The Annals of Statistics
2015, Vol. 43, No. 6, 2384–2411
DOI: 10.1214/15-AOS1338

SUBSAMPLING BOOTSTRAP OF COUNT FEATURES OF NETWORKS[1]

BY SHARMODEEP BHATTACHARYYA[*,†] AND PETER J. BICKEL[†]

University of California, Berkeley[*] and Oregon State University[†]

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 subsampling methods for finding empirical distribution of count features or "moments" (Bickel, Chen and Levina [*Ann. Statist.* **39** (2011) 2280–2301]) and smooth functions of these features for the networks. Using these methods, we cannot only estimate the 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 estimates of variance of the count features as well as show their efficacy through simulation. We also use the method on some real network data for estimation of variance and expectation of some count features.

1. Introduction. The study of networks has received recent increased attention, not only in social sciences, mathematics and statistics, but also in physics and computer science. With the information boom, a huge number of network data sets have appeared. In biology, gene regulation networks, protein–protein interaction networks, neural networks, ecological and epidemiological networks have become increasingly important. In social media, the Facebook, Twitter and Linkedin networks have come into prominence. Information networks have arisen in connection with text mining. Technological networks such as the Internet and many other networks related to Internet have also become objects of study.

In this paper, we consider a nonparametric formulation for network models where node labels carry no information. The model was proposed in Bickel and Chen [5] and has its origins in the works of Aldous [1] and Hoover [17]. Exchangeable probability models on infinite networks have a general representation based on the results of Aldous [1], Hoover [17], Kallenberg [18] and Diaconis and Janson [11]. The result is analogous to de Finetti's theorem. Note that numerical representation of networks come in the form of the adjacency matrix A, where $A_{ij} = 1$ if there is an edge from node i to j and 0 otherwise. We assume $A_{ii} = 0$;

Received February 2014; revised April 2015.

[1]Supported in part by NSF Grant DMS-11-60319.

MSC2010 subject classifications. Primary 62F40, 62G09; secondary 62D05.

Key words and phrases. Networks, subsampling, bootstrap, count features, model-based sampling.

that is, there are no self-loops. It is natural to assume exchangeable property for probability distribution on unlabeled random networks, which means that the probability distribution on the set of all adjacency matrices $\mathcal{L}\{[Aij], i, j \geq 1\}$ satisfy $\mathcal{L}\{[A_{ij}]\} = \mathcal{L}\{[A_{\sigma_i\sigma_j}]\}$, where σ is an arbitrary permutation function on node indices. Such exchangeable probability distributions on random infinite binary arrays can be characterized as follows: for $i, j \geq 1$,

$$\begin{aligned} \alpha, \xi_i, \eta_{ij} &\overset{\text{i.i.d.}}{\sim} U(0,1), \\ A_{ij} &= f(\alpha, \xi_i, \xi_j, \eta_{ij}), \end{aligned}$$

where, $f : [0,1]^4 \to [0,1]$ is a measurable function, symmetric in its second and third arguments and $\eta_{ij} = \eta_{ji}$. α, as in de Finetti's theorem, corresponds to the mixing distribution and is not identifiable. This representation is not unique, and f is not identifiable. These distributions can be parametrized through the function

$$h(u, v) = P[A_{ij} = 1 | \xi_i = u, \xi_j = v].$$

The function h is still not unique, but it can be shown that if two functions h_1 and h_2 define the same distribution $\mathcal{L}$, they can be related through a measure-preserving transformation. This leads to the Bickel and Chen [5] characterization of "nonparametric" unlabeled graph models, which is closely related to Lovász's notion of "graphons" [21]. The model will be described in more detail in Section 2. Other researchers have also studied similar, general classes of models, such as the *latent space models* of Hoff, Raftery and Handcock [15] and the *inhomogeneous random graph models* of Bollobás, Janson and Riordan [8]. Many previously studied probability models for networks fall into this class. The class includes the stochastic block models (Holland, Laskey and Leinhardt [16], Nowicki and Snijders [25]) and the configuration model (Chung and Lu [9]). Dynamically defined models such as the "preferential attachment" models (which seem to have been first mentioned by Yule in the 1920s and given its modern name by Barabási and Albert [2]) can also be thought of in this way if the dynamical construction process continues forever, producing an infinite graph. More details are given in Section 5.

Motifs or count statistics are the main statistics that we consider in this paper. Count statistics can be defined as smooth functions of counts of subgraphs in the network. Counts of special subgraphs have been extensively used in the network literature for analyzing network behavior [3, 23, 27]. The count statistics have appeared earlier under the names *motif* counts in biology [24] and *subgraph* counts in probability [21]. It also follows from the work of Lovász [21], Diaconis and Janson [11] and in part from Bickel and Chen [5] that there is a unique set of statistics whose joint distribution characterize the probability distribution on unlabeled networks. These statistics, called *empirical moments* by Bickel, Chen and Levina [6] are the counts of subgraphs in the network. The subgraphs most used are small cycles like triad, tetrad and small acyclic graphs.

The expectation and variances of count statistics can, in principle, be computed (Picard et al. [26]) and more usefully be asymptotically approximated [6]. Under appropriate conditions, normalized count statistics have limiting Gaussian distribution. They have many uses [3, 31, 32], particularly in distinguishing between the mechanisms generating different graphs as well as providing characterization of network distributions. The general asymptotic Gaussian distribution of count statistics was provided in Theorem 1 of [6] with an expression for the asymptotic mean and variance; however, the paper provided no way to calculate the quantities.

Motifs or count statistics have been used in testing equality of features of networks and finding confidence intervals of the count features [22, 28]. However, a major stumbling block in their use has been the calculation of motifs that have even moderately large number of vertices (i.e., more than five) and even more challenging problem of finding estimates of their variances. Finding the correct count statistics or motifs 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 the number of vertices in the network is even in thousands, the computation becomes difficult; if it is in millions, the computation becomes infeasible. Using subsampling methods to calculate the count statistics, we can greatly reduce the computational burden of computing the statistics and inference using them.

In the statistical literature on networks, some work has been done on devising sampling designs to select network samples. Various sampling designs have been proposed in the statistical and computer science literature to derive *representative* samples of a given network; see [19, 20] and [29]. Many of these sampling designs have been analyzed from the design-based sampling point of view [13, 30]. Some of these methods have been analyzed from a model-based sampling point of view, where mostly the *exponential random graph model (ERGM)* has been considered as the model generating the network, and a likelihood-based approach has been taken for inference [14]. As a result, only parametric inference was possible. On the other hand, our approach is not restricted to parametric models as we try to estimate the certain functionals of the underlying nonparametric generating model, using the samples obtained from the network data.

1.1. *Contribution and structure of our work.* We use subsampling-based bootstrap approaches to estimate the count statistics as well as find the approximate distributions for such count statistics under the general model of Bickel and Chen [5].

Along with the bootstrap methods and their theoretical analysis, we give two examples where the use of count statistics provides some useful insights into the behavior of the networks. One of the two examples is the Jefferson High School network given in Bearman et al. [3], and the other example uses the Facebook collegiate networks provided in Traud et al. [19]. The high school network is a nice example where counts of specific types of subgraphs in the network and their confidence intervals based on different generating models give us useful insight into

the behavior of nodes in the network [3]. The Facebook collegiate networks are larger and denser networks, and calculation of count statistics for these networks would not be computationally feasible without the use of subsampling methods.

In Section 2 we outline our main results. In Section 3 we describe the bootstrap subsampling methods and the theoretical properties of each bootstrap estimator. We also indicate a method for estimating asymptotic variances of these estimators using bootstrap. Additionally, we give a theoretical comparison of the methods. In Section 4, we give the general theorem on asymptotic Gaussianity of bootstrap subsampling estimates count statistics and their variance. In Section 5 we perform a simulation study under two special cases of the general "nonparametric" model: the stochastic block model and the preferential attachment model, respectively. In Section 6 we apply our method to test hypotheses about the count statistics of real networks.

2. Model and statistics. We consider a random unlabeled graph G_n as the data. Let $V(G_n) = \{v_i, \dots, v_n\}$ denote the vertices of G_n and $E(G_n)$ denote the set of edges of G_n. Thus the number of vertices in G_n is $|V(G_n)| = n$. We shall only consider undirected, unweighted graphs in this paper. For the sake of notational simplicity, we may denote G_n by G.

As usual we suppose the network is represented by an adjacency matrix $A_{n\times n}$ whose elements are $A_{ij} \in \{0, 1\}$,

$$A_{ij} = \begin{cases} 1, & \text{if node } i \text{ links to node } j, \\ 0, & \text{otherwise.} \end{cases}$$

A finite sample version of the Aldous–Hoover representation for exchangeable adjacency matrices $A_{n\times n}$ becomes, for $i, j \in \{1, \dots, n\}$,

$$\begin{aligned} \alpha, \xi_i, \eta_{ij} &\overset{\text{i.i.d.}}{\sim} U(0, 1), \\ A_{ij} &= f_n(\alpha, \xi_i, \xi_j, \eta_{ij}), \end{aligned}$$

where, $f_n : [0, 1]^4 \to [0, 1]$ is a measurable function, symmetric in its second and third arguments and $\eta_{ij} = \eta_{ji}$. Note that this is not a representation of all exchangeable probability distributions on finite networks.

Bickel and Chen [5] considered a special form of the general Aldous–Hoover representation,

$$h_n(u, v) \equiv \mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v).$$

The above-mentioned form can be simplified by decoupling n from the contribution of (ξ_i, ξ_j). Thus h_n is modeled as product of a scale function in terms of n, ρ_n, defined as

$$\rho_n = \int_0^1 \int_0^1 h_n(u, v)\, du\, dv$$

and a bivariate function independent of n, the *latent variable density*, $w(\xi_i, \xi_j)$. We call the resulting model a *nonparametric latent variable model*, and the model equation described in Bickel, Chen and Levina [6] becomes, for $i, j \in \{1, \dots, n\}$, $\xi_i \overset{\text{i.i.d.}}{\sim} U(0,1)$ and

$$\mathbb{P}(A_{ij} = 1 | \xi_i = u, \xi_j = v) = h_n(u,v) = \rho_n w(u,v) \mathbf{1}\big(w \le \rho_n^{-1}\big), \tag{2.1}$$

where $w(u,v) \ge 0$, symmetric, $0 \le u, v \le 1$, $\int\int w(u,v)\,du\,dv = 1$, $0 < \rho_n < 1$ and we define expected degree $\lambda_n = n\rho_n$.

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 $R \subseteq S$. The integral parameter corresponding to a subgraph R is defined as $P(R)$,

$$P(R) = \mathbb{E}\Bigg\{ \prod_{(i,j)\in E(R)} h(\xi_i, \xi_j) \prod_{(i,j)\in E(\bar{R})} \big(1 - h(\xi_i, \xi_j)\big) \Bigg\}, \tag{2.2}$$

where $\bar{R}$ is a subgraph of K_p (K_p is a complete graph on p vertices) with $V(\bar{R}) = \{i, j : (i,j) \notin E(R), i, j \in V(R)\}$ and $E(\bar{R}) = \{(i,j) : (i,j) \notin E(R), i \in V(R), j \in V(R)\}$.

Now, the empirical statistic corresponding to $P(R)$, which is the count statistics for subgraph R, is

$$\hat{P}(R) = \frac{1}{\binom{n}{p} |\operatorname{Iso}(R)|} \sum_{S \subseteq K_n, S \cong R} \mathbf{1}(S \subseteq G), \tag{2.3}$$

where $\operatorname{Iso}(R)$ is the group of isomorphisms of R, and K_n is the complete graph on n vertices.

We also have from [6]

$$\mathbb{E}\big(\hat{P}(R)\big) = P(R).$$

Examples of subgraphs and corresponding count statistics include the following:

EXAMPLE 1. $R =$ edge is a subgraph with two vertices and one edge connecting them, so $\hat{P}(R) = \frac{1}{n(n-1)} \sum_{i=1}^n D_i$, where $D_i =$ degree of v_i, $v_i \in V(G)$. $P(R) = \int_0^1 h_n(u,v)\,du\,dv$.

EXAMPLE 2. $R =$ triangle is a 3-clique subgraph, so $\hat{P}(R) = \frac{1}{\binom{n}{3}}$ total number of unique 3-clique subgraphs in G_n,

$$P(R) = \int_0^1 \int_0^1 \int_0^1 h_n(u,v) h_n(v,s) h_n(s,u)\,du\,dv\,ds.$$

EXAMPLE 3. We define a smooth function of counts of triangles and "V's," known as *transitivity*, T_{Tr}, as

$$\hat{P}_{\mathrm{Tr}} = \frac{\hat{\rho}_n^{-3}\hat{P}(R_1)}{\hat{\rho}_n^{-3}\hat{P}(R_1) + \hat{\rho}_n^{-2}\hat{P}(R_2)},$$

where R_1 is a triangle or a 3-cycle, and R_2 is a "V" or a path with three vertices and $\hat{\rho}_n = \hat{P}(\text{edge})$.

EXAMPLE 4. $R = p$-cycle is a cyclic subgraph with $|V(R)| = p$, $|E(R)| = p$, and R is a *ring* containing all p vertices. Triangle is a 3-cycle. $P(R) = \int_0^1 \cdots \int_0^1 h_n(u_1, u_2) \cdots h_n(u_{p-1}, u_p) h_n(u_p, u_1)\, du_1 \cdots du_p$.

DEFINITION 5 (Wheels). A (k, l)-wheel is an acyclic graph with $kl + 1$ vertices and kl edges and "hub" vertex (say, $\{1\}$), isomorphic to the graph with edges $\{((1, 2), (2, 3), \ldots, (k, k + 1))((1, k + 2), (k + 2, k + 3), \ldots, (2k, 2k + 1)), \ldots, ((1, (l - 1)k + 2), ((l - 1)k + 2, (l - 1)k + 3), \ldots, (lk, lk + 1))\}$.

Edges, "V," "W" are examples of (k, l)-wheels. An edge is a $(1, 1)$-wheel, a "V" is a $(1, 2)$-wheel and a "W" is a $(2, 2)$-wheel.

Now, as $\rho_n \to 0$, which is the case for graphs which are *not* fully dense, that is, $|E(G_n)| = O_P(n^2)$, $P(R) \to 0$ as well as its estimator $\hat{P}(R) \xrightarrow{P} 0$ and the asymptotics on $(\hat{P}(R) - P(R))$ become uninformative. So, in order to get a proper analysis of the behavior of $\hat{P}(R)$ in relation to $P(R)$, we have to appropriately normalize both $P(R)$ and $\hat{P}(R)$. The normalized versions of parameter $P(R)$ are defined as

$$\tilde{P}(R) = \rho_n^{-e} P(R), \tag{2.4}$$

where $e \equiv |E(R)|$. Then we define the corresponding normalized statistic to be

$$\hat{T}(R) = \hat{\rho}^{-e} \hat{P}(R), \tag{2.5}$$

where

$$\hat{\rho} = \frac{\bar{D}}{n - 1}, \tag{2.6}$$

where $D_i =$ degree of v_i, $v_i \in V(G_n)$ for $i = 1, \ldots, n$ and $\bar{D} = \frac{1}{n}\sum_{i=1}^n D_i$. Now the investigation on asymptotic behavior of $\sqrt{n}(\hat{T}(R) - \tilde{P}(R))$ is possible, as both terms are asymptotically nonzero quantities. This investigation was done in [6].

We wish to approximate the statistic $\hat{P}(R)$ and functional $\mathrm{Var}(\hat{P}(R))$ by nonparametric bootstrap. We consider two bootstrap procedures:

(I) the *uniform subsampling* bootstrap procedure and
(II) the *subgraph subsampling* bootstrap procedure.

How we get the bootstrap estimates will be discussed in next section, and we will state theorems justifying the use of these bootstrap estimations in next two sections.

2.1. *Bootstrap and model-based sampling*. Our work can be viewed from two different perspectives. The first perspective is that of the *bootstrap*. In *nonparametric 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 approximate estimates of functionals of the underlying population model generating the given network. Note that here we are interested in the *subsampling*, not the *resampling* of a network. Our use of the bootstrap corresponds to Efron's [12] use of bootstrap for approximations made by Monte Carlo quantities, which, in principle, could be calculated using data such as the empirical variance of complicated estimates.

The second perspective is that of *sampling*. In 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. We use subsampling bootstrap or sampling of network data to get estimates for population quantity (count statistics) and underlying probability model (integral parameter).

3. Bootstrap methods. We consider two different bootstrap methods. Both of the methods of bootstrap consider finding subsamples from the whole network given as the data. In the following subsections, we shall define each of these subsampling bootstrap methods. We shall also compare the theoretical performance between the two bootstrap schemes.

Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. 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 of vertices is the subnetwork we consider. This is a vertex subsampling or induced network sampling scheme. Given subnetwork size m and number of bootstrap iterates B, the *uniform subsampling* bootstrap scheme has the following steps:

(1) For the bth 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 $\hat{P}_{b1}(R)$, given by formula

$$\hat{P}_{b1}(R) = \frac{1}{\binom{m}{p}|\operatorname{Iso}(R)|} \sum_{S\subseteq K_m, S\cong R} \mathbf{1}(S \subseteq H). \tag{3.1}$$

The *uniform subsampling* bootstrap estimate of $\hat{P}(R)$ is given by

$$\bar{P}_{B1}(R) = \frac{1}{B}\sum_{b=1}^{B} \hat{P}_{b1}(R). \tag{3.2}$$

The uniform subsampling bootstrap scheme is the network version of the common subsampling bootstrap scheme seen in Bickel et al. [7]. Note that there are other ways of forming uniformly subsampled bootstrap estimates, as mentioned in [7]; however, we just mention one in this discourse.

For the bootstrap method, we prove a theorem of following type:

THEOREM 1. *Suppose R is fixed acyclic or p-cycle with $|V(R)| = p$ and $|E(R)| = e$:*

(i) *given G, $\hat{P}_{b1}(R)$ is an unbiased estimate of $\hat{P}(R)$;*
(ii) *given G, $\mathrm{Var}(\rho_n^{-e}\hat{P}_{b1}(R)|G) = O(\frac{1}{m^p\rho_n^e} \vee \frac{1}{m})$;*
(iii) *also, if $B \to \infty$, $n \to \infty$, $m \to \infty$, $m/n \to 0$ and $B(m^p\rho_n^e \wedge m) > O(n)$, under G generated from (2.1),*

$$\sqrt{n}\big(\rho_n^{-e}\bar{P}_{B1}(R) - \rho_n P(R)\big) \xrightarrow{P} 0. \tag{3.3}$$

PROOF. The proof is given in Appendix A2 in [4]. □

3.2. *Subgraph subsampling bootstrap.* In the *subgraph subsampling* bootstrap scheme, we use an enumeration scheme to find all possible subgraphs R of size $|V(R)| = p$ in the graph G. Then we convert the enumeration scheme into a sampling scheme by selecting each subgraph R of size p of G with a fixed probability and counting the number of sampled subgraphs. The enumeration scheme was proposed by Wernicke et al. [33]. A random version of the enumeration scheme was also proposed in [33]. 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. [33], which we shall henceforth call ESU. The enumeration algorithm is a *breadth-first search* algorithm. The enumeration scheme creates a forest of tree structures such that each tree corresponds to one vertex of the network G, and each leaf of each tree is a size-p subgraph [we have $|V(R)| = p$] of G. 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_1). We get such a particular fixed ordering of the vertices of the network with v_1 getting lowest order value and subsequently, searched vertices getting higher order values. The ordering is described in the algorithm ASSIGN ORDER or AO 1, where, given any set of vertices $\mathcal{V}$, we denote the set of vertices connected to $\mathcal{V}$, that is, the *neighbors* of $\mathcal{V}$, by $N(\mathcal{V})$. Also, based on the ordering defined by AO, we denote $v_i \succ v_j$, if v_i has a higher order than v_j.

Algorithm 1 ASSIGNORDER(G, p)

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

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

The enumeration algorithm starts with an available vertex of lowest possible order (where order is specified by Algorithm AO 1), say v_1. We construct a tree with the vertex v_1 as the root node. We consider v_1 as the "parent" node and neighbors of v_1, which have higher order than v_1, as its "children." In the next step, the "children" node becomes the "parent" node in the tree and has its own neighbors, which have higher order than the nodes that have already come into the tree as their "children." We define $N_{\text{excl}}(v, \mathcal{V})$ (v is a vertex, and $\mathcal{V}$ is a set of vertices) for $N(v) \setminus \mathcal{V}$. The tree is allowed to grow up to a height p if we are counting size-p subgraphs. Thus 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 v_1. For each vertex, we have such a tree, and over counting is averted as we maintain the order of vertices assigned by Algorithm AO 1 while forming the trees. So, with the help of the particular ordering of vertices, each of the size-p subgraphs ($|V(R)| = p$) is counted only once.

The randomized enumeration Algorithm RAND-ESU 2 also creates a forest of tree structures such that each tree corresponds to one vertex of the network G, and each leaf of each tree is a size-p subgraph [we have $|V(R)| = p$] of G. However, only a random selection of leaves of each tree is present in RAND-ESU with uniform probability of selection of each leaf. The random enumeration algorithm starts with an available vertex of lowest possible order (where, order is specified by Algorithm AO 1), say v_1, chosen with probability q_1. We construct a tree with the vertex v_1 as the root node. We consider v_1 as the "parent" node and neighbors of v_1, which have a higher order than v_1 as its "children" and each "child" is selected with probability q_2 independently. In the next step, the "children" nodes become the "parent" nodes in the tree and has their own neighbors, which have higher order than the nodes that have already come into the tree, as their "children," and each "child" is selected with probability q_3. The tree is allowed to grow up to a height

Algorithm 2 RANDOMIZEDENUMERATESUBGRAPH(G, p)

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 $d = 1, \ldots, p$.

Output: $\mathcal{S}_p^R$ = A sample of subgraphs, R of G, such that $|R| = p$.

1: **for** each vertex $v \in V$ **do**
2: $V_{\text{Extension}} \leftarrow \{u \in N(\{v\}) : u \succ v\}$
3: $d \leftarrow 1$
4: **With probability** q_d **Call** RandExtendSubgraph($\{v\}$, $V_{\text{Extension}}$, v, d})
5: **end for**
6: **function** RANDEXTENDSUBGRAPH(V_{Subgraph}, $V_{\text{Extension}}$, v, d)
7: **Input:** Graphs V_{Subgraph}, $V_{\text{Extension}}$ and vertex v.
8: **Output:** A sample of subgraphs, R of G, such that $|V(R)| = p$ and v is a vertex of R.
9: **if** $|V_{\text{Subgraph}}| = p$ **then**
10: **return** Subgraph of G induced by V_{Subgraph}
11: **else**
12: **while** $V_{\text{Extension}} \neq \phi$ **do**
13: Remove an arbitrarily chosen vertex w from $V_{\text{Extension}}$
14: $V'_{\text{Extension}} \leftarrow V_{\text{Extension}} \cup \{u \in N_{\text{excl}}(w, V_{\text{Subgraph}}) : u \succ v\}$
15: $d \leftarrow |V_{\text{Subgraph}}| + 1$
16: **With probability** q_d **Call** RandExtendSubgraph($V_{\text{Subgraph}} \cup \{w\}$, $V'_{\text{Extension}}$, v, d)
17: **end while**
18: **end if**
19: **return**
20: **end function**

p if we are counting size-p subgraphs, and at step d, the probability of selection is q_d. 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. So, with the help of the particular ordering of vertices, a subsample of the size-p subgraphs ($|V(R)| = p$) is obtained. The pseudo-code is given in Algorithm 2.

The ordering is needed for success of the ESU algorithm and its randomized counterpart 2. We formally state the subsampling algorithm, RAND-ESU 2 in this paper with an extra set of parameters $(q_1, \ldots, q_p)$. The enumeration version can be found in [33].

From the sampling scheme RAND-ESU we have a sample $\mathcal{S}_p^R$ 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 $\mathcal{S}_p$, then we can try to calculate the *inclusion probability* of each item in the sample $\mathcal{S}_p^R$.

The item $S \in \mathcal{S}_p$ is a subgraph of G induced by the set of vertices $\{w_1, \ldots, w_p\}$, where we take that $w_{i+1} \succ w_i$, $i = 1, \ldots, p-1$. Thus:

$$\begin{aligned}\pi \equiv \text{Inclusion probability of } S &= \mathbb{P}\big[(w_1, \ldots, w_p) \text{ is selected}\big] \\ &= \mathbb{P}\big[w_p | (w_1, \ldots, w_{p-1}) \text{ is selected}\big] \\ &\quad \times \mathbb{P}\big[(w_1, \ldots, w_{p-1}) \text{ is selected}\big] \\ &= q_p \cdot \mathbb{P}\big[(w_1, \ldots, w_{p-1}) \text{ is selected}\big] \\ &= q_p \cdot q_{p-1} \cdot \mathbb{P}\big[(w_1, \ldots, w_{p-2}) \text{ is selected}\big] \\ &= \cdots = q_p \cdot q_{p-1} \cdots q_1 = \prod_{d=1}^{p} q_d.\end{aligned}$$

So, each item $S \in \mathcal{S}_p$ has an inclusion probability π to be in the sample $\mathcal{S}_p^R$.

In Theorem 2 of [33] it was proved that the output of the ESU algorithm $\mathcal{S}_p$ contains all subgraphs R of G, such that $|V(R)| = p$, exactly once. Thus we can write statistic (2.3) for a specific subgraph R with $|V(R)| = p$ in the following way:

$$\hat{P}(R) = \frac{1}{\binom{n}{p}} \sum_{S \in \mathcal{S}_p} \mathbf{1}(S \cong R). \tag{3.4}$$

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 $\hat{P}(R)$ given G. To meet this goal we use a sampling version of the enumeration scheme ESU.

Now we have a sampling scheme by which we select a sample $\mathcal{S}_p^R$ from the population $\mathcal{S}_p$, where each element of $\mathcal{S}_p$ has probability of inclusion of π. Thus we can define a *Horvitz–Thompson* estimator (for reference, see Chapter 6.2 of [29]) of $\hat{P}(R)$ based on $\mathcal{S}_p^R$ as

$$\hat{P}_{b2}(R) = \frac{1}{(\prod_{d=1}^{p} q_d)\binom{n}{p}} \sum_{S \in \mathcal{S}_p^R} \mathbf{1}(S \cong R). \tag{3.5}$$

Now if we repeat the same procedure B number of times, each time getting independent copies of $\mathcal{S}_p^R$ with replacement from $\mathcal{S}_p$, we can get the *subgraph subsampling* bootstrap estimate,

$$\bar{P}_{B2}(R) = \frac{1}{B} \sum_{b=1}^{B} \hat{P}_{b2}(R). \tag{3.6}$$

For the bootstrap method, we prove a theorem of following type:

THEOREM 2. *Suppose R is fixed acyclic or p-cycle with $|V(R)| = p$ and $|E(R)| = e$:*

(i) *given G, $\hat{P}_{b2}(R)$ is an unbiased estimate of $\hat{P}(R)$;*

(ii) *given* G, $\operatorname{Var}(\rho_n^{-e}\hat{P}_{b2}(R)|G) = O((\frac{1}{q_1}-1)\frac{1}{n} + \frac{1}{n\rho_n^{e-p+1}}\cdot\prod_{d=2}^{p}\frac{1}{\lambda_n q_d})$;

(iii) *for* $B\to\infty$ *and* $q_d\to 0$ *for all* $d=1,\dots,p$ *such that* $\frac{1}{B}(\frac{1}{q_1}-1)\to 0$ *and* $B\prod_{d=2}^{p} q_d \geq \frac{1}{n^{p-1}\rho_n^e}$ *and* $n\to\infty$, $\lambda_n\to\infty$, *and under* G *generated from* (2.1),

$$\sqrt{n}\big(\rho_n^{-e}\bar{P}_{B2}(R) - \rho_n^{-e}P(R)\big) \xrightarrow{P} 0. \tag{3.7}$$

PROOF. The proof is given in Appendix A3 in [4]. □

Note that the main reason for taking repeated independent samples, $\mathcal{S}_p^R$ for this case, is to reduce the variance of the bootstrap estimates and to make the estimates more stable.

3.3. *Estimation of variance and covariance.* We first start with the situation when the source of variation is only the randomness coming from sampling from the underlying model (2.1). We denote $\operatorname{Var}[\rho^{-e}\hat{P}(R)]$ as $\sigma^2(R;\rho)$ and $\operatorname{Cov}(\rho^{-e_1}\hat{P}(R_1), \rho^{-e_2}\hat{P}(R_2))$ as $\sigma(R_1,R_2;\rho)$. Note that, $e_1 = |E(R_1)|$, $e_2 = |E(R_2)|$, $p_1 = |V(R_1)|$ and $p_2 = |V(R_2)|$.

PROPOSITION 6. *For connected subgraphs* R, R_1 *and* R_2 *of* G, *we have that*

$$\begin{aligned}
\sigma^2(R;\rho) &= \frac{1}{(\rho^e\binom{n}{p}|\operatorname{Iso}(R)|)^2}\sum_{\substack{W:W=S\cup T\\ S,T\cong R, S\cap T\neq\varnothing}}\mathbb{E}\Big[\sum_{W\subseteq K_n}\mathbf{1}(W\subseteq G)\Big]\\
&\quad -\Big(1-\frac{((n-p)!)^2}{n!(n-2p)!}\Big)\big(\tilde{P}(R)\big)^2,\\
\sigma(R_1,R_2;\rho) &= \frac{1}{(\rho^{e_1+e_2}\binom{n}{p_1}\binom{n}{p_2}|\operatorname{Iso}(R_1)||\operatorname{Iso}(R_2)|)}\\
&\quad\times\sum_{\substack{W:W=S\cup T,\\ S\cong R_1, T\cong R_2, S\cap T\neq\varnothing}}\mathbb{E}\Big[\sum_{W\subseteq K_n}\mathbf{1}(W\subseteq G)\Big]\\
&\quad -\Big(1-\frac{(n-p_1)!(n-p_2)!}{n!(n-p_1-p_2)!}\Big)\tilde{P}(R_1)\tilde{P}(R_2).
\end{aligned}$$

PROOF. The proof is given in Appendix B1 in [4]. □

Note that if we take $k=|V(W)|$ and $e_W\equiv|E(W)|$, then $k=p,\dots,2p-1$ and each term of sum on the RHS of the previous equation is

$$\begin{aligned}
\frac{1}{(\rho^e\binom{n}{p}|\operatorname{Iso}(R)|)^2}\mathbb{E}\Big[\sum_{W\subseteq K_n}\mathbf{1}(W\subseteq H)\Big] &= \frac{\rho^{e_W}\binom{n}{k}|\operatorname{Iso}(W)|}{(\rho^e\binom{n}{p}|\operatorname{Iso}(R)|)^2}\tilde{P}(W)\\
&= O\big(n^{k-2p}\rho^{e_W-2e}\big).
\end{aligned} \tag{3.8}$$

We can analyze each such term separately:

(1) If $k = |V(W)| = 2p - 1$, then W is a connected graph, with $e_W = 2e$. Thus we have that the main leading term equals $O(\frac{1}{n})$.

(2) In the case $k = |V(W)| < (2p - 1)$:

- If R is acyclic, then $e_W - 2e \le k - 2p - 1$ since $e = p - 1$, so $O(n^{k-2p} \times \rho^{e_W - 2e}) = o(n^{-1})$ if $\lambda_n = n\rho_n \to \infty$.
- If R is a p-cycle, $e_W - 2e = k - 2p < 0$ if $k = |V(W)| = p$ and $e_W - 2e \le k - 2p - 1$ if $p < k < (2p - 1)$, so $O(n^{k-2p}\rho^{e_W-2e}) = O(\lambda_n^{-p}) + o(n^{-1})$ if $\lambda_n \to \infty$.
- If R is any other cyclic graph, $O(n^{k-2p}\rho^{e_W-2e}) = O(n^{-c}\rho^{-d})$, where $0 < c \le p$ and $0 < d \le c(c-1)/2$ for each c. So, in order to have $n^{-c}\rho^{-d} \le Mn^{-1}$, the worst rate that λ_n can have is $\lambda_n = O(n^{1-2/p})$.

For connected and acyclic or p-cycle R, R_1 and R_2, we get that

$$\sigma^2(R; \rho) = O\left(\frac{1}{n} \vee \frac{1}{\lambda_n^p}\right),$$

$$\sigma(R_1, R_2; \rho) = O\left(\frac{1}{n}\right).$$

So, for calculation of variance, if R is acyclic or p-cycle, we only estimate the count of the features which are $W = S \cup T$ and $|V(W)| = 2p - 1$ and $|V(W)| = p$. Thus using the expansion given in Proposition 6, the empirical estimator of $\sigma^2(R; \rho)$ is defined as

$$\hat{\sigma}^2(R) = \frac{1/(1-x)}{(\hat{\rho}_n^e \binom{n}{p} |\operatorname{Iso}(R)|)^2} \sum_{\substack{W \subseteq K_n : W = S \cup T, \\ S, T \cong R, |S \cap T| = 1, p}} \mathbf{1}(W \subseteq G) - \frac{x\hat{\rho}_n^{-2e}\hat{P}(R)^2}{(1-x)}, \tag{3.9}$$

where $x = (1 - \frac{((n-p)!)^2}{n!(n-2p)!})$, and using the expansion given in Proposition 6, the empirical estimator of $\sigma(R_1, R_2; \rho)$ is defined as

$$\begin{aligned}\hat{\sigma}(R_1, R_2) = {} & \frac{1/(1-y)}{(\hat{\rho}_n^{e_1+e_2}\binom{n}{p_1}\binom{n}{p_2}|\operatorname{Iso}(R_1)||\operatorname{Iso}(R_2)|)} \\ & \times \sum_{\substack{W \subseteq K_n, W = S \cup T, \\ S \cong R_1, T \cong R_2, |S \cap T| = 1}} \mathbf{1}(W \subseteq G) - \frac{y\hat{\rho}_n^{-(e_1+e_2)}\hat{P}(R_1)\hat{P}(R_2)}{(1-y)},\end{aligned} \tag{3.10}$$

where $y = (1 - \frac{(n-p_1)!(n-p_2)!}{n!(n-p_1-p_2)!})$.

$\hat{\sigma}^2(R)$ and $\hat{\sigma}(R_1, R_2)$ become consistent estimates of $\sigma^2(R; \rho)$ and $\sigma(R_1, R_2; \rho)$ as well as $\sigma^2(R; \hat{\rho})$ and $\sigma(R_1, R_2; \hat{\rho})$, respectively.

LEMMA 7. *As $\lambda_n \to \infty$ and $n \to \infty$, if R, R_1, R_2 is connected acyclic or p-cycle, then additionally $\lambda_n^p \geq O(n)$,*

$$\frac{\hat{\sigma}^2(R)}{\sigma^2(R;\hat{\rho})} \xrightarrow{P} 1, \tag{3.11}$$

$$\frac{\hat{\sigma}(R_1,R_2)}{\sigma(R_1,R_2;\hat{\rho})} \xrightarrow{P} 1. \tag{3.12}$$

PROOF. The proof is given in Appendix B2 in [4]. □

Now we can see that $\hat{\sigma}^2(R)$ and $\hat{\sigma}(R_1,R_2)$ are nothing but count statistics on the statistic $W = S \cup T$. So, using bootstrap methods, we define a bootstrap-based estimate of $\hat{\sigma}^2(R)$, for $i = 1, 2$,

$$\begin{aligned}\hat{\sigma}_{Bi}^2(R) = &\sum_{\substack{W=S\cup T, S, T\cong R,\\ |S\cap T|=1,p}} \frac{(\hat{\rho}_n^{eW}\binom{n}{p_W}|\operatorname{Iso}(R)|)}{(1-x)(\hat{\rho}_n^{e}\binom{n}{p}|\operatorname{Iso}(R)|)^2}\bar{P}_{Bi}(W)\\ &- \frac{x\hat{\rho}_n^{-2e}\bar{P}_{Bi}(R)^2}{(1-x)},\end{aligned} \tag{3.13}$$

where $x = (1 - \frac{((n-p)!)^2}{n!(n-2p)!})$. A bootstrap-based estimate of $\hat{\sigma}(R_1,R_2)$ is

$$\begin{aligned}&\hat{\sigma}_{Bi}(R_1,R_2)\\ &= \sum_{\substack{W=S\cup T, S\cong R_1,\\ T\cong R_2, |S\cap T|=1}} \frac{(\hat{\rho}_n^{eW}\binom{n}{p_W}|\operatorname{Iso}(W)|)}{(1-y)(\hat{\rho}_n^{e_1+e_2}\binom{n}{p_1}\binom{n}{p_2}|\operatorname{Iso}(R_1)||\operatorname{Iso}(R_2)|)}\bar{P}_{Bi}(W)\\ &\quad - \frac{y\hat{\rho}_n^{-(e_1+e_2)}\bar{P}_{Bi}(R_1)\bar{P}_{Bi}(R_2)}{(1-y)},\end{aligned} \tag{3.14}$$

where $y = (1 - \frac{(n-p_1)!(n-p_2)!}{n!(n-p_1-p_2)!})$ and $\bar{P}_{Bi}(W)$ $(i = 1, 2)$ are bootstrap count statistics estimates, defined in equations (3.2) and (3.6).

LEMMA 8. *As $\lambda_n \to \infty$, $n \to \infty$, $B \to \infty$ and under the conditions of Theorems 1 and 2, if R, R_1 and R_2 are acyclic or p-cycle, then additionally $\lambda_n^p \geq O(n)$,*

$$\frac{\hat{\sigma}_{Bi}^2(R)}{\sigma^2(R;\hat{\rho})} \xrightarrow{P} 1 \qquad \text{for } i = 1, 2, \tag{3.15}$$

$$\frac{\hat{\sigma}_{Bi}(R_1,R_2)}{\sigma(R_1,R_2;\hat{\rho})} \xrightarrow{P} 1 \qquad \text{for } i = 1, 2. \tag{3.16}$$

PROOF. The proof is given in Appendix B3 in [4]. □

3.4. *Comparison of the bootstrap methods*. The variance of each of the subsampling bootstrap methods, just on the basis of the randomness generated from the bootstrap sampling, is given in Theorems 1 and 2. Also, the worst-case computational complexity of finding count statistics for subgraphs R of size p, for the *uniform subsampling* bootstrap, becomes $O(Bm^p)$, whereas for the *subgraph subsampling* bootstrap scheme, the worst-case complexity is $O(B\prod_{d=1}^{p}(nq_d))$. Now the question of balancing computational complexity and statistical stability become important.

For dense networks, say when $\rho_n = n^{-\varepsilon}$ with $\varepsilon > 0$ small (say between $0 < \varepsilon < 1/2$), we also have $\lambda_n = n^{1-\varepsilon}$:

- For *uniform subsampling* from Theorem 1, we get that $\mathrm{Var}(\rho^{-e}\bar{P}_{B1}(R)) = O(\frac{1}{n^{1+2\varepsilon}})$ with $m = n^{\varepsilon}$ and $B = n^{1+\varepsilon}$. The worst-case computational cost becomes $O(n^{1+(p+1)\varepsilon})$.
- For *subgraph subsampling* from Theorem 2, we get that $\mathrm{Var}(\rho^{-e}\bar{P}_{B2}(R)) = O(\frac{1}{n^{1+\varepsilon}})$ for p-cycle R and $O(\frac{1}{n^{1+2\varepsilon}})$ for acyclic R with $q_d = O(\frac{1}{n^{1-\varepsilon}})$ for $d = 2, \ldots, p$ and $B = n^{2\varepsilon}$. The worst-case computational complexity becomes $O(n^{1+(p+1)\varepsilon})$.

Thus in *dense* networks, both the subsampling bootstrap methods can achieve low enough bootstrap variance for low computational cost. In fact, the gain in computational complexity is quite astonishing as polynomial complexity gets reduced to near-linear complexity. The *uniform subsampling* bootstrap is a better choice for its ease of use and marginally smaller variance for p-cycle R. However, since m has to be greater than p, for large R, the benefit of using the *uniform subsampling* bootstrap starts to reduce, and in these cases, the *subgraph subsampling* bootstrap might be a better choice.

For the sparse case, say when $\rho_n = n^{\varepsilon-1}$ with $\varepsilon > 0$ small (say between $0 < \varepsilon < 1/2$), we also have $\lambda_n = n^{-\varepsilon}$:

- For *uniform subsampling* from Theorem 1, we get that $\mathrm{Var}(\rho^{-e}\bar{P}_{B1}(R)) = O(\frac{1}{n^2})$ for acyclic R and $O(\frac{1}{n^{1+\varepsilon}})$ for p-cycle R with $m = n^{1-\varepsilon}$ and $B = n^{1+\varepsilon}$. The worst-case computational cost becomes $O(n^{p-((p-1)\varepsilon-1)})$.
- For *subgraph subsampling* from Theorem 2, we get that $\mathrm{Var}(\rho^{-e}\bar{P}_{B2}(R)) = O(\frac{1}{n^{2-\varepsilon}})$ for acyclic R and $O(\frac{1}{n^{1+\varepsilon}})$ for p-cycle R with $q_d = O(\frac{1}{n^{\varepsilon}})$ for $d = 1, \ldots, p$ and $B = n$. The worst-case computational complexity becomes $O(n^{p-(p\varepsilon-1)})$.

Thus in *sparse* networks, the computational advantage of using the subsampling bootstrap starts to reduce, especially for small subgraphs R. However, for large subgraphs R, there is still a computational advantage to using subsampling bootstrap methods. The *subgraph subsampling* bootstrap scheme is a better choice in this case as it has smaller variance for similar computational complexity.

But for *sparse* graphs, the methods still remain polynomial in worst-case complexity, and for large p and n, the methods become numerically infeasible. In those cases, it becomes more of a *detection* problem than a *counting* problem, and a fundamentally different approach will be required for feasible inference.

4. Theoretical results. In this section, we shall try to provide asymptotic distribution for normalized bootstrap estimates of count statistics. We define normalized bootstrap estimates of count statistic for subgraph R from (3.2) and (3.6) by

$$\hat{T}_{Bi}(R) = \hat{\rho}_n^{-e} \bar{P}_{Bi}(R), \tag{4.1}$$

where $i = 1, 2$ for the two different bootstrap schemes. By obtaining an estimate of the asymptotic variance of $\rho^{-e}\hat{P}(R)$, we can estimate its asymptotic distribution and thus construct hypothesis tests based on the asymptotic distribution. We combine the results obtained in Section 3 to prove Theorem 3.

THEOREM 3. *Suppose R is fixed, acyclic or p-cycle with $|V(R)| = p$ and $|E(R)| = e$ and $\int_0^\infty \int_0^\infty w^{2e}(u, v)\, du\, dv < \infty$. Under the conditions defined in Theorems 1 and 2, for $i = 1, 2$, if $\lambda_n (\equiv n\rho_n) \to \infty$ and $B \to \infty$,*

$$\sqrt{n}\big(\hat{T}_{Bi}(R) - \tilde{P}(R)\big) \xrightarrow{P} 0, \tag{4.2}$$

$$\sqrt{n}\left(\frac{\hat{T}_{Bi}(R) - \tilde{P}(R)}{\hat{\sigma}_{Bi}(R)}\right) \xrightarrow{w} N(0, 1). \tag{4.3}$$

If for fixed, acyclic or p-cycle subgraphs $(R_1, \ldots, R_k)$, we define, $\mathbf{T}_{Bi}(\mathbf{R}) = (\hat{T}_{Bi}(R_1), \ldots, \hat{T}_{Bi}(R_k))$ and $\mathbf{P}(\mathbf{R}) = (\tilde{P}(R_1), \ldots, \tilde{P}(R_k))$

$$\sqrt{n}\big((\mathbf{T}_{Bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R}))^T \hat{\Sigma}_{Bi}^{-1/2}(\mathbf{R})(\mathbf{T}_{Bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R}))\big) \xrightarrow{w} N(\mathbf{0}, \mathbf{I}), \tag{4.4}$$

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

4.1. *Proof of Theorem 3.* The proof follows from the lemma and theorems of the previous section. Since we have $\sqrt{n}$-consistent bootstrap estimators of $\rho^{-e}\bar{P}_{Bi}(R)$ for $i = 1, 2$. Now, from the Theorem 1(a) in [6], we know that as $\lambda_n \to \infty$ if $\hat{\rho}_n = \frac{\bar{D}}{n-1}$, as defined in (2.6),

$$\frac{\hat{\rho}_n}{\rho_n} \xrightarrow{P} 1,$$

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

Now, we define the bootstrap estimates in (4.1). Thus we get by applying Slutsky's Theorem that

$$\sqrt{n}\big(\hat{T}_{Bi}(R) - \tilde{P}(R)\big) \xrightarrow{P} 0 \qquad \text{for } i = 1, 2.$$

The statement about bootstrap estimate of variance follows from Lemma 8 and the definitions of bootstrap variance in the form of equation (3.13).

Thus we have $\sqrt{n}$-consistent bootstrap estimators, $\hat{T}_{Bi}(R)$ (for $i = 1, 2$) of $\hat{T}(R)$ and consistent estimators, $\hat{\sigma}^2_{Bi}(R)$ (for $i = 1, 2$) of $\sigma^2(R; \rho)$. Also from Theorem 1 of [6], we have, for subgraphs $R_1, \ldots, R_k$ of G_n,

$$\sqrt{n}\big(\big(\hat{T}(R_1), \ldots, \hat{T}(R_k)\big) - \big(\tilde{P}(R_1), \ldots, \tilde{P}(R_k)\big)\big) \xrightarrow{w} N\big(\mathbf{0}, \Sigma(\mathbf{R})\big).$$

Thus we can combine the results from Theorems 1 and 2 with the above theorem, using Slutsky and convergence type theorems, to get the symptomatic normality behavior of $\hat{T}_{Bi}(R)$. As $n \to \infty$, $\lambda_n \to \infty$, and under the conditions of Theorems 1 and 2, if we define $\mathbf{T}_{Bi}(\mathbf{R}) = (\hat{T}_{Bi}(R_1), \ldots, \hat{T}_{Bi}(R_k))$ and $\mathbf{P}(\mathbf{R}) = (\tilde{P}(R_1), \ldots, \tilde{P}(R_k))$

$$\sqrt{n}\big(\big(\mathbf{T}_{Bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R})\big)\hat{\Sigma}^{-1/2}_{Bi}(\mathbf{R})\big(\mathbf{T}_{Bi}(\mathbf{R}) - \mathbf{P}(\mathbf{R})\big)\big) \xrightarrow{w} N(\mathbf{0}, \mathbf{I}) \qquad \text{for } i = 1, 2$$

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

5. Simulation results. We apply the two representative bootstrap subsampling schemes for simulated datasets to determine their performance. We generate data from two different simulation models. Both models are special cases of the nonparametric model described in [5]. The two models that we consider are the following:

- the stochastic block model and
- the preferential attachment model.

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

5.1. *Count statistics.* In these simulations, the main class of acyclic features we consider are (k, l)-wheels. We also consider the count of the cyclic patterns such as triads or triangles or 3-cycles and tetrads or *quadrilaterals* or 4-cycles. We also consider a smooth function of counts of triangle and $(1, 2)$-wheel, known as *transitivity*, $\hat{P}_{\mathrm{Tr}}$, defined in Example 3.

5.2. *Stochastic block model.* Let w correspond to a K-block model defined by parameters $\theta = (\boldsymbol{\pi}, \rho_n, S)$, where π_a is the probability of a node being assigned to block a as before, and

$$\mathbf{F}_{ab} = \mathbb{P}(A_{ij} = 1 | i \in a, j \in b) = s_n S_{ab}, \qquad 1 \le a, b \le K,$$

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

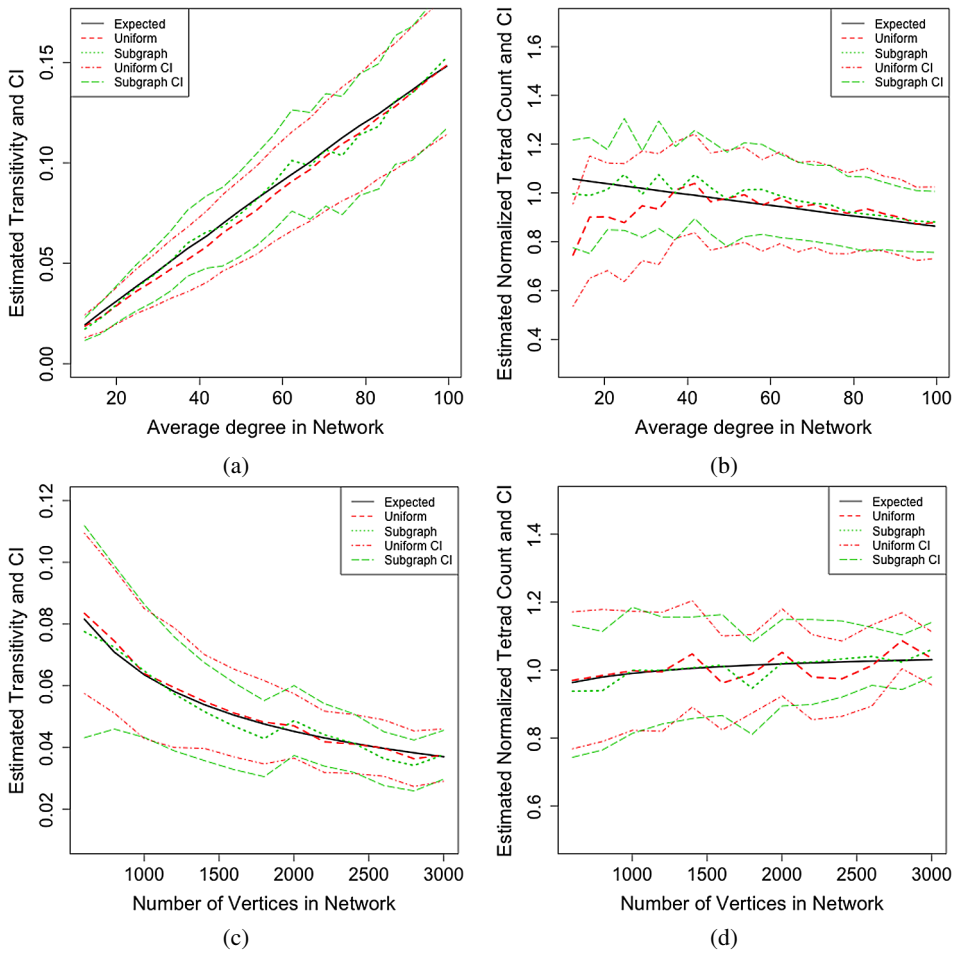

FIG. 1. *Stochastic block model*: *For* $n = 1000$, *we vary average degree* (λ_n) *and* (a) *plot estimated normalized tetrad count and* (b) *plot estimated transitivity and their* 95% *confidence interval* (*CI*), *where CI is estimated using bootstrap estimates of variance of the estimators*. *For* $\nu = 0.5$, *we vary* n, *and* (c) *plot estimated normalized tetrad count and* (d) *plot estimated transitivity and their* 95% *confidence interval* (*CI*). *We use different colors to indicate different bootstrap subsampling schemes and graph parameters*.

We consider a stochastic block model with $K = 2$, $S = \begin{pmatrix} 0.4 & 0.5 \\ 0.4 & 0.7 \end{pmatrix}$, $s_n = \frac{5\nu\sqrt{n}}{n}$ and $\boldsymbol{\pi} = (0.5, 0.5)$. Thus we get $\rho_n = \boldsymbol{\pi}^T \mathbf{F} \boldsymbol{\pi}$. First, we keep $n = 1000$ fixed and vary ν such that ρ_n varies from 10 to 100. Second, we vary ν fixed at 0.5 and vary $n = 500$ to 3000.

In the following figures, we try to see the behavior of mean and variances of the count statistics. In Figure 1(a)–(d), we compare the asymptotic 95% confidence interval of $\tilde{P}(R)$, where R is a 4-cycle or tetrad and $\mathbb{E}(\hat{T}_{\mathrm{Tr}})$, using bootstrap mean and variance estimates, as considered in Theorem 3. The bootstrap estimate of

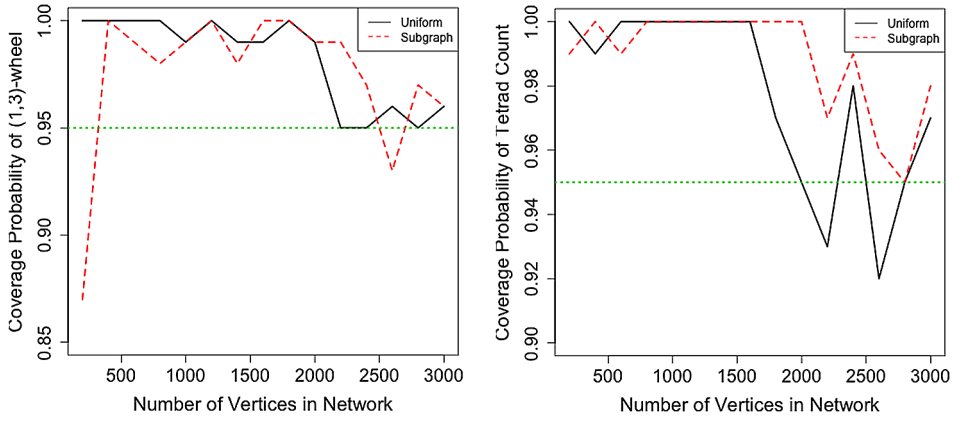

FIG. 2. *Stochastic block model*: *For* $\nu = 0.5$, *we vary* n *and* (a) *plot estimated coverage probability of* 95% *CI for* $(1, 3)$*-wheel count and* (b) *plot estimated coverage probability of* 95% *CI for normalized tetrad count. We use different colors to indicate different bootstrap subsampling schemes*.

asymptotic variance of $\hat{T}_{\mathrm{Tr}}^{Bi}$ is obtained from the bootstrap estimates of $\hat{\sigma}_{Bi}^2(R_1)$, $\hat{\sigma}_{Bi}^2(R_2)$ and $\hat{\sigma}_{Bi}(R_1, R_2)$ by using Delta method and using the Theorem 3.

We also try to see the estimated coverage probabilities of bootstrap estimated confidence intervals for $\tilde{P}(R)$. In Figure 2(a)–(b), we plot estimated coverage probabilities of asymptotic 95% confidence interval for $\tilde{P}(R)$, where R is a $(1, 3)$-wheel and a 4-cycle. We keep ν fixed and vary n from 200 to 3000. Estimated coverage probabilities start becoming close to 0.95 at around $n = 2000$.

In Figure 3, we compare the mean of the bootstrap estimates with the parameter $\tilde{P}(R)$. In Figure 3(a), we keep n fixed but vary λ_n from λ_n 10 to 100 by varying ν, and in Figure 3(b), we keep ν fixed and vary n from 500 to 3000. Thus we get reasonable estimates of integral parameters of graph as we vary the average degree and number of vertices of the graph.

In Figure 4, we compare the variance of the bootstrap estimates, based on bootstrap iterations for both the bootstrap schemes. We see that bootstrap variance is usually lower for the *subgraph subsampling* scheme as we increase the number of vertices of the graph for different count statistics.

5.3. *Preferential attachment model*. In the preferential attachment model, given k initial vertices, $k + 1$th vertex attach to one of the preceding k vertices with probability proportional to degree. Now we have degree of vertex v, defined as D_v and $\bar{D} = \frac{1}{n}\sum_{v=1} D_v$. Also, we have

$$\tau(v) \simeq \frac{D_v}{\bar{D}}.$$

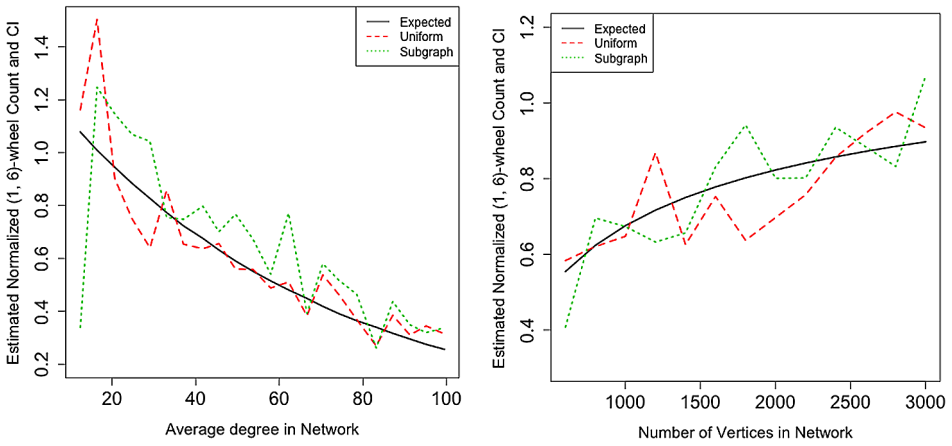

FIG. 3. *Stochastic block model*: *For* $n = 1000$, *we vary average degree* (λ_n) *and* (a) *plot estimated normalized* $(1, 6)$*-wheel count. For* $\nu = 0.5$, *we vary* n *and* (b) *plot estimated normalized* $(1, 6)$*-wheel count. We use different colors to indicate different bootstrap subsampling schemes and graph parameters*.

Thus following equation (2.1), we have the probability of edge formation as

$$w(u, v) = \frac{\tau(u)}{T(u)}\mathbf{1}(u \le v) + \frac{\tau(v)}{T'(u)}\mathbf{1}(v \le u),$$

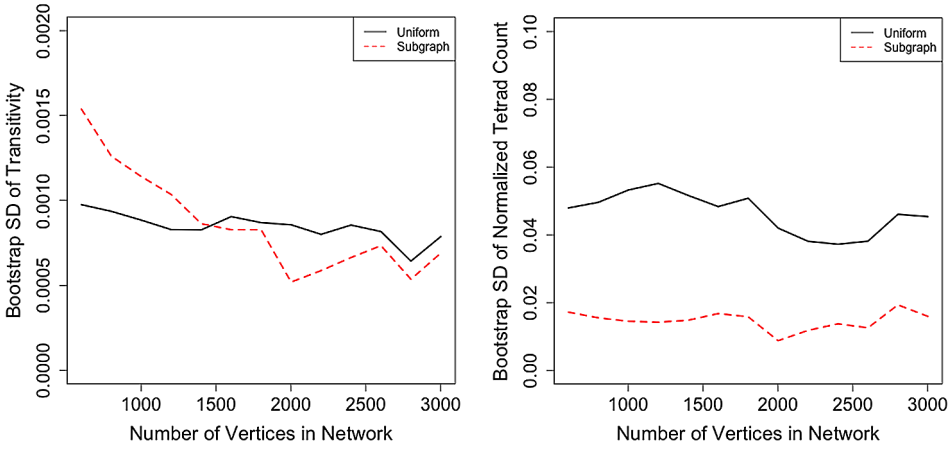

FIG. 4. *Stochastic block model*: *For* $\nu = 0.5$, *we vary the number of vertices* (n) *and plot* (a) *bootstrap variance of estimated transitivity and* (b) *bootstrap variance of normalized tetrad count. We use different colors to indicate different bootstrap subsampling schemes*.

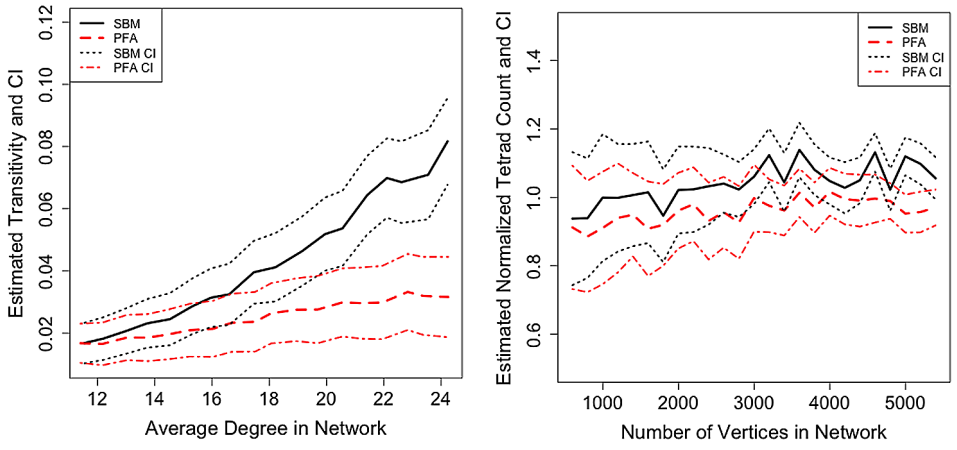

FIG. 5. (a) *For $n = 1000$ we vary λ_n, and we plot estimated transitivity $\hat{T}_{\mathrm{Tr}}^{B2}$ and their* 95% *confidence interval* (*CI*), *where CI is estimated using bootstrap estimates of variance of the estimators*. (b) *We vary n, and we plot estimated normalized tetrad count, $\hat{T}_{B2}(R)$, $R =$ tetrad and their* 95% *confidence interval* (*CI*).

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

$$\tau(u) = \int_0^1 w(u, v)\,dv.$$

Now the preferential attachment model can be defined by the following formula on w:

$$w(u, v) = \frac{\tau(u)}{\int_u^1 \tau(s)\,ds}\mathbf{1}(u \le v) + \frac{\tau(v)}{\int_v^1 \tau(s)\,ds}\mathbf{1}(v \le u).$$

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* $F \equiv \tau^{-1}$.

We simulate networks from both stochastic block models and preferential attachment models, and then we try to compare the distribution of count statistics of the graph for two different networks. In Figure 5(a) we vary the parameters of SBM as $\mathbf{F} = \mu F^{(1)} + (1 - \mu)F^{(2)}$, where $F^{(1)} = \mathrm{Diag}(0.035, 0.065)$ and $F^{(2)} = 0.001\mathbf{1}_2$. We increase μ to increase λ_n and SBM have more pronounced *cluster* structure. We keep the average degree, λ_n, of the two simulated networks the same, and then we try to get the asymptotic distribution of the transitivity

statistic, $\hat{T}_{\mathrm{Tr}}$, for the two cases for each λ_n. We see here that for low λ_n, we cannot statistically distinguish between the transitivity of networks generated from two different models, but they become statistically distinguishable as average degree, λ_n and μ, increase. In Figure 5(b), take SBM as in Section 5.2 and PFA as in Section 5.3 keeping the average degree, λ_n, of the two simulated networks the same, and vary n, and we can statistically distinguish the normalized tetrad count of networks between the two different models for large n based on *subgraph subsampling* scheme.

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, is a small social network, whereas the second one, Facebook college social network, has a greater number of nodes and links. For both cases we use a *subgraph subsampling* bootstrap scheme.

6.1. *High school network.* In this application, we try to quantitatively verify some of the hypotheses mentioned by the authors of [3] 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 [3] 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 of whether the absence of short cycles in this graph is significant or not. We consider a very simple model for the data.

We consider that 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

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{12} & P_{22} \end{pmatrix},$$

where $P_{ab} =$ the 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 $P_{11} = 0$, $P_{12} = 0.0058$ and $P_{22} = 0.000025$. The probability of belonging to the two blocks is $(0.497, 0.503)$.

(b) Preferential attachment model with $\rho = \frac{\lambda_n}{n}$, where $\lambda_n =$ the 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. For both models, we can find $\tilde{P}(R)$, where $R =$ the cycles of size 3 and 4 based on the parameters defined for models in (a) and (b) and using equation (2.4), and we

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

Subgraph	Normalized count	Standard deviation	Count (SBM)	Count (PFA)
(1, 2)-wheel	2.27	0.17	1.01	2.97
3-cycle	1.31	0.1	0.01	1.04
4-cycle	9.47	3.16	0.63	3.06

shall call it $\tilde{P}_0(R)$. Also, for the network, the unknown integral parameter for the subgraph R is $\tilde{P}(R)$. Formally, the hypothesis becomes

$$H_0 : \tilde{P}(R) = \tilde{P}_0(R) \quad \text{vs} \quad \tilde{P}(R) < \tilde{P}_0(R)$$

for each R and for each model (a) and (b). We use the results of Theorem 3 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. Thus 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 1 from [6] would have sufficed, but we give the example as an example of the use of count statistics and their quantitative behavior. In [3], simulation-based tests were used.

Comparison of count statistics in the social network literature has been based on parametric simulation [19] or data bank related tests [32]. In these tests, the networks are generated from either a random graph model or from a data bank of networks (as in [10]). Permutation of nodes' block identity-based tests are used for fitting block models [32]. We use asymptotically Gaussian tests based on nonparametric exchangeable models for comparing graphs. The hypothetical model we consider is nonparametric and thus more general than simulation-based tests on specific random graph models or data bank-based tests. Permutation of nodes' block identity-based tests seem to function more as measures of goodness of fit of the block model assignment of the particular graph.

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. [31]. 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. 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 having nodes coming exclusively from two different class years and the other having nodes coming exclusively from two different majors. We now try to find which of the two networks is more "clustered" by comparing the *transitivity* of the two networks. We can repeat the same exercise for any two covariates and choose a subset of their levels. For the two networks, the unknown integral parameter for transitivity is $\tilde{P}_{\mathrm{Tr}}^1$ and $\tilde{P}_{\mathrm{Tr}}^2$, respectively. Formally, the hypothesis becomes

$$H_0 : \tilde{P}_{\mathrm{Tr}}^1 = \tilde{P}_{\mathrm{Tr}}^2 \quad \text{vs} \quad \tilde{P}_{\mathrm{Tr}}^1 \neq \tilde{P}_{\mathrm{Tr}}^2.$$

The second question can also be answered in a spirit similar to the first. We consider the full collegiate network of two different colleges (Caltech and Princeton). Then, we try to compare the transitivity of these two collegiate networks. For the two networks, the unknown integral parameter for transitivity is $\tilde{P}_{\mathrm{Tr}}^1$ and $\tilde{P}_{\mathrm{Tr}}^2$, respectively. Formally, the hypothesis becomes

$$H_0 : \tilde{P}_{\mathrm{Tr}}^1 = \tilde{P}_{\mathrm{Tr}}^2 \quad \text{vs} \quad \tilde{P}_{\mathrm{Tr}}^1 \neq \tilde{P}_{\mathrm{Tr}}^2.$$

These comparisons could, in principle, be possible using the results given in Bickel et al. [6], but they are computationally intractable. Using bootstrap estimators, we can estimate the variance of the estimators and thus perform hypothesis testing in reasonable time.

In Tables 2, 3 and 4, we present an excerpt of the result of our analysis and answer both of the questions. These results give a better understanding about the network statistics reported in [31], like those reported in Table 3.1 of [31]. Using the numerical comparison of the transitivity values reported in the table of [31] alone can be statistically unreliable, without a proper testing of whether the difference in values for different networks is statistically significant. Such comparison statements are now possible to make with the methods proposed in this paper.

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

	Class year (CY)	**Dormitory (DM)**	**Major (MJ)**
Estimated transitivity	0.15	0.22	0.12

Now, without finding the bootstrap estimate of count statistics and its variance, finding the asymptotic distribution of these count statistics will not be possible. Thus with the help of the bootstrap-based estimates, we can perform hypothesis testing on the count statistics and provide the estimates of their asymptotic distribution.

7. Conclusion and future works. In this paper, we have considered two known subsampling schemes of networks and have tried to show situations where they are applicable to finding the asymptotic distribution of certain *count statistics* of the network. We have showed that the normalized bootstrap subsample estimates of the count statistics and their smooth functions have asymptotic normal distribution. We have proposed bootstrap schemes by which we can efficiently compute the asymptotic mean and variance of these count statistics. We have also showed that the *subgraph sampling* bootstrap scheme seems most stable, and we recommend using this scheme as bootstrap subsampling scheme in most cases.

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) distinguishing between the count statistics of two different networks;
(b) distinguishing between parts of same network;
(b) testing whether a network has been generated from a specific model by comparing the empirical and population versions of the count statistic;
(c) testing how close parameters of two different network models can become.

TABLE 3
The difference between class year and dorm is not significant, but the difference between dorm and major is significant by an asymptotic normal test at 5% *level. The data was presented in Traud et al.* [31]

Difference	**CY and DM**	**DM and MJ**
Estimated	0.07	0.1
Estimated SD	0.05	0.035

TABLE 4
The difference of transitivity between two networks is not significant by an asymptotic normal test at 5% level. Therefore Network 1 cannot be said to be more "clusterable." The data was presented in Traud et al. [31]

	Network 1	**Network 2**
Estimated transitivity	0.29	0.16
Estimated difference	0.13	
Estimated difference SD	0.11	

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

7.1. *Future works*. One natural generalization could be the use of a bootstrap scheme to get asymptotic distribution of global statistics, such as graph cut, conductance, functionals of graphon (nonintegral 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 our bootstrap subsampling scheme to estimate such global characteristics of the networks.

Acknowledgments. We thank Aiyou Chen, Dave Choi and Liza Levina for helpful discussions and comments.

SUPPLEMENTARY MATERIAL

Supplement to "Subsampling bootstrap of count features of networks" (DOI: 10.1214/15-AOS1338SUPP; .pdf). In the Supplement, we prove Theorems 1, 2, Proposition 6, Lemmas 7 and 8.

REFERENCES

[1] ALDOUS, D. J. (1981). Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.* **11** 581–598. MR0637937

[2] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634

[3] BEARMAN, P. S., MOODY, J. and STOVEL, K. (2004). Chains of affection: The structure of adolescent romantic and sexual Networks1. *American Journal of Sociology* **110** 44–91.
[4] BHATTACHARYYA, S. and BICKEL, P. J. (2015). Supplement to "Subsampling bootstrap of count features of networks." DOI:10.1214/15-AOS1338SUPP.
[5] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068–21073.
[6] 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
[7] BICKEL, P. J., GÖTZE, F. and VAN ZWET, W. R. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statist. Sinica* **7** 1–31. MR1441142
[8] BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31** 3–122. MR2337396
[9] CHUNG, F. and LU, L. (2002). Connected components in random graphs with given expected degree sequences. *Ann. Comb.* **6** 125–145. MR1955514
[10] DAVIS, J. A. and LEINHARDT, S. (1972). The structure of positive interpersonal relations in small groups. In *Sociological Theories in Progress*, *Vol.* 2 (J. Berger, M. Zelditch and B. Anderson, eds.) 218–251. Houghton-Mifflin, New York.
[11] DIACONIS, P. and JANSON, S. (2008). Graph limits and exchangeable random graphs. *Rend. Mat. Appl.* (7) **28** 33–61. MR2463439
[12] EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26. MR0515681
[13] FRANK, O. (2005). Models and methods in social network analysis. In *Network sampling and model fitting* (P. J. Carrington, J. Scott and S. S. Wasserman, eds.) 31–56. Cambridge Univ. Press, Cambridge.
[14] HANDCOCK, M. S. and GILE, K. J. (2010). Modeling social networks from sampled data. *Ann. Appl. Stat.* **4** 5–25. MR2758082
[15] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. MR1951262
[16] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5** 109–137. MR0718088
[17] HOOVER, D. N. (1979). Relations on probability spaces and arrays of random variables. Institute for Advanced Study, Princeton, NJ.
[18] KALLENBERG, O. (2005). *Probabilistic Symmetries and Invariance Principles*. Springer, New York. MR2161313
[19] KOLACZYK, E. D. (2009). *Statistical Analysis of Network Data*: *Methods and Models*. Springer, New York. MR2724362
[20] LESKOVEC, J., KLEINBERG, J. and FALOUTSOS, C. (2005). Graphs over time: Densification laws, shrinking diameters and possible explanations. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* 177–187. ACM, New York.
[21] LOVÁSZ, L. (2012). *Large Networks and Graph Limits*. *American Mathematical Society Colloquium Publications* **60**. Amer. Math. Soc., Providence, RI. MR3012035
[22] MIDDENDORF, M., ZIV, E. and WIGGINS, C. H. (2005). Inferring network mechanisms: The *Drosophila melanogaster* protein interaction network. *Proc. Natl. Acad. Sci. USA* **102** 3192–3197.
[23] MILO, R., ITZKOVITZ, S., KASHTAN, N., LEVITT, R., SHEN-ORR, S., AYZENSHTAT, I., SHEFFER, M. and ALON, U. (2004). Superfamilies of evolved and designed networks. *Science* **303** 1538–1542.
[24] 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.

[25] NOWICKI, K. and SNIJDERS, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *J. Amer. Statist. Assoc.* **96** 1077–1087. MR1947255

[26] PICARD, F., DAUDIN, J.-J., KOSKAS, M., SCHBATH, S. and ROBIN, S. (2008). Assessing the exceptionality of network motifs. *J. Comput. Biol.* **15** 1–20. MR2383618

[27] PRZYTYCKA, T. M. (2006). An important connection between network motifs and parsimony models. In *Research in Computational Molecular Biology* 321–335. Springer, Berlin.

[28] SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia* coli. *Nat. Genet.* **31** 64–68.

[29] THOMPSON, S. K. (2012). *Sampling*, 3rd ed. Wiley, Hoboken, NJ. MR2894042

[30] THOMPSON, S. K. and FRANK, O. (2000). Model-based estimation with link-tracing sampling designs. *Survey Methodology* **26** 87–98.

[31] TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Rev.* **53** 526–543. MR2834086

[32] WASSERMAN, S. and FAUST, K. (1994). *Social Network Analysis*: *Methods and Applications* **8**. Cambridge University Press, Cambridge.

[33] WERNICKE, S. (2006). Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **3** 347–359.

DEPARTMENT OF STATISTICS
OREGON STATE UNIVERSITY
44 KIDDER HALL
CORVALLIS, OREGON 97331
USA
E-MAIL: bhattash@science.oregonstate.edu

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: bickel@stat.berkeley.edu