Dirichlet Meets Horvitz and Thompson: Estimating Homophily in Large Networks via Sampling

Hamed Ajorlou*, Gonzalo Mateos*, and Luana Ruiz†
*Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY
†Dept. of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD

***Abstract*—Assessing homophily in large-scale networks is central to understanding structural regularities in graphs, and thus inform the choice of models (such as graph neural networks) adopted to learn from network data. Evaluation of smoothness metrics requires access to the entire network topology and node features, which may be impractical in several large-scale, dynamic, resource-limited, or privacy-constrained settings. In this work, we propose a sampling-based framework to estimate homophily via the Dirichlet energy (Laplacian-based total variation) of graph signals, leveraging the Horvitz-Thompson (HT) estimator for unbiased inference from partial graph observations. The Dirichlet energy is a so-termed total (of squared nodal feature deviations) over graph edges; hence, estimable under general network sampling designs for which edge-inclusion probabilities can be analytically derived and used as weights in the proposed HT estimator. We establish that the Dirichlet energy can be consistently estimated from sampled graphs, and empirically study other heterophily measures as well. Experiments on several heterophilic benchmark datasets demonstrate the effectiveness of the proposed HT estimators in reliably capturing homophilic structure (or lack thereof) from sampled network measurements.**

***Index Terms*—Network Sampling, Dirichlet Energy, Horvitz-Thompson Estimator, Homophily, Graphon.**

I. Introduction

Homophily is a cardinal principle at the heart of several graph-based statistical learning tasks, including nearest-neighbor prediction, semi-supervised learning, and topology inference [1]–[3]. Accordingly, assessing homophily characteristics (i.e., the extent to which edges preferentially connect nodes with similar attributes or labels) of network datasets is central to understanding structural regularities in graphs, and, e.g., inform the choice of learning architectures such as graph neural networks (GNNs) [4]. In fact, a recent trend is to develop models for heterophilous data as well [2], [5], [6].

Problem statement. Consider a weighted and undirected graph $G = (\mathcal{V}, \mathcal{E})$, with $n = |\mathcal{V}|$ nodes, adjacency matrix $\mathbf{A} = [A_{ij}] \in \mathbb{R}_+^{n\times n}$, and combinatorial graph Laplacian $\mathbf{L} = \text{diag}(\mathbf{A1}) - \mathbf{A}$. Let $\mathbf{X}_n \in \mathbb{R}^{n\times f}$ be a graph signal matrix, where row vector $\mathbf{x}_i^\top$ collects the f features at node $i \in \mathcal{V}$. A workhorse homophily (i.e., signal smoothness) metric is the Dirichlet energy or Laplacian-based total variation defined as

$$\text{TV}_G(\mathbf{X}_n) := \text{trace}(\mathbf{X}_n^\top \mathbf{L}\mathbf{X}_n) = \sum_{(i,j)\in\mathcal{E}} A_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2, \quad (1)$$

where smaller $\text{TV}_G(\mathbf{X}_n)$ values are indicative of homophilous feature characteristics in the data. The evaluation of $\text{TV}_G(\mathbf{X}_n)$ relies on the implicit assumption that the graph dataset is observed in its entirety; however, it is often the case that relational information is only acquired from a portion of the complex system of interest. Network *sampling* is arguably the rule rather than the exception in statistical network analysis [1, Ch. 5], either due to uncontrollable factors behind the data acquisition process; or, by design in large-scale, dynamic, resource-limited, or privacy-constrained settings.

Suppose that instead of fully observing G, we take measurements that effectively produce a sampled graph $G^* = (\mathcal{V}^*, \mathcal{E}^*)$ and signals $\mathbf{X}_{n^*}$, $n^* \leq n$. Consequently, it will be typically impossible to exactly recover $\text{TV}_G(\mathbf{X}_n)$ from the partial information in G^*. Instead, we study the *problem* of developing useful homophily estimates of $\text{TV}_G(\mathbf{X}_n)$, say $\widehat{\text{TV}}$, from G^* and $\mathbf{X}_{n^*}$ under various sampling designs [1], [7], [8].

Contributions in context. Sampling has been widely studied in graph signal processing, but often with the objective of reconstructing bandlimited graph signals from limited nodal samples (yet G is known in its entirety) [9]. Recent efforts have explored efficient training of GNNs from intentionally sampled homophilous graphs [10]. Our distinct goal is to estimate a global structural characteristic of a network (here homophily) in an unbiased fashion, and with quantifiable performance.

In Section II, we start by establishing a fundamental property of the estimation problem, namely that the Dirichlet energy is a so-termed *testable* graph parameter under induced subgraph sampling. As introduced in [11], testability of a network summary statistic is tantamount to the existence of an estimator satisfying a weak form of consistency as the sample size increases. The upshot is that, for testable parameters, one would expect that simple (e.g., plug-in $\widehat{\text{TV}} = \text{TV}_{G^*}(\mathbf{X}_{n^*})$) estimators will be accurate under induced subgraph sampling. For our technical arguments, we extend the seminal characterizations of testability based solely on convergence of graph sequences to graph limits (i.e., graphons) [11], to also accommodate sequences of convergent graph signals [12]. Although the testability result has merit to ensure the estimation task is feasible under induced subgraph sampling, other sampling designs could be of interest as well. And it is not uncommon for these to produce unequal probability sampling of nodes or edges, challenging the reliability of *biased* plug-in estimators [13, Ch. 3]. To bridge this gap, in Section III we propose an homophily estimation framework by

leveraging the Horvitz-Thompson (HT) estimator for unbiased inference from incomplete graph measurements. The Dirichlet energy (1) is a *total* (of squared nodal feature deviations) over graph edges. Thus, estimable under general network sampling designs for which edge-inclusion probabilities can be analytically derived and used as weights in the proposed HT estimator. Interestingly, the variance of $\widehat{\mathrm{TV}}$ can be readily estimated from the sample G^* and $\mathbf{X}_{n^*}$.

We experimentally verify the unbiasedness of the HT estimator and examine how different sampling rates influence its variance (Section IV). To gain further insight into the behavior of the estimator across a range of practical scenarios, we compare multiple network sampling designs and heterophily measures beyond (1). All in all, the main contributions of this work can be summarized as follows:

- We establish that the Dirichlet energy is a testable graph parameter under induced subgraph sampling.
- We develop a novel HT estimator for unbiased estimation of homophily measures under general sampling designs.
- We conduct a comprehensive experimental evaluation using several heterophilic benchmark datasets.

II. Testability of the Dirichlet Energy

This section leverages results from the theory of graph limits to establish that the Dirichlet energy is a *testable* graph (signal) parameter. Following [11, Def. 2.11], a graph parameter φ is said to be testable if for every $\varepsilon > 0$ there exists a sample size $n^*(\epsilon)$ such that for any graph G with $n \geq n^*(\epsilon)$, an estimate $\widehat{\varphi}(G^*)$ computed from a uniformly sampled induced subgraph G^* satisfies

$$\mathrm{P}\left(|\varphi(G) - \widehat{\varphi}(G^*)| > \varepsilon \right) \leq \varepsilon.$$

Accordingly, testability of φ implies the existence of an estimator $\widehat{\varphi}(G^*)$ that satisfies a weak form of consistency.

A key result [11, Prop. 2.12(a)] asserts that φ is testable if and only if $\varphi(G_i)$ converges for every convergent graph sequence $G_1, G_2, \ldots$, which naturally leads to graphons and the theory of graph limits [11], [14]. Thus, a suitable notion of continuity of φ suffices. To establish the testability of the Dirichlet energy (a more general parameter than the graph signal-agnostic $\varphi(G)$ studied in [11]), we first recall the graph limit framework and explain how it naturally extends to continuous domain representations of graph signals [12].

A *graphon* is a symmetric measurable function $\mathbf{W} : [0,1]^2 \mapsto \mathbb{R}$ that arises as the limit of a sequence of dense graphs w.r.t. the cut metric. The cut norm, which underlies the notion of continuity in graph limit theory, is defined by

$$\|\mathbf{W}\|_\square := \sup_{S,T \subseteq [0,1]} \Big| \iint_{S \times T} \mathbf{W}(u,v)\, du\, dv \Big|.$$

It measures the maximum discrepancy of $\mathbf{W}$ over measurable cuts $S \times T$. Similarly, graph signals on finite graphs admit continuous domain counterparts on the graphon, represented as measurable functions $\mathbf{X} : [0,1] \mapsto \mathbb{R}^f$. Every finite graph–signal pair $(G, \mathbf{X}_n)$ admits a representation on a continuous domain via an associated step graphon and step signal. To make this construction explicit, partition $[0,1]$ into intervals

$$I_i = \left[\tfrac{i-1}{n}, \tfrac{i}{n} \right), \qquad i = 1, \ldots, n,$$

and define the associated step graphon and step signal

$$\mathbf{W}_G(u,v) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij}\, \mathbb{I}\{u \in I_i\} \mathbb{I}\{v \in I_j\}, \tag{2}$$

$$\mathbf{X}_G(u) = \sum_{i=1}^{n} \mathbf{x}_i\, \mathbb{I}\{u \in I_i\}, \tag{3}$$

where $\mathbb{I}\{\cdot\}$ is an indicator. Thus, for any $u \in I_i$ and $v \in I_j$,

$$\mathbf{W}_G(u,v) = A_{ij}, \qquad \mathbf{X}_G(u) = \mathbf{x}_i, \qquad \mathbf{X}_G(v) = \mathbf{x}_j.$$

In the limit of large graphs ($n \to \infty$), the step graphon and step signal converge to $(\mathbf{W}, \mathbf{X})$ [12]. Now, the Dirichlet energy can be naturally extended to the graphon functional

$$\Phi(\mathbf{W}, \mathbf{X}) = \iint_{[0,1]^2} \mathbf{W}(u,v)\, \|\mathbf{X}(u) - \mathbf{X}(v)\|^2\, du\, dv. \tag{4}$$

The following lemma shows that evaluating the continuous functional defined in (4) on a step graphon–signal pair recovers the normalized Dirichlet energy of $\mathbf{X}_n$ in the corresponding finite graph G. Proofs are omitted due to lack of space, and will be reported in the extended journal version of this paper.

Lemma 1. *Consider the step graphon–signal pair* $(\mathbf{W}_G, \mathbf{X}_G)$ *in* (2)-(3)*. Then, the graphon functional* Φ *in* (4) *satisfies*

$$\begin{aligned} \Phi(\mathbf{W}_G, \mathbf{X}_G) &= \iint_{[0,1]^2} \mathbf{W}_G(u,v)\, \|\mathbf{X}_G(u) - \mathbf{X}_G(v)\|^2\, du\, dv \\ &= \frac{1}{n^2} \sum_{(i,j) \in \mathcal{E}} A_{ij}\, \|\mathbf{x}_i - \mathbf{x}_j\|^2 \\ &= \frac{1}{n^2} \mathrm{TV}_G(\mathbf{X}_n). \end{aligned} \tag{5}$$

Based on the fundamental result of Borgs et al. [14, Proposition 5.10] a graph parameter φ is *testable* if and only if it admits a graphon extension $\Phi : \mathcal{W} \mapsto \mathbb{R}$ on the space of graphons $\mathcal{W}$ that is continuous w.r.t. the rectangle norm, equivalently the cut norm. Thus, the key steps to establish testability of the Dirichlet energy is to show that it admits a graphon extension Φ (cf. Lemma 1) and that Φ is Lipschitz continuous w.r.t. the cut norm, as we show next.

Theorem 1. *The graphon functional* Φ *in* (4) *is Lipschitz continuous under graphon–signal pair convergence. Specifically, if* $\|\mathbf{W}_G - \mathbf{W}\|_\square \to 0$ *and* $\|\mathbf{X}_G - \mathbf{X}\|_2 \to 0$*, then*

$$\Phi(\mathbf{W}_G, \mathbf{X}_G) \to \Phi(\mathbf{W}, \mathbf{X}), \quad n \to \infty. \tag{6}$$

Because induced subgraph sampling for a large enough sample size n^* yields sufficiently accurate approximations G^* to G in terms of cut distance [13, p. 47], then reliable homophily estimates are feasible for (even graph signal-dependent) continuous metrics such as (1) and its associated graphon functional (4) as per Theorem 1. This naturally justifies the adoption of a simple plug-in estimator.

III. Horvitz–Thompson Estimator for Homophily

To estimate homophily under *general* sampling designs, we bring to bear ideas that (in network settings) can be traced back to the foundational work by O. Frank [15]; see also [1, Ch. 5]. The challenge here is that network sampling often violates the standard assumption of a sample comprising i.i.d. observations from the population graph and, in fact, it often yields unequal probability sampling (see also Section III-A). A key requirement is that the network statistic of interest can be expressed as a total (or average) over sampled units, here unordered node pairs $\mathcal{V} \times \mathcal{V}$ or directly edges in $\mathcal{E}$ [cf. (1)].

Consider a random sampling design applied to G, which results in a sampled graph $G^* = (\mathcal{V}^*, \mathcal{E}^*)$ and signals $\mathbf{X}_{n^*}$, $n^* \leq n$. Let $\pi_{ij} := \mathrm{P}\left((i,j) \in \mathcal{E}^*\right)$ be the *inclusion probability* of edge (i,j), meaning the probability that edge (i,j) is sampled. Suppose that for the sampled edges $(i,j) \in \mathcal{E}^*$ we observe (without error) the squared variation $V_{ij} := A_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2$. In this context, we propose a HT estimator [16] for the Dirichlet energy (1), namely

$$\widehat{\mathrm{TV}}_{\mathrm{HT}} := \sum_{(i,j)\in\mathcal{E}^*} \frac{A_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\pi_{ij}} = \sum_{(i,j)\in\mathcal{E}^*} \frac{V_{ij}}{\pi_{ij}}, \tag{7}$$

which is *unbiased* for arbitrary sampling designs if $\pi_{ij} > 0$ for all edges. Indeed, using inclusion probabilities π_{ij}^{-1} as weights is essential for unbiasedness since for the plug-in estimator $\widehat{\mathrm{TV}}_{G^*}(\mathbf{X}_{n^*})$ one has

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathrm{TV}}_{G^*}(\mathbf{X}_{n^*})\right] &= \mathbb{E}\left[\sum_{(i,j)\in\mathcal{E}^*} A_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right] \\
&= \mathbb{E}\left[\sum_{(i,j)\in\mathcal{E}} V_{ij}\mathbb{I}\left\{(i,j)\in\mathcal{E}^*\right\}\right] \\
&= \sum_{(i,j)\in\mathcal{E}} V_{ij}\pi_{ij}.
\end{aligned} \tag{8}$$

From (8), $\mathbb{E}\left[\widehat{\mathrm{TV}}_{\mathrm{HT}}\right] = \mathrm{TV}_G(\mathbf{X}_n)$ and unbiasedness follows. Notice that as it is customary in statistical sampling theory which relies on design-based inference, the only randomness is due to the sampling of nodes and edges. Graph data, in particular the V_{ij}, are asummed error free.

The variance of $\widehat{\mathrm{TV}}_{\mathrm{HT}}$ can be estimated from G^* using

$$\mathrm{var}\left[\widehat{\mathrm{TV}}_{\mathrm{HT}}\right] = \sum_{(i,j)\in\mathcal{E}^*}\sum_{(k,l)\in\mathcal{E}^*} V_{ij}V_{kl}\left(\frac{1}{\pi_{ij}\pi_{kl}} - \frac{1}{\pi_{ijkl}}\right), \tag{9}$$

where now $\pi_{ijkl} := \mathrm{P}\left((i,j) \text{ and } (k,l) \in \mathcal{E}^*\right)$ is the *joint* inclusion probability of a pair of edges (i,j) and (k,l). Apparently, applicability of the HT framework hinges on three essential requirements: (i) the network statistic is expressible as a total over sampled units; (ii) the sampling design makes it feasible to observe the unit attribute of interest (here V_{ij}); and (iii) the (joint) inclusion probabilities can be calculated to form the estimator and evaluate its variance. Next, we outline canonical network sampling designs and comment on (iii).

A. Graph sampling designs

We consider a couple workhorse network sampling designs such as induced subgraph sampling (where nodes are first selected via Bernoulli sampling or simple random sampling without replacement) and traceroute sampling; see e.g., [8], [17]. Under Bernoulli sampling (BS), each node is retained in $\mathcal{V}^*$ independently with probability p. Induced subragph sampling dictates that an edge $(i,j) \in \mathcal{E}$ is observed (and included in $\mathcal{E}^*$) only when both $i,j \in \mathcal{V}^*$. This yields the uniform edge inclusion probability $\pi_{ij} = p^2$, for all $(i,j) \in \mathcal{E}^*$. Alternatively, in simple random sampling (SRS) one selects n^* nodes uniformly at random and without replacement (as for testability in Section II). Therefore, the node inclusion probability is $\pi_i = \frac{n^*}{n}$ and for induced subgraph sampling the edge inclusion probability is $\pi_{ij} = \frac{n^*(n^*-1)}{n(n-1)}$. Again, notice that inclusion probabilities do not depend on specific units.

In contrast, traceroute sampling traverses and samples edges along shortest paths between n_{S} randomly chosen sources and n_{T} randomly chosen targets. Edges that appear frequently on shortest paths are more likely to be sampled, leading to non-uniform inclusion probabilities that depend on the topology of the graph. This is precisely an instance of unequal probability sampling. While it is challenging to derive the edge inclusion probabilities exactly, one can obtain the approximation

$$\pi_{ij} \approx 1 - \exp\Big(-b_{ij}\,\frac{n_{\mathrm{S}}n_{\mathrm{T}}}{n^2}\Big),$$

where b_{ij} denotes the betweeness centrality of edge (i,j) [1, Ch. 5]. These three sampling designs illustrate the transition from uniform and readily obtained inclusion probabilities to highly heterogeneous ones, which can be intractable to compute – a staple of network sampling.

In our ensuing numerical experiments, we will examine several of these canonical network sampling designs, as well as other homophily measures beyond $\mathrm{TV}_G(\mathbf{X})$ [2, Sec. 3]. These analyses will demonstrate the robustness of the proposed estimator and reveal trade-offs related to the feasibility of evaluating inclusion probabilities and estimators' variances.

IV. Numerical Evaluation

We assess the performance of the HT estimator for different homophily metrics using a variety of network datasets. The code used to reproduce all the results we report is publicly available on GitHub[1], and the interested reader is referred therein for additional experiments and implementation details.

Experimental setup. Each graph considered in our experiments is equipped with categorical node attributes, represented as one-hot encoded vectors collected in $\mathbf{X}_n \in \mathbb{R}^{n\times f}$, where f is the number of node-types. Given a graph G and its node features $\mathbf{X}_n$, the task is to estimate the global homophily level from only a sampled subset of its edges.

To generate partial observations, we consider sampling schemes with both equal and unequal edge inclusion probabilities. Induced subgraph sampling (where nodes are selected

[1] https://github.com/hamedajorlou/Homophily-HT-Estimation

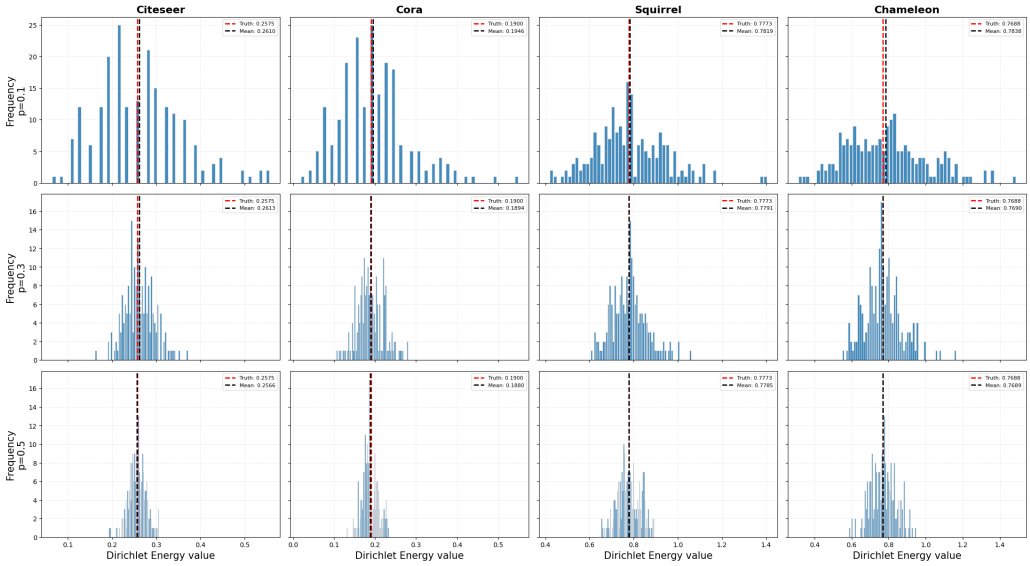

Fig. 1: Dispersion analysis of $\widehat{\mathrm{TV}}_{\mathrm{HT}}$. Nodes are sampled independently with probability $p = \{0.1, 0.3, 0.5\}$ (BS), and edges are selected using induced subgraph sampling. Histogram of estimates over $T = 200$ realizations, for varying p and different datasets. Unbiasedness is well supported in all cases, while the estimator variance grows with smaller p.

TABLE I: Estimation of different homophily measures under SRS-based induced subgraph sampling. Ground truth (GT) values and HT estimates of Dirichlet energy, edge homophily and node homophily are shown for samples containing 30% of the total node counts, averaged over $T = 200$ realizations.

	Size		Dirichlet energy			Edge homophily			Node homophily		
Dataset	Nodes	Edges	GT	Est	Bias	GT	Est	Bias	GT	Est	Bias
Amazon	334863	925872	0.6196	0.6200	0.0004	0.3804	0.3808	0.0005	0.3757	0.3760	0.0003
Citeseer	3327	4676	0.2575	0.2564	-0.0010	0.7425	0.7575	0.0150	0.7102	0.7070	-0.0032
Cora	2708	5429	0.1900	0.1902	0.0002	0.8100	0.8097	-0.0002	0.8252	0.8187	-0.0064
Cornell	183	298	0.8679	0.8817	0.0138	0.1321	0.1698	0.0376	0.1160	0.1411	0.0252
Pubmed	19717	44338	0.1976	0.1981	0.0005	0.8024	0.8021	-0.0003	0.7924	0.7971	0.0047
Wisconsin	251	450	0.7940	0.7991	0.0052	0.2060	0.2822	0.0762	0.1665	0.2025	0.0360
Question	4897	15362	0.1604	0.1590	-0.0014	0.8396	0.8410	0.0014	0.8980	0.8887	-0.0093
Squirrel	5201	217073	0.7773	0.7775	0.0002	0.2227	0.2239	0.0013	0.2176	0.2189	0.0013
Chameleon	2277	31371	0.7688	0.7633	-0.0056	0.2312	0.2332	0.0020	0.2469	0.2442	-0.0028
Karate club	34	78	0.1082	0.1062	-0.0013	0.8918	0.8882	-0.0036	0.8882	0.8728	-0.0154

using either BS or SRS) yields an equal–probability design, while traceroute sampling represents the unequal-probability regime. Each sampling method produces an observed subgraph G^* with attributes $\mathbf{X}_{n^*}$, from which we compute an homophily estimate $\widehat{\mathrm{TV}}_{\mathrm{HT}}$ using the inclusion probabilities in Section III-A. Performance is evaluated over $T = 200$ realizations by reporting the bias $(\widehat{\mathrm{TV}}_{\mathrm{HT}} - \mathrm{TV}_G(\mathbf{X}_n))$. All Dirichlet energy values are normalized to the interval $[0, 1]$.

Datasets and preliminary observations. Our experiments use the heterophily benchmark datasets introduced and standardized in recent evaluations of non-homophilous graphs [18]–[21]. The collection includes graphs such as Chameleon, Cora, Citeseer, Cornell, Amazon, Wisconsin, Squirrel, and Karate Club, which cover a wide range of network sizes and homophily levels. This diversity provides an ideal setting to assess homophily estimation under different sampling designs. Across all datasets, initial experiments confirm that the HT estimator recovers homophily metrics with negligible bias. We discuss these findings through several test cases presented next.

Test Case 1 - Dispersion analysis under different p. In this experiment, we study how p influences the variability of the HT estimator under BS sampling of nodes. We examine the sampling rates $p = \{0.1, 0.3, 0.5\}$, and recall the edge inclusion probability $\pi_{ij} = p^2$. Notice that $\widehat{\mathrm{TV}}_{\mathrm{HT}} = p^{-2}\widehat{\mathrm{TV}}_{G^*}(\mathbf{X}_{n^*})$, so the HT estimator is a scaled-up correction of the plug-in estimator. As shown in the histograms in Fig. 1, unbiasedness is supported for all values of p, but the dispersion changes significantly. For small sampling rates, the estimator exhibits a wide spread due to the limited number of observed V_{ij} values (even with the correction of p^{-2}). As the number of sampled edges grows, the distribution of estimates concentrates around the ground-truth value.

Test Case 2 - Different homophily measures. Here we evaluate the performance of the HT estimator under induced subgraph sampling. A fixed number of nodes is selected

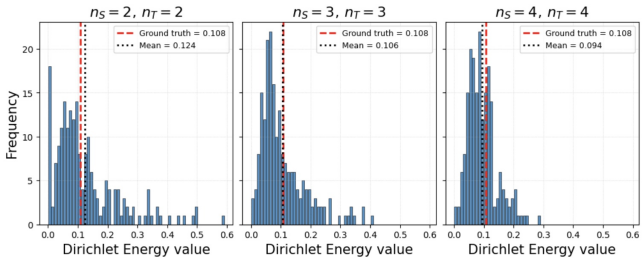

Fig. 2: Homophily estimation under traceroute sampling. Shortest paths are traced between randomly selected source and target nodes on the Karate Club network. Histograms of the estimated Dirichlet energy for different source and target sampling rates, obtained for $T = 200$ realizations.

uniformly at random without replacement (SRS), and all edges between the selected nodes are observed. Table I reports the results obtained by sampling 30% of the total nodes, averaged over $T = 200$ realizations. Across all datasets, the estimator remains effectively unbiased: the estimated Dirichlet energy, edge homophily where V_{ij} is replaced by $A_{ij}\mathbb{I}\{\mathbf{x}_j \equiv \mathbf{x}_j\}$, and node homophily [2] closely match their ground-truth values, with biases typically in the order of 10^{-3} or smaller. Even with a rather small sample size of edges, the consistency of the results across citation graphs, web networks, and smaller benchmark graphs confirms that the HT estimator performs reliably under induced subgraph sampling.

Test Case 3 - Traceroute sampling. To close, we study the behavior of the HT estimator under traceroute sampling, a setting in which edges are discovered along shortest paths between randomly selected source and target nodes. We consider the Zachary Karate Club network using three distinct source–target pair sampling rates, to illustrate the estimator's behavior under different path configurations. Across all cases, the estimator remains unbiased, confirming that the weights correctly compensate for unequal inclusion probabilities. The recovered homophily metrics remain stable across all three source–target settings, corroborating that the HT estimator is capable of handling strongly heterogeneous sampling designs, provided that the inclusion probabilities are well approximated.

V. Conclusions

This work introduced a unified framework for estimating homophily-related structural metrics from sampled graphs. Using the HT methodology, we derived unbiased estimators for the Dirichlet energy (and other heterophily measures expressible as edge totals or averages) under both unform and unequal-probability sampling designs. We also established that the Dirichlet energy is a testable (i.e., weakly consistent) graph parameter under induced subgraph sampling.

Our reproducible numerical experiments using benchmark graph datasets confirmed that the novel HT estimators remain unbiased when the population graphs are sampled via induced subgraph or traceroute sampling. These findings underscore the importance of accounting for the sampling design when inferring structural properties from partial network measurements, and the pitfalls of plug-in estimators. Future work includes developing variance-reduction strategies, extending the analysis to dynamic graphs, and examining testability for other heterophily-related network statistics.

References

[1] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*, Springer, 2009.

[2] S. Luan et al., "The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges," 2024.

[3] G. Mateos, S. Segarra, A. G. Marques, and A. Ribeiro, "Connecting the dots: Identifying network structure via graph signal processing," vol. 36, no. 3, pp. 16–43, 2019.

[4] F. Gama, E. Isufi, G. Leus, and A. Ribeiro, "Graphs, convolutions, and neural networks: From graph filters to graph neural networks," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 128–138, 2020.

[5] A. Raghuvanshi, G. Mateos, and S. P. Chepuri, "Task-driven heterophilic graph structure searning," in *Proc. Asilomar Conf. on Signals, Systems, Computers*, 2025, pp. 1–5.

[6] L. Liang, X. Hu, Z. Xu, Z Song, and I King, "Predicting global label relationship matrix for graph neural networks under heterophily," in *Adv. Neural. Inf. Process. Syst. (NeurIPS)*, 2023, vol. 36.

[7] N. K. Ahmed, J. Neville, and R. Kompella, "Network sampling: From static to streaming graphs," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 2, pp. 7:1–7:33, 2014.

[8] O. Frank, "Estimation of the number of vertices of different degrees in a graph," *J. Stat. Planning and Inference*, vol. 4, pp. 45–50, 1980.

[9] Y. Tanaka, C. Eldar, Y., A. Ortega, and G. Cheung, "Sampling signals on graphs: From theory to applications," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 14–30, 2020.

[10] H. Li, H. Wang, and L. Ruiz, "Graph sampling for scalable and expressive graph neural networks on homophilic graphs," in *Proc. of European Signal Process. Conf. (EUSIPCO)*, 2025, pp. 2397–2401.

[11] C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, "Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing," *Advances in Mathematics*, vol. 219, no. 6, pp. 1801–1851, 2008.

[12] L. Ruiz, L. F. O. Chamon, and A. Ribeiro, "Graphon signal processing," *IEEE Trans. Signal Process.*, vol. 69, pp. 4961–4976, 2021.

[13] E. Kolaczyk, *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*, SemStat Elements. Cambridge University Press, 2017.

[14] C. Borgs, J. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi, "Counting graph homomorphisms," in *Topics in Discrete Mathematics*, vol. 26 of *Algorithms and Combinatorics*, pp. 315–371. Springer, 2006.

[15] Ove Frank, *Network Sampling and Model Fitting*, p. 31–56, Structural Analysis in the Social Sciences. Cambridge, 2005.

[16] D. G. Horvitz and D. J. Thompson, "A generalization of sampling without replacement from a finite universe," *J. Am. Stat. Assoc.*, vol. 47, pp. 663–685, 1952.

[17] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore, "On the bias of traceroute sampling: Or, power law degree distributions in regular graphs," *Journal of the ACM*, vol. 56, no. 4, pp. 21:1–21:28, 2009.

[18] D. Lim, F. Hohne, X. Li, S. L. Huang, V. Gupta, O. Bhalerao, and S. N. Lim, "Large scale learning on non homophilous graphs: New benchmarks and strong simple methods," in *Adv. Neural. Inf. Process. Syst. (NeurIPS)*, 2021, vol. 34, pp. 20887–20902.

[19] Y. Sun, H. Deng, Y. Yang, C. Wang, J. Xu, R. Huang, L. Cao, Y. Wang, and L. Chen, "Beyond homophily: Structure aware path aggregation graph neural network," in *Proceedings of the Thirty First International Joint Conference on Artificial Intelligence*, 2022, pp. 2233–2240.

[20] O. Platonov, D. Kuznedelev, M. Diskin, A. Babenko, and L. Prokhorenkova, "A critical look at the evaluation of GNNs under heterophily: Are we really making progress," in *Intl. Conf. on Learning Representations (ICLR)*, 2022.

[21] Z. Zhou, S. Zhou, B. Mao, X. Zhou, J. Chen, Q. Tan, D. Zha, Y. Feng, C. Chen, and C. Wang, "Opengsl: A comprehensive benchmark for graph structure learning," in *Adv. Neural. Inf. Process. Syst. (NeurIPS)*, 2024, vol. 36.