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

Hamed Ajorlou†, Gonzalo Mateos†, and Luana Ruiz*

†Dept. of Electrical and Computer Eng., University of Rochester, Rochester, NY

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

Motivation and context

- **Homophily**: tendency of nodes with **similar attributes** to connect
- underpins key graph learning tasks such as **nearest-neighbor prediction**, **semi-supervised learning**, and **topology inference**
  - ⇒ Such insights guide **better and more tailored** model(e.g., GNN) design
  - ⇒ Recent works increasingly target **heterophilous data** [2], [3]
- **Limitation**: Often only a **sample** of the network is observed
- **This work**: estimate several homophily metrics from sampled graph data
  - ⇒ We address **testability** and validate **unbiasedness** on real data

Preliminaries and notation

- Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an undirected graph with $|\mathcal{V}| = N$ nodes. Its adjacency matrix is $\mathbf{A} \in \mathbb{R}^{N\times N}$. Each node $i \in \mathcal{V}$ is assigned a **one–hot encoded** feature vector $\mathbf{x}_i \in \mathbb{R}^d$, where d is the number of node types. Stacking all feature vectors yields the **feature matrix** $\mathbf{X} \in \mathbb{R}^{N\times d}$
- The Laplacian of $\mathcal{G}$ is defined as

$$\mathbf{L} := \mathrm{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$$

- The **Dirichlet energy** of $\mathcal{G}$ with respect to node features $\mathbf{X}$ is given by

$$\mathrm{TV}_{\mathcal{G}}(\mathbf{X}) := \mathrm{trace}(\mathbf{X}^\top \mathbf{L}\mathbf{X}) = \sum_{(i,j)\in\mathcal{E}} \mathbf{A}_{ij}\,\|\mathbf{x}_i - \mathbf{x}_j\|^2$$

  - ⇒ Can be expressed as a total over edges
- Let $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ be a random subgraph of $\mathcal{G}$ obtained under a **prescribed sampling scheme**. For each edge $(i,j) \in \mathcal{E}$, denote by

$$\pi_{ij} := \mathbb{P}\left[(i,j) \in \mathcal{E}^*\right]$$

  its **inclusion probability**.

Problem formulation

- **Setup:** Given a weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with adjacency $\mathbf{A}$ and feature matrix $\mathbf{X} \in \mathbf{R}^{N\times d}$, the Dirichlet energy quantifies **homophily**
- Many times, the complete graph is not accessible
  - ⇒ **Dynamic** environments
  - ⇒ **Resource** limitations
  - ⇒ **Privacy** constraints
- **Problem:** Given a sample $\mathcal{G}^* = (\mathcal{V}^*, \mathcal{E}^*)$ of $\mathcal{G}$ obtained using a sampling design with edge inclusion probabilities $\pi_{ij} = ((i,j) \in \mathcal{E}^*)$, estimate $\mathrm{TV}_{\mathcal{G}}(\mathbf{X})$ (and related metrics) **unbiasedly** from $(\mathcal{G}^*, \mathbf{X}^*)$
- Other homophily metrics can be considered and estimated with HT estimator:

$$\text{Edge Homophily} = \frac{\sum_{(u,v)\in\mathcal{E}} \mathbf{A}_{ij}\,\mathbb{I}\{\mathbf{x}_i = \mathbf{x}_j\}}{\sum_{(i,j)\in\mathcal{E}} \mathbf{A}_{ij}}$$

$$\text{Meta-path based Label Homophily}(\mathrm{MLH}_L) = \frac{\sum_{i\neq j} (\mathbf{A}^L)_{ij}\,\mathbb{I}\{\mathbf{x}_i = \mathbf{x}_j\}}{\sum_{i\neq j} (\mathbf{A}^L)_{ij}}$$

  - ⇒ Here, $\mathbb{I}\{\cdot\}$ denotes the indicator function
- **Challenges:**
  - ⇒ Inclusion probabilities may be **unequal** and hard to compute
  - ⇒ Estimator variance depends on the **sampling design**

Network sampling designs

A sampling design produces $\to \mathcal{V}^* \subseteq \mathcal{V}$ and the corresponding induced edge set

$$\mathcal{E}^* = \{(i,j) \in \mathcal{E} : \ i,j \in \mathcal{V}^*\}.$$

Inclusion probabilities: $\pi_i = \mathbb{P}(i \in \mathcal{V}^*)$, $\pi_{ij} = \mathbb{P}((i,j) \in \mathcal{E}^*)$, π_{ijkl} for pairs of edges.

Bernoulli sampling: Each node is included independently with probability p

- Node inclusion: $\pi_i = p$
- Edge inclusion: $\pi_{ij} = p^2$
- Joint edge inclusion: $\pi_{ijkl} = p^4$
  - ⇒ Explicit control over expected sample size
  - ⇒ Produces random sample sizes

Induced subgraph sampling: A fixed number n of vertices is drawn uniformly without replacement, and only edges between sampled vertices are kept.

- Node inclusion: $\pi_i = \frac{n}{N}$
- Edge inclusion: $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$
  - ⇒ Guarantees exact sample size n
  - ⇒ Preserves induced structure among sampled vertices

Horvitz–Thompson (HT) estimator

- **HT estimator**: Takes into account the inclusion probability π_{ij} of each sampled edge [1]
  - ⇒ Applies to estimation of network **totals** (or **averages**)

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

- ⇒ The HT estimator is **unbiased**. **Variance** given by

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

- **Variance estimation:** $\uparrow \pi \ \Rightarrow \downarrow \mathrm{Var} \ \Rightarrow$ **Tighter histogram**

Testability

- **Testability** of the Dirichlet energy statistic can be established under sampling
- A network statistic η is **testable** if for every $\epsilon > 0$ there exists a sample size n such that for any graph $\mathcal{G}$ with $N \geq n$, an estimate $\hat{\eta} = \eta(\mathcal{G}_n^*)$ from a sampled $\mathcal{G}_n^*$ satisfies

$$\mathbb{P}[|\eta(\mathcal{G}) - \hat{\eta}| > \varepsilon] \leq \varepsilon$$

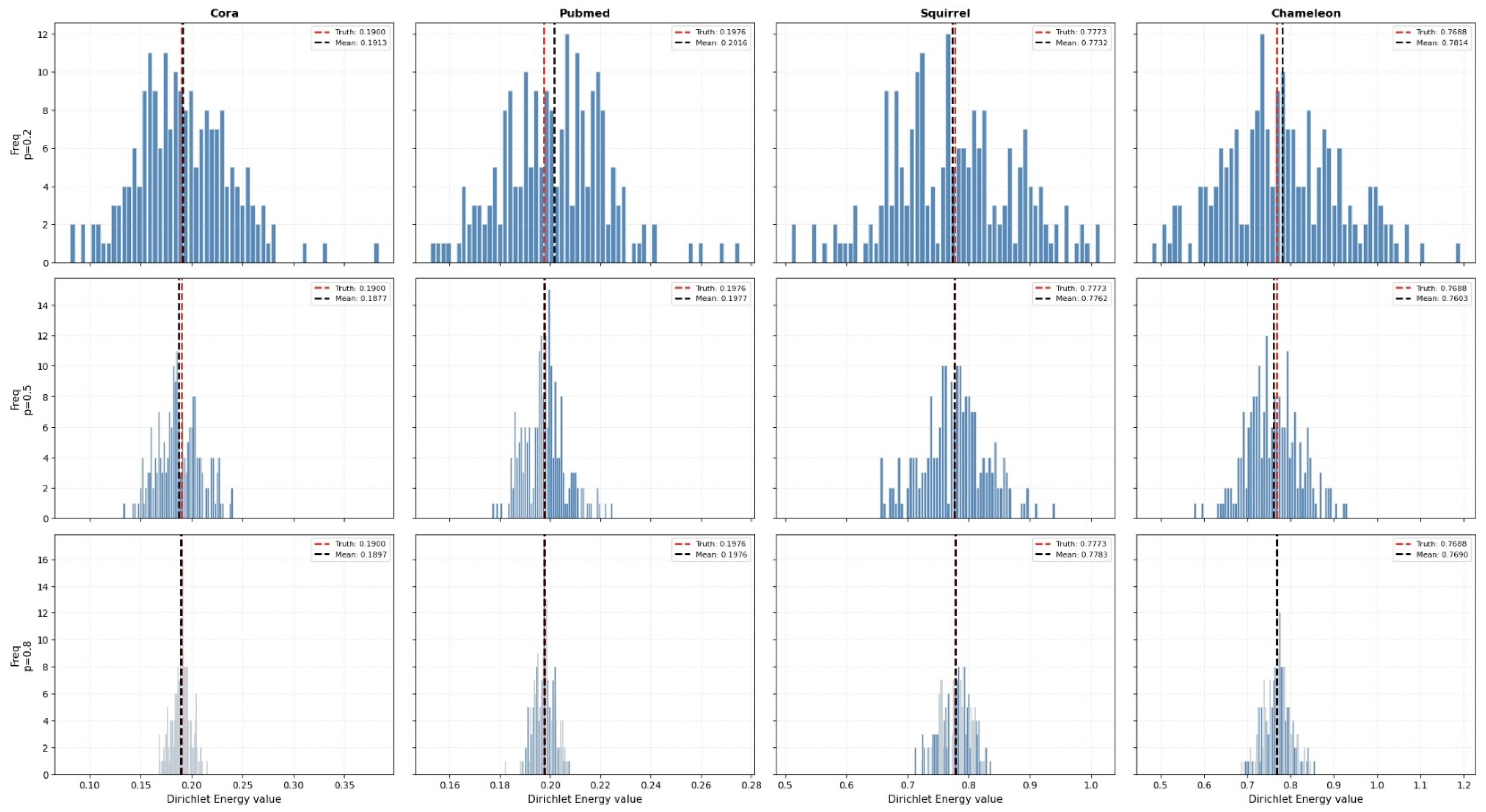
Implications

- Testability ⇔ existence of a **weakly consistent** estimator
- Foundation for studying **feasibility** of inference tasks on large graphs

Evaluation

- Estimation of homophily metrics under sampling **is evaluated** across several datasets
  - ⇒ how variance of the measures change
  - ⇒ unbiasedness of the estimation

Estimation of Dirichlet energy under Bernoulli sampling



- p = $\{0.2, 0.5, 0.8\}$, 200 realizations, datasets = $\{$*Cora*, *Pubmed*, *Squirrel*, *Chameleon*$\}$
  - ⇒ Higher sampling rate **shrinks variance of estimates**
  - ⇒ Unbiasedness well supported, even for small samples
  - ⇒ Sampling schemes can be tailored for specific cases

Induced subgraph sampling: Additional metrics and datasets

	Size		Dirichlet Energy			Edge Homophily			MLH		
Dataset	Nodes	Edges	GT	Est	Bias	GT	Est	Bias	GT	Est	Bias
Citeseer	3327	4676	0.2575	0.2560	-0.0015	0.7425	0.7598	0.0173	0.7547	0.7545	-0.0002
Cora	2708	5278	0.1900	0.1898	-0.0002	0.8100	0.8102	0.0002	0.7795	0.7732	-0.0063
Cornell	183	280	0.8679	0.8591	-0.0088	0.1321	0.1688	0.0366	0.2586	0.2907	0.0321
Wisconsin	251	466	0.7940	0.8135	0.0195	0.2060	0.2794	0.0734	0.3012	0.3120	0.0108
Amazon	24492	93050	0.6196	0.6193	-0.0003	0.3804	0.3801	-0.0002	0.3988	0.3981	-0.0007
Squirrel	5201	198493	0.7773	0.7817	0.0043	0.2227	0.2234	0.0007	0.2291	0.2287	-0.0004
Chameleon	2277	31421	0.7688	0.7689	0.0001	0.2312	0.2345	0.0034	0.2676	0.2701	0.0024

Table: Sampling: Induced Subgraph Sampling, $n = 0.2 \times N$, 200 realizations

Link to the Github & future directions

- Testability: explore graph limit models and sampling designs
- Test unequal probability sampling designs
- Extensive numerical experiments

References

[1] C. Borgs, J. Chayes, L. LovÃ¡sz, V. SÃ^{3}s, 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.

[2] 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)*, vol. 36, 2023.

[3] S. Luan *et al.*, *The heterophilic graph learning handbook: Benchmarks, models, theoretical analysis, applications and challenges*, 2024. arXiv: 2407.09618 [cs.LG].