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 subgraphs) 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 \( |V| \) vertices. We study two sampling models:

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

Subgraph Counts

Two fundamental quantities govern the distribution of the sampled graph \( G' \).

- **Subgraph sampling.** Let \( s(g,G) \) denote the number of induced subgraphs of \( G \) that are isomorphic to \( g \). e.g., \( \square \square \square \square \square \). Then \( \mathbb{P}(G' \cong g) = s(g,G)p^{|g|}(1-p)^{|G|\setminus|g|} \), where \( w(g) \) is number of vertices in \( g \).

- **Neighborhood sampling.** Let \( N(h,G) \) denote the number of ways that a bicolor \( g \) appears (isomorphic as a vertex-colored graph) in \( G \), e.g., \( \square \square \square \square \square \). Then \( \mathbb{P}(G' \cong g) = N(g,G)p^{|g|}(1-p)^{|G|\setminus|g|} \), where \( w(g) \) is number of black vertices in \( g \).

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 \( h \) and \( k \) vertices, there exists a pair of connected graphs \( H \) and \( H' \) such that \( s(h, H) \neq s(h, H') \) and \( s(g, H) = s(g, H') \) for all connected \( g \) with \( w(g) \leq k \).

- **Example for \( h = \bigcirc \).**

\[
\begin{align*}
H = \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc \bigcirc & \quad \text{and} \quad H' = \bigcirc \bigcirc \bigcirc \\
& \Rightarrow \text{TV}(\mathcal{L}(H'), \mathcal{L}(H)) = O(p^2).
\end{align*}
\]

Minimax Lower Bounds

- For subgraph sampling, the optimal sampling ratio \( p \) is \( p = (\max \{\sqrt{w(h)}^2, w(h)\}) \), which only depends on the size of the motif but not its actual topology.
- For neighborhood sampling, we achieve the sampling ratio \( p^{\star} = (\max \{\sqrt{w(h)}^2, w(h)\}) \), which again only depends on the size of \( h \).

Sample Complexity

- For subgraph sampling, the optimal sampling ratio \( p \) is \( p = (\max \{\sqrt{w(h)}^2, w(h)\}) \), which only depends on the size of the motif but not its actual topology.

- **For neighborhood sampling, we achieve the sampling ratio \( p^{\star} = (\max \{\sqrt{w(h)}^2, w(h)\}) \), which again only depends on the size of \( h \).**

Optimal Estimators: Achievability

- For parent graph \( G \) with maximal degree bounded by \( d \), for any motif \( h \) (connected subgraph) on \( k \) vertices, estimate \( s = \hat{s}(h,G) \) with a multiplicative error of \( e \).

Subgraph sampling. Horvitz-Thompson (HT) type estimator

\[
\hat{s}_H \equiv \frac{s(h,G)}{p^{\star}}.
\]

Neighborhood sampling. Use HT when \( s \leq 1/d \). However, HT is suboptimal for \( p > 1/d \). Tailored estimator is a linear combination of the \( N(h,G) \) and improves the HT estimator by incorporating the colors of the vertices to reduce (or eliminate) correlation.

- **Example.** Edge count estimator has form

\[
\hat{s}_N = (1-\epsilon)p^{\star} \cdot N(h,G) + \beta \cdot s(h,G),
\]

with \( \epsilon = \frac{1-dp^{\star}}{1-dp} \) and \( \beta = \frac{1-d(1-p^{\star})}{1-d(1-p)} \) optimized to reduce variance.

- Adaptive estimator available with smaller guarantees - do not need to know \( d \) a priori.

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).

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