SUBSAMPLING BOOTSTRAP OF COUNT FEATURES OF NETWORKS

By SHARMODEEP BHATTACHARYYA*† and PETER J. BICKEL‡†

University of California, Berkeley‡†

Abstract: Analysis of stochastic models of networks is quite important in light of the huge influx of network data in social, information and bio sciences. But a proper statistical analysis of features of different stochastic models of networks is still underway. We propose bootstrap methods for finding empirical distribution of count features or ‘moments’ (Bickel, Chen & Levina, AoS, 2011) and smooth functions of these for the networks. Using these methods, we can not only estimate variance of count features but also get good estimates of such feature counts, which are usually expensive to compute numerically in large networks. In our paper, we prove theoretical properties of the bootstrap variance estimates of the count features as well as show their efficacy through simulation. We also use the method on publicly available Facebook network data for estimate of variance and expectation of some count features.

1. Introduction. The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists, computer scientists and mathematicians. With the information boom, a huge number of network data sets have come into prominence. In biology - gene transcription networks, protein-protein interaction network, in social media - Facebook, Twitter, Linkedin networks, information networks arising in connection with text mining, technological networks such as the Internet, ecological and epidemiological networks and many others have come to the forefront. Although the study of networks has a long history in physics and mathematics literature and informal methods of analysis have arisen in many fields of application, statistical inference on network models as opposed to descriptive statistics, empirical modeling and some Bayesian approaches [25] [21] [16] has not been addressed extensively in literature. A mathematical and systematic study of statistical inference on network models has only started in recent years.

*Supported in part by NSF grant
†Supported in part by NSF grant

AMS 2000 subject classifications: Primary 62F40, 62G09; secondary 62D05

Keywords and phrases: Networks, Subsampling, Bootstrap, Count features, Model-based sampling
Frequentist statistical inference involves proposing random models, fitting the proposed model to the data, checking goodness of fit in nonparametric context and given a good fit constructing tests and confidence statements about features of the model. Systematic analysis of complex models can only be done asymptotically and validated by simulation and network models are no exception. Much recent analysis has focussed on block models [17] and exponential random graph models (ERGM) [14]. The block models in their first incarnation did not fit large graphs well, for instance, their exponential degree distribution did not fit empirically observed degree distributions which often seemed to be of power law type [3] [26]. But they serve as shown by Bickel and Chen (2009) [6] the role of histograms. They also have until recently (Amini et. al. (2012) [2], Daudin et. al. (2008) [11]) proved hard to fit well. Nevertheless their analysis proceeds apace [24] [7] [28] [10]. For ERGM also there has been some work done on likelihood inference [29] [19] for these models. But recently Bhamidi et al [5] has shown some issues with these models, as most of the time these models fail to represent real-world network properties. We will not dwell on these issues.

It follows from the work of Lovász [23] [9] and Aldous [1] and Hoover [18] that there is a representation of all probability models on \( n \) vertices which can be embedded in an infinite vertex model with natural invariance properties. This leads to the Bickel and Chen (2009) [6] characterization of “nonparametric” unlabeled graph models which is closely related to Lovász’s notion of “graphons”. It also follows from the work of Lovász [23], Diaconis and Janson [12] and in part from Bickel and Chen [6] that there is a unique set of statistics whose joint distribution characterize the probability distribution on graphs. These statistics, called “empirical moments” by Bickel, Chen and Levina [7], have appeared in various literatures earlier under the names of “motif” counts in biology [20], “subgraph” counts in probability [23]. Examples are the number of edges, the number of ‘V’s, the number of triangles contained in the observed graph.

The expectation and variances of the quantities can, in principle be computed (Picard et.al. [27]) and more usefully be asymptotically approximated [7] and under appropriate conditions these have limiting Gaussian distribution. They have many uses [33] [32] [4], particularly in distinguishing between the mechanisms generating different graphs as well as providing characterization, but in an outward form of the probability distribution.

A major stumbling block in their use has been the calculation of motifs that have more than 4 or 5 members. They have been used in testing equality of two distributions of count statistics and finding confidence intervals. Another problem that arises in dealing with the count statistics for
large numbers is actually computing the count statistics. Finding the correct count statistics is a computationally hard problem for large networks as the complexity of finding the count of a subgraph is polynomial in terms of number of vertices and when number of vertices is in millions, the computation becomes infeasible.

In the statistical literature on networks, some work has been done on devising sampling designs to select network samples. Various sampling designs has been proposed at different points in the statistical literature to derive meaningful samples of a given network. Kolackzyk (2009) and [21] contains a nice summary of network sampling designs. Examples of such sampling designs include random node selection, induced and incident sampling, star and snowball sampling, link-tracing sampling, random walks, forest fire and several modifications of the stated methods [21] [22] [30]. Many of these sampling designs have been analyzed from design-based sampling point of view [31] [13]. There has also been work done on analyzing some of these methods from model-based sampling point of view, where, mostly the exponential random graph model (ERGM) was considered as the model generating the network and a likelihood-based approach was taken for inference [15]. As a result only parametric inference was possible in those approaches. On the other hand, our approach is not restricted to parametric models as we try to estimate the certain functionals of the underlying generating model, using the samples obtained from the random population network. So, in our work, we consider a “nonparametric” model as the underlying model in our analysis and try to see both theoretically and by examples how some of these sampling schemes perform in estimating count features and their asymptotic variances.

1.1. Contribution and Structure of Our Work. We use subsampling based bootstrap approaches to estimate the count statistics as well as find approximate distribution for such count statistics under the general model of Bickel and Chen [6]. We also state certain properties under which a network sampling design becomes adaptive to the network model for count statistics. By adaptive, we mean here that the network sampling design produces subsamples of network, such that the count statistics obtained from the subsampled network, becomes consistent to the original count statistics for the whole network.

We also apply our bootstrap method in simulated networks as well as two real-life networks. In simulation, we use two different models, stochastic block models [17] and preferential attachment models [3]. We try to compare the performance of different bootstrap methods for each of the simulated
networks as well as to compare the networks generated by these two models using some well-known descriptive statistics of networks [21]. One of the real-life networks is the Jefferson high-school network given in Bearman et al. (2004) [4] and the others are the Facebook collegiate networks provided in Traud et al. (2010) [32]. For the high-school network, we try to answer the question whether the number of small-cycles in the network is small. For the Facebook collegiate networks, we try to decide whether the node covariates given for the network have any potential clustering power. We also try to distinguish two different networks based on the partitioning properties. The test-statistics that we use in these comparisons is network transitivity, which has been argued to indicate network clustering capability [21].

In section 2 we outline our main results. In section 3 we describe the bootstrap subsampling methods and the theoretical properties of each bootstrap estimators. We also indicate a method for estimating asymptotic variances of these estimators using bootstrap. We also give a theoretical comparison of the methods. In section 5 we perform simulation under two special cases of the general “nonparametric” model: stochastic block model and preferential attachment model respectively. Under each of these cases we try to estimate count statistics and their variances with the the help of the bootstrap subsampling schemes and we compare the empirical performance of the three proposed bootstrap subsampling schemes as well as perform tests for model mis-specification. In 6 we apply our method to test hypotheses about the count statistics of the real network.

2. Main Results. Let us consider that a random graph $G_n$ as the data. Let $V(G_n) = \{v_1, \ldots, v_n\}$ denote the vertices of $G_n$ and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of $G_n$. So, the number of vertices in $G_n$ is $|V(G_n)| = n$ and number of edges of $G_n$ is $|E(G_n)| = m$. Let the adjacency matrix of $G_n$ be denoted by $A_{n \times n}$. For sake of notational simplicity, from here onwards we shall denote $G_n$ by $G$ having $n$ vertices unless specifically mentioned.

We consider a general non-parametric model, as described in Bickel, Chen and Levina (2011) [7], generates the random data network $G$. The general non-parametric model can be described by the following equation -

\[
\Pr(A_{ij} = 1 | \xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v) 1(w \leq \rho_n^{-1}),
\]

where, $w(u, v) \geq 0$, symmetric, $0 \leq u, v \leq 1$, $\rho_n \to 0$. This model assumes exchangeability.

The graph statistics that we are concerned with, are count statistics of subgraphs. Let $R$ be a subgraph of $G$, with $V(R) \subseteq V(G)$ and $E(R) \subseteq E(G)$.
We have \(|V(R)| = p\) and \(E(R) = e\). For notation, if two graphs \(R\) and \(S\) are equivalent, we denote them by \(R \cong S\) and if \(R\) is a subgraph of \(S\), we denote them by \(R \subseteq S\). Now, the empirical statistic of our concern is

\[
T_G(R) = \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S \subseteq K_n, S \cong R} 1(S \subseteq G)
\]

where, \(Iso(R)\) is the group of Isomorphisms of \(R\) and \(K_n\) is the complete graph on \(n\) vertices.

The population version of the sample statistic \(T_G(R)\) can be defined as

\[
P(R) = E \left\{ \prod_{(i,j) \in R} h(\xi_i, \xi_j) \prod_{(i,j) \notin R} h(\xi_i, \xi_j) \right\}
\]

where, \(\bar{R} = \{(i,j) \notin R, i \in V(G), j \in V(G)\}\). Evidently, we have

\[
E(T_G(R)) = P(R)
\]

If we define normalized versions of parameter \(P(R)\) as

\[
\hat{P}(R) = \rho^{-e} P(R)
\]

where, \(e \equiv \vert E(R) \vert\), then, we can define the corresponding normalized statistic to be

\[
\hat{T}_G(R) = \hat{\rho}^{-e} T_G(R)
\]

where,

\[
\hat{\rho} = \frac{\bar{D}}{n - 1}
\]

where, \(D_i = \text{degree of } v_i, v_i \in V(G_n)\) for \(i = 1, \ldots, n\) and \(\bar{D} = \frac{1}{n} \sum_{i=1}^{n} D_i\).

We wish to approximate the functionals \(E(T_G(R))\) and \(\text{Var}(T_G(R))\) by nonparametric bootstrap. Let us consider the bootstrap estimate of \(\hat{T}_G(R)\) to be \(\hat{T}_b(R)\) and bootstrap estimate of \(\text{Var}(T_G(R))\) to be \(\hat{\sigma}_b^2(R)\). We consider \(b\) as the bootstrap repetition or resampling parameter. How, we get the bootstrap estimates will be discussed in next section. But, for such a bootstrap estimates, we can state the general theorem that we proved -

**Theorem 2.1.** Suppose \(R\) is fixed, acyclic with \(|V(R)| = p\) and \(\int_0^\infty \int_0^\infty \lambda^{2|R|}(u,v)du dv < \infty\). Then if \(\lambda_n \to \infty\) and \(b \to \infty\)

\[
\sqrt{n} \left( \frac{\hat{T}_b(R) - \hat{P}(R)}{\hat{\sigma}_b(R)} \right) \xrightarrow{\mathcal{D}} N(0, 1)
\]
If for fixed, acyclic subgraphs \((R_1, \ldots, R_k)\), we define, \(T_b(R) = (\hat{T}_b(R_1), \ldots, \hat{T}_b(R_k))\) and \(P(R) = (\hat{P}(R_1), \ldots, \hat{P}(R_k))\)

\[
\sqrt{n} \left( \left( T_b(R) - P(R) \right)^T \hat{\Sigma}_b^{-1/2}(R) \left( T_b(R) - P(R) \right) \right) \xrightarrow{w} N(0, I)
\]

where, \([\Sigma_b]_{st} = \hat{\sigma}_b(R_s, R_t), s, t = 1, \ldots, k\) and if \(R_s = R_t\), \(\sigma_b(R_s, R_t) = \hat{\sigma}_b^2(R)\). These results also hold for subgraphs \(R\), which are \(k\)-cycles.

Note that the above theorem is the master theorem. The proof of this theorem depends upon how we obtain the bootstrap estimates \(\hat{T}_b(R)\) and \(\hat{\sigma}_b^2(R)\). We consider two bootstrap procedures.

(I) uniform subsampling bootstrap procedure
(II) sampling-based bootstrap procedure.

All these three bootstrap procedures have different small-sample behavior. That means depending on \(\lambda_n\) and size of \(R\) for a given network \(G\), the efficiency of the bootstrap estimates differ.

We also considered another bootstrap procedure, which was a variant of the common snowball sampling. However, we do not discuss that method in the main discourse, instead, we relegate discussion on that method to the Appendix, since the method performs poorly for all types of graphs compared to the other two methods, both theoretically and empirically.

For each of the bootstrap methods, we prove a theorem of following type

**Theorem 2.2.** Suppose \(R\) is fixed, acyclic with \(|V(R)| = p\), then, if \(b \to \infty\),

\[
\sqrt{n} \left( \hat{T}_b(R) - \hat{T}_G(R) \right) \xrightarrow{P} 0
\]

Also, if \(n \to \infty\) and \(\lambda_n \to \infty\) and under certain conditions depending on the bootstrap method

\[
\frac{\hat{\sigma}_b^2(R)}{\sigma^2(R)} \xrightarrow{P} 1
\]

where, \(\sigma^2(R)\) is the asymptotic variance of \(\hat{T}_G(R)\) as defined in Theorem 1 of [7]. These results also hold for subgraphs \(R\), which are \(k\)-cycles.

We shall prove Theorem 2.2 for each of the three bootstrap cases in Section 3. Then, we shall use the Theorem 2.2 to prove the general theorem 2.1 in Section 4.3.
2.1. Bootstrap and Model-based Sampling. Our work can be viewed from two different perspectives. The first perspective is that of bootstrap. In non-parametric bootstrap, we use resamples or subsamples of the data, where the data comes from an unknown distribution, to find the functionals of the unknown distribution. In our situation also, we have a network that has been generated from an underlying probability model. We want to subsample networks from our given network and use those subsampled networks to find estimates of functionals of the underlying population model generating the given network. Note that here we are interested in subsampling not resampling a network. This is precisely because one of our goals is to reduce the burden of performing computation on the large original network.

The second perspective is that of model-based sampling. In model-based sampling, we consider that the population, from which the sample is selected according to some sampling design, is a realization of a probabilistic event. So, in our case, we consider the given network as the population and it is generated from an underlying probability model.

3. Bootstrap Methods. We propose three different bootstrap methods. Each of them are subsampling approaches of bootstrap. In the following subsections, we shall define each of these subsampling bootstrap methods. We shall also compare the theoretical performance between the three bootstrap schemes.

Let us consider that we have a random graph $G_n$ as the data with $|V(G_n)| = n$ and number of edges of $G_n$ is $|E(G_n)| = m$. Let the adjacency matrix of $G_n$ be denoted by $A_{n \times n}$. For sake of notational simplicity, from here onwards we shall denote $G_n$ by $G$ having $n$ vertices unless specifically mentioned. Let $R$ be a subgraph of $G$, with $V(R) \subseteq V(G)$ and $E(R) \subseteq E(G)$. We have $|V(R)| = p$ and $|E(R)| = e$.

3.1. Uniform Subsampling Bootstrap. In the uniform subsampling bootstrap scheme at each bootstrap iteration a subset of vertices of the full network $G$ is selected without replacement and the graph induced by the selected subset is the subsample we consider. This is a vertex subsampling or induced network sampling scheme. The full bootstrap procedure given the subsample size, $m$ and number of bootstrap iterates, $B$, is as follows –

1. For $b^{th}$ iterate of the bootstrap, $b = 1, \ldots, B$,
2. Choose $m$ vertices without replacement from $V(G)$ and form the induced subgraph of $G$ based on the selected vertices. Denote the graph formed by $H$. 
3. Calculate $T_{b_1}(R)$, given by formula

$$T_{b_1}(R) = \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H)$$

The bootstrap estimate of $T_G(R)$ is given by

$$\tilde{T}_{b_1}(R) = \frac{1}{B} \sum_{b=1}^{B} T_{b_1}(R)$$

The uniform subsampling bootstrap scheme is the network version of the common subsampling bootstrap scheme seen in Bickel et. al. [8]. Note that, there are other ways of forming uniformly subsampled bootstrap estimates as mentioned in [8], however, we just mention one of them in this discourse. The properties of the bootstrap estimator $\tilde{T}_{b_1}(R)$ is given is Lemma 1.

**Lemma 1.** The estimator $\tilde{T}_{b_1}(R)$ has the following properties

(i) Given $G$, $\tilde{T}_{b_1}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $B \to \infty$, $n \to \infty$, $m \to \infty$ and $m/n \to 0$,

$$\sqrt{n}(\rho - \varepsilon \tilde{T}_{b_1}(R) - \rho - \varepsilon T_G(R)) \overset{P}{\to} 0$$

**Proof.** (i) Now, let us try to find the expectation of $T_{b_1}(R)$ under the sampling distribution conditional on the given data $G$.

$$E_b \left[ \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H) \mid G \right]$$

$$= \frac{1}{\binom{m}{p} |Iso(R)|} E \left[ \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H) \mid G \right]$$

$$= \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{H \subseteq G, |H|=m} \frac{1}{\binom{n}{m}} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H)$$

$$= \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{S \subseteq K_n, H \supseteq S, H \subseteq G \mid |H|=m} \frac{1}{\binom{n}{m}} 1(S \subseteq G)$$

$$= \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{S \subseteq K_n, S \supseteq R} \frac{1}{\binom{n}{m}} \frac{1}{\binom{n-p}{m-p}} 1(S \subseteq G)$$

$$= \frac{1}{\binom{m}{p} |Iso(R)|} \sum_{S \subseteq K_n, S \supseteq R} 1(S \subseteq G)$$
So, we have,

$$E_b[\tilde{T}_{b_1}(R) | G] = E_b[T_{b_1}(R) | G] = T_G(R)$$

(ii) Here, we use properties of the underlying model. Let us condition on $\xi = \{\xi_1, \ldots, \xi_n\}$ and the whole graph $G$ separately. Now, conditioning on $\xi$, we get the main term of $T_G(R)$ to be,

$$\begin{equation}
E(\hat{P}(R) | \xi) = \frac{1}{(n)! |Iso(R)|} \sum_{S \subseteq K_n, S \subseteq R} \left( \prod_{(i,j) \in E(S)} w(\xi_i, \xi_j) \right) + O(n^{-1}\lambda_n).
\end{equation}
$$

We shall use the same decomposition as used in [7] of $\langle \rho_n^{-e} \tilde{T}_{b_1}(R) - \tilde{P}(R) \rangle$ into

$$\langle \rho_n^{-e} \tilde{T}_{b_1}(R) - \tilde{P}(R) \rangle = \rho_n^{-e} \left( \tilde{T}_{b_1} - E_b[T_{b_1}(R) | G] \right) + \rho_n^{-e} \left( T_G(R) - E(T_G(R) | \xi) \right) + E(\hat{P}(R) | \xi) \rho_n^{-e} - \tilde{P}(R)$$

Let us define,

$$U_3 = E(\hat{P}(R) | \xi) \rho_n^{-e} - \tilde{P}(R)$$

$$U_2 = \rho_n^{-e} \left( T_G(R) - E(T_G(R) | \xi) \right)$$

$$U_1 = \rho_n^{-e} \left( \tilde{T}_{b_1} - E_b[T_{b_1}(R) | G] \right)$$

Now, it is easy to see that

$$\text{Var}(\rho_n^{-e} \tilde{T}_{b_1}(R)) = E(\text{Var}(\rho_n^{-e} \tilde{T}_{b_1}(R) | G) + \text{Var}(E(\rho_n^{-e} \tilde{T}_{b_1}(R) | G))$$

$$= E(\text{Var}(\rho_n^{-e} \tilde{T}_{b_1}(R) - T_G(R) | G) + \text{Var}(T_G(R))$$

$$= E(\text{Var}(U_1 | G)) + E(\text{Var}(T_G(R) | \xi)) + E(\text{Var}(E(T_G(R) | \xi))$$

$$= E(\text{Var}(U_1 | G)) + E(\text{Var}(U_2 | \xi)) + E(U_3)$$

We shall try to see the behavior of $\text{Var}(U_1 | G) = \text{Var}_b[\rho_n^{-e} \tilde{T}_{b_1}(R) | G]$. Now,

$$\text{Var}_b[\rho_n^{-e} \tilde{T}_{b_1}(R) | G] = \rho_n^{-2e} \frac{1}{B^2} \left( \sum_{b=1}^{B} \text{Var}_b[T_{b_1}(R)] + \sum_{b,b'=1; b \neq b'}^{B} \text{Cov}_b[T_{b_1}(R), T_{b'}(R)] \right)$$

Now, the formula for $\text{Var}_b[T_{b_1}(R)] = O(\frac{1}{m})$ and $\text{Cov}_b[T_{b_1}(R), T_{b'}(R)] = O(\frac{1}{m})$ for acyclic and $k$-cycle $R$ is given in Appendix A1. If we consider the uniform probability for bootstrap to be $\gamma$, then, $B = O(\gamma n^p)$. 

Note that, if $E(H_b) \cap E(H_{b'}) = \phi$, then, $\text{Cov}_b(T_{b1}(R), T_{b'1}(R)) = 0$. The number of pairs such that $E(H_b) \cap E(H_{b'}) \neq \phi$ is $O(m^2 \gamma^2 n^{2m-2})$. Also, the number of edges for the leading term in the covariance is equal to or more than $2e$. So,

$$\mathbb{E}(\text{Var}(U_1|G)) = O\left(\frac{m^2 \gamma^2 n^{2m-2}}{m^2 \gamma^2 n^{2m}}\right) = O\left(\frac{m}{n^2}\right) = o(n^{-1})$$

The last equality follows since we have $m/n \to 0$ as $n \to \infty$.

Now, by proof of Theorem 1 in [7], we have,

$$\text{Var}(U_2) = o(n^{-1})$$
$$\text{Var}(U_3) = o(n^{-1})$$

So, we get, $\text{Var}(\rho^{-e}T_{b1}(R)) = o(n^{-1})$. Since, we already know $\sqrt{n}$-consistency of $(\rho_n^{-e}T_{G}(R) - \hat{P}(R))$, this proves the $\sqrt{n}$-consistency of $\rho_n^{-e}T_{b1}(R)$ to $\rho_n^{-e}T_{G}(R)$.

The variance of $T_{b1}(R)$ given $G$ can also be calculated and is given in the Appendix A1.

3.2. Sampling based bootstrap. In this bootstrap scheme we use an enumeration scheme of finding all possible subgraph $R$ in the graph $G$ and convert the enumeration scheme into a sampling scheme. The enumeration scheme was proposed by Wernicke et. al. (2006) [34]. A random version of the enumeration scheme was also proposed in the paper [34]. We use the random version of the enumeration scheme to form our sampling scheme.

Let us first discuss the enumeration scheme of Wernicke et al [34], which we shall henceforth call EnumerateSubgraph or ESU. The enumeration algorithm is a breadth-first search algorithm. The algorithm strives to create a forest of tree structures such that each leaf of each tree is a size-$p$ subgraph (we have, $|R| = p$). Since, the counting scheme follows a breadth-first search route, before performing the ESU algorithm, we need an ordering of the vertices based on breadth-first search of the graph starting from any particular vertex (say $v$). So, we get a particular fixed ordering of the vertices of the network with $v$ getting lowest order value and subsequently searched vertices getting higher order values. The ordering is described in the algorithm AssignOrder or AO. Based on that ordering of $G$ we perform ESU.

When counting, ESU algorithm creates a forest of tree structures such that each tree represents one vertex of the network and each leaf of each
tree is a size-$p$ subgraph (we have, $|R| = p$). We start with an available vertex of lowest possible order, say $u$. We construct a tree with the vertex $u$ as the root node. We consider $u$ as the “parent” node and neighbors of $u$, which have a higher order than $u$ as its “children”. In the next step, the “children” node become the “parent” node in the tree and has its own neighbors as their “children”. The tree is allowed to grow unto a height $p$, if we are counting size-$p$ subgraphs. So, we can see that each leaf of the tree represents a collection of $p$ nodes coming from the path connecting the leaf to the root. For each vertex, we have such a tree and over counting is averted as we maintain the order while forming the trees. So, with the help of the particular ordering of vertices, each of the size-$p$ subgraphs ($|R| = p$) is counted only once.

Algorithm 3.1 AssignOrder($G, p$)

**Input:** A graph $G = (V, E)$, where $|V(G)| = n$.

**Output:** A vector $\sigma = (\sigma(1), \ldots, \sigma(n))$, where, $\sigma$ is some permutation of $\{1, \ldots, n\}$ and $\sigma(i)$ is associated with vertex $v_{\sigma(i)} \in V(G)$ for all $i = 1, \ldots, n$.

1: $\sigma_1 \leftarrow 1$
2: $V \leftarrow \{v_1\}$
3: $i \leftarrow 1$
4: while $|V| < n$ do
5:   Denote $k \leftarrow |N(V)|$ and $\{v_{h_1}, \ldots, v_{h_k}\} = N(V)$
6:   Define $\sigma(i + j) \leftarrow h_j$ for $j = 1, \ldots, k$.
7:   $i \leftarrow i + k$
8:   $V \leftarrow V \cup N(V)$
9: end while

Once, we have the ordering $\sigma$ for $G$, we define, $v_i \succ v_j$ if $\sigma^{-1}(i) > \sigma^{-1}(j)$, where, $v_i, v_j \in V(G)$ and $i, j = 1, \ldots, n$ with $i \neq j$. This ordering is needed for success of the ESU algorithm and its randomized counterpart 3.2. We shall only formally state the randomized version of the algorithm, RAND-ESU 3.2 in this paper. The enumeration version can be found in [34].

In Theorem 2 of [34] it was proved that the output of ESU algorithm, $S_p$ contains all subgraphs $R$ of $G$, such that $|R| = p$, exactly once. So, we can write the statistic (2.2) for a specific subgraph $R$ with $|R| = p$ in the following way

\[
T_G(R) = \frac{1}{\binom{n}{p}|Iso(R)|} \sum_{S \in S_p} 1(S \cong R)
\]

Essentially we have a normalized population total in terms of sampling theory. Our goal is to form a sampling design and devise a corresponding sampling estimator of $T_G(R)$ given $G$. To meet that goal we use a sampling version of the enumeration scheme ESU.
Now, the sampling version of the ESU algorithm has an extra set of parameters \((q_1, \ldots, q_p)\). We shall call the new algorithm as \textsc{RandomizedEnumerateSubgraph} or \textsc{Rand–ESU}.

**Algorithm 3.2 RandomizedEnumerateSubgraph\((G, p)\)**

\begin{algorithmic}[1]
\STATE **Input:** A graph \(G = (V, E)\), an integer \(p\) and an vector \((q_1, \ldots, q_p)\), where, \(1 \leq p \leq |V|\) and \(q_d \leq 1\) for all \(d = 1, \ldots, p\).
\STATE **Output:** \(S^R_p\) = A sample of subgraphs, \(R\) of \(G\), such that \(|R| = p\).
\STATE 1: for each vertex \(v \in V\) do
\STATE 2: \(V_{\text{Extension}} \leftarrow \{u \in N(\{v\}) : u \succ v\}\)
\STATE 3: With probability \(q_1\) Call RandExtendSubgraph(\(\{v\}, V_{\text{Extension}}, v\))
\STATE 4: end for
\end{algorithmic}

**Algorithm 3.3 RandExtendSubgraph\((V_{\text{Subgraph}}, V_{\text{Extension}}, v)\)**

\begin{algorithmic}[1]
\STATE **Input:** Graphs \(V_{\text{Subgraph}}, V_{\text{Extension}}\) and vertex \(v\).
\STATE **Output:** A sample of subgraphs, \(R\) of \(G\), such that \(|R| = p\) and \(v\) is a vertex of \(R\).
\STATE 1: if \(|V_{\text{Subgraph}}| = p\) then
\STATE 2: return Subgraph of \(G\) induced by \(V_{\text{Subgraph}}\)
\STATE 3: else
\STATE 4: while \(V_{\text{Extension}} \neq \emptyset\) do
\STATE 5: Remove an arbitrarily chosen vertex \(w\) from \(V_{\text{Extension}}\)
\STATE 6: \(V'_{\text{Extension}} \leftarrow V_{\text{Extension}} \cup \{u \in N_{\text{excl}}(w, V_{\text{Subgraph}}) : u \succ v\}\)
\STATE 7: \(d \leftarrow |V_{\text{Subgraph}}| + 1\)
\STATE 8: With probability \(q_d\) Call RandExtendSubgraph\((V_{\text{Subgraph}} \cup \{w\}, V'_{\text{Extension}}, v)\)
\STATE 9: end while
\STATE 10: end if
\STATE 11: return
\end{algorithmic}

From the sampling scheme \textsc{Rand–ESU} we have a sample \(S^R_p\) of size \(p\) subgraphs of \(G\). Now, if we consider each item to be one size \(p\) subgraph of \(G\), that is, an element of \(S_p\), then, we can try to calculate the inclusion probability of each item in the sample \(S^R_p\).

An item, \(S \in S_p\) is a subgraph of \(G\) induced by the set of vertices \(\{w_1, \ldots, w_p\}\), where, we take that \(w_{i+1} \succ w_i, i = 1, \ldots, p - 1\). So,

\[
\pi \equiv \text{Inclusion Probability of } S = \mathbb{P}[(w_1, \ldots, w_p) \text{ is selected}] = \mathbb{P}[w_p | (w_1, \ldots, w_{p-1}) \text{ is selected}] \cdot \mathbb{P}[(w_1, \ldots, w_{p-1}) \text{ is selected}]
\]

\[
= q_p \cdot \mathbb{P}[(w_1, \ldots, w_{p-1}) \text{ is selected}]
\]

\[
= q_p \cdot q_{p-1} \cdot \mathbb{P}[(w_1, \ldots, w_{p-2}) \text{ is selected}]
\]

\[
= \cdots = q_p \cdot q_{p-1} \cdots q_1 = \prod_{d=1}^{p} q_d
\]
So, each item $S \in S_p$ has an inclusion probability $\pi$ to be in the sample $S_p^R$. So, we can define a Horvitz-Thompson estimator of $T_G(R)$ based on $S_p^R$ as

$$\hat{T}_{b2}(R) = \frac{1}{\left(\prod_{d=1}^{p} q_d\right) |Iso(R)|} \sum_{S \in S_p^R} 1(S \simeq R)$$

For variance calculation, we also need the joint inclusion probability of two items, $S, S' \in S_p$, which are subgraphs of $G$ induced by the set of vertices \{$w_1, \ldots, w_p$\} and \{$w'_1, \ldots, w'_p$\} respectively, where, we take that $w_{i+1} \succ w_i$ and $w'_{i+1} \succ w'_i$, $i = 1, \ldots, p - 1$. So,

$$\pi_{SS'} \equiv \text{Inclusion Probability of } S \text{ and } S' = P[ (w_1, \ldots, w_p) \text{ is selected}&(w'_1, \ldots, w'_p) \text{ is selected}]$$

$$= \prod_{d=1}^{p} (q_d)^{z_{1d}} \prod_{d=1}^{p} (q_d^2)^{z_{2d}}$$

where,

$$z_{1d} = \begin{cases} 1(w_d = w'_d), & \text{for } d = 1 \\ 1((w_d, w_{d-1}) = (w'_d, w'_{d-1})), & \text{for } d = 2, \ldots, p \end{cases}$$

$$z_{2d} = \begin{cases} 1(w_d \neq w'_d), & \text{for } d = 1 \\ 1((w_d, w_{d-1}) \neq (w'_d, w'_{d-1})), & \text{for } d = 2, \ldots, p \end{cases}$$

Now, let us try to see the properties of the bootstrap estimator $\hat{T}_{b2}(R)$ -

**Lemma 2.** The estimator $\hat{T}_{b2}(R)$ has the following properties

(i) Given $G$, $\hat{T}_{b2}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $B \to \infty$, $q_1 = 1$ or $q_1 \to 1$ and $q_d \to 0$ and $nq_d \to \infty$ for $d = 2, \ldots, p$ and $n \to \infty$,

$$\sqrt{n}(\rho^e \hat{T}_{b2}(R) - \rho^e T_G(R)) \xrightarrow{P} 0$$

**Proof.** (i) according to the sampling theory [30], we have that $\hat{T}_{b2}(R)$ is an unbiased estimator of $T_G(R)$ given the network $G$.

(ii) The variance of $\hat{T}_{b2}(R)$ coming from the bootstrap sampling only is given by

(3.6)

$$\text{Var}_b[\hat{T}_{b2}(R)] = \frac{1}{N^2} \left[ \frac{1 - \pi}{\pi} \sum_{S \in S_p} 1(S \simeq R) + \sum_{S, S' \in S_p, S \neq S'} \frac{\pi_{SS'}}{\pi^2} 1(S \simeq R, S' \simeq R) \right]$$
where,
\[ N = \binom{n}{p} |\text{Iso}(R)| \]

We shall use the same decomposition as used in [7] of \((\rho_n^{-e} \hat{T}_{b2}(R) - \hat{P}(R))\) into
\[
(\rho_n^{-e} \hat{T}_{b2}(R) - \hat{P}(R)) = \rho_n^{-e} (\hat{T}_{b2} - \mathbb{E}_b[T_{b1}(R)|G])
+ \rho_n^{-e} (T_G(R) - \mathbb{E}(T_G(R)|\xi))
+ \mathbb{E} (\hat{P}(R)|\xi) \rho_n^{-e} - \hat{P}(R)
\]

Let us define,
\[
U_3 = \mathbb{E} (\hat{P}(R)|\xi) \rho_n^{-e} - \hat{P}(R)
U_2 = \rho_n^{-e} (T_G(R) - \mathbb{E}(T_G(R)|\xi))
U_1 = \rho_n^{-e} (\hat{T}_{b2} - \mathbb{E}_b[T_{b2}(R)|G])
\]

Now, it is easy to see that
\[
\text{Var}(\hat{T}^b(R)) = \mathbb{E} (\text{Var}(T_{b2}(R)|G) + \text{Var}(\mathbb{E}(T_{b2}(R)|G)))
= \mathbb{E} (\text{Var}(\hat{T}_{b2}(R) - T_R(G)|G) + \text{Var}(T_G(R)))
= \mathbb{E} (\text{Var}(U_1|G)) + \mathbb{E} (\text{Var}(T_G(R)|\xi)) + \text{Var}(\mathbb{E}(T_G(R)|\xi))
= \mathbb{E} (\text{Var}(U_1|G)) + \mathbb{E} (\text{Var}(U_2|\xi)) + \text{Var}(U_3)
\]

We shall try to see the behavior of \(\text{Var}(U_1|G) = \text{Var}_b[T_{b3}(R)|G]\). From the formula of \(\text{Var}_b[T_{b3}(R)|G]\), we see that, the covariance terms vanishes when \(\pi_{SS'} = \pi^2\). Now, if \(q_1 = 1\), then, \(\pi_{SS'} = \pi^2\) if \(E(S) \cap E(S') = \phi\). The number of pairs such that \(E(S) \cap E(S') \neq \phi\) is \(O(p^2n^{2p-2})\). So,
\[
\mathbb{E} (\text{Var}(U_1|G)) = O\left(\frac{p^2n^{2p-2}}{N}\right) = O(p^2/n^2) = o(n^{-1})
\]

Now, the condition of \(q_1 = 1\) is a bit restrictive. In stead, if we have \(q_1 \to 1\) as \(n \to \infty\), then, the highest order term of covariance term comes from the case when \(E(S) \cap E(S') \neq \phi\) but the root nodes are
same that is \( w_1 = w'_1 \). So, for some constant \( C > 0 \),

\[
\frac{1}{N^2} \sum_{S,S' \in S_p, S \neq S'} \pi_S \pi_{S'} - \frac{\pi^2}{2} \mathbf{1}(S \cong R, S' \cong R) \leq \frac{C}{N^2} \sum_{S,S' \in S_p, S \neq S'} \frac{q_1 - q_1^2}{q_1^2} \mathbf{1}(S \cong R, S' \cong R)
\]

\[
= O \left( \left( \frac{1}{q_1} - 1 \right) \frac{n^{2p-1}}{n^{2p}} \right)
\]

\[
= O \left( \left( \frac{1}{q_1} - 1 \right) \frac{1}{n} \right)
\]

\[
= o(n^{-1})
\]

Now, for the variance term to vanish we need the conditions \( q_1 = 1 \) or \( q_1 \rightarrow 1 \) and \( q_d \rightarrow 0 \) and \( nq_d \rightarrow \infty \) for \( d = 2, \ldots, p \) as \( n \rightarrow \infty \). Under these conditions, we have

\[
\frac{1}{N^2} \frac{1 - \pi}{\pi} \sum_{S \in S_p} \mathbf{1}(S \cong R) = \left( \frac{1}{\pi} - 1 \right) O \left( \frac{n^p}{n^{2p}} \right)
\]

\[
= O \left( \frac{1}{n^{2p} \pi} \right)
\]

\[
= O \left( \frac{1}{n} \cdot \prod_{d=2}^{p} \frac{1}{nq_d} \right)
\]

\[
= o(n^{-1})
\]

So, we have,

\[
\text{Var}(U_1) = o(n^{-1})
\]

Now, by proof of Theorem 1 in [7], we have,

\[
\text{Var}(U_2) = o(n^{-1})
\]

\[
\text{Var}(U_3) = o(n^{-1})
\]

So, we get, \( \text{Var}(\rho^{-e}\hat{t}_{k_2}(R)) = o(n^{-1}) \). Since, we already know \( \sqrt{n} \)-consistency of \( (\rho_n^{-e}T_G(R) - \hat{P}(R)) \), this proves the \( \sqrt{n} \)-consistency of \( \rho_n^{-e}\hat{t}_{k_2}(R) \) to \( \rho_n^{-e}T_G(R) \).

\[\square\]

3.3. Comparison of the Bootstrap Methods. Among the two bootstrap methods, the uniform subsampling scheme works for dense graphs only. If the subgraph pattern becomes large or if the original network is sparse,
then the subgraph size also need to be large leading to slowing of the bootstrap scheme. However for dense graphs the uniform subsampling bootstrap scheme is fast and accurate.

The sampling-based bootstrap scheme is more accurate and stable, that means it has less bootstrap variance than the uniform subsampling bootstrap scheme. But, for dense graphs it becomes slower than uniform subsampling bootstrap to maintain its low bootstrap variance.

So, if speed is your concern, uniform subsampling bootstrap might be better choice, however, if you want a more reliable estimate then, sampling based bootstrap would be a better choice.

4. Theoretical Results. In this section, we try to give an estimate of asymptotic variance of \( \rho^{-e}T_G(R) \), \( \sigma^2(R) \), which is defined in Theorem 1 of [7]. By obtaining an estimate of the asymptotic variance of \( \rho^{-e}T_G(R) \), we can estimate its asymptotic distribution and thus construct hypothesis tests based on the asymptotic distribution. We combine the results obtained in Section 2 to prove Theorem 2.1

4.1. Estimation of Variance and Covariance. We shall try to find the variance of the statistic \( \rho^{-e}T_G(R) \) and then, using it give an estimate of the variance of the statistic \( \hat{T}_G(R) \). The source of variation here is the randomness coming from sampling from the underlying model (2.1).

\[
\text{Var} \left[ \rho^{-e}T_G(R) \right] = \text{Var} \left[ \sum_{S \subseteq K_n, S \subseteq R} \frac{1(S \subseteq H)}{\rho^e(n_p)|Iso(R)|} \right] = \frac{1}{(\rho^e(n_p)|Iso(R)|)^2} \mathbb{E} \left[ \sum_{S \subseteq K_n, S \subseteq R} 1(S \subseteq H) \right]^2 - \left( \hat{P}(R) \right)^2
\]

If \( R \) is a connected subgraph and \( W = S \cup T \) and \( k = |W| \) and \( e_W \equiv |E(W)| \), then, \( k = p, \ldots, 2p - 1 \) and each term of sum in RHS (4.1)

\[
\frac{1}{(\rho^e(n_p)|Iso(R)|)^2} \mathbb{E} \left[ \sum_{W \subseteq K_n} 1(W \subseteq H) \right] = \frac{\rho^{e_W}(n_p)|Hom(W)|}{(\rho^e(n_p)|Iso(R)|)^2} \hat{P}(W) = O(n^{k-2p}\rho^{e_W-2e})
\]

If \( |W| = |S \cup T| = 2p - 1 \), then, \( W \) is a connected graph, with \( e_W = 2e \). So, we have the main leading term equals \( O\left(\frac{1}{n}\right) \) for acyclic \( R \). For, any other
W, such that, |W| = k < 2p − 1 then the Eq (4.1) becomes

\[
\frac{1}{\binom{n}{p}} \mathbb{E} \left[ \sum_{W \subseteq K_n} 1(W \subseteq H) \right] = \frac{\rho^{e_W(n)}}{\binom{n}{p} |Iso(R)|^2} \hat{P}(W) = O(n^{k-2p} \rho^{e_W-2e}) = o(n^{-1})
\]

under the condition that \( \rho \) decreases at a rate slower than \( O(n^{-1}) \), that means except when we are in the constant degree case. The condition is equivalent to stating that \( \lambda_n \rightarrow \infty \).

So, for calculation of variance, we only estimate the count of the features which are \( W = S \cup T \) and \( |W| = 2p - 1 \), that means \( S \) and \( T \) have only one node in common. So, the estimator of variance becomes -

\[
(4.2) \quad \hat{\sigma}^2(R) = \frac{1}{\binom{n}{p} |Iso(R)|^2} \sum_{W \subseteq K_n, W = S \cup T, S, T \sim R, \|S \cap T\| = 1} 1(W \subseteq G)
\]

and we have

\[
\mathbb{E} \hat{\sigma}^2(R) = \text{Var} [\rho^{-e} T_G(R)] + o(n^{-1})
\]

Similarly, for calculation of covariance between two count statistics, \( T_G(R_1) \) and \( T_G(R_2) \), we only estimate the count of the features which are \( W = S \cup T \) and \( |S \cap T| = 1 \), \( S \sim R_1, T \sim R_2 \), that means \( S \) and \( T \) have only one node in common. So, the estimator of covariance becomes -

\[
(4.3) \quad \hat{\sigma}(R_1, R_2) = \frac{1}{\binom{n}{p} |Hom(R_1)|} \frac{1}{\binom{n}{p} |Hom(R_2)|} \sum_{W \subseteq K_n, W = S \cup T, S \sim R_1, T \sim R_2, \|S \cap T\| = 1} 1(W \subseteq G)
\]

where \( e_{R_1} = |E(R_1)| \) and \( e_{R_2} = |E(R_2)| \). So, we have

\[
\mathbb{E} \hat{\sigma}(R_1, R_2) = \text{Cov} [\rho^{-e_{R_1}} T_G(R), \rho^{-e_{R_2}} T_G(R_2)] + o(n^{-1})
\]

Now, from the Theorem 1(a) in [7], we know that as \( \lambda_n \rightarrow \infty \), if \( \hat{\rho}_n = \frac{D}{n-1} \) as defined in (2.3),

\[
\frac{\hat{\rho}_n}{\rho_n} \xrightarrow{P} 1
\]

So, using the estimate \( \hat{\rho}_n \), we define the estimate of variance -

\[
(4.4) \quad \hat{\sigma}^2(R) = \frac{1}{\left( \hat{\rho}_n \binom{n}{p} |Iso(R)| \right)^2} \sum_{W \subseteq K_n, W = S \cup T, S \sim R_1, T \sim R_2, \|S \cap T\| = 1} 1(W \subseteq G)
\]
and the estimate of covariance is -

\[ \hat{\sigma}(R_1, R_2) = \frac{1}{(\mathbb{P}^{\mathcal{R}^1}(\mathcal{R}^2 \mid \mathcal{H}om(R_1)))} \frac{1}{(\mathbb{P}^{\mathcal{R}^2}(\mathcal{R}^1 \mid \mathcal{H}om(R_2)))} \sum_{W \subseteq K_n, W = S \cup T, S \cong R_1, T \cong R_2, |S \cap T| = 1} 1(W \subseteq G) \]

So, \( \hat{\sigma}^2(R) \) and \( \hat{\sigma}(R_1, R_2) \) become consistent estimates of \( \text{Var}[\rho^{-e}T_G(R)] \) and \( \text{Cov}[\rho^{-eR_1}T_G(R_1), \rho^{-eR_2}T_G(R_2)] \) respectively and consequently a consistent estimate of \( \text{Var}[\hat{T}_G(R)] \) and \( \text{Cov}[\hat{T}_G(R_1), \hat{T}_G(R_2)] \) respectively.

**Lemma 3.** As \( \lambda_n \to \infty \) and \( n \to \infty \),

\[ \frac{\hat{\sigma}^2(R)}{\text{Var}[\rho^{-e}T_G(R)]} \xrightarrow{P} 1 \]

\[ \frac{\hat{\sigma}(R_1, R_2)}{\text{Cov}[\hat{T}_G(R_1), \hat{T}_G(R_2)]} \xrightarrow{P} 1 \]

**Proof.** The proof follows from previous discussion. \( \square \)

Now, we can see that \( \hat{\sigma}^2(R) \) and \( \hat{\sigma}(R_1, R_2) \) are nothing but count statistics on the statistic \( W = S \cup T \), where, \( S, T \cong R \) and \( |S \cap T| = 1 \). So, using bootstrap methods, we can get an estimate of \( \hat{\sigma}^2(R) \) -

\[ \hat{\sigma}^2_b(R) = \left( \frac{\hat{\rho}_n^{eW}(\mathcal{R}^2_{2p-1})|Iso(R)|}{\hat{\rho}_n^e(\mathcal{R}^1_p)|Iso(R)|} \right)^2 \hat{T}_{bi}(W) \text{ for } i = 1, 2. \]

and an estimate of \( \hat{\sigma}(R_1, R_2) \) -

\[ \hat{\sigma}_{bi}(R_1, R_2) = \left( \frac{\hat{\rho}_n^{eW}(\mathcal{R}^2_{2p-1})|Iso(R)|}{\hat{\rho}_n^{eR_1}(\mathcal{R}^2_{2p-1})|\mathcal{H}om(R_1)|} \left( \frac{\hat{\rho}_n^{eR_2}(\mathcal{R}^2_{2p-1})|\mathcal{H}om(R_2)|}{\hat{\rho}_n^e(\mathcal{R}^1_p)|\mathcal{H}om(R_1)|} \right) \right)^2 \hat{T}_{bi}(W) \text{ for } i = 1, 2. \]

where, \( W = S \cup T \) with \( S, T \cong R \) and \( |S \cap T| = 1 \) and \( |V(W)| = 2p - 1 \) and \( e_W = |E(W)| \). \( \hat{T}_{bi}(W) \) \( (i = 1, 2) \) are bootstrap count statistics estimates, defined in Eq (3.2) and (3.5). Also from Lemma 1 and Lemma 2, we get the \( \sqrt{n} \)-consistency of the bootstrap count estimates \( \hat{T}_{bi}(W) \) \( (i = 1, 2, 3) \). So, we can now combine the Lemma 1 and 2 and Lemma 3, to see that \( \hat{\sigma}^2_b(R) \) and \( \hat{\sigma}_{bi}(R_1, R_2) \) \( (i = 1, 2) \) are consistent estimators of \( \sigma^2(R) \) and \( \sigma(R_1, R_2) \) respectively by Slutsky’s Theorem and Convergence of Types Theorem. So, we get an estimate of variance, \( \hat{\sigma}^2_b(R) \) \( (i = 1, 2) \) with the property
Lemma 4. As \( \lambda_n \to \infty, n \to \infty \) and under conditions of Lemma 1 and Lemma 2,

\[
\frac{\hat{\sigma}^2_{bi}(R)}{\sigma^2(R)} \xrightarrow{P} 1 \text{ for } i = 1, 2
\]

\[
\frac{\hat{\sigma}_{bi}(R_1, R_2)}{\sigma(R_1, R_2)} \xrightarrow{P} 1 \text{ for } i = 1, 2
\]

Proof. The proof follows from previous discussion.

4.2. Proof of Theorem 2.2. The proof of the internal theorem follows from the lemmas of previous section. Since, we have \( \sqrt{n} \)-consistent bootstrap estimators of \( \rho^{-e}\tilde{T}_b(R) \). Now, from the Theorem 1(a) in [7], we know that as \( \lambda_n \to \infty \), if \( \hat{\rho}_n = \frac{\hat{D}}{n-1} \) as defined in (2.3),

\[
\frac{\hat{\rho}_n}{\rho_n} \xrightarrow{P} 1 \quad \sqrt{n} \left( \frac{\hat{\rho}_n}{\rho_n} - 1 \right) \xrightarrow{w} N(0, \sigma^2)
\]

Now, we can define the bootstrap estimates as -

\[
\hat{T}_bi(R) = \hat{\rho}^{-e}\tilde{T}_bi(R) \text{ for } i = 1, 2, 3.
\]

So, we get by applying Slutsky’s Theorem that

\[
\sqrt{n} \left( \hat{T}_bi(R) - \hat{T}_G(R) \right) \xrightarrow{P} 0 \text{ for } i = 1, 2, 3.
\]

The statement about bootstrap estimate of variance follows from Lemma 4 and the definitions of bootstrap variance in the form of equation (4.8).

4.3. Proof of Theorem 2.1. The proof of the main theorem follows from the lemma of Section 4.1 and Theorem 2.2. We have \( \sqrt{n} \)-consistent bootstrap estimators, \( \hat{T}_b(R) \) (for \( i = 1, 2, 3 \)) of \( \hat{T}_G(R) \) and consistent estimators, \( \hat{\sigma}^2_{bi}(R) \) (for \( i = 1, 2 \)) of \( \sigma^2(R) \). Also from Theorem 1 of [7], we have, for subgraphs \( R_1, \ldots, R_k \) of \( G_n \),

\[
\sqrt{n} \left( \left( \hat{T}_G(R_1), \ldots, \hat{T}_G(R_k) \right) - \left( \hat{P}(R_1), \ldots, \hat{P}(R_k) \right) \right) \xrightarrow{w} N(0, \Sigma(R))
\]

So, we can combine the result from Theorem 2.2 with the above theorem, using Slutsky and convergence of types theorem, to get the symptomatic normality behavior of \( \hat{T}_b(R) \). As \( n \to \infty, \lambda_n \to \infty \) and under
conditions of Lemma 1, Lemma 7 and Lemma 2, if we define, \(T_{bi}(R) = (\hat{T}_{bi}(R_1), \ldots, \hat{T}_{bi}(R_k))\) and \(P(R) = (\hat{P}(R_1), \ldots, \hat{P}(R_k))\)

\[
\frac{1}{\sqrt{n}} \left( (T_{bi}(R) - P(R)) \Sigma_{bi}^{-1/2}(R) (T_{bi}(R) - P(R)) \right) \overset{w}{\rightarrow} N(0, I)
\]

for \(i = 1, 2, 3\)

where, \([\Sigma_{bi}]_{st} = \hat{\sigma}_{bi}(R_s, R_t), s, t = 1, \ldots, k\) and if \(R_s = R_t = R\), \(\sigma_{bi}(R_s, R_t) = \hat{\sigma}_{bi}^2(R)\) for \(i = 1, 2, 3\).

### 5. Simulation Results

We apply the three representative bootstrap subsampling schemes for simulated datasets to find out their performances. We generate data from two different simulation models. Both models are special cases of the nonparametric model described in [6]. The two models that we consider are:

- **Stochastic block model**
- **Preferential attachment model**

For each of the models, we try to find estimate of the count statistics features and their confidence intervals through bootstrap subsampling. The features that we consider are generalized \((k, l)\)-wheels, triangles and a smooth function of them, transitivity.

#### 5.1. Count Statistics

The main class of acyclic features we consider are generalized \((k, l)\)-wheels.

**Definition 5 (Wheels).** A \((k, l)\)-wheel is an acyclic graph with \(kl + 1\) vertices and \(kl\) edges isomorphic to the graph with edges \(\{(1, 2), \ldots, (k, k + 1), (1, k + 2), \ldots, (2k, 2k + 1), \ldots, (1, (l - 1)k + 2), \ldots, (lk, lk + 1)\}\).

In other words a \((k, l)\)-wheel is a subgraph \(R\), such that it contains

- i) A “hub” vertex
- ii) \(l\) spokes from hub
- iii) Each spoke has \(k\) connected vertices.

Edges, ‘V’, ‘W’ are all examples of \((k, l)\)-wheels. An edge is a \((1, 1)\)-wheel, a ‘V’ is a \((1, 2)\)-wheel and a ‘V’ is a \((2, 2)\)-wheel.

**Definition 6 (Generalized Wheels).** A generalized \((k, l)\)-wheel, where \(k = (k_1, \ldots, k_t), l = (l_1, \ldots, l_t)\) are vectors and the \(k_j\)’s are distinct integers, is the union \(R_1 \cup \ldots \cup R_t\), where \(R_j\) is a \((k_j, l_j)\)-wheel, \(j = 1, \ldots, t\) and the wheels \(R_1, \ldots, R_t\) share a common hub but all their spokes are disjoint.
A \((k, l)\)-wheel has a total of \(p = \sum_j k_j l_j + 1\) vertices and \(\sum_j k_j l_j\) edges. The following picture is an example of \(((2, 1), (1, 1, 1))\)-wheel and it is an union of two ‘V’s, where the common vertex is the hub of one ‘V’ and leaf of the other ‘V’.

In these simulations, we consider counts of simple \((k, l)\)-wheels such as \((1, 2)\) and \((1, 3)\). We also consider the count of the cyclic pattern such as triangle and quadrilaterals or 4-cycles. We consider a smooth function of counts of triangle and ‘V’s, known as transitivity, \(T_{Tr}\), defined as

\[
T_{Tr} = \frac{\hat{P}(R_1)}{\hat{P}(R_2) + \hat{P}(R_2)}
\]

where, \(R_1\) is a triangle or a 3-cycle and \(R_2\) is a ‘V’ or a \((1, 2)\)-wheel.

5.2. Stochastic Block Model. Let \(w\) correspond to a \(K\)-block model defined by parameters \(\theta = (\pi, \rho_n, S)\), where \(\pi_a\) is the probability of a node being assigned to block \(a\) as before, and

\[
F_{ab} = \mathbb{P}(A_{ij} = 1 | i \in a, j \in b) = \rho_n S_{ab}, \quad 1 \leq a, b \leq K.
\]

and the probability of node \(i\) to be assigned to block \(a\) to be \(\pi_a\) \((a = 1, \ldots, K)\).

We consider a stochastic block model with \(K = 2\). We consider the parameter matrix \(F = \lambda F^{(1)} + (1 - \lambda) F^{(2)}\), where, \(F^{(1)}_{2 \times 2} = \text{Diag}(0.0525, 0.0975)\) and \(F^{(2)}_{2 \times 2} = 0.015 J_2\), where, \(J_2\) is a \(2 \times 2\) matrix of all 1’s. So, we get \(\rho_n = \pi^T F \pi\). We now, vary \(\lambda\) to get different combinations of \(F\) as well as \(\rho_n\).

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary \(\lambda_n\) and \(n\) for the model. In Figure 1(a), we compare the mean and variance of the normalized count statistic, \(\hat{T}_{G}(R)\), where, \(R\) is \((1, 2)\)-wheel. We find the bootstrap estimator \(\hat{T}_{b}(R)\) for all three bootstrap schemes - \(i = 1, 2, 3\) and we also find the corresponding estimates of variance by bootstrap, \(\hat{\sigma}_{b}^2(R)\), for all three bootstrap schemes - \(i = 1, 2, 3\). We then plot the plot the estimator \(\hat{T}_{b}(R)\) along with the asymptotic 95% confidence interval using the asymptotic normality result of Theorem 2.1 and the bootstrap estimates of variance \(\hat{\sigma}_{b}^2(R)\). In Figure 1(b) we have a similar plot, but instead of \(T_{G}(R)\), we use the statistic \(T_{Tr}\). We find the bootstrap estimate of \(T_{Tr}\) in the form,

\[
\hat{T}_{Tr}^b = \frac{\hat{T}_{b}(R_1)}{\hat{T}_{b}(R_2) + \hat{T}_{b}(R_2)}
\]
where, $\hat{T}_b$ is the bootstrap estimate of count statistic $T_G(R)$ and $R_1$ is a triangle or a 3-cycle and $R_2$ is a ‘V’ or a (1,2)-wheel. The bootstrap estimate of asymptotic variance of $\hat{T}^b_{\hat{T}_b}$ is obtained from the bootstrap estimates of $\hat{\sigma}^2_b(R_1)$, $\hat{\sigma}^2_b(R_2)$ and $\hat{\sigma}_b(R_1, R_2)$ by using Delta method and using the Theorem 2.1.

In Figure 1, for $n = 500$, we vary average degree $(\lambda_n)$ and (a) Plot estimated normalized (1, 2)-wheel count (b) Plot estimated Transitivity and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use different colors to indicate different bootstrap subsampling schemes.

In Figure 2, we see that the variance of the bootstrap estimates, arising solely from bootstrap iterations and not from model-based iterations, decrease, as we increase the number of subsamples in the Uniform Subsampling Scheme. In Figure 3, we compare the variance of the bootstrap estimates, based on bootstrap iterations for the three different bootstrap schemes. We see that bootstrap variance is universally low for the Sampling-based Scheme as we vary average degree, $\lambda_n$ of the graph. However, the bootstrap variance of the Non-uniform Snowball Sampling Scheme, decrease, as average degree $\lambda_n$ increase. We expect such a behavior, as, $\lambda_n$ increase, the bootstrap subsample in Non-uniform Snowball Sampling Scheme becomes large and stable. So, based on simulations, we recommend Sampling-based Scheme for bootstrap.

In Figure 4, we try to see the behavior of mean and variances of the count statistics, (1, 3)-wheels and 4-cycles. We use only Sampling-based Scheme for bootstrap in this case. Like in Figure 1, we plot the plot the estimator $\hat{T}_{I_2}(R)$ along with the asymptotic 95% confidence interval using the asymptotic normality result of Theorem 2.1 and the bootstrap estimates of variance.
Figure 2. For \( n = 500 \) and \( \lambda_n = 19.875 \), we vary the subsample size of the Uniform subsampling scheme and plot the bootstrap variance of bootstrap estimators of Uniform subsampling scheme.

Figure 3. For \( n = 500 \), we vary average degree \( (\lambda_n) \) and plot (a) bootstrap variance of estimated normalized \((1,2)\)-wheel count (b) bootstrap variance of normalized \((1,3)\)-wheel count. We use different colors to indicate different bootstrap subsampling schemes.
\( \hat{\sigma}_b^2(R) \), for \( R \) as \((1,3)\)-wheel in Figure 4(a) and or \( R \) as 4-cycle in Figure 4(b).

\[ \tau(v) \simeq \frac{D_v}{D} \]

So, following Eq. (2.1), we have the probability of edge formation as

\[ w(u, v) = \frac{\tau(u)}{T(u)} 1(u \leq v) + \frac{\tau(v)}{T'(u)} 1(v \leq u) \]

where, \( T(u) = \int_u^1 \tau(s)ds \) and \( T'(v) = 1 - T(v) \) and

\[ \tau(u) = \int_0^1 w(u, v)dv \]

So, the preferential attachment model can be defined by the following formula on \( w \)

\[ w(u, v) = \frac{\tau(u)}{\int_u^1 \tau(s)ds} 1(u \leq v) + \frac{\tau(v)}{\int_v^1 \tau(s)ds} 1(v \leq u) \]
Thus, for
\[ w(u, v) = (1 - u)^{-1/2}(1 - v)^{-1/2} \]
we have,
\[ \tau(v) = c(1 - v)^{-1/2} \]
which is equivalent to power law of degree distribution \( F \equiv \tau^{-1} \).

In the following figure, we try to see the behavior of mean and variances of the count statistics, as we vary \( \lambda_n \) for the model. In Figure 5(a), we compare the mean and variance of the normalized count statistic, \( \hat{T}_{G}(R) \), where, \( R \) is \((1, 2)\)-wheel. We find the bootstrap estimator \( \hat{T}_{bi}(R) \) for all three bootstrap schemes - \( i = 1, 2, 3 \) and we also find the corresponding estimates of variance by bootstrap, \( \hat{\sigma}^2_{bi}(R) \), for all three bootstrap schemes - \( i = 1, 2, 3 \). We then plot the plot the estimator \( \hat{T}_{bi}(R) \) along with the asymptotic 95% confidence interval using the asymptotic normality result of Theorem 2.1 and the bootstrap estimates of variance \( \hat{\sigma}^2_{bi}(R) \). In Figure 1(b) we have a similar plot, but instead of \( T_{G}(R) \), we use the statistic \( T_{Tr} \). We find the bootstrap estimate of \( T_{Tr} \) and the bootstrap estimate of asymptotic variance of \( \hat{\sigma}^2_{Tr} \) is obtained from the bootstrap estimates of \( \hat{\sigma}^2_{R1}(R), \hat{\sigma}^2_{R2}(R) \) and \( \hat{\sigma}_{R1, R2} \) by using Delta method and using the Theorem 2.1.

![Figure 5](image-url)

**Figure 5.** For \( n = 500 \), we vary average degree \( \lambda_n \) and (a) Plot estimated normalized \((1, 2)\)-wheel count (b) Plot estimated Transitivity and their 95% Confidence Interval (CI), where, CI is estimated using bootstrap estimates of variance of the estimators. We use different colors to indicate different bootstrap subsampling schemes.
5.4. Comparison Between Stochastic Block Model and Preferential Attachment Model. We simulate networks from both stochastic block model and preferential attachment model and then, we try to compare the distribution of a statistic of the graph for two different networks. As a statistic, we use transitivity here. In Figure 6, compare the bootstrap estimated mean and variance of transitivity of the networks simulated from the two different models. We keep the average degree, $\lambda_n$, of the two simulated networks same and then, we try to get the asymptotic distribution of the transitivity statistic for the two cases for each $\lambda_n$. We see here that, for low average degree, we can not statistically distinguish between the transitivity of networks generated from two different models, but, they become statistically distinguishable as average degree, $\lambda_n$, increases.

6. Real Data Examples. Social networks recently has become quite large after the introduction of social networking sites. We consider two different social networks as a platform for our experiments. The first one, High School Romantic Relations data (HS) is a small social network, whereas the second one, Facebook College Social Network (FB) has greater number of nodes and links. The Facebook data was presented in [32]. In this dataset, the Facebook social network of different colleges are represented. In a network, the nodes are people of the colleges and a link represents online friendship between the nodes. The High School data was presented in [4]. In this network, each student is a node and a link between students indicate that they had romantic relations.
6.1. *High School Network*. In this application, we try to quantitatively verify some of the hypothesis mentioned by the authors in the paper [4] when presenting the data. The network here is formed by students of Jefferson High school as nodes and if two students have romantic relations then there exists a link between those two nodes. In the paper, [4], where the data was presented, an observation was made about the dearth of short cycles in the network. Our application here is trying to answer the question whether the absence of short cycles in this graph is significant or not. We consider a very simple model for the data.

We consider the data has been generated from two different models -

(a) Stochastic block model with two blocks (Male and Female) and the connection probability matrix is given by

\[
P = \begin{pmatrix}
\hat{P}_{11} & \hat{P}_{12} \\
\hat{P}_{12} & \hat{P}_{22}
\end{pmatrix}
\]

where, \(\hat{P}_{ab}\) = Average number of edges between blocks \(a\) and \(b\) in the network, where, \(a, b = 1, 2\) are the two blocks with Male = 1 and Female = 2. In this network, we have \(\hat{P}_{11} = 0\), \(\hat{P}_{12} = 0.0058\) and \(\hat{P}_{22} = 0.000025\). The probability of belonging to the two blocks are (0.497, 0.503).

(b) Preferential attachment model with \(\rho = \frac{\lambda_n}{n}\), where, \(\lambda_n\) = Average degree of the network = 1.66 and \(n\) is the number of nodes.

Now, for these two simple models, we can theoretically find the normalized count of small cycles. Then, we can perform a hypothesis test to find out whether the number of small cycles we see in this network is significantly small or not. We use the results of Theorem 1 to form the asymptotic test. The results are given in Table 1. We see in the results that, according to the two simple models, it is extremely unlikely for 3-cycles and 4-cycles to occur in the graph. In fact, the original network has *too many* 4-cycles short cycles not *too few*. This is an interesting observation coming out of our simple exploratory analysis. So, our simple models do not capture the probabilistic mechanism of the original network correctly and we need to analyze the short cycles in the network more closely to understand their formation.

Note that, this is a very small and sparse network. For this network, the use of Theorem 2 from [7] would have sufficed, but we give the example as an example of the use of count statistics and their quantitative behavior. Here in the paper [4] permutation tests were used. We use asymptotic Gaussian tests
<table>
<thead>
<tr>
<th>Subgraph</th>
<th>Normalized Count</th>
<th>Standard Deviation</th>
<th>Count (SBM)</th>
<th>Count (PFA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)-wheel</td>
<td>2.27</td>
<td>0.17</td>
<td>1.01</td>
<td>2.97</td>
</tr>
<tr>
<td>3-cycle</td>
<td>1.31</td>
<td>0.1</td>
<td>0.01</td>
<td>1.04</td>
</tr>
<tr>
<td>4-cycle</td>
<td>9.47</td>
<td>3.16</td>
<td>0.63</td>
<td>3.06</td>
</tr>
</tbody>
</table>

Table 1: The normalized subgraph counts, their standard deviation and the expected counts from the stochastic block model (SBM) and preferential attachment model (PFA) for the whole high school network.

and we can directly answer the questions without the possible awkwardness of permutation tests.

6.2. Facebook Network. In this application, we try to quantitatively analyze the behavior of some of the known descriptive statistics for Facebook collegiate networks. The networks were presented in the paper by Traud et al. (2011) [32]. The network is formed by Facebook users acting as nodes and if two Facebook users are “friends” there is an edge between the corresponding nodes. Along with the network structure, we also have the data on covariates of the nodes. Each node has covariates: gender, class year, and data fields that represent (using anonymous numerical identifiers) high school, major, and dormitory residence. We try to answer two very basic questions quantitatively for these networks-

1. Can the node covariates act as cluster identifiers?
2. Can two college networks be distinguishable in terms of some basic descriptive statistics?

In order to address the first question, we consider the network of a specific college (Caltech). We consider the covariates class year, major and dormitory residence as our covariates of interest. Note that each of these covariates are district covariates. We take the induced network created by levels of each of these covariates and try to see if those networks have different clustering properties. For example, consider class year and major as the covariates of interest. We consider the nodes belonging two different class years and find their induced network from the whole collegiate network. Similarly, we consider the nodes belonging two different majors and find their induced network from the whole collegiate network. Now, we have two different networks, one of which has nodes coming exclusively from two different class years and another has nodes coming exclusively from two different majors. We now try to find which of two networks is more “clustered” by comparing transitivity of the two networks. We can repeat the same exercise for any two covariates and choosing a subset of their levels.
The second question can also be answered in the similar spirit as the first one. We consider the full collegiate network of two different colleges (Caltech and Princeton). Then, we try to compare the mean degree and transitivity of these two collegiate networks.

These comparisons could in principle be possible using the results given in Bickel et. al. (2011) [7], but computationally intractable. Using bootstrap estimators, we can estimate the variance of the estimators and thus perform hypothesis testing in reasonable time.

In Tables 6.2, 3 and 4, we present an excerpt of the result of our analysis and the answer the both the questions. The results indicate that some conclusions in [32] are questionable.

<table>
<thead>
<tr>
<th>Class Year(CY)</th>
<th>Dormitory(DM)</th>
<th>Major(MJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.22</td>
<td>0.12</td>
</tr>
</tbody>
</table>

**Table 2**
Transitivity of induced networks formed by considering only two levels of a specific covariate of a specific collegiate network.

<table>
<thead>
<tr>
<th>Difference</th>
<th>CY and DM</th>
<th>DM and MJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated</td>
<td>0.07</td>
<td>0.1</td>
</tr>
<tr>
<td>Estimated SD</td>
<td>0.05</td>
<td>0.035</td>
</tr>
</tbody>
</table>

**Table 3**
The Difference between Class Year and Dorm is not significant but difference between Dorm and Major is significant by asymptotic normal test at 5% level. The data was presented in Traud et. al. (2011) SIAM Review.

<table>
<thead>
<tr>
<th>Estimated Transitivity</th>
<th>Network 1</th>
<th>Network 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.29</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimated Difference</th>
<th>Network 1</th>
<th>Network 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.13</td>
<td>0.11</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimated Difference SD</th>
<th>Network 1</th>
<th>Network 2</th>
</tr>
</thead>
</table>

**Table 4**
The Difference of transitivity between two networks is not significant by asymptotic normal test at 5% level. Therefore Network 1 can not be said to be more 'clusterable'. The data was presented in Traud et. al. (2011) SIAM Review.

Now, without finding the bootstrap estimate of count statistics and its variance, finding the asymptotic distribution of these count statistics will not have been possible. So, now, with the help of the bootstrap based estimates we can perform hypothesis testing on the count statistics and provide their estimates of their asymptotic distribution.
7. Conclusion and Future Works. In this paper, we have considered three known subsampling schemes of networks and tried to show situations, where, they are applicable to find the asymptotic distribution of certain local statistics of the network. We consider the count of fixed subgraphs of the network as local statistics and call them count statistics. These have also been referred to as motif counts. We showed that the bootstrap subsample estimates of the count statistics and their smooth functions have asymptotic normal distribution. We proposed bootstrap schemes by which we could efficiently compute the asymptotic mean and variance of these count statistics. We also showed that the Sampling based bootstrap subsampling scheme seemed most stable and we recommend that scheme for use as bootstrap subsampling scheme.

We also use the estimated asymptotic mean and variances of the count statistics to construct hypothesis tests. These hypothesis tests can serve several purposes, such as

(a) Distinguish between the count statistics of two different networks
(b) Distinguish between parts of same network
(b) Testing whether a network has been generated from a specified model, by comparing the empirical and population version of the count statistic.
(c) Testing how close parameters of two different network models can become.

All of these different qualitative tests can be made quantitative by using hypothesis tests using the count statistics. We showed during simulations, that transitivity of networks from stochastic block models becomes easier to differentiate from transitivity of preferential attachment model as average degree grows. Similarly, in real networks, such as Facebook collegiate network, we show that certain covariate based subnetworks have more ‘cluster’ structure than others. We were also able to show that even in large networks conclusions based on means only as opposed to confidence statements using variances could be unreliable.

7.1. Future Works. Here we used bootstrap subsampling scheme to estimate local statistics only. But, one natural generalization can be use of bootstrap scheme to get asymptotic distribution of global statistics - such as graph cut, conductance, functionals of graphon (non-integral functionals) and such parameters. Sample and bootstrap estimates of such parameters are sometimes obtainable, but their theoretical properties are still unknown. It would be a nice future endeavor to extend bootstrap subsampling scheme to estimate such global characteristics of the networks.
References.


formatics 20 1746–1758.

Appendix.

A1. Variance of $\hat{T}_{b1}(R)$. The variance of $T^b(R)$ is

$$
\text{Var}_b \left[ \frac{1}{(\binom{m}{p})|\text{Iso}(R)|} \sum_{S \subseteq K_m, S \not\cong R} 1(S \subseteq H) \bigg| G \right] = \left( \frac{1}{(\binom{m}{p})|\text{Iso}(R)|} \right)^2 \text{Var}_b \left[ \sum_{S \subseteq K_m, S \not\cong R} 1(S \subseteq H) \bigg| G \right]
$$

$$
\text{Var}_b \left[ \sum_{S \subseteq K_m, S \not\cong R} 1(S \subseteq H) \bigg| G \right] = \mathbb{E}_b \left[ \left( \sum_{S \subseteq K_m, S \not\cong R} 1(S \subseteq H) \right)^2 \bigg| G \right] - \left( \mathbb{E}_b \left[ \sum_{S \subseteq K_m, S \not\cong R} 1(S \subseteq H) \bigg| G \right] \right)^2
$$
\[
\mathbb{E}_b \left[ \left( \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H) \right)^2 \mid G \right] = \mathbb{E}_b \left[ \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H) \mid G \right] + \mathbb{E}_b \left[ \sum_{S,T \subseteq K_m, S \ni T, S \ni R} 1(S,T \subseteq H) \mid G \right] \\
= I + II \quad \text{(Suppose)}
\]

Thus,

\[
I = \sum_{S \subseteq K_n, S \ni R} \frac{(n-p)}{\binom{n}{m}} 1(S \subseteq G)
\]

\[
II = \mathbb{E}_b \left[ \sum_{S,T \subseteq K_m, S \ni T, S \ni R} 1(S,T \subseteq H) \mid G \right]
\]

Now, a host of subgraphs can be formed by the intersection of two copies of \( R \). The number of intersected vertices can range from 0 to \( p - 1 \). Let us consider, that for number of vertices in intersection as \( k \) \( (k = 1, \ldots, (p - 1)) \), the number of graph structures that can be formed is \( g_k \) and we represent that graph structure by \( W_{jk} \), where, \( j = 1, \ldots, g_k \). Thus,

\[
II = \sum_{k=0}^{p-1} \sum_{j=1}^{g_k} \sum_{S \subseteq K_n, S \ni W_{jk}} \frac{(n-(2p-k))}{\binom{n}{m}} 1(S \subseteq G)
\]

So,

\[
\mathbb{E}_b \left[ \left( \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H) \right)^2 \mid G \right] = \sum_{S \subseteq K_n, S \ni R} \frac{(n-p)}{\binom{n}{m}} 1(S \subseteq G) \]

\[
+ \sum_{k=0}^{p-1} \sum_{j=1}^{g_k} \sum_{S \subseteq K_n, S \ni W_{jk}} \frac{(n-(2p-k))}{\binom{n}{m}} 1(S \subseteq G)
\]

\[
\text{Var}_b \left[ \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H) \mid G \right] = \sum_{S \subseteq K_n, S \ni R} \frac{(n-p)}{\binom{n}{m}} 1(S \subseteq G) \]

\[
+ \sum_{k=0}^{p-1} \sum_{j=1}^{g_k} \sum_{S \subseteq K_n, S \ni W_{jk}} \frac{(n-(2p-k))}{\binom{n}{m}} 1(S \subseteq G) - \left( \sum_{S \subseteq K_n, S \ni R} \frac{(n-p)}{\binom{n}{m}} 1(S \subseteq G) \right)^2
\]
A2. Degree-based non-uniform sampling bootstrap. In the non-uniform subsampling bootstrap scheme at each bootstrap iteration a subset of vertices of the full network \( G \) is selected by looking at the neighborhood graph of a randomly selected vertex and the graph induced by the selected subset is the subsample we consider. This is a vertex subsampling scheme and is a variant of common snowball sampling scheme. The full bootstrap procedure given the neighborhood size, \( d \) and number of bootstrap iterates, \( B \), is as follows –

1. For \( b \)th iterate of the bootstrap, \( b = 1, \ldots, B \),
2. Fix \( d = \) Depth of \( R \).
3. Choose \( i \in V(G) \), as the central vertex. Form the graph \( H \) as the neighborhood graph with \( i \) as root and depth \( d \). Let us denote \( |V(H)| \) by \( m \).
4. Calculate \( B_2(R) \), given by formula

\[
B_2(R) = \frac{n}{m} \frac{1}{\text{Is}(R)} \sum_{S \subseteq K_m, S \subseteq R} 1(S \subseteq H, i \in S)
\]
The bootstrap estimate of $T_G(R)$ is given by

$$\tilde{T}_{b3}(R) = \frac{1}{B} \sum_{b=1}^{B} T_{b3}(R)$$

In this sampling, all subgraphs $H$ do not have the same probability of being sampled. Actually some subgraphs may have zero probability of being selected. In uniform sampling bootstrap, we were considering all possible subgraphs with vertex count equal to $m$. In degree-based bootstrap we consider all subgraph $H$, which are $d$-neighborhood graph $a i \in V(G)$. But, unlike all possible subgraphs with vertex count equal to $m$, all $d$-neighborhood graph are not distinct and so we have considered that there are $n_R$ of them, where, $n_R \leq n$.

**Lemma 7.** The estimator $\tilde{T}_{b3}(R)$ has the following properties

(i) Given $G$, $\tilde{T}_{b3}(R)$ is an unbiased estimate of $T_G(R)$.

(ii) As $b \to \infty$, $\sqrt{n} \left( \rho^{-e} \tilde{T}_{b3}(R) - \rho^{-e} \hat{T}_G(R) \right) \overset{P}{\to} 0$

**Proof.** (i) Now, let us try to find the expectation of $T^b(R)$ under the sampling distribution conditional on the given data $G$.

$$\mathbb{E}_b \left[ \frac{n}{m(n)} \sum_{S \subseteq K_m, S \neq R} \mathbb{I}(S \subseteq H, i \in S) \right]$$

$$= \frac{n}{m(n)} \sum_{S \subseteq K_m, S \neq R} \mathbb{P}(H) \sum_{S \subseteq K_m, S \neq R} \mathbb{I}(S \subseteq H, i \in S)$$
Then, RHS is

\[
E_b \left[ \frac{n}{m(p)} |Iso(R)| \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H, i \in S) \bigg| G \right]
\]

\[
= \frac{n}{m(p)} |Iso(R)| \sum_{H \subseteq G} \frac{1}{n} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H, i \in S)
\]

\[
= \frac{n}{m(p)} |Iso(R)| \sum_{H \subseteq G_d} \frac{1}{n} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq G, i \in S)
\]

\[
= \frac{n}{m(p)} |Iso(R)| \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq G)
\]

So, we have,

\[
E_b[T_{b3}(R)|G] = T_G(R)
\]

(ii) Here, we use properties of the underlying model. Let us condition on \( \xi = \{\xi_1, \ldots, \xi_n\} \) and the whole graph \( G \) separately. Now, conditioning on \( \xi \), we get the main term of \( T_G(R) \) to be,

(7.3)

\[
E(\hat{P}(R)|\xi) = \frac{1}{n} \sum_{S \subseteq K_m, S \supseteq R} \left( \prod_{(i,j) \in E(S)} w(\xi_i, \xi_j) \right) + O(n^{-1}\lambda_n).
\]

We shall use the same decomposition as used in [7] of \( (\rho_n^{-e} \tilde{T}_{b3}(R) - \tilde{P}(R)) \) into

\[
(\rho_n^{-e} \tilde{T}_{b3}(R) - \tilde{P}(R)) = \rho_n^{-e} \left( \tilde{T}_{b3} - E_b[T_{b3}(R)|G] \right)
\]

\[
+ \rho_n^{-e} \left( T_G(R) - E(T_G(R)|\xi) \right)
\]

\[
+ E(\hat{P}(R)|\xi) \rho_n^{-e} - \tilde{P}(R)
\]

Let us define,

\[
U_3 = E(\hat{P}(R)|\xi) \rho_n^{-e} - \tilde{P}(R)
\]

\[
U_2 = \rho_n^{-e} \left( T_G(R) - E(T_G(R)|\xi) \right)
\]

\[
U_1 = \rho_n^{-e} \left( \tilde{T}_{b3} - E_b[T_{b3}(R)|G] \right)
\]
Now, it is easy to see that
\[
\text{Var}(\rho^{-e}\tilde{T}_{b3}(R)) = \mathbb{E}(\text{Var}(\rho^{-e}\tilde{T}_{b3}(R)|G) + \text{Var}(\mathbb{E}(\rho^{-e}\tilde{T}_{b3}(R)|G))
\]
\[
= \mathbb{E}(\text{Var}(\rho^{-e}\tilde{T}_{b2}(R) - T_G(R)|G) + \text{Var}(T_G(R))
\]
\[
= \mathbb{E}(\text{Var}(U_1|G)) + \mathbb{E}(\text{Var}(T_G(R)|\xi)) + \text{Var}(\mathbb{E}(T_G(R)|\xi))
\]
\[
= \mathbb{E}(\text{Var}(U_1|G)) + \mathbb{E}(\text{Var}(U_2|\xi)) + \text{Var}(U_3)
\]
We shall try to see the behavior of \(\text{Var}(U_1|G) = \text{Var}_b[\rho^{-e}\tilde{T}_{b3}(R)|G]\).

Now,
\[
\text{Var}_b[\rho^{-e}\tilde{T}_{b3}(R)|G] = \rho^{-2e} \frac{1}{B^2} \left( \sum_{b=1}^{B} \text{Var}_b[T_{b3}(R)] + \sum_{b,b'=1,b\neq b'} \text{Cov}_b(T_{b3}(R), T_{b3}(R)) \right)
\]

Now, the formula for \(\text{Var}_b[T_{b3}(R)]\) is given in Appendix A3. Note that, if \(E(H_b) \cap E(H_{b'}) = \phi\), then, \(\text{Cov}_b(T_{b3}(R), T_{b3}(R')) = 0\). The number of pairs such that \(E(H_b) \cap E(H_{b'}) \neq \phi\) is depends on the density of the nodes and size of \(R\). If we have \(\lambda|V(R)|\) and if \(\lambda|V(R)| = o(n)\) and also, the number of edges for the leading term in the covariance is equal to or more than \(2e\). So,
\[
\mathbb{E}(\text{Var}(U_1|G)) = o(n^{-1})
\]

Now, by proof of Theorem 1 in [7], we have,
\[
\text{Var}(U_2) = o(n^{-1})
\]
\[
\text{Var}(U_3) = o(n^{-1})
\]
So, we get, \(\text{Var}(\rho^{-2e}\tilde{T}_{b3}(R)) = o(n^{-1})\). Since, we already know \(\sqrt{n}\)-consistency of \((\rho_n^{-e}T_G(R) - P(R))\), this proves the \(\sqrt{n}\)-consistency of \(\rho_n^{-e}\tilde{T}_{b3}(R)\) to \(\rho_n^{-e}T_G(R)\).

\[\square\]

The variance of \(\tilde{T}_{b2}(R)\) given \(G\) can also be calculated and is given in the Appendix.

**A3. Variance of \(\tilde{T}_{b2}(R)\).** The variance of \(T^b(R)\) is
\[
\text{Var}_b \left[ \frac{n_R}{\binom{n}{p}|\text{Iso}(R)|} \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H, i \text{ central vertex of } S) \right] \bigg| G
\]
\[
= \left( \frac{n_R}{\binom{n}{p}|\text{Iso}(R)|} \right)^2 \text{Var}_b \left[ \sum_{S \subseteq K_m, S \supseteq R} 1(S \subseteq H, i \text{ central vertex of } S) \right] \bigg| G
\]
Thus, 
\[
I = \sum_{S \subseteq K_n, S \not\subset R} \frac{1}{n_R} 1(S \subseteq G)
\]

\[
II = \mathbb{E}_b \left[ \sum_{S \subseteq K_n, S \not\subset R} 1(S, T \subseteq H, i \text{ central vertex of } S, T) \bigg| G \right]
\]

Now, a host of subgraphs can be formed by the intersection of two copies of \( R \), having the same central vertex \( i \). The number of intersected vertices can range from 1 to \( d \), the depth of the graph \( R \). Let us consider, that for number of vertices in intersection as \( k \) \((k = 1, \ldots, (p - 1))\), the number of graph structures that can be formed is \( d_k \) and we represent that graph structure by \( R_{jk} \), where, \( j = 1, \ldots, d_k \). Thus,
\[
II = \sum_{k=1}^{d} \sum_{j=1}^{d_k} \frac{1}{n_R} 1(S \subseteq G)
\]
So,

\[ E_b \left[ \left( \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H) \right)^2 \right| G \right] = \sum_{S \subseteq K_n, S \ni R} \frac{1}{n_R} 1(S \subseteq G) + \sum_{k=1}^{d} \sum_{S \subseteq K_n, S \ni R_{jk}} \frac{1}{n_{R_{jk}}} 1(S \subseteq G) \]

\[ \text{Var}_b \left[ \sum_{S \subseteq K_m, S \ni R} 1(S \subseteq H, i \text{ central vertex of } S) \right| G \right] \]

\[ = \sum_{S \subseteq K_n, S \ni R} \frac{1}{n_R} 1(S \subseteq G) + \sum_{k=1}^{d} \sum_{S \subseteq K_n, S \ni R_{jk}} \frac{1}{n_{R_{jk}}} 1(S \subseteq G) \]

\[ - \left( \sum_{S \subseteq K_n, S \ni R} \frac{1}{n_R} 1(S \subseteq G) \right)^2 \]

So,

\[ \text{Var}_b [T^k(R)] = \left( \frac{n_R}{(n_p)|\text{Iso}(R)|} \right)^2 \left[ \sum_{S \subseteq K_n, S \ni R} \frac{1}{n_R} 1(S \subseteq G) \right] \]

\[ - \left( \frac{n_R}{(n_p)|\text{Iso}(R)|} \right)^2 \left[ \left( \sum_{S \subseteq K_n, S \ni R} \frac{1}{n_R} 1(S \subseteq G) \right)^2 \right] \]

\[ + \left( \frac{n_R}{(n_p)|\text{Iso}(R)|} \right)^2 \left[ \sum_{k=1}^{d} \sum_{S \subseteq K_n, S \ni R_{jk}} \frac{1}{n_{R_{jk}}} 1(S \subseteq G) \right] \]