The Annals of Statistics
2022, Vol. 50, No. 2, 987–1011
https://doi.org/10.1214/21-AOS2134

MOTIF ESTIMATION VIA SUBGRAPH SAMPLING: THE FOURTH-MOMENT PHENOMENON

BY BHASWAR B. BHATTACHARYA[1,a], SAYAN DAS[2,b] AND SUMIT MUKHERJEE[3,c]

[1]*Department of Statistics, University of Pennsylvania,* [a]*bhaswar@wharton.upenn.edu*
[2]*Department of Mathematics, Columbia University,* [b]*sayan.das@columbia.edu*
[3]*Department of Statistics, Columbia University,* [c]*sm3949@columbia.edu*

Network sampling is an indispensable tool for understanding features of large complex networks where it is practically impossible to search over the entire graph. In this paper, we develop a framework for statistical inference for counting network motifs, such as edges, triangles and wedges, in the widely used subgraph sampling model, where each vertex is sampled independently, and the subgraph induced by the sampled vertices is observed. We derive necessary and sufficient conditions for the consistency and the asymptotic normality of the natural Horvitz–Thompson (HT) estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts based on the sampled graph. In particular, we show that the asymptotic normality of the HT estimator exhibits an interesting fourth-moment phenomenon, which asserts that the HT estimator (appropriately centered and rescaled) converges in distribution to the standard normal whenever its fourth-moment converges to 3 (the fourth-moment of the standard normal distribution). As a consequence, we derive the exact thresholds for consistency and asymptotic normality of the HT estimator in various natural graph ensembles, such as sparse graphs with bounded degree, Erdős–Rényi random graphs, random regular graphs and dense graphons.

1. Introduction. One of the main challenges in network analysis is that the observed network is often a sample from a much larger (parent) network. This is generally due to the massive size of the network or the inability to access parts of the network, making it practically impossible to search/query over the entire graph. The central statistical question in such studies is to estimate global features of the parent network, that accounts for the bias and variability induced by the sampling paradigm. The study of network sampling began with the results of Frank [22, 23] and Capobianco [13], where methods for estimating features such as connected components and graph totals were studied (see [49] for a more recent survey of these results). Network sampling has since then emerged as an essential tool for estimating features of large complex networks, with applications in social networks [32, 41, 58], protein interaction networks [53, 57], internet and communication networks [29] and socioeconomic networks [3, 4] (see [18, 38, 39] for a detailed discussion of different network sampling techniques and their applications).

Counting motifs (patterns of subgraphs) [44, 50] in a large network, which encode important structural information about the geometry of the network, is an important statistical and computational problem. In this direction, various sublinear time algorithms based on edge and degree queries have been proposed for testing and estimating properties such as the average degree [21, 25], triangles [7, 20], stars [2], general subgraph counts [27] and expansion properties [26]. These results are, however, all based on certain adaptive queries, which are

Received November 2020; revised August 2021.
MSC2020 subject classifications. 62G05, 62G20, 05C30.
Key words and phrases. Fourth moment phenomenon, Motif counting, network analysis, asymptotic inference, Stein's method, random graphs.

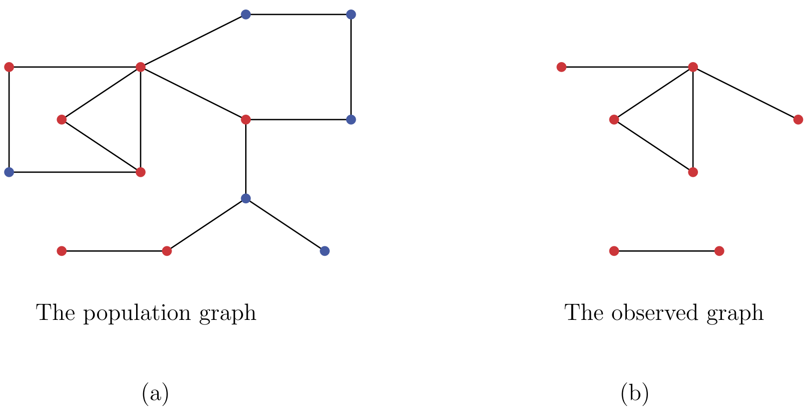

FIG. 1. *The subgraph sampling scheme*: (*a*) *The population graph and the vertices sampled* (*colored in red*), *and* (*b*) *the observed graph.*

unrealistic in applications where the goal is to estimate features of the network based on a single sampled graph [3, 14]. In this framework, estimating features such as the degree distribution [59], the number of connected components [37] and the number of motifs [36] have been studied recently, under various sampling schemes and structural assumptions on the parent graph.

In this paper we consider the problem of motif estimation, that is, counting the number of copies of a fixed graph $H = (V(H), E(H))$ (for example, edges, triangles, and wedges) in a large parent graph G_n in the most popular and commonly used subgraph sampling model, where each vertex of G_n is sampled independently with probability $p_n \in (0, 1)$ and the subgraph induced by these sampled vertices is observed. Here, the natural Horvitz–Thompson (HT) estimator obtained by weighting the number of copies of H in the observed network by $p_n^{-|V(H)|}$ (the inverse probability of sampling a subset of size $|V(H)|$ in the graph G_n) is unbiased for the true motif count. Very recently, Klusowski and Yu [36] showed that the HT estimator (for induced subgraph counts) is minimax rate optimal in the subgraph sampling model for classes of graphs with maximum degree constraints. Given this result, it becomes imperative to develop a framework for statistical inference for the motif counts in the subgraph sampling model. In this paper we derive precise conditions for the consistency and the asymptotic normality of the HT estimator, which can be used for constructing confidence intervals and hypothesis testing for the motif counts in the subgraph sampling model. The results give a complete characterization of the asymptotics of the HT estimator, thus providing a mathematical framework for evaluating its performance in different examples. We begin by formally describing the subgraph sampling model and the motif estimation problem in Section 1.1. A summary of the results obtained is given in Section 1.2.

1.1. *The subgraph sampling model.* Suppose $G_n = (V(G_n), E(G_n))$ is a simple, labeled and undirected graph with vertex set $V(G_n) = \{1, 2, \dots, |V(G_n)|\}$ and edge set $E(G_n)$. We denote by $A(G_n) = ((a_{ij}))_{i,j \in V(G_n)}$ the adjacency matrix of G_n, that is, $a_{ij} = 1$ whenever there is an edge between (i, j) and zero otherwise. In the *subgraph sampling model*, each vertex of the graph G_n is sampled independently with probability $p_n \in (0, 1)$, and we observe the subgraph induced by the sampled vertices. The parameter p_n is referred to as the *sampling ratio* of the graph G_n. In the survey sampling literature, this sampling scheme is also referred to as the Poisson sampling plan (see Tillé [56] and the references therein). The sampling scheme is illustrated in Figure 1, where the population graph and the vertices sampled (colored in red) are shown in the left and the observed graph is shown in the right.

Having observed this sampled subgraph, our goal is to estimate the number of copies of a fixed connected graph $H = (V(H), E(H))$ in the parent graph G_n. Formally, the number of copies of H in G_n is given by

$$N(H, G_n) := \frac{1}{|\mathrm{Aut}(H)|} \sum_{s \in V(G_n)_{|V(H)|}} \prod_{(i,j) \in E(H)} a_{s_i s_j}, \tag{1.1}$$

- $V(G_n)_{|V(H)|}$ is the set of all $|V(H)|$-tuples $s = (s_1, \dots, s_{|V(H)|}) \in V(G_n)^{|V(H)|}$ with distinct indices.[1] Thus, the cardinality of $V(G_n)_{|V(H)|}$ is $\frac{|V(G_n)|!}{(|V(G_n)|-|V(H)|)!}$.
- $\mathrm{Aut}(H)$ is the *automorphism group* of H, that is, the number permutations σ of the vertex set $V(H)$ such that $(x, y) \in E(H)$ if and only if $(\sigma(x), \sigma(y)) \in E(H)$.

Let X_v be the indicator of the event that the vertex $v \in V(G_n)$ is sampled under subgraph sampling model. Note that $\{X_v\}_{v \in V(G_n)}$ is a collection of i.i.d. $\mathrm{Ber}(p_n)$ variables. For $s \in V(G_n)_{|V(H)|}$, denote

$$X_s := X_{s_1} X_{s_2} \dots X_{s_{|V(H)|}} := \prod_{u=1}^{|V(H)|} X_{s_u} \quad \text{and} \quad M_H(s) := \prod_{(i,j) \in E(H)} a_{s_i s_j}.$$

Then the number of copies of H in the sampled subgraph is given by

$$T(H, G_n) := \frac{1}{|\mathrm{Aut}(H)|} \sum_{s \in V(G_n)_{|V(H)|}} M_H(s) X_s. \tag{1.2}$$

Note that $\mathbb{E}[T(H, G_n)] = p_n^{|V(H)|} N(H, G_n)$, hence

$$\widehat{N}(H, G_n) := \frac{1}{p_n^{|V(H)|}} T(H, G_n) \tag{1.3}$$

is a natural unbiased estimator for the parameter $N(H, G_n)$. This is referred to in the literature as the Horvitz–Thompson (HT) estimator of the motif count $N(H, G_n)$ [36], since it uses inverse probability weighting to achieve unbiasedness [33].

1.2. *Summary of results.* In this paper, we develop a framework for statistical inference for the motif counts using the HT estimator in the subgraph sampling model. The following is a summary of the results obtained:

- To begin with, we establish a necessary and sufficient condition for the consistency of the HT estimator, that is, conditions under which $\widehat{N}(H, G_n)/N(H, G_n)$ converges to 1 in probability. To this end, we introduce the notion of local count function, which counts the number of copies of H incident on a fixed subset of vertices, and show that the precise condition for the consistency of the HT estimator is to ensure that subsets of vertices with "high" local counts have asymptotically negligible contribution to the total count $N(H, G_n)$ (Theorem 2.1).
- To derive the asymptotic normality of the HT estimator, we consider the rescaled statistic

$$Z(H, G_n) := \frac{\widehat{N}(H, G_n) - N(H, G_n)}{\sqrt{\mathrm{Var}[\widehat{N}(H, G_n)]}}. \tag{1.4}$$

[1] For a set S, the set S^N denotes the N-fold Cartesian product $S \times S \times \cdots \times S$.

Using the Stein's method for normal approximation, we derive an explicit rate of convergence (in the Wasserstein's distance) between $Z(H, G_n)$ and the standard normal distribution. As a consequence, we show that $Z(H, G_n) \overset{D}{\to} N(0, 1)$, whenever the fourth moment $\mathbb{E}[Z(H, G_n)] \to 3$ (the fourth moment of $N(0, 1)$) (see Theorem 2.3 for details). This is an example of the celebrated *fourth-moment phenomenon*, which initially appeared in the asymptotics of multiple stochastic integrals (Wiener chaos) in the seminal papers [45, 48] and has, since then, emerged as the driving condition for the asymptotic normality of various nonlinear functions of random fields [46]. In the present context of motif estimation, we show that the asymptotic normality of $Z(H, G_n)$ is a consequence of a more general central limit theorem (CLT) for random multilinear forms in Bernoulli variables, a result which might be of independent interest (Theorem A.3).

- Next, we discuss how the CLT for $Z(H, G_n)$ can be used to compute a confidence interval for the motif count $N(H, G_n)$. Toward this, we provide an unbiased estimate of the variance of $Z(H, G_n)$ that is consistent whenever the CLT error term for $Z(H, G_n)$ goes to zero, which can be used to construct an asymptotically valid confidence interval for $N(H, G_n)$ (Proposition 2.4).
- We then derive a necessary and sufficient condition for the asymptotic normality of $Z(H, G_n)$. For this, we need to weaken the fourth-moment condition $\mathbb{E}[Z(H, G_n)^4] \to 3$, which although sufficient, is not always necessary for the asymptotic normality of $Z(H, G_n)$. In particular, there are graph sequences for which $Z(H, G_n) \overset{D}{\to} N(0, 1)$, even though the fourth-moment condition fails (Example D.4). Instead, we show that the asymptotic normality of $Z(H, G_n)$ is characterized by a *truncated fourth-moment* condition. More precisely, $Z(H, G_n)$ converges in distribution to $N(0, 1)$ if and only if the second and fourth moments of an appropriate truncation of $Z(H, G_n)$, based on the local count functions, converges to 1 and 3, respectively (Theorem 2.5).
- As a consequence of the above results, we derive the exact thresholds for consistency and asymptotic normality of the HT estimator in various natural graph ensembles, such as sparse graphs with bounded degree (Proposition 2.6), Erdős–Rényi random graphs (Theorem 2.8), random regular graphs (Corollary 2.10) and graphons (Proposition 2.12). In each of these cases, there is a threshold (which depends on the graph parameters) such that if the sampling ratio p_n is much larger than this threshold, then the HT estimator is consistent and asymptotically normal, whereas if p_n is of the same order as the threshold, the HT estimator is neither consistent nor asymptotic normal. In particular, for the Erdős–Rényi graph, the threshold for consistency and asymptotic normality depends on the well-known balancedness coefficient of the graph H (Definition 2.7), and is related to the threshold for the occurrence of H is the sampled random graph.

These results provide a comprehensive characterization of the asymptotics of the HT estimator for the motif counts in the subgraph sampling model, which can be used to validate its performance in various applications. The formal statements of the results and their various consequences are given below in Section 2.

1.3. *Asymptotic notation.* Throughout we will use the following standard asymptotic notations. For two positive sequences, $\{a_n\}_{n\geq 1}$ and $\{b_n\}_{n\geq 1}$, $a_n = O(b_n)$ means $a_n \leq C_1 b_n$, $a_n = \Omega(b_n)$ means $a_n \geq C_2 b_n$, and $a_n = \Theta(b_n)$ means $C_2 b_n \leq a_n \leq C_1 b_n$, for all n large enough and positive constants C_1, C_2. Similarly, $a_n \lesssim b_n$ means $a_n = O(b_n)$, and $a_n \gtrsim b_n$ means $a_n = \Omega(b_n)$, and subscripts in the above notation, for example, $\lesssim_\square$ or $\gtrsim_\square$, denote that the hidden constants may depend on the subscripted parameters. Moreover, $a_n \ll b_n$ means $a_n = o(b_n)$, and $a_n \gg b_n$ means $b_n = o(a_n)$. Finally, for a sequence of random variables $\{X_n\}_{n\geq 1}$ and a positive sequence $\{a_n\}_{n\geq 1}$, the notation $X_n = O_P(a_n)$ means X_n/a_n is stochastically bounded, that is, $\lim_{M\to\infty} \lim_{n\to\infty} \mathbb{P}(|X_n/a_n| \leq M) = 1$, and $X_n = \Theta_P(a_n)$ will mean $X_n = O_P(a_n)$ and $\lim_{\delta\to 0} \lim_{n\to\infty} \mathbb{P}(|X_n/a_n| \geq \delta) = 1$.

2. Statements of the main results. In this section, we state our main results. Throughout we will assume that there exists $\kappa \in (0, 1)$ such that

$$p_n \leq 1 - \kappa, \tag{2.1}$$

for all $n \geq 1$. This is to rule out the degenerate case when we observe nearly the whole graph, in which case the estimation problem becomes trivial. The rest of this section is organized as follows: The necessary and sufficient condition for the consistency of the HT estimator is discussed in Section 2.1. The precise conditions for the asymptotic normality of the HT estimator and construction of confidence intervals are given in Section 2.2. Finally, in Section 2.3 we compute the thresholds for consistency and asymptotic normality for various graph ensembles.

2.1. *Consistency of the HT estimator.* In this section, we obtain the precise conditions for consistency of the HT estimator $\widehat{N}(H, G_n)$, for any fixed connected motif H and any sequence of graphs $\{G_n\}_{n\geq 1}$, such that $N(H, G_n) > 0$ for all $n \geq 1$. To state our results precisely, we need a few definitions. For an ordered tuple $\boldsymbol{s} \in V(G_n)_{|V(H)|}$ with distinct entries, denote by $\overline{\boldsymbol{s}}$ the (unordered) set formed by the entries of $\boldsymbol{s}$ (e.g., if $\boldsymbol{s} = (4, 2, 5)$, then $\overline{\boldsymbol{s}} = \{2, 4, 5\}$). For any nonempty set $A \subset V(G_n)$ with $1 \leq |A| \leq |V(H)|$, define the *local count function* of H on the set A as follows:

$$t_H(A) := \frac{1}{|\operatorname{Aut}(H)|} \sum_{\boldsymbol{s} \in V(G_n)_{|V(H)|} : \overline{\boldsymbol{s}} \supseteq A} M_H(\boldsymbol{s}), \tag{2.2}$$

where the sum is over of all ordered $\boldsymbol{s} \in V(G_n)_{|V(H)|}$ such that the set $\overline{\boldsymbol{s}}$ contains all the elements of A. In other words, $t_H(A)$ counts the number of copies of H in G_n that passes through a given set A of distinct vertices.

EXAMPLE 2.1. To help parse the above definition, we compute $t_H(A)$ in a few examples. For this, fix vertices $u, v, w \in V(G_n)$.

- If $H = K_2$ is an edge, then

$$t_{K_2}(\{v\}) = \frac{1}{2} \sum_{u \in V(G_n)} \{a_{uv} + a_{vu}\} = \sum_{u \in V(G_n)} a_{uv},$$

is the degree of vertex v in G_n. On the other hand, $t_{K_2}(\{u, v\}) = \frac{a_{uv}+a_{vu}}{2} = a_{uv}$.

- If $H = K_{1,2}$ is a 2-star (wedge), then

$$t_{K_{1,2}}(\{v\}) = \sum_{\substack{1 \leq u_1 < u_2 \leq |V(G_n)| \\ u_1, u_2 \neq v}} (a_{vu_1} a_{u_1 u_2} + a_{u_2 v} a_{vu_1} + a_{u_1 u_2} a_{u_2 v}),$$

$$t_{K_{1,2}}(\{u, v\}) = \sum_{\substack{1 \leq w \leq |V(G_n)| \\ w \neq u, v}} (a_{vu} a_{uw} + a_{wv} a_{vu} + a_{uw} a_{wv}),$$

$$t_{K_{1,2}}(\{u, v, w\}) = a_{vu} a_{uw} + a_{wv} a_{vu} + a_{uw} a_{wv}.$$

- If $H = K_3$ is a triangle, then

$$t_{K_3}(\{v\}) = \sum_{\substack{1 \leq u_1 < u_2 \leq |V(G_n)| \\ u_1, u_2 \neq v}} a_{vu_1} a_{u_1 u_2} a_{vu_2}, \qquad t_{K_3}(\{u, v\}) = \sum_{\substack{1 \leq w \leq |V(G_n)| \\ w \neq u, v}} a_{vu} a_{uw} a_{vw},$$

counts the number of triangles in G_n which passes through the vertex v, and the edge (u, v), respectively. Finally, $t_{K_3}(\{u, v, w\}) = a_{vu} a_{uw} a_{vw}$.

Our first result gives a necessary and sufficient condition for the consistency of the HT estimator $\widehat{N}(H, G_n)$ (recall (1.3)). Note that, since the parameter being estimated $N(H, G_n)$ can grow to infinity with n, consistency is defined in terms of the ratio of the estimator to the true parameter converging to 1. More formally, given a sequence of graphs $\{G_n\}_{n\geq 1}$ the HT estimator $\widehat{N}(H, G_n)$ is said to be *consistent* for the true motif count $N(H, G_n)$, if

$$\frac{\widehat{N}(H, G_n)}{N(H, G_n)} \xrightarrow{P} 1,$$

as $n \to \infty$.

THEOREM 2.1. *Suppose $G_n = (V(G_n), E(G_n))$ is a sequence of graphs, with $|V(G_n)| \to \infty$ as $n \to \infty$, and H is a fixed connected graph. Then, given a sampling ratio $p_n \in (0, 1)$ which satisfies* (2.1), *the HT estimator $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$ if and only if the following holds: For all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{N(H, G_n)} \sum_{\substack{A\subset V(G_n)\\ 1\leq|A|\leq|V(H)|}} t_H(A)\mathbf{1}\{t_H(A) > \varepsilon p_n^{|A|} N(H, G_n)\} = 0. \tag{2.3}$$

REMARK 2.1. Note that since every term in the sum in (2.3) is nonnegative, (2.3) is equivalent to

$$\lim_{n\to\infty} \frac{1}{N(H, G_n)} \sum_{\substack{A\subset V(G_n)\\ |A|=s}} t_H(A)\mathbf{1}\{t_H(A) > \varepsilon p_n^{s} N(H, G_n)\} = 0, \tag{2.4}$$

for all $\varepsilon > 0$ and all $1 \leq s \leq |V(H)|$. To understand the implications of the condition in (2.3) (or equivalently, (2.4)) note that

$$\begin{aligned}\sum_{\substack{A\subset V(G_n)\\ 1\leq|A|\leq|V(H)|}} t_H(A) &= \sum_{K=1}^{|V(H)|} \sum_{\substack{A\subset V(G_n)\\ |A|=K}} \frac{1}{|\mathrm{Aut}(H)|} \sum_{\boldsymbol{s}:\overline{\boldsymbol{s}}\supseteq A} M_H(\boldsymbol{s}) \\ &= \sum_{K=1}^{|V(H)|} \sum_{\boldsymbol{s}\in V(G_n)_{|V(H)|}} \frac{1}{|\mathrm{Aut}(H)|} M_H(\boldsymbol{s}) \sum_{\substack{A\subseteq\overline{\boldsymbol{s}}\\ |A|=K}} 1 \\ &= \sum_{K=1}^{|V(H)|} N(H, G_n)\binom{|V(H)|}{K} = (2^{|V(H)|} - 1)N(H, G_n).\end{aligned} \tag{2.5}$$

Hence, (2.3) demands that the contribution to $N(H, G_n)$ coming from subsets of vertices with “high” local counts is asymptotically negligible.

The proof of Theorem 2.1 is given in Section 2.1. To show (2.3) is sufficient for consistency, we define a truncated random variable $T_\varepsilon^+(H, G_n)$ (see (3.2)), which is obtained by truncating the HT estimator whenever the local counts functions are large, more precisely, if $t_H(A) > \varepsilon p_n^{|A|} N(H, G_n)$. Then the proof involves two steps: (1) showing that the difference between $T_\varepsilon^+(H, G_n)$ and $T(H, G_n)$ is asymptotically negligible whenever (2.3) holds (Lemma 3.1), and (2) a second-moment argument to show that $T_\varepsilon^+(H, G_n)$ concentrates around its expectation. For the necessity, assuming condition (2.3) does not hold, an application of the well-known Fortuin–Kasteleyn–Ginibre (FKG) correlation inequality [30], Chapter 2, shows that with positive probability no $|V(H)|$-tuple with “high” local count functions is observed. Moreover, conditional on this event, there is a positive probability (bounded

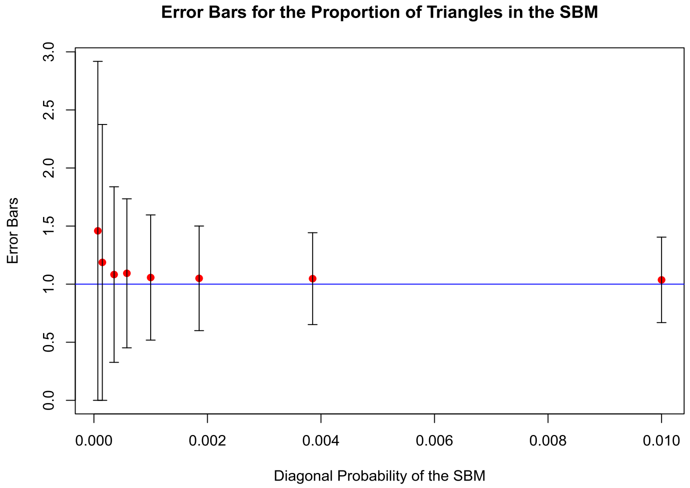

FIG. 2. *Error bars for* $\widehat{N}(K_3, G_n)/N(K_3, G_n)$ *in a* 2-*block stochastic block model on* $n = 10{,}000$ *vertices and equal block size, with off-diagonal probability* 0.5 *and diagonal probability varying between* 0 *and* 0.01 (*shown along the horizontal axis*).

away from 0) that the HT estimator is atypically small. This implies that the (unconditional) probability of the HT estimator being atypically small is also bounded away from zero, which shows the inconsistency of the HT estimator.

In Section 2.3, we will use Theorem 2.1 to derive the precise thresholds for consistency of the HT estimator for many natural classes of graph ensembles. The condition in (2.4) simplifies for specific choices of the motif H, as illustrated for the number of edges ($H = K_2$) in the example below.

EXAMPLE 2.2. Suppose $H = K_2$ is an edge. Then $N(K_2, G_n) = |E(G_n)|$ is the number of edges in G_n and, recalling the calculations in Example 2.1, the assumption in (2.4) is equivalent to the following two simultaneous conditions: For all $\varepsilon > 0$,

$$\lim_{n\to\infty} p_n^2 |E(G_n)| = \infty \quad \text{and}$$

$$\lim_{n\to\infty} \frac{1}{|E(G_n)|} \sum_{v=1}^{|V(G_n)|} d_v \mathbf{1}\{d_v > \varepsilon p_n |E(G_n)|\} = 0, \tag{2.6}$$

where d_v is the degree of the vertex v in G_n. Note that first condition requires that the expected number of edges in the sampled graph goes to infinity, and the second condition ensures that the fraction of edges incident on vertices with "high" degree (greater than $\varepsilon p_n |E(G_n)|$) is small. In Example D.1, we construct a sequence of graphs $\{G_n\}_{n\geq 1}$ for which $p_n^2 |E(G_n)| \to \infty$, but the HT estimator $\widehat{N}(K_2, G_n)$ is inconsistent, illustrating the necessity of controlling the number of edges incident on the high-degree vertices, as in the second condition of (2.6). The condition in (2.4) can be similarly simplified for $H = K_{1,2}$ and $H = K_3$ using the calculations in Example 2.1.

Figure 2 shows the empirical 1-standard deviation error bars for estimating the number of triangles in a 2-block stochastic block model (SBM) with equal block sizes, where edges between vertices in the same block are present independently with probability $a \in (0, 1)$ and edges between vertices in different blocks are present independently with probability $b \in (0, 1)$. Here, fixing $a, b \in (0, 1)$ we consider a realization of G_n from a stochastic block model on $n = 10{,}000$ vertices with equal block sizes and diagonal probability a and off-diagonal probability $b = 0.5$, and sampling ratio $p_n = 0.03$. Figure 2 then shows the empirical 1-standard deviation error bars of $\widehat{N}(K_3, G_n)/N(K_3, G_n)$ over 1000 repetitions, for a range of 8 values of a between 0 and 0.01 (as shown along the horizontal axis). Note that as a increases, the sizes of the error bars decrease, that is, $\widehat{N}(K_3, G_n)$ becomes a more accurate estimator of $N(K_3, G_n)$. This is because one of the conditions that determine the consistency of $\widehat{N}(K_3, G_n)$ is that the expected number of triangles in the sampled graph diverges, that is, $\mathbb{E}[T(K_3, G_n)] = p_n^3\mathbb{E}[N(K_3, G_n)]$ (which is obtained by taking $s = 3$ in (2.4)). Now, as a increases, $\mathbb{E}[N(K_3, G_n)]$, which is the expected number of triangles in the SBM, increases, hence $\mathbb{E}[T(K_3, G_n)]$ increases, improving the accuracy of $\widehat{N}(K_3, G_n)$ for estimating $N(K_3, G_n)$.

2.1.1. *A simpler variance condition.* In this section, we discuss a simpler sufficient condition for the consistency of the HT estimator, arising from the direct application of Chebyshev's inequality, which will be useful in applications. To this end, note that

$$\lim_{n\to\infty} \frac{1}{N(H, G_n)^2} \sum_{\substack{A\subset V(G_n)\\ 1\le|A|\le|V(H)|}} \frac{t_H(A)^2}{p_n^{|A|}} = 0 \tag{2.7}$$

is a sufficient condition for (2.3), since

$$t_H(A)\mathbf{1}\{t_H(A) > \varepsilon p_n^{|A|} N(H, G_n)\} \le \frac{t_H(A)^2}{\varepsilon p_n^{|A|} N(H, G_n)}.$$

The condition in (2.7), which does not require any truncations, is often easier to verify, as will be seen in the examples discussed below. To derive (2.7) without using (2.3), use Chebyshev's inequality to note that a straightforward sufficient condition for the consistency of the estimate $\widehat{N}(H, G_n)$ is that $\operatorname{Var}[\widehat{N}(H, G_n)] = o(\widehat{N}(H, G_n)^2)$. This last condition is equivalent to (2.7), as can be seen by invoking Lemma C.1 to get the estimate

$$\operatorname{Var}(\widehat{N}(H, G_n)) = \Theta\biggl(\sum_{\substack{A\subset V(G_n)\\ 1\le|A|\le|V(H)|}} \frac{t_H(A)^2}{p_n^{|A|}}\biggr).$$

Even though the variance condition (2.7) is natural and often easier to verify, it is not necessary for consistency, as shown in the example below.

EXAMPLE 2.3 (The variance condition is not necessary for consistency). Let $H = K_2$ be the edge, and G_n be the disjoint union of an a_n-star K_{1,a_n} and b_n disjoint edges, with $a_n \ll b_n \ll a_n^{3/2}$. Then

$$\begin{aligned} |V(G_n)| &= a_n + 1 + 2b_n = (1 + o(1))2b_n, \\ N(H, G_n) = |E(G_n)| &= a_n + b_n = (1 + o(1))b_n. \end{aligned} \tag{2.8}$$

In this case, the HT estimator is consistent whenever the sampling probability p_n satisfies $\frac{1}{\sqrt{b_n}} \ll p_n \ll a_n^2/b_n$. To see this, note that $p_n^2|E(G_n)| = (1 + o(1))p_n^2 b_n \gg 1$, that is, the first condition in (2.6) holds. Also, fixing $\varepsilon > 0$ and noting that $p_n|E(G_n)| = (1 + o(1))p_n b_n \gg 1$ implies, for all n large only the central vertex of the a_n-star satisfies the $d_v > \varepsilon p_n|E(G_n)|$ cutoff. Hence,

$$\sum_{v=1}^{|V(G_n)|} d_v \mathbf{1}\{d_v > \varepsilon p_n|E(G_n)\} = a_n = o(b_n),$$

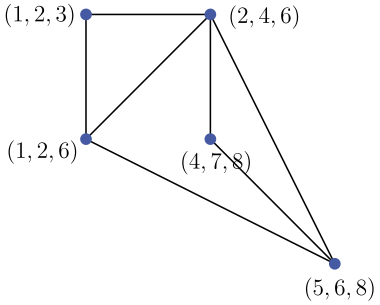

FIG. 3. *The graph* $\mathcal{G}(\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4, \boldsymbol{s}_5)$ *as in Definition* 2.2 *with* $\boldsymbol{s}_1 = (1, 2, 3)$, $\boldsymbol{s}_2 = (1, 2, 6)$, $\boldsymbol{s}_3 = (4, 7, 8)$, $\boldsymbol{s}_4 = (2, 4, 6)$ *and* $\boldsymbol{s}_5 = (5, 6, 8)$.

verifying second condition in (2.6). However, since

$$\frac{1}{p_n|E(G_n)|^2} \sum_{v=1}^{|V(G_n)|} d_v^2 = \frac{1}{p_n b_n^2}\big(a_n^2 + a_n + b_n\big) = \big(1 + o(1)\big)\frac{a_n^2}{p_n b_n^2} \to \infty,$$

the variance condition (2.7) does not hold. Thus for this example one needs the full strength of Theorem 2.1 to show that the HT estimator is consistent.

2.2. *Asymptotic normality of the HT estimator.* In this section, we determine the precise conditions under which the HT estimator is asymptotically normal. For this, recall the definition of $Z(H, G_n)$ from (1.4),

$$Z(H, G_n) := \frac{\widehat{N}(H, G_n) - N(H, G_n)}{\sqrt{\operatorname{Var}[\widehat{N}(H, G_n)]}} = \frac{T(H, G_n) - p_n^{|V(H)|} N(H, G_n)}{\sigma(H, G_n)}, \tag{2.9}$$

where $\sigma(H, G_n)^2 := \operatorname{Var}[T(H, G_n)]$. To begin with, one might wonder whether the conditions which ensure the consistency of $\widehat{N}(H, G_n)$ is enough to imply the asymptotic normality of $Z(H, G_n)$. However, it is easy to see that this is not the case. In fact, there are examples where $\widehat{N}(H, G_n)$ is consistent, but $Z(H, G_n)$ has a non-Gaussian limiting distribution (see Example D.2 in Appendix D). Hence, to establish the asymptotic normality of $Z(H, G_n)$ additional conditions are needed. To state our result, we need the following definition.

DEFINITION 2.2. Fix $r \geq 1$. Given a collection of r tuples $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_r\}$ from $V(G_n)_{|V(H)|}$, let $\mathcal{G}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r)$ be the simple graph with vertex set $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\}$, with an edge between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ whenever $\overline{\boldsymbol{s}}_i \cap \overline{\boldsymbol{s}}_j \neq \varnothing$ (see Figure 3 for an illustration). We will say the collection $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\}$ is *connected*, if the graph $\mathcal{G}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r)$ is connected. The set of all r tuples $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\}$ in $V(G_n)_{|V(H)|}$ such that the collection $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\}$ is connected will be denoted by $\mathcal{K}_{n,r}$.

Now, denote by W_n the random variable

$$W_n := \sum_{\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3, \boldsymbol{s}_4\} \in \mathcal{K}_{n,4}} |Y_{\boldsymbol{s}_1} Y_{\boldsymbol{s}_2} Y_{\boldsymbol{s}_3} Y_{\boldsymbol{s}_4}|, \tag{2.10}$$

where $Y_{\boldsymbol{s}} := \frac{1}{|\operatorname{Aut}(H)|} \prod_{(i,j) \in E(H)} a_{s_i s_j} (X_{\boldsymbol{s}} - p_n^{|V(H)|})$. In the following theorem, we give a quantitative error bound (in terms of the Wasserstein distance) between $Z(H, G_n)$ and the standard normal distribution $N(0, 1)$, in terms of the expected value of the random variable

W_n. To this end, recall that the Wasserstein distance between random variables $X \sim \mu$ and $Y \sim \nu$ on $\mathbb{R}$ is defined as

$$\operatorname{Wass}(X, Y) = \sup\left\{\left|\int f \, \mathrm{d}\mu - \int f \, \mathrm{d}\nu\right| : f \text{ is 1-Lipschitz}\right\},$$

where a function $f : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz if $|f(x) - f(y)| \le |x - y|$, for all $x, y \in \mathbb{R}$.

THEOREM 2.3. *Fix a connected graph H, a network $G_n = (V(G_n), E(G_n))$ and a sampling ratio p_n, which satisfies* (2.1). *Then*

$$\operatorname{Wass}\bigl(Z(H, G_n), N(0, 1)\bigr) \lesssim \frac{|V(H)|}{(1-\kappa)^3} \cdot \sqrt{\frac{\mathbb{E}[W_n]}{\sigma(H, G_n)^4}}, \tag{2.11}$$

where $Z(H, G_n)$ and W_n are as defined in (2.9) *and* (2.10), *respectively. Moreover, if $p_n \in (0, \frac{1}{20}]$, then $\frac{\mathbb{E}[W_n]}{\sigma(H,G_n)^4} \le \mathbb{E}[Z(H, G_n)^4] - 3$ and, as a consequence,*

$$\operatorname{Wass}\bigl(Z(H, G_n), N(0, 1)\bigr) \lesssim \bigl|V(H)\bigr| \cdot \sqrt{\mathbb{E}\bigl[Z(H, G_n)^4\bigr] - 3}, \tag{2.12}$$

The proof of this result is given in Appendix A.2. In addition to giving an explicit rate of convergence between $Z(H, G_n)$ and $N(0, 1)$, Theorem 2.3 shows that for p_n small enough, the asymptotic normality of the (standardized) HT estimator exhibits a curious fourth-moment phenomenon, that is, $Z(H, G_n) \xrightarrow{D} N(0, 1)$ whenever $\mathbb{E}[Z(H, G_n)^4] \to 3$ (the fourth moment of the standard normal distribution). The proof uses Stein's method for normal approximation [5, 17, 54] and is a consequence of more general result about the asymptotic normality and the fourth-moment phenomenon of certain random multilinear forms in Bernoulli variables, which might be of independent interest (Theorem A.3).

REMARK 2.2. The fourth-moment phenomenon was first discovered by Nualart and Peccati [48], who showed that the convergence of the first, second and fourth moments to 0, 1 and 3, respectively, guarantees asymptotic normality for a sequence of multiple stochastic Wiener–Itô integrals of fixed order. Later, Nourdin and Peccati [45] provided an error bound for the fourth-moment theorem of [48]. The fourth-moment phenomenon has since then emerged as a unifying principle governing the central limit theorems for various nonlinear functionals of random fields [9, 43, 47]. We refer the reader to the book [46] for an introduction to the topic and the website https://sites.google.com/site/malliavinstein/home for a list of the recent results. The result in Theorem 2.3 is an example of the fourth-moment phenomenon in the context of motif estimation. In fact, the result in Section A on the asymptotic normality of general random multilinear forms suggests that the fourth-moment phenomenon is more universal, and we expect it to emerge in various other combinatorial estimation problems, where counting statistics similar to $T(H, G_n)$ arise naturally.

REMARK 2.3. Note that the result in (2.12) requires an upper bound on the sampling ratio $p_n \le \frac{1}{20}$. This condition ensures that the leading order of the central moments of $T(H, G_n)$ is the same as the leading order of its raw moments (as shown in Lemma A.2), a fact which is used to estimate the error terms arising from the Stein's method calculations. Interestingly, it is, in fact, necessary to assume an upper bound on p_n for the limiting normality and the fourth-moment phenomenon of the HT estimator to hold (see Example D.3 in Appendix D). This example constructs a sequence of graphs $\{G_n\}_{n\ge1}$ for which if p_n is chosen large enough, then $\mathbb{E}[Z(K_2, G_n)^4] \to 3$, but $Z(K_2, G_n)$ does not converge to $N(0, 1)$. However, in applications, where it is natural to chose $p_n \ll 1$ to have any significant reduction in the size of the sampled graph, the fourth-moment phenomenon always holds.

We now discuss how the results above can be used to construct asymptotically valid confidence intervals for the parameter $N(H, G_n)$. To this end, we need to consistently estimate $\sigma(H, G_n)^2$, the variance of $T(H, G_n)$. The following result shows that it is possible to consistently estimate $\sigma(H, G_n)^2$ whenever the error term in (2.11) goes to zero, which combined with the asymptotic normality of $Z(H, G_n)$ gives a confidence for $N(H, G_n)$ with asymptotic coverage probability $1 - \alpha$.

PROPOSITION 2.4. *Fix a connected graph H, a network $G_n = (V(G_n), E(G_n))$ and a sampling ratio p_n, which satisfies* (2.1). *Suppose $\mathbb{E}[W_n] = o(\sigma(H, G_n)^4)$, where W_n is as defined in* (2.10). *Then the following hold, as $n \to \infty$:*

(a) *The HT estimator $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$.*

(b) *Let*

$$\widehat{\sigma}(H, G_n)^2 := \sum_{\substack{s_1, s_2 \in V(G_n)_{|V(H)|} \\ \overline{s}_1 \cap \overline{s}_2 \neq \varnothing}} M_H(s_1) M_H(s_2) \big(X_{s_1} - p_n^{|V(H)|}\big)\big(X_{s_2} - p_n^{|V(H)|}\big).$$

Then $\widehat{\sigma}(H, G_n)^2$ is a consistent estimate of $\sigma(H, G_n)^2$, that is, $\frac{\widehat{\sigma}(H,G_n)^2}{\sigma(H,G_n)^2} \xrightarrow{P} 1$.

(c) *Let $\widehat{\sigma}(H, G_n)_+ := \sqrt{\max(0, \widehat{\sigma}(H, G_n)^2)}$. Then, as $n \to \infty$,*

$$\mathbb{P}\left(N(H, G_n) \in \left[\widehat{N}(H, G_n) - z_{\frac{\alpha}{2}} \frac{\widehat{\sigma}(H, G_n)_+}{p_n^{|V(H)|}}, \widehat{N}(H, G_n) + z_{\frac{\alpha}{2}} \frac{\widehat{\sigma}(H, G_n)_+}{p_n^{|V(H)|}}\right]\right) \to 1 - \alpha,$$

where $z_{\frac{\alpha}{2}}$ is the $(1 - \frac{\alpha}{2})$th quantile of the standard normal distribution $N(0, 1)$.

The proof of this result is given in Appendix A.3. The proof of (a) entails showing that $\sigma(H, G_n)^2 = o((\mathbb{E}[T(H, G_n)])^2)$. This is a consequence of the assumption $\mathbb{E}[W_n] = o(\sigma(H, G_n)^4)$ and the more general bound $\sigma(H, G_n)^6 \lesssim_H \mathbb{E}[W_n](\mathbb{E}[T(H, G_n)])^2$, which can be proved by expanding out the terms and an application of the Hölder's inequality. For (b), note that $\widehat{\sigma}(H, G_n)^2$ is an unbiased estimate of $\sigma(H, G_n)^2$; hence, to prove the consistency of $\widehat{\sigma}(H, G_n)^2$ it suffices to show that $\mathrm{Var}[\widehat{\sigma}(H, G_n)^2] = o(\sigma(H, G_n)^4)$, under the given assumptions. Finally, (c) is an immediate consequence of (b) and the asymptotic normality of $Z(H, G_n)$ proved in Theorem 2.3.

Given the result in Theorem 2.3, it is now natural to wonder whether the convergence of the fourth moment $\mathbb{E}[Z(H, G_n)^4] \to 3$ is necessary for the asymptotic normality of $Z(H, G_n)$. This however turns out to be not the case. In fact, Example D.4 gives a sequence of graphs $\{G_n\}_{n \geq 1}$ for which $Z(K_2, G_n)$ is asymptotic normal, but $\mathbb{E}[Z(K_2, G_n)^4] \nrightarrow 3$, showing that the (untruncated) fourth-moment condition is not necessary for the asymptotic normality of the HT estimator. As we will see, in this example the graph G_n has a few "high" degree vertices, which forces $\mathbb{E}[Z(H, G_n)^4]$ to diverge. However, the existence of a "small" number of high-degree vertices does not affect the distribution of the rescaled statistic. This suggests that, as in the case of consistency in Theorem 2.1, to obtain the precise condition for the asymptotic normality of $Z(H, G_n)$ we need to appropriately truncate the graph G_n, by removing a small number of hubs with "high" local count functions, and consider the moments of the truncated statistic. Toward this end, fix $M > 0$ and define the event

$$\mathcal{C}_M(A) = \big\{t_H(A)^2 > M p_n^{2|A| - 2|V(H)|} \mathrm{Var}[T(H, G_n)]\big\}, \tag{2.13}$$

and $\mathcal{C}_M(s)^c = \bigcap_{A \subseteq s : A \neq \varnothing} \mathcal{C}_M(A)^c$. (For any set A, A^c denotes the complement of A.) Then consider the truncated statistic,

$$T_M^\circ(H, G_n) := \frac{1}{|\mathrm{Aut}(H)|} \sum_{s \in V(G_n)_{|V(H)|}} M_H(s) X_s \mathbf{1}\{\mathcal{C}_M(s)^c\}, \tag{2.14}$$

and define

$$Z_M^\circ(H,G_n):=\frac{T_M^\circ(H,G_n)-\mathbb{E}[T_M^\circ(H,G_n)]}{\sigma(H,G_n)}. \tag{2.15}$$

The following theorem gives a necessary and sufficient condition for asymptotic normality for $Z(H,G_n)$ in the terms of the second and fourth moments of the truncated statistic (2.14).

THEOREM 2.5. *Suppose $G_n=(V(G_n),E(G_n))$ is a sequence of graphs, with $|V(G_n)|\to\infty$ and H is a fixed connected graph. Then, given a sampling ratio $p_n\in(0,\frac{1}{20}]$, the rescaled statistic $Z(H,G_n)\xrightarrow{D}N(0,1)$ if and only if*

$$\begin{aligned}&\limsup_{M\to\infty}\limsup_{n\to\infty}\big|\mathbb{E}[Z_M^\circ(H,G_n)^2]-1\big|=0 \quad \textit{and}\\ &\limsup_{M\to\infty}\limsup_{n\to\infty}\big|\mathbb{E}[Z_M^\circ(H,G_n)^4]-3\big|=0,\end{aligned} \tag{2.16}$$

holds simultaneously.

This result shows that the asymptotic normality of $Z(H,G_n)$ is characterized by a truncated fourth-moment phenomenon, more precisely, the convergence of the second and fourth moments of $Z_M^\circ(H,G_n)$ to 1 and 3, respectively. Note that the second-moment condition in (2.16) ensures that $\operatorname{Var}[T_M^\circ(H,G_n)]=(1+o(1))\operatorname{Var}[T(H,G_n)]$. Hence, the fourth-moment condition in (2.16) and the Theorem 2.3 implies that

$$\frac{T_M^\circ(H,G_n)-\mathbb{E}[T_M^\circ(H,G_n)]}{\sqrt{\operatorname{Var}[T_M^\circ(H,G_n)]}}\xrightarrow{D}N(0,1).$$

Therefore, to establish the sufficiency of the conditions in (2.16), it suffices to show that the difference between $T(H,G_n)$ and $T_M^\circ(H,G_n)$ scaled by $\operatorname{Var}[T(H,G_n)]$ is small, which follows from the properties of the truncation event (2.13) (see Lemma C.2). To prove that (2.16) is also necessary for the asymptotic normality of $Z(H,G_n)$, we show all moments of $Z_M^\circ(H,G_n)$ are bounded (Lemma C.3), which combined with the fact that $T(H,G_n)-T_M^\circ(H,G_n)\xrightarrow{P}0$ and uniform integrability, implies the desired result (see Appendix C.2 for details).

2.3. *Thresholds for consistency and normality.*

In this section, we apply the results above to derive the thresholds for consistency and asymptotic normality of the HT estimator in various natural graph ensembles. Throughout this section, we will assume that $p_n\in(0,\frac{1}{20}]$.

2.3.1. *Bounded degree graphs.*

We begin with graphs which have bounded maximum degree. Toward this, denote by d_v the degree of the vertex v in $G_n=(V(G_n),E(G_n))$, and let $\Delta(G_n)=\max_{v\in V(G_n)}d_v$ be the maximum degree of the graph G_n.

PROPOSITION 2.6 (Bounded degree graphs). *Suppose $\{G_n\}_{n\ge1}$ is a sequence of graphs with bounded maximum degree, that is, $\Delta:=\sup_{n\ge1}\Delta(G_n)=O(1)$. Then for any connected graph H the following hold:*

(a) *If $p_n^{|V(H)|}N(H,G_n)\gg1$, then the HT estimator $\widehat{N}(H,G_n)$ is consistent for $N(H,G_n)$, and the rescaled statistic $Z(H,G_n)\xrightarrow{D}N(0,1)$. Moreover,*

$$\operatorname{Wass}\big(Z(H,G_n),N(0,1)\big)\lesssim_{\Delta,H}\sqrt{\frac{1}{p_n^{|V(H)|}N(H,G_n)}}.$$

(b) *If $p_n^{|V(H)|}N(H, G_n) = O(1)$, then the HT estimator $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$ and the rescaled statistic $Z(H, G_n)$ is not asymptotically normal.*

Recall that $\mathbb{E}[T(H, G_n)] = p_n^{|V(H)|}N(H, G_n)$. Therefore, in other words, the result above shows that the HT estimator is consistent and asymptotic normal in bounded degree graphs whenever the expected number of copies of H in the sampled graph diverges, whereas it is inconsistent whenever the expected number copies remains bounded. The proof of Proposition 2.6 is given in Appendix B.1. For (a), using Proposition 2.4, it is suffices to bound $\frac{1}{\sigma(H,G_n)^4}\mathbb{E}[W_n]$. This involves, recalling the definition of W_n from (2.10), bounding the number of copies of various subgraphs in G_n obtained by the union of 4 isomorphic copies H, which in this case can be estimated using the maximum degree bound on G_n. For (b), we show that whenever $\mathbb{E}[T(H, G_n)] = p_n^{|V(H)|}N(H, G_n) = O(1)$, there is a positive chance that $T(H, G_n)$ is zero, which immediately rule out consistency and normality.

2.3.2. *Erdős–Rényi random graphs.* We now derive the thresholds for consistency and asymptotic normality in various random graph models. We begin with the Erdős–Rényi model $G_n \sim \mathcal{G}(n, q_n)$, which is a random graph on n vertices where each edge is present independently with probability $q_n \in (0, 1)$. Here, the location of the phase transition is related to the notion of balancedness of a graph.

DEFINITION 2.7 ([34], Chapter 3). For a fixed connected graph H, define

$$m(H) = \max_{H_1 \subseteq H} \frac{|E(H_1)|}{|V(H_1)|},$$

where the maximum is over all nonempty subgraphs H_1 of H. The graph H is said to be *balanced*, if $m(H) = \frac{|E(H)|}{|V(H)|}$, and *unbalanced* otherwise.

THEOREM 2.8 (Erdős–Rényi graphs). *Let $G_n \sim \mathcal{G}(n, q_n)$ be an Erdős–Rényi random graph with edge probability $q_n \in (0, 1)$. Then for any connected graph H the following hold*:

(a) *If $np_nq_n^{m(H)} \gg 1$, then the HT estimator $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$, and the rescaled statistic $Z(H, G_n) \xrightarrow{D} N(0, 1)$. Moreover,*

$$\mathrm{Wass}\big(Z(H, G_n), N(0, 1)\big) = O_P\big(\big(np_nq_n^{m(H)}\big)^{-\frac{1}{2}}\big).$$

(b) *If $np_nq_n^{m(H)} = O(1)$, then $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$, and $Z(H, G_n)$ is not asymptotically normal.*

The proof of this result is given in Appendix B.2. Here, to estimate W_n, we first take expectation over the randomness of the graph, and then use an inductive counting argument (Lemma B.2) combined with a second-moment calculation, to obtain the desired bound.

REMARK 2.4. To interpret the threshold in Theorem 2.8, recall that $nq_n^{m(H)}$ is the threshold for the occurrence of H in the random graph $\mathcal{G}(n, q_n)$ [34], Theorem 3.4. More precisely, whenever $nq_n^{m(H)} = O(1)$ the number of copies of H in $\mathcal{G}(n, q_n)$ is $O_P(1)$, whereas if $nq_n^{m(H)} \gg 1$, the number of copies of H in G_n diverges. In this case, conditional on the set of sampled vertices S, the observed graph behaves like the Erdős–Rényi model $\mathcal{G}(|S|, q_n)$. As a result, since $S \sim \mathrm{Bin}(n, p_n)$, the observed graph (unconditionally) looks roughly like the model $\mathcal{G}(np_n, q_n)$. Therefore, Theorem 2.8 essentially shows that the HT estimator is consistent and asymptotically normal whenever the number of copies of H in sampled graph

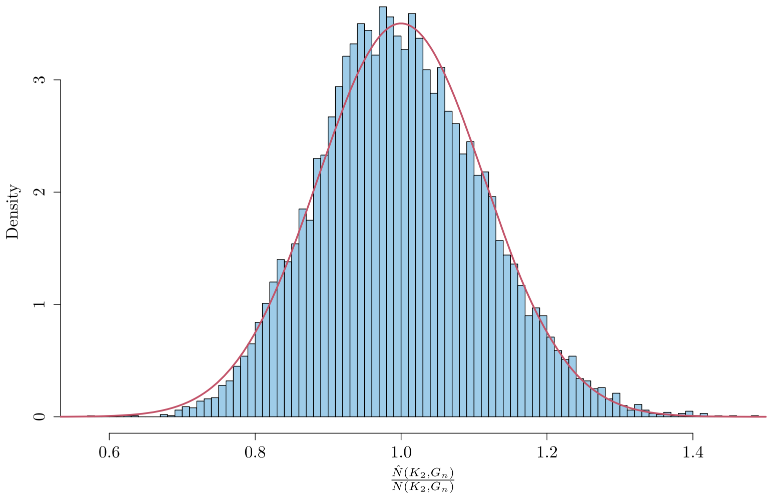

FIG. 4. *Histogram of* $\widehat{N}(K_2, G_n)/N(K_2, G_n)$ *in the Erdős–Rényi random graph* $G_n \sim \mathcal{G}(10{,}000, 0.5)$ *with sampling ratio* $p_n = 0.03$ *over* 10,000 *replications*, *and the limiting normal density* (*plotted in red*).

diverges (which happens if $np_nq_n^{m(H)} \to \infty$), whereas it is inconsistent whenever the number of copies of H is bounded in probability. The histogram in Figure 4 illustrates the asymptotic normality of the HT estimator for the number of edges ($H = K_2$). Here, we fix a realization of the Erdős–Rényi random graph $G_n \sim \mathcal{G}(n, q_n)$, with $n = 10{,}000$ and $q_n = \frac{1}{2}$, choose the sampling ratio $p_n = 0.03$, and plot the histogram of $\widehat{N}(K_2, G_n)/N(K_2, G_n)$ over 10,000 replications. Note that, as expected, the histogram is centered around 1, with the red curve showing the limiting normal density.

Note that Theorem 2.8 above gives a CLT for $\widehat{N}(H, G_n)$ centered around $N(H, G_n)$, when $np_nq_n^{m(H)} \gg 1$. However, since G_n is a random graph $N(H, G_n)$ is itself random, and it is natural to wonder whether one can obtain a CLT for $\widehat{N}(H, G_n)$ centered around $\mathbb{E}[N(H, G_n)]$, where the expectation is taken with respect to the randomness of G_n. This question is not just specific to the Erdős–Rényi model, it arises whenever G_n is generated from any underlying stochastic model. To address this issue, suppose $\{G_n\}_{n\geq 1}$ is a sequence of random graphs (from some generative model) and define

$$\mathcal{A}(H, G_n) := \frac{\widehat{N}(H, G_n) - \mathbb{E}[N(H, G_n)]}{\sqrt{\mathrm{Var}[\widehat{N}(H, G_n)]}}, \tag{2.17}$$

where the expectation and the variance above are taken over both the randomness of the sampling scheme and the graph G_n. Note that

$$\mathcal{A}(H, G_n) = \sqrt{\frac{\mathrm{Var}_{G_n}[\widehat{N}(H, G_n)]}{\mathrm{Var}[\widehat{N}(H, G_n)]}} \cdot Z(H, G_n) + \sqrt{\frac{\mathrm{Var}[N(H, G_n)]}{\mathrm{Var}[\widehat{N}(H, G_n)]}} \cdot \mathcal{E}(H, G_n), \tag{2.18}$$

where

$$Z(H, G_n) := \frac{\widehat{N}(H, G_n) - N(H, G_n)}{\sqrt{\mathrm{Var}_{G_n}[\widehat{N}(H, G_n)]}} \quad \text{and}$$

$$\mathcal{E}(H, G_n) := \frac{N(H, G_n) - \mathbb{E}[N(H, G_n)]}{\sqrt{\mathrm{Var}[N(H, G_n)]}}, \tag{2.19}$$

with $\mathbb{E}_{G_n}$ and Var_{G_n} denoting the conditional expectation and conditional variance taken conditionally on the random graph G_n. Recall that Theorem 2.3 deals with the CLT of $Z(H, G_n)$ conditional on the graph G_n (often known as a *quenched* CLT in the language of statistical physics). Given this result, to obtain a CLT for $\mathcal{A}(H, G_n)$ (i.e., an *annealed* CLT in statistical physics terminology), we would need to show a CLT for $\mathcal{E}(H, G_n)$ and establish that the conditional variance $\mathrm{Var}_{G_n}[\widehat{N}(H, G_n)]$ is consistent for its expectation (see Lemma B.3 for the formal statement). In particular, for the Erdős–Rényi (ER) model $G(n, q_n)$ both these results can be easily established and we have the following result.

COROLLARY 2.9 (Erdős–Rényi graphs (annealed version)). *Let $G_n \sim \mathcal{G}(n, q_n)$ be an Erdős–Rényi random graph with edge probability $q_n \in (0, 1)$. Then for any connected graph H the following hold*:

(a) *If $np_nq_n^{m(H)} \gg 1$, then the HT estimator $\widehat{N}(H, G_n)$ is consistent for $\mathbb{E}[N(H, G_n)]$ and $\mathcal{A}(H, G_n) \xrightarrow{D} N(0, 1)$.*

(b) *If $np_nq_n^{m(H)} = O(1)$, then $\widehat{N}(H, G_n)$ is not consistent for $\mathbb{E}[N(H, G_n)]$, and $\mathcal{A}(H, G_n)$ is not asymptotically normal.*

The proof of Corollary 2.9 is given in Appendix B.3. This is a consequence of a more general result (see Lemma B.3) about the CLT of $\mathcal{A}(H, G_n)$ (when G_n is generated according to some stochastic model). In particular, in Lemma B.3 we show that $\mathcal{A}(H, G_n) \xrightarrow{D} N(0, 1)$ whenever the following conditions hold: (a) conditional on the graph sequence $\{G_n\}_{n\geq 1}$, $Z(H, G_n) \xrightarrow{D} N(0, 1)$, (b) $\mathcal{E}(H, G_n) \xrightarrow{D} N(0, 1)$, and (c) $\mathrm{Var}_{G_n}[\widehat{N}(H, G_n)]$ is consistent for its expected value $\mathbb{E}[\mathrm{Var}_{G_n}[\widehat{N}(H, G_n)]]$. These conditions can be easily verified for the Erdős–Rényi model $\mathcal{G}(n, q_n)$ whenever $np_nq_n^{m(H)} \gg 1$, which establishes the result in Corollary 2.9(1).

REMARK 2.5. The normality condition (assumption (b)) on $\mathcal{E}(H, G_n)$ in Lemma B.3 can be removed if instead of assumption (c) the following stronger condition holds:

$$\frac{\mathrm{Var}[\widehat{N}(H, G_n)]}{\mathbb{E}[\mathrm{Var}[\widehat{N}(H, G_n)|G_n]]} \xrightarrow{P} 1. \tag{2.20}$$

This is because (2.20) implies $\mathrm{Var}[N(H, G_n)] \ll \mathrm{Var}[\widehat{N}(H, G_n)]$, hence, recalling (2.18), the CLT of $\mathcal{A}(H, G_n)$ follows from the conditional CLT of $Z(H, G_n)$, since $\mathcal{E}(H, G_n)$ is bounded in probability. In the Erdős–Rényi model, there is a regime of the parameters p_n, q_n where (2.20) holds. There is also a regime where $\mathrm{Var}[N(H, G_n)]$ and $\mathrm{Var}[\widehat{N}(H, G_n)]$ are of the same order (i.e., (2.20) does not hold), where one needs to invoke Lemma B.3 to establish the CLT of $\mathcal{A}(H, G_n)$. (Recall that unlike (2.20), assumption (c) in Lemma B.3 holds in the full range of parameters in Erdős–Rényi model.) Nevertheless, condition (2.20) broadens the scope of our results and can be useful in other random graph models.

2.3.3. *Random regular graphs.* As a corollary to Theorem 2.8, we can also derive the threshold for random regular graphs. To this end, denote by $\mathscr{G}_{n,d}$ the collection of all simple d-regular graphs on n vertices, where $1 \leq d \leq n - 1$ is such that nd is even.

COROLLARY 2.10 (Random regular graphs). *Suppose G_n is a uniform random sample from $\mathscr{G}_{n,d}$ and $H = (V(H), E(H))$ is a connected graph with maximum degree $\Delta(H)$.*

(a) *If $d \gg 1$, then setting $q_n = d/n$ the following hold*:

- *If $np_nq_n^{m(H)} \gg 1$, then $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$, and $Z(H, G_n)$ converges in distribution to $N(0, 1)$.*
- *If $np_nq_n^{m(H)} = O(1)$, then $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$, and $Z(H, G_n)$ is not asymptotically normal.*

(b) *If $d = O(1)$, then assuming $\Delta(H) \leq d$, the following hold*:

- *If $|E(H)| = |V(H)| - 1$, then $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$ and $Z(H, G_n)$ converges in distribution to $N(0, 1)$ if and only if $np_n^{|V(H)|} \gg 1$.*
- *If $|E(H)| \geq |V(H)|$, then $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$, and $Z(H, G_n)$ is not asymptotically normal, irrespective of the value of p_n.*

It is well known that the typical behavior of the number of small subgraphs in a random d-regular graph asymptotically equals to that in a Erdős–Rényi graph $\mathcal{G}(n, q_n)$, with $q_n = d/n$, whenever $d \gg 1$ [35, 40]. As a result, the threshold for consistency and asymptotic normality for random d-regular graphs obtained in Corollary 2.10 above, match with the threshold for Erdős–Rényi graphs obtained in Theorem 2.8 with $q_n = d/n$, whenever $d \gg 1$. However, this analogy with the Erdős–Rényi model is no longer valid when $d = O(1)$. In this case, to compute the threshold we invoke Proposition 2.6 instead, which deals with the case of general bounded degree graphs. Note that here it suffices to assume $\Delta(H) \leq d$, since $N(H, G_n) = 0$ whenever $\Delta(H) > d$. Therefore, assuming $\Delta(H) \leq d$, there are two cases: (1) $|E(H)| = |V(H)| - 1$ (i.e., H is a tree) and (2) $|E(H)| \geq |V(H)|$ (i.e., H has a cycle). In the second case, it can be easily shown that $N(H, G_n) = O_P(1)$; hence, by Proposition 2.6(b) consistency and asymptotic normality does not hold. On the other hand, in the first case, by a inductive counting argument, it can be shown that $N(H, G_n) = \Theta_P(n)$. Hence, by Proposition 2.6(a), the threshold for consistency and asymptotic normality is $np_n^{|V(H)|} \gg 1$. The details of the proof are given in Appendix B.4.

2.3.4. *Graphons.* In this section, we apply our results for dense graph sequences. The asymptotics of dense graphs can be studied using the framework of graph limit theory (graphons), which was developed by Borgs et al. [11, 12] (for a detailed exposition see the book by Lovász [42]), and commonly appears in various popular models for network analysis (see [1, 10, 15, 16, 18, 24, 55] and the references therein). For a detailed exposition of the theory of graph limits, refer to Lovász [42]. Here, we recall the basic definitions about the convergence of graph sequences. If F and G are two graphs, then define the homomorphism density of F into G by

$$t(F, G) := \frac{|\hom(F, G)|}{|V(G)|^{|V(F)|}},$$

where $|\hom(F, G)|$ denotes the number of homomorphisms of F into G. In fact, $t(F, G)$ is the proportion of maps $\phi : V(F) \to V(G)$, which define a graph homomorphism.

To define the continuous analogue of graphs, consider $\mathscr{W}$ to be the space of all measurable functions from $[0, 1]^2$ into $[0, 1]$ that satisfy $W(x, y) = W(y, x)$, for all $x, y \in [0, 1]$. For a simple graph F with $V(F) = \{1, 2, \dots, |V(F)|\}$, let

$$t(F, W) = \int_{[0,1]^{|V(F)|}} \prod_{(i,j)\in E(F)} W(x_i, x_j)\, \mathrm{d}x_1\, \mathrm{d}x_2 \cdots \mathrm{d}x_{|V(F)|}.$$

DEFINITION 2.11 ([11, 12, 42]). A sequence of graphs $\{G_n\}_{n\geq 1}$ is said to *converge to* W if for every finite simple graph F,

$$\lim_{n\to\infty} t(F, G_n) = t(F, W).$$

The limit objects, that is, the elements of $\mathscr{W}$, are called *graph limits* or *graphons*. A finite simple graph $G = (V(G), E(G))$ can also be represented as a graphon in a natural way. Define

$$W^G(x, y) = \mathbf{1}\{(\lceil |V(G)|x\rceil, \lceil |V(G)|y\rceil) \in E(G)\},$$

that is, partition $[0, 1]^2$ into $|V(G)|^2$ squares of side length $1/|V(G)|$, and let $W^G(x, y) = 1$ in the (i, j)th square if $(i, j) \in E(G)$, and 0 otherwise.

The following result gives the threshold for consistency and asymptotic normality of the HT estimator for a sequence of graphs $\{G_n\}_{n\geq 1}$ converging to a graphon W.

PROPOSITION 2.12 (Graphons). *Fix a connected graph H and suppose $G_n = (V(G_n), E(G_n))$ is a sequence of graphs converging to a graphon W such that $t(H, W) > 0$. Then the following hold*:

(a) *If $|V(G_n)|p_n \gg 1$, then the HT estimator $\widehat{N}(H, G_n)$ is consistent for $N(H, G_n)$ and the rescaled statistic $Z(H, G_n)$ is asymptotically normal. Moreover,*

$$\mathrm{Wass}(Z(H, G_n), N(0, 1)) \lesssim_H (|V(G_n)|p_n)^{-\frac{1}{2}}.$$

(b) *If $|V(G_n)|p_n = O(1)$, then the HT estimator $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$ and the rescaled statistic $Z(H, G_n)$ is not asymptotically normal.*

Note that the assumption $t(H, W) > 0$ ensures that the density of the graph H in the graphon W is positive, which can be equivalently reformulated as $N(H, G_n) = \Theta(|V(G_n)|^{|V(H)|})$. In fact, as will be evident from the proof, the result above holds for any sequence of graphs with $N(H, G_n) = \Theta(|V(G_n)|^{|V(H)|})$.

2.4. *Organization.* The rest of the article is organized as follows. The proof of Proposition 2.4 is given Section 3. Consequences of our results and future directions are discussed in Section 4. The proofs of the remaining results are given in the Supplementary Material [8].

3. Proof of Theorem 2.1. In this section, we prove the necessary and sufficient condition for the consistency of the estimate $\widehat{N}(H, G_n)$. We start with a few definitions. Fix an $\varepsilon > 0$. For each set $A \subset V(G_n)$ and each $\boldsymbol{s} \in V(G_n)_{|V(H)|}$, define the following events:

$$\mathscr{B}_{n,\varepsilon}(A) := \{t_H(A) > \varepsilon p_n^{|A|} N(H, G_n)\}, \qquad \mathscr{B}_{n,\varepsilon}(\boldsymbol{s})^c := \bigcap_{A: A\subseteq \boldsymbol{s}, A\neq\varnothing} \mathscr{B}_{n,\varepsilon}(A)^c. \tag{3.1}$$

Consider the following truncation of $T(H, G_n)$ (recall (1.2)):

$$T_\varepsilon^+(H, G_n) = \frac{1}{|\mathrm{Aut}(H)|} \sum_{\boldsymbol{s}\in V(G_n)_{|V(H)|}} \prod_{(i,j)\in E(H)} a_{s_i s_j} X_{\boldsymbol{s}} \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s})^c\}. \tag{3.2}$$

Moreover, let $N_\varepsilon^+(H, G_n) := \frac{1}{p_n^{|V(H)|}} \mathbb{E}[T_\varepsilon^+(H, G_n)]$ be the truncation of the true motif count $N(H, G_n)$. This truncation has the following properties:

LEMMA 3.1. *Define*

$$M_n := \sum_{\substack{A\subset V(G_n)\\ 1\leq |A|\leq |V(H)|}} t_H(A)\mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}.$$

Then the following hold:

(a) $\frac{M_n}{2^{|V(H)|}-1} \leq N(H, G_n) - N_\varepsilon^+(H, G_n) \leq M_n$.

(b) $\mathbb{P}(T(H,G_n) \neq T_\varepsilon^+(H,G_n)) \leq \frac{M_n}{\varepsilon N(H,G_n)}$.

PROOF. Note that

$$\Delta_n := N(H,G_n) - N_\varepsilon^+(H,G_n) = \frac{1}{|\operatorname{Aut}(H)|} \sum_{\boldsymbol{s} \in V(G_n)_{|V(H)|}} \prod_{(i,j)\in E(H)} a_{s_i s_j} \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s})\}.$$

Since $\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}) = \bigcup_{A: A \subseteq \boldsymbol{s}, A \neq \varnothing} \mathscr{B}_{n,\varepsilon}(A)$ is the union of $2^{|V(H)|} - 1$ many sets, applying the elementary inequality

$$\frac{1}{m} \sum_{r=1}^{m} \mathbf{1}\{B_r\} \leq \mathbf{1}\left\{\bigcup_{r=1}^{m} B_r\right\} \leq \sum_{r=1}^{m} \mathbf{1}\{B_r\},$$

for any finite collection of sets $B_1, B_2, \ldots, B_m$, gives

$$\frac{M_n}{2^{|V(H)|} - 1} \leq \Delta_n \leq M_n,$$

with

$$\begin{aligned} M_n &= \frac{1}{|\operatorname{Aut}(H)|} \sum_{\boldsymbol{s} \in V(G_n)_{|V(H)|}} \sum_{\substack{A \subseteq \boldsymbol{s} \\ 1 \leq |A| \leq |V(H)|}} \prod_{(i,j)\in E(H)} a_{s_i s_j} \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\} \\ &= \sum_{\substack{A \subset V(G_n) \\ 1 \leq |A| \leq |V(H)|}} t_H(A) \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}, \end{aligned}$$

where last equality follows by interchanging the order of the sum and recalling the definition of $t_H(A)$ in (2.2). This proves the result in (a).

We now proceed to prove (b). For any $A \subset V(G_n)$, define $X_A := \prod_{u \in A} X_u$. Hence, recalling definitions (1.2) and (3.2) gives

$$\begin{aligned} \mathbb{P}(T(H,G_n) \neq T_\varepsilon^+(H,G_n)) &\leq \mathbb{E}[T(H,G_n) - T_\varepsilon^+(H,G_n)] \\ &\leq \sum_{\substack{A \subset V(G_n) \\ 1 \leq |A| \leq |V(H)|}} \mathbb{P}(X_A = 1) \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\} \\ &\leq \sum_{\substack{A \subset V(G_n) \\ 1 \leq |A| \leq |V(H)|}} p_n^{|A|} \mathbf{1}\{t_H(A) > \varepsilon p_n^{|A|} N(H,G_n)\} \\ &\leq \frac{1}{\varepsilon N(H,G_n)} \sum_{\substack{A \subset V(G_n) \\ 1 \leq |A| \leq |V(H)|}} t_H(A) \mathbf{1}\{t_H(A) > \varepsilon p_n^{|A|} N(H,G_n)\} \\ &\leq \frac{M_n}{\varepsilon N(H,G_n)} \end{aligned}$$

This completes the proof of (b). □

PROOF OF THEOREM 2.1 (SUFFICIENCY). Recall that condition (2.3) assumes $\frac{M_n}{N(H,G_n)} \to 0$, where M_n is as defined above in Lemma 3.1. Therefore, Lemma 3.1 and the condition in (2.3) together implies that

$$\text{(3.3)} \quad \frac{\mathbb{E}[T_\varepsilon^+(H,G_n)]}{\mathbb{E}[T(H,G_n)]} = \frac{N_\varepsilon^+(H,G_n)}{N(H,G_n)} \to 1 \quad \text{and} \quad \mathbb{P}(T(H,G_n) = T_\varepsilon^+(H,G_n)) \to 1,$$

as $n \to \infty$, for every fixed $\varepsilon > 0$. Now, write

$$\frac{\widehat{N}(H, G_n)}{\mathbb{E}[\widehat{N}(H, G_n)]} = \frac{T(H, G_n)}{\mathbb{E}[T(H, G_n)]} = \frac{T(H, G_n)}{T_\varepsilon^+(H, G_n)} \cdot \frac{T_\varepsilon^+(H, G_n)}{\mathbb{E}[T_\varepsilon^+(H, G_n)]} \cdot \frac{\mathbb{E}[T_\varepsilon^+(H, G_n)]}{\mathbb{E}[T(H, G_n)]}.$$

Note that, by (3.3), the first and the third ratios in the RHS above converge to 1 in probability for every fixed ε. Therefore, to prove the consistency of $\widehat{N}(H, G_n)$ it suffices to show that the ratio $\frac{T_\varepsilon^+(H,G_n)}{\mathbb{E}[T_\varepsilon^+(H,G_n)]} \xrightarrow{P} 1$, as $n \to \infty$ followed by $\varepsilon \to 0$. This follows by the using Chebyshev's inequality if we show that

$$\lim_{\varepsilon \to 0} \lim_{n \to \infty} \frac{\operatorname{Var}[T_\varepsilon^+(H, G_n)]}{(\mathbb{E}[T_\varepsilon^+(H, G_n)])^2} = 0. \tag{3.4}$$

To this effect, we have

$$\begin{aligned}
&\operatorname{Var}\bigl[T_\varepsilon^+(H, G_n)\bigr] \\
&\quad = \frac{1}{|\operatorname{Aut}(H)|^2} \sum_{\substack{\boldsymbol{s}_1, \boldsymbol{s}_2 \in V(G_n)_{|V(H)|} \\ \overline{\boldsymbol{s}}_1 \cap \overline{\boldsymbol{s}}_2 \neq \varnothing}} \operatorname{Cov}(X_{\boldsymbol{s}}, X_{\boldsymbol{s}_2}) M_H(\boldsymbol{s}_1) M_H(\boldsymbol{s}_2) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_2)^c\bigr\}.
\end{aligned}$$

Now, if $|\overline{\boldsymbol{s}}_1 \cap \overline{\boldsymbol{s}}_2| = K$, then $\operatorname{Cov}[X_{\boldsymbol{s}_1}, X_{\boldsymbol{s}_2}] = p_n^{2|V(H)|-K} - p_n^{2|V(H)|} \leq p_n^{2|V(H)|-K}$. Thus,

$$\begin{aligned}
&\operatorname{Var}\bigl[T_\varepsilon^+(H, G_n)\bigr] \\
&\quad \leq \frac{1}{|\operatorname{Aut}(H)|^2} \sum_{K=1}^{|V(H)|} p_n^{2|V(H)|-K} \\
&\qquad \times \sum_{\substack{\boldsymbol{s}_1, \boldsymbol{s}_2 \in V(G_n)_{|V(H)|} \\ K = |\overline{\boldsymbol{s}}_1 \cap \overline{\boldsymbol{s}}_2|}} M_H(\boldsymbol{s}_1) M_H(\boldsymbol{s}_2) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_2)^c\bigr\}.
\end{aligned} \tag{3.5}$$

We now focus on the inner sum in the right-hand side of (3.5). Note that

$$\begin{aligned}
&\sum_{\substack{\boldsymbol{s}_1, \boldsymbol{s}_2 \in V(G_n)_{|V(H)|} \\ K = |\overline{\boldsymbol{s}}_1 \cap \overline{\boldsymbol{s}}_2|}} M_H(\boldsymbol{s}_1) M_H(\boldsymbol{s}_2) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_2)^c\bigr\} \\
&\quad = \sum_{\substack{A \subset V(G_n) \\ |A| = K}} \sum_{\substack{\boldsymbol{s}_1, \boldsymbol{s}_2 \in V(G_n)_{|V(H)|} \\ \overline{\boldsymbol{s}}_1 \cap \overline{\boldsymbol{s}}_2 = A}} M_H(\boldsymbol{s}_1) M_H(\boldsymbol{s}_2) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_2)^c\bigr\} \\
&\quad \leq \sum_{\substack{A \subset V(G_n) \\ |A| = K}} \sum_{\boldsymbol{s}_1 : \overline{\boldsymbol{s}}_1 \supseteq A} \sum_{\boldsymbol{s}_2 : \overline{\boldsymbol{s}}_2 \supseteq A} M_H(\boldsymbol{s}_1) M_H(\boldsymbol{s}_2) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_2)^c\bigr\}.
\end{aligned} \tag{3.6}$$

The argument inside the sum now separates out. Therefore, applying the fact

$$\sum_{\boldsymbol{s}_1 : \overline{\boldsymbol{s}}_1 \supseteq A} M_H(\boldsymbol{s}_1) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s}_1)^c\bigr\} \leq \sum_{\boldsymbol{s}_1 : \overline{\boldsymbol{s}}_1 \supseteq A} M_H(\boldsymbol{s}_1) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(A)^c\bigr\} = \bigl|\operatorname{Aut}(H)\bigr| t_H(A) \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(A)^c\bigr\},$$

it follows from (3.5) and (3.6) that

$$\operatorname{Var}\bigl[T_\varepsilon^+(H, G_n)\bigr] \leq \sum_{K=1}^{|V(H)|} p_n^{2|V(H)|-K} \sum_{\substack{A \subset V(G_n) \\ |A| = K}} t_H(A)^2 \mathbf{1}\bigl\{\mathscr{B}_{n,\varepsilon}(A)^c\bigr\}$$

(since $t_H(A) \le \varepsilon p_n^{|A|} N(H, G_n)$ on $\mathscr{B}_{n,\varepsilon}(A)^c$)

$$\begin{aligned}
&\le \varepsilon N(H, G_n) \sum_{K=1}^{|V(H)|} p_n^{2|V(H)|} \sum_{\substack{A \subset V(G_n)\\ |A|=K}} t_H(A)\mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)^c\}\\
&\le \varepsilon p_n^{2|V(H)|} N(H, G_n) \sum_{K=1}^{|V(H)|} \sum_{\substack{A \subset V(G_n)\\ |A|=K}} t_H(A)\\
&\varepsilon p_n^{2|V(H)|} N(H, G_n) \sum_{K=1}^{|V(H)|} \binom{|V(H)|}{K} N(H, G_n)\\
&= \varepsilon p_n^{2|V(H)|} \big(2^{|V(H)|} - 1\big) N(H, G_n)^2,
\end{aligned}$$

where the last line uses (2.5). Since (3.4) is immediate from this, we have verified sufficiency. □

PROOF OF THEOREM 2.1 (NECESSITY). We will show the contrapositive statement, that is, if (2.3) fails, then $\widehat{N}(H, G_n)$ is not consistent for $N(H, G_n)$. Toward this, assume (2.3) fails. Define

$$E_1 := \big\{X_{\boldsymbol{s}} = 0 \text{ for all } \boldsymbol{s} \in V(G_n)_{|V(H)|} \text{ with } \mathbf{1}\big\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s})^c\big\} = 0\big\}, \tag{3.7}$$

and, for $1 \le K \le |V(H)|$, let

$$E_{2,K} = \bigg\{ X_A := \prod_{u \in A} X_u = 0 \text{ for all } A \subset V(G_n) \text{ where } |A| = K \text{ and } \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\} = 1 \bigg\}.$$

Take any $\boldsymbol{s} \in V(G_n)_{|V(H)|}$ with $\mathbf{1}\{\mathscr{B}_{n,\varepsilon}(\boldsymbol{s})^c\} = 0$. By definition (recall (3.1)), this implies $\mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\} = 1$ for some $A \subseteq \boldsymbol{s}$, $A \neq \varnothing$. In particular, under the event $\bigcap_{K=1}^{|V(H)|} E_{2,K}$, we have $X_A = 0$, forcing $X_{\boldsymbol{s}} = 0$. Hence, $E_1 \supset \bigcap_{K=1}^{|V(H)|} E_{2,K}$. Note that

$$E_{2,K} = \bigcap_{K=1}^{|V(H)|} \bigcap_{\substack{A \subset V(G_n): |A|=K\\ \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}=1}} \{X_A = 0\},$$

and the event $\{X_A = 0\}$ is a decreasing event, for all $A \subset V(G_n)$ with $1 \le |A| \le |V(H)|$.[2] Then the FKG inequality between decreasing events for product measures on $\{0, 1\}^{|V(G_n)|}$ [30], Chapter 2, gives

$$\begin{aligned}
\mathbb{P}(E_1) \ge \mathbb{P}\bigg(\bigcap_{K=1}^{|V(H)|} E_{2,K}\bigg) &\ge \prod_{K=1}^{|V(H)|} \prod_{\substack{A \subset V(G_n): |A|=K\\ \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}=1}} \mathbb{P}(X_A = 0)\\
&\ge \prod_{K=1}^{|V(H)|} \big(1 - p_n^K\big)^{\sum_{A \subset V(G_n): |A|=K} \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}}
\end{aligned}$$

[2] An event $\mathcal{D} \subseteq \{0, 1\}^{|V(G_n)|}$ is said to be *decreasing* if for two vectors $\boldsymbol{x} = (x_a)_{a \in V(G_n)} \in \{0, 1\}^{|V(G_n)|}$ and $\boldsymbol{y} = (y_a)_{a \in V(G_n)} \in \{0, 1\}^{|V(G_n)|}$, with $\{a : y_a = 1\} \subseteq \{a : x_a = 1\}$, $\boldsymbol{x} \in \mathcal{D}$ implies $\boldsymbol{y} \in \mathcal{D}$. Then the FKG inequality states that if $\mathcal{D}_1, \mathcal{D}_2 \subseteq \{0, 1\}^{|V(G_n)|}$ are two decreasing events, $\mathbb{P}(\mathcal{D}_1 \cap \mathcal{D}_2) \ge \mathbb{P}(\mathcal{D}_1)\mathbb{P}(\mathcal{D}_2)$ (see [30], Chapter 2).

Now, since p_n is bounded away from 1 (recall (2.1)), there exists a constant $c > 0$ such that $\log(1 - p_n^K) > -cp_n^K$, for all $1 \leq K \leq |V(H)|$. Hence,

$$\begin{aligned}
\mathbb{P}(E_1) &\geq \exp\left(-c \sum_{K=1}^{|V(H)|} p_n^{|K|} \sum_{A \subset V(G_n):|A|=K} \mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\}\right) \\
&\geq \exp\left(-\frac{c}{\varepsilon N(H, G_n)} \sum_{K=1}^{|V(H)|} \sum_{A \subset V(G_n):|A|=K} t_H(A)\right) \qquad (3.8) \\
&\geq e^{-\frac{c(2^{|V(H)|}-1)}{\varepsilon}},
\end{aligned}$$

where the last step uses (2.5). Now, since (2.3) does not hold, there exists $\varepsilon > 0$ and $\delta \in (0, 1)$ such that

$$\limsup_{n\to\infty} \frac{1}{N(H, G_n)} \sum_{K=1}^{|V(H)|} \sum_{A \subset V(G_n):|A|=K} t_H(A)\mathbf{1}\{\mathscr{B}_{n,\varepsilon}(A)\} > \left(2^{|V(H)|} - 1\right)\frac{2\delta}{1+\delta}.$$

From Lemma 3.1, it follows that along a subsequence, $N(H, G_n) - N_\varepsilon^+(H, G_n) > \frac{2\delta}{1+\delta} N(H, G_n)$, that is, $(1 + \delta)N_\varepsilon^+(H, G_n) < (1 - \delta)N(H, G_n)$. Thus, by Markov inequality, along a subsequence

$$\begin{aligned}
&\mathbb{P}\left(T_\varepsilon^+(H, G_n) \geq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\right) \\
&\quad\leq \mathbb{P}\left(T_\varepsilon^+(H, G_n) \geq (1 + \delta)p_n^{|V(H)|}N_\varepsilon^+(H, G_n)\right) \qquad (3.9) \\
&\quad\leq \frac{1}{1+\delta}.
\end{aligned}$$

Also, observe that $\{T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\}$ is a decreasing event, because if $\boldsymbol{X} = (X_a)_{a\in V(G_n)} \in \{T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\}$ then any vector $\boldsymbol{X}' = (X'_a)_{a\in V(G_n)}$ obtained changing a subset of the ones in $\boldsymbol{X}$ to zeros does not increase the value of $T_\varepsilon^+(H, G_n)$, and hence, $\boldsymbol{X}' \in \{T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\}$. Similarly, E_1 (recall definition in (3.7)) is a decreasing event. Hence, by the FKG inequality,

$$\begin{aligned}
\mathbb{P}\left(T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)|E_1\right) &\geq \mathbb{P}\left(T_\varepsilon^+(H, G_n)\right. \qquad (3.10) \\
&\quad\left.\leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\right).
\end{aligned}$$

This implies

$$\begin{aligned}
&\mathbb{P}\left(\widehat{N}(H, G_n) \leq (1 - \delta)N(H, G_n)\right) \\
&\qquad\geq \mathbb{P}\left(T(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)|E_1\right)\mathbb{P}(E_1) \\
&\left(\text{since } T(H, G_n) \geq T_\varepsilon^+(H, G_n)\right) \geq \mathbb{P}\left(T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)|E_1\right)\mathbb{P}(E_1) \\
&\left(\text{by (3.10)}\right) \geq \mathbb{P}\left(T_\varepsilon^+(H, G_n) \leq (1 - \delta)p_n^{|V(H)|}N(H, G_n)\right)\mathbb{P}(E_1) \\
&\qquad\geq \frac{\delta}{1+\delta}\mathbb{P}(E_1),
\end{aligned}$$

where the last step uses (3.9). This is a contradiction to the consistency of $\widehat{N}(H, G_n)$, since $\liminf_{n\to\infty} \mathbb{P}(E_1) > 0$ by (3.8), completing the proof of the desired result. □

4. Discussions and future directions. The theme that emerges from the examples considered in the paper is that in most of the natural network models, the HT estimator $\widehat{N}(H, G_n)$ is consistent and asymptotically normal whenever the expected number of copies of H in the sampled graph diverges, and inconsistent and not asymptotically normal otherwise. For dense graphs (graphons), this implies sampling at rate $p_n \gg 1/|V(G_n)|$ ensures that the HT estimator is consistent and asymptotically normal. For sparser graphs, one needs to sample at rate $p_n \gg N(H, G_n)^{-\frac{1}{|V(H)|}}$, which can be much larger, depending on the magnitude of $N(H, G_n)$. In particular, this implies that there is a nontrivial sampling rate beyond which the HT estimator is consistent for sparser graphs (even for bounded degree graphs), as soon as the number of copies of H in G_n is diverging. An interesting question is whether under this assumption ($N(H, G_n) \to \infty$), it is possible to improve the estimation accuracy of $N(H, G_n)$ using other sampling strategies, such as neighborhood sampling [32, 36, 38], snowball sampling [28] or random walk based exploration methods [41, 51]. However, not much is known about the asymptotic fluctuations of the resulting estimates in these sampling models. In fact, it has been shown recently in [36] that the natural inverse probability weighted estimator might not be minimax optimal in the neighborhood sampling scheme. Therefore, it is encouraging to see that the HT estimator in the simple (albeit idealized) subgraph sampling model provides consistent and asymptotically exact confidence intervals for large classes of natural network models. These results are the first steps toward understanding properties of more practical (and complicated) models for network sampling, and will provide useful benchmarks for comparing the performances of different estimates arising from other sampling schemes.

From a computational perspective, the subgraph sampling scheme has time complexity $O(|V(G_n)|)$. Since on average the sampled graph as $O(p_n|V(G_n)|)$ vertices, one way to reduce the computational cost is to sample without replacement a uniform random subset of size $N = p_n|V(G_n)|$ from $V(G_n)$, and then consider the induced graph as before. This can be done in $O(N \log N)$ time [31, 52], which is faster whenever $N \ll |V(G_n)|$ (up to a logarithmic factor). In certain situations, the asymptotic properties of the HT estimator in the sampling without replacement model should be the same as that in the subgraph sampling model with sampling probability $p_n = N/|V(G_n)|$. For example, we conjecture that using [19], Theorem 4, one should be able to derive consistency of the HT estimator in the sampling without replacement model, at least for certain regimes of p_n. In a similar manner, using the asymptotic normality for the HT estimator in the subgraph sampling model along with the conditional approach in [6], one should be able verify a similar result for the sampling without replacement model in certain regimes of p_n as well. The exact detection boundary of the sampling without replacement model seems to be an interesting question for possible future research.

Acknowledgments. The authors thank Sohom Bhattacharya for pointing out [35], and Jason Klusowski for helpful discussions. The authors also thank the Associate Editor and the anonymous referees for their detailed and thoughtful comments, which greatly improved the quality and the presentation of the paper.

Funding. The first author was supported by NSF CAREER Grant DMS-2046393 and a Sloan Research Fellowship.

The third author was supported by NSF Grant DMS-1712037.

SUPPLEMENTARY MATERIAL

Supplement to "Motif estimation via subgraph sampling: The fourth-moment phenomenon" (DOI: 10.1214/21-AOS2134SUPP; .pdf). Proofs of the main results and additional examples are given in the supplementary materials.

REFERENCES

[1] AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

[2] 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

[3] 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. https://doi.org/10.1038/nature10736

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

[5] BARBOUR, A. D. and CHEN, L. H. Y., eds. (2005). *An Introduction to Stein's Method. Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore* **4**. World Scientific, Hackensack, NJ. MR2235447 https://doi.org/10.1142/9789812567680

[6] BERTAIL, P., CHAUTRU, E. and CLÉMENÇON, S. (2017). Empirical processes in survey sampling with (conditional) Poisson designs. *Scand. J. Stat.* **44** 97–111. MR3619696 https://doi.org/10.1111/sjos.12243

[7] BHATTACHARYA, A., BISHNU, A., GHOSH, A. and MISHRA, G. (2019). Triangle estimation using tripartite independent set queries. In 30*th International Symposium on Algorithms and Computation* (*ISAAC* 2019). Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. MR4042092

[8] BHATTACHARYA, B. B, DAS, S. and MUKHERJEE, S. (2022). Supplement to "Motif estimation via subgraph sampling: The fourth-moment phenomenon." https://doi.org/10.1214/21-AOS2134SUPP

[9] BHATTACHARYA, B. B., DIACONIS, P. and MUKHERJEE, S. (2017). Universal limit theorems in graph coloring problems with connections to extremal combinatorics. *Ann. Appl. Probab.* **27** 337–394. MR3619790 https://doi.org/10.1214/16-AAP1205

[10] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 2280–2301. MR2906868 https://doi.org/10.1214/11-AOS904

[11] 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

[12] BORGS, C., CHAYES, J. T., LOVÁSZ, L., SÓS, V. T. and VESZTERGOMBI, K. (2012). Convergent sequences of dense graphs II. Multiway cuts and statistical physics. *Ann. of Math.* (2) **176** 151–219. MR2925382 https://doi.org/10.4007/annals.2012.176.1.2

[13] 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

[14] CHANDRASEKHAR, A. and LEWIS, R. (2011). Econometrics of sampled networks.

[15] CHATTERJEE, S. and DIACONIS, P. (2013). Estimating and understanding exponential random graph models. *Ann. Statist.* **41** 2428–2461. MR3127871 https://doi.org/10.1214/13-AOS1155

[16] CHATTERJEE, S., DIACONIS, P. and SLY, A. (2011). Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21** 1400–1435. MR2857452 https://doi.org/10.1214/10-AAP728

[17] CHEN, L. H. Y. and SHAO, Q.-M. (2004). Normal approximation under local dependence. *Ann. Probab.* **32** 1985–2028. MR2073183 https://doi.org/10.1214/009117904000000450

[18] CRANE, H. (2018). *Probabilistic Foundations of Statistical Network Analysis. Monographs on Statistics and Applied Probability* **157**. CRC Press, Boca Raton, FL. MR3791467

[19] DIACONIS, P. and FREEDMAN, D. (1980). Finite exchangeable sequences. *Ann. Probab.* **8** 745–764. MR0577313

[20] EDEN, T., LEVI, A., RON, D. and SESHADHRI, C. (2017). Approximately counting triangles in sublinear time. *SIAM J. Comput.* **46** 1603–1646. MR3709896 https://doi.org/10.1137/15M1054389

[21] FEIGE, U. (2006). On sums of independent random variables with unbounded variance and estimating the average degree in a graph. *SIAM J. Comput.* **35** 964–984. MR2203734 https://doi.org/10.1137/S0097539704447304

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

[23] 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

[24] GAO, C. and MA, Z. (2021). Minimax rates in network analysis: Graphon estimation, community detection and hypothesis testing. *Statist. Sci.* **36** 16–33. MR4194201 https://doi.org/10.1214/19-STS736

[25] 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

[26] GOLDREICH, O. and RON, D. (2011). On testing expansion in bounded-degree graphs. In *Studies in Complexity and Cryptography. Miscellanea on the Interplay Between Randomness and Computation* 68–75. Springer.

[27] GONEN, M., RON, D. and SHAVITT, Y. (2011). Counting stars and other small subgraphs in sublinear-time. *SIAM J. Discrete Math.* **25** 1365–1411. MR2837605 https://doi.org/10.1137/100783066

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

[29] GOVINDAN, R. and TANGMUNARUNKIT, H. (2000). Heuristics for Internet map discovery. In *Proceedings IEEE INFOCOM* 2000. *Conference on Computer Communications*. *Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies* **3** 1371–1380. IEEE, New York.

[30] GRIMMETT, G. (1999). *Percolation*, 2nd ed. *Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **321**. Springer, Berlin. MR1707339 https://doi.org/10.1007/978-3-662-03981-6

[31] GUPTA, P. and BHATTACHARJEE, G. P. (1984). An efficient algorithm for random sampling without replacement. *Int. J. Comput. Math.* **16** 201–209. MR0775085 https://doi.org/10.1080/00207168408803438

[32] 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

[33] 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

[34] JANSON, S., ŁUCZAK, T. and RUCINSKI, A. (2000). *Random Graphs*. *Wiley-Interscience Series in Discrete Mathematics and Optimization* **45**. Wiley Interscience, New York. MR1782847 https://doi.org/10.1002/9781118032718

[35] KIM, J. H., SUDAKOV, B. and VU, V. (2007). Small subgraphs of random regular graphs. *Discrete Math.* **307** 1961–1967. MR2320201 https://doi.org/10.1016/j.disc.2006.09.032

[36] KLUSOWSKI, J. M. and WU, Y. (2018). Counting motifs with graph sampling. In *Conference on Learning Theory* 1966–2011.

[37] KLUSOWSKI, J. M. and WU, Y. (2020). Estimating the number of connected components in a graph via subgraph sampling. *Bernoulli* **26** 1635–1664. MR4091087 https://doi.org/10.3150/19-BEJ1147

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

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

[40] KRIVELEVICH, M., SUDAKOV, B., VU, V. H. and WORMALD, N. C. (2001). Random regular graphs of high degree. *Random Structures Algorithms* **18** 346–363. MR1839497 https://doi.org/10.1002/rsa.1013

[41] LESKOVEC, J. and FALOUTSOS, C. (2006). Sampling from large graphs. In *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 631–636.

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

[43] MARINUCCI, D. and PECCATI, G. (2011). *Random Fields on the Sphere*: *Representation*, *Limit Theorems and Cosmological Applications*. *London Mathematical Society Lecture Note Series* **389**. Cambridge Univ. Press, Cambridge. MR2840154 https://doi.org/10.1017/CBO9780511751677

[44] MILO, R., SHEN-ORR, S., ITZKOVITZ, S., KASHTAN, N., CHKLOVSKII, D. and ALON, U. (2002). Network motifs: Simple building blocks of complex networks. *Science* **298** 824–827.

[45] NOURDIN, I. and PECCATI, G. (2009). Stein's method on Wiener chaos. *Probab. Theory Related Fields* **145** 75–118. MR2520122 https://doi.org/10.1007/s00440-008-0162-x

[46] NOURDIN, I. and PECCATI, G. (2012). *Normal Approximations with Malliavin Calculus*: *From Stein's Method to Universality*. *Cambridge Tracts in Mathematics* **192**. Cambridge Univ. Press, Cambridge. MR2962301 https://doi.org/10.1017/CBO9781139084659

[47] NOURDIN, I., PECCATI, G. and REINERT, G. (2010). Invariance principles for homogeneous sums: Universality of Gaussian Wiener chaos. *Ann. Probab.* **38** 1947–1985. MR2722791 https://doi.org/10.1214/10-AOP531

[48] NUALART, D. and PECCATI, G. (2005). Central limit theorems for sequences of multiple stochastic integrals. *Ann. Probab.* **33** 177–193. MR2118863 https://doi.org/10.1214/009117904000000621

[49] OVE, F. (2005). Network sampling and model fitting. In *Structural Analysis in the Social Sciences* 31–56. Cambridge Univ. Press, Cambridge.

[50] PRŽULJ, N., CORNEIL, D. G. and JURISICA, I. (2004). Modeling interactome: Scale-free or geometric? *Bioinformatics* **20** 3508–3515.

[51] RIBEIRO, B. and TOWSLEY, D. (2010). Estimating and sampling graphs with multidimensional random walks. In *Proceedings of the* 10*th ACM SIGCOMM Conference on Internet Measurement* 390–403.

[52] ROUZANKIN, P. S. and VOYTISHEK, A. V. (1999). On the cost of algorithms for random selection. *Monte Carlo Methods Appl*. **5** 39–54. MR1684992 https://doi.org/10.1515/mcma.1999.5.1.39

[53] RUAL, J.-F., VENKATESAN, K., HAO, T., HIROZANE-KISHIKAWA, T., DRICOT, A., LI, N., BERRIZ, G. F., GIBBONS, F. D., DREZE, M. et al. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437** 1173–1178.

[54] STEIN, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California*, *Berkeley*, *Calif.*, 1970/1971), *Vol. II*: *Probability Theory* 583–602. MR0402873

[55] TANG, M., SUSSMAN, D. L. and PRIEBE, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *Ann. Statist*. **41** 1406–1430. MR3113816 https://doi.org/10.1214/13-AOS1112

[56] TILLÉ, Y. (2006). *Sampling Algorithms*. *Springer Series in Statistics*. Springer, New York. MR2225036

[57] UETZ, P., GIOT, L., CAGNEY, G., MANSFIELD, T. A., JUDSON, R. S., KNIGHT, J. R., LOCKSHON, D., NARAYAN, V., SRINIVASAN, M. et al. (2000). A comprehensive analysis of protein–protein interactions in saccharomyces cerevisiae. *Nature* **403** 623–627.

[58] YANG, J. and LESKOVEC, J. (2015). Defining and evaluating network communities based on ground-truth. *Knowl*. *Inf*. *Syst*. **42** 181–213.

[59] ZHANG, Y., KOLACZYK, E. D. and SPENCER, B. D. (2015). Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Ann*. *Appl*. *Stat*. **9** 166–199. MR3341112 https://doi.org/10.1214/14-AOAS800