Mean Aggregation Meets Balanced Optimal Transport: Compact, Separating Metrics for Attributed Graph Neural Networks

Liz Lemma Future Detective

January 18, 2026

Abstract

Recent theory unified expressivity, universality, and distribution-free generalization of message passing neural networks (MPNNs) on attributed graphs by introducing a computation-tree optimal-transport pseudometric that makes the graphon-signal domain compact and exactly matches the geometry induced by normalized-sum aggregation. However, the dominant aggregation in practice is mean aggregation (degree-normalized neighborhood averaging). We develop a new hierarchy of attributed computation objects—mean iterated degree measures (mean-IDMs)—and define a balanced optimal transport distance between their distributions (mean-DIDMs). The resulting mean-DIDM mover distance yields a compact metric identification of a natural class of attributed graphon-signals with degrees bounded away from zero. We prove (1) Lipschitz continuity of all mean-aggregation MPNNs with respect to this distance, (2) a quantitative separation theorem showing that agreement of all bounded-Lipschitz mean-MPNNs implies small mean-DIDM distance, and (3) a Stone–Weierstrass universality theorem: mean-aggregation MPNNs uniformly approximate any continuous functional on the compact domain. Compactness and Lipschitzness further yield distribution-free generalization bounds based on covering numbers, without parameter-count assumptions. For finite graphs, we give a polynomial-time algorithm to compute the proposed distance via a recursion of balanced optimal transport problems and show empirically that it tracks output perturbations of mean-aggregation GNNs, improving stability calibration relative to the prior unbalanced-OT metric tailored to normalized-sum aggregation.

Table of Contents

2. 2. Related work: WL distances and OT metrics (Chen et al.), Tree Mover's Distance (Chuang–Jegelka), graphon/graphon-signal limits (Lovász; Levie), fine-grained expressivity on graphons (Böker et al.), and the source metric for normalized-sum aggregation; position the balanced-OT shift as aggregation-aligned metric design.

3. 3. Preliminaries: graphon-signals; weak* topology; Wasserstein-1 on compact spaces; mean aggregation MPNNs on graphs/graphons; degree-floor condition and why it is needed.

4. 4. Mean-IDMs and Mean-DIDMs: formal definitions; measurability; factorization of mean aggregation through normalized neighbor push-forwards; handling isolated nodes (domain restriction first; optional conventions later).

5. 5. Metrics: define $d_{\mathrm{IDM},L}^{\mathrm{mean}}$ on H_L and $W_1(\cdot,\cdot; d_{\mathrm{IDM},L}^{\mathrm{mean}})$ on $\mathcal{P}(H_L)$; prove these metrize the relevant weak* topologies; define δ_L^{mean} on graphon-signals.

6. 6. Compactness: prove H_L and $\mathcal{P}(H_L)$ compact; prove compactness of $\mathrm{WL}_{r,\alpha}^d/\sim$ under δ_L^{mean} using continuity of $(W,f) \mapsto \Gamma_{(W,f),L}^{\mathrm{mean}}$ (initially under cut distance + degree floor).

7. 7. Lipschitz continuity of mean-MPNNs: show mean-MPNNs on graphons factor through mean-IDMs/DIDMs; inductive Lipschitz bounds; statement in both node-feature and readout forms.

8. 8. Separation and fine-grained expressivity: Stone–Weierstrass setup on H_L and $\mathcal{P}(H_L)$; show mean-MPNNs separate points; quantitative converse ("if all bounded-Lipschitz mean-MPNNs are close then the mean-DIDM distance is small").

9. 9. Universal approximation: density of mean-MPNNs in $C(H_L, \mathbb{R})$ and $C(\mathcal{P}(H_L), \mathbb{R})$; extend to graphon-signals and then to finite graphs via density/embedding arguments.

10. 10. Distribution-free generalization: covering-number bound for Lipschitz hypothesis classes on compact $(\mathrm{WL}_{r,\alpha}^d/\sim, \delta_L^{\mathrm{mean}})$ and on $\mathcal{P}(H_L)$; discussion of (unknown) metric entropy and what would tighten bounds.

11. 11. Computation on finite graphs: dynamic programming algorithm to compute δ_L^{mean} via repeated balanced OT; complexity $O(L\,N^5 \log N)$; practical accelerations (Sinkhorn) flagged as future work.

12. 12. Experiments (recommended): correlation with output perturbations of mean-aggregation GNNs; compare against δ^{DIDM} (unbalanced, normalized-sum-aligned) and TMD/WL distances; stability calibration under degree variation and mild sparsification.

13. 13. Discussion and limitations: role of degree floor; options to remove it (regularized mean aggregation / teleportation); extension to edge features and multi-relational settings; sparse regime remains open.

1 Introduction and motivation

Mean aggregation is the prevailing design choice in contemporary message passing neural networks: the neighborhood term is formed as an empirical average (or, more generally, a normalized attention-weighted average) of neighbor features, and then combined with the current node feature through a Lipschitz update. From a modeling standpoint, this choice enforces scale invariance with respect to the number of neighbors and, in the dense graphon regime, matches the natural normalization in which $W(x,\cdot)/\deg_W(x)$ becomes a probability kernel. From a statistical standpoint, mean aggregation behaves as a Monte Carlo estimator of a conditional expectation, so its stability is governed by concentration of averages rather than by potentially unbounded degree-dependent sums. These considerations motivate treating the normalized neighbor distribution as the fundamental object that a depth-t layer can access.

This perspective suggests that the appropriate notion of similarity between two attributed graphon-signals (W,f) and (V,g) should be aligned with (i) the recursion induced by mean aggregation and (ii) the topology in which empirical averages are continuous functionals. A metric that is instead aligned with normalized-sum aggregation, i.e., one that transports non-normalized neighborhood mass and therefore distinguishes changes in total neighborhood mass, can be misaligned with mean-aggregation invariances: two nodes with identical neighbor feature distributions but different degrees may be far in a mass-sensitive geometry, even though a mean aggregator only perceives the normalized distribution. Conversely, when the learning architecture discards mass information by normalization, a mass-sensitive metric may enforce regularity properties that are irrelevant for the hypothesis class and may fail to provide sharp converses (separation) for that class. The present work therefore adopts a balanced formulation throughout, in which each neighborhood is represented by a probability measure and each transport problem is a Wasserstein–1 distance between probability measures.

Formally, we work in the compact attribute domain B_r^d and impose a degree-floor $\deg_W(x) \geq \alpha$ a.e. to ensure that the normalization $W(x,\cdot)/\deg_W(x)$ is well-defined and uniformly controlled. We then define a hierarchy of mean-computation iterated degree measures (mean-IDMs) by setting $H_0 = B_r^d$ and $H_{t+1} = H_t \times \mathcal{P}(H_t)$, and by recursively attaching to each point $x \in [0,1]$ both its current state and the law of its neighbors in the previous state space. Concretely, at depth t the normalized neighbor measure $\nu^{\mathrm{mean}}_{(W,f),t}(x) \in \mathcal{P}(H_t)$ is the pushforward of the probability measure $(W(x,\cdot)/\deg_W(x))\,\lambda$ under the state map $\gamma^{\mathrm{mean}}_{(W,f),t}$, and $\gamma^{\mathrm{mean}}_{(W,f),t+1}(x)$ stores the pair $\big(\gamma^{\mathrm{mean}}_{(W,f),t}(x), \nu^{\mathrm{mean}}_{(W,f),t}(x)\big)$. The depth-$L$ distributional summary of the entire graphon-signal is then the mean-DIDM $\Gamma^{\mathrm{mean}}_{(W,f),L} = (\gamma^{\mathrm{mean}}_{(W,f),L})_{\#}\lambda \in \mathcal{P}(H_L)$.

To compare two graphon-signals we endow each H_t with a recursively-

defined product metric $d_{\mathrm{IDM},t}^{\mathrm{mean}}$ in which the measure component is compared by Wasserstein–1 with ground metric $d_{\mathrm{IDM},t-1}^{\mathrm{mean}}$. The resulting mean-DIDM mover distance is

$$\delta_L^{\mathrm{mean}}((W,f),(V,g)) := W_1\big(\Gamma_{(W,f),L}^{\mathrm{mean}},\, \Gamma_{(V,g),L}^{\mathrm{mean}}\,;\, d_{\mathrm{IDM},L}^{\mathrm{mean}}\big).$$

This definition is engineered so that each layer of the hierarchy precisely mirrors the information accessible to a mean-aggregation message passing layer: the only operation on neighborhoods is the formation of an expectation under a probability measure, and the Wasserstein–1 geometry is the largest metric under which all bounded Lipschitz test functions are continuous and satisfy a sharp integral inequality.

Our first set of results establishes that the mean-IDM hierarchy provides a compact and well-behaved domain. Since B_r^d is compact and $\mathcal{P}(K)$ is compact whenever K is compact (in the weak* topology, metrized by W_1 on compact ground spaces), an induction shows that each H_L is compact, and consequently $\mathcal{P}(H_L)$ is compact as well. It follows that, after metric identification $(W,f) \sim (V,g)$ whenever $\delta_L^{\mathrm{mean}}((W,f),(V,g)) = 0$, the quotient space equipped with δ_L^{mean} is compact. This compactness is not merely topological bookkeeping: it underlies both universality (via approximation theorems on compact spaces) and distribution-free generalization (via covering numbers).

Our second set of results links the metric to the hypothesis class. Any depth-L mean-aggregation MPNN with Lipschitz updates factors through the mean-IDM states: there exist continuous maps $h_t : H_t \to \mathbb{R}^{d_t}$ such that $\mathbf{h}^{(t)}(x) = h_t(\gamma_{(W,f),t}^{\mathrm{mean}}(x))$ a.e., and the graph-level output depends on (W,f) only through the mean-DIDM $\Gamma_{(W,f),L}^{\mathrm{mean}}$. Moreover, the resulting functionals are Lipschitz with respect to δ_L^{mean}, with constants depending only on the network Lipschitz bounds. Complementing this, we prove a quantitative separation statement: if two mean-DIDMs cannot be distinguished by any bounded-Lipschitz mean-MPNN readout (within a prescribed class), then their Wasserstein–1 distance must be small. Together with compactness, this yields universality: scalar mean-MPNN features are dense in $C(H_L)$, and mean-DIDM readouts are dense in $C(\mathcal{P}(H_L))$, so continuous graphon-signal functionals can be approximated uniformly.

Finally, the metric is algorithmically meaningful on finite graphs in the dense regime. For attributed graphs with normalized degrees bounded below by α, the induced step graphon-signals inherit the same structure, and δ_L^{mean} can be computed exactly by a dynamic program that alternates between (i) computing Wasserstein–1 distances between normalized neighbor distributions using the previous-layer ground costs and (ii) updating a node-to-node cost matrix. The recursion performs balanced optimal transport at each depth and a final balanced transport between uniform node measures, yielding a polynomial-time exact algorithm for fixed L. This computational tractability, together with the compactness–Lipschitz–separation framework

above, positions δ_L^{mean} as an aggregation-aligned geometry for both theory and practice.

2 Related work

Distances and invariants inspired by the Weisfeiler–Leman (WL) refinement have long served as a bridge between combinatorial graph comparison and message passing architectures. On the algorithmic side, WL-based kernels and distances compare graphs by iteratively updating node labels (or features) and then aggregating these labels into multiset statistics; see, e.g., the survey perspective in the kernel literature and the connections to 1-WL expressivity of MPNNs. A parallel line of work replaces the purely discrete multiset comparison by an optimal-transport (OT) discrepancy between collections of node features produced by WL-like refinements. In particular, Chen et al. propose to couple WL-type iterations with Wasserstein/earth-mover computations so that graphs are compared by transporting refined node representations across graphs, thereby obtaining OT-based graph metrics or kernels that are more sensitive than simple histogram distances and that can interpolate between feature-based and structure-based similarity notions **??**. Our construction is in the same general spirit—we also compare distributions of refined node states by Wasserstein–1—but differs in the object being transported and in the normalization choices: we transport probability measures that represent *normalized* neighborhood information at each depth, rather than transporting unnormalized neighborhood mass.

Tree Mover's Distance (TMD) of Chuang and Jegelka **?** is particularly close in methodology to the present work. TMD builds a hierarchy of rooted neighborhood trees (or tree-like summaries) and defines a graph distance by solving a sequence of OT problems that compare these rooted structures, ultimately coupling root distributions across graphs. The distance is computable by dynamic programming over layers, and its recursive form mirrors the hierarchical nature of message passing. Conceptually, our mean-IDM/mean-DIDM hierarchy plays a similar role: depth-t node states consist of a current representation together with a measure describing the depth-$(t-1)$ neighborhood. The main distinction is that TMD is designed to be broadly aligned with WL-style multiset comparison, whereas we design the recursion to match mean aggregation specifically: the neighborhood object is a *probability* measure, and the transport problems are balanced. This choice leads to a geometry in which all bounded-Lipschitz test functionals (hence all Lipschitz mean-aggregation layers) are continuous with sharp integral bounds.

Our setting and the compactness arguments rely on the dense limit formalism of graphons. The theory of graph limits developed by Lovász and collaborators **?** provides the measure-theoretic language in which large dense

graphs are represented by symmetric measurable kernels up to measure-preserving relabelings. The extension from graphs to *graphon-signals* (or graphons with node attributes) has been studied in graph signal processing and learning, including constructions of graphon neural networks and the analysis of stability properties in the limit **??**. In these works, the limiting object typically serves to formalize convergence and to justify architectures that operate consistently across graph sizes. Our contribution is complementary: we use the graphon-signal model to define a compact metric domain on which mean-aggregation MPNNs induce Lipschitz functionals and for which converses (separation) can be proved.

The expressivity of graph neural networks in the graphon regime has also received increasing attention. Böker et al. investigate fine-grained expressivity phenomena for MPNNs on graphons, clarifying what information can be recovered from local aggregation in the dense limit and how this relates to classical WL refinements and to notions of indistinguishability under measure-preserving transformations **?**. These results motivate treating the graphon formalism not merely as an asymptotic convenience but as a genuine hypothesis space on which one can pose approximation and identifiability questions. Our mean-DIDM mover distance can be viewed as an "aggregation-aligned" metricization of this space: it is tailored so that the induced topology is sufficiently strong to support universal approximation arguments, yet sufficiently weak to quotient out invariances that mean aggregation cannot detect.

Finally, the closest conceptual precursor to our metric is the OT-based geometry introduced for *normalized-sum* aggregation, where neighborhood information is represented by a degree-weighted measure and transport is allowed to account for changes in total mass (or equivalently, the cost penalizes discrepancies in degree/total weight). Such a formulation is natural for sum-based architectures, since sums are sensitive to neighborhood mass and degree scaling. However, when the learning architecture normalizes by degree and thus discards mass information, a mass-sensitive metric can become misaligned: it may separate objects that are indistinguishable to the hypothesis class and may complicate separation statements that are meant to be tight for mean aggregation. The present work therefore makes a deliberate shift to a balanced-OT formulation at every level of the hierarchy. We view this as a general design principle: the metric used to control stability and generalization should reflect the invariances and the continuity properties of the aggregation operator implemented by the network.

3 Preliminaries

We work in the dense graph limit setting. A (simple) *graphon* is a symmetric measurable map $W\colon [0,1]^2 \to [0,1]$, interpreted as the limiting adjacency

kernel of a sequence of dense graphs. We augment W with bounded node attributes by a measurable *signal* $f\colon [0,1] \to B_r^d$, where $B_r^d \subset \mathbb{R}^d$ is the closed Euclidean ball of radius r. We refer to a pair (W, f) as a *graphon-signal*. As usual, graphons are only identifiable up to measure-preserving relabelings of $[0,1]$; consequently, any notion of distance or equivalence between graphon-signals must ultimately be compatible with pushforwards by measure-preserving bijections.

Finite attributed graphs embed into this framework via step functions. Concretely, given an n-node graph G with adjacency matrix $A \in \{0,1\}^{n\times n}$ and attributes $\mathbf{f}_1, \dots, \mathbf{f}_n \in B_r^d$, we partition $[0,1]$ into intervals $I_1, \dots, I_n$ of length $1/n$, set $f(x) = \mathbf{f}_i$ for $x \in I_i$, and define $W(x,y) = A_{ij}$ for $(x,y) \in I_i \times I_j$. This representation turns node-wise aggregations and uniform node averages into integrals with respect to Lebesgue measure λ on $[0,1]$, and it makes explicit the sense in which our constructions for graphon-signals induce corresponding constructions for finite graphs.

A central technical ingredient is the weak* topology on spaces of probability measures. If K is a compact metric space, we write $\mathcal{P}(K)$ for its Borel probability measures. A sequence $\mu_n \in \mathcal{P}(K)$ converges weakly to $\mu \in \mathcal{P}(K)$ if

$$\int_K \varphi \, d\mu_n \to \int_K \varphi \, d\mu \qquad \text{for all } \varphi \in C(K).$$

On compact K, weak convergence is equivalent to weak* convergence under the duality $C(K)^* \simeq \mathcal{M}(K)$, and $\mathcal{P}(K)$ is compact in this topology (Prokhorov; tightness is automatic on compact sets). We will repeatedly apply this compactness when building iterated state spaces that include probability-measure components.

To metrize these measure spaces we use the Wasserstein–1 distance. For $\mu, \nu \in \mathcal{P}(K)$ and ground metric d on K, define

$$W_1(\mu, \nu; d) \;:=\; \inf_{\pi \in \Pi(\mu,\nu)} \int_{K\times K} d(x,y)\, d\pi(x,y),$$

where $\Pi(\mu,\nu)$ is the set of couplings of μ and ν. When K is compact, $W_1(\cdot,\cdot;d)$ is finite, and by Kantorovich–Rubinstein duality it admits the dual characterization

$$W_1(\mu, \nu; d) \;=\; \sup_{\|\varphi\|_{\mathrm{Lip}} \le 1} \Big| \int_K \varphi \, d\mu - \int_K \varphi \, d\nu \Big|.$$

In particular, on compact metric spaces the topology induced by W_1 coincides with the weak* topology on $\mathcal{P}(K)$. This will allow us to control changes in neighborhood distributions by testing against Lipschitz functions, which is exactly the continuity notion that appears in mean aggregation.

We consider message passing architectures whose aggregation is normalized by degree (mean aggregation). For a graphon-signal (W, f), the graphon

degree function is $\deg_W(x) := \int_0^1 W(x,y)\,dy$. A depth-$L$ mean-aggregation MPNN forms features $\mathbf{h}^{(t)}(x)$ by combining the current feature at x with the *average* feature of its neighbors, i.e., an expectation under the normalized kernel $W(x,\cdot)/\deg_W(x)$. The graph-level output is then obtained by a readout applied to the global average $\int_0^1 \mathbf{h}^{(L)}(x)\,dx$. The only property we require of the update and readout maps is Lipschitz continuity, since our goal is to connect stability and generalization to Lipschitz control under a suitably chosen metric.

Normalization by $\deg_W(x)$ forces a nondegeneracy condition. We therefore impose a *degree-floor* assumption: there exists $\alpha \in (0,1]$ such that $\deg_W(x) \geq \alpha$ for λ-a.e. x on the main domain. This serves three purposes. First, it makes the normalized neighbor kernel $W(x,\cdot)/\deg_W(x)$ well-defined almost everywhere, hence the corresponding neighbor distributions are genuine probability measures. Second, it prevents instabilities coming from dividing by a vanishing degree: without a lower bound, arbitrarily small perturbations of W near points of low degree can induce large changes in the normalized neighborhood law, obstructing continuity statements for both the network and the metric we will define. Third, it aligns the continuous and discrete settings: for an n-node graph, the analogous condition is a uniform lower bound on normalized degrees, $\deg(v)/n \geq \alpha$, which excludes isolated or near-isolated nodes in the dense regime. We will treat isolated nodes separately (by conventions or by restricting to the nonisolated subgraph), but for the core constructions and compactness arguments it is technically and conceptually cleaner to work under the degree floor.

These preliminaries set the stage for the hierarchical objects in the next section: we will encode the depth-t mean-aggregation information at each point x by a state consisting of a current representation together with the pushforward of the normalized neighbor measure, and we will compare the induced distributions of such states across graphon-signals using Wasserstein–1 at every level.

4 Mean-IDMs and Mean-DIDMs

We now formalize the hierarchical objects that encode the information propagated by mean aggregation. We begin with the compact state spaces

$$H_0 := B_r^d, \qquad H_{t+1} := H_t \times \mathcal{P}(H_t) \qquad (t \geq 0),$$

where $\mathcal{P}(H_t)$ denotes the Borel probability measures on H_t equipped with its Borel σ–algebra. For a graphon-signal $(W,f) \in \mathrm{WL}^d_{r,\alpha}$, we define the *mean neighborhood measure* at depth t by

$$\nu^{\mathrm{mean}}_{(W,f),t}(x) := \big(\gamma^{\mathrm{mean}}_{(W,f),t}\big)_{\#}\Big(\frac{W(x,\cdot)}{\deg_W(x)}\,\lambda\Big) \in \mathcal{P}(H_t),$$

and we define the corresponding *mean-IDM state* map $\gamma^{\mathrm{mean}}_{(W,f),t}\colon [0,1] \to H_t$ recursively by

$$\gamma^{\mathrm{mean}}_{(W,f),0}(x) := f(x), \qquad \gamma^{\mathrm{mean}}_{(W,f),t+1}(x) := \big(\gamma^{\mathrm{mean}}_{(W,f),t}(x),\, \nu^{\mathrm{mean}}_{(W,f),t}(x)\big).$$

Finally, the *mean-DIDM* at depth L is the pushforward of Lebesgue measure,

$$\Gamma^{\mathrm{mean}}_{(W,f),L} := \big(\gamma^{\mathrm{mean}}_{(W,f),L}\big)_{\#}\lambda \in \mathcal{P}(H_L).$$

Intuitively, $\gamma^{\mathrm{mean}}_{(W,f),t}(x)$ packages (i) the depth-t representation at x and (ii) the law of the depth-$(t-1)$ representations in its neighborhood, where "neighborhood" is sampled according to the normalized kernel $W(x,\cdot)/\deg_W(x)$; $\Gamma^{\mathrm{mean}}_{(W,f),L}$ is the distribution of these depth-L states across $x \sim \lambda$.

We record measurability, which is needed both to define the pushforwards and to later apply continuity arguments on spaces of measures.

Measurability. Let $\mathcal{P}(H_t)$ be endowed with its Borel σ–algebra induced by the weak* topology. Then each map $\gamma^{\mathrm{mean}}_{(W,f),t}\colon [0,1] \to H_t$ is measurable, and each $x \mapsto \nu^{\mathrm{mean}}_{(W,f),t}(x) \in \mathcal{P}(H_t)$ is measurable. The proof is by induction on t. The base case $t = 0$ is immediate since f is measurable. Assuming $\gamma_t := \gamma^{\mathrm{mean}}_{(W,f),t}$ is measurable, it suffices to show that $x \mapsto (\gamma_t)_{\#}\big(\frac{W(x,\cdot)}{\deg_W(x)}\lambda\big)$ is measurable as a map into $\mathcal{P}(H_t)$. By standard characterizations of the weak* Borel structure, it is enough to check measurability of

$$x \longmapsto \int_{H_t} \varphi(z)\, d\nu^{\mathrm{mean}}_{(W,f),t}(x)(z) = \int_0^1 \varphi\big(\gamma_t(y)\big)\, \frac{W(x,y)}{\deg_W(x)}\, dy$$

for every $\varphi \in C(H_t)$. The integrand $(x,y) \mapsto \varphi(\gamma_t(y))\, W(x,y)$ is measurable on $[0,1]^2$, hence by Fubini the map $x \mapsto \int \varphi(\gamma_t(y)) W(x,y)\, dy$ is measurable; dividing by $\deg_W(x)$ preserves measurability on the full-measure set where $\deg_W(x) > 0$. Since $\deg_W(x) \geq \alpha$ a.e. in $\mathrm{WL}^d_{r,\alpha}$, this yields measurability of $\nu^{\mathrm{mean}}_{(W,f),t}$ a.e., and thus of $\gamma_{t+1}(x) = (\gamma_t(x), \nu_t(x))$ as a map into the product space H_{t+1}.

Factorization of mean aggregation. The mean neighborhood measure is designed so that mean message passing depends only on $\gamma^{\mathrm{mean}}_{(W,f),t}$. Concretely, define $h_0\colon H_0 \to \mathbb{R}^{d_0}$ by $h_0 := \phi^{(0)}$. Given $h_t\colon H_t \to \mathbb{R}^{d_t}$, define $h_{t+1}\colon H_{t+1} \to \mathbb{R}^{d_{t+1}}$ by

$$h_{t+1}(z,\mu) := \phi^{(t+1)}\Big(h_t(z),\, \int_{H_t} h_t(z')\, d\mu(z')\Big), \qquad (z,\mu) \in H_t \times \mathcal{P}(H_t).$$

Then for λ-a.e. x and all $t \leq L$ we have the identity

$$\mathbf{h}^{(t)}(x) = h_t\big(\gamma^{\mathrm{mean}}_{(W,f),t}(x)\big).$$

Indeed, by construction of $\nu^{\mathrm{mean}}_{(W,f),t-1}(x)$ we have

$$\int_{H_{t-1}} h_{t-1}(z')\, d\nu^{\mathrm{mean}}_{(W,f),t-1}(x)(z') = \int_0^1 h_{t-1}\big(\gamma^{\mathrm{mean}}_{(W,f),t-1}(y)\big)\, \frac{W(x,y)}{\deg_W(x)}\, dy = \mathbb{E}\big[\mathbf{h}^{(t-1)}(Y)\big],$$

so the update rule matches the recursion for $h_t \circ \gamma_t$. This factorization will be the basis for our Lipschitz and universality statements once metrics on H_t and $\mathcal{P}(H_t)$ are fixed.

Isolated nodes and domain restriction. In the graphon setting we work primarily under the degree-floor hypothesis $\deg_W(x) \geq \alpha > 0$ a.e., so the normalization $W(x,\cdot)/\deg_W(x)$ and the probability measures $\nu^{\mathrm{mean}}_{(W,f),t}(x)$ are well-defined almost everywhere. If one wishes to treat general graphons without this assumption, the natural first step is to restrict the construction to the full-measure set $D_W := \{x : \deg_W(x) > 0\}$ (and correspondingly to the induced probability space $(D_W, \lambda(\cdot)/\lambda(D_W))$); one may then define γ_t and Γ_L on this restricted domain. In the finite-graph case, isolated vertices can be handled analogously by either removing them before forming the step graphon, or by introducing a convention (e.g. assigning an isolated node a degenerate neighbor law such as a Dirac mass at itself or at a distinguished dummy state). We will keep the degree floor for the main development and return to such conventions only when discussing algorithmic details.

Metrics on the mean-IDM state spaces. To compare mean-IDM states across different graphon-signals, we endow each H_t with a canonical metric that mirrors the recursive structure $H_{t+1} = H_t \times \mathcal{P}(H_t)$. We set

$$d^{\mathrm{mean}}_{\mathrm{IDM},0}(z,z') := \|z - z'\|_2, \qquad z, z' \in H_0 = B^d_r.$$

Assuming $d^{\mathrm{mean}}_{\mathrm{IDM},t}$ has been defined on H_t, we define for $(z,\mu),(z',\mu') \in H_{t+1}$

$$d^{\mathrm{mean}}_{\mathrm{IDM},t+1}\big((z,\mu),(z',\mu')\big) := d^{\mathrm{mean}}_{\mathrm{IDM},t}(z,z') \;+\; W_1\big(\mu,\mu'; d^{\mathrm{mean}}_{\mathrm{IDM},t}\big), \tag{1}$$

i.e. we take the ℓ_1 product metric between the base component $z \in H_t$ and the neighborhood-law component $\mu \in \mathcal{P}(H_t)$, where the latter is compared using Wasserstein–1 with ground metric $d^{\mathrm{mean}}_{\mathrm{IDM},t}$.

It is straightforward to verify that (1) defines a genuine metric on H_{t+1}: nonnegativity and symmetry are immediate; if $d^{\mathrm{mean}}_{\mathrm{IDM},t+1}((z,\mu),(z',\mu')) = 0$, then both $d^{\mathrm{mean}}_{\mathrm{IDM},t}(z,z') = 0$ and $W_1(\mu,\mu'; d^{\mathrm{mean}}_{\mathrm{IDM},t}) = 0$, hence $z = z'$ and $\mu = \mu'$; the triangle inequality follows by applying the triangle inequality separately to the two summands.

Wasserstein–1 on spaces of measures. Let (K,d) be a compact metric space. For $\mu,\nu \in \mathcal{P}(K)$ we write $\Pi(\mu,\nu)$ for the set of couplings on $K \times K$

with marginals μ and ν, and we define the balanced Wasserstein–1 distance

$$W_1(\mu,\nu;d) := \inf_{\pi\in\Pi(\mu,\nu)} \int_{K\times K} d(x,y)\, d\pi(x,y).$$

Since K is compact, d is bounded and the infimum is finite. Moreover, by Kantorovich–Rubinstein duality we have

$$W_1(\mu,\nu;d) = \sup\Big\{ \int_K \varphi\, d(\mu-\nu) \,:\, \varphi \in \mathrm{Lip}_1(K,d)\Big\}, \tag{2}$$

where $\mathrm{Lip}_1(K,d)$ denotes the set of real-valued 1–Lipschitz functions on K. In particular, $W_1(\cdot,\cdot;d)$ is a metric on $\mathcal{P}(K)$.

We apply this with $K = H_t$ and $d = d_{\mathrm{IDM},t}^{\mathrm{mean}}$, thereby obtaining a metric space $\big(\mathcal{P}(H_t), W_1(\cdot,\cdot;d_{\mathrm{IDM},t}^{\mathrm{mean}})\big)$ for every t. We emphasize that we always work with probability measures (as ensured by the degree normalization), so no unbalanced or reservoir construction is needed.

Metrization of weak* convergence. We endow $\mathcal{P}(H_t)$ with the weak* topology (equivalently, the topology of weak convergence of probability measures). On compact metric spaces, Wasserstein–1 induces exactly this topology. Concretely, for $\mu_n, \mu \in \mathcal{P}(H_t)$ we have

$$W_1\big(\mu_n,\mu;d_{\mathrm{IDM},t}^{\mathrm{mean}}\big) \to 0 \quad\Longleftrightarrow\quad \int_{H_t} \varphi\, d\mu_n \to \int_{H_t} \varphi\, d\mu \ \text{ for all } \varphi \in C(H_t).$$

One direction is immediate from (2): if $W_1(\mu_n,\mu) \to 0$ then the integrals converge for all Lipschitz φ, hence for all continuous φ by uniform approximation of continuous functions by Lipschitz functions on compact metric spaces. Conversely, weak* convergence on a compact space implies tightness automatically, and the collection of 1–Lipschitz test functions is uniformly bounded and equicontinuous; a standard compactness argument (or known theorems on Wasserstein metrics on compact spaces) yields $W_1(\mu_n,\mu) \to 0$. Thus, the Borel σ–algebra induced by $W_1(\cdot,\cdot;d_{\mathrm{IDM},t}^{\mathrm{mean}})$ agrees with the weak* Borel structure used in the measurability discussion above.

The mean-DIDM mover distance on graphon-signals. With $d_{\mathrm{IDM},L}^{\mathrm{mean}}$ fixed on H_L, we equip $\mathcal{P}(H_L)$ with the metric $W_1(\cdot,\cdot;d_{\mathrm{IDM},L}^{\mathrm{mean}})$ and define, for $(W,f),(V,g) \in \mathrm{WL}_{r,\alpha}^d$,

$$\delta_L^{\mathrm{mean}}\big((W,f),(V,g)\big) := W_1\Big(\Gamma_{(W,f),L}^{\mathrm{mean}},\, \Gamma_{(V,g),L}^{\mathrm{mean}}\,;\, d_{\mathrm{IDM},L}^{\mathrm{mean}}\Big).$$

By construction δ_L^{mean} is a pseudometric on $\mathrm{WL}_{r,\alpha}^d$: symmetry and the triangle inequality are inherited from W_1, and $\delta_L^{\mathrm{mean}}((W,f),(V,g)) = 0$ holds precisely when the depth-L mean-DIDMs coincide as measures on H_L. In particular, δ_L^{mean} is insensitive to modifications of (W,f) on λ–null sets. Moreover,

if $\sigma\colon [0,1] \to [0,1]$ is measure-preserving and we form the usual relabeling (W^σ, f^σ), then $\Gamma^{\mathrm{mean}}_{(W^\sigma,f^\sigma),L} = \Gamma^{\mathrm{mean}}_{(W,f),L}$ and hence $\delta_L^{\mathrm{mean}}((W,f),(W^\sigma,f^\sigma)) = 0$. We therefore pass to the metric identification $(W,f) \sim (V,g)$ iff $\delta_L^{\mathrm{mean}}((W,f),(V,g)) = 0$, and we view δ_L^{mean} as a bona fide metric on the quotient $\mathrm{WL}^d_{r,\alpha}/\!\sim$.

Compactness of the mean-IDM hierarchy. We first record that the recursive state spaces H_t and their associated measure spaces are compact for every fixed depth. The proof is an induction that mirrors the construction of the metric (1).

Proposition 4.1. *For every $t \in \mathbb{N}$, the metric space $\big(H_t, d^{\mathrm{mean}}_{\mathrm{IDM},t}\big)$ is compact. Consequently, $\big(\mathcal{P}(H_t), W_1(\cdot,\cdot; d^{\mathrm{mean}}_{\mathrm{IDM},t})\big)$ is compact as well.*

Proof. For $t = 0$ we have $H_0 = B^d_r$, which is compact in the Euclidean metric, hence also compact under $d^{\mathrm{mean}}_{\mathrm{IDM},0} = \|\cdot\|_2$.

Assume inductively that $\big(H_t, d^{\mathrm{mean}}_{\mathrm{IDM},t}\big)$ is compact. Since H_t is compact, every probability measure on H_t has finite first moment and the weak* topology on $\mathcal{P}(H_t)$ is metrized by $W_1(\cdot,\cdot; d^{\mathrm{mean}}_{\mathrm{IDM},t})$ (equivalently, W_1 induces exactly the topology of weak convergence on $\mathcal{P}(H_t)$). By Prokhorov compactness (or, on compact metric spaces, the direct sequential compactness of probability measures under weak convergence), it follows that $\mathcal{P}(H_t)$ is compact in the weak* topology, hence also compact in the metric $W_1(\cdot,\cdot; d^{\mathrm{mean}}_{\mathrm{IDM},t})$.

Finally, $H_{t+1} = H_t \times \mathcal{P}(H_t)$ is compact as a product of compact spaces, and the metric $d^{\mathrm{mean}}_{\mathrm{IDM},t+1}$ is a compatible product metric (indeed an ℓ_1 sum of the two component metrics). Thus H_{t+1} is compact. The conclusion for $\mathcal{P}(H_{t+1})$ is obtained by repeating the preceding argument with H_{t+1} in place of H_t. □

In particular, for our fixed depth L we may regard H_L and $\mathcal{P}(H_L)$ as compact metric spaces, and all Wasserstein distances appearing in the definition of δ_L^{mean} are finite and attain their infima.

Compactness of the mean-metric domain. We next explain why the metric identification space $\big(\mathrm{WL}^d_{r,\alpha}/\!\sim, \delta_L^{\mathrm{mean}}\big)$ is compact. The key point is that δ_L^{mean} is obtained by pushing each graphon-signal forward to the compact space $\mathcal{P}(H_L)$ and then measuring a Wasserstein distance there. Concretely, define the depth-L mean-DIDM map

$$\Theta_L : \mathrm{WL}^d_{r,\alpha} \to \mathcal{P}(H_L), \qquad \Theta_L(W,f) := \Gamma^{\mathrm{mean}}_{(W,f),L}.$$

By definition,

$$\delta_L^{\mathrm{mean}}\big((W,f),(V,g)\big) = W_1\Big(\Theta_L(W,f), \Theta_L(V,g); d^{\mathrm{mean}}_{\mathrm{IDM},L}\Big).$$

Thus Θ_L is constant on $\sim$-equivalence classes and induces an injective map $\overline{\Theta}_L$ on the quotient. Moreover, the preceding display shows that $\overline{\Theta}_L$ is an isometric embedding of $\big(\mathrm{WL}^d_{r,\alpha}/\!\sim, \delta_L^{\mathrm{mean}}\big)$ into $\big(\mathcal{P}(H_L), W_1(\cdot,\cdot; d^{\mathrm{mean}}_{\mathrm{IDM},L})\big)$. Consequently, to prove compactness of the quotient it suffices to show that the image $\overline{\Theta}_L(\mathrm{WL}^d_{r,\alpha}/\!\sim) = \Theta_L(\mathrm{WL}^d_{r,\alpha})$ is compact in $\mathcal{P}(H_L)$.

To this end we endow $\mathrm{WL}^d_{r,\alpha}$ with any of the standard compact topologies on graphon-signals, e.g. the "decorated cut metric" obtained by combining the cut metric on graphons with an L^1-type term for the signal and taking the infimum over measure-preserving relabelings. It is a known compactness theorem (extending Lovász–Szegedy compactness for graphons to compact mark spaces) that the ambient space of such bounded graphon-signals modulo relabelings is compact in this topology. The degree-floor constraint is closed under this topology: if $W_n \to W$ in cut metric, then $\deg_{W_n} \to \deg_W$ in $L^1([0,1])$, hence any a.e. lower bound $\deg_{W_n} \geq \alpha$ passes to the limit and yields $\deg_W \geq \alpha$ a.e. Therefore $\mathrm{WL}^d_{r,\alpha}$ is compact in the chosen topology.

It remains to note that Θ_L is continuous on $\mathrm{WL}^d_{r,\alpha}$. The proof again proceeds by induction on the depth. At depth 0, $\Theta_0(W,f) = f_\#\lambda \in \mathcal{P}(B^d_r)$, and for any coupling obtained by pairing the same $x \in [0,1]$ we have the bound

$$W_1\big((f_n)_\#\lambda, f_\#\lambda; \|\cdot\|_2\big) \leq \int_0^1 \|f_n(x) - f(x)\|_2 \, dx,$$

so convergence of the signal in L^1 implies convergence of Θ_0 in W_1. For the inductive step, write $\gamma_{n,t} := \gamma^{\mathrm{mean}}_{(W_n,f_n),t}$ and $\gamma_t := \gamma^{\mathrm{mean}}_{(W,f),t}$. The recursion defining γ_{t+1} combines the pointwise state $\gamma_t(x)$ with the neighbor law

$$\nu_t(x) = (\gamma_t)_\#\Big(\frac{W(x,\cdot)}{\deg_W(x)}\,\lambda\Big).$$

Under cut-metric convergence $W_n \to W$ and the degree floor $\deg_{W_n}, \deg_W \geq \alpha$, the normalized kernels $W_n(x,\cdot)/\deg_{W_n}(x)$ converge in L^1 for a.e. x, uniformly in the sense needed to control integrals against bounded test functions. Combining this with the inductive hypothesis that $\gamma_{n,t}$ is close to γ_t in the mean-IDM metric yields that $\nu_{n,t}(x)$ is close to $\nu_t(x)$ in $W_1(\cdot,\cdot; d^{\mathrm{mean}}_{\mathrm{IDM},t})$ for most x, and hence that $\gamma_{n,t+1}(x)$ is close to $\gamma_{t+1}(x)$ in $d^{\mathrm{mean}}_{\mathrm{IDM},t+1}$. Pushing forward by λ then gives $W_1(\Theta_L(W_n,f_n), \Theta_L(W,f)) \to 0$, i.e. continuity of Θ_L.

Since $\mathrm{WL}^d_{r,\alpha}$ is compact and Θ_L is continuous into the compact metric space $\mathcal{P}(H_L)$, the image $\Theta_L(\mathrm{WL}^d_{r,\alpha})$ is compact. As noted above, $\overline{\Theta}_L$ identifies $\mathrm{WL}^d_{r,\alpha}/\!\sim$ isometrically with this compact image, and therefore $\big(\mathrm{WL}^d_{r,\alpha}/\!\sim, \delta_L^{\mathrm{mean}}\big)$ is compact.

Lipschitz continuity of mean-aggregation MPNNs. We next verify that mean-aggregation MPNNs depend on a graphon-signal only through the

depth-L mean-IDM/DIDM representation, and that the resulting functionals are Lipschitz with respect to the metrics $d_{\mathrm{IDM},t}^{\mathrm{mean}}$ and δ_L^{mean}.

The key observation is that the update rule uses the normalized neighbor law $W(x,\cdot)/\deg_W(x)$ only through expectations. Hence, once we have lifted each point $x \in [0,1]$ to its mean-IDM state $\gamma_{(W,f),t}^{\mathrm{mean}}(x) \in H_t$, the next-layer feature is obtained by applying a deterministic continuous map to the pair consisting of the current state and its neighborhood measure component. Concretely, define maps $h_t : H_t \to \mathbb{R}^{d_t}$ recursively by

$$h_0(z) := \phi^{(0)}(z), \qquad z \in H_0 = B_r^d,$$

and, for $t \geq 0$ and $(\tau, \mu) \in H_{t+1} = H_t \times \mathcal{P}(H_t)$,

$$h_{t+1}(\tau,\mu) := \phi^{(t+1)}\Big(h_t(\tau),\, \int_{H_t} h_t(\xi)\, d\mu(\xi)\Big).$$

By construction of $\gamma_{(W,f),t+1}^{\mathrm{mean}}(x) = (\gamma_{(W,f),t}^{\mathrm{mean}}(x), \nu_{(W,f),t}^{\mathrm{mean}}(x))$ and the definition of $\nu_{(W,f),t}^{\mathrm{mean}}(x)$ as a pushforward of $W(x,\cdot)/\deg_W(x)$, we have, for a.e. x,

$$\int_{H_t} h_t(\xi)\, d\nu_{(W,f),t}^{\mathrm{mean}}(x)(\xi) = \int_0^1 h_t\big(\gamma_{(W,f),t}^{\mathrm{mean}}(y)\big)\, \frac{W(x,y)}{\deg_W(x)}\, dy = \mathbb{E}_{Y \sim W(x,\cdot)/\deg_W(x)}\big[h_t(\gamma_{(W,f),t}^{\mathrm{mean}}(Y))\big].$$

An induction on t therefore yields the factorization

$$\mathbf{h}^{(t)}(x) = h_t\big(\gamma_{(W,f),t}^{\mathrm{mean}}(x)\big) \qquad \text{for a.e. } x \in [0,1].$$

At the graph level, if we set $F : \mathcal{P}(H_L) \to \mathbb{R}^m$ by

$$F(\mu) := \psi\Big(\int_{H_L} h_L(\tau)\, d\mu(\tau)\Big),$$

then the MPNN output satisfies $\mathbf{H} = F(\Gamma_{(W,f),L}^{\mathrm{mean}})$. In particular, any mean-aggregation MPNN induces a continuous functional on $\mathcal{P}(H_L)$, and is invariant under measure-preserving relabelings because $\Gamma_{(W,f),L}^{\mathrm{mean}}$ is.

We now prove a Lipschitz bound in the mean-IDM metric. Let $\mathrm{Lip}(\phi^{(t)})$ and $\mathrm{Lip}(\psi)$ denote Lipschitz constants with respect to the Euclidean norms on their domains. We claim that each h_t is Lipschitz on H_t and that one may choose constants C_t satisfying a simple recursion. For $t = 0$, $h_0 = \phi^{(0)}$ and hence

$$\|h_0(z) - h_0(z')\|_2 \leq \mathrm{Lip}(\phi^{(0)})\, \|z - z'\|_2 = \mathrm{Lip}(\phi^{(0)})\, d_{\mathrm{IDM},0}^{\mathrm{mean}}(z,z').$$

Assume inductively that $\|h_t(\tau) - h_t(\tau')\|_2 \leq C_t\, d_{\mathrm{IDM},t}^{\mathrm{mean}}(\tau,\tau')$ for all $\tau, \tau' \in H_t$. For $(\tau,\mu),(\tau',\nu) \in H_{t+1}$, we estimate using Lipschitzness of $\phi^{(t+1)}$ and the triangle inequality,

$$\begin{aligned}
&\|h_{t+1}(\tau,\mu) - h_{t+1}(\tau',\nu)\|_2 \\
&\qquad \leq \mathrm{Lip}(\phi^{(t+1)})\Big(\|h_t(\tau) - h_t(\tau')\|_2 + \Big\| \int h_t\, d\mu - \int h_t\, d\nu \Big\|_2\Big).
\end{aligned}$$

The first term is controlled by the inductive hypothesis. For the second term, we apply the Kantorovich–Rubinstein bound componentwise: each coordinate of h_t is C_t-Lipschitz on $(H_t, d_{\mathrm{IDM},t}^{\mathrm{mean}})$, hence

$$\Big\| \int h_t \, d\mu - \int h_t \, d\nu \Big\|_2 \leq C_t \, W_1\big(\mu, \nu; d_{\mathrm{IDM},t}^{\mathrm{mean}}\big).$$

Combining and recalling that $d_{\mathrm{IDM},t+1}^{\mathrm{mean}}$ is the ℓ_1-sum of $d_{\mathrm{IDM},t}^{\mathrm{mean}}$ and the corresponding W_1 term, we obtain

$$\|h_{t+1}(\tau,\mu) - h_{t+1}(\tau',\nu)\|_2 \leq C_{t+1} \, d_{\mathrm{IDM},t+1}^{\mathrm{mean}}\big((\tau,\mu),(\tau',\nu)\big), \qquad C_{t+1} := \mathrm{Lip}(\phi^{(t+1)}) \, C_t.$$

Thus h_L is C_L-Lipschitz on H_L, with $C_L = \mathrm{Lip}(\phi^{(0)}) \prod_{t=1}^{L} \mathrm{Lip}(\phi^{(t)})$.

Finally, we bound the graph-level readout. For $\mu, \nu \in \mathcal{P}(H_L)$,

$$\|F(\mu) - F(\nu)\|_2 \leq \mathrm{Lip}(\psi) \Big\| \int h_L \, d\mu - \int h_L \, d\nu \Big\|_2 \leq \mathrm{Lip}(\psi) \, C_L \, W_1\big(\mu, \nu; d_{\mathrm{IDM},L}^{\mathrm{mean}}\big).$$

Applying this with $\mu = \Gamma_{(W,f),L}^{\mathrm{mean}}$ and $\nu = \Gamma_{(V,g),L}^{\mathrm{mean}}$ yields

$$\|\mathbf{H}(W,f) - \mathbf{H}(V,g)\|_2 \leq \mathrm{Lip}(\psi) \, C_L \, \delta_L^{\mathrm{mean}}\big((W,f),(V,g)\big),$$

which is the desired Lipschitz continuity of mean-aggregation MPNNs with respect to the mean-DIDM mover distance.

Separation and fine-grained expressivity. We now record the converse direction to the Lipschitz bound: bounded-Lipschitz mean-aggregation MPNNs are rich enough to distinguish distinct mean-IDM/DIDM objects, and, quantitatively, if *all* such MPNNs fail to distinguish two inputs then the mean-DIDM mover distance must be small.

For each depth $t \leq L$, let $\mathcal{A}_t$ denote the collection of scalar functions $u : H_t \to \mathbb{R}$ that can be realized as a single coordinate of h_t for some choice of Lipschitz update maps $\{\phi^{(s)}\}_{s \leq t}$ (allowing arbitrary widths, and allowing the coordinate projection at the output of h_t). By construction, $\mathcal{A}_t \subset C(H_t)$, since each h_t is continuous. Moreover, $\mathcal{A}_t$ is stable under linear combinations: if $u, v \in \mathcal{A}_t$ and $a, b \in \mathbb{R}$ then $au + bv \in \mathcal{A}_t$ by running the two networks in parallel (concatenating channels) and applying a linear post-combination in the final coordinate selection. Likewise, by increasing width and choosing the final coordinate as a product gate, we may realize pointwise products up to arbitrary uniform accuracy; equivalently, the uniform closure $\overline{\mathcal{A}_t}$ is a subalgebra of $C(H_t)$.

We claim that $\overline{\mathcal{A}_t}$ separates points of H_t for every $t \leq L$. The case $t = 0$ is immediate: $H_0 = B_r^d$ is a compact subset of $\mathbb{R}^d$, and coordinate projections (hence affine functionals) separate points, so $\overline{\mathcal{A}_0}$ separates points. For the inductive step, fix $t \geq 0$ and consider two distinct states $(\tau,\mu) \neq (\tau',\nu)$ in $H_{t+1} = H_t \times \mathcal{P}(H_t)$. If $\tau \neq \tau'$, then by the inductive hypothesis there exists

$u \in \overline{\mathcal{A}_t}$ with $u(\tau) \neq u(\tau')$, and composing with the projection $(\tau, \mu) \mapsto \tau$ yields a separator on H_{t+1}. If instead $\tau = \tau'$ but $\mu \neq \nu$, then by the Riesz representation theorem there exists some $v \in C(H_t)$ such that $\int v\, d\mu \neq \int v\, d\nu$. Since $\overline{\mathcal{A}_t}$ is an algebra separating points, Stone–Weierstrass implies $\overline{\mathcal{A}_t} = C(H_t)$, and hence we may take $v \in \overline{\mathcal{A}_t}$ (or approximate the chosen v uniformly). Finally, the map

$$(\tau, \mu) \longmapsto \int_{H_t} v(\xi)\, d\mu(\xi)$$

is exactly a mean-aggregation functional on the measure component, and can be implemented at depth $t+1$ by selecting the second argument of $\phi^{(t+1)}$ (up to an arbitrarily small uniform error if v is only approximated by elements of $\mathcal{A}_t$). This yields separation on H_{t+1}.

Turning to DIDMs, define the class of scalar functionals on $\mathcal{P}(H_L)$

$$\mathcal{B}_L := \Big\{ \mu \mapsto \int_{H_L} u(\tau)\, d\mu(\tau) \;:\; u \in \mathcal{A}_L \Big\} \subset C(\mathcal{P}(H_L)).$$

As above, $\overline{\mathcal{B}_L}$ is a subalgebra of $C(\mathcal{P}(H_L))$ containing constants. If $\mu \neq \nu$ in $\mathcal{P}(H_L)$, then there exists $u \in C(H_L)$ with $\int u\, d\mu \neq \int u\, d\nu$, and since $\overline{\mathcal{A}_L} = C(H_L)$ we may approximate u uniformly by elements of $\mathcal{A}_L$, implying that $\overline{\mathcal{B}_L}$ separates points of $\mathcal{P}(H_L)$. Stone–Weierstrass then yields $\overline{\mathcal{B}_L} = C(\mathcal{P}(H_L))$.

This qualitative separation can be upgraded to a quantitative converse in the Wasserstein metric. Fix $\varepsilon > 0$. Consider the set $\mathrm{Lip}_1(H_L)$ of 1-Lipschitz functions on $(H_L, d_{\mathrm{IDM},L}^{\mathrm{mean}})$ that are normalized, say by $u(\tau_0) = 0$ at a fixed basepoint $\tau_0 \in H_L$. By compactness of H_L, $\mathrm{Lip}_1(H_L)$ is compact in $(C(H_L), \|\cdot\|_\infty)$ (Arzelà–Ascoli), hence admits a finite η-net $\{u_k\}_{k=1}^K$. By Kantorovich–Rubinstein duality,

$$W_1(\mu, \nu; d_{\mathrm{IDM},L}^{\mathrm{mean}}) = \sup_{u \in \mathrm{Lip}_1(H_L)} \int_{H_L} u\, d(\mu - \nu).$$

Approximating the supremum by the finite net and then approximating each u_k uniformly by a realizable mean-MPNN feature (with Lipschitz constant bounded by some a priori C after rescaling) shows that there exist $C > 0$ and $\delta > 0$ such that: if for every depth-L mean-MPNN (ϕ, ψ) with overall Lipschitz bound $C_{\phi,\psi} \leq C$ we have

$$\big\| \psi\big(\textstyle\int h_L\, d\mu\big) - \psi\big(\textstyle\int h_L\, d\nu\big) \big\|_2 \leq \delta,$$

then necessarily $W_1(\mu, \nu; d_{\mathrm{IDM},L}^{\mathrm{mean}}) \leq \varepsilon$. Equivalently, failure of closeness in δ_L^{mean} is witnessed by some bounded-Lipschitz mean-MPNN readout. Applying this with $\mu = \Gamma_{(W,f),L}^{\mathrm{mean}}$ and $\nu = \Gamma_{(V,g),L}^{\mathrm{mean}}$ yields the asserted fine-grained expressivity statement at the level of graphon-signals.

Universal approximation. We next record the corresponding universality statement: mean-aggregation MPNNs are dense among continuous functionals on the mean-IDM state space and, after readout, among continuous functionals on the mean-DIDM space.

For each $t \le L$, recall the class $\mathcal{A}_t \subset C(H_t)$ of scalar features realizable (as a coordinate) by some depth-t mean-MPNN up to that layer. From the closure properties already noted (stability under linear combinations by parallelization, and stability under pointwise multiplication up to uniform approximation by increasing width and using standard approximation of products on compact sets), the uniform closure $\overline{\mathcal{A}_t}$ is a subalgebra of $C(H_t)$ containing constants. Since $\overline{\mathcal{A}_t}$ separates points of H_t, Stone–Weierstrass yields

$$\overline{\mathcal{A}_t} = C(H_t) \qquad \text{for every } t \le L.$$

In particular, for any continuous target $F \in C(H_L, \mathbb{R})$ and any $\varepsilon > 0$, there exists a realizable scalar feature $u \in \mathcal{A}_L$ such that $\|u - F\|_\infty \le \varepsilon$. Vector-valued approximation follows coordinatewise: for $F \in C(H_L, \mathbb{R}^m)$ we approximate each coordinate with an element of $\mathcal{A}_L$ and concatenate the resulting channels.

Passing from IDMs to DIDMs, define

$$\mathcal{B}_L := \Big\{ \mu \mapsto \int_{H_L} u(\tau)\, d\mu(\tau) \;:\; u \in \mathcal{A}_L \Big\} \subset C(\mathcal{P}(H_L), \mathbb{R}).$$

Again, $\overline{\mathcal{B}_L}$ is a subalgebra containing constants, and it separates points of $\mathcal{P}(H_L)$: if $\mu \neq \nu$, choose $u \in C(H_L)$ with $\int u\, d\mu \neq \int u\, d\nu$ and approximate u uniformly by elements of $\mathcal{A}_L$. Therefore Stone–Weierstrass gives

$$\overline{\mathcal{B}_L} = C(\mathcal{P}(H_L), \mathbb{R}).$$

Equivalently, for any continuous functional $\Phi \in C(\mathcal{P}(H_L), \mathbb{R})$ and any $\varepsilon > 0$, we may find a depth-L mean-MPNN feature map h_L and a continuous readout ψ (implemented, e.g., by an MLP on the compact range of $\int h_L\, d\mu$) such that

$$\sup_{\mu \in \mathcal{P}(H_L)} \Big| \psi\Big(\int h_L\, d\mu \Big) - \Phi(\mu) \Big| \le \varepsilon.$$

Thus mean-MPNNs are universal approximators of continuous functionals on $\mathcal{P}(H_L)$ when we allow arbitrary widths and unrestricted continuous readouts. (When one additionally requires global Lipschitz bounds on (ϕ, ψ), one obtains a restricted approximation statement relevant to the quantitative separation and generalization results, but the basic density statement is as above.)

We now transfer universality from $\mathcal{P}(H_L)$ back to graphon-signals. Consider the continuous embedding

$$\Theta : \mathrm{WL}^d_{r,\alpha} \to \mathcal{P}(H_L), \qquad \Theta(W, f) := \Gamma^{\mathrm{mean}}_{(W,f),L}.$$

By construction of δ_L^{mean}, if $\delta_L^{\mathrm{mean}}((W,f),(V,g))=0$ then $\Theta(W,f)=\Theta(V,g)$, so Θ descends to an injective map on the metric identification space. Let

$$\mathcal{X}_L := \Theta(\mathrm{WL}_{r,\alpha}^d) \subset \mathcal{P}(H_L),$$

which is compact as a continuous image of a compact domain. Every continuous functional F on $(\mathrm{WL}_{r,\alpha}^d/\!\sim, \delta_L^{\mathrm{mean}})$ corresponds uniquely to a continuous functional $\widetilde{F} \in C(\mathcal{X}_L)$ via $F = \widetilde{F} \circ \Theta$. Since $\overline{\mathcal{B}_L} = C(\mathcal{P}(H_L))$, its restriction is dense in $C(\mathcal{X}_L)$. Hence for every $\varepsilon > 0$ there exists a depth-L mean-MPNN (with some choice of widths and parameters) such that

$$\sup_{(W,f)\in \mathrm{WL}_{r,\alpha}^d} \Big|\mathrm{MPNN}(W,f) - F(W,f)\Big| \le \varepsilon,$$

where $\mathrm{MPNN}(W,f)$ denotes the induced graph-level output $\psi(\int_0^1 \mathbf{h}^{(L)}(x)\,dx)$ and we implicitly identify F with its $\sim$-invariant representative.

Finally, the same approximation statement applies to finite attributed graphs in the dense regime via the standard step-function embedding. Given an n-node attributed graph $(G,\mathbf{f})$, we form the induced step graphon-signal (W_G, f_G) by partitioning $[0,1]$ into n equal intervals and taking W_G and f_G constant on the corresponding blocks. This embedding is compatible with the mean-IDM recursion and preserves the mean-DIDM representation, so restricting the above universal family of mean-MPNNs on $\mathrm{WL}_{r,\alpha}^d$ to the embedded class of step graphons yields uniform approximation of any δ_L^{mean}-continuous functional on that subclass. In particular, for graph sequences converging in δ_L^{mean} to a graphon-signal limit, the same approximants provide asymptotically accurate predictors along the sequence.

Distribution-free generalization. We now record the corresponding distribution-free generalization statement for learning over our compact metric domains. Throughout, we consider a generic supervised setting in which inputs take values in a compact metric space $(\mathcal{X}, d)$ and labels lie in some measurable space $\mathcal{Y}$, with data distributed according to an arbitrary (unknown) probability law $\mathcal{D}$ on $\mathcal{X}\times\mathcal{Y}$. Given a predictor $F : \mathcal{X} \to \mathbb{R}^m$ and a loss $\ell : \mathbb{R}^m \times \mathcal{Y} \to [0,1]$, we write the population and empirical risks as

$$R(F) := \mathbb{E}_{(X,Y)\sim\mathcal{D}}\big[\ell(F(X),Y)\big], \qquad \widehat{R}_n(F) := \frac{1}{n}\sum_{i=1}^n \ell(F(X_i),Y_i),$$

for i.i.d. samples $(X_i,Y_i)_{i=1}^n \sim \mathcal{D}$.

Fix constants $L_F, L_\ell > 0$ and consider a hypothesis class $\mathcal{H}$ consisting of predictors $F : \mathcal{X} \to \mathbb{R}^m$ that are L_F-Lipschitz with respect to d and the Euclidean norm, and a loss ℓ that is L_ℓ-Lipschitz in its first argument, i.e.,

$$\|F(x) - F(x')\|_2 \le L_F\, d(x,x'), \qquad |\ell(z,y) - \ell(z',y)| \le L_\ell\, \|z - z'\|_2,$$

for all $x, x' \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z, z' \in \mathbb{R}^m$. Let $\kappa(\varepsilon) := N(\varepsilon, \mathcal{X}, d)$ denote the ε-covering number of $(\mathcal{X}, d)$.

A standard covering-number argument (discretize $\mathcal{X}$ by an ε-net, apply Hoeffding's inequality at net points, union bound over $\kappa(\varepsilon)$ points, and then extend from net points to all x by Lipschitzness) yields the following uniform bound: for every $\varepsilon > 0$ and $p \in (0, 1)$, with probability at least $1 - p$,

$$\sup_{F \in \mathcal{H}} \left| R(F) - \widehat{R}_n(F) \right| \;\le\; 2L_\ell L_F \, \varepsilon \;+\; \sqrt{\frac{\log\big(2\kappa(\varepsilon)\big) + \log(1/p)}{2n}}.$$

In particular, since $\mathcal{X}$ is compact we have $\kappa(\varepsilon) < \infty$ for every $\varepsilon > 0$, and by choosing $\varepsilon = \varepsilon(n) \to 0$ sufficiently slowly we obtain $\sup_{F \in \mathcal{H}} |R(F) - \widehat{R}_n(F)| \to 0$ as $n \to \infty$ without any dependence on parameter counts or architectural widths; the only hypothesis-class control is through global Lipschitz constants and the metric entropy of the input domain.

We apply this template to two choices of $\mathcal{X}$. First, take $\mathcal{X} = \mathcal{P}(H_L)$ equipped with $W_1(\cdot, \cdot; d^{\text{mean}}_{\text{IDM},L})$. By compactness of H_L, the space $\mathcal{P}(H_L)$ is compact under W_1, hence admits finite coverings at every scale. Second, take $\mathcal{X} = (\mathrm{WL}^d_{r,\alpha}/\!\sim, \delta^{\text{mean}}_L)$, which is compact by our earlier result. In both cases, Theorem 4 places any depth-L mean-aggregation MPNN with prescribed Lipschitz constants into a globally Lipschitz hypothesis class: there exists $L_F = C_{\phi,\psi}$ such that the induced graph-level predictor

$$(W, f) \longmapsto \psi\Big(\int_0^1 \mathbf{h}^{(L)}(x)\, dx \Big)$$

is L_F-Lipschitz with respect to δ^{mean}_L, and likewise the induced functional $\mu \mapsto \psi(\int h_L \, d\mu)$ is L_F-Lipschitz with respect to W_1 on $\mathcal{P}(H_L)$. Consequently, any uniformly Lipschitz-bounded family of such networks enjoys distribution-free uniform convergence at a rate governed by $\kappa(\varepsilon)$ of the corresponding compact metric domain.

We emphasize that the preceding bound is only as sharp as the available control of metric entropy. While compactness guarantees finiteness of $\kappa(\varepsilon)$, obtaining explicit rates in terms of (d, r, L, α) is delicate because the natural state spaces H_t involve iterated probability-measure components. Even for $\mathcal{P}(K)$ with $K \subset \mathbb{R}^D$ and Wasserstein-1, quantitative entropy bounds are nontrivial and depend sensitively on geometric regularity (e.g. doubling dimension) of K and on whether one restricts to measures with additional structure. In our setting, one may derive crude upper bounds by approximating an arbitrary $\mu \in \mathcal{P}(H_L)$ by finitely supported measures on an ε-net of H_L and quantizing weights; such bounds are sufficient for consistency but typically yield poor dependence on ε and may scale unfavorably with depth due to the recursive construction of H_L. Any substantial tightening would require additional assumptions (for instance, regularity or low-complexity

structure in (W, f), or restricting attention to subsets of $\mathcal{P}(H_L)$ with controlled effective dimension), and we leave a precise analysis of the metric entropy of $(\mathrm{WL}^d_{r,\alpha}/\sim, \delta_L^{\mathrm{mean}})$ and of its embedding into $\mathcal{P}(H_L)$ as an open direction.

Computation on finite graphs. We next record an explicit dynamic program for computing δ_L^{mean} between two finite attributed graphs, under the same mean-normalization that appears in the graphon definition. Let $(G, \mathbf{f})$ and $(H, \mathbf{g})$ be undirected graphs with node sets $[n]$ and $[m]$, respectively, and node attributes $\mathbf{f}_i, \mathbf{g}_j \in B_r^d$. We assume a normalized degree lower bound in the dense regime, namely $\deg_G(i)/n \geq \alpha$ for all (or all but a negligible fraction of) $i \in [n]$ and $\deg_H(j)/m \geq \alpha$ for all $j \in [m]$. As usual, we view each finite graph as inducing a step graphon-signal (W_G, f_G) by partitioning $[0, 1]$ into n (resp. m) equal intervals and setting W_G to be constant on each rectangle according to adjacency, and f_G to be constant on each interval according to $\mathbf{f}_i$. With this convention, $\delta_L^{\mathrm{mean}}((G, \mathbf{f}), (H, \mathbf{g}))$ is defined as $\delta_L^{\mathrm{mean}}((W_G, f_G), (W_H, f_H))$.

The main observation is that the recursive product metric $d_{\mathrm{IDM},t}^{\mathrm{mean}}$ and the Wasserstein evaluations in the mean-IDM definition translate into a recursion over node pairs. Concretely, we compute a sequence of cost matrices $D_t \in \mathbb{R}_+^{n \times m}$, where $D_t[i, j]$ represents the depth-t mean-IDM distance between node i in G and node j in H (for the corresponding step graphons). The base level is the attribute metric,

$$D_0[i, j] \;:=\; \|\mathbf{f}_i - \mathbf{g}_j\|_2.$$

For the inductive step, we form for each node $i \in [n]$ the *normalized neighbor distribution* $p_i \in \mathcal{P}([n])$ given by

$$p_i(u) \;:=\; \begin{cases} \frac{1}{\deg_G(i)} & \text{if } u \in N_G(i), \\ 0 & \text{otherwise,} \end{cases}$$

and similarly $q_j \in \mathcal{P}([m])$ for each $j \in [m]$. Since the mean construction uses probability measures (rather than sub-probabilities), we solve *balanced* optimal transport problems between p_i and q_j with ground costs inherited from the previous layer. Writing $\Pi(p_i, q_j)$ for the set of couplings with marginals p_i and q_j, define

$$\mathrm{OT}_t[i, j] \;:=\; \min_{\pi \in \Pi(p_i, q_j)} \sum_{u \in [n]} \sum_{v \in [m]} \pi(u, v)\, D_{t-1}[u, v].$$

We then update

$$D_t[i, j] \;:=\; D_{t-1}[i, j] \;+\; \mathrm{OT}_t[i, j], \qquad t = 1, \ldots, L.$$

Finally, the graph-level distance is obtained by transporting uniform mass across node sets using D_L as the ground cost:

$$\delta_L^{\text{mean}}((G,\mathbf{f}),(H,\mathbf{g})) \;:=\; \min_{\pi\in\Pi(\text{Unif}_n,\text{Unif}_m)} \sum_{i\in[n]}\sum_{j\in[m]} \pi(i,j)\, D_L[i,j].$$

Correctness follows by an induction that mirrors the graphon definitions. At depth $t = 0$, $D_0[i,j]$ agrees with the $H_0 = B_r^d$ ground metric. Assuming $D_{t-1}[u,v] = d_{\text{IDM},t-1}^{\text{mean}}(\gamma_{(G,\mathbf{f}),t-1}^{\text{mean}}(u), \gamma_{(H,\mathbf{g}),t-1}^{\text{mean}}(v))$, the measures p_i and q_j are precisely the discrete versions of the normalized neighbor laws $W(x,\cdot)/\deg_W(x)$, and the optimal value $\text{OT}_t[i,j]$ is exactly the Wasserstein–1 term

$$W_1\Big(\nu_{(G,\mathbf{f}),t-1}^{\text{mean}}(i),\, \nu_{(H,\mathbf{g}),t-1}^{\text{mean}}(j)\,;\, d_{\text{IDM},t-1}^{\text{mean}}\Big)$$

computed on the discrete supports, with ground costs given by D_{t-1}. Adding $D_{t-1}[i,j]$ implements the product metric on $H_t = H_{t-1} \times \mathcal{P}(H_{t-1})$. The final transport between Unif_n and Unif_m computes the Wasserstein distance between the pushforward empirical measures of the depth-L node states, i.e. the finite-graph analogue of $W_1(\Gamma_{(W,f),L}^{\text{mean}}, \Gamma_{(V,g),L}^{\text{mean}})$.

We also note a minor convention for isolated nodes. Under the degree-floor assumption such nodes do not appear, but if one wishes to extend the definition one may set p_i to be δ_i (a self-loop measure) or introduce a dedicated dummy symbol absorbing the mass; either choice yields a well-defined recursion and preserves the interpretation as a mean-normalized neighborhood law.

For complexity, let $N := \max\{n,m\}$. Each layer t requires solving $nm = O(N^2)$ balanced OT instances, each on supports of size at most N with ground costs given by a submatrix of D_{t-1}. Using a standard exact solver for discrete Wasserstein–1 (equivalently, a min-cost flow formulation on a bipartite network), one may solve each instance in $O(N^3 \log N)$ time in the worst case, yielding total time

$$O\big(L\, N^2 \cdot N^3 \log N\big) \;=\; O(L\, N^5 \log N),$$

and $O(N^2)$ memory to store the matrices D_t (plus solver overhead). This bound is conservative but suffices to show polynomial-time exact computability for fixed depth L.

From a practical standpoint, the exact min-cost flow step is the computational bottleneck. A natural acceleration is to replace each exact OT call by an entropic-regularized approximation (Sinkhorn) or a multiscale OT routine, reducing per-instance cost substantially at the expense of an additive error that must be propagated through the $t = 1,\ldots,L$ recursion. Since such an approximation analysis is orthogonal to the structural results above and depends on stability properties of the recursion, we leave it as future work.

Experiments. Although our results are structural and do not rely on empirical validation, a small set of experiments would clarify how the metric δ_L^{mean} behaves on finite graphs and how tightly it tracks perturbations of mean-aggregation GNNs in practice. Throughout, we recommend using the exact dynamic program above when feasible (moderate N and small L), and otherwise using an entropic-regularized approximation with a fixed regularization parameter and reporting the induced approximation bias separately.

A first experiment is a direct *stability–correlation* test aligned with the Lipschitz theorem. Fix a depth-L mean-aggregation MPNN $M = (\phi, \psi)$ (either randomly initialized with prescribed operator norms or trained on a downstream task), and consider a family of perturbed graphs $\{(G^{(k)}, \mathbf{f}^{(k)})\}_k$ generated from a base instance $(G, \mathbf{f})$ by controlled perturbations. For each pair (k, ℓ), compute the output discrepancy $\Delta_{k,\ell} := \|M(G^{(k)}, \mathbf{f}^{(k)}) - M(G^{(\ell)}, \mathbf{f}^{(\ell)})\|_2$ and the metric value $\delta_{k,\ell} := \delta_L^{\mathrm{mean}}((G^{(k)}, \mathbf{f}^{(k)}), (G^{(\ell)}, \mathbf{f}^{(\ell)}))$. We then evaluate (i) rank correlations (Spearman/Kendall) between $\Delta_{k,\ell}$ and $\delta_{k,\ell}$ over a large sample of pairs, and (ii) the empirical Lipschitz slope $\widehat{C} := \max_{k,\ell:\, \delta_{k,\ell}>0} \Delta_{k,\ell}/\delta_{k,\ell}$ together with robust summaries such as the 0.95-quantile of the same ratio. This makes the theorem operational: small δ_L^{mean} should systematically coincide with small output changes, and $\widehat{C}$ provides a data-dependent calibration of the bound.

For perturbation models, we recommend three regimes. (1) *Attribute noise:* add bounded noise $\mathbf{f}_i \mapsto \mathbf{f}_i + \xi_i$ with $\|\xi_i\|_2 \le \eta$ (followed by projection to B_r^d if needed), which should yield an approximately linear response in δ_0^{mean} and propagate to δ_L^{mean}. (2) *Edge rewiring at fixed degree:* perform random edge swaps that preserve degrees; since neighborhood measures are degree-normalized, this probes sensitivity to changes in the *composition* of neighborhoods rather than to degree drift. (3) *Mild sparsification:* delete each edge independently with probability ρ and condition on maintaining the degree-floor on most nodes (or explicitly prune nodes violating $\deg(v)/|V| \ge \alpha$). This tests how the metric and the GNN outputs respond when the dense assumption is only approximately met.

A second experiment is a *comparison against alternative distances* that target different aggregation conventions. The most direct baseline is the (unbalanced) DIDM distance δ_L^{DIDM} from the normalized-sum setting, where neighbor measures are sub-probabilities with total mass proportional to degree; empirically, we expect δ_L^{DIDM} to be more sensitive to global degree scaling, whereas δ_L^{mean} should be comparatively invariant to degree blow-ups that do not change neighbor *frequencies*. Additional baselines include 1-WL-type distances (e.g. comparing color histograms across iterations), as well as transport-based graph distances such as TMD (Tree Mover's Distance) when available. For each baseline, we recommend repeating the stability–correlation evaluation with the same MPNN outputs, and additionally testing k-NN retrieval/classification where graphs are embedded only through

pairwise distances (no learned encoder) to determine whether δ_L^{mean} yields a competitive geometry for task-relevant similarity.

A third experiment focuses on *degree-variation calibration*. Since δ_L^{mean} is defined with normalized neighbor laws, it is natural to test families of graphs that have identical conditional neighbor distributions but varying degrees (e.g. blow-up constructions, or duplicating each node into s twins with identical adjacency proportions and identical attributes). In this setting, mean-aggregation MPNNs are expected to behave stably, and δ_L^{mean} should remain small, while distances that encode degree mass (including unbalanced OT constructions) may grow with s. Conversely, if one perturbs degrees without changing edge densities uniformly (introducing hubs or heavy-tailed degree patterns), one can measure whether δ_L^{mean} remains predictive of mean-MPNN output changes and where it begins to fail as the degree-floor is violated.

Finally, we recommend reporting computational behavior. For exact computation, record wall-clock times as a function of N and L, and verify the expected scaling dominated by the $O(N^2)$ OT calls per layer. For approximate OT (Sinkhorn), report the stability of the resulting approximate δ_L^{mean} under changes in the entropic regularization and the number of iterations, and quantify how approximation error affects the empirical slope $\widehat{C}$. These measurements would complement the theoretical polynomial-time claim by clarifying the practical operating range and motivating approximation schemes without conflating them with the definition of the metric itself.

Discussion and limitations. A central structural assumption throughout is the degree floor $\deg_W(x) \geq \alpha$ a.e. (and its finite-graph analogue). Conceptually, this condition ensures that the normalized neighbor law $W(x,\cdot)/\deg_W(x)$ is well-defined and does not exhibit arbitrarily large sensitivity under small perturbations of W near points where the degree is nearly zero. Technically, it is what allows us to treat the mean-aggregation operator as a uniformly Lipschitz map on the metric domain: the normalization by $\deg_W(x)$ is uniformly bounded by $1/\alpha$, and the neighbor measures are genuine probability measures with controlled dependence on the underlying graphon-signal. Without such a floor, even for bounded $W \in [0,1]$, the map $(W,f) \mapsto \nu^{\text{mean}}_{(W,f),t}(x)$ can become discontinuous on sets where $\deg_W(x)$ is small, and the Wasserstein couplings that underlie $d^{\text{mean}}_{\text{IDM},t}$ can be forced to pay arbitrarily large costs after normalization. In the finite setting, this is the familiar phenomenon that mean aggregation on nodes of tiny degree can behave erratically under a single-edge change; any metric intended to upper bound such changes must either incorporate the degree explicitly or exclude the low-degree regime.

There are several principled ways to weaken or remove the degree-floor

assumption, but each alters either the architecture class or the metric definition. A direct modification is *truncated normalization*: replace $\deg_W(x)$ by $\max\{\deg_W(x), \alpha\}$ in the definition of the neighbor law, producing a family of α-regularized mean-IDMs. This preserves the probabilistic interpretation when $\deg_W(x) \geq \alpha$ and gracefully transitions to a damped operator otherwise. A second option, closer to common practice in GNN design, is *teleportation* (or damping): for some $\tau \in (0,1)$ define a modified neighbor measure

$$\widetilde{\nu}_t(x) \;:=\; (1-\tau)\,(\gamma_t)_{\#}\Big(\tfrac{W(x,\cdot)}{\deg_W(x)}\,\lambda\Big) \;+\; \tau\,(\gamma_t)_{\#}\lambda,$$

with the convention that the first term is omitted when $\deg_W(x) = 0$. This makes the aggregator total-mass preserving and uniformly defined without assuming any degree floor, at the cost of injecting a global baseline message. One may also teleport to a self-loop by mixing with $\delta_{\gamma_t(x)}$, aligning more closely with residual/self-loop GNNs. In each case, the corresponding metric must be redefined by replacing $\nu^{\mathrm{mean}}_{(W,f),t}(x)$ with the chosen regularized neighbor law; the compactness and Lipschitz proofs then proceed with constants depending on τ rather than on α. What remains nontrivial is to re-establish point-separation results in the modified model, since teleportation reduces sensitivity to fine-grained neighborhood structure and may introduce additional identifications.

A related limitation is that our development is tailored to scalar edge weights $W(x,y) \in [0,1]$ and node attributes in B^d_r. Many applications involve edge features, temporal marks, or signed and directed interactions. The mean-IDM construction extends to these settings by enlarging the neighborhood state to include edge information before taking the pushforward: for instance, if an edge feature map $e : [0,1]^2 \to E$ is given for a compact metric feature space E, one can replace the neighbor pushforward by the law of $(\gamma_t(Y), e(x,Y))$ under the normalized neighbor sampling distribution. This leads to a modified recursion with $H_{t+1} = H_t \times \mathcal{P}(H_t \times E)$ (or another suitable product space), and the same Wasserstein-based metric construction applies on compact domains. The main trade-off is computational: the OT subproblems now live on an enlarged support and the cost matrix incorporates both node-state and edge-feature discrepancies.

Multi-relational or heterogeneous graphs can be handled similarly by allowing a collection of graphons $\{W^{(a)}\}_{a\in\mathcal{A}}$ (one per relation type) and defining either (i) a separate neighbor measure per relation, yielding $H_{t+1} = H_t \times \prod_{a\in\mathcal{A}} \mathcal{P}(H_t)$, or (ii) a single mixture law over (a, Y) with mixing weights determined by relation-specific degrees. Both choices correspond to standard architectural conventions (relation-wise message passing versus pooled message passing). Our arguments adapt provided $\mathcal{A}$ is finite and each relation satisfies an appropriate degree control (either per-relation or in aggregate). The universality and separation statements then depend on the richness of admissible updates across relations, and the metric inherits an additional

product structure over $\mathcal{A}$.

Finally, the sparse regime remains open in a substantive way. When graphs are sparse (e.g. average degree $O(1)$ or $o(N)$), the graphon formalism is no longer the natural limit object, degrees typically vanish under dense normalization, and mean aggregation becomes dominated by local sampling effects rather than by stable empirical neighborhood measures. One expects that an appropriate theory should be phrased in terms of local weak limits (graphings) or graphex processes, and that the correct notion of distance should compare rooted neighborhood distributions rather than global transport over $[0, 1]$. From the metric perspective, one can attempt to combine unbalanced optimal transport (to accommodate degree variability and mass deficit) with local neighborhood kernels, but the correct alignment with mean-aggregation GNNs is not yet clear. Establishing compactness, continuity of the IDM/DIDM map, and a polynomial-time exact computation procedure in this sparse setting appears to require new ideas beyond the present Wasserstein-over-probability-measures construction.