

A GENERATIVE MODEL FOR CONTROLLABLE FEATURE HETEROPHILY IN GRAPHS

Haoyu Wang¹, Renyuan Ma², Gonzalo Mateos³, and Luana Ruiz¹

¹ Dept. of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD, USA

² Dept. of Statistics and Data Science, Yale University, New Haven, CT, USA

³ Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA

ABSTRACT

We introduce a principled generative framework for graph signals that enables explicit control of feature heterophily, a key property underlying the effectiveness of graph learning methods. Our model combines a Lipschitz graphon-based random graph generator with Gaussian node features filtered through a smooth spectral function of the rescaled Laplacian. We establish new theoretical guarantees: (i) a concentration result for the empirical heterophily score; and (ii) almost-sure convergence of the feature heterophily measure to a deterministic functional of the graphon degree profile, based on a graphon-limit law for polynomial averages of Laplacian eigenvalues. These results elucidate how the interplay between the graphon and the filter governs the limiting level of feature heterophily, providing a tunable mechanism for data modeling and generation. We validate the theory through experiments demonstrating precise control of homophily across graph families and spectral filters.

Index Terms— graph generative models, homophily, graphons

1. INTRODUCTION

The success of many graph information processing problems, including node-level tasks in graph machine learning [1, 2] and network topology inference [3–5], hinges on the alignment between graph topology and node features, often summarized by the notion of homophily or heterophily.

We develop a generative framework for graphs and node features (i.e., graph signals) that allows explicit control of *feature heterophily* in the range from homophily to heterophily. Feature heterophily extends the usual notion of homophily (nodes with the same labels are more likely to establish relational ties) [6] to arbitrary feature vectors. To our knowledge, this is the first result characterizing feature heterophily in graphon-based generative models, showing it can be systematically controlled via interaction between a graphon and a graph filter. While the analysis is related to notions of label heterophily and signal smoothness, the novelty lies in modeling features as distinct graph signals—which may combine stationary components with feature-dependent mean shifts—rather than repeated realizations.

Our model combines graphs sampled from a graphon [7–10] with features generated by a stationary graph signal model, where white Gaussian noise is transformed by a polynomial graph filter [11]. From a signal processing (SP) perspective, our results provide a rigorous foundation for studying feature heterophily in graph-signal pairs, complementing literature focusing on label heterophily. By enabling explicit control of heterophily, the framework supports systematic study of effects on graph filtering and GNNs, of interaction with label homophily, and design of mitigation strategies when heterophily is detrimental.

Our theoretical contributions are twofold. First, we establish concentration of the empirical feature heterophily around a deterministic spectral quantity, showing that randomness in feature generation becomes negligible as the graph grows. Second, we prove almost-sure convergence of the heterophily score to a deterministic functional of the graphon degree profile, obtained through a graphon-limit law for polynomials of Laplacian eigenvalues. Together, these results reveal a simple and interpretable mechanism for heterophily control in graph generation via a combined graphon and stationary signal model. They further show that, in this combined model, heterophily is governed jointly by the graphon and the polynomial filter giving rise to the features. Notably, the dependence on the graphon is only through its degree function.

We validate both the concentration and convergence results with numerical experiments on synthetic graphons and filters, confirming the predicted behavior across graph families. Beyond validating the theory, these experiments highlight the potential practical impact of our framework as a controllable benchmark generator—by varying the underlying graphon and spectral filter, one can systematically produce graph–feature pairs spanning a wide range of homophily levels. In this sense, our approach plays a role analogous to synthetic benchmarking environments such as GraphWorld [12], but we distinctly offer a principled mechanism grounded in graphon limits and stationary signal models.

2. PRELIMINARIES

2.1. Graph and graph signals

A graph $G = (V, E)$ consists of two components: a set of vertices or nodes V , and a set of edges $E \subseteq V \times V$. Graphs can be categorized as either directed or undirected depending on the structure of their edge set E . A graph is undirected if and only if for any two nodes $u, v \in V$, $(u, v) \in E$ also implies $(v, u) \in E$, and both correspond to the same undirected edge. In this work, we restrict attention to undirected graphs.

Let $|V| = n$ be the number of nodes and $m = |E|$ the number of edges in G . The *adjacency matrix* is the $n \times n$ matrix \mathbf{A} with entries

$$\mathbf{A}_{ij} = \begin{cases} 1 & \text{if } (i, j) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

The *degree matrix* \mathbf{D} is the $n \times n$ diagonal matrix with $\mathbf{D}_{ii} = \sum_{j=1}^n \mathbf{A}_{ij}$. The *graph Laplacian* is then defined as $\mathbf{L} = \mathbf{D} - \mathbf{A}$. Since G is undirected, \mathbf{A} and \mathbf{L} are symmetric.

A *graph signal* is a vector $x \in \mathbb{R}^n$, where x_i corresponds to the signal value at node i [13, 14]. More generally, graph signals can have multiple features, represented as matrices $X \in \mathbb{R}^{n \times d}$, with d

denoting the feature dimension. In this paper we focus on a stationary signal model for graph signals originally introduced in [11]; see also [15].

Definition 1 (Stationary graph signals [11]). *Let $e_1, \dots, e_n \sim \mathcal{N}(0, I_d)$ be i.i.d. Gaussian vectors, and define the initial feature matrix $X_0 = \frac{1}{\sqrt{d}} [e_1, \dots, e_n]^\top \in \mathbb{R}^{n \times d}$. Let G_n be a graph with n nodes sampled from any random graph model. Denote $\mathbf{L}_n = \mathbf{D}_n - \mathbf{A}_n$ its unnormalized Laplacian, and define the rescaled Laplacian $\mathcal{L}_n = \mathbf{L}_n/n$. The observed feature matrix is then generated by*

$$X = \sum_{k=0}^{K-1} a_k \mathcal{L}_n^k X_0 = f(\mathcal{L}_n) X_0,$$

where $f(\cdot)$ is a so-called linear shift-invariant filter with $K-1$ coefficients.

This model is important for two main reasons. First, it reflects the idea that node features are not arbitrary but are *propagated by the structure of the sampled graph*: the polynomial filter $f(\mathcal{L}_n)$ spreads and smooths the initial white noise features X_0 according to the connectivity in G_n . Second, the parameterization by powers of \mathcal{L}_n allows us to control the level of feature smoothness (and thus feature homophily), linking the spectral properties of the Laplacian to the statistical properties of the features.

2.2. Graphons

A *graphon* is a symmetric, bounded, and measurable function $W : [0, 1]^2 \rightarrow [0, 1]$. Graphons serve both as generative models and as limits of dense graph sequences, capturing the asymptotic behavior of graphs as their size increases. Here we are primarily interested in the use of graphons as generative models. Given a graphon W , a random graph G_n with n nodes is sampled by drawing independent latent variables $u_1, \dots, u_n \sim \text{Unif}[0, 1]$, and forming edges between nodes i and j independently with probability $W(u_i, u_j)$. This yields a random graph $G_n \sim \mathcal{G}_W(n)$ whose structure reflects the properties of W .

An important class of graphons is given by *step-function graphons*, which correspond exactly to (dense) stochastic block models (SBMs). Such a graphon, denoted $W_{\mathcal{P}}$, is defined as

$$W_{\mathcal{P}}(x, y) = P_{ij} \quad \text{if } x \in I_i, y \in I_j,$$

where $\{I_1, \dots, I_r\}$ is a partition of $[0, 1]$ with $\mu(I_i) = n_i/n$. Step-function or SBM graphons are the canonical example of graphons of finite rank. Though the graphon acts as the limit of graphs with infinitely many nodes, it may still have low intrinsic dimensionality, which in the case of SBMs is given by the finite number of communities.

We state the following theorem, due to [16], proving convergence of the Laplacian spectrum of $G_n \sim \mathcal{G}_W(n)$.

Theorem 1 (Laplacian spectral convergence for $G_n \sim \mathcal{G}_W(n)$ [16]). *Let $G_n \sim \mathcal{G}_W(n)$ be sequence of random graphs. Denote its adjacency and degree matrices by \mathbf{A}_n and \mathbf{D}_n , respectively, and let $\mathbf{L}_n = \mathbf{D}_n - \mathbf{A}_n$ be the (unnormalized) graph Laplacian. Define the rescaled Laplacian as $\mathcal{L}_n = \mathbf{L}_n/n$. Then, as $n \rightarrow \infty$, the empirical spectral distribution of \mathcal{L}_n converges weakly (almost surely) to the spectral distribution of the graphon Laplacian operator*

$$(\mathcal{L}_W f)(x) = f(x) - \int_0^1 W(x, y) f(y) dy, \quad f \in L^2([0, 1]).$$

In particular, the eigenvalues of \mathcal{L}_n converge to the eigenvalues of \mathcal{L}_W .

This key result underscores many other important results such as the convergence of spectral operators parametrized by the graph Laplacian, like graph filters [17, 18] and graph neural networks [2, 19, 20], and the convergence of graph sampling sets [21, 22]. We will leverage it to prove consistency of the feature heterophily of graphs sampled from a graphon, which we discuss next.

2.3. Feature heterophily

Given a graph with a signal corresponding to a d -dimensional feature vector per node, its feature heterophily is defined as follows [23].

Definition 2 (Feature heterophily). *Given a graph $G_n = (V, E)$ with feature matrix $X \in \mathbb{R}^{n \times d}$, its feature heterophily is defined as*

$$h_{G_n} = \frac{1}{n} \text{Tr}(\mathcal{L}_n X X^\top), \quad (1)$$

where $\mathcal{L}_n = \mathbf{L}_n/n$ and $\mathbf{L}_n = \mathbf{D}_n - \mathbf{A}_n$ is the unnormalized Laplacian.

The notion of feature heterophily generalizes to arbitrary data the usual definition of homophily as a measure of the connection strength between nodes sharing the same categorical labels. To see this, note that \mathbf{L}_n can be written equivalently as $\mathbf{L}_n = \mathbf{B}^\top \mathbf{B}$, where $\mathbf{B} \in \mathbb{R}^{|E| \times n}$ is the signed incidence matrix of the graph. Substituting this into (1), we obtain $h_{G_n} = \frac{1}{n^2} \|\mathbf{B}X\|_F^2$. Since each row of $\mathbf{B}X$ corresponds to the difference $X_i - X_j$ for an edge $(i, j) \in E$, this expression simplifies to

$$h_{G_n} = \frac{1}{2n^2} \sum_{(i,j) \in E} \|X_i - X_j\|_2^2, \quad (2)$$

recovering an expression that depends on the dissimilarity of the signal values at the end-nodes of every edge.

Smaller values of the feature heterophily h_{G_n} correspond to *homophily*—connected nodes have similar features—while larger values correspond to *heterophily*—connected nodes have dissimilar features. Hence, h_{G_n} quantifies the alignment between the graph topology and the feature variation over the nodes.

3. THEORETICAL RESULTS

We consider stochastic graphs drawn from a graphon model, augmented with node features generated by a stationary signal model, and analyze how their interaction with the underlying graph shapes the feature heterophily of the network. Then, we study the consistency of feature heterophily across graphs produced by this combined model, supporting its use as a generative framework for controllable feature heterophily.

Heterophily concentration for fixed G_n . Given a fixed graph $G_n \sim \mathcal{G}_W(n)$ sampled from a graphon W , our first result establishes the concentration of the feature heterophily under the stationary signal model. Before stating it, we require a mild assumption on the number of features d .

Assumption 3.1 (Polynomial regime). *The number of nodes n and the feature dimension d grow proportionally such that $d^{1/\alpha} \leq n \leq d^\alpha$, for some fixed $\alpha > 1$.*

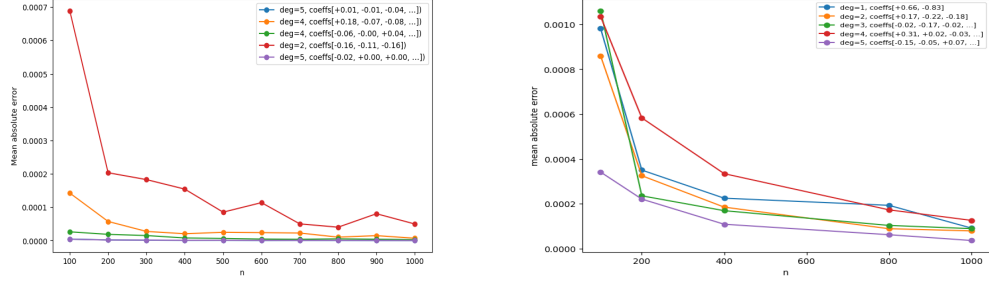


Fig. 1. (Left) Validation of Proposition 3.1: concentration of h_{G_n} around its expected limit. (Right) Validation of Theorem 2: convergence of h_{G_n} to its graphon limit.

This assumption is standard in high-dimensional asymptotic analysis and ensures well-behaved limits. The polynomial regime $d^{1/\alpha} \leq n \leq d^\alpha$ guarantees that the feature dimension and the number of nodes are comparable in scale, leading to deviations that shrink on the order of $1/\sqrt{N}$, up to logarithmic factors.

Proposition 3.1 (Concentration of feature heterophily). *Suppose Assumption 3.1 holds. Let \mathcal{L}_n be the rescaled Laplacian of a random graph (such as one sampled from a graphon model), independent of the Gaussian initialization X_0 . Let $N = \max\{n, d\}$. Then, for any $\epsilon > 0$, there exist a universal constant $C > 0$ and constants $C_P, c_\alpha > 0$ depending only on α and f such that*

$$\mathbb{P}\left(|h_{G_n} - n^{-1} \text{Tr}(f(\mathcal{L}_n)\mathcal{L}_n f(\mathcal{L}_n))| \geq \frac{C_P(\log N)^{1+\epsilon}}{\sqrt{N}}\right) \leq CN^{-c_\alpha}(\log N)^\epsilon.$$

Proof. The proof follows from Hanson-Wright inequality and is left empty for space reasons. \square

Proposition 3.1 has two main implications. First, it shows that the random heterophily score h_{G_n} , which depends on the Gaussian initialization X_0 , is tightly concentrated around its expectation¹, which is given by the deterministic spectral quantity $\mu_n = \frac{1}{n} \text{Tr}(f(\mathcal{L}_n)\mathcal{L}_n f(\mathcal{L}_n))$. Thus, the observed feature heterophily is essentially determined by the graph spectrum and the filter f in the stationary model.

Second, as $n \rightarrow \infty$, Proposition 3.1 implies (in soft- O notation) $|h_{G_n} - \mu_n| = \tilde{O}(\frac{1}{\sqrt{N}})$ with high probability, where \tilde{O} hides poly-logarithmic factors. In fact, we can say something stronger: since the deviation probabilities in Proposition 3.1 are summable in N , the Borel–Cantelli lemma implies that $\lim_{N \rightarrow \infty} h_{G_n} - \mu_n = 0$ almost surely. Therefore, on large-scale graphs h_{G_n} is eventually pinned to its deterministic spectral counterpart μ_n , providing a rigorous justification for analyzing homophily through the graphon limit. Since the convergence is not only in probability but also almost surely, the concentration result is essentially tight and robust for asymptotic analysis.

Heterophily convergence as $n \rightarrow \infty$. Next, we study the asymptotic behavior of the random quantity μ_n , which depends on the graphon model W , to establish feature heterophily convergence under the combined graphon and stationary signal model. From now on, we denote $\lambda_1^{(n)}, \dots, \lambda_n^{(n)}$ the non-decreasing Laplacian eigenvalues of $G_n \sim \mathcal{G}_W(n)$. We implicitly assume that the number of

¹Here the expectation is taken over the randomness in X_0 , while the graph G_n (and thus \mathcal{L}_n) stays fixed. The independence between G_n and X_0 is crucial, as it ensures that deviations of h_{G_n} come only from Gaussian randomness, which can be controlled using matrix concentration arguments.

features $d(n)$ is a function of n that satisfies the proportional regime from Assumption 3.1, and further, that the graphon W is Lipschitz as in Assumption 3.2 below.

Assumption 3.2 (Lipschitz graphon). *The graphon $W : [0, 1]^2 \rightarrow [0, 1]$ is symmetric and L -Lipschitz in the product metric. I.e., for all $(x, y), (x', y') \in [0, 1]^2$,*

$$|W(x, y) - W(x', y')| \leq L(|x - x'| + |y - y'|).$$

Theorem 2 (Almost-sure convergence of feature heterophily h_{G_n}). *Under Assumptions 3.1 and 3.2 and the hypotheses of Proposition 3.1, we have*

$$h_{G_n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \int_0^1 \delta(x) f(\delta(x))^2 dx,$$

where $\delta(x) = \int_0^1 W(x, y) dy$ is the degree function of W .

Proof. The proof relies on the following lemma.

Lemma 3.1. *For any polynomial filter P ,*

$$\frac{1}{n} \sum_{i=1}^n P(\lambda_i^{(n)}) \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \int_0^1 P(\delta(x)) dx.$$

Proof of Lemma 3.1. By linearity, it suffices to consider $P(t) = t^m$ with $m \geq 1$ (the case $m = 0$ is trivial). For any fixed $m \geq 1$, $\frac{1}{n} \sum_{i=1}^n (\lambda_i^{(n)})^m = \frac{1}{n} \text{Tr}(\mathcal{L}^m) = \frac{1}{n^{m+1}} \text{Tr}((D - A)^m)$. Expanding $(D - A)^m$, the all- D word contributes $\frac{1}{n^{m+1}} \text{Tr}(D^m) = \frac{1}{n} \sum_{i=1}^n \left(\frac{d_i}{n}\right)^m$. For any “mixed” word $w(D, A)$ of length m containing at least one A (where the number of A ’s is $r \geq 1$), cyclicity of the trace allows us to write $w = D^{\alpha_1} A D^{\alpha_2} \dots A D^{\alpha_r}$ with $\sum \alpha_s = m - r$. Expanding the trace yields $\text{Tr} w = \sum_{i_1, \dots, i_r=1}^n (d_{i_1}^{\alpha_1} A_{i_1 i_2} \dots A_{i_r i_1} d_{i_1}^{\alpha_r})$. Since $d_i \leq n$ and $A_{ij} \in \{0, 1\}$, each summand is bounded by n^{m-r} . Summing over the n^r tuples of indices gives $|\text{Tr} w(D, A)| \leq n^r \cdot n^{m-r} = n^m$. After dividing by n^{m+1} , all mixed words vanish as $O(1/n)$. Therefore

$$\frac{1}{n} \sum_{i=1}^n (\lambda_i^{(n)})^m = \frac{1}{n} \sum_{i=1}^n \left(\frac{d_i}{n}\right)^m + o(1). \quad (3)$$

Define $m_i := \sum_{j \neq i} W(x_i, x_j)$. We claim that

$$\max_{1 \leq i \leq n} \left| \frac{d_i}{n} - \delta(x_i) \right| \xrightarrow{\text{a.s.}} 0. \quad (4)$$

Conditional on $X = (x_j)$, d_i is a sum of independent Bernoullis with mean m_i . Hoeffding and a union bound yield

$$\mathbb{P}\left(\max_i \left| \frac{d_i}{n} - \frac{m_i}{n} \right| > \varepsilon \mid X\right) \leq 2n e^{-2\varepsilon^2 n}.$$

The RHS is summable and independent of X , hence by Borel–Cantelli, $\max_i \left| \frac{d_i}{n} - \frac{m_i}{n} \right| \rightarrow 0$ a.s.

Set $Z_n := \sup_{x \in [0,1]} \left| \frac{1}{n} \sum_{j=1}^n W(x, x_j) - \delta(x) \right|$. Fix $\varepsilon > 0$, let $\eta := \varepsilon/(4L)$, and let \mathcal{G} be a grid on $[0, 1]$ with mesh η and size $|\mathcal{G}| \leq 3/\eta$. For each x choose $u \in \mathcal{G}$ with $|x - u| \leq \eta$. Next we observe $|\delta(u) - \delta(x)| = \left| \int_0^1 (W(u, y) - W(x, y)) dy \right| \leq \int_0^1 L|u - x| dy = L|u - x|$.

By triangle inequality and the L -Lipschitz property of W in its first argument and of δ ,

$$\left| \frac{1}{n} \sum_{j=1}^n W(x, x_j) - \delta(x) \right| \leq \left| \frac{1}{n} \sum_{j=1}^n W(u, x_j) - \delta(u) \right| + 2L\eta. \quad (5)$$

Thus $Z_n \leq \max_{u \in \mathcal{G}} \left| \frac{1}{n} \sum_{j=1}^n W(u, x_j) - \delta(u) \right| + \frac{\varepsilon}{2}$. For fixed $u \in \mathcal{G}$, the variables $W(u, x_j) \in [0, 1]$ are i.i.d. with mean $\delta(u)$, so Hoeffding yields $\mathbb{P}\left(\left| \frac{1}{n} \sum_{j=1}^n W(u, x_j) - \delta(u) \right| > \frac{\varepsilon}{2}\right) \leq 2e^{-\frac{\varepsilon^2}{2} n}$.

A union bound over \mathcal{G} gives $\mathbb{P}(Z_n > \varepsilon) \leq \frac{24L}{\varepsilon} e^{-\frac{\varepsilon^2}{2} n}$, which is summable; hence $Z_n \rightarrow 0$ a.s. Moreover, $\left| \frac{m_i}{n} - \delta(x_i) \right| \leq Z_n + \frac{1}{n} \rightarrow 0$ a.s. Combining previous results yields (4). Finally, since $t \mapsto t^m$ is m -Lipschitz on $[0, 1]$,

$$\left| \frac{1}{n} \sum_{i=1}^n \left(\frac{d_i}{n}\right)^m - \frac{1}{n} \sum_{i=1}^n \delta(x_i)^m \right| \leq m \max_i \left| \frac{d_i}{n} - \delta(x_i) \right| \xrightarrow{\text{a.s.}} 0. \quad (6)$$

By the strong law of large numbers,

$$\frac{1}{n} \sum_{i=1}^n \delta(x_i)^m \xrightarrow{\text{a.s.}} \int_0^1 \delta(x)^m dx.$$

From (3), (6), and earlier equations, we conclude

$$\frac{1}{n} \sum_{i=1}^n (\lambda_i^{(n)})^m \xrightarrow{\text{a.s.}} \int_0^1 \delta(x)^m dx.$$

By linearity, the lemma follows for all polynomials P . \square

Using this lemma, Theorem 2 follows immediately by choosing $P(x) = x f(x)^2$, applying the concentration result from Proposition 3.1, and observing that $\frac{1}{n} \text{Tr}(f(\mathcal{L}_n) \mathcal{L}_n f(\mathcal{L}_n)) = \frac{1}{n} \sum_{i=1}^n \lambda_i^{(n)} f(\lambda_i^{(n)})^2$. \square

By establishing convergence of h_{G_n} on sequences of graphs converging to a graphon, Theorem 2 implies consistency of the feature heterophily across graphs and feature vectors generated from the combined graphon and stationary signal model. This provides theoretical support for using this combined model as a principled generative framework for graphs and graph signals that allows explicit control of homophily, which is key to several information processing problems on graphs, including node-level tasks in graph machine learning [1] and network topology inference [3].

Another key takeaway from this theorem is that the limiting feature heterophily, besides depending on the function f (a parameter of the stationary signal model), depends on the graphon –not

through the function W or its eigenvalues– but through its degree profile δ . This is a consequence of two properties of our combined model. First, stationarity implies the ability to jointly diagonalize the Laplacian and the feature covariance, so that h_{G_n} can be computed from the product of their eigenvalues. Second, due to the inherent edge density associated with graphons, the Laplacian eigenspectrum is localized around the node degrees [16].

Finally, we remark that by the spectral theorem applied to the matrix \mathcal{L} , the filter f is only evaluated in the range of the spectrum of \mathcal{L} , which is in $[0, 2]$. Thus, leveraging the fact that any continuous function can be written as a sum of polynomials uniformly over the closed interval, Theorem 2 can readily be extended to continuous filters over the closed interval $[0, 2]$.

We conclude with two examples particularizing Theorem 2 to an Erdős–Rényi (ER) and an SBM graphon respectively.

Example 3.1 (ER graphon). Let $W(x, y) = p$ for some $p \in (0, 1)$. Then $\delta(x) = p$ for all $x \in [0, 1]$, and

$$h_{G_n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} p f(p)^2.$$

Example 3.2 (SBM graphon). Consider an r -block SBM graphon with block proportions $\alpha_1, \dots, \alpha_r$ and probability matrix $[P_{ij}]$. The degree function is piecewise constant:

$$\delta(x) = \delta_i := \sum_{j=1}^r \alpha_j P_{ij}, \quad \text{for } x \in I_i.$$

Then, $h_{G_n} \xrightarrow[n \rightarrow \infty]{\text{a.s.}} \sum_{i=1}^r \alpha_i \delta_i f(\delta_i)^2$.

4. NUMERICAL RESULTS

In this section, we provide empirical validation of the main theoretical results developed above. Specifically, we verify both the concentration result (Proposition 3.1) and the almost-sure convergence of the feature heterophily (Theorem 2).

4.1. Experimental Setup

We sample graphs from a fixed Lipschitz graphon and generate node features according to the stationary signal model. For each graph size n , we compute the empirical heterophily score h_{G_n} and compare it with the theoretical targets described in our results. The validation is performed over multiple trials, with random polynomial filters f of varying degrees. All experiments use independent graph samples and Gaussian noise features.

4.2. Results

Figure 1 (left) illustrates the empirical validation of the concentration in Proposition 3.1. The plot shows that the deviation between h_{G_n} and its expectation shrinks as n grows, consistent with the theoretical $O\left(\frac{(\log N)^{1+\epsilon}}{\sqrt{N}}\right)$ bound.

Figure 1 (right) validates Theorem 2, showing that the heterophily score h_{G_n} converges almost surely to the deterministic integral $\int_0^1 \delta(x) f(\delta(x))^2 dx$ as $n \rightarrow \infty$. The error decays rapidly with n , in line with the theoretical prediction and the Borel–Cantelli argument ensuring almost-sure convergence. These experiments confirm that both the concentration bound and the almost-sure convergence result hold in practice, thus supporting the theoretical framework.

5. REFERENCES

- [1] Jiong Zhu, Yujun Yan, Lingxiao Zhao, Mark Heimann, Leman Akoglu, and Danai Koutra, "Beyond homophily in graph neural networks: Current limitations and effective designs," *Advances in neural information processing systems*, vol. 33, pp. 7793–7804, 2020.
- [2] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th Int. Conf. Learn. Representations*, Toulon, France, 24-26 Apr. 2017, Assoc. Comput. Linguistics.
- [3] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," *IEEE Signal Process. Mag.*, vol. 36, no. 3, pp. 16–43, May 2019.
- [4] S. Segarra, A. G. Marques, G. Mateos, and A. Ribeiro, "Network topology inference from spectral templates," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 3, no. 3, pp. 467–483, August 2017.
- [5] R. Shafipour, S. Segarra, A. G. Marques, and G. Mateos, "Identifying the topology of undirected networks from diffused non-stationary graph signals," *IEEE Open J. Signal Process.*, vol. 2, pp. 171–189, April 2021.
- [6] Miller McPherson, Lynn Smith-Lovin, and James M Cook, "Birds of a feather: Homophily in social networks," *Annual review of sociology*, vol. 27, no. 1, pp. 415–444, 2001.
- [7] L. Lovász, *Large Networks and Graph Limits*, vol. 60, American Mathematical Society, 2012.
- [8] C. Borgs and J. Chayes, "Graphons: A nonparametric method to model, estimate, and design algorithms for massive networks," in *Proc. of the 2017 ACM Conf. on Economics and Computation*, 2017, pp. 665–672.
- [9] J. Eldridge, M. Belkin, and Y. Wang, "Graphons, mergeons, and so on!," *Conf. Neural Inform. Process. Syst.*, vol. 29, 2016.
- [10] E. M. Airolidi, T. B. Costa, and S. H. Chan, "Stochastic block-model approximation of a graphon: Theory and consistent estimation," in *27th Conf. Neural Inform. Process. Syst.*, 2013, pp. 692–700.
- [11] A. G. Marques, S. Segarra, G. Leus, and A. Ribeiro, "Stationary graph processes and spectral estimation," *IEEE Trans. Signal Process.*, vol. 65, pp. 5911–5926, 2017.
- [12] John Palowitch, Anton Tsitsulin, Brandon Mayer, and Bryan Perozzi, "GraphWorld: Fake graphs bring real insights for gnns," in *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, 2022, pp. 3691–3701.
- [13] D. I Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 83–98, May 2013.
- [14] A. Ortega, P. Frossard, J. Kovačević, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proc. IEEE*, vol. 106, no. 5, pp. 808–828, 2018.
- [15] Nathanaël Perraudin and Pierre Vandergheynst, "Stationary signal processing on graphs," *IEEE Trans. Signal Process.*, vol. 65, no. 13, pp. 3462–3477, 2017.
- [16] R. Vizuete, F. Garin, and P. Frasca, "The Laplacian spectrum of large graphs sampled from graphons," *IEEE Trans. Netw. Sci. Eng.*, vol. 8, no. 2, pp. 1711–1721, 2021.
- [17] S. Segarra, A. G. Marques, G. Leus, and A. Ribeiro, "Interpolation of graph signals using shift-invariant graph filters," in *23rd Eur. Signal Process. Conf.*, Nice, France, 31 Aug.-4 Sep. 2015, IEEE.
- [18] J. Du, J. Shi, S. Kar, and J. M. F. Moura, "On graph convolution for graph CNNs," in *2018 IEEE Data Sci. Workshop*, Lausanne, Switzerland, 4-6 June 2018, pp. 239–243, IEEE.
- [19] F. Gama, A. G. Marques, G. Leus, and A. Ribeiro, "Convolutional neural network architectures for signals supported on graphs," *IEEE Trans. Signal Process.*, vol. 67, pp. 1034–1049, 2018.
- [20] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Conf. Neural Inform. Process. Syst.*, Barcelona, Spain, 5-10 Dec. 2016, NeurIPS Foundation.
- [21] Thien Le, Luana Ruiz, and Stefanie Jegelka, "A Poincaré inequality and consistency results for signal sampling on large graphs," in *The Twelfth International Conference on Learning Representations*, 2024.
- [22] A. Parada-Mayorga and A. Ribeiro, "Sampling and uniqueness sets in graphon signal processing," *IEEE Transactions on Signal Processing*, 2025.
- [23] Haolin Li, Haoyu Wang, and Luana Ruiz, "Graph sampling for scalable and expressive graph neural networks on homophilic graphs," in *Proceedings of the 2025 European Signal Processing Conference (EUSIPCO)*, Palermo, Italy, September 2025.