Estimating the Number of Connected Components in a Graph via Subgraph Sampling

Jason M. Klusowski Yihong Wu

Working paper: February 28, 2017

Contents

1 Notation

Fix a simple graph $G = (V, E)$. Throughout this document, we use the following notation. Let N or $\mathsf{v}(G)$ be the number of vertices and $\mathsf{e}(G)$ the number of edges. Let $\mathsf{cc}(G)$ is the number of components in G, Let $\mathsf{ind}(H, G)$ is the number of induced subgraphs of G that are isomorphic to H

2 Main results

The main results are summarized below in terms of the normalized minimax mean square error. We focus on the subgraph sampling model, that is, a subset of vertices is sampled at random and the induced subgraph is observed. More specifically, we consider the Bernoulli sampling model, where each vertex is sampled independently with probability p. Similar results can be obtained for the uniformly sampling model, where S is drawn uniformly at random from all subsets of $V(G)$ of cardinality n, when we identify $p = n/N$.

The estimator $\widehat{\mathsf{cc}}$ is a function of the subgraph induced by the sampled vertices. The goal is to estimate $\mathsf{cc}(G)$ uniformly for all G in a class of graphs. It turns out there are two major obstructions for subgraph sampling (a) high-degree vertices, and (b) long induced cycles, either of which implies high sample complexity. This motivates us to consider class of graphs defined by their maximal degree and maximal length of induced cycles (chordality). The major class of graphs we study is the so-called *chordal graphs*, which include *forests* and *disjoint union of cliques* as special cases, the two model that was studied in Frank's original 1978 paper [2].

Definition 1. *A graph G is chordal if it does not contain induced cycles of length four or above, i.e., $\mathsf{ind}(C_k, G) = 0$ for $k \geq 4$.*

Theorem 1 (Chordal graphs)**.** *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with clique number ω and maximum degree d. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \frac{\mathbb{E}_G(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \leq \frac{1}{Np^\omega} \vee \frac{2d}{Np^{\omega-1}}.$$

Conversely, if the maximal degree d is a constant, then

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \frac{\mathbb{E}_G(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \asymp \frac{1}{Np^\omega}.$$

Theorem 2 (Forests)**.** *Let $\mathcal{F}(N, d)$ denote the collection of all forests on N vertices with maximum degree d. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{F}(d)} \frac{\mathbb{E}_G(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \asymp \left(\frac{1}{Np^2} \vee \frac{d}{Np} \right) \wedge 1.$$

Theorem 3 (Cliques)**.** *Let $\mathcal{C}(N)$ denote the collection of all graphs on N vertices consisting of disjoint unions of cliques. Let $N > 4$. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{C}(N)} \frac{\mathbb{E}_G(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \leq (N/4)^{-\min\{1, \frac{p}{2q-p}\}}.$$

The above theorems can be summarized in terms of the *sample complexity*, i.e., the minimum sample size that allows an estimator $\mathsf{cc}(G)$ within $\pm\epsilon N$ with probability, say, 0.99, uniformly for all graphs in a given class. Here the sample size is understood as the average number of sampled vertices $n = pN$. We have the following characterization:

- Forests:
$$n = \Theta\left(\max\left\{ \frac{d}{\epsilon^2}, \frac{\sqrt{N}}{\epsilon} \right\} \right)$$
- Chordal graphs:
$$n = \Theta_d(N^{\frac{d}{d+1}} \epsilon^{-\frac{2}{d+1}})$$
provided d is bounded.
- Cliques:
$$n = \Theta\left(\frac{N}{\log N} \log \frac{1}{\epsilon} \right)$$
provided $\epsilon \geq N^{-1/2+\Omega(1)}$. The lower bound part of this statement follows from [7], which shows the optimality of Theorem 3.

3 Chordal graphs

Definition 2. *A perfect elimination ordering (PEO) of a graph G of size N is an ordering of the vertices $(v_1, v_2, \ldots, v_N)$ such that, for each i, $G[N[v_i] \cap \{v_1, \ldots, v_i\}]$ is a clique.*

A classical result of Dirac states that the existence of PEO is in fact the defining property of chordal graphs (cf. e.g. [6, Theorem 5.3.17]).

Theorem 4. *A graph is chordal if and only if it admits a perfect elimination ordering.*

Recall that $\mathsf{ind}(K_i, G)$ denotes the number of cliques of size i in G. For any chordal graph G, it turns out the number of components can be expressed as a linear combination of clique counts (see, e.g., [6, Exercise 5.3.22, p. 231]). The next lemma presents this result together with a sandwich bound.

Lemma 1. *For any chordal graph G,*

$$\mathsf{cc}(G) = \sum_{i \geq 1} (-1)^{i+1} \mathsf{ind}(K_i, G). \tag{1}$$

Furthermore, for any $r \geq 1$,

$$\sum_{i=1}^{2r} (-1)^{i+1} \mathsf{ind}(K_i, G) \leq \mathsf{cc}(G) \leq \sum_{i=1}^{2r-1} (-1)^{i+1} \mathsf{ind}(K_i, G). \tag{2}$$

Before proceeding to the construction of the estimator and its analysis, it is instructive to give a proof of Lemma 1, which illustrates how to count cliques in chordal graphs using vertex elimination. The same technique will be applied in bounding the variance of the estimator.

Proof of Lemma 1. Since G is chordal, by Theorem 4, it has a perfect elimination ordering $(v_1, \ldots, v_N)$. Denote

$$C_j = N(v_j) \cap \{v_1, \ldots, v_{j-1}\}, \quad c_j = |C_j|. \tag{3}$$

Since the neighbors of v_j among $v_1, v_2, \ldots, v_{j-1}$ form a clique, we obtain $\binom{c_j}{i-1}$ new cliques of size i when we adjoin the vertex v_j to the induced subgraph spanned by $v_1, v_2, \ldots, v_{j-1}$. Thus,

$$\mathsf{ind}(K_i, G) = \sum_{j=1}^{N} \tbinom{c_j}{i-1}.$$

Moreover, note that

$$\mathsf{cc}(G) = \sum_{j=1}^{N} \mathbb{I}\{c_j = 0\}.$$

Hence, it follows that

$$\begin{aligned}
\sum_{i=1}^{2r-1}(-1)^{i+1}\mathsf{ind}(K_i,G) &= \sum_{i=1}^{2r-1}(-1)^{i+1}\sum_{j=1}^{N}\tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N}\sum_{i=1}^{2r-1}(-1)^{i+1}\tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N}\sum_{i=0}^{2(r-1)}(-1)^{i}\tbinom{c_j}{i} \\
&= \sum_{j=1}^{N}[\tbinom{c_j-1}{2(r-1)}\mathbb{I}\{c_j\neq 0\}+\mathbb{I}\{c_j=0\}] \\
&\geq \sum_{j=1}^{N}\mathbb{I}\{c_j=0\} \\
&= \mathsf{cc}(G),
\end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^{2r}(-1)^{i+1}\mathsf{ind}(K_i,G) &= \sum_{i=1}^{2r}(-1)^{i+1}\sum_{j=1}^{N}\tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N}\sum_{i=1}^{2r}(-1)^{i+1}\tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N}\sum_{i=0}^{2r-1}(-1)^{i}\tbinom{c_j}{i} \\
&= \sum_{j=1}^{N}[-\tbinom{c_j-1}{2r-1}\mathbb{I}\{c_j\neq 0\}+\mathbb{I}\{c_j=0\}] \\
&\leq \sum_{j=1}^{N}\mathbb{I}\{c_j=0\} \\
&= \mathsf{cc}(G).
\end{aligned}$$

□

The subgraph count identity (1) suggests the following unbiased estimator

$$\widehat{\mathsf{cc}} = -\sum_{i\geq 1}\left(-\frac{1}{p}\right)^{i}\mathsf{ind}(K_i,\tilde{G}). \tag{4}$$

Using elementary enumerative combinatorics, in particular, the vertex elimination structure of chordal graphs, we prove the following bound on its performance.

Theorem 5. *Let G be a chordal graph on N vertices. Suppose $S \subset V(G)$ is generated by Bernoulli(p) sampling and let $\tilde{G} = G[S]$. Then $\widehat{\mathsf{cc}}$ defined in (4) is an unbiased estimator of*

$\mathsf{cc}(G)$. *Furthermore,*

$$\mathsf{Var}(\widehat{\mathsf{cc}}) \le \frac{N}{p^\omega} + \frac{2Nd}{p^{\omega-1}}, \tag{5}$$

where $\omega = \omega(G)$ is the size of the largest clique in G.

Lemma 2. *Let*

$$f(k) = \left(-\frac{q}{p}\right)^k. \tag{6}$$

Let S be a subset of a collection $\mathcal{S}$ of non-empty subsets of $V(G)$. Define $b_S = \sum_{v\in S} b_v$, where $\{b_v\}_{v\in V(G)}$ is a sequence of independent Bern(p) *random variables. Then, for any $S, T \in \mathcal{S}$,*

$$\mathbb{E}[f(b_S)f(b_T)] = \mathbb{I}\{S = T\}(q/p)^{|S|},$$

and any sequence of numbers $\{\alpha_S\}_{S\in\mathcal{S}}$,

$$\mathsf{Var}\sum_{S\in\mathcal{S}} \alpha_S f(b_S) = \sum_{S\in\mathcal{S}} \alpha_S^2 \left(\frac{q}{p}\right)^{|S|}.$$

Remark 1. *The function $f(S) = (-\frac{q}{p})^{b_S}$ is exactly the (unnormalized) orthogonal basis for the binomial measure that is used in Boolean function analysis [4, Definition 8.40].*

Proof of Theorem 5. For a chordal graph G on N vertices, let $(v_1, \dots, v_N)$ be a perfect elimination ordering of G. Recall from (3) that c_j denotes the number of neighbors of v_j among $v_1, \dots, v_{j-1}$. That is, $c_j = \sum_{k=1}^{j-1} \mathbb{I}\{v_k \sim v_j\}$. We also let $\tilde{c}_j$ denote the empirical version; that is

$$\tilde{c}_j = b_j \sum_{k=1}^{j-1} b_k \mathbb{I}\{v_k \sim v_j\},$$

where $b_k = \mathbb{I}\{v_k \in S\}$. Note that $c_j|\{b_j = 1\} \sim \text{Binom}(c_j, p)$ and

$$\mathsf{ind}(K_i, \tilde{G}) = \sum_{j=1}^{N} b_j \tbinom{\tilde{c}_j}{i-1}.$$

and hence

$$\begin{aligned}
\widehat{\mathsf{cc}} &= -\sum_{i=1}^{N} (-\frac{1}{p})^i \mathsf{ind}(K_i, \tilde{G}) \\
&= -\sum_{i=1}^{N} (-\frac{1}{p})^i \sum_{j=1}^{N} b_j \tbinom{\tilde{c}_j}{i-1} \\
&= -\sum_{j=1}^{N} b_j \sum_{i=1}^{N} (-\frac{1}{p})^i \tbinom{\tilde{c}_j}{i-1} \\
&= \frac{1}{p} \sum_{j=1}^{N} b_j \sum_{i=0}^{N-1} (-\frac{1}{p})^i \tbinom{\tilde{c}_j}{i} \\
&= \frac{1}{p} \sum_{j=1}^{N} b_j f(\tilde{c}_j).
\end{aligned}$$

where the function f is defined in (6). To show the variance bound, we note that

$$\mathsf{Var}(\widehat{\mathsf{cc}}) = \frac{1}{p^2}\sum_{j=1}^{N}\mathsf{Var}[b_j f(\tilde{c}_j)] + \frac{2}{p^2}\sum_{j<i}\mathrm{Cov}[b_j f(\tilde{c}_j), b_i f(\tilde{c}_i)]. \tag{7}$$

It is easy to verify that

$$\mathsf{Var}[b_j f(\tilde{c}_j)] = \begin{cases} p\left(\frac{q}{p}\right)^{c_j} & \text{if } c_j > 0 \\ pq & \text{if } c_j = 0 \end{cases}. \tag{8}$$

Since $c_j \le \omega - 1$, it follows that $\mathsf{Var}[b_j f(\tilde{c}_j)] \le p\left(\frac{q}{p}\right)^{\omega-1}$. Moreover, let $A_j = N[v_j] \cap \{v_1, \ldots, v_{j-1}\}$. Then by Lemma 2,

$$\mathrm{Cov}[b_j f(\tilde{c}_j), b_i f(\tilde{c}_i)] = \begin{cases} p^2\left(\frac{q}{p}\right)^{c_j} & \text{if } A_j = A_i \neq \emptyset \\ -pq\left(\frac{q}{p}\right)^{c_j} & \text{if } A_j = A_i \setminus \{v_j\} \neq \emptyset \text{ and } v_j \sim v_i \\ 0 & \text{otherwise} \end{cases}.$$

Thus,

$$\sum_{j<i}\mathrm{Cov}[b_j f(\tilde{c}_j), b_i f(\tilde{c}_i)] \le \sum_{j<i: A_j = A_i \neq \emptyset} p^2\left(\frac{q}{p}\right)^{c_j} \le Ndp^2\left(\frac{q}{p}\right)^{\omega-1}, \tag{9}$$

where the last step follows the fact that $\#\{(i,j) : i > j,\ A_j = A_i \neq \emptyset\} \le Nd$. Finally, combining (7), (8) and (9) yields

$$\mathsf{Var}(\widehat{\mathsf{cc}}) \le \frac{N}{p}\left(\frac{q}{p}\right)^{\omega-1} + 2Nd\left(\frac{q}{p}\right)^{\omega-1}$$

and hence the desired (5). □

Proof of Lemma 2. The second identity follows from the first since $f(N_S)$ and $f(N_T)$ have zero mean. For the first conclusion, assume without loss of generality that $T \subseteq S$. We note that $N_S + N_T = N_{S\setminus T} + 2N_{S\cap T}$, where $N_{S\setminus T}$ and $N_{S\cap T}$ are independent binomial distributed random variables. In particular, $N_{S\setminus T} \sim \mathrm{Binom}(|S \setminus T|, p)$ and thus if $S \neq T$,

$$\begin{aligned}\mathbb{E}[f(N_S)f(N_T)] &= \mathbb{E}[(-\frac{q}{p})^{N_S+N_T}] \\ &= \mathbb{E}[(-\frac{q}{p})^{N_{S\setminus T}+N_{S\cap T}}] \\ &= \mathbb{E}[(-\frac{q}{p})^{N_{S\setminus T}}]\mathbb{E}[(-\frac{q}{p})^{N_{S\cap T}}].\end{aligned}$$

Finally, note that $\mathbb{E}[(-\frac{q}{p})^{N_{S\setminus T}}] = 0$. If $S = T$,

$$\begin{aligned}\mathbb{E}[f(N_S)f(N_T)] &= \mathbb{E}[(-\frac{q}{p})^{N_S+N_T}] \\ &= \mathbb{E}[(-\frac{q}{p})^{2N_S}] \\ &= \left(\frac{q}{p}\right)^{2|S|}.\end{aligned}$$

□

3.1 Forests

Since forests are chordal graphs with clique number $\omega = 2$, we immediately get the following result from Theorem 5.

Theorem 6. *Let G be a forest on N vertices with maximum degree d. Define*

$$\widehat{\mathsf{cc}} = \mathsf{v}(\tilde{G})/p - \mathsf{e}(\tilde{G})/p^2.$$

Then $\widehat{\mathsf{cc}}$ is an unbiased estimator of $\mathsf{cc}(G)$ and

$$\mathsf{Var}(\widehat{\mathsf{cc}}) \le \frac{N}{p^2} + \frac{2Nd}{p}.$$

3.2 Cliques

If the parent graph G consists disjoint union of cliques, so does the sampled graph $\tilde{G}$. Then the estimator (10) can be rewritten as

$$\widehat{\mathsf{cc}} = \sum_{r=1}^{N} \left(1 - \left(-\frac{q}{p}\right)^r\right) \widehat{\mathsf{cc}}_r = \mathsf{cc}(\tilde{G}) - \sum_{r=1}^{N} \left(-\frac{q}{p}\right)^r \widehat{\mathsf{cc}}_r, \tag{10}$$

where $\widehat{\mathsf{cc}}_r$ is the number of components in the sampled graph $\tilde{G}$ that have r vertices. This coincides with the unbiased estimator proposed by Frank [2] for cliques, which is, in turn, based on the estimator of Goodman [3].

The estimator (4) can also be written as $\widehat{\mathsf{cc}} = \sum_{k=1}^{\mathsf{cc}(G)} [1 - (-\frac{q}{p})^{\tilde{N}_k}]$, where $\tilde{N}_k$ is the number of vertices in the k^{th} component in $\tilde{G}$. Using this, it is easy to prove the following theorem.

Theorem 7. *Let G be a union of disjoint cliques with clique number ω. Then $\widehat{\mathsf{cc}}$ is an unbiased estimator of $\mathsf{cc}(G)$ and*

$$\mathsf{Var}(\widehat{\mathsf{cc}}) = \sum_{r=1}^{N} \left(\frac{q}{p}\right)^r \mathsf{cc}_r \le \frac{N}{p^\omega},$$

where cc_r is the number of connected components in G of size r.

4 Smoothing

4.1 Cliques

When the sampling ratio p is small, the coefficients of the unbiased estimator (10) grow exponentially and result in large variance. Using a technique known as *smoothing* introduced in [5], we modify this estimator to achieve a near-optimal bias-variance tradeoff. The key observation is the following: consider the truncated version of the unbiased estimator (10) by summing over all clique size up to ℓ. In view of the sandwich bound in (2), if ℓ is odd (resp. even), the estimator is positively (resp. negatively) biased. By averaging over all truncated versions according to an appropriately chosen distribution, the bias cancels each other and is dramatically reduced. To this end, consider a discrete random variable $L \in \mathbb{N}$ and define the following estimator by truncating (10) at the random location L and average over its distribution:

$$\widehat{\mathsf{cc}}_L \triangleq \mathsf{cc}(\tilde{G}) - \mathbb{E}_L \left[\sum_{r=1}^{L} (-\tfrac{q}{p})^r \widehat{\mathsf{cc}}_r\right],$$

which can be simplified to the following "smoothed" estimator:

$$\widehat{\mathsf{cc}}_L = \mathsf{cc}(\tilde{G}) - \sum_{r=1}^{N} \left(-\frac{q}{p}\right)^r \mathbb{P}(L \geq r)\widehat{\mathsf{cc}}_r. \tag{11}$$

Therefore effectively smoothing acts as soft truncation by introducing a tail probability that modulates the exponential growth of the original coefficients. The variance can then be bounded simply by the ℓ_∞-norm of the coefficients.

Theorem 8. *Suppose the parent graph is a disjoint union of cliques and S is generated by Bernoulli(p). Assume that $p < 1/2$ and $N > 4$. Let $L \sim Pois(\lambda)$, where $\lambda = \frac{p}{2q-p}\log(\frac{N}{4})$. Then*

$$\frac{\mathbb{E}_G(\widehat{\mathsf{cc}}_L - \mathsf{cc}(G))^2}{N^2} \leq (N/4)^{-\frac{p}{2q-p}}.$$

Remark 2. *The estimator $\widehat{\mathsf{cc}}_L$ is valid as long as N is known.*

Remark 3. *Because $\widehat{\mathsf{cc}}_L$ is not guaranteed to satisfy $0 \leq \widehat{\mathsf{cc}}_L \leq N$, we might want to redefine it as $\widehat{\mathsf{cc}} = (\widehat{\mathsf{cc}}_L \vee 0) \wedge N$. Alternatively, since $\mathsf{cc}(\tilde{G}) \leq \mathsf{cc}(G)$, we can consider $\widehat{\mathsf{cc}} = (\widehat{\mathsf{cc}}_L \vee \mathsf{cc}(\tilde{G})) \wedge N$.*

Proof. The bias of this estimator is seen to be

$$\widehat{\mathsf{cc}}_L - \mathsf{cc}(G) = \sum_{k=1}^{\mathsf{cc}(G)} \mathbb{E}[\mathbb{P}(L < \tilde{N}_k)(-\frac{q}{p})^{\tilde{N}_k}].$$

Note that $\mathbb{P}(L < r) = \sum_{i=0}^{r-1} \mathbb{P}(L = i)$ so that

$$\sum_{r=1}^{N} \mathbb{P}(L < r) \left(-\frac{q}{p}\right)^r \mathbb{P}(\tilde{N}_k = r) = \sum_{i=0}^{N-1} \mathbb{P}(L = i) \sum_{r=i+1}^{N} \left(-\frac{q}{p}\right)^r \mathbb{P}(\tilde{N}_k = r)$$

Since $\tilde{N}_k \sim \mathrm{Bin}(N_k, p)$, it follows that

$$\sum_{r=i+1}^{N} \left(-\frac{q}{p}\right)^r \mathbb{P}(\tilde{N}_k = r) = q^{N_k} \sum_{r=i+1}^{N} \binom{N_k}{r} (-1)^r.$$

We also have the identity

$$\sum_{r=i+1}^{N} \binom{N_k}{r} (-1)^r = (-1)^{i+1} \binom{N_k - 1}{i}.$$

Putting these facts together, we have

$$\begin{aligned} |\widehat{\mathsf{cc}}_L - \mathsf{cc}(G)| &\leq \max_{1 \leq k \leq \mathsf{cc}(G)} \left| \mathbb{E}_L \left[(-1)^L \binom{N_k - 1}{L} \right] \right| \sum_{k=1}^{K} q^{N_k} \\ &\leq \sqrt{N} \max_{1 \leq k \leq \mathsf{cc}(G)} \left| \mathbb{E}_L \left[(-1)^L \binom{N_k - 1}{L} \right] \right| \sqrt{\sum_{k=1}^{K} q^{N_k}}, \end{aligned}$$

since $\mathsf{cc}(G) \leq N$. For the variance of $\widehat{\mathsf{cc}}_L$, note that $\widehat{\mathsf{cc}}_L = \sum_{k=1}^{K} W_k$, where $W_k = 1 - \mathbb{P}(L \geq \tilde{N}_k)(-\frac{q}{p})^{\tilde{N}_k}$. The W_k are independent random variables and hence

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) = \sum_{k=1}^{K} \mathsf{Var}(W_k) \leq \sum_{k=1}^{K} \mathbb{E}W_k^2.$$

Also,

$$W_k^2 \leq \max_{1 \leq r \leq N} (1 - \mathbb{P}(L \geq r) \left(-\frac{q}{p}\right)^r)^2 \mathbb{I}\{\tilde{N}_k \geq 1\}.$$

This means that

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \leq \max_{1 \leq r \leq N} (1 - \mathbb{P}(L \geq r) \left(-\frac{q}{p}\right)^r)^2 \sum_{k=1}^{\mathsf{cc}(G)} (1 - q^{N_k}).$$

If $p \leq 1/2$,

$$\begin{aligned}
\mathbb{P}(L \geq r)(\frac{q}{p})^r &= \sum_{i=r}^{\infty} \mathbb{P}(L = i)(\frac{q}{p})^r \\
&\leq \sum_{i=r}^{\infty} \mathbb{P}(L = i)(\frac{q}{p})^i \\
&\leq \sum_{i=0}^{\infty} \mathbb{P}(L = i)(\frac{q}{p})^i \\
&= \mathbb{E}_L(\frac{q}{p})^L.
\end{aligned}$$

Thus, it follows that

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \leq 4(\mathbb{E}_L(\frac{q}{p})^L)^2 \sum_{k=1}^{\mathsf{cc}(G)} (1 - q^{N_k}).$$

We have shown that

$$\begin{aligned}
\mathbb{E}(\widehat{\mathsf{cc}}_L - \mathsf{cc}(G))^2 \leq 4(\mathbb{E}_L(\frac{q}{p})^L)^2 \sum_{k=1}^{\mathsf{cc}(G)} (1 - q^{N_k}) + \\
N \max_{1 \leq k \leq \mathsf{cc}(G)} \left| \mathbb{E}_L \left[(-1)^L \binom{N_k - 1}{L} \right] \right|^2 \sum_{k=1}^{\mathsf{cc}(G)} q^{N_k}.
\end{aligned}$$

If $L \sim \text{Pois}(\lambda)$, then $\mathbb{E}_L(\frac{q}{p})^L = e^{\lambda(\frac{q}{p}-1)}$ and

$$\begin{aligned}
\mathbb{E}_L \left[(-1)^L \binom{N_k - 1}{L} \right] &= e^{-\lambda} \sum_{i=0}^{N_k - 1} \binom{N_k - 1}{i} \frac{(-\lambda)^i}{i!} \\
&= e^{-\lambda} L_{N_k - 1}(\lambda),
\end{aligned}$$

where L_m is the Laguerre polynomial of degree m. It is a classical fact [1] that $|L_m(x)| \leq e^{x/2}$ for all $m \geq 0$ and $x \geq 0$. Thus it follows that $\left|\mathbb{E}_L\left[(-1)^L\binom{N_k-1}{L}\right]\right| \leq e^{-\lambda/2}$. This means that the bound on $\mathbb{E}(\widehat{\mathsf{cc}}_L - \mathsf{cc}(G))^2$ reduces to

$$4e^{2\lambda(\frac{q}{p}-1)}\sum_{k=1}^{\mathsf{cc}(G)}(1-q^{N_k}) + Ne^{-\lambda}\sum_{k=1}^{\mathsf{cc}(G)}q^{N_k}.$$

If we set $4e^{2\lambda(\frac{q}{p}-1)} = Ne^{-\lambda}$. It follows that

$$\mathbb{E}\frac{(\widehat{\mathsf{cc}}_L - \mathsf{cc}(G))^2}{N^2} \leq e^{-\lambda}.$$

The solution to λ produces the desired bound. □

4.2 Chordal graphs

Lemma 3. *Let $(v_1, \ldots, v_N)$ be a perfect elimination ordering of a chordal graph G on N vertices with maximal degree d and maximum clique size ω. Let $A_j = N[v_j] \cap \{v_1, \ldots, v_j\}$. Then,*

$$\sum_{j<i}\mathbb{I}\{A_i \cap A_j \neq \emptyset\} \leq Nd\omega.$$

Proof of Lemma 3. Let $\alpha_j = |A_j|$. We will prove that for a fixed j,

$$\sum_{i=j+1}^{N}\mathbb{I}\{A_i \cap A_j \neq \emptyset\} \leq (d+1-\alpha_j)\alpha_j \leq d\omega.$$

We will prove this by contradiction. Suppose

$$\sum_{i=j+1}^{N}\mathbb{I}\{A_i \cap A_j \neq \emptyset\} \geq (d+1-\alpha_j)\alpha_j + 1$$

Then at least $(d+1-\alpha_j)\alpha_j + 1$ of the A_i have nonempty intersection with A_j. Note that $\deg(v) \geq \alpha_j - 1$ for each $v \in A_j$ by definition (since the vertices in A_j form a clique of size α_j in G). By the pigeonhole principle, there is at least one vertex u in A_j such that $\deg(u) \geq (\alpha_j - 1) + (d - \alpha_j + 2) = d + 1$. This is a contradiction to the bounded degree assumption. Thus,

$$\sum_{i=j+1}^{N}\mathbb{I}\{A_i \cap A_j \neq \emptyset\} \leq (d+1-\alpha_j)\alpha_j,$$

and the conclusion follows by noting that $1 \leq \alpha_j \leq d$. □

Lemma 4. *Let G be a graph with maximum degree d. Then*

$$\mathsf{Var}(\mathsf{ind}(K_i, \tilde{G})) \leq N\sum_{r=1}^{i}(dp)^{2i-r} \lesssim \begin{cases} N(dp)^i & \text{if } p \leq 1/d \\ N(dp)^{2i-1} & \text{if } p > 1/d \end{cases}.$$

Theorem 9. *Let $L \sim Poisson(\lambda)$ and let*

$$\widehat{\mathsf{cc}}_L = \frac{1}{p}\sum_{j=1}^{N} b_j(-\frac{q}{p})^{\tilde{c}_j}\mathbb{P}(L \geq \tilde{c}_j),$$

and

$$\widetilde{\mathsf{cc}}_L = -\sum_{i=1}^{N}(-\frac{1}{p})^i \mathsf{ind}(K_i, \tilde{G})\mathbb{P}(L \geq i-1).$$

Then

$$|\mathbb{E}[\widehat{\mathsf{cc}}_L - K]| \leq Ne^{-\lambda/2},$$
$$|\mathbb{E}[\widetilde{\mathsf{cc}}_L - K]| \leq Ne^{-\lambda/2},$$

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \lesssim Ne^{2\lambda(\sqrt{\frac{d}{p}}-1)},$$

and

$$\mathsf{Var}(\widetilde{\mathsf{cc}}_L) \lesssim Nd\omega e^{2\lambda(\frac{q}{p}-1)}.$$

Moreover,

$$\frac{\mathbb{E}(\widehat{\mathsf{cc}}_L - \mathsf{cc}(G))^2}{N^2} \lesssim (\frac{d\omega}{N})^{\frac{p}{2q-p}}.$$

and there exists a universal constant $c > 0$ such that

$$\frac{\mathbb{E}(\widetilde{\mathsf{cc}}_L - \mathsf{cc}(G))^2}{N^2} \lesssim N^{-c(\sqrt{\frac{p}{d}}\wedge\frac{1}{d})}.$$

Proof. Note that

$$
\begin{aligned}
\mathbb{E}[\widetilde{\mathsf{cc}}_L - \mathsf{cc}(G)] &= \sum_{i=1}^{d+1} \mathbb{P}(L < i-1)(-\frac{1}{p})^i \mathbb{E}\mathsf{ind}(K_i, \tilde{G}) \\
&= \sum_{i=1}^{d+1} \mathbb{P}(L < i-1)(-1)^i \mathsf{ind}(K_i, \tilde{G}) \\
&= \sum_{i=1}^{d+1} \mathbb{P}(L < i-1)(-1)^i \sum_{j=1}^{N} \tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N} \sum_{i=1}^{d+1} \mathbb{P}(L < i-1)(-1)^i \tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N} \sum_{i=1}^{d+1} \sum_{\ell=0}^{i-2} \mathbb{P}(L = \ell)(-1)^i \tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N} \sum_{\ell=0}^{d-1} \mathbb{P}(L = \ell) \sum_{i=\ell+2}^{d+1} (-1)^i \tbinom{c_j}{i-1} \\
&= \sum_{j=1}^{N} \sum_{\ell=0}^{d-1} \mathbb{P}(L = \ell)(-1)^\ell \tbinom{c_j - 1}{\ell} \\
&= e^{-\lambda} \sum_{j=1}^{N} L_{c_j - 1}(\lambda)
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}[\mathsf{cc}(G) - \widehat{\mathsf{cc}}_L] &= \frac{1}{p} \sum_{j=1}^{N} \mathbb{E}[b_j (-\frac{q}{p})^{\tilde{c}_j} \mathbb{P}(L < \tilde{c}_j)] \\
&= \sum_{j=1}^{N} \sum_{i=0}^{c_j} \tbinom{c_j}{i} p^i q^{c_j - i} (-\frac{q}{p})^i \mathbb{P}(L < i) \\
&= \sum_{j=1}^{N} q^{c_j} \sum_{i=0}^{c_j} \tbinom{c_j}{i} (-1)^i \mathbb{P}(L < i) \\
&= \sum_{j=1}^{N} q^{c_j} \sum_{i=0}^{c_j} \tbinom{c_j}{i} (-1)^i \sum_{\ell=0}^{i-1} \mathbb{P}(L = \ell) \\
&= \sum_{j=1}^{N} q^{c_j} \sum_{\ell=0}^{c_j - 1} \mathbb{P}(L = \ell) \sum_{i=\ell+1}^{c_j} \tbinom{c_j}{i} (-1)^i \\
&= \sum_{j=1}^{N} q^{c_j} \sum_{\ell=0}^{c_j - 1} \mathbb{P}(L = \ell) \tbinom{c_j - 1}{\ell} (-1)^{\ell+1} \\
&= -e^{-\lambda} \sum_{j=1}^{N} q^{c_j} L_{c_j - 1}(\lambda),
\end{aligned}
$$

where L_{c_j-1} is the Laguerre polynomial of order $c_j - 1$. It is a classical fact [1] that $|L_n(x)| \le e^{x/2}$ for all $m \ge 0$ and $x \ge 0$. Thus we see that $|\mathbb{E}[\widetilde{\mathsf{cc}}_L - \widehat{\mathsf{cc}}]| \le Ne^{-\lambda/2}$ and $|\mathbb{E}[\widehat{\mathsf{cc}}_L - \widehat{\mathsf{cc}}]| \le Ne^{-\lambda/2}$.

The variance calculation for $\widetilde{\mathsf{cc}}_L$ uses the inequality

$$\mathsf{Var}(\widetilde{\mathsf{cc}}_L) \le (\sum_{i=1}^{d+1} \mathbb{P}(L \ge i-1)(\frac{1}{p})^i \sqrt{\mathsf{Var}(\mathsf{ind}(K_i, \tilde{G}))})^2.$$

Using Lemma 4, we can further bound the variance of $\widetilde{\mathsf{cc}}_L$ in the two regimes $p \le 1/d$ or $p > 1/d$. □

Remark 4. *In the case that the parent graph is a union of cliques, $\widehat{\mathsf{cc}}_L$ can be written as*

$$\frac{1}{p}\sum_{k=1}^{\mathsf{cc}(G)}\sum_{j=1}^{\tilde{N}_k}(-\frac{q}{p})^{j-1}\mathbb{P}(L \ge j-1) = \frac{1}{p}\sum_{k=1}^{\mathsf{cc}(G)}\sum_{j=0}^{\tilde{N}_k-1}(-\frac{q}{p})^{j}\mathbb{P}(L \ge j),$$

and hence

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \le \frac{1}{p^2}\sum_{k=1}^{\mathsf{cc}(G)}(\sum_{j=0}^{N_k-1}(\frac{q}{p})^j\mathbb{P}(L \ge j)\sqrt{\mathsf{Var}(\mathbb{I}\{j \le \tilde{N}_k - 1\})})^2.$$

By Markov's inequality,

$$\mathsf{Var}(\mathbb{I}\{j \le \tilde{N}_k - 1\}) \le \mathbb{P}(\tilde{N}_k \ge j+1) \le \frac{\mathbb{E}\tilde{N}_k}{j+1} = \frac{pN_k}{j+1}.$$

Thus,

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \le \frac{1}{p}\sum_{k=1}^{K} N_k(\sum_{j=0}^{N_k-1}(\frac{q}{p})^j\frac{\mathbb{P}(L \ge j)}{\sqrt{j+1}})^2.$$

Next, observe that

$$\sum_{j=0}^{\infty}(\frac{q}{p})^j\mathbb{P}(L \ge j) = \frac{\frac{q}{p}e^{\lambda(\frac{q}{p}-1)} - 1}{\frac{q}{p} - 1},$$

and so the variance is bounded by a constant multiple of $N\frac{e^{2\lambda(\frac{q}{p}-1)}}{p}$. This variance bound is very similar to the other smoothed estimator (11). In fact both achieve the rate in Theorem 8.

5 Lower bounds

5.1 General strategy

We begin with a lemma.

Lemma 5. *Suppose H is a disconnected subgraph of G and the number of vertices in H is v. Then $\mathsf{ind}(H, G)$ can be expressed as a polynomial, independent of G, in $\mathsf{ind}(g, G)$ where either g is connected and $v(g) \le v$ or g is disconnected and $v(g) \le v - 1$. Consequently, $\mathsf{ind}(H, G)$ can be expressed as a polynomial, independent of G, in $\mathsf{ind}(g, G)$ where g is connected and $v(g) \le v$.*

Proof. We use Kocay's Vertex Lemma which says that if $\mathcal{H}$ is a collection of graphs, then

$$\prod_{h\in\mathcal{H}} \mathsf{ind}(h,G) = \sum_{g} a_g \mathsf{ind}(g,G),$$

where the sum runs over all graphs g such that $v(g) \le \sum_{h\in\mathcal{H}} v(h)$ and a_g is the number of decompositions of $V(g)$ into $\cup_{h\in\mathcal{H}} V(h)$ such that $g[V(h)] \equiv h$.

In particular, if $\mathcal{H}$ consists of the connected components of H, then the only disconnected g with $v(g) = v$ satisfying the above decomposition property is $g = H$. For this g, $a_g = 1$. Thus, either g is connected and $v(g) \le v$ or g is disconnected and $v(g) \le v-1$.

Let $S(v)$ be the statement about v that for any disconnected subgraph H of G with at most v vertices, $\mathsf{ind}(H,G)$ can be expressed as a polynomial, independent of G, in $\mathsf{ind}(g,G)$ where g is connected and $v(g) \le v$. The base case $S(1)$ is clearly true. Next, suppose $S(v)$ is true. We will show that $S(v+1)$ is also true. By the first assertion of the lemma, $\mathsf{ind}(H,G)$ can be expressed as a polynomial, independent of G, in $\mathsf{ind}(h,G)$ where either h is connected and $v(h) \le v+1$ or h is disconnected and $v(h) \le v$. By $S(v)$, each $\mathsf{ind}(h,G)$ with h disconnected and $v(h) \le v$ can be expressed as a polynomial, independent of G, in $\mathsf{ind}(g,G)$ where g is connected and $v(g) \le v$. Thus, we can express $\mathsf{ind}(H,G)$ as a polynomial, independent of G, in terms of $\mathsf{ind}(g,G)$ where g is connected and $v(g) \le v+1$. □

Corollary 1. *Suppose H and H' are two graphs in which $\mathsf{ind}(h,H) = \mathsf{ind}(h,H')$ for all connected h with $v(h) \le v$. Then $\mathsf{ind}(h,H) = \mathsf{ind}(h,H')$ for all h with $v(h) \le v$.*

Next we give a general lower bound involving a pair of graphs with matching lower-order subgraph counts.

Theorem 10. *Let $\mathcal{G}$ be a class of graphs on N vertices that is closed under disjoint graph unions. Let H and H' be two graphs in $\mathcal{G}$ with $m \in \mathbb{N}$ vertices such that*

$$\mathsf{ind}(h,H) = \mathsf{ind}(h,H')$$

for all connected *graphs h with at most $k \in [m]$ vertices. Suppose $m \le N/500$ and $TV(P_{\tilde{H}}, P_{\tilde{H'}}) \le 1/500$. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}} \mathbb{P}(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \ge \Delta) \ge 0.075.$$

where

$$\Delta = \frac{|\mathsf{cc}(H) - \mathsf{cc}(H')|}{8}\Big(\sqrt{\frac{N}{m\,TV(P_{\tilde{H}}, P_{\tilde{H'}})}} \wedge \frac{N}{m}\Big).$$

Moreover,

$$TV(P_{\tilde{H}}, P_{\tilde{H'}}) \le \sum_{v=k+1}^{m} \tbinom{m}{v} p^v q^{m-v}.$$

Proof. Let $M = N/m$ and $G = G_1 + G_2 + \cdots + G_M$, where $G_i = \begin{cases} H & \text{with probability } \alpha \\ H' & \text{with probability } 1-\alpha \end{cases}$. Assume without loss of generality that $\mathsf{cc}(H) > \mathsf{cc}(H')$. Note that $\mathbb{E}_\alpha \mathsf{cc}(G) = M[\alpha \mathsf{cc}(H) + (1-\alpha)\mathsf{cc}(H')]$. Let $\alpha_0 = 1/2$ and $\alpha_1 = 1/2 + \delta$, where $\delta \in [0, 1/2]$. Define $\Delta \triangleq [\mathbb{E}_{\alpha_1}\mathsf{cc}(G) - \mathbb{E}_{\alpha_0}\mathsf{cc}(G)]/4 = M(\delta/4)[\mathsf{cc}(H) - \mathsf{cc}(H')]$.

Let g_i be the portion of $\tilde{G}$ that comes from G_i. Then $g_i = h + (m - v(h))K_1$, where h contains no isolated vertices. In this case,

$$\mathbb{P}(g_i = h + (m - v(h))K_1 | G_i \sim H) = \sum_{0 \le j \le m - v(h)} \mathsf{ind}(h + jK_1, H) p^{v(h)+j} q^{m - v(h) - j},$$

and

$$\mathbb{P}(g_i = h + (m - v(h))K_1 | G_i \sim H') = \sum_{0 \le j \le m - v(h)} \mathsf{ind}(h + jK_1, H') p^{v(h)+j} q^{m - v(h) - j}.$$

Let $P = \mathcal{L}(g_i | G_i \sim H)$ and $P' = \mathcal{L}(g_i | G_i \sim H')$ so that $P_\alpha \triangleq \mathcal{L}(g_i) = \alpha P + (1 - \alpha)P'$.

We desire a method to lower bound the minimax mean square error of the functional $G \mapsto \mathsf{cc}(G)$. To apply the method of two fuzzy hypotheses, we must ensure that there exists $L > 0$ and $\beta_0, \beta_1 \in (0, 1)$ such that

$$\mathbb{P}_{\alpha_0}(\mathsf{cc}(G) \le L) = \mathbb{P}(\mathrm{Binom}(M, \alpha_0) \le L) \ge 1 - \beta_0,$$

and

$$\mathbb{P}_{\alpha_1}(\mathsf{cc}(G) \ge L + 2\Delta) = \mathbb{P}(\mathrm{Binom}(M, \alpha_1) \ge L + 2\Delta) \ge 1 - \beta_1.$$

Note that $\mathsf{cc}(G) \sim |\mathsf{cc}(H) - \mathsf{cc}(H')|\mathrm{Bin}(M, \alpha) + \mathsf{cc}(H')M$. Choose $L = \mathsf{cc}(H)(1/2 + \delta/4)M + \mathsf{cc}(H')(1/2 - \delta/4)M$ so that

$$\mathbb{P}_{\alpha_0}(\mathsf{cc}(G) \le L) = \mathbb{P}(\mathrm{Binom}(M, \alpha_0) \le M\alpha_0(1 + \delta/2)) \ge 1 - e^{-\delta^2 M/12}.$$

and

$$\mathbb{P}_{\alpha_1}(\mathsf{cc}(G) \ge L + 2\Delta) = \mathbb{P}(\mathrm{Binom}(M, \alpha_1) \ge M\alpha_1(1 - \delta/4)) \ge 1 - e^{-\delta^2 M/32}.$$

The first observation we make is that $(g_1, g_2, \ldots, g_M)' \overset{iid}{\sim} P_\alpha^{\otimes M}$. By Theorem 2.15(i) in [cite Tsybakov],

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}} \mathbb{P}(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \ge \Delta) \ge \frac{1 - \mathrm{TV}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) - \beta_0 - \beta_1}{2}.$$

To ensure that

$$\mathrm{TV}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) < 1 - \beta_0 - \beta_1,$$

we use the fact that

$$\begin{aligned}
\mathrm{TV}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) &\le 1 - \frac{1}{2}\exp\{-\chi^2(P_{\alpha_0}^{\otimes M} || P_{\alpha_1}^{\otimes M})\} \\
&= 1 - \frac{1}{2}\exp\{-(1 + \chi^2(P_{\alpha_0} || P_{\alpha_1}))^M + 1\}
\end{aligned}$$

and

$$\begin{aligned}
\chi^2(P_{\alpha_0} || P_{\alpha_1}) &= \chi^2(\frac{P + P'}{2} + \delta(P - P') || \frac{P + P'}{2}) \\
&= \delta^2 \int \frac{(P - P')^2}{\frac{P + P'}{2}} \\
&\le 4\delta^2 \mathrm{TV}(P, P').
\end{aligned}$$

In summary,

$$\mathrm{TV}(P_{\alpha_0}, P_{\alpha_1}) \leq 1 - \frac{1}{2}\exp\{-(1+4\delta^2\mathrm{TV}(P,P'))^M + 1\}.$$

Choosing $\delta = \frac{1}{2} \wedge (\sqrt{\frac{1}{4M\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}})$, $M \geq 500$, and $\mathrm{TV}(P,P') \leq 1/500$ produces

$$\frac{1}{2}\exp\{-(1+4\delta^2\mathrm{TV}(P,P'))^M + 1\} - e^{-\delta^2 M/32} - e^{-\delta^2 M/12} > 0.15.$$

By Corollary 1,

$$\mathsf{ind}(h, H) = \mathsf{ind}(h, H'),$$

for all h with $v(h) \leq k$ (h not necessarily connected). We can calculate $\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = \mathrm{TV}(P,P')$ explicitly as

$$\sum_h |\sum_{j \leq m - v(h)} [\mathsf{ind}(h + jK_1, H) - \mathsf{ind}(h + jK_1, H')] p^{v(h)+j} q^{m-v(h)-j}|,$$

where the outer sum runs over all h that do not contain any isolated vertices. Since $\mathsf{ind}(h + jK_1, H) = \mathsf{ind}(h + jK_1, H')$ for all h with $v(h) + j \leq k$, it follows that $\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})$ is in fact equal to

$$\frac{1}{2}\sum_h |\sum_{k+1-v(h) \leq j \leq m-v(h)} [\mathsf{ind}(h + jK_1, H) - \mathsf{ind}(h + jK_1, H')] p^{v(h)+j} q^{m-v(h)-j}|.$$

The triangle inequality can be used to further bound $\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})$ by

$$\frac{1}{2}\sum_h \sum_{k+1-v(h) \leq j \leq m-v(h)} [\mathsf{ind}(h + jK_1, H) + \mathsf{ind}(h + jK_1, H')] p^{v(h)+j} q^{m-v(h)-j}.$$

This sum is seen to be exactly equal to

$$\frac{1}{2}\sum_{g: v(g) \geq k+1} [\mathsf{ind}(g, H) + \mathsf{ind}(g, H')] p^{v(g)} q^{m-v(g)} = \sum_{v=k+1}^{m} \binom{m}{v} p^v q^{m-v},$$

since

$$\begin{aligned}
\sum_{g: v(g)=v} \mathsf{ind}(g, H) &= \sum_{g: v(g)=v} [\sum_{|S|=v} \mathbb{I}\{H[S] \sim g\}] \\
&= \sum_{|S|=v} [\sum_{g: v(g)=v} \mathbb{I}\{H[S] \sim g\}] \\
&= \sum_{|S|=v} 1 \\
&= \binom{m}{v}.
\end{aligned}$$

□

5.2 Lower bound for forests

Theorem 11. *Let $\mathcal{T}$ denote the collection of all trees on N vertices with maximum degree d. Suppose we sample vertices according to a Bernoulli(p) model. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{T}} \frac{\mathbb{E}(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \asymp (\frac{d}{Np} \vee \frac{1}{Np^2}) \wedge 1.$$

Suppose we sample vertices according to the SRS(v) model. Then

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{T}} \mathbb{E}\frac{(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \asymp (\frac{d}{v} \vee \frac{N}{v^2}) \wedge 1.$$

Proof. We will prove the lower bound via Theorem 10. Let H consist of $d+1$ isolated vertices. Let H' be a d-star. In the notation of Theorem 10, $k = 1$, $m = d+1$, and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = d$. Note that if h is an isolated vertex, then

$$\mathbb{P}(g_i = h + (m - v(h))K_1 | G_i \sim H) = 1$$

and

$$\mathbb{P}(g_i = h + (m - v(h))K_1 | G_i \sim H') = q + pq^w.$$

Moreover, if h is an ℓ-star with $\ell \geq 1$, then

$$\mathbb{P}(g_i = h + (m - v(h))K_1 | G_i \sim H') = \tbinom{d}{\ell} p^{\ell+1} q^{d-\ell}.$$

Thus,

$$2\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) = 1 - (q + pq^d) + \sum_{\ell=1}^{d} \tbinom{d}{\ell} p^{\ell+1} q^{d-\ell} = 2p(1 - q^d).$$

Thus, by Theorem 10,

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}} \mathbb{P}(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \geq \Delta) \geq c.$$

where

$$\Delta \asymp \left(\sqrt{\frac{Nd}{p(1-q^d)}}\right) \wedge N.$$

If $p > 1/d$, then $\sqrt{\frac{p(1-q^d)}{Nd}} \approx \sqrt{\frac{p}{Nd}}$. If $p \leq 1/d$, then $\sqrt{\frac{p(1-q^d)}{Nd}} \approx \sqrt{\frac{p^2}{N}}$.

□

5.3 Lower bound for graphs with cycles

Theorem 12. *Let $\mathcal{G}$ denote the collection of all r-chordal graphs on N vertices, $r \geq 4$ and $m = 2r - 1$. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}} \mathbb{E}\frac{(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \gtrsim \frac{1}{N2^m p^r} \wedge \frac{1}{m^2}.$$

Proof. We will prove the lower bound via Theorem 10. Let $H = C_r + P_{r-1}$ and $H' = P_{2r-1}$. Note that $\mathsf{ind}(P_i, H) = \mathsf{ind}(P_i, H') = 2r - i$ for $i = 1, 2, \ldots, r-1$. Since paths of length at most $r-1$ are the only connected subgraphs of H and H', Corollary 1 implies that H and H' have matching subgraph counts up to order $r - 1$.

In the notation of Theorem 10, $k = r-1$, $m = 2r-1$, and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = 1$. By Theorem 10, there exists a universal positive constant c such that

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}} \mathbb{P}(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \geq \Delta) \geq c.$$

where

$$\Delta \asymp |\mathsf{cc}(H) - \mathsf{cc}(H')| (\sqrt{\frac{N}{m\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m}) = \sqrt{\frac{N}{m\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m}.$$

$$\begin{aligned} \mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) &\leq \sum_{v=r}^{m} \tbinom{m}{v} p^v q^{m-v} \\ &\leq (4p)^r. \end{aligned}$$

The desired lower bound on the NMSE follows from Markov's inequality.

□

5.4 Lower bound for chordal graphs

Lemma 6. *There exist two chordal graphs H and H' on $(d+1)2^{d-1}$ vertices with maximum degree d such that $\mathsf{ind}(h, H) = \mathsf{ind}(h, H')$ for all h with $v(h) \leq d$ and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = 1$.*

Proof. We construct two chordal graphs H and H' with the desired properties. If d is even, we set

$$H = \tbinom{d+1}{d+1} K_{d+1} + \tbinom{d+1}{d-1} K_{d-1} + \tbinom{d+1}{d-3} K_{d-3} + \tbinom{d+1}{d-5} K_{d-5} + \cdots + \tbinom{d+1}{1} K_1,$$

and

$$H' = \tbinom{d+1}{d} K_d + \tbinom{d+1}{d-2} K_{d-2} + \tbinom{d+1}{d-4} K_{d-4} + \tbinom{d+1}{d-6} K_{d-6} + \cdots + \tbinom{d+1}{2} K_2.$$

If d is odd, we set

$$H = \tbinom{d+1}{d+1} K_{d+1} + \tbinom{d+1}{d-1} K_{d-1} + \tbinom{d+1}{d-3} K_{d-3} + \tbinom{d+1}{d-5} K_{d-5} + \cdots + \tbinom{d+1}{2} K_2,$$

and

$$H' = \tbinom{d+1}{d} K_d + \tbinom{d+1}{d-2} K_{d-2} + \tbinom{d+1}{d-4} K_{d-4} + \tbinom{d+1}{d-6} K_{d-6} + \cdots + \tbinom{d+1}{1} K_1.$$

Thus,

$$|\mathsf{cc}(H) - \mathsf{cc}(H')| = |\sum_{k=1}^{d+1} (-1)^k \tbinom{d+1}{k}| = 1,$$

and for $i = 1, 2, \ldots, d$,

$$|\mathsf{ind}(K_i, H) - \mathsf{ind}(K_i, H')| = |\sum_{k=i}^{d+1} (-1)^k \tbinom{d+1}{k} \tbinom{k}{i}| = 0.$$

Thus for $i = 1, 2, \ldots, d$,

$$\mathsf{ind}(K_i, H) = \mathsf{ind}(K_i, H') = \frac{1}{2} \sum_{k=i}^{d+1} \tbinom{d+1}{k} \tbinom{k}{i} = 2^{d-i} \tbinom{d+1}{i}.$$

In particular, the number of vertices in H and H' is $(d+1)2^{d-1}$. Note that the complete graphs with at most d vertices are the only connected vertex induced subgraphs of H and H' with at most d vertices. The claim follows from Corollary 1. □

Theorem 13. *Let $\mathcal{G}$ denote the collection of all chordal graphs with maximum degree d and set $m = (d+1)2^{d-1}$. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}} \mathbb{E}\frac{(\widehat{\mathsf{cc}} - \mathsf{cc}(G))^2}{N^2} \gtrsim \frac{1}{N2^m p^{d+1}} \wedge \frac{1}{m^2}.$$

Proof. We will prove the lower bound via Theorem 10.

By Lemma 6, there exist two chordal graphs H and H' on $(d+1)2^{d-1}$ vertices with clique number $d+1$, $\mathsf{ind}(h, H) = \mathsf{ind}(h, H')$ for all h with $v(h) \le d$ and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = 1$. In the notation of Theorem 10, $k = d$, $m = (d+1)2^{d-1}$, and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = 1$.

By Theorem 10, there exists a universal positive constant c such that

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}} \mathbb{P}(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \ge \Delta) \ge c.$$

where

$$\Delta \asymp |\mathsf{cc}(H) - \mathsf{cc}(H')|(\sqrt{\frac{N}{m\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m}) = \sqrt{\frac{N}{m\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'})}} \wedge \frac{N}{m}.$$

$$\begin{aligned}\mathrm{TV}(P_{\tilde{H}}, P_{\tilde{H}'}) &\le \sum_{v=d+1}^{m} \tbinom{m}{v} p^v q^{m-v} \\ &\le p^{d+1}2^m.\end{aligned}$$

The desired lower bound on the NMSE follows from Markov's inequality. □

References

[1] Milton Abramowitz and Irene A. Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *National Bureau of Standards Applied Mathematics Series*. For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 1964. 10, 13

[2] Ove Frank. Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Statist.*, 5(4):177–188, 1978. 2, 7

[3] Leo A. Goodman. On the estimation of the number of classes in a population. *Ann. Math. Statistics*, 20:572–579, 1949. 7

[4] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, 2014. 5

[5] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Estimating the number of unseen species: A bird in the hand is worth $\log n$ in the bush. *Preprint*, 2016. 7

[6] Douglas B West. *Introduction to Graph Theory*. Prentice Hall, 2 edition, 2000. 3

[7] Yihong Wu and Pengkun Yang. Sample complexity of the distinct element problem. *arxiv preprint arxiv:1612.03375*, Apr 2016. 2