Bernoulli **26**(3), 2020, 1635–1664
https://doi.org/10.3150/19-BEJ1147

Estimating the number of connected components in a graph via subgraph sampling

JASON M. KLUSOWSKI[1] and YIHONG WU[2]

[1]*Department of Statistics, Rutgers University – New Brunswick, 110 Frelinghuysen Road, Piscataway, NJ, 8019, USA. E-mail: jason.klusowski@rutgers.edu*
[2]*Department of Statistics and Data Science, Yale University, 24 Hillhouse Avenue, New Haven, CT 06511, USA. E-mail: yihong.wu@yale.edu*

Learning properties of large graphs from samples has been an important problem in statistical network analysis since the early work of Goodman (*Ann. Math. Stat.* **20** (1949) 572–579) and Frank (*Scand. J. Stat.* **5** (1978) 177–188). We revisit a problem formulated by Frank (*Scand. J. Stat.* **5** (1978) 177–188) of estimating the number of connected components in a large graph based on the subgraph sampling model, in which we randomly sample a subset of the vertices and observe the induced subgraph. The key question is whether accurate estimation is achievable in the *sublinear* regime where only a vanishing fraction of the vertices are sampled. We show that it is impossible if the parent graph is allowed to contain high-degree vertices or long induced cycles. For the class of chordal graphs, where induced cycles of length four or above are forbidden, we characterize the optimal sample complexity within constant factors and construct linear-time estimators that provably achieve these bounds. This significantly expands the scope of previous results which have focused on unbiased estimators and special classes of graphs such as forests or cliques.

Both the construction and the analysis of the proposed methodology rely on combinatorial properties of chordal graphs and identities of induced subgraph counts. 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.

Keywords: chordal graph; minimax lower bound; network sampling; perfect elimination ordering; subgraph counts; subgraph sampling model

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) to more global features like the number of connected components – is an important task in network analysis. For example, the global clustering coefficient of a graph (i.e., the fraction of closed triangles) is a measure of the tendency for nodes to cluster together and a key quantity used to study cohesion in various networks [44]. To learn these graph properties, applied researchers typically collect data from a random sample of nodes to construct a representation of the true network. We refer to these problems collectively as *statistical inference on sampled networks*, where the goal is to infer properties of the parent network (population) from a subsampled version. Below we mention a few examples that arise in various fields of study.

- Sociology: Social networks of the Hadza hunter-gatherers of Tanzania were studied in [12] by surveying 205 individuals in 17 Hadza camps (from a population of 517). Another study

[11] of farmers in Ghana used network data from a survey of 180 households in three villages from a population of 550 households.
- Economics and business: Low sampling ratios have been used in applied economics (such as 30% in [18]), particularly for large scale studies [2,20]. A good overview of various experiments in applied economics and their corresponding sampling ratios can be found in [8], Appendix F, p. 11. Word of mouth marketing in consumer referral networks was studied in [51] using 158 respondents from a potential subject pool of 238.
- Genomics: The authors of [55] use protein-protein interaction data and demonstrate that it is possible to arrive at a reliable statistical estimate for the number of interactions (edges) from a sample containing approximately 1500 vertices.
- World Wide Web and Internet: Informed random IP address probing was used in [30] in an attempt to obtain a router-level map of the Internet.

As mentioned earlier, a primary concern of these studies is how well the data represent the true network and how to reconstruct the relevant properties of the parent graphs from samples. These issues and how they are addressed broadly arise from two perspectives:

- The full network is unknown due to the lack of data, which could arise from the underlying experimental design and data collection procedure, see, for example, historical or observational data. In this case, one needs to construct statistical estimators (i.e., functions of the sampled graph) to conduct sound inference. These estimators must be designed to account for the fact that the sampled network is only a partial observation of the true network, and thus subject to certain inherent biases and variability.
- The full network is either too large to scan or too expensive to store. In this case, approximation algorithms can overcome such computational or storage issues that would otherwise be unwieldy. For example, for massive social networks, it is generally impossible to enumerate the whole population. Rather than reading the entire graph, query-based algorithms randomly (or deterministically) sample parts of the graph or adaptively explore the graph through a random walk [3]. Some popular instances of traversal based procedures are snowball sampling [29] and respondent-driven sampling [54]. Indeed, sampling (based on edge and degree queries) is a commonly used primitive to speed up computation, which leads to various sublinear-time algorithms for testing or estimating graph properties such as the average degree [26], triangle and more general subgraph counts [1,15], expansion properties [27]; we refer the readers to the monograph [24].

Learning properties of graphs from samples has been an important problem in statistical network analysis since the early work of Goodman [28] and Frank [22]. Estimation of various properties such as graph totals [21] and connectivity [7,22] has been studied in a variety of sample models. However, most of the analysis has been confined to obtaining unbiased estimators for certain classes of graphs and little is known about their optimality. The purpose of this paper is to initiate a systematic study of statistical inference on sampled networks, with the goal of determining their statistical limits in terms of minimax risks and sample complexity, achieved by computationally efficient procedures.

As a first step towards this goal, in this paper we focus on a representative problem introduced in [22], namely, estimating the number of connected components in a graph from a partial sample of the population network. In fact, the techniques developed in this paper are also useful for estimating other graph statistics such as motif counts, which were studied in the companion paper [37].

Before we proceed, let us emphasize that the objective of this paper is *not* testing whether the graph is connected, which is a property too fragile to test on the basis of a small sampled graph; indeed, missing a single edge can destroy the connectivity. Instead, our goal is to estimate the number of connected components with an optimal additive accuracy. Thus, naturally, it is applicable to graphs with a large number of components.

We study the problem of estimating the number of connected components for two reasons. First, it encapsulates many challenging aspects of statistical inference on sampled graphs, and we believe the mathematical framework and machinery developed in this paper will prove useful for estimating other graph properties as well. Second, the number of connected components is a useful graph property that quantifies the connectivity of a network. In addition, it finds use in data-analytic applications related to determining the number of classes in a population [28]. Another example is the recent work [10], which studies the estimation of the number of documented deaths in the Syrian Civil War from a subgraph induced by a set of vertices obtained from an adaptive sampling process (similar to subgraph sampling). There, the goal is to estimate the number of unique individuals in a population, which roughly corresponds to the number of connected components in a network of duplicate records connected by shared attributes.

Next we discuss the sampling model, which 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 [13,19] for a list of techniques and [31,40,41] for comprehensive reviews). For simplicity, this paper focuses on the simplest sampling model, namely, *subgraph sampling*, where we randomly sample a subset of the vertices and observe their induced subgraph; in other words, only the edges between the sampled vertices are revealed. Results on estimating motif counts for the related neighborhood sampling model can be found in the companion paper [37]. One of the earliest works that adopts the subgraph sampling model is by Frank [22], which is the basis for the theory developed in this paper. Drawing from previous work on estimating population total using vertex sampling [21], Frank obtained unbiased estimators of the number of connected components and performance guarantees (variance calculations) for graphs whose connected components are either all trees or all cliques. Extensions to more general graphs are briefly discussed, although no unbiased estimators are proposed. This generality is desirable since it is more realistic to assume that the objects in each class (component) are in between being weakly and strongly connected to each other, corresponding to having the level of connectivity between a tree and clique. While the results of Frank are interesting, questions of their generality and optimality remain open and we therefore address these matters in the sequel. Specifically, the main goals of this paper are as follows:

- Characterize the sample complexity, that is, the minimal sample size to achieve a given accuracy, as a function of graph parameters.
- Devise computationally efficient estimators that provably achieve the optimal sample complexity bound.

Of particular interest is the *sublinear regime*, where only a vanishing fraction of the vertices are sampled. In this case, it is impossible to reconstruct the entire graph, but it might still be possible to accurately estimate the desired graph property.

The problem of estimating the number of connected components in a large graph has also been studied in the computer science literature, where the goal is to design randomized algorithms with sublinear (in the size of the graph) time complexity. The celebrated work [9] proposed a randomized algorithm to estimate the number of connected components in a general graph (motivated by computing the weight of the minimum spanning tree) within an additive error of ϵN for graphs with N vertices and average degree d_{avg}, with runtime $O(\frac{d_{\mathrm{avg}}}{\epsilon^2}\log\frac{d_{\mathrm{avg}}}{\epsilon})$. 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 a priori. The runtime was further improved to $O(\epsilon^{-2}\log\frac{1}{\epsilon})$ by modifying the stopping criterion [4]. In these algorithms, the breadth first search may visit many of the edges and explore a larger fraction of the graph at each round. From an applied perspective, such traversal based procedures can be impractical or impossible to implement in many statistical applications due to limitations inherent in the experimental design and it is more realistic to treat the network data as a random sample from a parent graph.

Finally, let us compare, conceptually, the framework in the present paper with the work on *model-based* network analysis, where networks are modeled as random graphs drawn from specific generative models, such as the stochastic block model [32], graphons [23], or exponential random graph models [33] (cf. the recent survey [41]), and performance analysis of statistical procedures for parameter estimation or clustering are carried out for these models. In contrast, in network sampling we adopt a *design-based* framework [31], where the graph is assumed to be deterministic and the randomness comes from the sampling process.

Organization

The paper is organized as follows. In Section 2, we formally define the estimation problem, the subgraph sampling model, and describe what classes of graphs we will be focusing on. To motivate our attention on specific classes of graphs (chordal graphs with maximum degree constraints), we show that in the absence of such structural assumptions, sublinear sample complexity is impossible in the sense that at least a constant faction of the vertices need to be sampled. Section 3 introduces the definition of chordal graphs and states our main results in terms of the minimax risk and sample complexity. In Section 4, after introducing the relevant combinatorial properties of chordal graphs, we define the estimator of the number of connect components and provide its statistical guarantees. We also propose a heuristic for constructing an estimator on non-chordal graphs. In Section 5, we develop a general strategy for proving minimax lower bound for estimating graph properties and particularize it to obtain matching lower bounds for the estimator constructed in Section 4.

The supplemental article [38] contains some of the technical proofs, additional results for the uniform sampling model and for forests and graphs with long cycles, and a numerical study of the proposed estimators on synthetic and real data for various graphs (see Appendix 6, Appendix 7, and Appendix 8, respectively).

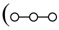
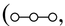
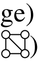
Notations

We use standard big-O notations, e.g., for any positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ or $a_n \lesssim b_n$ if $a_n \le Cb_n$ for some absolute constant $C > 0$, $a_n = o(b_n)$ or $a_n \ll b_n$ or if $\lim a_n/b_n = 0$. Furthermore, the subscript in $a_n = O_r(b_n)$ means $a_n \le C_r b_n$ for some constant C_r depending on the parameter r only. For positive integer k, let $[k] = \{1, \dots, k\}$. Let $\mathrm{Bern}(p)$ denote the Bernoulli distribution with mean p and $\mathrm{Bin}(N, p)$ the binomial distribution with N trials and success probability p.

Next, we introduce some graph-theoretic notations that will be used throughout the paper. Let $G = (V, E)$ be a simple undirected graph. Let $\mathsf{e} = \mathsf{e}(G) = |E(G)|$ denote the number of edges, $\mathsf{v} = \mathsf{v}(G) = |V(G)|$ 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 $N_G(u) = \{v \in V(G) : \{u, v\} \in E(G)\}$.

Two graphs G and G' are isomorphic, denoted by $G \simeq G'$, if there exists a bijection between the vertex sets of G and G' that preserves adjacency, that is, if there exists a bijective function $g : V(G) \to V(G')$ such that $\{g(u), g(v)\} \in E(G')$ if and only if $\{u, v\} \in E(G)$. The disjoint union of two graphs G and G', denoted $G + G'$, is the graph whose vertex (resp. edge) set is the disjoint union of the vertex (resp. edge) sets of G and of G'. For brevity, we denote by kG to the disjoint union of k copies of G.

We use the notation K_n, P_n, and C_n to denote the complete graph, path graph, and cycle graph on n vertices, respectively. Let $K_{n,n'}$ denote the complete bipartite graph with nn' edges and $n + n'$ vertices. Let S_n denote the star graph $K_{1,n}$ on $n + 1$ vertices.

We need two types of subgraph counts: Denote by $\mathsf{s}(H, G)$ (resp. $\mathsf{n}(H, G)$) the number of vertex (resp. edge) induced subgraphs of G that are isomorphic to H.[1] For example, $\mathsf{s}($, $) = 2$ and $\mathsf{n}($, $) = 8$. Let $\omega(G)$ denote the clique number, that is, the size of the largest clique in G.

2. Model

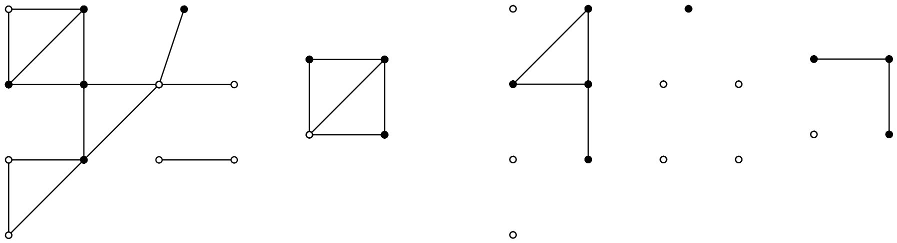
2.1. Subgraph sampling model

To fix notations, let $G = (V, E)$ be a simple, undirected graph on N vertices. In the subgraph sampling model, we sample a set of vertices denoted by $S \subset V$, and observe their induced subgraph, denoted by $G[S] = (S, E[S])$, where the edge set is defined as $E[S] = \{\{i, j\} \in E : i, j \in S\}$. See Figure 1 for an illustration. To simplify notations, we abbreviate the sampled graph $G[S]$ as $\widetilde{G}$.

According to how the set S of sampled vertices is generated, there are two variations of the subgraph sampling model [22]:

---

[1] The subgraph counts are directly related to the graph homomorphism numbers [43], Section 5.2. Denote by $\mathsf{inj}(H, G)$ the number of injective homomorphisms from H to G and $\mathsf{ind}(H, G)$ the number of injective homomorphisms that also preserve non-adjacency. Then $\mathsf{ind}(H, G) = \mathsf{s}(H, G)\mathsf{aut}(H)$ and $\mathsf{inj}(H, G) = \mathsf{n}(H, G)\mathsf{aut}(H)$, where $\mathsf{aut}(H)$ denotes the number of automorphisms (i.e. isomorphisms to itself) for H.



(a) Parent graph G with the set of sampled vertices S shown in black.

(b) Subgraph induced by sampled vertices $\widetilde{G} = G[S]$. Non-sampled vertices are shown as isolated vertices.

Figure 1. Subgraph sampling.

- *Uniform sampling*: Exactly n vertices are chosen uniformly at random without replacement from the vertex set V. In this case, the probability of observing a subgraph isomorphic[2] to H with $\mathsf{v}(H) = n$ is equal to

$$\mathbb{P}[\widetilde{G} \simeq H] = \frac{\mathsf{s}(H,G)}{\binom{N}{n}}. \tag{1}$$

- *Bernoulli sampling*: Each vertex is sampled independently with probability p, where p is called the *sampling ratio*. Thus, the sample size $|S|$ is distributed as $\mathrm{Bin}(N, p)$, and the probability of observing a subgraph isomorphic to H is equal to

$$\mathbb{P}[\widetilde{G} \simeq H] = \mathsf{s}(H,G)p^{\mathsf{v}(H)}(1-p)^{\mathsf{v}(G)-\mathsf{v}(H)}. \tag{2}$$

The relation between these two models is analogous to that between sampling without replacements and sampling with replacements. In the sublinear sampling regime where $n \ll N$, they are nearly equivalent. For technical simplicity, we focus on the Bernoulli sampling model and we refer to $n \triangleq pN$ as the *effective sample size*. Extensions to the uniform sampling model will be discussed in Section 7.1 of Appendix 7.

A number of previous work on subgraph sampling is closely related with the theory of graph limits [6], which is motivated by the so-called property testing problems in graphs [24]. According to [6], Definition 2.11, a graph parameter f is "testable" if for any $\epsilon > 0$, there exists a sample size n such that for any graph G with at least n vertices, there is an estimator $\widehat{f} = \widehat{f}(\widetilde{G})$ such that $\mathbb{P}[|f(G) - \widehat{f}| > \epsilon] < \epsilon$. In other words, testable properties can be estimated with sample complexity that is *independent* of the size of the graph. Examples of testable properties include the edge density $\mathsf{e}(G)/\binom{\mathsf{v}(G)}{2}$ and the density of maximum cuts $\frac{\mathsf{MaxCut}(G)}{\mathsf{v}(G)^2}$, where $\mathsf{MaxCut}(G)$ is the size of the maximum edge cut-set in G [25]; however, the number of connected components

---

[2] Note that it is sufficient to describe the sampled graph up to isomorphism since the property cc we want to estimate is invariant under graph isomorphisms.

$\mathsf{cc}(G)$ or its normalized version $\frac{\mathsf{cc}(G)}{\mathsf{v}(G)}$ are not testable.[3] Instead, our focus is to understand the dependency of sample complexity of estimating $\mathsf{cc}(G)$ on the graph size N as well as other graph parameters. It turns out for certain classes of graphs, the sample complexity grows *sublinearly* in N, which is the most interesting regime.

2.2. Classes of graphs

Before introducing the classes of graphs we consider in this paper, we note that, unless further structures are assumed about the parent graph, estimating many graph properties, including the number of connected components, has very high sample complexity that scales linearly with the size of the graph. Indeed, there are two main obstacles in estimating the number of connected components in graphs, namely, *high-degree vertices* and *long induced cycles*. If either is allowed to be present, we will show that even if we sample a constant faction of the vertices, any estimator of $\mathsf{cc}(G)$ has a worst-case additive error that is almost linear in the network size N. Specifically,

- For any sampling ratio p bounded away from 1, as long as the maximum degree is allowed to scale as $\Omega(N)$, even if we restrict the parent graph to be acyclic, the worst-case estimation error for any estimator is $\Omega(N)$.
- For any sampling ratio p bounded away from 1/2, as long as the length of the induced cycles is allowed to be $\Omega(\log N)$, even if we restrict the parent graph to have maximum degree 2, the worst-case estimation error for any estimator is $\Omega(\frac{N}{\log N})$.

The precise statements follow from the minimax lower bounds in Theorem 14 and Theorem 13 of Appendix 7. Below we provide an intuitive explanation for each scenario.

For the first claim involving large degree, consider a pair of acyclic graphs G and G', where G is the star graph on N vertices and G' consisting of N isolated vertices. Note that as long as the center vertex in G is not sampled, the sampling distributions of G and G' are identical. This implies that the total variation between the sampled graph under G and G' is at most p. Since the numbers of connected components in G and G' differ by $N-1$, this leads to a minimax lower bound for the estimation error of $\Omega(N)$ whenever p is bounded away from one.

The effect of long induced cycles is subtler. The key observation is that a cycle and a path (or a cycle versus two cycles) locally look exactly the same. Indeed, let G (resp. G') consists of $N/(2r)$ disjoint copies of the smaller graph H (resp. H'), where H is a cycle of length $2r$ and H' consists of two disjoint cycles of length r. Both G and G' have maximum degree 2 and contain induced cycles of length at most $2r$. The local structure of G and G' is the same (e.g., each connected subgraph with at most $r-1$ vertices appears exactly N times in each graph) and the sampled versions of H and H' are identically distributed provided at most $r-1$ vertices are sampled. Thus, we must sample at least r vertices (which occurs with probability at most $e^{-r(1-2p)^2}$) for the distributions to be different. By a union bound, it can be shown that the total variation between the sampled graphs $\widetilde{G}$ and $\widetilde{G}'$ is $O((N/r)e^{-r(1-2p)^2})$. Thus, whenever the

---

[3] To see this, recall from [6], Theorem 6.1(b), an equivalent characterization of f being testable is that for any $\epsilon > 0$, there exists a sample size n such that for any graph G with at least n vertices, $|f(G) - \mathbb{E}f(\widetilde{G})| < \epsilon$. This is violated for star graphs $G = S_N$ as $N \to \infty$.

sampling ratio p is bounded away from $1/2$, choosing $r = \Theta(\log N)$ leads to a near-linear lower bound $\Omega(\frac{N}{\log N})$.

The difficulties caused by high-degree vertices and long induced cycles motivate us to consider classes of graphs defined by two key parameters, namely, the maximum degree d and the length of the longest induced cycles c. The case of $c = 2$ corresponds to forests (acyclic graphs), which have been considered by Frank [22]. The case of $c = 3$ corresponds to *chordal graphs*, that is, graphs without induced cycle of length four or above, which is the focus of this paper. It is well known that various computation tasks that are intractable in the worst case, such as maximal clique and graph coloring, are easy for chordal graphs; it turns out that the chordality structure also aids in both the design and the analysis of computationally efficient estimators which provably attain the optimal sample complexity.

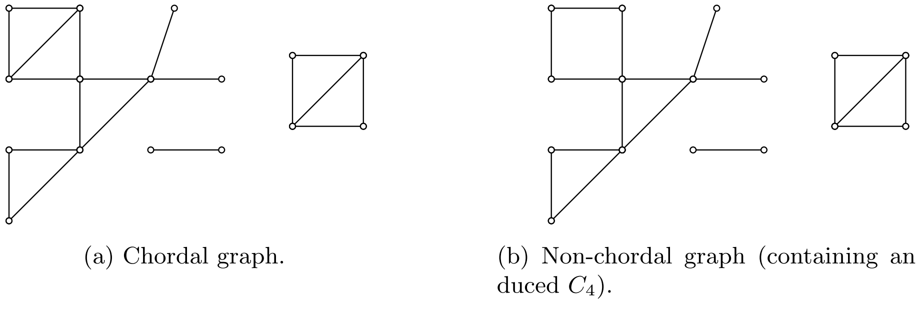
3. Main results

This section summarizes our main results in terms of the minimax risk of estimating the number of connected components over various class of graphs. As mentioned before, for ease of exposition, we focus on the Bernoulli sampling model, where each vertex is sampled independently with probability p. Similar conclusions can be obtained for the uniform sampling model upon identifying $p = n/N$, as given in Section 7.1.

When p grows from 0 to 1, an increasing fraction of the graph is observed and intuitively the estimation problem becomes easier. Indeed, all forthcoming minimax rates are inversely proportional to powers of p. Of particular interest is whether accurate estimation in the sublinear sampling regime, that is, $p = o(1)$. The forthcoming theory will give explicit conditions on p for this to hold true.

As mentioned in the previous section, the main class of graphs we study is the so-called *chordal graphs* (see Figure 2 for an example):

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



(a) Chordal graph.

(b) Non-chordal graph (containing an induced C_4).

Figure 2. Examples of chordal and non-chordal graphs both with three connected components.

We emphasize that chordal graphs are allowed to have arbitrarily long cycles but no induced cycles longer than three. The class of chordal graphs encompasses *forests* and *disjoint union of cliques* as special cases, the two models that were studied in Frank's original paper [22]. In addition to constructing estimators that adapt to larger collections of graphs (for which forests and unions of cliques are special cases), we also provide theoretical analysis and optimality guarantees – elements that were not considered in past work.

Next, we characterize the rate of the minimax mean-squared error for estimating the number of connected components in a chordal graph, which turns out to depend on the number of vertices, the maximum degree, and the clique number. The upper and lower bounds differ by at most a multiplicative factor depending only on the clique number. To simplify the notation, henceforth we denote $q = 1 - p$.

Theorem 1 (Chordal graphs). *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with maximum degree and clique number at most d and $\omega \geq 2$, respectively. Then*

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

where the lower bound holds provided that $p \leq p_0$ for some constant $p_0 < \frac{1}{2}$ that only depends on ω. Furthermore, if $p \geq 1/2$, then for any ω,

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d,\omega)} \mathbb{E}_G \left|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\right|^2 \leq Nq(d+1). \tag{3}$$

Specializing Theorem 1 to $\omega = 2$ yields the minimax rates for estimating the number of trees in forests for small sampling ratio p. The next theorem shows that the result holds verbatim even if p is arbitrarily close to 1, and, consequently, shows minimax rate-optimality of the bound in (3). The lower bound component is proved in Section 7.3 of Appendix 7.

Theorem 2 (Forests). *Let $\mathcal{F}(N, d) \triangleq \mathcal{G}(N, d, 2)$ denote the collection of all forests on N vertices with maximum degree at most d. Then for all $0 \leq p \leq 1$ and $1 \leq d \leq N$,*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{F}(N,d)} \mathbb{E}_G \left|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\right|^2 \asymp \left(\frac{Nq}{p^2} \vee \frac{Nqd}{p}\right) \wedge N^2. \tag{4}$$

The upper bounds in the previous results are achieved by unbiased estimators. As (3) shows, they work well even when the clique number ω grow with N, provided we sample more than half of the vertices; however, if the sample ratio p is below $\frac{1}{2}$, especially in the sublinear regime of $p = o(1)$ that we are interested in, the variance is exponentially large. To deal with large d and ω, we must give up unbiasedness to achieve a good bias-variance tradeoff. Such biased estimators, obtained using the smoothing technique introduced in [49], lead to better performance as quantified in the following theorem. The proofs of these bounds are given in Theorem 7 and Theorem 9.

Theorem 3 (Chordal graphs). *Let $\mathcal{G}(N, d)$ denote the collection of all chordal graphs on N vertices with maximum degree at most d. Then, for any $p < 1/2$,*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{G}(N,d)} \mathbb{E}_G \big|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\big|^2 \lesssim N^2 \big(N/d^2\big)^{-\frac{p}{2-3p}}.$$

Finally, for the special case of graphs consisting of disjoint union of cliques, as the following theorem shows, there are enough structures so that we no longer need to impose any condition on the maximal degree. Similar to Theorem 3, the achievable scheme is a biased estimator, significantly improving the unbiased estimator in [22,28] which has exponentially large variance.

Theorem 4 (Cliques). *Let $\mathcal{C}(N)$ denote the collection of all graphs on N vertices consisting of disjoint unions of cliques. Then, for any $p < 1/2$,*

$$\inf_{\widehat{\mathsf{cc}}} \sup_{G \in \mathcal{C}(N)} \mathbb{E}_G \big|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\big|^2 \le N^2 (N/4)^{-\frac{p}{2-3p}}.$$

Alternatively, the above results are summarized in Table 1 in terms of the *sample complexity*, that is, the minimum sample size that allows an estimator $\mathsf{cc}(G)$ within an additive error of ϵN with probability, say, at least 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:

A consequence of Theorem 2 is that if the effective sample size n scales as $O(\max(\sqrt{N}, d))$, for the class of forests $\mathcal{F}(N, d)$ the worse-case estimation error for any estimator is $\Omega(N)$, which is within a constant factor to the trivial error bound when no samples are available. Conversely, if $n \gg \max(\sqrt{N}, d)$, which is sublinear in N as long as the maximal degree satisfies $d = o(N)$, then it is possible to achieve a non-trivial estimation error of $o(N)$. More generally for chordal graphs, Theorem 1 implies that if $n = O(\max(N^{\frac{\omega-1}{\omega}}, d^{\frac{1}{\omega-1}} N^{\frac{\omega-2}{\omega-1}}))$, the worse-case estimation error in $\mathcal{G}(N, d, \omega)$ for any estimator is at least $\Omega_\omega(N)$,

Table 1. Sample complexity for various classes of graphs

Graph	Sample complexity n
Chordal	$\Theta_\omega\left(\max\left\{N^{\frac{\omega-2}{\omega-1}} d^{\frac{1}{\omega-1}} \epsilon^{-\frac{2}{\omega-1}}, N^{\frac{\omega-1}{\omega}} \epsilon^{-\frac{2}{\omega}}\right\}\right)$
Forest	$\Theta\left(\max\left\{\frac{d}{\epsilon^2}, \frac{\sqrt{N}}{\epsilon}\right\}\right)$
Cliques	$\Theta\left(\frac{N}{\log N} \log \frac{1}{\epsilon}\right)$, $\epsilon \ge N^{-1/2+\Omega(1)}$*

*The lower bound part of this statement follows from [60], Section 3, which shows the optimality of Theorem 4.

4. Algorithms and performance guarantees

In this section, we propose estimators which provably achieve the upper bounds presented in Section 3 for the Bernoulli sampling model. In Section 4.1, we highlight some combinatorial properties and characterizations of chordal graphs that underpin both the construction and the analysis of the estimators in Section 4.2. The special case of disjoint unions of cliques is treated in Section 4.3, where the estimator of Frank [22] is recovered and further improved. Analogous results for the uniform sampling model are given in Section 7.1 of Appendix 7. Finally, in Section 4.4, we discuss a heuristic to generalize the methodology to non-chordal graphs.

4.1. Combinatorial properties of chordal graphs

In this subsection, we discuss the relevant combinatorial properties of chordal graphs which aid in the design and analysis of our estimators. We start by introducing a notion of vertex elimination ordering.

Definition 2. *A perfect elimination ordering (PEO) of a graph G on N vertices is a vertex labelling $\{v_1, v_2, \dots, v_N\}$ such that, for each j, $N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$ is a clique.*

In other words, if one eliminates the vertices sequentially according to a PEO starting from the last vertex, at each step, the neighborhood of the vertex to be eliminated forms a clique; see Figure 3 for an example. A classical result of Dirac asserts that the existence of a PEO is in fact the defining property of chordal graphs (cf. e.g., [58], Theorem 5.3.17).

Theorem 5. *A graph is chordal if and only if it admits a PEO.*

In general a PEO of a chordal graph is not unique; however, it turns out that the size of each neighborhood in the vertex elimination process is unique up to permutation, a fact that we will exploit later on. The next lemma makes this claim precise. For brevity, we defer its proof to Appendix 6.

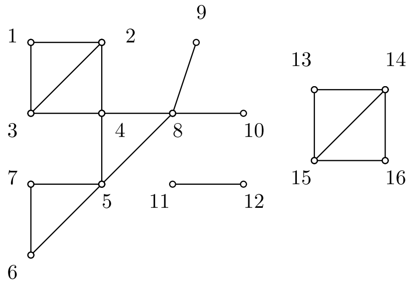


Figure 3. A chordal graph G with PEO labelled. In this example, $\mathsf{cc}(G) = 3 = 16 - 19 + 6 = \mathsf{s}(K_1, G) - \mathsf{s}(K_2, G) + \mathsf{s}(K_3, G)$.

Lemma 1. *Let* $\{v_1, \dots, v_N\}$ *and* $\{v'_1, \dots, v'_N\}$ *be two PEOs of a chordal graph* G. *Let* c_j *and* c'_j *denote the cardinalities of* $N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$ *and* $N_G(v'_j) \cap \{v'_1, \dots, v'_{j-1}\}$, *respectively. Then there is a bijection* $\sigma : [N] \to [N]$ *such that* $\mathsf{c}_{\sigma(j)} = \mathsf{c}'_j$ *for all* j.

Recall that $\mathsf{s}(K_i, G)$ denotes the number of cliques of size i in G. For any chordal graph G, it turns out that the number of components can be expressed as an alternating sum of clique counts (cf. e.g., [58], Exercise 5.3.22, p. 231); see Figure 3 for an example. Instead of the topological proof involving properties of the clique simplex of chordal graphs [14,46], in the next lemma we provide a combinatorial proof together with a sandwich bound. The main purpose of this exposition is to explain how to enumerate cliques in chordal graphs using vertex elimination, which plays a key role in analyzing the statistical estimator developed in the next subsection.

Lemma 2. *For any chordal graph* G,

$$\mathsf{cc}(G) = \sum_{i \ge 1} (-1)^{i+1} \mathsf{s}(K_i, G). \tag{5}$$

Furthermore, for any $r \ge 1$,

$$\sum_{i=1}^{2r} (-1)^{i+1} \mathsf{s}(K_i, G) \le \mathsf{cc}(G) \le \sum_{i=1}^{2r-1} (-1)^{i+1} \mathsf{s}(K_i, G). \tag{6}$$

Proof. Since G is chordal, by Theorem 5, it has a PEO $\{v_1, \dots, v_N\}$. Define

$$C_j \triangleq N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}, \quad \mathsf{c}_j \triangleq |C_j|. \tag{7}$$

Since the neighbors of v_j among $v_1, \dots, v_{j-1}$ 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 subgraph induced by $v_1, \dots, v_{j-1}$. Thus,

$$\mathsf{s}(K_i, G) = \sum_{j=1}^{N} \binom{\mathsf{c}_j}{i-1}. \tag{8}$$

Moreover, note that $\mathsf{cc}(G) = \sum_{j=1}^{N} \mathbb{1}\{\mathsf{c}_j = 0\}$. Hence, it follows that

$$\begin{aligned} &\sum_{i=1}^{2r-1} (-1)^{i+1} \mathsf{s}(K_i, G) \\ &\quad = \sum_{i=1}^{2r-1} (-1)^{i+1} \sum_{j=1}^{N} \binom{\mathsf{c}_j}{i-1} = \sum_{j=1}^{N} \sum_{i=1}^{2r-1} (-1)^{i+1} \binom{\mathsf{c}_j}{i-1} \\ &\quad = \sum_{j=1}^{N} \sum_{i=0}^{2(r-1)} (-1)^{i} \binom{\mathsf{c}_j}{i} \end{aligned}$$

$$= \sum_{j=1}^{N} \left(\binom{\mathsf{c}_j - 1}{2(r-1)} \mathbb{1}\{\mathsf{c}_j \neq 0\} + \mathbb{1}\{\mathsf{c}_j = 0\} \right)$$

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

and

$$\sum_{i=1}^{2r} (-1)^{i+1} \mathsf{s}(K_i, G) = \sum_{i=1}^{2r} (-1)^{i+1} \sum_{j=1}^{N} \binom{\mathsf{c}_j}{i-1} = \sum_{j=1}^{N} \sum_{i=1}^{2r} (-1)^{i+1} \binom{\mathsf{c}_j}{i-1}$$

$$= \sum_{j=1}^{N} \sum_{i=0}^{2r-1} (-1)^{i} \binom{\mathsf{c}_j}{i} = \sum_{j=1}^{N} \left(-\binom{\mathsf{c}_j - 1}{2r-1} \mathbb{1}\{\mathsf{c}_j \neq 0\} + \mathbb{1}\{\mathsf{c}_j = 0\} \right)$$

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

□

4.2. Estimators for chordal graphs

4.2.1. *Bounded clique number*: *Unbiased estimators*

In this subsection, we consider unbiased estimation of the number of connected components in chordal graphs. As we will see, unbiased estimators turn out to be minimax rate-optimal for chordal graphs with bounded clique size. The subgraph count identity (5) suggests the following unbiased estimator

$$\widehat{\mathsf{cc}} = -\sum_{i \geq 1} \left(-\frac{1}{p} \right)^{i} \mathsf{s}(K_i, \widetilde{G}). \tag{9}$$

Indeed, since the probability of observing any given clique of size i is p^i, (9) is clearly unbiased in the same spirit of the Horvitz-Thompson estimator [34]. In the case where the parent graph G is a forest, (9) reduces to the estimator $\widehat{\mathsf{cc}} = \mathsf{v}(\widetilde{G})/p - \mathsf{e}(\widetilde{G})/p^2$, as proposed by Frank [22].

A few comments about the estimator (9) are in order. First, it is completely adaptive to the parameters ω, d and N, since the sum in (9) terminates at the clique number of the subsampled graph. Second, it can be evaluated in time that is linear in $\mathsf{v}(\widetilde{G}) + \mathsf{e}(\widetilde{G})$. Indeed, the next lemma gives a simple formula for computing (9) using the PEO. Since a PEO of a chordal graph G can be found in $O(\mathsf{v}(G) + \mathsf{e}(G))$ time [52,56] and any induced subgraph of a chordal graph remains chordal, the estimator (9) can be evaluated in linear time.[4] Recall that $q = 1 - p$.

---

[4] The algorithm in [56] is implemented in `R` using the `max_cardinality()` function in the graph package `igraph`.

Lemma 3. *Let* $\{\widetilde{v}_1, \dots, \widetilde{v}_m\}$, $m = |S|$, *be a PEO of* $\widetilde{G}$. *Then*

$$\widehat{\mathsf{cc}} = \frac{1}{p} \sum_{j=1}^{m} \left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_j}, \tag{10}$$

where $\widetilde{\mathsf{c}}_j \triangleq |N_{\widetilde{G}}(\widetilde{v}_j) \cap \{\widetilde{v}_1, \dots, \widetilde{v}_{j-1}\}|$ *can be calculated from* $\widetilde{G}$ *in linear time.*

Proof. Because the subsampled graph $\widetilde{G}$ is also chordal, by (8), we have $\mathsf{s}(K_i, \widetilde{G}) = \sum_{j=1}^{m} \binom{\widetilde{\mathsf{c}}_j}{i-1}$. Thus, (9) can also be written as

$$\begin{aligned}
\widehat{\mathsf{cc}} &= -\sum_{i=1}^{m} \left(-\frac{1}{p}\right)^{i} \mathsf{s}(K_i, \widetilde{G}) = -\sum_{i=1}^{m} \left(-\frac{1}{p}\right)^{i} \sum_{j=1}^{m} \binom{\widetilde{\mathsf{c}}_j}{i-1} \\
&= -\sum_{j=1}^{m} \sum_{i=1}^{m} \left(-\frac{1}{p}\right)^{i} \binom{\widetilde{\mathsf{c}}_j}{i-1} = \frac{1}{p} \sum_{j=1}^{m} \sum_{i=0}^{m-1} \left(-\frac{1}{p}\right)^{i} \binom{\widetilde{\mathsf{c}}_j}{i} \\
&= \frac{1}{p} \sum_{j=1}^{m} \left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_j}.
\end{aligned}$$

□

In addition to the aforementioned computational advantages of using (10) over (9), let us also describe why (10) is more numerically stable. Both estimators are equal to an alternating sum of the form $\sum_i a_i (-1/p)^{b_i}$. In (10), $a_i = q^{b_i}/p$, whereas $a_i = -\mathsf{s}(K_{b_i}, \widetilde{G})$ in (9), which can be as large as $O(N 2^{\omega})$ in magnitude. Thus, when $\widetilde{G}$ is sufficiently dense, computation of (9) involves adding and subtracting extremely large numbers – making it prone to integer overflow and suffer from loss of numerical precision. For example, double-precision floating-point arithmetic (e.g., used in `R`) gives from 15 to 17 significant decimal digits precision. In our experience, this tends to be insufficient for most mid-sized, real-world networks (see Appendix 8) and the estimator (9) outputs wildly imprecise numbers.

Using elementary enumerative combinatorics, in particular, the vertex elimination structure of chordal graphs, the next theorem provides a performance guarantee for the estimator (9) in terms of a variance bound and a high-probability bound, which, in particular, settles the upper bound of the minimax mean squared error in Theorem 1 and Theorem 2.

Theorem 6. *Let* G *be a chordal graph on* N *vertices with maximum degree and clique number at most* d *and* $\omega \geq 2$, *respectively. Suppose* $\widetilde{G}$ *is generated by the* Bern(p) *sampling model. Then* $\widehat{\mathsf{cc}}$ *defined in* (9) *is an unbiased estimator of* $\mathsf{cc}(G)$. *Furthermore,*

$$\mathsf{Var}[\widehat{\mathsf{cc}}] \leq N\left(\frac{q}{p} + d\right)\left(\left(\frac{q}{p}\right)^{\omega-1} \vee \frac{q}{p}\right) \leq \frac{N}{p^{\omega}} + \frac{Nd}{p^{\omega-1}}, \tag{11}$$

and for all $t \geq 0$,

$$\mathbb{P}\left[\left|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\right| \geq t\right] \leq 2\exp\left\{-\frac{8p^{\omega}t^2}{25(d\omega+1)(N+t/3)}\right\}. \tag{12}$$

To prove Theorem 6, we start by presenting a useful lemma. Note that Lemma 3 states that $\widehat{\mathsf{cc}}$ is a linear combination of $(-q/p)^{\widetilde{\mathsf{c}}_j}$; here $\widetilde{\mathsf{c}}_j$ is computed using a PEO of the sampled graph, which itself is random. The next result allows us rewrite the same estimator as a linear combination of $(-q/p)^{\widehat{\mathsf{c}}_j}$, where $\widehat{\mathsf{c}}_j$ depends on the PEO of the parent graph (which is deterministic). Note that this is only used in the course of analysis since the population level PEO is not observed. This representation is extremely useful in analyzing the performance of $\widehat{\mathsf{cc}}$ and its biased variant in Section 4.2.2. More generally, we have the following result, which we prove in Appendix 6.

Lemma 4. *Let* $\{v_1, \dots, v_N\}$ *be a PEO of* G *and let* $\{\widetilde{v}_1, \dots, \widetilde{v}_m\}$, $m = |S|$, *be a PEO of* $\widetilde{G}$. *Furthermore, let* $\widehat{\mathsf{c}}_j = |N_{\widetilde{G}}(v_j) \cap \{v_1, \dots, v_{j-1}\}|$ *and* $\widetilde{\mathsf{c}}_j = |N_{\widetilde{G}}(\widetilde{v}_j) \cap \{\widetilde{v}_1, \dots, \widetilde{v}_{j-1}\}|$. *Let* $\widehat{\mathsf{g}} = \widehat{\mathsf{g}}(\widetilde{G})$ *be a linear estimator of the form*

$$\widehat{\mathsf{g}} = \sum_{j=1}^{m} g(\widetilde{\mathsf{c}}_j). \tag{13}$$

Then

$$\widehat{\mathsf{g}} = \sum_{j=1}^{N} b_j g(\widehat{\mathsf{c}}_j),$$

where $b_j \triangleq \mathbb{1}\left\{v_j \in S\right\}$.

We also need a couple of ancillary results whose proofs are also given in Appendix 6:

Lemma 5 (Orthogonality). *Let*[5]

$$f(k) = \left(-\frac{q}{p}\right)^k, \quad k \geq 0. \tag{14}$$

Let $\{b_v : v \in V\}$ *be independent* Bern(p) *random variables. For any* $S \subset V$, *define* $N_S = \sum_{v\in S} b_v$. *Then*

$$\mathbb{E}\left[f(N_S)f(N_T)\right] = \mathbb{1}\{S = T\}(q/p)^{|S|}.$$

In particular, $\mathbb{E}[f(N_S)] = 0$ *for any* $S \neq \varnothing$.

---

[5] In fact, the function $f(N_S) = (-\frac{q}{p})^{N_S}$ is the (unnormalized) orthogonal basis for the binomial measure that is used in the analysis of Boolean functions [48], Definition 8.40.

Lemma 6. *Let $\{v_1, \dots, v_N\}$ be a PEO of a chordal graph G on N vertices with maximum degree and clique number at most d and ω, respectively. Let $C_j \triangleq N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}$. Then*[6]

$$\bigl|\{(i, j) : i \neq j, C_j = C_i \neq \varnothing\}\bigr| \le N(d-1). \tag{15}$$

Furthermore, let

$$A_j = \{v_j\} \cup C_j. \tag{16}$$

Then for each $j \in [N]$,

$$\bigl|\{i \in [N] : i \neq j, A_i \cap A_j \neq \varnothing\}\bigr| \le d\omega. \tag{17}$$

Proof of Theorem 6. For a chordal graph G on N vertices, let $\{v_1, \dots, v_N\}$ be a PEO of G. Recall from (7) that C_j denote the set of neighbors of v_j among $v_1, \dots, v_{j-1}$ and c_j denotes its cardinality. That is,

$$\mathsf{c}_j = \bigl|N_G(v_j) \cap \{v_1, \dots, v_{j-1}\}\bigr| = \sum_{k=1}^{j-1} \mathbb{1}\{v_k \sim v_j\}.$$

As in Lemma 4, let $\widehat{\mathsf{c}}_j$ denote the sample version, i.e.,

$$\widehat{\mathsf{c}}_j \triangleq \bigl|N_{\widetilde{G}}(v_j) \cap \{v_1, \dots, v_{j-1}\}\bigr| = b_j \sum_{k=1}^{j-1} b_k \mathbb{1}\{v_k \sim v_j\},$$

where $b_k \triangleq \mathbb{1}\{v_k \in S\} \overset{\text{i.i.d.}}{\sim} \text{Bern}(p)$. By Lemma 3 and Lemma 4, $\widehat{\mathsf{cc}}$ can be written as

$$\widehat{\mathsf{cc}} = \frac{1}{p} \sum_{j=1}^{m} f(\widetilde{\mathsf{c}}_j) = \frac{1}{p} \sum_{j=1}^{N} b_j f(\widehat{\mathsf{c}}_j), \tag{18}$$

where the function f is defined in (14).

To show the variance bound (11), we note that

$$\mathsf{Var}[\widehat{\mathsf{cc}}] = \frac{1}{p^2} \sum_{j=1}^{N} \mathsf{Var}\bigl[b_j f(\widehat{\mathsf{c}}_j)\bigr] + \frac{1}{p^2} \sum_{j \neq i} \mathsf{Cov}\bigl[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{c}_i)\bigr]. \tag{19}$$

Note that $\widehat{\mathsf{c}}_j \mid \{b_j = 1\} \sim \text{Bin}(\mathsf{c}_j, p)$. Using Lemma 5, it is straightforward to verify that

$$\mathsf{Var}\bigl[b_j f(\widehat{\mathsf{c}}_j)\bigr] = \begin{cases} p\left(\dfrac{q}{p}\right)^{\mathsf{c}_j} & \text{if } \mathsf{c}_j > 0, \\ pq & \text{if } \mathsf{c}_j = 0. \end{cases} \tag{20}$$

---

[6] The bound in (15) is almost optimal, since the left-hand side is equal to $N(d-2)$ when G consists of $N/(d+1)$ copies of stars S_d.

Since $\mathsf{c}_j \leq \omega - 1$, it follows that

$$\mathsf{Var}\big[b_j f(\widehat{\mathsf{c}}_j)\big] \leq p\left[\left(\frac{q}{p}\right)^{\omega-1} \vee \frac{q}{p}\right]. \tag{21}$$

The covariance terms are less obvious to bound; but thanks to the orthogonality property in Lemma 5, many of them are zero or negative. Let $N_C \triangleq \sum b_j \mathbb{1}\left\{v_j \in C\right\}$. For any j, since $v_j \notin C_j$ by definition, applying Lemma 5 yields

$$\mathbb{E}\big[b_j f(\widehat{\mathsf{c}}_j)\big] = p\mathbb{E}\big[f(N_{C_j})\big] = p\mathbb{1}\left\{C_j = \varnothing\right\}. \tag{22}$$

Without loss of generality, assume $j < i$. By the definition of C_j, we have $v_i \notin C_j$. Next, we consider two cases separately:

Case I: $v_j \notin C_i$. If either C_j or C_i is nonempty, Lemma 5 yields

$$\begin{aligned}\mathsf{Cov}\big[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{\mathsf{c}}_i)\big] &\overset{(22)}{=} \mathbb{E}\big[b_i b_j f(\widehat{\mathsf{c}}_j) f(\widehat{\mathsf{c}}_i)\big] \\ &= p^2\mathbb{E}\big[f(N_{C_j}) f(N_{C_i})\big] \\ &= p^2\mathbb{1}\{C_j = C_i\}\left(\frac{q}{p}\right)^{\mathsf{c}_j}.\end{aligned}$$

If $C_j = C_i = \varnothing$, then $\mathsf{Cov}[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{\mathsf{c}}_i)] = \mathsf{Cov}[b_j, b_i] = 0$.

Case II: $v_j \in C_i$. Then $\mathbb{E}[b_i f(\widehat{\mathsf{c}}_i)] = 0$ by (22). Using Lemma 5 again, we have

$$\begin{aligned}\mathsf{Cov}\big[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{\mathsf{c}}_i)\big] &= p\mathbb{E}\left[b_j\left(-\frac{q}{p}\right)^{b_j}\right]\mathbb{E}\big[f(N_{C_j}) f(N_{C_i\setminus\{v_j\}})\big] \\ &= -pq\mathbb{E}\big[f(N_{C_j}) f(N_{C_i\setminus\{v_j\}})\big] \\ &= -pq\mathbb{1}\big\{C_j = C_i\setminus\{v_j\}\big\}\left(\frac{q}{p}\right)^{\mathsf{c}_j}.\end{aligned}$$

To summarize, we have shown that

$$\mathsf{Cov}\big[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{\mathsf{c}}_i)\big] = \begin{cases} p^2\left(\dfrac{q}{p}\right)^{\mathsf{c}_j} & \text{if } C_j = C_i \neq \varnothing, \\ -pq\left(\dfrac{q}{p}\right)^{\mathsf{c}_j} & \text{if } C_j = C_i\setminus\{v_j\} \text{ and } v_j \in C_i, \\ 0 & \text{otherwise.}\end{cases}$$

Thus,

$$\begin{aligned}\sum_{j\neq i}\mathsf{Cov}\big[b_j f(\widehat{\mathsf{c}}_j), b_i f(\widehat{c}_i)\big] &\le \sum_{j\neq i: C_j = C_i \neq \varnothing} p^2\left(\frac{q}{p}\right)^{\mathsf{c}_j} \\ &\overset{(15)}{\le} N(d-1)p^2\left[\left(\frac{q}{p}\right)^{\omega-1} \vee \frac{q}{p}\right]. \end{aligned} \tag{23}$$

Finally, combining (19), (21) and (23) yields the desired (11).

The high-probability bound (12) for $\widehat{\mathsf{cc}}$ follows from the concentration inequality in Lemma 10 in Appendix 6. To apply this result, note that $\widehat{\mathsf{cc}}$ is a sum of dependent random variables

$$\widehat{\mathsf{cc}} = \sum_{j\in[N]} Y_j, \tag{24}$$

where $Y_j = \frac{1}{p} b_j f(\widehat{\mathsf{c}}_j)$ satisfies $\mathbb{E}[Y_j] = 0$ for $\mathsf{c}_j > 0$ and $|Y_j| \le b \triangleq (\frac{1}{p})^\omega$ almost surely. Also, $S \triangleq \sum_{j\in[N]} \mathsf{Var}[Y_j] \le N(\frac{1}{p})^\omega$ by (20). To control the dependency between $\{Y_j\}_{j\in[N]}$, note that $\widehat{\mathsf{c}}_j = b_j \sum_{k: v_k\in C_j} b_k$. Thus, Y_j only depends on $\{b_k : k \in A_j\}$, where $A_j = \{v_j\} \cup C_j$. Define a dependency graph Γ, where $V(\Gamma) = [N]$ and

$$E(\Gamma) = \big\{\{i, j\} : i \neq j, A_i \cap A_j \neq \varnothing\big\}.$$

Then Γ has maximum degree bounded by $d\omega$, by Lemma 6. ☐

4.2.2. *Unbounded clique number*: *Smoothed estimators*

Up to this point, we have only considered unbiased estimators of the number of connected components. If the sample ratio p is at least $\frac{1}{2}$, Theorem 1 implies its variance is

$$\mathsf{Var}[\widehat{\mathsf{cc}}] \le N(d+1),$$

regardless of the clique number ω of the parent graph. However, if the clique number ω grows with N, for small sampling ratio p the coefficients of the unbiased estimator (9) are as large as $\frac{1}{p^\omega}$ which results in exponentially large variance. Therefore, in order to deal with graphs with large cliques, we must give up unbiasedness to achieve better bias-variance tradeoff. Using a technique known as *smoothing* introduced in [49], next we modify the unbiased estimator to achieve a good bias-variance tradeoff.

To this end, consider a discrete random variable $L \in \mathbb{N}$ independent of everything else. Define the following estimator by discarding those terms in (10) for which $\widetilde{\mathsf{c}}_j$ exceeds L, and then averaging over the distribution of L. In other words, let

$$\widehat{\mathsf{cc}}_L \triangleq \mathbb{E}_L\left[\frac{1}{p}\sum_{j=1}^{m}\left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_j}\mathbb{1}\{\widetilde{\mathsf{c}}_j \le L\}\right] = \frac{1}{p}\sum_{j=1}^{m}\left(-\frac{q}{p}\right)^{\widetilde{\mathsf{c}}_j}\mathbb{P}\big[L \ge \widetilde{\mathsf{c}}_j\big]. \tag{25}$$

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 by the maximum magnitude of the coefficients in (25). Like (9), (25) can be computed in linear time.

The next theorem bounds the mean-square error of $\widehat{\mathsf{cc}}_L$, which implies the minimax upper bound previously announced in Theorem 3. Its proof is somewhat technical and so we defer it to Appendix 6.

Theorem 7. *Let* $L \sim \mathrm{Poisson}(\lambda)$ *with* $\lambda = \frac{p}{2-3p}\log(\frac{Np}{1+d\omega})$. *If the maximum degree and clique number of* G *is at most* d *and* ω, *respectively, then when* $p < 1/2$,

$$\mathbb{E}_G\left|\widehat{\mathsf{cc}}_L - \mathsf{cc}(G)\right|^2 \leq 2N^2\left(\frac{Np}{1+d\omega}\right)^{-\frac{p}{2-3p}}.$$

4.3. Unions of cliques

If the parent graph G consists of disjoint union of cliques, so does the sampled graph $\widetilde{G}$. Counting cliques in each connected components, we can rewrite the estimator (9) as

$$\widehat{\mathsf{cc}} = \sum_{r\geq 1}\left(1-\left(-\frac{q}{p}\right)^r\right)\widetilde{\mathsf{cc}}_r = \mathsf{cc}(\widetilde{G}) - \sum_{r\geq 1}\left(-\frac{q}{p}\right)^r\widetilde{\mathsf{cc}}_r, \tag{26}$$

where $\widetilde{\mathsf{cc}}_r$ is the number of components in the sampled graph $\widetilde{G}$ that have r vertices. This coincides with the unbiased estimator proposed by Frank [22] for cliques, which is, in turn, based on the estimator of Goodman [28]. The following theorem, whose proof is given in Appendix 6, provides an upper bound on its variance, recovering the previous result in [22], Corollary 11.

Theorem 8. *Let* G *be a disjoint union of cliques with clique number at most* ω. *Then* $\widehat{\mathsf{cc}}$ *is an unbiased estimator of* $\mathsf{cc}(G)$ *and*

$$\mathbb{E}_G\left|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\right|^2 = \mathsf{Var}[\widehat{\mathsf{cc}}] = \sum_{r=1}^{N}\left(\frac{q}{p}\right)^r \mathsf{cc}_r \leq N\left(\left(\frac{q}{p}\right)^{\omega} \wedge \frac{q}{p}\right),$$

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

Theorem 8 implies that as long as we sample at least half of the vertices, that is, $p \geq \frac{1}{2}$, for any G consisting of disjoint cliques, the unbiased estimator (26) satisfies

$$\mathbb{E}_G\left|\widehat{\mathsf{cc}} - \mathsf{cc}(G)\right|^2 \leq N,$$

regardless of the clique size. However, if $p < 1/2$, the variance can be exponentially large in N. Next, we use the smoothing technique again to obtain a biased estimator with near-optimal performance. To this end, consider a discrete random variable $L \in \mathbb{N}$ and define the following estimator by truncating (26) at the random location L and average over its distribution:

$$\widetilde{\mathsf{cc}}_L \triangleq \mathsf{cc}(\widetilde{G}) - \mathbb{E}_L\left[\sum_{r=1}^{L}\left(-\frac{q}{p}\right)^r \widetilde{\mathsf{cc}}_r\right] = \mathsf{cc}(\widetilde{G}) - \sum_{r\geq 1}\left(-\frac{q}{p}\right)^r \mathbb{P}[L \geq r]\,\widetilde{\mathsf{cc}}_r. \tag{27}$$

The following result, proved in Appendix 6, bounds the mean squared error of $\widetilde{\mathsf{cc}}_L$ and, consequently, bounds the minimax risk in Theorem 4. It turns out that the smoothed estimator (27) with appropriately chosen parameters is nearly optimal. In fact, Theorem 9, whose proof is given in Appendix 6, gives an upper bound on the sampling complexity (see Table 1), which, in view of [60], Theorem 4, is seen to be optimal.

Theorem 9. *Let G be a disjoint union of cliques. Let $L \sim Pois(\lambda)$ with $\lambda = \frac{p}{2-3p}\log(N/4)$. If $p < 1/2$, then*

$$\mathbb{E}_G\left|\widetilde{\mathsf{cc}}_L - \mathsf{cc}(G)\right|^2 \leq N^2(N/4)^{-\frac{p}{2-3p}}.$$

Remark 1. *Alternatively, we could specialize the estimator $\widehat{\mathsf{cc}}_L$ in (25) that is designed for general chordal graphs to the case when G is a disjoint union of cliques; however, the analysis is less clean and the results are slightly weaker than Theorem 9.*

4.4. Non-chordal graphs

A general graph can always be made 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). Without loss of generality, triangulations do not affect the number of connected components, since the operation can be applied to each component.

In view of the various estimators and their performance guarantees developed so far for chordal graphs, a natural question to ask is how one might generalize those to non-chordal graphs. One heuristic is to first triangulate the subsampled graph and then apply the estimator such as (10) and (25) that are designed for chordal graphs. Suppose a triangulation operation commutes with subgraph sampling in distribution,[7] then the modified estimator would inherit all the performance guarantees proved for chordal graphs; unfortunately, this does not hold in general. Thus, so far our theory does not readily extend to non-chordal graphs. Nevertheless, the empirical performance of this heuristic estimator is competitive with $\widehat{\mathsf{cc}}$ in both performance (see Figure 10) and computational efficiency. Indeed, there are polynomial time algorithms that add at most $8k^2$

---

[7] By “commute in distribution” we mean the random graphs $\mathsf{TRI}(\widetilde{G})$ and $\widetilde{\mathsf{TRI}(G)}$ have the same distribution. That is, the triangulated sampled graph is statistically identical to a sampled version of a triangulation of the parent graph.

edges if at least k edges must be added to make the graph chordal [47].[8] In view of the theoretical guarantees in Theorem 6, it is better to be conservative with adding edges so as the maximal degree d and the clique number ω are kept small.

It should be noted that blindly applying estimators designed for chordal graphs to the subsampled non-chordal graph without triangulation leads to nonsensical estimates. Thus, preprocessing the graph appears to be necessary for producing good results. We will leave the task of rigorously establishing these heuristics for future work.

5. Lower bounds

5.1. General strategy

Next, we give a general lower bound for estimating additive graph properties (e.g. the number of connected components, subgraph counts) under the Bernoulli sampling model. The proof uses the method of two fuzzy hypotheses [57], Theorem 2.15, which, in the context of estimating graph properties, entails constructing a pair of random graphs whose properties have different average values, and the distributions of their subsampled versions are close in total variation, which is ensured by matching lower-order subgraph counts or sampling certain configurations on their vertices. The utility of this result is to use a pair of smaller graphs (which can be found in an ad hoc manner) to construct a bigger pair of graphs on N vertices and produce a lower bound that scales with N. The proof of Theorem 10 is furnished in Appendix 6.

Theorem 10. *Let f be a graph parameter that is invariant under isomorphisms and* additive *under disjoint union, that is, $f(G+H) = f(G) + f(H)$ [43], p. 41. Let $\mathcal{G}$ be a class of graphs with at most N vertices. Let m and $M = N/m$ be integers. Let H and H' be two graphs with m vertices. Assume that any disjoint union of the form $G_1 + \cdots + G_M$ is in $\mathcal{G}$ where G_i is either H or H'. Suppose $M \geq 300$ and $\mathrm{TV}(P, P') \leq 1/300$, where P (resp. P') denote the distribution of the isomorphism class of the sampled graph $\widetilde{H}$ (resp. $\widetilde{H}'$). Let $\widetilde{G}$ denote the sampled version of G under the Bernoulli sampling model with probability p. Then*

$$\inf_{\widehat{f}} \sup_{G \in \mathcal{G}} \mathbb{P}\left[\left|\widehat{f}(\widetilde{G}) - f(G)\right| \geq \Delta\right] \geq 0.01, \tag{28}$$

where $\Delta \triangleq \frac{|f(H)-f(H')|}{8}\left(\sqrt{\frac{N}{m\mathrm{TV}(P,P')}} \wedge \frac{N}{m}\right)$.

5.2. Bounding total variations between sampled graphs

The application of Theorem 10 relies on the construction of a pair of small graphs H and H' whose sampled versions are close in total variation. To this end, we provide two schemes to bound $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'})$ from above.

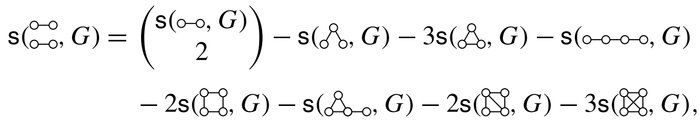
---

[8] An implementation of graph triangulation `R` is provided by the `is_chordal()` function in the package `igraph` [35].

5.2.1. *Matching subgraphs*

Since $\mathtt{cc}(G)$ is invariant with respect to isomorphisms, it suffices to describe the sampled graph $\widetilde{G}$ up to isomorphisms. It is well known that a graph G can be determined up to isomorphisms by its homomorphism numbers that count the number of ways to embed a smaller graph in G. Among various versions of graph homomorphism numbers (cf. [43], Section 5.2) the one that is most relevant to the present paper is $\mathsf{s}(H, G)$, which, as defined in Section 1, is the number of *vertex-induced* subgraphs of G that are isomorphic to H. Specifically, the relevance of induced subgraph counts to the subgraph sampling model is two-fold:

- The list of vertex-induced subgraph counts $\{\mathsf{s}(H, G) : v(H) \le N\}$ determines G up to isomorphism and hence constitutes a sufficient statistic for $\widetilde{G}$. In fact, it is further sufficient to summarize $\widetilde{G}$ into the list of numbers:[9] $\{\mathsf{s}(H, \widetilde{G}) : v(H) \le N, H \text{is connected}\}$, since the count of any disconnected subgraph is a fixed polynomial of connected subgraph counts. This is a well-known result in the theory of graph reconstruction [16,39,59]. For example, for any graph G, we have $\mathsf{s}(\circ\circ, G) = \binom{\mathsf{s}(\circ, G)}{2} - \mathsf{s}(\circ\!\!-\!\!\circ, G)$ and



  which can be obtained by counting pairs of vertices or edges in two different ways, respectively. See [45], Section 2, for more examples.
- Under the Bernoulli sampling model, the probabilistic law of the isomorphism class of the sampled graph is a polynomial in the sampling ratio p, with coefficients given by the induced subgraph counts. Indeed, recall from (2) that $\mathbb{P}[\widetilde{G} \simeq H] = \mathsf{s}(H, G)p^{\mathsf{v}(H)}(1 - p)^{\mathsf{v}(G)-\mathsf{v}(H)}$. Therefore two graphs with matching subgraph counts for all (connected) graphs of n vertices are statistically indistinguishable unless more than n vertices are sampled.

We begin with a refinement of the classical result that says disconnected subgraphs counts are fixed polynomials of connected subgraph counts. Below we provide a more quantitative version by showing that only those connected subgraphs which contain no more vertices than the disconnected subgraph involved. The proofs of the next set of results are given in Appendix 6.

Lemma 7. *Let H be a disconnected graph of v vertices. Then for any G, $\mathsf{s}(H, G)$ can be expressed as a polynomial, independent of G, in $\{\mathsf{s}(g, G) : g$ is connected and $v(g) \le v\}$.*

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

---

[9] This statistic cannot be further reduced because it is known that the connected subgraphs counts do not fulfill any predetermined relations in the sense that the closure of the range of their normalized version (subgraph densities) has nonempty interior [16].

Lemma 8. *Let* H *and* H' *be two graphs on* m *vertices. If*

$$\mathsf{s}(h, H) = \mathsf{s}\big(h, H'\big) \tag{29}$$

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

$$\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq \mathbb{P}\big[\mathrm{Bin}(m, p) \geq k+1\big] \leq \binom{m}{k+1} p^{k+1}. \tag{30}$$

Furthermore, if $p \leq (k+1)/m$, *then*

$$\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq \exp\left\{-\frac{2(k+1-pm)^2}{m}\right\}. \tag{31}$$

In Figure 6, we give an example of two graphs H and H' on 8 vertices that have matching counts of connected subgraphs with at most 4 vertices. Thus, by Lemma 8, they also have matching counts of *all* subgraphs with at most 4 vertices, and if $p \leq 5/8$, then $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq e^{-\frac{25}{4}(1-\frac{8p}{5})^2}$.

5.2.2. *Labeling-based coupling*

It is well known that for any probability distributions P and P', the total variation is given by $\mathrm{TV}(P, P') = \inf \mathbb{P}\big[X \neq X'\big]$, where the infimum is over all couplings, that is, joint distributions of X and X' that are marginally distributed as P and P', respectively. There is a natural coupling between the sampled graphs $\widetilde{H}$ and $\widetilde{H}'$ when we define the parent graph H and H' on the same set of labelled vertices. In some of the applications of Theorem 10, the constructions of H and H' are such that if certain configurations of the vertices are included or excluded in the sample, the resulting graphs are isomorphic. This property allows us to bound the total variation between the sampled graphs as follows.

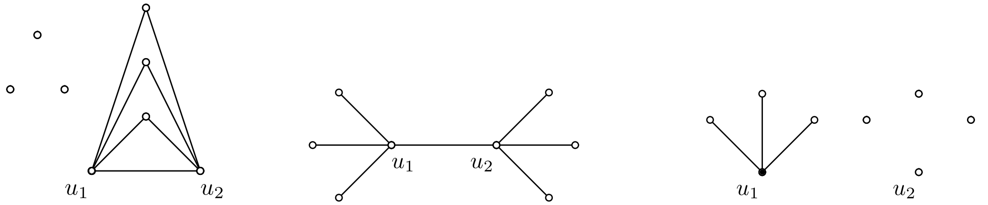
Lemma 9. *Let* H *and* H' *be graphs defined on the same set of vertices* V. *Let* U *be a subset of* V *and suppose that for any* $u \in U$, *we have* $H[V \setminus \{u\}] \simeq H'[V \setminus \{u\}]$. *Then, the total variation* $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'})$ *can be bounded by the probability that every vertex in* U *is sampled, viz.,*

$$\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq 1 - \mathbb{P}\big[\widetilde{H} \simeq \widetilde{H}'\big] \leq p^{|U|}.$$

If, in addition, $H[U] \simeq H'[U]$, *then the total variation* $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'})$ *can be bounded by the probability that every vertex in* U *is sampled and at least one vertex in* $V \setminus U$ *is sampled, viz.,*

$$\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p^{|U|}\big(1 - (1-p)\big)^{|V|-|U|}.$$

In Figure 4, we give an example of two graphs H and H' satisfying the assumption of Lemma 9. In this example, $|U| = 2$, and $|V| = 8$. Note that if any of the vertices in U are removed along with all their incident edges, then the resulting graphs are isomorphic. Also, since $H[U] \simeq H'[U]$, Lemma 9 implies that $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p^2(1-(1-p)^6)$.



(a) The graph H.

(b) The graph H'.

(c) The resulting graph when u_1 is sampled but not u_2.

Figure 4. Example where $U = \{u_1, u_2\}$ is an edge. If any of these vertices are not sampled and all incident edges are removed, the resulting graphs are isomorphic.

In the remainder of the section, we apply Theorem 10, Lemma 8, and Lemma 9 to derive lower bounds on the minimax risk for graphs that contain cycles and general chordal graphs, respectively. The main task is to handcraft a pair of graphs H and H' that either have matching counts of small subgraphs *or* for which certain configurations of their vertices induce subgraphs that are isomorphic.

5.3. Lower bound for chordal graphs

Theorem 11 (Chordal graphs). *Let $\mathcal{G}(N, d, \omega)$ denote the collection of all chordal graphs on N vertices with maximum degree and clique number at most d and $\omega \geq 2$, respectively. Assume that $p < \frac{1}{2^{\omega}100}$. Then*

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

Proof. There are two different constructions we give, according to whether $d \geq 2^{\omega}$ or $d < 2^{\omega}$.

Case I: $d \geq 2^{\omega}$

For every $\omega \geq 2$ and $m \in \mathbb{N}$, we construct a pair of graphs H and H', such that

$$v(H) = v\left(H'\right) = \omega - 1 + m2^{\omega-2} \tag{32}$$

$$d_{\max}(H) = d_{\max}\left(H'\right) = m2^{\omega-3} + \omega - 2, \quad \omega \geq 3 \tag{33}$$

$$d_{\max}(H) = 0, \qquad d_{\max}\left(H'\right) = m, \quad \omega = 2 \tag{34}$$

$$\mathsf{cc}(H) = m + 1, \qquad \mathsf{cc}\left(H'\right) = 1 \tag{35}$$

$$\left|\mathsf{s}(K_{\omega}, H) - \mathsf{s}\left(K_{\omega}, H'\right)\right| = m \tag{36}$$

Fix a set of $\omega - 1$ vertices U that forms a clique. We first construct H. For every subset $S \subset U$ such that $|S|$ is even, let V_S be a set of m distinct vertices such that the neighborhood of every

$v \in V_S$ is given by $\partial v = S$. Let the vertex set $V(H)$ be the union of U and all V_S such that $|S|$ is even. In particular, because of the presence of $S = \varnothing$, H always has exactly m isolated vertices (unless $\omega = 2$, in which case H consists of $m + 1$ isolated vertices). Repeat the same construction for H' with $|S|$ being odd. Then both H are H' are chordal and have the same number of vertices as in (32), since

$$\begin{aligned} v(H) &= \omega - 1 + m \sum_{0 \le i \le \omega-1,\ i \text{ even}} \binom{\omega - 1}{i} = v(H') \\ &= \omega - 1 + m \sum_{0 \le i \le \omega-1,\ i \text{ odd}} \binom{\omega - 1}{i} \end{aligned}$$

which follows from the binomial summation formula. Similarly, (33)–(36) can be readily verified.

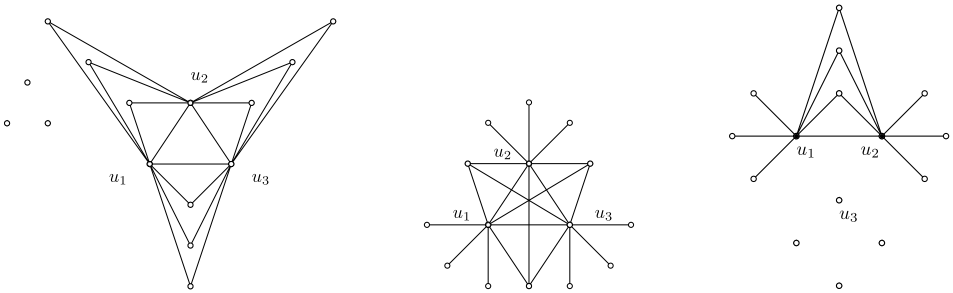
We also have that

$$\begin{aligned} \mathsf{s}(K_i, H) &= \binom{\omega - 1}{i} + m \sum_{0 \le j \le \omega-1, j \text{ even}} \binom{\omega - 1}{j} \binom{j}{i-1} \\ &= \mathsf{s}(K_i, H') = \binom{\omega - 1}{i} + m \sum_{0 \le j \le \omega-1, j \text{ odd}} \binom{\omega - 1}{j} \binom{j}{i-1} \\ &= \binom{\omega - 1}{i} + m \binom{\omega - 1}{i-1} 2^{\omega-1-i}, \end{aligned}$$

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, we first assume that H and H' are defined on the same set of labelled vertices V. The key observation is the following: by construction, $H[U] \simeq H'[U]$ (since U induces a clique) and, furthermore, failing to sample any vertex in U results in an isomorphic graph, i.e., $H[V \setminus \{u\}] \simeq H'[V \setminus \{u\}]$ for any $u \in U$. Indeed, the structure of the induced subgraph $H[V \setminus \{u\}]$ can be described as follows. First, let U form a clique. Next, for every nonempty subset $S \subset U \setminus \{u\}$, attach a set of m distinct vertices (denoted by V_S) so that the neighborhood of every $v \in V_S$ is given by $\partial v = S$. Finally, add $m + 1$ isolated vertices. See Figure 4 ($\omega = 3$) and Figure 5 ($\omega = 4$) for illustrations of this property and the iterative nature of this construction, in the sense that the construction of H (resp. H') for $\omega = k + 1$ can be obtained from the construction of H (resp. H') for $\omega = k$ by adding another vertex u to U such that $\partial u = U$ and then adjoining m distinct vertices to every even (resp. odd) cardinality set $S \subset U$ containing u.

Thus by Lemma 9, $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \le p^{|U|}\left(1 - (1 - p)^{|V|-|U|}\right) = p^{\omega-1}(1 - (1 - p)^{m2^{\omega-2}})$. According to (33), we choose $m = \lfloor (d - \omega + 2)2^{-\omega+3} \rfloor \ge d2^{-\omega+2}$ if $\omega \ge 3$ and $m = d$ if $\omega = 2$. Then we have, $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) = p^{\omega-1}(1 - (1 - p)^d) \le p^{\omega-1}(pd \wedge 1)$. The condition on p ensures



(a) The graph H.

(b) The graph H'.

(c) The resulting graph when u_1 and u_2 are sampled but not u_3.

Figure 5. Example for $\omega = 4$ and $m = 3$, where $U = \{u_1, u_2, u_3\}$ form a triangle. If any one or two (as shown in the figure) of these vertices are not sampled and all incident edges are removed, the resulting graphs are isomorphic.

that $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \le p < 1/300$. In view of Theorem 10 and (35), we have

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

provided $d \ge 2^\omega$.

Case II: $d \le 2^\omega$

In this case, the previous construction is no longer feasible and we must construct another pair of graphs with a smaller maximum degree. To this end, we consider graphs H and H' consisting of disjoint cliques of size at most $\omega \ge 2$, such that

$$\begin{aligned} v(H) = v\bigl(H'\bigr) &= \omega 2^{\omega-2}, \\ d_{\max}(H) = d_{\max}\bigl(H'\bigr) &= \omega - 1, \\ \bigl|\mathsf{cc}(H) - \mathsf{cc}\bigl(H'\bigr)\bigr| &= 1. \end{aligned} \tag{37}$$

If ω is odd, we set

$$\begin{aligned} H &= \binom{\omega}{\omega} K_\omega + \binom{\omega}{\omega-2} K_{\omega-2} + \cdots + \binom{\omega}{3} K_3 + \binom{\omega}{1} K_1 \\ H' &= \binom{\omega}{\omega-1} K_{\omega-1} + \binom{\omega}{\omega-3} K_{\omega-3} + \cdots + \binom{\omega}{4} K_4 + \binom{\omega}{2} K_2. \end{aligned} \tag{38}$$

If ω is even, we set

$$\begin{aligned} H &= \binom{\omega}{\omega} K_\omega + \binom{\omega}{\omega-2} K_{\omega-2} + \cdots + \binom{\omega}{4} K_4 + \binom{\omega}{2} K_2 \\ H' &= \binom{\omega}{\omega-1} K_{\omega-1} + \binom{\omega}{\omega-3} K_{\omega-3} + \cdots + \binom{\omega}{3} K_3 + \binom{\omega}{1} K_1. \end{aligned} \tag{39}$$

For example, for $\omega = 3$, (38) becomes $H =$ $+ 3 \times \circ$ and $H' = 3 \times$ ∘–∘ ; for $\omega = 4$, (39) becomes $H =$ $+ 6 \times$ ∘–∘ and $H' = 4 \times$ $+ 4 \times \circ$.

Next we verify that H and H' have matching subgraph counts. Indeed, for $i = 1, 2, \ldots, \omega - 1$, $\mathsf{s}(K_i, H) - \mathsf{s}(K_i, H') = \sum_{k=i}^{\omega} (-1)^k \binom{\omega}{k}\binom{k}{i} = 0$ and $\mathsf{s}(K_i, H) = \mathsf{s}(K_i, H') = \frac{1}{2}\sum_{k=i}^{\omega} \binom{\omega}{k}\binom{k}{i} = 2^{\omega-1-i}\binom{\omega}{i}$. Hence, H and H' contain matching number of cliques up to size $\omega - 1$. Note that the only connected induced subgraphs of H and H' with at most $\omega - 1$ vertices are cliques. Consequently, by (30), $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq \binom{\omega 2^{\omega-2}}{\omega} p^\omega$ and together with Theorem 10 and (37), we have

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

where the last inequality follows from the current assumption that $d \leq 2^\omega$. The condition on p ensures that $\mathrm{TV}(P_{\widetilde{H}}, P_{\widetilde{H}'}) \leq p 2^{\omega-2} < 1/300$. □

Acknowledgment

The authors are grateful to Ben Rossman for fruitful discussions and Richard Stanley for helpful comments on (5).

This research was supported in part by the NSF Grant IIS-1447879, CCF-1527105, an NSF CAREER award CCF-1651588, and an Alfred Sloan fellowship.

Supplementary Material

Supplement to "Estimating the number of connected components in a graph via subgraph sampling" (DOI: 10.3150/19-BEJ1147SUPP; .pdf). Proofs of supporting lemmas and additional results and numerical experiments.

References

[1] Aliakbarpour, M., Shankha Biswas, A., Gouleakis, T., Peebles, J., Rubinfeld, R. and Yodpinyanee, A. (2018). Sublinear-time algorithms for counting star subgraphs via edge sampling. *Algorithmica* **80** 668–697. MR3757567 https://doi.org/10.1007/s00453-017-0287-3

[2] Bandiera, O. and Rasul, I. (2006). Social networks and technology adoption in northern Mozambique. *Econ. J.* **116** 869–902.

[3] Ben-Hamou, A., Oliveira, R.I. and Peres, Y. (2018). Estimating graph parameters via random walks with restarts. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* 1702–1714. Philadelphia, PA: SIAM. MR3775899 https://doi.org/10.1137/1.9781611975031.111

[4] Berenbrink, P., Krayenhoff, B. and Mallmann-Trenn, F. (2014). Estimating the number of connected components in sublinear time. *Inform. Process. Lett.* **114** 639–642. MR3230913 https://doi.org/10.1016/j.ipl.2014.05.008

[5] Abramowitz, M. and Stegun, I.A. (eds) (1992). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover.

[6] Borgs, C., Chayes, J.T., Lovász, L., Sós, V.T. and Vesztergombi, K. (2008). Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing. *Adv. Math.* **219** 1801–1851. MR2455626 https://doi.org/10.1016/j.aim.2008.07.008

[7] Capobianco, M. (1972). Estimating the connectivity of a graph. In *Graph Theory and Applications* (*Proc. Conf., Western Michigan Univ., Kalamazoo, Mich.*, 1972; *Dedicated to the Memory of J. W. T. Youngs*) *Lecture Notes in Math.* **303** 65–74. MR0332542

[8] Chandrasekhar, A. and Lewis, R. (2011). Econometrics of sampled networks. Unpublished manuscript.

[9] Chazelle, B., Rubinfeld, R. and Trevisan, L. (2005). Approximating the minimum spanning tree weight in sublinear time. *SIAM J. Comput.* **34** 1370–1379. MR2165745 https://doi.org/10.1137/S0097539702403244

[10] Chen, B., Shrivastava, A. and Steorts, R.C. (2018). Unique entity estimation with application to the Syrian conflict. *Ann. Appl. Stat.* **12** 1039–1067. MR3834294 https://doi.org/10.1214/18-AOAS1163

[11] Conley, T.G. and Udry, C.R. (2010). Learning about a new technology: Pineapple in Ghana. *Am. Econ. Rev.* **100** 35–69.

[12] Apicella, C.L., Marlowe, F.W., Fowler, J.H. and Christakis, N.A. (2012). Social networks and cooperation in hunter-gatherers. *Nature* **481** 497–501.

[13] Cormode, G. and Duffield, N. (2014). Sampling for big data: A tutorial. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1975–1975. ACM.

[14] Dohmen, K. (2013). Lower bounds for the probability of a union via chordal graphs. *Electron. Commun. Probab.* **18** no. 70, 4. MR3101635 https://doi.org/10.1214/ECP.v18-2357

[15] Eden, T., Levi, A., Ron, D. and Seshadhri, C. (2015). Approximately counting triangles in sublinear time. In 2015 *IEEE 56th Annual Symposium on Foundations of Computer Science – FOCS 2015* 614–633. Los Alamitos, CA: IEEE Computer Soc. MR3473331

[16] Erdős, P., Lovász, L. and Spencer, J. (1979). Strong independence of graphcopy functions. In *Graph Theory and Related Topics* (*Proc. Conf., Univ. Waterloo, Waterloo, Ont.*, 1977) 165–172. New York: Academic Press. MR0538044

[17] Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Magy. Tud. Akad. Mat. Kut. Intéz. Közl.* **5** 17–61. MR0125031

[18] Fafchamps, M. and Lund, S. (2003). Risk-sharing networks in rural Philippines. *J. Dev. Econ.* **71** 261–287.

[19] Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 631–636. New York: ACM.

[20] Feigenberg, B., Field, E.M. and Pande, R. (2010). Building social capital through microfinance Technical report, National Bureau of Economic Research.

[21] Frank, O. (1977). Estimation of graph totals. *Scand. J. Stat.* **4** 81–89. MR0458659

[22] Frank, O. (1978). Estimation of the number of connected components in a graph by using a sampled subgraph. *Scand. J. Stat.* **5** 177–188. MR0515656

[23] Gao, C., Lu, Y. and Zhou, H.H. (2015). Rate-optimal graphon estimation. *Ann. Statist.* **43** 2624–2652. MR3405606 https://doi.org/10.1214/15-AOS1354

[24] Goldreich, O. (2017). *Introduction to Property Testing*. Cambridge: Cambridge Univ. Press. MR3837126 https://doi.org/10.1017/9781108135252

[25] Goldreich, O., Goldwasser, S. and Ron, D. (1998). Property testing and its connection to learning and approximation. *J. ACM* **45** 653–750. MR1675099 https://doi.org/10.1145/285055.285060

[26] Goldreich, O. and Ron, D. (2008). Approximating average parameters of graphs. *Random Structures Algorithms* **32** 473–493. MR2422391 https://doi.org/10.1002/rsa.20203

[27] Goldreich, O. and Ron, D. (2011). On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography*. *Lecture Notes in Computer Science* **6650** 68–75. Heidelberg: Springer. MR2844253 https://doi.org/10.1007/978-3-642-22670-0_9

[28] Goodman, L.A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Stat.* **20** 572–579. MR0032165 https://doi.org/10.1214/aoms/1177729949

[29] Goodman, L.A. (1961). Snowball sampling. *Ann. Math. Stat.* **32** 148–170. MR0124140 https://doi.org/10.1214/aoms/1177705148

[30] Govindan, R. and Tangmunarunkit, H. (2000). Heuristics for Internet map discovery. In *INFOCOM* 2000. *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies*. *Proceedings*. *IEEE* **3** 1371–1380. IEEE.

[31] Handcock, M.S. and Gile, K.J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082 https://doi.org/10.1214/08-AOAS221

[32] Holland, P.W., Laskey, K.B. and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[33] Holland, P.W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *J. Amer. Statist. Assoc.* **76** 33–65. MR0608176

[34] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460

[35] igraph: Network analysis and visualization, 2019. https://cran.r-project.org/web/packages/igraph.

[36] Janson, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures Algorithms* **24** 234–248. MR2068873 https://doi.org/10.1002/rsa.20008

[37] Klusowski, J.M. and Wu, Y. (2018). Counting motifs with graph sampling. In *Proceedings of the* 31*st Conference on Learning Theory* (S. Bubeck, V. Perchet and P. Rigollet, eds.). *Proceedings of Machine Learning Research* **75** 1966–2011. PMLR, 06–09, 2018.

[38] Klusowski, J.M. and Wu, Y. (2020). Supplement to "Estimating the number of connected components in a graph via subgraph sampling." https://doi.org/10.3150/19-BEJ1147SUPP.

[39] Kocay, W.L. (1982). Some new methods in reconstruction theory. In *Combinatorial Mathematics*, *IX* (*Brisbane*, 1981). *Lecture Notes in Math.* **952** 89–114. Berlin: Springer. MR0674132

[40] Kolaczyk, E.D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. *Springer Series in Statistics*. New York: Springer. MR2724362 https://doi.org/10.1007/978-0-387-88146-1

[41] Kolaczyk, E.D. (2017). *Topics at the Frontier of Statistics and Network Analysis*. *SemStat Elements*. Cambridge: Cambridge Univ. Press. (Re)visiting the foundations. MR3702038 https://doi.org/10.1017/9781108290159

[42] Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data/ca-CondMat.html.

[43] Lovász, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Providence, RI: Amer. Math. Soc. MR3012035 https://doi.org/10.1090/coll/060

[44] Luce, R.D. and Perry, A.D. (1949). A method of matrix analysis of group structure. *Psychometrika* **14** 95–116. MR0035974 https://doi.org/10.1007/BF02289146

[45] McKay, B.D. and Radziszowski, S.P. (1997). Subgraph counting identities and Ramsey numbers. *J. Combin. Theory Ser. B* **69** 193–209. MR1438619 https://doi.org/10.1006/jctb.1996.1741

[46] McMahon, E.W., Shimkus, B.A. and Wolfson, J.A. (2003). Chordal graphs and the characteristic polynomial. *Discrete Math.* **262** 211–219. MR1951389 https://doi.org/10.1016/S0012-365X(02)00500-9

[47] Natanzon, A., Shamir, R. and Sharan, R. (2000). A polynomial approximation algorithm for the minimum fill-in problem. *SIAM J. Comput.* **30** 1067–1079. MR1786752 https://doi.org/10.1137/S0097539798336073

[48] O'Donnell, R. (2014). *Analysis of Boolean Functions*. New York: Cambridge Univ. Press. MR3443800 https://doi.org/10.1017/CBO9781139814782

[49] Orlitsky, A., Suresh, A.T. and Wu, Y. (2016). Optimal prediction of the number of unseen species. *Proc. Natl. Acad. Sci. USA* **113** 13283–13288. MR3582444 https://doi.org/10.1073/pnas.1607774113

[50] Polyanskiy, Y., Theertha Suresh, A. and Wu, Y. (2017). Sample complexity of population recovery. In *Proceedings of Conference on Learning Theory* (*COLT*). Amsterdam, Netherland. Available at arXiv:1702.05574.

[51] Reingen, P.H. and Kernan, J.B. (1986). Analysis of referral networks in marketing: Methods and illustration. *J. Mark. Res.* 370–378.

[52] Rose, D.J., Tarjan, R.E. and Lueker, G.S. (1976). Algorithmic aspects of vertex elimination on graphs. *SIAM J. Comput.* **5** 266–283. MR0408312 https://doi.org/10.1137/0205021

[53] Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D.S., Zhang, L.V., Wong, S.L., Franklin, G., Li, S., Albala, J.S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R.S., Vandenhaute, J., Zoghbi, H.Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M.E., Hill, D.E., Roth, F.P. and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437** 1173–1178. https://doi.org/10.1038/nature04209

[54] Salganik, M.J. and Heckathorn, D.D. (2004). Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Method.* **34** 193–240.

[55] Stumpf, M.P.H., Thorne, T., de Silva, E., Stewart, R., Jun An, H., Lappe, M. and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105** 6959–6964.

[56] Tarjan, R.E. and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. *SIAM J. Comput.* **13** 566–579. MR0749707 https://doi.org/10.1137/0213035

[57] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation*. *Springer Series in Statistics*. New York: Springer. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats. MR2724359 https://doi.org/10.1007/b13794

[58] West, D.B. (1996). *Introduction to Graph Theory*. Upper Saddle River, NJ: Prentice Hall, Inc. MR1367739

[59] Whitney, H. (1932). The coloring of graphs. *Ann. of Math.* (2) **33** 688–718. MR1503085 https://doi.org/10.2307/1968214

[60] Wu, Y. and Yang, P. (2016). Sample complexity of the distinct element problem. Preprint. Available at arXiv:1612.03375.

Received August 2018 and revised June 2019