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

Jason M. Klusowski Yihong Wu

Working paper: October 11, 2017

Contents

Abstract

Since the early work of Goodman (1949) [27] and Frank (1978) [23], learning properties of large graphs from samples has emerged as an important problem in statistical network analysis. We revisit a problem formulated by Frank of estimating the numbers of connected components in a graph of N vertices. The estimation procedure is based on the subgraph sampling model, where sample each vertex independently with probability p and observe the subgraph induced these sampled vertices. The key question is whether it is possible to achieve accurate estimation by sampling a vanishing fraction of the vertices. We show that it is possible by accessing only sublinear number of samples if the graph does not contains high-degree vertices or long induced cycles; otherwise it is impossible. We obtain optimal sample complexity bounds for several classes of graphs including forests, cliques, and chordal graphs.

The methodology relies on topological identities of graph homomorphism numbers. They, in turn, also play a key role in proving minimax lower bounds based on construction of random instances of graphs with matching structures of small subgraphs. We perform a numerical study of the estimators on simulated data and discuss extensions to general graphs.

1 Introduction

Counting the number of features in a graph – ranging from basic local structures like motifs or graphlets (e.g., edges, triangles, wedges, stars, cycles, cliques, clustering coefficients), or other more global features like the number of connected components – is an important statistical and computational problem. For instance, applied researchers seek to capture from such features the interactions and relationships between groups and individuals. In doing so, they typically collect data from a random sample of nodes to construct a representation of the true network. This setting is largely due to cost and time constraints (e.g., in-person interviews that are in remote locations) or an inability to gain access the full population (e.g., historical data). Most of the problems encountered in practice are based on the desire to infer global properties of the parent network (population) from the sampled version. For example, [3] studied the social networks of the Hadza hunter-gatherers of Tanzania by surveying 205 individuals in 17 Hadza camps (from a population of 517), each camp consisting of approximately 30 people. Another study [12] of famers in Ghana used network data from a survey of 180 households in three villages (from a population of 550 households). Even lower sampling ratios have been used in other works (such as 30% in [17]), particularly for large scale studies [4, 19]. A good overview of various experiments and their corresponding sampling ratios can be found in [10, Appendix F, p. 11].

These coverage problems and how they are addressed broadly arise from two different perspectives.

- If the full network is either too large to scan or expensive to store, approximation algorithms can overcome such computational or storage issues that would otherwise be unwieldy. For example, if the network is extremely large, it is generally impossible to directly enumerate the whole population. Rather than reading the entire graph, these algorithms randomly (or deterministically) query parts of the full graph or explore the graph through a random process that evolves over time [5]. Some popular instances of traversal based procedures are snowball sampling [28] and respondent-driven sampling [41].
- If the full network is unknown due to lack of data (which could arise from the underlying experimental design and data collection procedure), one must construct statistical estimators (i.e., functions of the sampled graph) to conduct inference. These estimators must be designed to account for the fact that the sampled network is only a representation of the true network, and thus subject to certain inherent biases and variability.

Because of these aforementioned issues, it is desirable to design sublinear time (in the size of the graph) algorithms or estimators that are obtained from sublinear sampling complexities. Various algorithms based on edge and degree queries have been proposed to estimate average degree [18, 25], triangle counts [14], and other more general subgraph counts [26, 30] in sublinear time.

The problem of estimating the number of connected components in a graph has been considered in various settings and for various sampling models. Some real world examples include sampling of plants or animals from different species [42] and sampling of words from a text and estimating vocabulary [36]. Here it is desired to estimate the number of classes in the population from an observed sample. The objects in each class may be weakly or strongly related to each other, corresponding to a tree or clique, respectively. Even within each class, relationships may vary is strength, and thus a flexible model is required to adapt to such cases.

In [11], it was shown that the runtime to estimate the number of connected components in a general graph (motivated by calculating the weight of the minimum spanning tree) within an additive error of $N\epsilon$ is $d_{\text{avg}}\epsilon^{-2}\log\frac{d_{\text{avg}}}{\epsilon}$ for graphs on N vertices with average degree d_{avg}. Their method relies on data obtained from a random sample of vertices and then performing a breadth first search on each vertex which ends according to a random stopping criterion. The algorithm requires knowledge of the average degree d_{avg} and must therefore be known or estimated apriori. Subsequently, [6] noted that by modifying the stopping criterion, one could improve the runtime to $\epsilon^{-2}\log\frac{1}{\epsilon}$. Neither work shows the optimality of the algorithms

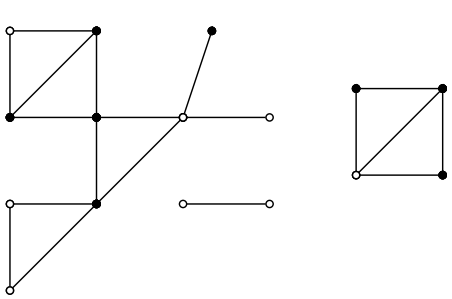
among the class of simple graphs (i.e., graphs without multiple edges or self-loops). In these algorithms, the breadth first search may visit many of the edges and explores larger fraction of the graph at each round. In real-world settings, such a sampling scheme may be impractical due to limitations inherent in the experimental design. Indeed, in applied settings, typically only a random sampled of nodes is taken and the connections between them observed. One of the earliest works in this direction is Frank [23], which is the basis for the theory developed in this paper. Drawing from previous experience on random vertex sampling to estimate certain graph quantities [20–22,24], he derives unbiased estimators and corresponding performance guarantees (i.e., variance bounds) for forests and unions of cliques. Extensions to more general graphs are briefly discussed, although no consistent estimators are proposed. While these results are interesting, questions of their generality and optimality remain open. We consider these issues in the sequel.

The paper is organized as follows. In Section 2, we formally describe the problem and sampling model. We motivate our focus on chordal graphs by showing that in the absence of such structural assumptions, the estimation problem becomes difficult. In Section 3, we define chordal graphs and their associated quantities and discuss the estimator of cc and its statistical properties. In Section 4, we develop a general minimax lower bound strategy for graph counts and use it to show that our estimator of cc from Section 3 is minimax rate optimal. Finally, in Section 5, we perform a numerical study of the estimators on simulated data for chordal and non-chordal graphs.

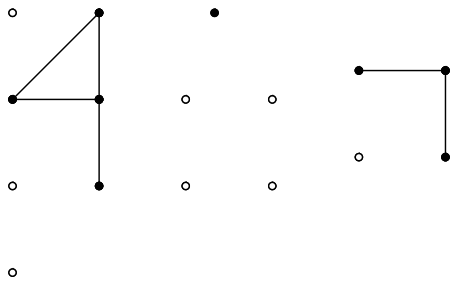
For subsequent sections, we establish some notation that will be used throughout the paper. Let $G = (\mathsf{V}, \mathsf{E})$ be a simple, undirected graph. Let $\mathsf{e} = \mathsf{e}(G)$ denote the number of edges, $\mathsf{v} = \mathsf{v}(G) = |\mathsf{V}(\mathsf{G})|$ denote the number of vertices, and $\mathsf{cc} = \mathsf{cc}(G)$ be the number of (connected) components in G. The neighborhood of a vertex u is denoted by $N_G[\mathsf{u}] = \{\mathsf{v} \in \mathsf{V}(\mathsf{G}) : (\mathsf{u}, \mathsf{v}) \in \mathsf{E}(\mathsf{G})\}$. Let $\mathsf{ind}(H, G)$ be the number of vertex induced subgraphs of G that are isomorphic to H. For graphs G and G', we use the the notation $G \cup G'$ to denote the graph union $(\mathsf{V}(G) \cup \mathsf{V}(G'), \mathsf{E}(G) \cup \mathsf{E}(G'))$. For brevity, we will write kG denote the graph union of k copies of G. We will use the notation K_θ, P_θ, and C_θ to denote the complete graph, path graph, and cycle graph on θ vertices, respectively. Let $K_{\omega,\omega'}$ denote the complete bipartite graph with $\theta\theta'$ edges and $\theta + \theta'$ vertices. Let S_θ denote the star graph $K_{1,\theta}$ on $\theta + 1$ vertices.

2 Model

The sampling model determines how reflective the data is of the population graph and therefore the quality of the estimation procedure. There are many ways to sample from a graph (see [13] for a list of techniques and [32–34] for a good overview), but here we will focus on two nearly equivalent sampling schemes [20–22, 24]. Fix a simple graph $G = (\mathsf{V}, \mathsf{E})$ on N vertices. For $\mathsf{S} \subset \mathsf{V}$, we denote by $G[\mathsf{S}]$ the vertex induced subgraph. If S represents a collection of vertices that are randomly sampled according to a sampling mechanism, we denote $G[\mathsf{S}]$ by $\widetilde{G}$. In this paper, we study the uniform sampling model: when S is generated from the outcome of N Bernoulli trials with success probability p, corresponding to deciding whether a vertex v belongs to the sample from the outcome of a coin $b_v \sim \mathrm{Bern}(p)$, where p is the *sampling rate* or *sampling probability*. Thus, $\mathsf{S} = \{v : b_v = 1\}$ and $|\mathsf{S}| \sim \mathrm{Bin}(N, p)$. A nearly equivalent counterpart of this model is when $n = pN$ members are chosen uniformly at random without replacement from V. Because of their likeness to each other, we will mainly focus on the Bernoulli sampling model.



(a) Sampled vertices S of G.



(b) Subgraph induced by sampled vertices $\widetilde{G} = G[\mathsf{S}]$.

Figure 1: We take a random sample of vertices, shown in black, from the graph on the left. The graph on the right is the subgraph induced by these vertices. Non-sampled vertices are observed as isolated vertices.

As we will see, it is natural to impose conditions on certain graph parameters. In our case, we consider two: the maximum degree at most d and clique number (size of the largest vertex induced clique) ω. The graph classes we consider involve boundedness assumptions on d or ω.

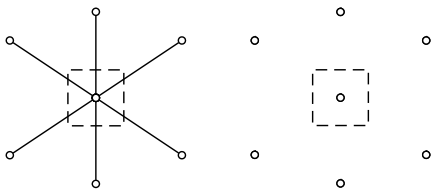
There are two main obstacles in estimating the number of connected components in graphs – large degrees and large induced cycles. If either is allowed to be arbitrarily large, we will show that root mean squared error is nearly a constant multiple of N (i.e., we cannot estimate cc with a sublinear number of samples).

- Accurate estimation is impossible if the parent graph is acyclic and has vertices with large degree.
- Accurate estimation is impossible if the parent graph contains large induced cycles and has maximum degree at most $d = 2$ (viz., graphs that are unions of cycle and path graphs).

Theorem 1. *Let $\mathcal{G}(N)$ denote the collection of all forests on N vertices. There exist universal constants $c > 0$ and $C > 0$ such that*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N)} \mathbb{P}_G(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| > cN) \geq C.$$

This result can be shown as follows. Let G denote the star graph on N vertices. Let G' denote the graph consisting of N isolated vertices. Note that as long as the center vertex in G is not sampled (the vertex enclosed in the dotted region in Fig. 2), the sampling distributions of G and G' are identical. More precisely, $d_{\mathrm{TV}}(P_{\widetilde{G}}, P_{\widetilde{G'}}) \leq p(1-q^{N-1}) < 1$. The gap between the number of connected components in G and G' is $N-1$. By LeCam's two point method for lower bounding the minimax probability [43, Theorem 2.2(i)], we have that $\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N)} \mathbb{P}_G(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| > cN) \gtrsim 1 - p(1-q^{N-1})$.



(a) The graph of G for $N = 6$

(b) The graph of G' for $N = 6$

Figure 2: The two graphs are isomorphic if the center vertex (enclosed in the dotted region) is not sampled and all incident edges are removed.

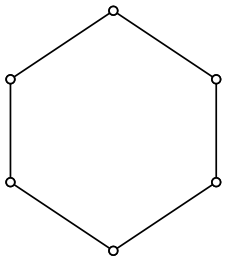
Theorem 2. *Let $\mathcal{G}(N,r)$ denote the collection of all graphs on N vertices with longest induced cycle less than r and maximum degree at most $d=2$. If $r \asymp \frac{\log N}{\log \frac{1}{p}}$, there exist universal constants $c>0$ and $C>0$ such that*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}(N,r)} \mathbb{P}_G\left(|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| > N\left(\frac{c\log\frac{1}{p}}{\log N}\right)\right) \geq C.$$

Thus if r grows like $(\log N)/(\log\frac{1}{p})$, we see that the minimax risk cannot decay faster than logarithmic in N. It is for this reason that we consider only *small* induced cycles. To give a high level intuition for this statement, let G denote the graph consisting of $N/(2r)$ copies of a cycle of length $2r$. Let G' denote the graph consisting of $N/(2r)$ copies of two cycles of length r. Then the sampling distributions of each of the $N/(2r)$ copies are identical provided at most $r-1$ vertices are sampled (this statement will be made more precise in Section 4). To see why this is plausible, note that each connected subgraph with $k<r$ vertices appears exactly N times in each graph. Using a union bound, we have that

$$\begin{aligned} d_{\mathrm{TV}}(P_{\widetilde{G}}, P_{\widetilde{G'}}) &\leq \mathbb{P}(\text{at least } r \text{ vertices are sampled from at least one of the } N/(2r) \text{ copies}) \\ &\leq \frac{N}{2r}\mathbb{P}(\mathrm{Bin}(2r,p)\geq r) \\ &\leq \frac{N(4p)^r}{2r}. \end{aligned}$$

The gap between the number of connected components in G and G' is $N/(2r)$. By LeCam's two point method for lower bounding the minimax probability [43, Theorem 2.2(i)], we have that $\inf_{\widehat{\mathsf{cc}}}\sup_{G\in\mathcal{G}(N,r)}\mathbb{P}_G\left(|\widehat{\mathsf{cc}}-\mathsf{cc}(G)|>\frac{cN}{r}\right) \gtrsim 1-\frac{N(4p)^r}{2r}$. Choosing $r\asymp\frac{\log N}{\log\frac{1}{p}}$ ensures that $\frac{N(4p)^r}{2r}<1$.



(a) One copy making G for $r=3$

(b) One copy making G' for $r=3$

Figure 3: Each connected subgraph with $k<3$ vertices (viz., isolated vertices and edges) appears exactly 6 times in each graph.

This motivates us to consider classes of graphs defined by their maximal degree and maximal length of induced cycles. Therefore, this paper focuses on the case when maximum induced cycle length is no more than three, exactly corresponding to chordal graphs.

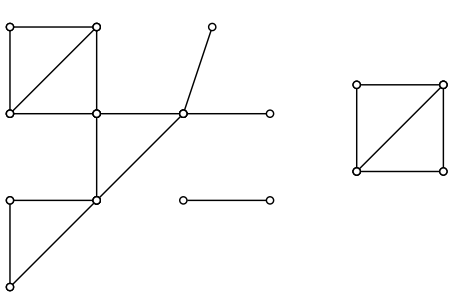
3 Algorithms and upper bound

The main results are summarized below in terms of the 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. For brevity, we write $q=1-p$. Similar results can be obtained for the uniformly sampling model, where S is drawn uniformly at random from all subsets of $\mathsf{V}(\mathsf{G})$ of cardinality n, when we identify $p=n/N$. As p grows from zero to one,

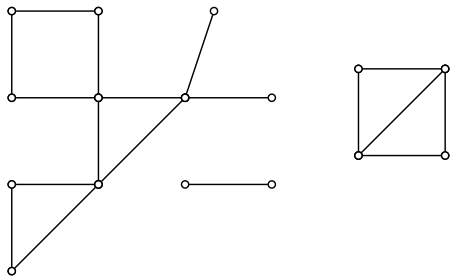
an increasing fraction of the graph is observed and intuitively the estimation problem should become easier. Indeed, all forthcoming minimax risk rates are inversely proportional to p. We are also interested in suitable ranges for p and part of the analysis will give explicit conditions on p for consistency.

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 [23]. There are a few extensions of his work that we provide. First, none of the analysis involves optimality and, second, we provide estimators that adapt to larger collections of graphs (for which forests and unions of cliques are special cases).

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* $k \geq 4$.



(a) An example of a chordal graph with $\mathsf{cc} = 3$



(b) An example of a non-chordal graph with $\mathsf{cc} = 3$

Figure 4: The graph on the left is chordal. The graph on the right is not chordal because it contains an induced cycle of length four.

Theorem 3 (Chordal graphs)**.** *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with clique number $\omega \geq 2$ and maximum degree at most d. Then*

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

Conversely,

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega-1}} \right) \wedge N^2 \right),$$

and if d is a constant, then

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

Theorem 4 (Forests)**.** *Let $\mathcal{F}(N, d) = \mathcal{G}(N, d, 2)$ denote the collection of all forests on N vertices with maximum degree at most d. Then*

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

Theorem 5 (Cliques)**.** *Let $\mathcal{C}(N)$ denote the collection of all graphs on N vertices consisting of disjoint unions of cliques. Let $N > 4$. Then*

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

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 [46], which shows the optimality of Theorem 5.

3.1 Chordal graphs

Definition 2. *A perfect elimination ordering (PEO) of a graph G of size N is an ordering of the vertices $\sigma : [N] \mapsto [N]$ such that, for each i, $N_G[v_{\sigma(i)}] \cap \{v_{\sigma(1)}, \ldots, v_{\sigma(i)}\}$ is a clique.*

A classical result of Dirac states that the existence of a PEO is in fact the defining property of chordal graphs (cf. e.g., [44, Theorem 5.3.17]).

Theorem 6. *A graph is chordal if and only if it admits a PEO.*

Theorem 7. *Let σ and σ' be any two PEOs of a chordal graph G. Let c_i and c'_i denote the cardinalities of $N_G[v_{\sigma(i)}] \cap \{v_{\sigma(1)}, \ldots, v_{\sigma(i)}\}$ and $N_G[v_{\sigma'(i)}] \cap \{v_{\sigma'(1)}, \ldots, v_{\sigma'(i)}\}$, respectively. Then there exists a permutation of the vertices τ such that $\mathsf{c}_{\tau(i)} = \mathsf{c}'_i$ for all i.*

Proof. By [44, Theorem 5.3.26], the chromatic polynomial of G is

$$\chi(G; x) = (x - \mathsf{c}_1)\cdots(x - \mathsf{c}_N) = (x - \mathsf{c}'_1)\cdots(x - \mathsf{c}'_N).$$

Since these factorizations are unique, it follows that there exists a permutation of the vertices τ such that $\mathsf{c}_{\tau(i)} = \mathsf{c}'_i$ for all i. □

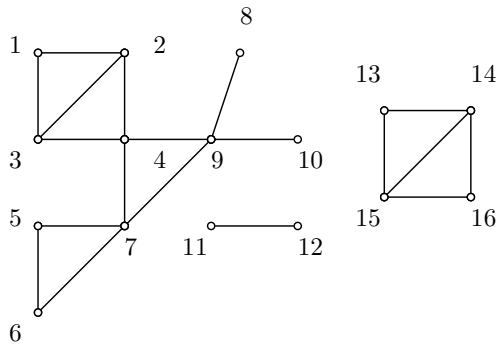


Figure 5: A chordal graph G with PEO $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16)$. In this example, $\mathsf{cc} = 3 = 16 - 19 + 6 = \mathsf{ind}(K_1, G) - \mathsf{ind}(K_2, G) + \mathsf{ind}(K_3, G)$.

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., [44, Exercise 5.3.22, p. 231]). The fact that one can infer global properties of a graph (such as number of connected components or communities) from local features has also be used in generative graph models [2, 7]. 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 6, it has a PEO $(v_1, \ldots, v_N)$. Denote

$$\mathsf{C}_j = N[v_j] \cap \{v_1, \ldots, v_j\}, \quad \mathsf{c}_j = |\mathsf{C}_j|. \tag{3}$$

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

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

Moreover, note that

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

Hence, it follows that

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

f and

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

□

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

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

Using elementary enumerative combinatorics, in particular, the vertex elimination structure of chordal graphs, we prove Theorem 8, which bounds its performance.

A few comments about (4) are in order. First, it is completely adaptive to ω, d and N. The sum in (4) terminates at the clique number of the subsampled graph and therefore one does not need to know ω. Second, it has low computational complexity. An algorithm in the celebrated work of [40] finds a PEO in $O(N+\mathsf{e})$ time. Because the subsampled graph $\widetilde{G}$ is also chordal, it too has a PEO. Thus, to calculate $\mathsf{ind}(K_i, \widetilde{G})$, we can first find a PEO $\sigma : [N] \mapsto [N]$ of $\widetilde{G}$ and then use the formula $\mathsf{ind}(K_i, \widetilde{G}) = \sum_{j=1}^{N} \binom{\widetilde{\mathsf{c}}_j}{i-1}$, where $\widetilde{\mathsf{c}}_j$ is the cardinality of $N[v_j] \cap \{v_1, \ldots, v_j\}$.

A general graph can always be modified to become chordal by adding edges. Such an operation is called a *chordal completion* or *triangulation* of a graph, henceforth denoted by TRI. There are many ways to triangulate a graph and this is typically done with the goal of minimizing some objective function (e.g., number of edges or the clique number). Efficient triangulations do not affect the number of connected components, since the operation can be applied to each component. Although we have not developed any theory for non-chordal graphs, it is almost certain that graphs encountered in practice contain induced cycles of length greater than three. One possible way to generalize our method is to first triangulate the subsampled graph and then apply the $\widehat{\mathsf{cc}}$ estimator. For this to work, we would require triangulation to commute with subgraph sampling in the sense that the law of $(\mathsf{TRI}(G))[\mathsf{S}]$ is the same as $\mathsf{TRI}(G[\mathsf{S}])$ (i.e., the triangulated sampled graph is the sampled graph from a parent graph triangulation); unfortunately, this does not hold in general. Thus, our theory does not readily extend to non-chordal graphs. Nevertheless, we modify the original estimator by first triangulating the subsampled graph $G[\mathsf{S}] \mapsto \mathsf{TRI}(G[\mathsf{S}])$ and then applying $\widehat{\mathsf{cc}}$ to this transformed data via $\widehat{\mathsf{cc}} = \widehat{\mathsf{cc}}(\mathsf{TRI}(G[\mathsf{S}]))$. This more robust estimator seems to be competitive with $\widehat{\mathsf{cc}}$ in both performance (see Fig. 11) and computational efficiency. Indeed, there are polynomial time algorithms that add at most $8k^2$ edges if at least k edges must be added to make the graph chordal [37]. It should be noted that blindly applying the $\widehat{\mathsf{cc}}$ estimator to the subsampled graph without triangulation leads to nonsensical outcomes. Thus, preprocessing the graph appears to be necessary for producing good outcomes. We will leave the task of rigorously establishing these heuristics for future consideration.

Theorem 8. *Let G be a chordal graph on N vertices. Suppose $S \subset \mathsf{V}(\mathsf{G})$ is generated by Bernoulli(p) sampling and let $\widetilde{G} = G[S]$. Then $\widehat{\mathsf{cc}}$ defined in (4) is an unbiased estimator of $\mathsf{cc}(G)$. Furthermore,*

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

and for all $t \geq 0$,

$$\mathbb{P}[|\widehat{\mathsf{cc}} - \mathsf{cc}(G)| \geq t] \leq 2\exp\left\{-\frac{8p^\omega t^2}{25(d\omega+1)(N+t/3)}\right\}, \tag{6}$$

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

Lemma 2. *Let*

$$f(k) = \left(-\frac{q}{p}\right)^k. \tag{7}$$

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

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

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

$$\mathsf{Var}\left[\sum_{\mathsf{S}\in\mathcal{S}}\alpha_{\mathsf{S}}f(N_{\mathsf{S}})\right]=\sum_{\mathsf{S}\in\mathcal{S}}\alpha_{\mathsf{S}}^2\left(\frac{q}{p}\right)^{|\mathsf{S}|}.$$

Remark 1. *The function* $f(N_{\mathsf{S}})=(-\frac{q}{p})^{N_{\mathsf{S}}}$ *is exactly the (unnormalized) orthogonal basis for the binomial measure that is used in Boolean function analysis [38, Definition 8.40].*

Proof of Theorem 8. For a chordal graph G on N vertices, let $(v_1,\dots,v_N)$ be a PEO of G. Recall from (3) that c_j denotes the number of neighbors of v_j among $v_1,\dots,v_j$. That is, $\mathsf{c}_j=\sum_{k=1}^{j-1}\mathbb{I}\{v_k\sim v_j\}$. We also let $\widetilde{\mathsf{c}}_j$ denote the empirical version; that is

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

where $b_k=\mathbb{I}\{v_k\in S\}$. Note that $\mathsf{c}_j|\{b_j=1\}\sim\mathrm{Bin}(\mathsf{c}_j,p)$ and

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

and hence

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

where the function f is defined in (7). To show the variance bound, we note that

$$\mathsf{Var}(\widehat{\mathsf{cc}})=\frac{1}{p^2}\sum_{j=1}^{N}\mathsf{Var}[b_jf(\widetilde{\mathsf{c}}_j)]+\frac{2}{p^2}\sum_{j<i}\mathrm{Cov}[b_jf(\widetilde{\mathsf{c}}_j),b_if(\widetilde{c}_i)].\tag{8}$$

It is easy to verify that

$$\mathsf{Var}[b_jf(\widetilde{\mathsf{c}}_j)]=\begin{cases}p\left(\frac{q}{p}\right)^{\mathsf{c}_j} & \text{if } \mathsf{c}_j>0\\ pq & \text{if } \mathsf{c}_j=0\end{cases}.\tag{9}$$

Since $\mathsf{c}_j\le\omega-1$, it follows that $\mathsf{Var}[b_jf(\widetilde{\mathsf{c}}_j)]\le p\left(\frac{q}{p}\right)^{\omega-1}$. Moreover, let $\mathsf{C}_j=N[v_j]\cap\{v_1,\dots,v_j\}$. Then by Lemma 2,

$$\mathrm{Cov}[b_jf(\widetilde{\mathsf{c}}_j),b_if(\widetilde{c}_i)]=\begin{cases}p^2\left(\frac{q}{p}\right)^{\mathsf{c}_j} & \text{if } \mathsf{C}_j=\mathsf{C}_i\neq\emptyset\\ -pq\left(\frac{q}{p}\right)^{\mathsf{c}_j} & \text{if } \mathsf{C}_j=\mathsf{C}_i\setminus\{v_j\}\neq\emptyset \text{ and } v_j\sim v_i\\ 0 & \text{otherwise}\end{cases}.$$

Thus,

$$\sum_{j<i}\mathrm{Cov}[b_jf(\widetilde{\mathsf{c}}_j),b_if(\widetilde{c}_i)]\le\sum_{j<i:\mathsf{C}_j=\mathsf{C}_i\neq\emptyset}p^2\left(\frac{q}{p}\right)^{\mathsf{c}_j}\le Ndp^2\left(\frac{q}{p}\right)^{\omega-1},\tag{10}$$

where the last step follows from the fact that $\#\{(i,j) : i > j,\ \mathsf{C}_j = \mathsf{C}_i \neq \emptyset\} \leq Nd$. Finally, combining (8), (9) and (10) yields

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

and hence the desired (5).

The concentration inequality (6) for $\widehat{\mathsf{cc}}$ follows from Theorem 2.3 in [29]. In their notation, $\mathcal{A} = \mathsf{V}(\mathsf{G})$ and $\widehat{\mathsf{cc}} = \sum_{\alpha\in\mathcal{A}} Y_\alpha$, where $Y_\alpha = \frac{1}{p} b_\alpha f(\widetilde{\omega_\alpha})$. The dependency graph Γ for $\{Y_\alpha\}_{\alpha\in\mathcal{A}}$ has maximum degree at most bounded by $d\omega$ (by Lemma 3) and hence $\chi^\star(\mathcal{A}) \leq d\omega + 1$. Also, $b = (\frac{1}{p})^\omega$ and $S = \sum_{\alpha\in\mathcal{A}} \mathsf{Var}[Y_\alpha] \leq N(\frac{1}{p})^\omega$. □

Remark 2. *According to [9], a graph parameter f is "testable" if there exists an estimator $\widehat{f}$ such that for each $\epsilon > 0$ and N sufficiently large, $\mathbb{P}(|f(G) - \widehat{f}(\widetilde{G})| > \epsilon) < \epsilon$. It is shown that this property holds if and only if $f(G_k)$ converges to $f(G)$ for every sequence of graphs $G_1, G_2, \ldots$ that converge to G in the cut-set norm. It follows from (6) that $\mathsf{cc}(G)/N$ is testable.*

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

$$\mathbb{E}[f(N_\mathsf{S})f(N_\mathsf{T})] = \mathbb{E}[(-\frac{q}{p})^{N_\mathsf{S}+N_\mathsf{T}}] = \mathbb{E}[(-\frac{q}{p})^{N_{\mathsf{S}\setminus\mathsf{T}}+N_{\mathsf{S}\cap\mathsf{T}}}] = \mathbb{E}[(-\frac{q}{p})^{N_{\mathsf{S}\setminus\mathsf{T}}}]\mathbb{E}[(-\frac{q}{p})^{N_{\mathsf{S}\cap\mathsf{T}}}].$$

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

$$\mathbb{E}[f(N_\mathsf{S})f(N_\mathsf{T})] = \mathbb{E}[(-\frac{q}{p})^{N_\mathsf{S}+N_\mathsf{T}}] = \mathbb{E}[(-\frac{q}{p})^{2N_\mathsf{S}}] = \left(\frac{q}{p}\right)^{2|\mathsf{S}|}.$$

□

For the closely related model where n vertices of G are selected uniformly at random, an analogous unbiased estimator of $\mathsf{cc}(G)$ is

$$\widehat{\mathsf{cc}} = \sum_{i\geq 1} \frac{\binom{N}{n}}{\binom{N-i}{n-i}} (-1)^{i+1}\, \mathsf{ind}(K_i, \widetilde{G}), \tag{11}$$

which can similarly be written as

$$\widehat{\mathsf{cc}} = \sum_{j\geq 1} b_j \sum_{i\geq 1} \frac{\binom{N}{n}\binom{\widetilde{\mathsf{c}}_j}{i-1}}{\binom{N-i}{n-i}} (-1)^{i+1}.$$

It does not appear that there is a closed form expression for each $\sum_{i\geq 1} \frac{\binom{N}{n}\binom{\widetilde{\mathsf{c}}_j}{i-1}}{\binom{N-i}{n-i}} (-1)^{i+1}$, nor is it clear how the cross covariance terms

$$\mathsf{Cov}\left[\sum_{i\geq 1} \frac{\binom{N}{n}\binom{\widetilde{\mathsf{c}}_j}{i-1}}{\binom{N-i}{n-i}} (-1)^{i+1}, \sum_{i\geq 1} \frac{\binom{N}{n}\binom{\widetilde{\mathsf{c}}_{j'}}{i-1}}{\binom{N-i}{n-i}} (-1)^{i+1}\right]$$

behave for $j \neq j'$. Part of this difficulty lies in the fact that $\widehat{\mathsf{c}}_j | \{b_j = 1\}$ follows a hypergeometric distribution with mass function

$$\mathbb{P}(\widehat{\mathsf{c}}_j = k | b_j = 1) = \frac{\binom{\mathsf{c}_j}{k}\binom{N-1-\mathsf{c}_j}{n-1-k}}{\binom{N-1}{n-1}},\ k = 0, 1, \ldots, \mathsf{c}_j.$$

Thus, it is not certain how the variances of (11) and (4) compare with each other.

3.1.1 Forests

Since forests are chordal graphs with clique number $\omega = 2$, we immediately get the following result from Theorem 8.

Theorem 9. *Let G be a forest on N vertices with maximum degree at most d. Define*

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

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

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

3.1.2 Cliques

If the parent graph G consists disjoint union of cliques, so does the sampled graph $\widetilde{G}$. Then the estimator (12) can be rewritten as

$$\widehat{\mathsf{cc}} = \sum_{r=1}^{N} \left(1 - \left(-\frac{q}{p}\right)^r\right) \widehat{\mathsf{cc}}_r = \mathsf{cc}(\widetilde{G}) - \sum_{r=1}^{N} \left(-\frac{q}{p}\right)^r \widehat{\mathsf{cc}}_r, \tag{12}$$

where $\widehat{\mathsf{cc}}_r$ is the number of components in the sampled graph $\widetilde{G}$ that have r vertices. This coincides with the unbiased estimator proposed by Frank [23] for cliques, which is, in turn, based on the estimator of Goodman [27].

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

Theorem 10. *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*

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

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

Proof. Observe that $\widehat{\mathsf{cc}} = \sum_{k=1}^{\mathsf{cc}(G)} [1 - (-\frac{q}{p})^{\widetilde{N}_k}]$ and $\{\widetilde{N}_k\}$ is an independent collection with $\widetilde{N}_k \sim \mathrm{Bin}(N_k, p)$. Thus,

$$\mathsf{Var}(\widehat{\mathsf{cc}}) = \sum_{k=1}^{cc(G)} \left(\frac{q}{p}\right)^{N_k} = \sum_{r=1}^{N} \left(\frac{q}{p}\right)^r \mathsf{cc}_r.$$

The upper bound follows from the fact that $\mathsf{cc}_r = 0$ for all $r > \omega$ and $\sum_{r=1}^{N} \mathsf{cc}_r = \mathsf{cc}(G) \le N$. □

3.2 Smoothing

3.2.1 Cliques

When the sampling ratio p is small, the coefficients of the unbiased estimator (12) grow exponentially and result in large variance. Using a technique known as *smoothing* introduced in [39], 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 (12) 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 (12) at the random location L and average over its distribution:

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

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

$$\widehat{\mathsf{cc}}_L = \mathsf{cc}(\widetilde{G}) - \sum_{r=1}^{N}\left(-\frac{q}{p}\right)^r \mathbb{P}(L \geq r)\widehat{\mathsf{cc}}_r. \tag{13}$$

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 11. *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*

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

Proof. The bias of this estimator is seen to be

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

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

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

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

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

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

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

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

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

Also,

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

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 $p\leq 1/2$,

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

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

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

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

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

The solution to λ produces the desired bound. □

3.2.2 Chordal graphs

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

$$\#\{(i,j)\in[N]\times[N] : i>j,\ \mathsf{C}_i\cap\mathsf{C}_j\neq\emptyset\}\leq Nd\omega.$$

Proof of Lemma 3. Let $\mathsf{c}_j = |\mathsf{C}_j|$. We will prove that for a fixed j,

$$\#\{i \in [N] : i > j,\ \mathsf{C}_i \cap \mathsf{C}_j \neq \emptyset\} \leq (d+1-\mathsf{c}_j)\mathsf{c}_j \leq d\omega. \tag{14}$$

We will prove this by contradiction. Suppose

$$\#\{i \in [N] : i > j,\ \mathsf{C}_i \cap \mathsf{C}_j \neq \emptyset\} \geq (d+1-\mathsf{c}_j)\mathsf{c}_j + 1$$

Then at least $(d+1-\mathsf{c}_j)\mathsf{c}_j + 1$ of the C_i have nonempty intersection with C_j. Note that $\deg(v) \geq \mathsf{c}_j - 1$ for each $v \in \mathsf{C}_j$ by definition (since the vertices in C_j form a clique of size c_j in G). By the pigeonhole principle, there is at least one vertex u in C_j such that $\deg(u) \geq (\mathsf{c}_j - 1) + (d - \mathsf{c}_j + 2) = d+1$. This is a contradiction to the bounded degree assumption. Thus, (14) holds and the conclusion follows from noting that $1 \leq \mathsf{c}_j \leq \omega$. □

Theorem 12. *Let $L \sim \text{Poisson}(\lambda)$ with $\lambda \asymp \frac{p}{2q-p}\log\left(\frac{N}{d\omega}\right)$. Let $\widehat{\sigma} : [N] \mapsto [N]$ be a PEO of $\widetilde{G}$ and let $\widehat{\mathsf{c}}_j$ be the cardinality of $N_{\widetilde{G}}[v_{\widehat{\sigma}}] \cap \{v_{\widehat{\sigma}(1)}, \ldots, v_{\widehat{\sigma}(i)}\}$. Define*

$$\widehat{\mathsf{cc}}_L = \frac{1}{p}\sum_{j\geq 1}\left(-\frac{q}{p}\right)^{\widehat{\mathsf{c}}_j}\mathbb{P}(L \geq \widehat{\mathsf{c}}_j).$$

Then

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

Remark 3. *Note that even if $\mathsf{d}(G)\omega(G) = o(N)$, the sampling complexity is sublinear.*

Proof. Let σ be a PEO of the parent graph G and let $\widetilde{\mathsf{c}}_j$ and c_j be the cardinalities of $N_{\widetilde{G}}[v_{\sigma(i)}] \cap \{v_{\sigma(1)}, \ldots, v_{\sigma(i)}\}$ and $N_G[v_{\sigma(i)}] \cap \{v_{\sigma(1)}, \ldots, v_{\sigma(i)}\}$, respectively. Note that σ is also a PEO of $\widetilde{G}$ and hence by Theorem 7, there exists a permutation of the vertices τ such that $\widetilde{\mathsf{c}}_{\tau(j)} = \widehat{\mathsf{c}}_j$ for all j. This means that we can rewrite $\widehat{\mathsf{cc}}_L$ as

$$\begin{aligned}
\widehat{\mathsf{cc}}_L &= \frac{1}{p}\sum_{j\geq 1}\left(-\frac{q}{p}\right)^{\widehat{\mathsf{c}}_j}\mathbb{P}(L \geq \widehat{\mathsf{c}}_j) \\
&= \frac{1}{p}\sum_{j\geq 1} b_{\tau(j)}\left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_{\tau(j)}}\mathbb{P}(L \geq \widetilde{\mathsf{c}}_{\tau(j)}) \\
&= \frac{1}{p}\sum_{j\geq 1} b_j\left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_j}\mathbb{P}(L \geq \widetilde{\mathsf{c}}_j),
\end{aligned}$$

where $\widetilde{\mathsf{c}}_j \mid \{b_j = 1\} \sim \text{Bin}(\mathsf{c}_j, p)$.

We compute the bias and variance of $\widehat{\mathsf{cc}}_L$ and then choose the λ that balances them in an optimal way. First, begin by noting that

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

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 $|\mathbb{E}[\widehat{\mathsf{cc}}_L - \widehat{\mathsf{cc}}]| \le Ne^{-\lambda/2}$. Writing $\widehat{\mathsf{cc}}_L = \frac{1}{p}\sum_{j=1}^{N} W_j$, where $W_j = b_j(-\frac{q}{p})^{\widetilde{\mathsf{c}}_j}\mathbb{P}(L \ge \widetilde{\mathsf{c}}_j)$, and using Lemma 3, we have that

$$\mathsf{Var}[\widehat{\mathsf{cc}}_L] \le N(p + 2p^2\mathsf{d}(G)\omega(G)) \sup_{1\le j\le N} \mathsf{Var}[W_j | b_j = 1].$$

Finally, note that

$$\mathsf{Var}[W_j|b_j=1] \le \mathbb{E}[W_j^2|b_j=1] \le [\|W_j|b_j=1\|_\infty]^2 \le \left[\mathbb{E}_L\left(\frac{q}{p}\right)^L\right]^2 = \exp\left\{2\lambda\left(\frac{q}{p}-1\right)\right\}.$$

Thus, $\mathbb{E}_G|\widehat{\mathsf{cc}}_L - \mathsf{cc}(G)|^2$ is bounded by a multiple of $Ne^{-\lambda} + \frac{\mathsf{d}(G)\omega(G)e^{2\lambda\left(\frac{q}{p}-1\right)}}{p}$. The choice of λ yields the desired bound. □

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

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

where $\widetilde{N}_k$ is the observed number of vertices in the k^{th} component of G. Hence

$$\mathsf{Var}(\widehat{\mathsf{cc}}_L) \le \frac{1}{p}\sum_{k=1}^{\mathsf{cc}(G)}\left(\sum_{j=0}^{N_k-1}\left(\frac{q}{p}\right)^j\mathbb{P}(L \ge j)\right)^2,$$

where N_k is the number of vertices in the k^{th} component of G. Next, observe that

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

and so the variance is bounded by a constant multiple of $\frac{Ne^{2\lambda(\frac{q}{p}-1)}}{p}$. This variance bound is very similar to the other smoothed estimator (13). In fact both achieve the rate in Theorem 11.

Remark 5. *An alternative to smoothing with respect to the $\widehat{c}_j$ is to smooth with respect to the size of the observed complete subgraphs. To be more concrete, let $L \sim \text{Poisson}(\lambda)$ and define*

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

Using arguments similar to those in bounding the bias of $\widehat{\mathsf{cc}}_L$, we can also bound the bias $|\mathbb{E}[\widetilde{\mathsf{cc}}_L - \mathsf{cc}(G)]|$ by $Ne^{-\lambda}$. Thus $\widetilde{\mathsf{cc}}_L$ and $\widehat{\mathsf{cc}}_L$ have similar biases. The variance of $\widetilde{\mathsf{cc}}_L$, however, is more difficult to control and seems to be larger in order, viz., $\frac{d\omega e^{\lambda c(\sqrt{\frac{\omega}{p}}\vee\omega)}}{p}$ for some universal constant $c > 0$.

4 Lower bounds

Since $\mathsf{cc}(G)$ is invariant with respect to isomorphisms, it suffices to describe the sampled graph $\widetilde{G}$ up to isomorphisms. In general, there are three key quantities that measure graph homomorphism numbers.

- $\mathsf{hom}(H, G)$ is the number of homomorphisms from H to G

- $\mathsf{inj}(H, G)$ is the number of edge-induced subgraphs of G that are isomorphic to H
- $\mathsf{ind}(H, G)$ is the number of vertex-induced subgraphs of G that are isomorphic to H

We will pay special attention to the number $\mathsf{ind}(H, G)$ because it records the number of ways H can appear in a random sample from G. The list of vertex-induced subgraphs $\{\mathsf{ind}(H, G) : v(H) \le N\}$ determines G up to isomorphism and hence $\{\mathsf{ind}(H, \widetilde{G}) : v(H) \le N\}$ is a sufficient statistic for $\tilde{G}$. The next lemma tell us that $\{\mathsf{ind}(H, \widetilde{G}) : v(H) \le N,\ H \text{ connected}\}$ is also sufficient. This is appealing from an analytical and computational perspective since counting disconnected subgraphs can be challenging.

4.1 General strategy

We begin with a lemma that has a long history in graph reconstruction theory [45], [16], [8], [31], [35]. It essentially says that disconnected subgraphs are functionally related to connected subgraph counts of the same order.

Lemma 4. *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 Theorem [31] 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 $\mathsf{V}(g)$ into $\cup_{h \in \mathcal{H}} \mathsf{V}(h)$ such that $g[\mathsf{V}(h)] \simeq h$.

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 13. *Let $\mathcal{G}$ be a class of graphs on N vertices that is closed under disjoint graph unions. Let f be a function that is linear in the connected components of $G \in \mathcal{G}$. Let H and H' be two graphs in $\mathcal{G}$ with $m \in \mathbb{N}$ vertices. Suppose $m \le N/500$ and $d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}}) \le 1/500$. Then*

$$\inf_{\hat{f}} \sup_{G \in \mathcal{G}} \mathbb{P}(|\hat{f} - f(G)| \ge \Delta) \ge 0.075.$$

where

$$\Delta = \frac{|f(H) - f(H')|}{8} \left(\sqrt{\frac{N}{m d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}})}} \wedge \frac{N}{m} \right).$$

Moreover, if

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

for all connected *graphs* h *with at most* $k \in [m]$ *vertices. Then*

$$d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}}) \leq \sum_{v=k+1}^{m} \binom{m}{v} p^v q^{m-v}.$$

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

Let g_i be the portion of $\widetilde{G}$ that comes from G_i. Since we know the labels $\{b_v\}$, we can decompose g_i via $g_i = h \cup (m - v(h))K_1$, where $b_v = 1$ for all $v \in \mathsf{V}(h)$ and $b_v = 0$ otherwise. In this case,

$$\mathbb{P}(g_i = h \cup (m - v(h))K_1 | G_i \sim H) = \mathsf{ind}(h, H) p^{v(h)} q^{m - v(h)},$$

and

$$\mathbb{P}(g_i = h \cup (m - v(h))K_1 | G_i \sim H') = \mathsf{ind}(h, H') p^{v(h)} q^{m - v(h)}.$$

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 $G \mapsto f(G)$. To apply the method of two fuzzy hypotheses [43, Theorem 2.15(i)], we must ensure that there exists $L > 0$ and $\beta_0, \beta_1 \in (0, 1)$ such that

$$\mathbb{P}_{\alpha_0}(f(G) \leq L) = \mathbb{P}(\mathrm{Bin}(M, \alpha_0) \leq L) \geq 1 - \beta_0,$$

and

$$\mathbb{P}_{\alpha_1}(f(G) \geq L + 2\Delta) = \mathbb{P}(\mathrm{Bin}(M, \alpha_1) \geq L + 2\Delta) \geq 1 - \beta_1.$$

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

$$\mathbb{P}_{\alpha_0}(f(G) \leq L) = \mathbb{P}(\mathrm{Bin}(M, \alpha_0) \leq M\alpha_0(1 + \delta/2)) \geq 1 - e^{-\delta^2 M/12}.$$

and

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

The first observation we make is that $(g_1, g_2, \ldots, g_M)' \overset{iid}{\sim} P_\alpha^{\otimes M}$. By [43, Theorem 2.15(i)],

$$\inf_{\hat{f}} \sup_{G \in \mathcal{G}} \mathbb{P}(|\hat{f} - f(G)| \geq \Delta) \geq \frac{1 - d_{\mathrm{TV}}(P_{\alpha_0}^{\otimes M}, P_{\alpha_1}^{\otimes M}) - \beta_0 - \beta_1}{2}.$$

To ensure that

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

we use the fact that

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

and

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

In summary,

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

Choosing $\delta = \frac{1}{2} \wedge \sqrt{\frac{1}{4M d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}})}}$, $M \ge 500$, and $d_{\mathrm{TV}}(P,P') \le 1/500$ produces

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

By Corollary 1,

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

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

$$\frac{1}{2}\sum_{h: v(h)\le m} |\mathsf{ind}(h,H) - \mathsf{ind}(h,H')| p^{v(h)} q^{m-v(h)}.$$

Since $\mathsf{ind}(h,H) = \mathsf{ind}(h,H')$ for all h with $v(h) \le k$, it follows that $d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}})$ is in fact equal to

$$\frac{1}{2}\sum_{h: k+1\le v(h)\le m} |\mathsf{ind}(h,H) - \mathsf{ind}(h,H')| p^{v(h)} q^{m-v(h)}.$$

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

$$\frac{1}{2}\sum_{h: k+1\le v(h)\le m} |\mathsf{ind}(h,H) + \mathsf{ind}(h,H')| p^{v(h)} q^{m-v(h)} = \sum_{v=k+1}^{m} \binom{m}{v} p^v q^{m-v}.$$

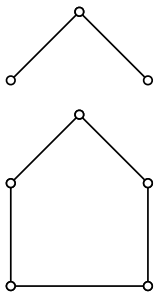
□

4.2 Lower bound for graphs with cycles

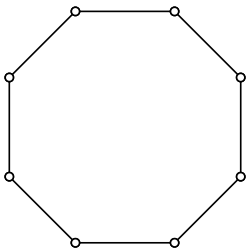
Theorem 14. *Let $\mathcal{G}(N,r)$ denote the collection of all graphs on N vertices with longest induced cycle at most r, $r \ge 4$. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G\in\mathcal{G}(N,r)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 \gtrsim \frac{N}{(4p)^r} \wedge \frac{N^2}{r^2}.$$

Proof. We will prove the lower bound via Theorem 13 with $m = 2(r-1)$. Let $H = C_r \cup P_{r-2}$ and $H' = P_{2(r-1)}$. Note that $\mathsf{ind}(P_i, H) = \mathsf{ind}(P_i, H') = 2r - 1 - i$ 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 $r-1$.



(a) The graph of H for $r = 5$



(b) The graph of H' for $r = 5$

Figure 6: Each connected subgraph with $k < r$ vertices appears exactly $9 - k$ times in each graph.

In the notation of Theorem 13, $k = r-1$, $m = 2(r-1)$, and $|\mathsf{cc}(H) - \mathsf{cc}(H')| = 1$. By Theorem 13, there exists a universal positive constant c such that

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

where

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

$$\begin{aligned} d_{\mathrm{TV}}(P_{\widetilde{H}}, P_{\widetilde{H'}}) &\leq \mathbb{P}(\mathrm{Bin}(2(r-1), p) \geq r) \\ &= \sum_{v=r}^{2(r-1)} \binom{2(r-1)}{v} p^v q^{2(r-1)-v} \\ &\leq (4p)^r. \end{aligned}$$

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

□

4.3 Lower bound for chordal graphs

Theorem 15 (Chordal graphs)**.** *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with clique number $\omega \geq 2$ and maximum degree at most d. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 \gtrsim \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{Np^{\omega-1}} \right) \wedge N^2 \right).$$

Consequently, if d is a constant, then

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 \gtrsim \left(\frac{N}{p^\omega} \wedge N^2 \right).$$

Proof. For every $\omega \geq 3$ and $m \in \mathbb{N}$, we construct a pair of graphs H and H', such that

$$v(H) = v(H') = \omega - 1 + m2^{\omega-2} \tag{15}$$

$$d_{\max}(H) = d_{\max}(H') = m2^{\omega-3} \tag{16}$$

$$\mathsf{cc}(H) = m + 1, \quad \mathsf{cc}(H') = 1 \tag{17}$$

$$|\mathsf{ind}(K_\omega, H) - \mathsf{ind}(K_\omega, H')| = m \tag{18}$$

Let $u = \omega - 1$. Let U be a set of u vertices that form a K_u. We first construct H. For every subset $\mathsf{S} \subset \mathsf{U}$ such that $|\mathsf{S}|$ is even, let V_S be a set of m distinct vertices such that the neighborhood of every $v \in \mathsf{V}_\mathsf{S}$ is given by $\partial v = \mathsf{S}$. Let the vertex set $\mathsf{V}(H)$ be the union of U and all V_S such that $|\mathsf{S}|$ is even. (If $\mathsf{S} = \emptyset$, then $|\mathsf{S}| = 0$ is even and hence H always has exactly m isolated vertices). Repeat the same construction for H' with $|\mathsf{S}|$ being odd. Then both H are H' are chordal and have the same number of vertices:

$$v(H) = \omega - 1 + m \sum_{0 \leq i \leq \omega-1,\ i \text{ even}} \binom{\omega-1}{i} = v(H') = \omega - 1 + m \sum_{0 \leq i \leq \omega-1,\ i \text{ odd}} \binom{\omega-1}{i}$$

which follows from the binomial summation formula. Similarly, (16)–(18) can be verified.

We also have that

$$\mathsf{ind}(K_i, H) = \binom{\omega-1}{i} + m \sum_{0 \leq j \leq \omega-1,\ j \text{ even}} \binom{\omega-1}{j} \binom{j}{i-1} =$$

$$\mathsf{ind}(K_i, H') = \binom{\omega-1}{i} + m \sum_{0 \leq j \leq \omega-1,\ j \text{ odd}} \binom{\omega-1}{j} \binom{j}{i-1} =$$

$$\binom{\omega-1}{i} + m \binom{\omega-1}{i-1} 2^{\omega-1-i},$$

for $i = 1, 2, \ldots, \omega - 1$. This follows from the fact that

$$\sum_{0 \le j \le \omega - 1} (-1)^j \binom{\omega - 1}{j} \binom{j}{i-1} = 0,$$

and

$$\sum_{0 \le j \le \omega - 1} \binom{\omega - 1}{j} \binom{j}{i-1} = \binom{\omega - 1}{i-1} 2^{\omega - i}.$$

To compute the total variation distance between the sampled graphs, the key observation is the following: for both H and H', if a given non-empty subset $\mathsf{U}' \subset \mathsf{U}$ of vertices is not sampled, upon deleting all edges incident to them, the resulting graph is isomorphic to the same graph, whose vertex set is the disjoint union $\mathsf{U}' \cup \bigcup_{S \subset \mathsf{U} \setminus \mathsf{U}'} \mathsf{V}_{\mathsf{S}'}$, where $\mathsf{V}_{\mathsf{S}'}$ consists of m distinct vertices whose neighbourhoods are S. Thus

$$d_{\mathrm{TV}}(\widetilde{H}, \widetilde{H}') \le \mathbb{P}\left[\text{all vertices in } \mathsf{U} \text{ are sampled}\right] \left(1 - (1-p)^{v(H) - |\mathsf{U}|}\right) = p^{\omega - 1}(1 - (1-p)^{m 2^{\omega - 2}})$$

Next assume that

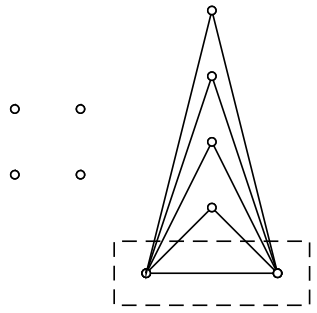
$$d \gtrsim 2^{\omega}.$$

According to (16), we choose $m = \lfloor d 2^{-\omega + 3} \rfloor \ge d 2^{-\omega + 2}$. Then we have

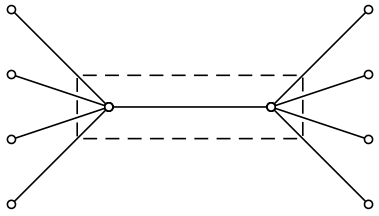
$$d_{\mathrm{TV}}(\widetilde{H}, \widetilde{H}') \le p^{\omega - 1}(1 - (1-p)^{2d}) \asymp p^{\omega - 1}(pd \wedge 1).$$

In view of Theorem 13 and (18), we have

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N, d, \omega)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 = \Theta_\omega \left(\left(\frac{N}{p^\omega} \vee \frac{Nd}{p^{\omega - 1}} \right) \wedge N^2 \right).$$

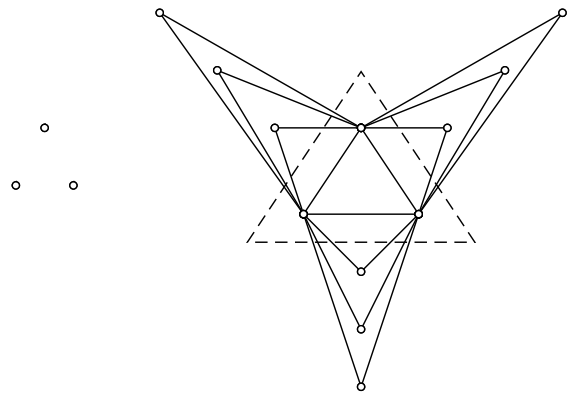


(a) The graph of H for $\omega = 3$ and $m = 4$

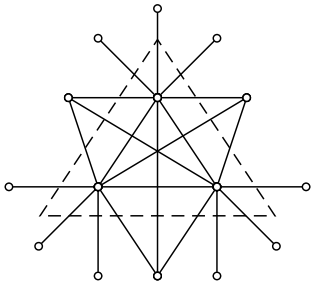


(b) The graph of H' for $\omega = 3$ and $m = 4$

Figure 7: The vertices enclosed in the dotted boundary make up the set U. In this case, $u = 2$, so they form a edge. If any one of these vertices is not sampled and all incident edges are removed, the graphs are isomorphic.



(a) The graph of H for $\omega = 4$ and $m = 3$



(b) The graph of H' for $\omega = 4$ and $m = 3$

Figure 8: The vertices enclosed in the dotted boundary make up the set U. In this case, $u = 3$, so they form a triangle. If any one or two of these vertices are not sampled and all incident edges are removed, the graphs are isomorphic.

□

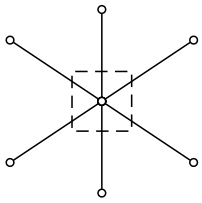
Remark 6. *This means that for consistent estimation for graphs in $\mathcal{G}(N, d, \omega)$, we need p to be greater than both $(\frac{1}{N})^{\frac{1}{\omega}}$ and $(\frac{d}{N})^{\frac{1}{\omega-1}}$.*

4.4 Lower bound for forests

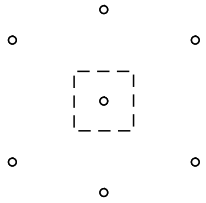
Theorem 16 (Forests)**.** *Let $\mathcal{F}(N, d) = \mathcal{G}(N, d, 2)$ denote the collection of all forests on N vertices with maximum degree at most d. Then*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{F}(N,d)} \mathbb{E}_G |\widehat{\mathsf{cc}} - \mathsf{cc}(G)|^2 \gtrsim \left(\frac{N}{p^2} \vee \frac{Nd}{p} \right) \wedge N^2.$$

Proof. This follows from Theorem 15 with $\omega = 2$.



(a) The graph of H for $\omega = 2$ and $m = 6$



(b) The graph of H' for $\omega = 2$ and $m = 6$

Figure 9: The two graphs are isomorphic if the center vertex (enclosed in the dotted boundary) is not sampled and all incident edges are removed.

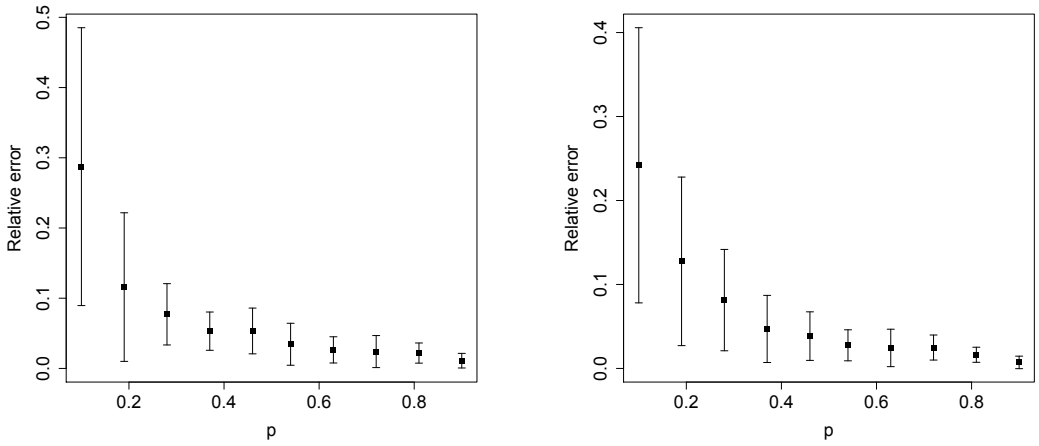
□

Remark 7. *This means that for consistent estimation for forests, we need p to be greater than both $1/\sqrt{N}$ and d/N.*

5 Experiments

In this section, we perform a simulation study of the estimator $\widehat{\mathsf{cc}}$ using synthetic data from various random graphs. The error bars in the following plots show the variability of the relative error $\frac{|\widehat{\mathsf{cc}}-\mathsf{cc}(G)|}{\mathsf{cc}(G)}$ over 20 independent experiments of subgraph sampling on a fixed parent graphs G. The solid black horizontal line shows the sample average and the whiskers show the mean $\pm$ the standard deviation. Note that as p increases, the sampling variability decreases monotonically. The estimates perform poorly for small p (e.g., $p < 0.2$), but becomes increasingly better as p grows. The decay of relative error also appears to be non-linear, changing from very large to moderate improvements as p varies from 0 to 1.

Our parent graphs are generated as follows. We first generate a random Erdös-Renyi graph $G(n, \delta)$ and then triangulate it by calculating the fill-in of edges required to make it chordal. In Fig. 10a, we take G to be a triangulated realization of $G(n, \delta)$ with $n = 2000$ and $\delta = 0.0005$. Note the critical threshold $\delta = \frac{\log n}{n}$ for G to be connected [15]. In other words, to make $\mathsf{cc}(G)$ large, we require that the edge connection probability δ be sufficiently small. In Fig. 10b, we take G to be a triangulated realization of 200 copies of $G(n, \delta)$ with $n = 100$ and $\delta = 0.2$. In accordance with Theorem 8, the better performance in Fig. 10b is due to moderately sized d and ω, and large $\mathsf{cc}(G)$.



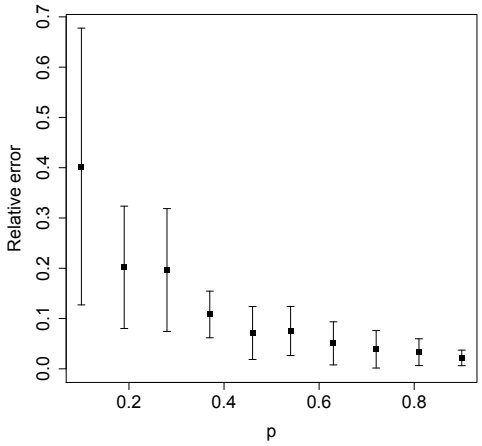
(a) Parent graph equal to a triangulated realization of $G(2000, 0.0005)$ with $d = 36$, $\omega = 5$, and $\mathsf{cc}(G) = 985$.

(b) Parent graph equal to a triangulated realization of 200 copies of $G(100, 0.2)$ with $d = 8$, $\omega = 4$, and $\mathsf{cc}(G) = 803$.

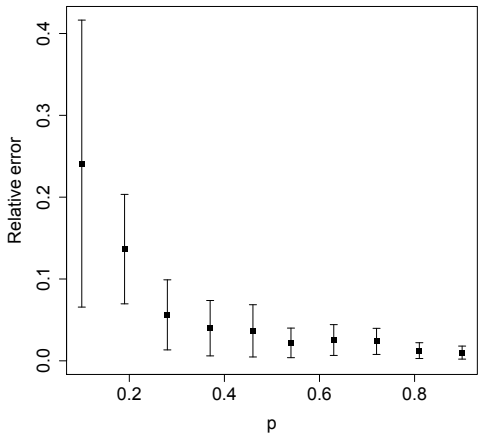
Figure 10

5.1 Non-chordal graphs

We also run the experiment when the parent graph is not chordal. Although we have not developed any theory for this setting, it is almost certain that the types of graphs encountered in practice contain induced cycles of length greater than three (e.g., sparse graphs). As mentioned earlier, one possible method is to first triangulate the subsampled graph and then apply the $\widehat{\mathsf{cc}}$ estimator. We modify the original estimator by first triangulating the subsampled graph $G[\mathsf{S}] \mapsto \mathsf{TRI}(G[\mathsf{S}])$ and then applying $\widehat{\mathsf{cc}}$ to this transformed data via $\widehat{\mathsf{cc}} = \widehat{\mathsf{cc}}(\mathsf{TRI}(G[\mathsf{S}]))$. The plots in Fig. 11 indicate that this is a reasonable strategy; in fact they are competitive with the errors in Fig. 10.



(a) Parent graph equal to a realization of $G(2000, 0.0005)$ with $d = 8$, $\omega = 3$, and $\mathsf{cc}(G) = 756$.



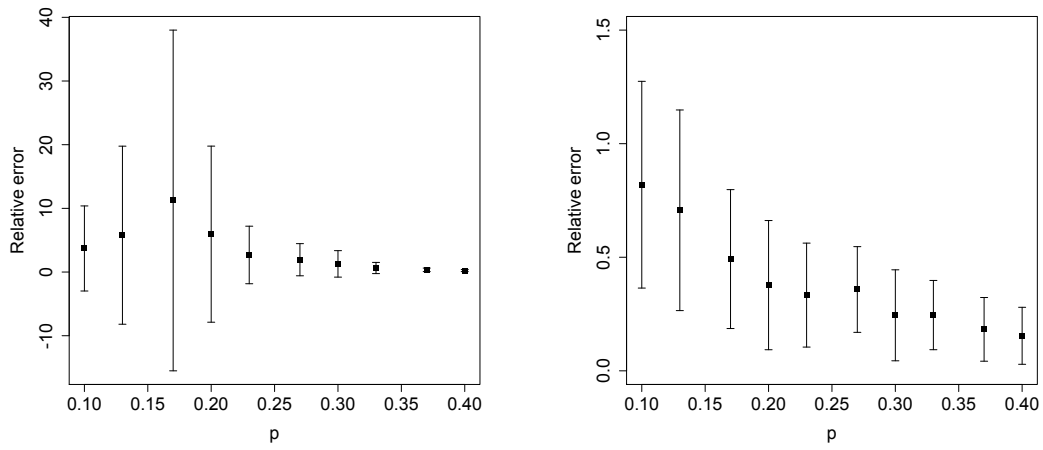
(b) Parent graph equal to a realization of 200 copies of $G(100, 0.2)$ with $d = 7$, $\omega = 4$, and $\mathsf{cc}(G) = 532$.

Figure 11

5.2 Smoothed estimator

We perform a simulation study of $\widehat{\mathsf{cc}}_L$ from Theorem 12. The parent graph is equal to a triangulated realization of $G(1000, 0.0015)$ with $d = 88$, $\omega = 15$, and $\mathsf{cc}(G) = 325$. The plots in Fig. 12b show that the sampling variability is significantly reduced for the smoothed estimator, particularly for small values of p (to show detail, the vertical axes are plotted on different scales). This behavior is in accordance with the upper bounds furnished in Theorem 8 and Theorem 12. Large values of ω inflate the variance of $\widehat{\mathsf{cc}}$ considerably by an exponential factor of $1/p^{\omega}$, whereas the effect of large ω on the variance of $\widehat{\mathsf{cc}}_L$ is polynomial, viz., $\omega^{\frac{p}{2q-p}}$. We chose the smoothing parameter λ to be $p \log N$, but other values that improve the performance can be chosen through cross-validation on various known graphs.

The non-monotone behavior of the relative error in Fig. 12a is likely due to a tradeoff between increasing p (which improves the accuracy) and increasing probability of observing a complete subgraph (which increases the variability, particularly in this case of large ω). Such behavior is apparent for moderate values of p (e.g., $p < 0.25$), but the effect of improved accuracy dominates for p near 1 and hence the relative error decreases to zero in that regime.



(a) Nonsmoothed $\widehat{\mathsf{cc}}$

(b) Smoothed $\widehat{cc}_L$

Figure 12: The parent graph is equal to a triangulated realization of $G(1000, 0.0015)$ with $d = 88$, $\omega = 15$, and $\mathsf{cc}(G) = 325$.

References

[1] Milton Abramowitz and Irene A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover Publications, Inc., New York, 1992. Reprint of the 1972 edition.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, and Sham M. Kakade. A tensor approach to learning mixed membership community models. *J. Mach. Learn. Res.*, 15:2239–2312, 2014.

[3] Coren L. Apicella, Frank W. Marlowe, James H. Fowler, and Nicholas A. Christakis. Social networks and cooperation in hunter-gatherers. *Nature*, 481(7382):497–501, 01 2012.

[4] Oriana Bandiera and Imran Rasul. Social networks and technology adoption in northern mozambique. *The Economic Journal*, 116(514):869–902, 2006.

[5] Anna Ben-Hamou, Roberto I Oliveira, and Yuval Peres. Estimating graph parameters via random walks with restarts. *arXiv preprint arXiv:1709.00869*, 2017.

[6] Petra Berenbrink, Bruce Krayenhoff, and Frederik Mallmann-Trenn. Estimating the number of connected components in sublinear time. *Inform. Process. Lett.*, 114(11):639–642, 2014.

[7] Peter J. Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *Ann. Statist.*, 39(5):2280–2301, 2011.

[8] Norman Biggs. On cluster expansions in graph theory and physics. *Quart. J. Math. Oxford Ser. (2)*, 29(114):159–173, 1978.

[9] 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. *Adv. Math.*, 219(6):1801–1851, 2008.

[10] Arun Chandrasekhar and Randall Lewis. Econometrics of sampled networks. *Unpublished manuscript*, 2011.

[11] Bernard Chazelle, Ronitt Rubinfeld, and Luca Trevisan. Approximating the minimum spanning tree weight in sublinear time. *SIAM J. Comput.*, 34(6):1370–1379, 2005.

[12] Timothy G. Conley and Christopher R. Udry. Learning about a new technology: Pineapple in ghana. *American Economic Review*, 100(1):35–69, March 2010.

[13] Graham Cormode and Nick Duffield. Sampling for big data: a tutorial. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1975–1975. ACM, 2014.

[14] Talya Eden, Amit Levi, Dana Ron, and C. Seshadhri. Approximately counting triangles in sublinear time. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*, pages 614–633. IEEE Computer Soc., Los Alamitos, CA, 2015.

[15] P. Erdös and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[16] Paul Erdös, László Lovász, and Joel Spencer. Strong independence of graphcopy functions. In *Graph theory and related topics (Proc. Conf., Univ. Waterloo, Waterloo, Ont., 1977)*, pages 165–172. Academic Press, New York-London, 1979.

[17] Marcel Fafchamps and Susan Lund. Risk-sharing networks in rural philippines. *Journal of development Economics*, 71(2):261–287, 2003.

[18] Uriel Feige. On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.*, 35(4):964–984, 2006.

[19] Benjamin Feigenberg, Erica M Field, and Rohini Pande. Building social capital through microfinance. Technical report, National Bureau of Economic Research, 2010.

[20] Ove Frank. Estimation of graph totals. *Scand. J. Statist.*, 4(2):81–89, 1977.

[21] Ove Frank. A note on Bernoulli sampling in graphs and Horvitz-Thompson estimation. *Scand. J. Statist.*, 4(4):178–180, 1977.

[22] Ove Frank. Survey sampling in graphs. *J. Stat. Plann. Inference*, 1(3):235–264, 1977.

[23] 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.

[24] Ove Frank. Sampling and inference in a population graph. *Internat. Statist. Rev.*, 48(1):33–41, 1980.

[25] Oded Goldreich and Dana Ron. Approximating average parameters of graphs. *Random Structures Algorithms*, 32(4):473–493, 2008.

[26] Mira Gonen, Dana Ron, and Yuval Shavitt. Counting stars and other small subgraphs in sublinear-time. *SIAM J. Discrete Math.*, 25(3):1365–1411, 2011.

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

[28] Leo A. Goodman. Snowball sampling. *Ann. Math. Statist.*, 32:148–170, 1961.

[29] Svante Janson. Large deviations for sums of partly dependent random variables. *Random Structures Algorithms*, 24(3):234–248, 2004.

[30] Jason M. Klusowski and Yihong Wu. Counting motifs in a graph via graph sampling. *Working paper*, 2017.

[31] W. L. Kocay. Some new methods in reconstruction theory. In *Combinatorial mathematics, IX (Brisbane, 1981)*, volume 952 of *Lecture Notes in Math.*, pages 89–114. Springer, Berlin-New York, 1982.

[32] Eric D Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Science & Business Media, 2009.

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

[34] Jure Leskovec and Christos Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, 2006.

[35] Brendan D. McKay and Stanisław P. Radziszowski. Subgraph counting identities and Ramsey numbers. *J. Combin. Theory Ser. B*, 69(2):193–209, 1997.

[36] Donald R. McNeil. Estimating an author's vocabulary. *J. Amer. Statist. Assoc.*, 68:92–96, 1973.

[37] Assaf Natanzon, Ron Shamir, and Roded Sharan. A polynomial approximation algorithm for the minimum fill-in problem. *SIAM J. Comput.*, 30(4):1067–1079, 2000.

[38] Ryan O'Donnell. *Analysis of Boolean functions*. Cambridge University Press, New York, 2014.

[39] 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.

[40] Donald J. Rose, R. Endre Tarjan, and George S. Lueker. Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.*, 5(2):266–283, 1976.

[41] Matthew J Salganik and Douglas D Heckathorn. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociological methodology*, 34(1):193–240, 2004.

[42] Carl J Schwarz and George AF Seber. Estimating animal abundance: review iii. *Statistical Science*, pages 427–456, 1999.

[43] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[44] Douglas B. West. *Introduction to graph theory*. Prentice Hall, Inc., Upper Saddle River, NJ, 1996.

[45] Hassler Whitney. The coloring of graphs. *Ann. of Math. (2)*, 33(4):688–718, 1932.

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