Counting Motifs with Graph Sampling

Jason M. Klusowski and Yihong Wu

Yale University, Department of Statistics and Data Science

Objectives

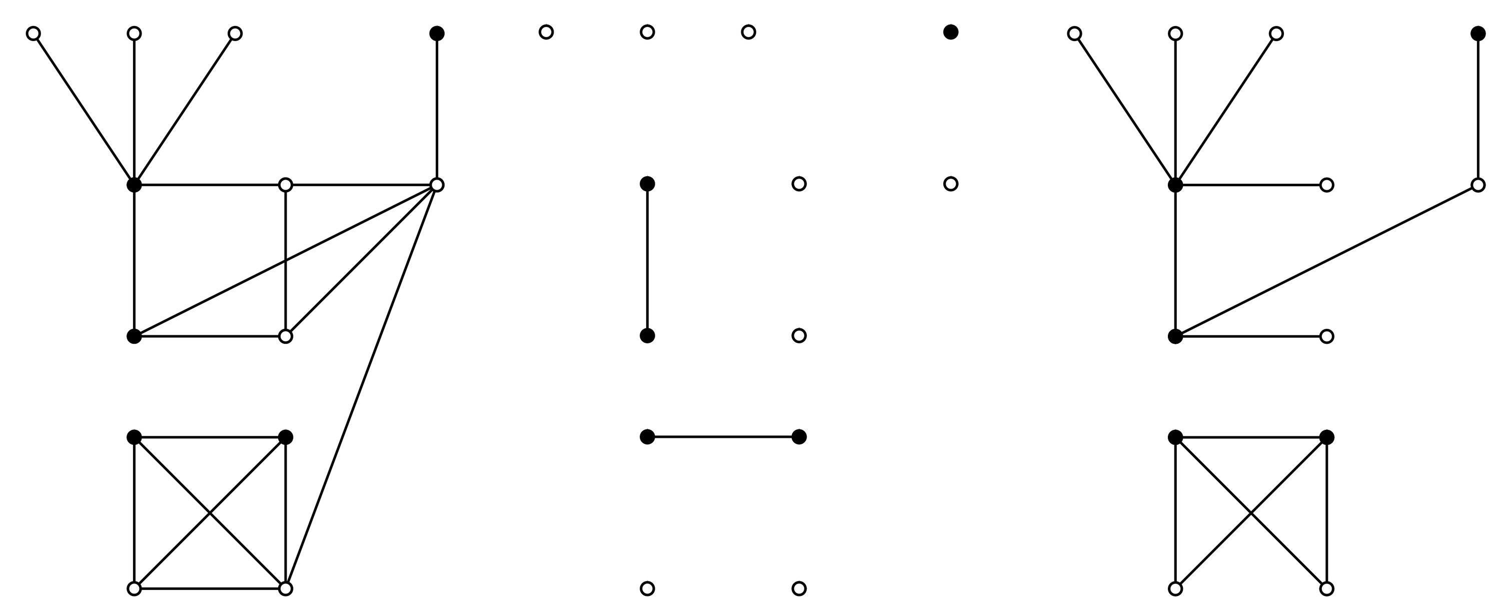
Develop a statistical theory for estimating motif counts (as induced subgraph) in a sampled graph. Focus on large graphs and sublinear sampling regime, where only a vanishing fraction of vertices are sampled.

- How does the sample complexity depend on the motif itself? For example, is estimating the count of open triangles as easy as estimating the closed triangles?
- How much of the graph must be observed to ensure accurate estimation?
- How much more informative is neighborhood sampling than subgraph sampling from the perspective of reducing the sample complexity?

Introduction

Fix a simple graph $G = (V, E)$ on $\mathsf{v}(G)$ vertices. We study two sampling models:

- **Subgraph sampling.** Sample vertices $S \subset V$ independently with probability p. Observe induced subgraph $G^\star \triangleq G[S]$.
- **Neighborhood sampling.** Observe $G[S]$ *and* edges between vertices in S and their neighbors. Sampled graph $G^\star$ is bicolored, with black (sampled) and white (unsampled) vertices.



(a) Parent graph. (b) Subgraph sampling. (c) Neighborhood sampling.

Subgraph Counts

Two fundamental quantities govern the distribution of the sampled graph $G^\star$.

- **Subgraph sampling.** Let $\mathsf{s}(g, G)$ denote the number of induced subgraphs of G that are isomorphic to g, e.g., $\mathsf{s}(\ ,\ ) = 2$. Then
$$\mathbb{P}[G^\star \simeq g] = \mathsf{s}(g, G) p^{\mathsf{v}(g)} (1-p)^{\mathsf{v}(G)-\mathsf{v}(g)},$$
where $\mathsf{v}(g)$ is number of vertices in g.
- **Neighborhood sampling.** Let $\mathsf{N}(h, G)$ denote the number of ways that a bicolored g appears (isomorphic as a vertex-colored graph) in G, e.g., $\mathsf{N}(\ ,\ ) = 2$. Then
$$\mathbb{P}[G^\star \simeq g] = \mathsf{N}(g, G) p^{\mathsf{v}_b(g)} (1-p)^{\mathsf{v}(G)-\mathsf{v}_b(g)},$$
where $\mathsf{v}_b(g)$ is number of black vertices in g.

Optimal Estimators: Achievability

- For parent graph G with maximal degree bounded by d, for any motif h (**connected** subgraph) on k vertices, estimate $s = \mathsf{s}(h, G)$ with a multiplicative error of ϵ.
- **Subgraph sampling.** Horvitz-Thompson (HT) type estimator
$$\hat{\mathsf{s}}_h \triangleq \mathsf{s}(h, G^\star)/p^{\mathsf{v}(h)}.$$
- **Neighborhood sampling.** Use HT when $p \le 1/d$. However, HT is suboptimal for $p > 1/d$. Tailored estimator is a linear combination of the $\mathsf{N}(\cdot, G)$ and improves the HT estimator by incorporating the colors of the vertices to reduce (or eliminate) correlation.
  - **Example.** Edge count estimator has form
$$\hat{\mathsf{e}} = \alpha \cdot \mathsf{N}(\ , G^\star) + \beta \cdot \mathsf{N}(\ , G^\star),$$
with $\alpha = \frac{1+dp}{p(2+(d-1)p)}$ and $\beta = \frac{1-d(1-2p)}{p(2+(d-1)p)}$ optimized to reduce variance.
  - Adaptive estimator available with similar guarantees – do not need to know d a priori.

Minimax Lower Bounds

- Construct random instances of graphs with matching structures of small subgraphs, akin to the method of moment matching.
- **Subgraph sampling.** For any connected graph h with k vertices, there exists a pair of connected graphs H and H', such that $\mathsf{s}(h, H) \neq \mathsf{s}(h, H')$ and $\mathsf{s}(g, H) = \mathsf{s}(g, H')$ for all connected g with $\mathsf{v}(g) \le k - 1$.
  - **Example for $h =$ and $k = 3$.**
$$H = \qquad H' = \qquad \Longrightarrow \qquad \mathrm{TV}(\mathcal{L}(H^\star), \mathcal{L}(H'^\star)) = O(p^3).$$
- **Neighborhood sampling.** For k-cliques $h = K_k$, there exists two graphs H and H' such that $\mathsf{s}(K_k, H) - \mathsf{s}(K_k, H') \ge 1$ and $\mathsf{N}(g, H) = \mathsf{N}(g, H')$ for all neighborhood subgraphs g such that $\mathsf{v}_b(g) \le k - 2$.
  - **Example for $h =$ and $k = 3$.**
$$H = \qquad H' = \qquad \Longrightarrow \qquad \mathrm{TV}(\mathcal{L}(H^\star), \mathcal{L}(H'^\star)) = O(p^2).$$
- Produce two graphs H and H' that are statistically indistinguishable unless at least k vertices (resp. $k-1$) are sampled with subgraph (resp. neighborhood) sampling, but have large separation in the number of motif h.

Sample Complexity

- For subgraph sampling, the optimal sampling ratio p is $\Theta_k(\max\{(s\epsilon^2)^{-\frac{1}{k}}, \frac{d^{k-1}}{s\epsilon^2}\})$, which only depends on the size of the motif but *not* its actual topology. Furthermore, HT type estimators are universally optimal for any connected motifs.
  - When $s = \mathsf{e}(G)$ is edge count, optimal sampling ratio scales as
$$\Theta\left(\max\left\{\frac{1}{\sqrt{s}\epsilon}, \frac{d}{s\epsilon^2}\right\}\right).$$
- For neighborhood sampling, we achieve the sampling ratio $O_k(\min\{(\frac{d}{s\epsilon^2})^{\frac{1}{k-1}}, \sqrt{\frac{d^{k-2}}{s\epsilon^2}}\})$, which again only depends on the size of h. Optimal for all motifs with at most 4 vertices and cliques of all sizes.
  - When $s = \mathsf{e}(G)$ is edge count, optimal sampling ratio scales as
$$\Theta\left(\min\left\{\frac{1}{\sqrt{s}\epsilon}, \frac{d}{s\epsilon^2}\right\}\right).$$

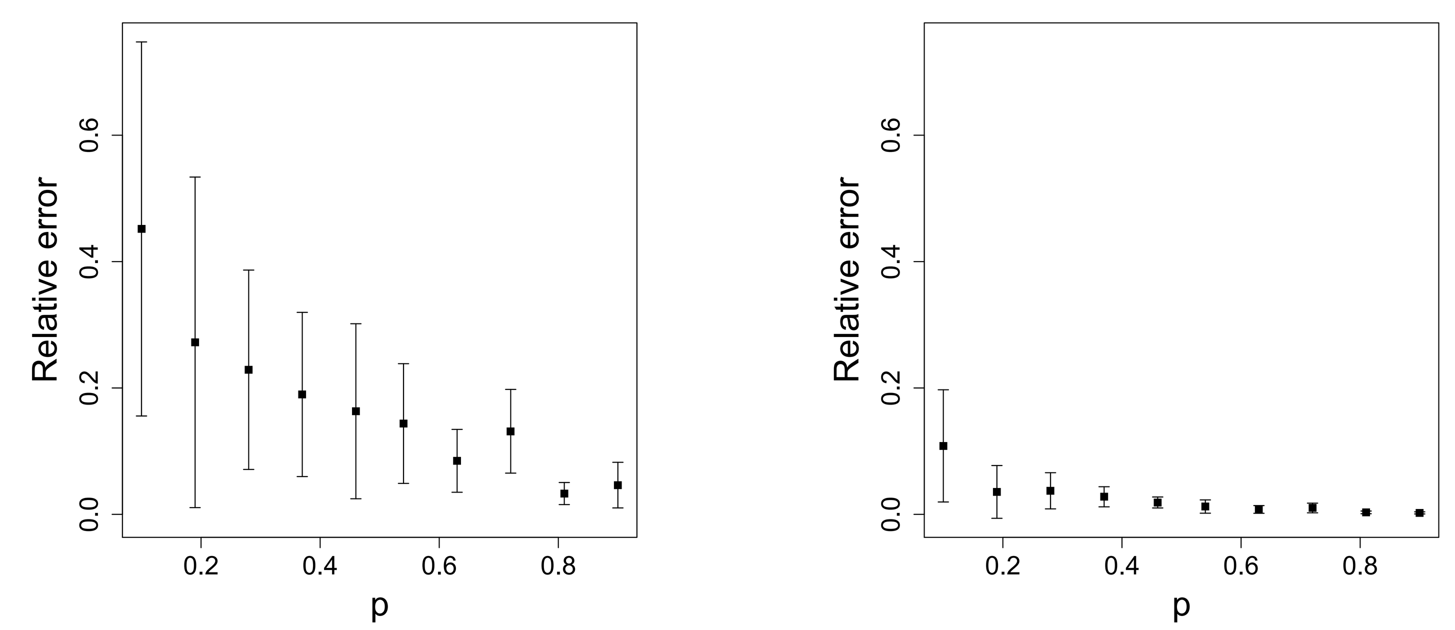
Additional Structure

To what extent does additional structures of the parent graph, e.g., tree or planarity, impact the sample complexity?

- Tree structure only marginally improves estimation of edges and wedges for both subgraph and neighborhood sampling.
- Planarity improves estimation for triangles for both sampling models.

Experiments

- Collaboration network between jazz musicians in 198 bands that performed between 1912 and 1940 [Gleiser-Danon, 2003]. Each node is a band and there is an edge between two bands if and only if at least one jazz musician has played in both bands.
- Estimators based on neighborhood sampling perform better than subgraph sampling (significantly less variability).



(a) Jazz network (subgraph sampling). (b) Jazz network (neighborhood sampling).

Figure: Relative error of estimating the edge count over 10 independent trials. The parent graph G is the jazz network with $d = 100$, $\mathsf{v}(G) = 198$, $\mathsf{e}(G) = 2742$.

Open Questions

- Determine optimal sample complexity in neighborhood sampling for general subgraph counts.
- Statistical limits of r-hop neighborhood sampling, where we observe a labeled radius-r ball rooted at a randomly chosen vertex [3]. Neighborhood sampling corresponds to $r = 1$.
- Statistical limits of counting *edge-induced* subgraphs.

References

[1] Jason M. Klusowski and Yihong Wu.
Counting motifs with graph sampling.
To appear in Conference on Learning Theory (COLT), 2018.

[2] Jason M. Klusowski and Yihong Wu.
Estimating the number of connected components in a graph via subgraph sampling.
arXiv preprint arXiv:1801.04339, 2018.

[3] László Lovász.
Large Networks and Graph Limits, volume 60.
American Mathematical Society, 2012.