \quad \theta^*_{ij} = {B^*_{z^*(i),z^*(j)}}^2.$$

It is easy to see that $\mathcal{L}_E(\hat{z}) = \frac{1}{n^2}\sum_{i,j}\left\|E_{ij} - \hat{\theta}_{ij}\right\|^2$, and that $\mathcal{L}_E(z^*) \leq \frac{1}{n^2}\sum_{i,j}\left\|E_{ij} - \theta^*_{ij}\right\|^2 = \mathcal{L}_{E,B^*}(z^*)$. Using Lemma B.1, and the fact that $\bar{z}$ is optimal, we have

$$\mathcal{L}_E(\hat{z}) \leq \mathcal{L}_E(\bar{z}) + D_n \leq \mathcal{L}_E(z^*) + D_n \leq \mathcal{L}_{E,B^*}(z^*) + D_n,$$

or equivalently,

$$\sum_{i,j}\left\|E_{ij} - \hat{\theta}_{ij}\right\|^2 \leq \sum_{i,j}\left\|E_{ij} - \theta^*_{ij}\right\|^2 + n^2 D_n. \tag{8}$$

From there, letting $\|E - \hat{\theta}\|^2 = \sum_{i,j}\left\|E_{ij} - \hat{\theta}_{ij}\right\|^2$, we use the sesquilinearity of the inner product to get

$$\|E - \hat{\theta}\|^2 = \|\hat{\theta} - \theta^*\|^2 + 2\langle\hat{\theta} - \theta^*, \theta^* - E\rangle + \|E - \theta^*\|^2, \tag{9}$$

which we can combine with Eq. (8) to obtain

$$\|\hat{\theta} - \theta^*\|^2 + 2\langle\hat{\theta} - \theta^*, \theta^* - E\rangle + \|E - \theta^*\|^2 \leq \|E - \theta^*\|^2 + n^2 D_n,$$

and thus

$$\frac{1}{n^2}\sum_{i,j}\left\|\hat{\theta}_{ij} - \theta^*_{ij}\right\|^2 \leq \frac{2}{n^2}\left\langle\theta^* - \hat{\theta}, E - \theta^*\right\rangle + D_n. \tag{10}$$

We bound the cross product term as in Gao et al. [8][3], to obtain that for any constant $C' > 0$, there exists a constant $\tilde{C} > 0$ only depending on C' such that

$$\sum_{i,j}\left\|\hat{\theta}_{ij} - \theta^*_{ij}\right\|^2 \leq 2\tilde{C}\sqrt{n\log k}\sqrt{\sum_{i,j}\left\|\hat{\theta}_{ij} - \theta^*_{ij}\right\|^2} + 4\tilde{C}^2\sqrt{k^2 + n\log k}\sqrt{n\log k} + D_n n^2,$$

[2]Note that this is not $B(z^*)$

[3]None of the concentration inequalities use explicitly the optimality of $\bar{z}$.

with probability at least $1 - \exp(-C' n \log k)$. Solving the above quadratic inequality, we get the following inequality:

$$\sqrt{\sum_{i,j} \left\| \hat{\theta}_{ij} - \theta^*_{ij} \right\|^2} \leq \frac{1}{2} \left(2\tilde{C} \sqrt{n \log k} + \sqrt{\left(20 \tilde{C}^2 \sqrt{k^2 + n \log k} \sqrt{n \log k} \right) + 4 D_n n^2} \right). \quad (11)$$

Since from Proposition 3.3 $D_n n^2 = n$, for some constant $C > 0$ only depending on $\tilde{C}$ we have

$$\sum_{i,j} \left\| \hat{\theta}_{ij} - \theta^*_{ij} \right\|^2 \leq C \left(k^2 + n \log k \right),$$

which implies Eq. (11) with probability at least $1 - \exp(-C' n \log k)$, concluding the proof for the SBM case.

In the graphon case, using a similar argument, we obtain that for any constant $C' > 0$, there exists a constant $C_1 > 0$ only depending on C' such that

$$\sum_{i,j} \left\| \hat{\theta}_{ij} - \theta^*_{ij} \right\|^2 \leq C_1 \left(n^2 \left(\frac{1}{k^2} \right)^{\alpha \wedge 1} + n^2 D_n + k^2 + n \log(k) \right),$$

with probability at least $1 - \exp(-C' n \log k)$. Since $D_n n^2 = C_2 n$ for some $C_2 > 0$, setting $k = n^\beta$ and minimizing the right hand side over $\beta \in (0, 1)$ yields:

$$\sum_{i,j} \left\| \hat{\theta}_{ij} - \theta^*_{ij} \right\|^2 \leq C \left(n^2 \cdot n^{-\frac{2\alpha}{\alpha+1}} + n \log n \right),$$

for some constant $C > 0$. This concludes the proof. □

Proof of Corollary 3.6. The proof follows the same idea as the proof of Theorem 3.5, but applying the concentration bounds found in Klopp et al. [23, Section 4]. More precisely, equation (24) of Klopp et al. [23], where the optimality of the least-squares estimator is used, is transformed to incorporate our optimality gap D_n.

Let $\theta^* \in [0,1]^{n \times n}$ be the matrix with entries $\theta^*_{ij} = \rho_n W^*(\xi_i, \xi_j)$ for $i \neq j$ and $\theta^*_{ii} = 0$, and write $E := E - \theta^*$ for the (centered) noise matrix. Let $\hat{\theta}_{ij} = \mathrm{B}(\hat{z})_{\hat{z}(i),\hat{z}(j)}$ be our estimator and $\bar{\theta}_{ij} = \bar{B}_{\bar{z}(i),\bar{z}(j)}$ the unrestricted block-constant least-squares oracle (i.e. $\bar{z} \in \arg\min_{z \in \mathcal{Z}_{n,k}} \mathcal{L}_E(\mathrm{B}(z))$). For any $z \in \mathcal{Z}_{n,k}$ and any symmetric $Q \in \mathbb{R}^{k \times k}_{\mathrm{sym}}$, let $\Theta_{z,Q}$ denote the symmetric matrix with off-diagonal entries $Q_{z(i),z(j)}$, and define the family

$$\mathcal{T}[k] := \left\{ \Theta_{z,Q} \,:\, z \in \mathcal{Z}_{n,k},\, Q \in \mathbb{R}^{k \times k}_{\mathrm{sym}} \right\},$$

of k-class block-constant matrices. Let $\Theta_* \in \mathcal{T}[k]$ be the best Frobenius-norm approximation of θ^* in $\mathcal{T}[k]$.

From Lemma B.1 and the optimality of $\bar{z}$, for any $\Theta \in \mathcal{T}[k]$,

$$\left\| E - \hat{\theta} \right\|_F^2 \;=\; n^2\, \mathcal{L}_E(\hat{z}) \;\leq\; n^2\, \mathcal{L}_E(\bar{z}) + n^2 D_n \;\leq\; \left\| E - \Theta \right\|_F^2 + n^2 D_n.$$

Expanding the squared Frobenius norm (similarly to Equation (9)) yields the analogue of Klopp et al. [23, eq. (24)] with our additive optimality gap D_n:

$$\left\| \hat{\theta} - \theta^* \right\|_F^2 \;\leq\; \left\| \theta^* - \Theta_* \right\|_F^2 + 2\left\langle \hat{\theta} - \theta^*,\, E \right\rangle + 2\left\langle \theta^* - \Theta_*,\, E \right\rangle + n^2 D_n. \quad (12)$$

From there, we can conclude exactly as in Klopp et al. [23].

□

Remark B.2. We note that in the sparse case, we have an unrestricted least square estimator, i.e. we do not constraint the size of the partitions nor the maximum value of the block matrix. Klopp et al. [23] also showed more results for constrained least square estimators. An interesting question is whether we could get such an estimator using a penalized version of the srGWB, where we add a penalty term to the objective that penalizes small partitions. We leave this question for future work.

C Scope of the theory beyond simple graphs

Our theoretical results are stated for simple graphs: binary, undirected, and without self-loops. The estimator itself is defined for any square matrix of edge values, and we apply it to weighted and decorated graphs in Section 4 and Section E.2. We summarise here which parts of the analysis extend directly and which require new arguments.

Directed graphs. The objective $\mathcal{L}_E(\pi, B)$ in (2) is a quadratic form in π and does not use the symmetry of E or B. For a non-symmetric E, node i appears both in the row and in the column slot, so Lemma 3.1 holds with the symmetrised marginal cost

$$\mathcal{J}_\pi(i,a) = \sum_{j=1}^{n}\sum_{b=1}^{k}\Big(\|E_{i,j} - B_{a,b}\|_2^2 + \|E_{j,i} - B_{b,a}\|_2^2\Big)\pi_{j,b},$$

and the proof of Proposition 3.3 goes through unchanged. The closed form (3) then returns a non-symmetric block matrix: the block class has k^2 free parameters instead of $k(k+1)/2$, and since both are of order k^2, only the constants in the rates of Theorem 3.5 change.

Self-loops. The absence of self-loops is used at a single place, when bounding the second-order term (D2) in the proof of Proposition 3.3, where the diagonal entries $E_{i,i}$ appear. With self-loops, $\|B_{a,b}\|_2^2$ is replaced there by $\|E_{i,i} - B_{a,b}\|_2^2$, so the gap bound holds with $\|B\|_\infty^2$ replaced by $\max_{i,a,b}\|E_{i,i} - B_{a,b}\|_2^2$. It remains of order $1/n$ whenever the edge values are bounded.

Bounded weights. If the edge values lie in a bounded set $\Omega \subset \mathbb{R}$, the block matrices of interest take values in the convex hull of Ω, and Lemma 3.1 and Proposition 3.3 hold verbatim, with $\|B\|_\infty$ bounded in terms of Ω. The rates of Theorem 3.5 carry over without major modifications, with constants depending on Ω.

Unbounded weights. For unbounded edge distributions (e.g. Poisson or Gaussian), Lemma 3.1 still holds, since it is a deterministic statement given E and B. The statistical analysis, however, requires a high-probability bound replacing the corresponding step in the proofs of Gao et al. [8]. We have not carried out this verification and leave it as an open extension; empirically, the estimator remains well-behaved in the Gaussian settings of Section E.2.

Decorated graphs. Decorated graphs cover more than weighted graphs (e.g. multiplex networks) and require a more careful treatment. For a finite decoration space Ω, the arguments of Dufour and Olhede [24] adapt directly and the theory carries over; we omit this extension for clarity and space. For general decoration spaces, one must first ensure that the limit object exists, and the estimation step no longer reduces to averaging: the object to be estimated for each pair of blocks is a probability measure on Ω , whereas the closed form (3) only returns its barycenter (see Dufour and Olhede [52] for more details).

D Algorithmic details

D.1 On implicit regularization toward deterministic couplings

Assel et al. [22] observed that srGW implementations on networks often converge to *deterministic* solutions, even under entropic regularization (e.g., Sinkhorn) which is designed to promote diffuse couplings. This can be partly explained by the fact that deterministic assignments are extreme points of the coupling polytope. We formalize this intuition with two results: (i) deterministic mappings always exist that are stationary points of (2) (Lemma 3.1), and (ii) the best deterministic solution incurs at most an $O(1/n)$ loss relative to the optimal probabilistic coupling (Proposition 3.3). Still, these results do not fully explain why gradient-based methods typically converge to deterministic solutions – which can arguably be seen as a form of *implicit regularization*, with gradient descent converging to more structured solutions. Empirically, the BCM solver returns near-fully-deterministic couplings at convergence: across our 10 synthetic settings, the fraction of nodes with non-trivial probabilistic mass on more than one block is typically zero (see Section E).

Algorithm 2 Conditional Gradient Solver for srGWB

Input: coupling $\pi^{(0)} \in \Pi_{[n],k}$, matrix $B \in [0,1]^{k \times k}$
for $t = 0, 1, 2, \ldots$ until convergence **do**

$$G^{(t)} \leftarrow \nabla_\pi \, \mathcal{L}_{E,B}(\pi^{(t)}).$$

$$\psi^{(t)} \leftarrow \operatorname*{argmin}_{\psi \in \Pi_{[n],k}} \langle \psi, G^{(t)} \rangle.$$

Choose step size $\gamma^{(t)} \in [0,1]$ by exact line search.

$$\pi^{(t+1)} \leftarrow (1 - \gamma^{(t)})\pi^{(t)} + \gamma^{(t)}\psi^{(t)}.$$

end for
Output: π

Self-pruning of the number of blocks. The free right marginal in (4) appears to promote simple solutions: blocks that do not improve the objective lose their mass, and the effective number of blocks $\widehat{k}$ adapts to the data (Section 3.4). We do not have a theoretical account of this behaviour. The resulting $\widehat{k}$ is not the output of an explicit penalty or criterion, and it therefore falls outside the usual model-selection framework. Adding an explicit sparsity-promoting regularisation would move in that direction, but it modifies the optimisation landscape and does not simplify the analysis (Section 5). The phenomenon is not specific to our setting: the tendency of semi-relaxed formulations to concentrate mass on a reduced set of target points has been observed in prior srGW work [18, 20–22]. A plausible explanation, which we cannot currently substantiate, is that this is a property of the optimisation dynamics rather than of the objective. The linear minimisation step of Algorithm 2 decouples across nodes and returns a vertex of $\Pi_{[n],k}$, i.e. a deterministic assignment; conditional-gradient iterates thus move along vertices of the coupling polytope and appear to select structured stationary points, in the spirit of implicit regularisation. We regard both the self-pruning of blocks and the convergence to deterministic couplings as open problems. Empirical frequencies of diffuse nodes across sparsity levels and initialisations are reported in Table 5.

E Additional numerical experiments

E.1 Simulated data

We consider synthetic experiments where graphs are generated from an unknown stochastic block model or continuous graphon. Given an observed network E on n nodes, our goal is to recover a low-dimensional block structure that approximates the underlying graphon using the semi-relaxed Gromov-Wasserstein barycenter. In our experiments, we consider the 10 different latent, unknown network structures, specifically 6 continuous graphons with varying smoothness properties, and 4 are SBMs, as defined in Table 2.

For each ground-truth structure, we sample graphs of increasing size $n = 200, 300, \ldots, 10000$ nodes. For each sampled graph, we apply our pipeline to recover a block structure with at most k blocks which seeks to minimize the semi-relaxed GW objective to the observed adjacency matrix. For the continuous graphons, we set $k \leq \lceil \sqrt{n} \rceil$ blocks and $k \leq 20$ for the SBMs. This effectively manifests a block estimator for the unobserved ground truth. To solve the associated optimization problem, we rely on the block-coordinate minimization algorithm (Algorithm 1) with a conditional gradient solver based on the Frank-Wolfe algorithm [18, 53]: see Remark E.1. Our implementation is based on the `pot` library [37]. Due to the non-convex nature of the objective, the initialization of the probabilistic labeling π plays a crucial role in the quality of the solution. We explored various initialization strategies, and found a clustering-based initialization to be the most effective for graphons, while for SBMs a product based is typically more accurate both in estimating connection probabilities and number of blocks; see discussion below.

We performed 20 independent (Monte Carlo) trials for each configuration (ground-truth structure and sample size), and report the MSE achieved by the recovered estimate for each trial in Figures 5 and 6.

ID	*Graphon Definition* $W(u,v)$
1.	$W(u,v) = u \cdot v$
2.	$W(u,v) = 0.5(u+v)$
3.	$W(u,v) = \exp\left(-\max(u,v)^{0.75}\right)$
4.	$W(u,v) = \lvert u-v \rvert$
5.	$W(u,v) = 1-\lvert u-v \rvert$
6.	$W(u,v) = 0.5 + 0.5 \cdot \sin(2\pi u) \cdot \sin(2\pi v)$

ID	ν	B
1.	$\begin{bmatrix}0.5 & 0.5\end{bmatrix}$	$\begin{bmatrix}0.2 & 0.8\\ 0.8 & 0.2\end{bmatrix}$
2.	$\begin{bmatrix}0.5 & 0.5\end{bmatrix}$	$\begin{bmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{bmatrix}$
3.	$\begin{bmatrix}0.3 & 0.7\end{bmatrix}$	$\begin{bmatrix}0.8 & 0.2\\ 0.2 & 0.8\end{bmatrix}$
4.	$\begin{bmatrix}0.2 & 0.5 & 0.3\end{bmatrix}$	$\begin{bmatrix}0.8 & 0.1 & 0.2\\ 0.1 & 0.7 & 0.3\\ 0.2 & 0.3 & 0.5\end{bmatrix}$

Table 2: Ground truths considered in our experiments: graphons (left) and SBMs (right).

While we set the maximum number of blocks k in our estimate to be controlled by the sample size, e.g., $k \leq \lceil \sqrt{n} \rceil$, the algorithm adaptively determines the effective number of blocks, often selecting fewer than the maximum allowed, as fewer blocks are often optimal for accurate approximation. This is however initialization sensitive: see Figures 7 to 9.

On the choice of initialization. In our experiments, we evaluated the impact of different initialization strategies on the performance of the relaxed least-squares estimator. Specifically, we considered a k-means-based, a spectral-clustering based one, and a product based one. The k-means initialization is based on clustering the observed nodes from the observed adjacency matrix to obtain an initial hard assignment of nodes to blocks. This hard assignment is then converted into a probabilistic labeling by setting $\pi_{i,a} = 1/n$ if node i is assigned to block a, and $\pi_{i,b} = 0$ for all other blocks. Similarly, the spectral-based initialization clusters the nodes with the usual spectral clustering. Differently, the product-based initialization initializes the right-marginal – encoding the distribution of weights across blocks of the barycenter—rather than the probabilistic labeling itself, which is then initialized as the product measure on the two marginals. Our experiments, as illustrated by Figures 8 and 9, indicate that, for Graphons, clustering-based initialization is significantly less prone to local minimum entrapment compared to the random and product-based strategies. This is evident from the consistently lower mean squared error (MSE) achieved across all ground-truth structures and sample sizes. For graphons, k-means initialization, in particular, effectively starts the optimization process favorably, and achieves the smallest loss. However, we observed that the solutions obtained using k-means initialization tend to be less sparse compared to those derived from random or product-based initializations, as appreciated by the colorscale in Figures 8 and 9 displaying the selected number of blocks by the algorithm. This suggests that while k-means initialization mitigates the risk of local minimum, it may favor solutions with higher block number, potentially overestimating the complexity of the underlying structure. In contrast, random and product initializations, despite being more susceptible to local minimum, occasionally yield sparser solutions that align more closely with the true block structure. Yet, when the data is generated from a block model, product-based initialization typically grants the fastest convergence.

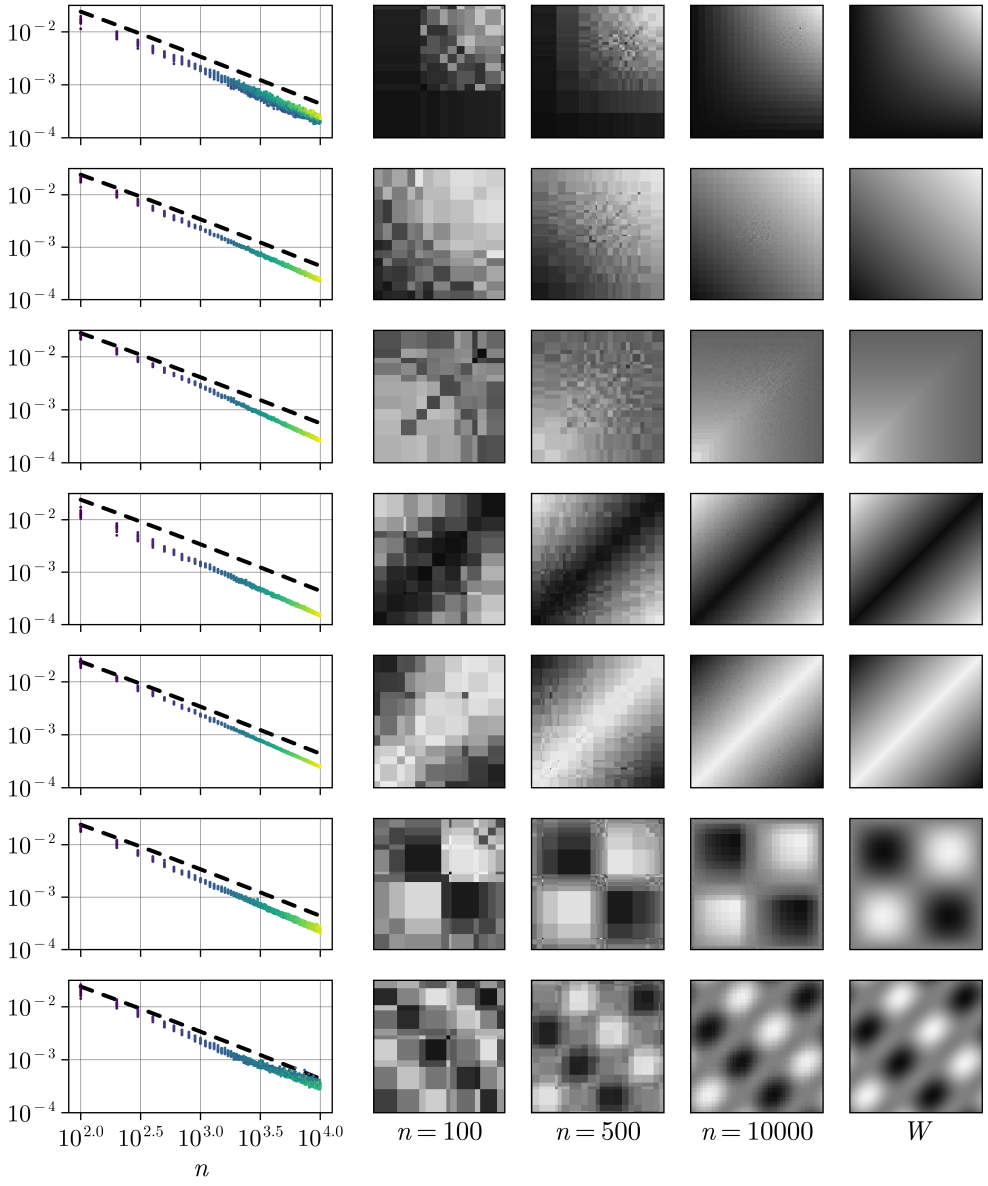


Figure 5: Left: Empirical performance of the relaxed least-squares estimator across different graphon ground-truth structures and sample sizes in 20 independent experiments. We evaluate the Gromov-Wasserstein distance of our estimator to the ground-truth for varying sample sizes n. Each subplot represents a different ground-truth structure, with theoretical rates indicated by bold lines. Right: instance of estimated block structure (aligned to the latent truth for ease of visualization) obtained from observed networks of size $n = 100, 500, 10000$, compared to the ground truth. The ordering reflects that of Table 2.

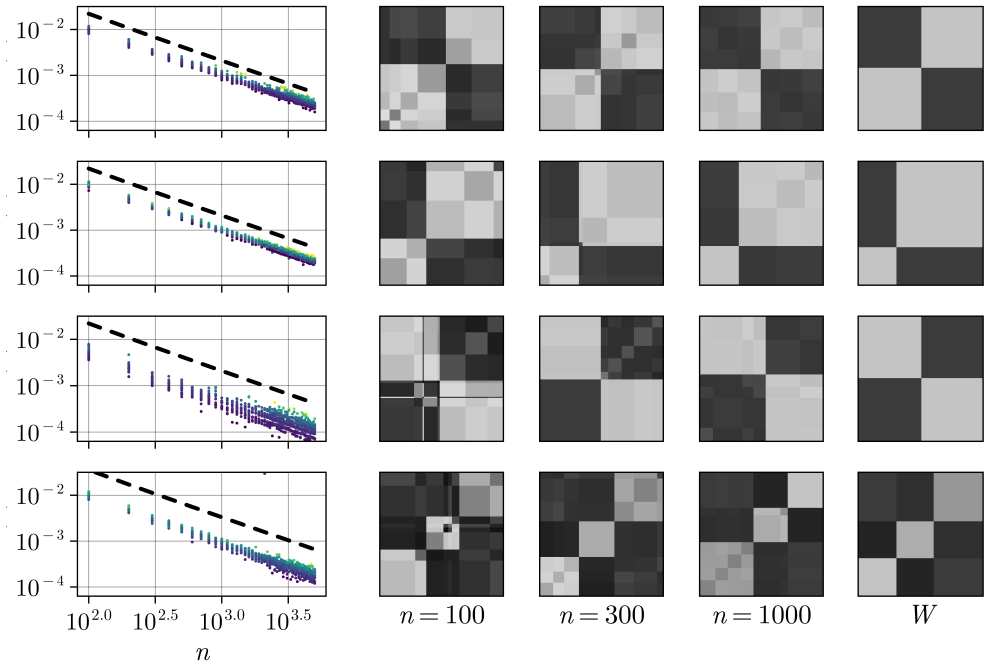


Figure 6: Left: Empirical performance of the relaxed least-squares estimator across different SBM ground-truth structures and sample sizes in 20 independent experiments. We evaluate the Gromov-Wasserstein distance of our estimator to the ground-truth for varying sample sizes n. Each subplot represents a different ground-truth structure, with theoretical rates indicated by bold lines. Right: instance of estimated block structure (aligned to the latent truth for ease of visualization) obtained from observed networks of size $n = 100, 500, 10000$, compared to the ground truth. The ordering reflects that of Table 2.

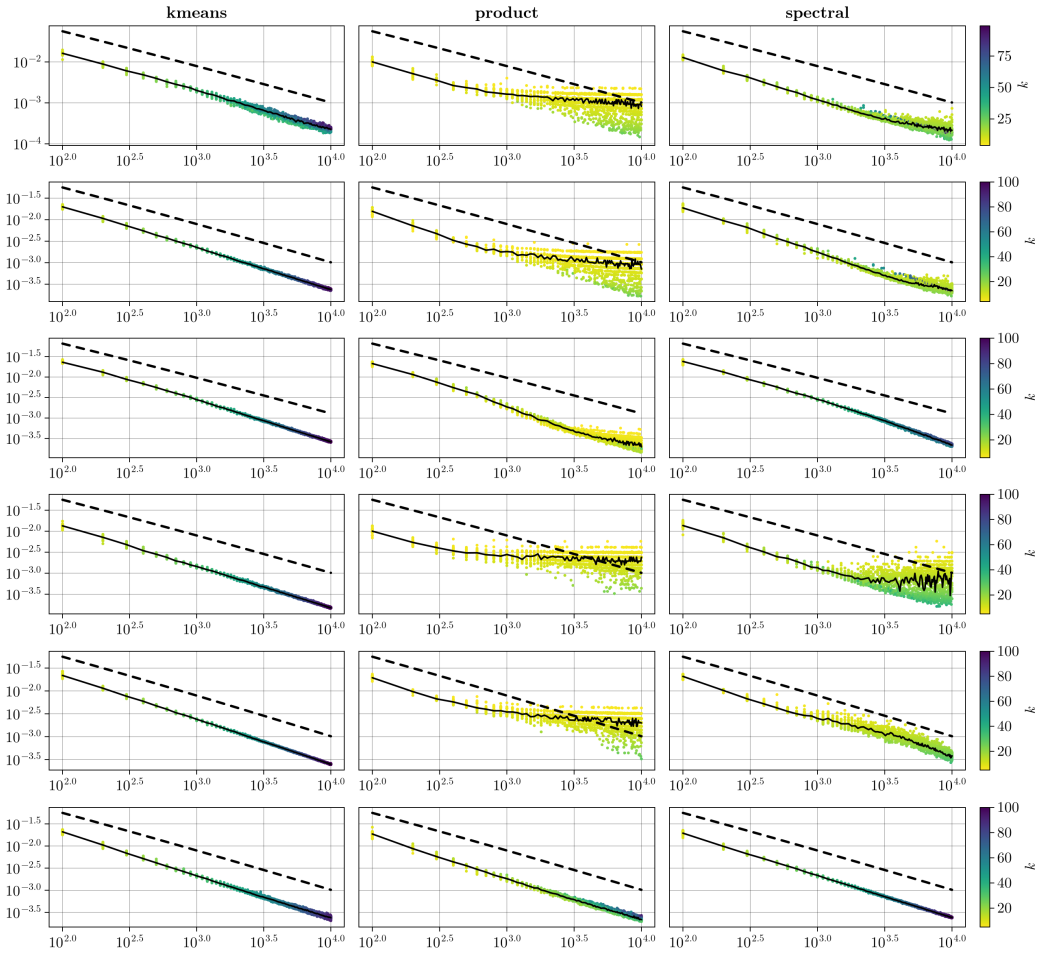


Figure 7: Empirical rates of estimation (GW distance) for graphons for increasing network size. Colorscale shows selected size of best approximating block model, with lighter colors indicating fewer blocks. Full lines represent the median loss and dashed lines represent theoretical rates. Column corresponds to initialization strategy (k-means, product, spectral). The ordering reflects that of Table 2.

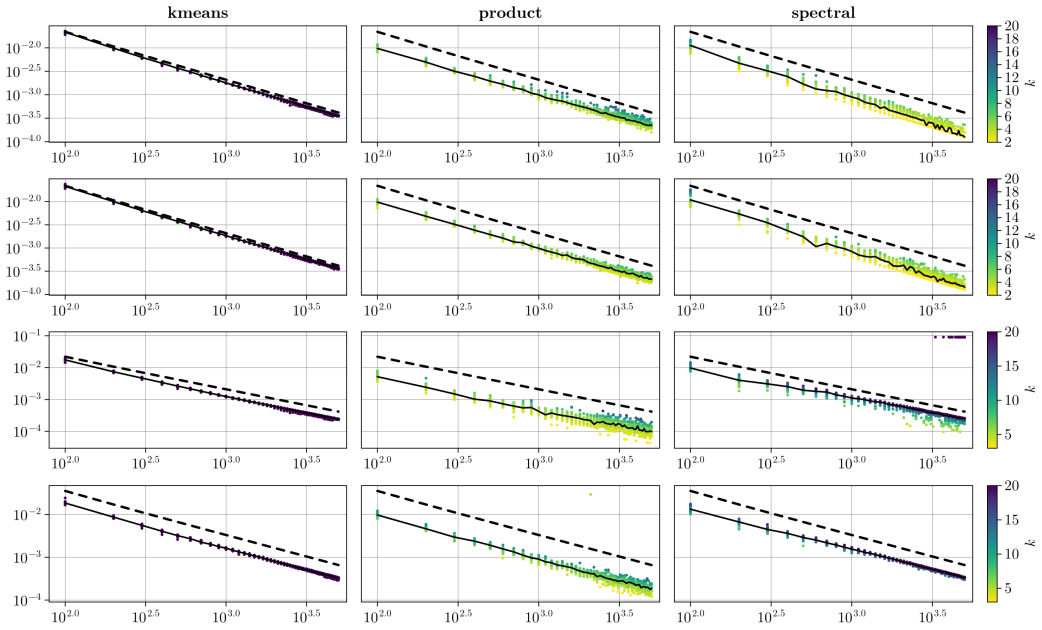


Figure 8: Empirical rates of estimation (GW distance) for SBMs for increasing network size and adaptive block numbers. Colorscale shows selected size of best approximating block model, with lighter colors indicating fewer blocks. Full lines represent median loss and dashed lines theoretical rate. Column corresponds to initialization strategy (k-means, product, spectral). The ordering reflects that of Table 2.

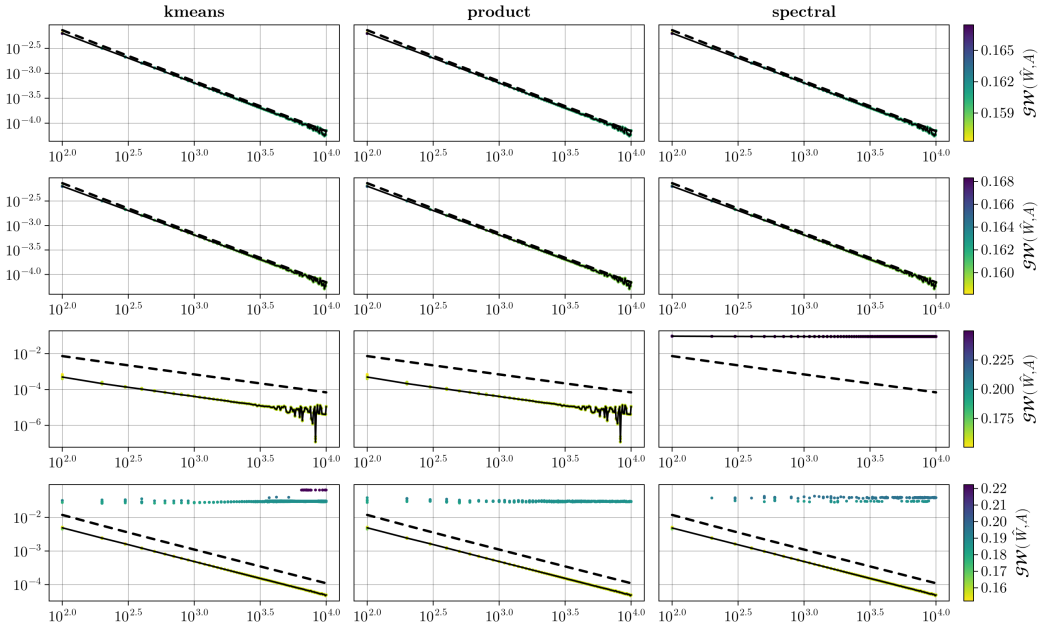


Figure 9: Empirical rates of estimation (GW distance) for SBMs for increasing network size and given (oracle) number of blocks. Full lines represent the median loss and dashed lines represent theoretical rates. The ordering reflects that of Table 2.

Algorithmic implementation, computational speed and comparison with existing methods. In the (SRGWB) step of Algorithm 1, we employ a revisited Frank-Wolfe method [53] to solve the constrained non-convex quadratic optimization problem over the probabilistic labeling π. Similarly to [18], we employ a conditional gradient solver which is known to converge to local stationary point on non-convex problems [19]. Since we consider a semi-relaxed problem – with barycenter marginal estimated in the process, rather than constrained – the cost of this operation is of order $O(nk)$, with possible parallelization on GPUs. Our implementation relies on the semi-relaxed Gromov-Wasserstein solver made available by the `POT` library (Python Optimal Transport [37]), and runs entirely on GPUs. We terminate the outer loop when the relative change in $\mathcal{L}_E$ drops below 10^{-6}; fewer than 50 outer iterations suffice for n up to 10^4 (runtime in Figure 10).

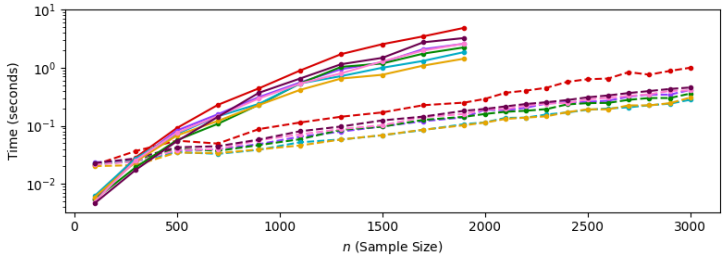


Figure 10: Runtime of our method with k-means initialization for graphons in Table 2. Lines represent the time on CPU (Mac M1) , the dashed lines GPU (Nvidia L40s GPU with a single node).

Figure 11 shows the comparison of our method to other approaches fitting piecewise constant graphons: the histogram estimator with greedy estimation [25], the variational approximation of the Maximum Likelihood Estimator (MLE) [29] (implemented in `sparseBM` [42]), and the Largest Gap (LG) method [41]. We also benchmark against methods that do not perform dimensionality reduction: the Sort and Smooth (SAS) method [28], the Neighbourhood Smoothing (NBDsmooth) method [39], and the Universal Singular Value Thresholding (USVT) method [26]. We highlight that while all the methods above apply to simple binary graphs, they do not directly extend to more complex network data, contrary to ours.

The srGW-based methods achieve a lower Gromov-Wasserstein distance to the truth while maintaining superior scalability and faster runtimes compared to the HistogramEstimator and `sparseBM`. These existing block-model methods exhibit significantly higher computational costs as the sample size n increases, often encountering convergence issues (often associated with log-likelihood maximization for binary variable). While USVT [26] remains competitive in specific regimes (e.g., when the graphon is rank 1 as is $G1$), it fails to accurately recover dissociative structures, such as those found in SBM 2. Furthermore, unlike our barycentric approach, USVT yields an $n \times n$ estimator that lacks the low-dimensional interpretability required to simplify and compress the underlying generative mechanism.

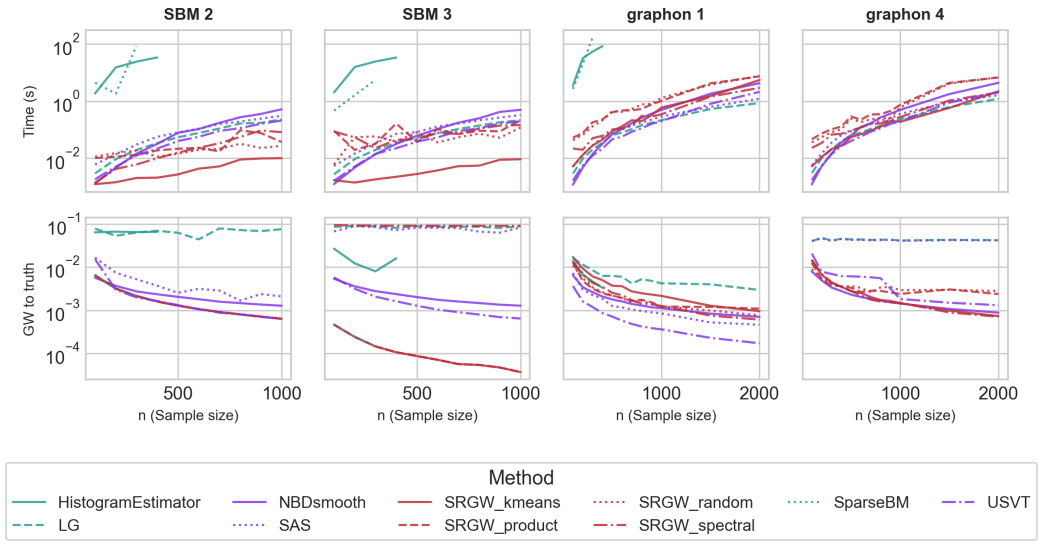


Figure 11: Average running time (seconds, CPU) for different methods. For Histogram and the `sparseBM`, graph sizes only go up to 300 as sparseBM would run into numerical instability and convergence issues for higher number of nodes and the computation times were too important. Red indicates our method with different initialization strategies (k-means, spectral, product), blue indicates other low-dimensional methods, and purple indicates methods that do not perform dimensionality reduction.

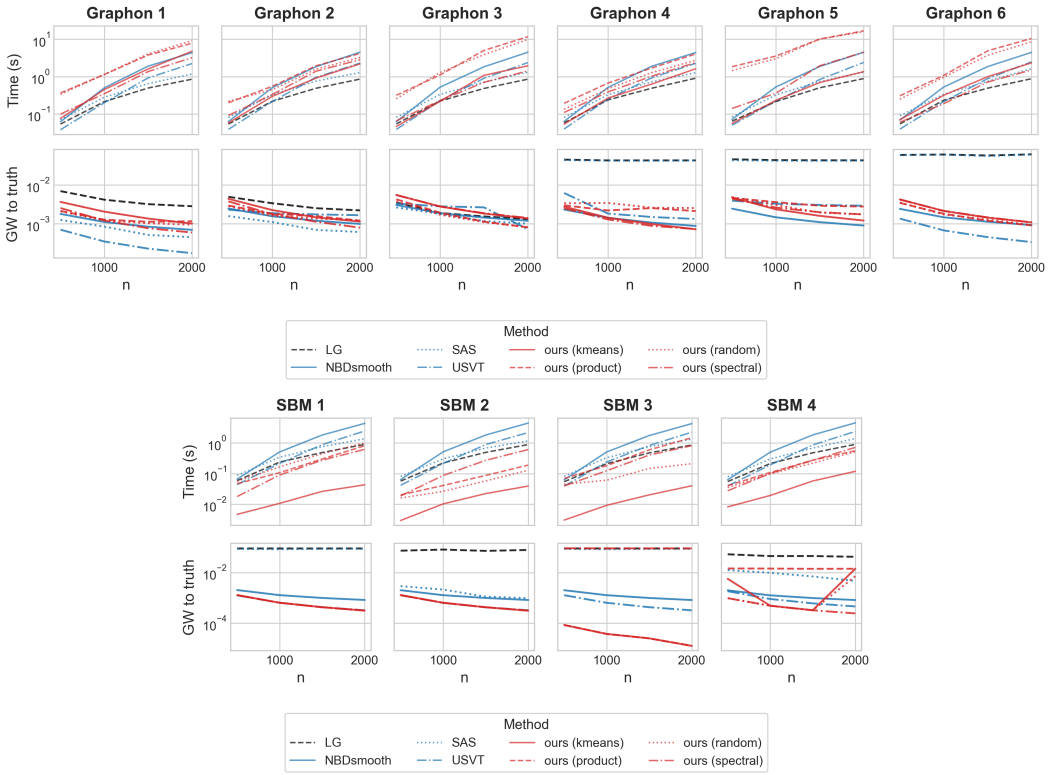


Figure 12: Complete results for all graphons (top) and SBMs (bottom) for all methods.

On the relation to structured Gromov–Wasserstein barycenter methods. A natural question is how our estimator relates to GW-barycenter methods for graphons, in particular the structured GW barycenter (SGWB) framework of Xu et al. [16]. Although SGWB and our estimator are both based on GW-type objectives, the aims are different: SGWB is designed for graph averaging through regularised barycenters, whereas our objective is recovery of the latent graphon mechanism with statistical guarantees. We discuss this relation here and explain why a direct head-to-head comparison on the recovery loss of Table 1 would not be informative.

Theoretical object vs. computed object. The theoretical analysis in Xu et al. [16] relates the cut distance between graphons to the GW distance between their step-function approximations, leveraging the weak regularity lemma. The estimator that is actually computed, however, is a regularised surrogate of this object: the public implementation solves the inner OT step with a proximal-point Sinkhorn iteration of the form $\mathrm{kernel} \propto \exp(-\mathrm{cost}/\beta)$, and adds a quadratic spectral smoothness penalty $\alpha\, D^\top D$ on the barycenter, with D a discrete second-difference operator.[4]Neither of these two regularisers is part of the cut-distance argument, and the diffuseness of the resulting transport plan is governed by the entropic parameter β rather than by the GW objective itself. As a consequence, the rates of convergence established for the unregularised barycenter do not transfer verbatim to the implemented estimator, and our Proposition 3.3 does not apply to β-regularised solutions, whose support spreads with β.

Hyperparameters and absence of an oracle-rate barycenter size. The implemented method requires the user to specify (α, β, γ) controlling the smoothness, entropic and Wasserstein regularisation strengths, together with two solver budgets and a Sinkhorn tolerance. We are not aware of a data-driven rule to pick (α, β) in our setting, where ground-truth assignments are unavailable. Furthermore, the barycenter size in the public implementation is set by a logarithmic heuristic of the largest input graph ($\approx \log_2 n$), which is independent of the smoothness exponent α of the unknown graphon and therefore does not realise the rate-optimal scaling $k \asymp n^{1/(\alpha \wedge 1+1)}$ used in Theorem 3.5.

E.2 Weighted networks

Our theory is developed for binary graphs, while the method applies to weighted ones (Section C), as in the real-data examples, where no ground truth is available. To isolate how edge weights affect performance, we consider synthetic weighted graphs with known ground truth. Latent positions are drawn as $\xi_i \sim \mathrm{Unif}[0,1]$ i.i.d., and edge weights are drawn independently given the latent positions, with mean $W(\xi_i, \xi_j)$, either from a Beta distribution,

$$E_{ij} \sim \mathrm{Beta}\big(\kappa\, W(\xi_i, \xi_j),\ \kappa\,(1 - W(\xi_i, \xi_j))\big),$$

which is bounded and covered by Section C, or from a Gaussian distribution, $E_{ij} \sim \mathcal{N}\big(W(\xi_i, \xi_j), \sigma^2\big)$, which is unbounded and outside the scope of our theory. We compare our estimator with USVT, the strongest competitor in Table 1, over 20 replicates per setting (Table 3).

Table 3: Weighted benchmark: error ($\times 10^{-4}$) of our estimator / USVT, mean$_{\pm \mathrm{std}}$ over 20 replicates. **Bold**: better of the two. Beta-distributed edges (WG4–WG6) are bounded; Gaussian edges (WG1–WG3, WG7, WG8) are unbounded.

ID	$W(u,v)$	Noise	$n = 500$	$n = 1000$	$n = 1500$	$n = 2000$
WG1	$e^{-(u-v)^2}$	$\mathcal{N}, \sigma = 0.1$	$4.32_{\pm 0.35}$ / $\mathbf{1.61}_{\pm 0.07}$	$1.92_{\pm 0.17}$ / $\mathbf{1.16}_{\pm 0.03}$	$1.18_{\pm 0.09}$ / $\mathbf{1.01}_{\pm 0.03}$	$\mathbf{0.82}_{\pm 0.08}$ / $0.94_{\pm 0.04}$
WG2	$2\min(u-\frac{1}{2}, v-\frac{1}{2})$	$\mathcal{N}, \sigma = 1$	$168_{\pm 10}$ / $\mathbf{158}_{\pm 21}$	$87.3_{\pm 4.0}$ / $\mathbf{78.9}_{\pm 1.7}$	$59.3_{\pm 1.3}$ / $\mathbf{53.2}_{\pm 1.2}$	$44.1_{\pm 1.3}$ / $\mathbf{42.9}_{\pm 5.1}$
WG3	$10\min(u-\frac{1}{2}, v-\frac{1}{2})$	$\mathcal{N}, \sigma = 2$	$\mathbf{666}_{\pm 80}$ / $32{,}087_{\pm 148}$	$\mathbf{348}_{\pm 46}$ / $32{,}057_{\pm 110}$	$\mathbf{239}_{\pm 22}$ / $32{,}028_{\pm 74}$	$\mathbf{170}_{\pm 15}$ / $32{,}015_{\pm 48}$
WG4	$1 - \lvert u - v \rvert$	Beta, $\kappa = 2$	$\mathbf{11.2}_{\pm 1.2}$ / $31.0_{\pm 0.9}$	$\mathbf{5.36}_{\pm 0.52}$ / $28.6_{\pm 0.7}$	$\mathbf{3.57}_{\pm 0.22}$ / $27.7_{\pm 0.6}$	$\mathbf{2.58}_{\pm 0.14}$ / $27.2_{\pm 0.6}$
WG5	$1 - \lvert u - v \rvert$	Beta, $\kappa = 4$	$\mathbf{8.40}_{\pm 0.94}$ / $29.3_{\pm 0.9}$	$\mathbf{4.02}_{\pm 0.35}$ / $27.7_{\pm 0.7}$	$\mathbf{2.60}_{\pm 0.17}$ / $27.1_{\pm 0.6}$	$\mathbf{1.92}_{\pm 0.19}$ / $26.8_{\pm 0.6}$
WG6	$e^{-\max(u,v)^{3/4}}$	Beta, $\kappa = 2$	$\mathbf{8.91}_{\pm 0.77}$ / $27.1_{\pm 1.6}$	$\mathbf{4.99}_{\pm 0.62}$ / $25.0_{\pm 1.0}$	$\mathbf{3.65}_{\pm 0.67}$ / $24.5_{\pm 1.0}$	$\mathbf{3.03}_{\pm 0.91}$ / $4.02_{\pm 0.09}$
WG7	$e^{-2\lvert u-v \rvert}$	$\mathcal{N}, \sigma = 1$	$\mathbf{153}_{\pm 17}$ / $204_{\pm 44}$	$\mathbf{63.6}_{\pm 7.6}$ / $119_{\pm 26}$	$\mathbf{38.2}_{\pm 4.9}$ / $82.7_{\pm 13.3}$	$\mathbf{27.1}_{\pm 2.9}$ / $60.5_{\pm 7.9}$
WG8	$e^{-(u-v)^2}$	$\mathcal{N}, \sigma = 3$	$\mathbf{1{,}937}_{\pm 65}$ / $84{,}092_{\pm 398}$	$\mathbf{1{,}129}_{\pm 27}$ / $84{,}303_{\pm 208}$	$\mathbf{828}_{\pm 20}$ / $84{,}331_{\pm 142}$	$\mathbf{664}_{\pm 31}$ / $84{,}348_{\pm 111}$

Our estimator is better on six of the eight settings, and its error decreases with n throughout. The two exceptions, WG1 and WG2, are smooth and close to low-rank, the regime in which USVT is close to an oracle (see the discussion of Table 1 in Section E.1); on WG1 the gap closes monotonically

[4]See `methods/learner.py` of `https://github.com/HongtengXu/SGWB-Graphon`

and reverses at $n = 2000$. On WG3 and WG8, which have the widest range of edge values, USVT does not converge at all: its error is flat in n (from 32,087 to 32,015 and from 84,092 to 84,348), consistent with its thresholding being calibrated to a known bounded range. On the Beta settings WG4 and WG5, USVT improves only marginally, while our error decreases by a factor of four.

E.3 Sparse networks

The experiments above consider dense networks, whereas many large real-world networks are sparse. Corollary 3.6 points to two effects of sparsity. On the optimisation side, the relaxation is not harmed: the optimality gap of Proposition 3.3 scales as $\|B\|_\infty^2/n$, which is of order ρ_n^2/n for a block matrix with entries of order ρ_n (whether $\mathrm{B}(\hat{\pi})$ satisfies this is the additive term left uncontrolled in Corollary 3.6). On the statistical side, the problem becomes harder Klopp et al. [23].

We generate graphs in the setting of Corollary 3.6, with $\theta^*_{ij} = \rho_n W(\xi_i, \xi_j)$ for the six graphons of Table 2, sparsity levels $\rho_n \in \{n^{-1/4},\, n^{-1/2},\, 5\log n/n\}$, network sizes $n \in \{500, 1250, 2000\}$, and 10 replicates per configuration. We use the sparsity-aware schedule $k \asymp (\rho_n n^2)^{1/4}$ suggested by Klopp et al. [23]. Since the raw risk shrinks trivially with ρ_n, we report the ratio of the loss of each estimator to that of the trivial estimator $\hat{\theta} \equiv 0$: lower is better, and a ratio of one means doing no better than predicting an empty graph. Table 4 reports the results for $\rho_n = n^{-1/4}$.

Table 4: Sparse regime, $\rho_n = n^{-1/4}$: ratio of the loss to that of the trivial estimator $\hat{\theta} \equiv 0$ (lower is better), ours / USVT, mean$_{\pm\text{std}}$ over 10 replicates. Ours uses the best initialisation per setting.

ID	$W(u,v)$	$n = 500$	$n = 1250$	$n = 2000$
G1	uv	$\mathbf{0.19}_{\pm 0.02}$ / $1.00_{\pm 0.00}$	$\mathbf{0.09}_{\pm 0.01}$ / $0.51_{\pm 0.51}$	$0.06_{\pm 0.01}$ / $\mathbf{0.02}_{\pm 0.00}$
G2	$\frac{1}{2}(u+v)$	$0.12_{\pm 0.01}$ / $\mathbf{0.04}_{\pm 0.00}$	$0.06_{\pm 0.01}$ / $\mathbf{0.02}_{\pm 0.00}$	$0.04_{\pm 0.01}$ / $\mathbf{0.02}_{\pm 0.00}$
G3	$e^{-\max(u,v)^{0.75}}$	$0.16_{\pm 0.01}$ / $\mathbf{0.04}_{\pm 0.00}$	$0.10_{\pm 0.01}$ / $\mathbf{0.03}_{\pm 0.00}$	$0.08_{\pm 0.01}$ / $\mathbf{0.02}_{\pm 0.00}$
G4	$\lvert u-v\rvert$	$\mathbf{0.18}_{\pm 0.03}$ / $1.00_{\pm 0.00}$	$\mathbf{0.08}_{\pm 0.01}$ / $0.29_{\pm 0.00}$	$\mathbf{0.06}_{\pm 0.01}$ / $0.29_{\pm 0.00}$
G5	$1-\lvert u-v\rvert$	$\mathbf{0.11}_{\pm 0.01}$ / $\mathbf{0.11}_{\pm 0.01}$	$\mathbf{0.06}_{\pm 0.00}$ / $0.10_{\pm 0.00}$	$\mathbf{0.04}_{\pm 0.00}$ / $0.10_{\pm 0.00}$
G6	$\frac{1}{2}+\frac{1}{2}\sin(2\pi u)\sin(2\pi v)$	$\mathbf{0.13}_{\pm 0.01}$ / $0.22_{\pm 0.01}$	$\mathbf{0.07}_{\pm 0.01}$ / $0.21_{\pm 0.01}$	$\mathbf{0.05}_{\pm 0.00}$ / $0.21_{\pm 0.00}$

Our estimator remains consistent under sparsity: its ratio decreases with n in every setting, and no configuration degenerates. The comparison with USVT is regime-dependent, in the way discussed for Table 1: USVT is competitive or better on the smooth graphons G2 and G3 and on G1 at $n = 2000$, but it collapses to the null estimator on G1 at $n = 500$ (and in about half of the replicates at $n = 1250$, as its standard deviation indicates) and on G4 at $n = 500$; on G4, G5 and G6 its ratio then stays essentially flat in n, while ours keeps decreasing.

Near-deterministic couplings across sparsity levels. Table 5 reports the fraction of nodes carrying mass on more than one block at convergence, over all runs of the sparse grid, for each sparsity level and initialisation scheme. The typical outcome is a fully deterministic coupling in every cell (the distributions are strongly right-skewed), and even in the worst cell fewer than 7% of the nodes are diffuse on average. The dependence on the sparsity level is initialisation-specific rather than intrinsic: k-means initialisation yields slightly more diffuse couplings in the two sparsest regimes, whereas product and spectral initialisations become essentially deterministic there.

Table 5: Fraction of nodes carrying mass on more than one block at convergence, over all runs of the sparse grid: mean$_{\pm\text{std}}$, with the share of runs having no diffuse node in parentheses.

ρ_n	k-means	product	spectral
$n^{-1/4}$	$0.001_{\pm 0.006}$ (98%)	$0.021_{\pm 0.064}$ (87%)	$0.004_{\pm 0.018}$ (81%)
$n^{-1/2}$	$0.050_{\pm 0.092}$ (64%)	$0.000_{\pm 0.000}$ (100%)	$0.000_{\pm 0.000}$ (89%)
$5\log n/n$	$0.068_{\pm 0.119}$ (64%)	$0.000_{\pm 0.000}$ (100%)	$0.000_{\pm 0.000}$ (93%)

E.4 Real data: EEG

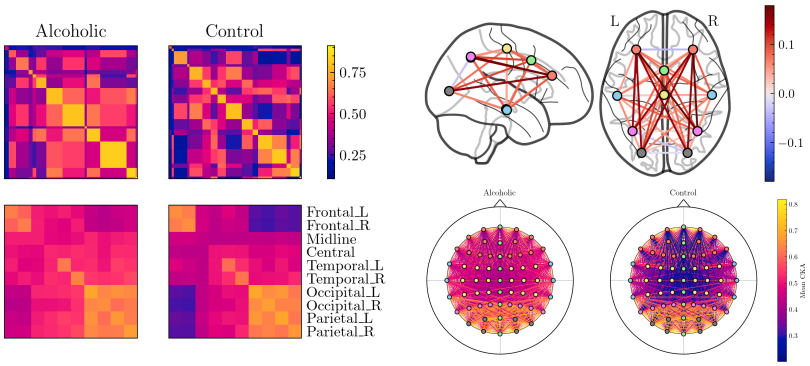


Figure 13: Low-dimensional co-activation structures estimated by the srGW barycenter for subjects exposed to a single stimulus.

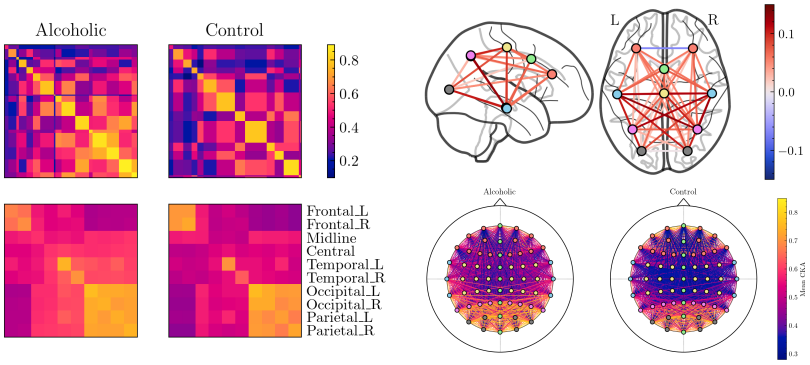


Figure 14: Low-dimensional co-activation structures estimated by the srGW barycenter for subjects exposed to two matching stimuli.

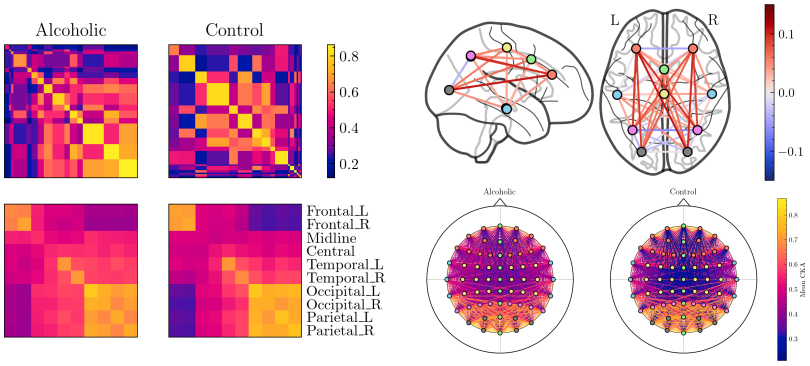


Figure 15: Low-dimensional co-activation structures estimated by the srGW barycenter for subjects exposed to two non-matching stimuli.

E.5 Real data: OpenFlight Airport Database

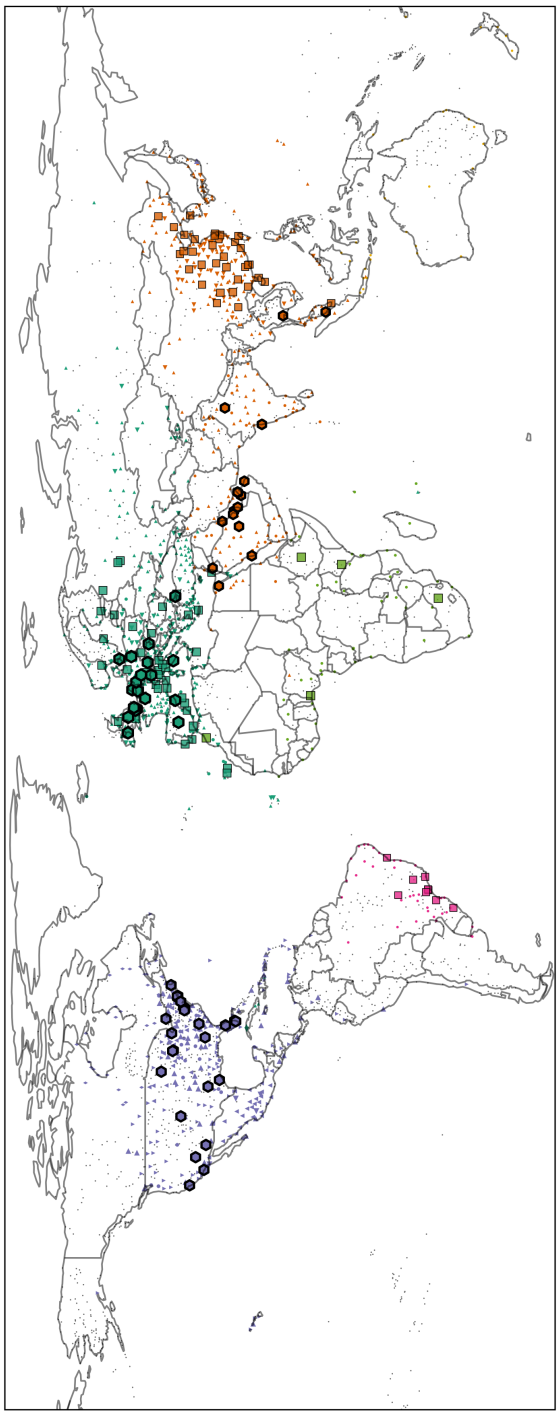


Figure 16: Airports location colored by inferred cluster membership, with node size proportional to cluster connectivity. Our method uses 23 clusters out of a maximum of 57. The biggest clusters in terms of number of airports correspond to mostly local airports, and account for close to half of the total number of airports (dark small nodes in the map).

E.6 Stability across initialisations and multi-start protocol

Since Eq. (4) is non-convex, we assess the sensitivity of the real-data results to the initialisation. On both real networks, the OpenFlight airport network ($k_{\max} = 57$) and the EEG co-activation barycenters ($k = 40$), we run each of the three initialisation schemes of Section 3.4 with 20 random seeds and record the objective value $\mathcal{L}_E(\hat{\pi})$ at convergence (Table 6). In the absence of ground truth, this objective is the natural relabelling-invariant quantity to compare across runs. On the airport network, the objective varies by less than 5% across seeds within each initialisation scheme; on the smaller EEG network the spread is larger, between 12% and 22%. k-means initialisation gives the smallest spread on both datasets and the lowest objective on the airport network; on EEG, spectral initialisation reaches the lowest single value, but with a spread almost twice as large.

Table 6: Objective value $\mathcal{L}_E(\hat{\pi})$ ($\times 10^{-3}$) at convergence on the real networks, over 20 seeds per initialisation scheme. Span = max − min.

	OpenFlight airports				EEG co-activation			
Init.	min	max	span	span/min	min	max	span	span/min
k-means	2.411	2.472	0.061	2.5%	11.912	13.381	1.469	12.3%
spectral	2.549	2.642	0.093	3.6%	11.012	13.446	2.433	22.1%
product	2.533	2.639	0.106	4.2%	11.356	13.777	2.422	21.3%

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three concrete claims—(i) formulating SBM/graphon estimation as a srGW barycenter problem, (ii) an $O(1/n)$ optimality gap bound and minimax-optimal estimation rates, and (iii) a GPU-accelerated block-coordinate algorithm—each of which is substantiated by Proposition 3.3, Theorem 3.5 and Corollary 3.6, and Algorithm 1 respectively. Theoretical results are developed for simple binary undirected graphs and this scope is stated explicitly in the Contributions paragraph.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses limitations and future directions. In particular: (1) adaptive, theoretically grounded selection of the number of blocks k is left open—adding sparsity-promoting regularisation to (4) substantially modifies the optimisation landscape and requires separate analysis; (2) theoretical results are developed for simple binary undirected graphs (scope stated explicitly in the Contributions), whereas the extension to decorated graphs used in Section 4 relies on Dufour and Olhede [24] but is not analysed here; (3) the choice of $k_{\max}$, initialisation sensitivity, and the gap between Frank–Wolfe stationarity and global optimality are discussed in Section 3.4 and Remark 3.4.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results (Lemma 3.1, Proposition 3.3, Theorem 3.5, Corollary 3.6, and the supporting Lemma B.1) are numbered and cross-referenced. Assumptions—Hölder smoothness for graphons, standard SBM parameter conditions, and the non-vanishing block-fraction assumption used in the sparse regime—are stated inline within each result. Complete proofs appear in Section B, and proof sketches are provided in the main text to convey intuition.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The algorithm is fully described in pseudocode (Algorithm 1), the implementation is based on the publicly available `POT` library [37], and the experimental protocol (20 Monte Carlo runs, $n \in \{100, \dots, 10{,}000\}$, $k \leq \lceil \sqrt{n} \rceil$ for graphons, three initialisation strategies) is described in Section 4 and Section E.1. The real datasets (EEG, OpenFlight) are publicly available.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Both real datasets used (EEG [43] and OpenFlight [47]) are publicly available. An repository has been submitted with our paper, and contains the relevant code: see [40].

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Section 4 and Section E.1 describe the generative models, sample sizes ($n \in \{100, 200, \ldots, 10{,}000\}$), block number bounds ($k \leq \lceil \sqrt{n} \rceil$ for graphons, $k \leq 20$ for SBMs), initialisation strategies (k-means, spectral, product), and the conditional gradient solver from `POT`. The benchmark baselines and their implementations are also cited.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All synthetic experiments are based on 20 independent Monte Carlo runs. Figures report the median loss over these runs (indicated in captions, e.g., Figures 2 and 7 to 9), and the spread across runs is visible in the figures. Table 1 reports medians; individual-run variability is shown in the appendix figures.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Figure 10 reports wall-clock runtimes for all graphon experiments on both CPU (Apple M1) and GPU (Nvidia L40s, single node). The algorithmic complexity is $O(nk)$ per iteration and is discussed in Section E.1.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper presents foundational methodology for network inference. No human subject data is collected, no sensitive data is scraped, and no harmful applications are enabled. The EEG dataset used is a pre-existing, publicly available benchmark.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper presents foundational methodology for latent network inference (estimation of SBMs and graphons from a single observed graph). It does not introduce a deployed system, dataset, or model targeted at a specific application, and we do not foresee specific positive or negative societal impacts that go beyond those generally associated with statistical methodology for network data.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pre-trained models, generative models, or scraped datasets. No high-risk assets are introduced.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The EEG dataset [43], the OpenFlight Airport Database [47], and the `POT` library [37] are all cited. All three are publicly available for academic use. Explicit licenses are not stated in the paper, but all are properly attributed.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new datasets, benchmarks, or pre-trained models.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or new human subject experiments. The EEG data used is a pre-existing, openly released dataset [43].

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human subject research is conducted. The EEG recordings analysed are from a pre-existing publicly available dataset; any original IRB approval is the responsibility of the original data collectors.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as any methodological component of this research. Any use was limited to writing assistance only and does not affect scientific content or results.

Acknowledgments and Disclosure of Funding

Funding: U.Naepels and L. V. Santoro were supported by a Swiss NSF grant to V.M. Panaretos.
Competing interests: The authors declare no competing interests.