Fast Patchwork Bootstrap for Quantifying Estimation Uncertainties in Sparse Random Networks

Yulia R. Gel
University of Texas at Dallas, USA
ygl@utdallas.edu

Vyacheslav Lyubchich
University of Maryland Center for Environmental Science, USA
lyubchic@umces.edu

L. Leticia Ramirez Ramirez
Centro de Investigacion en Matematicas, Mexico
leticia.ramirez@cimat.mx

ABSTRACT

We propose a new method of nonparametric bootstrap to quantify estimation uncertainties in large and possibly sparse random networks. The method is tailored for inference on functions of network degree distribution, under the assumption that both network degree distribution and network order are unknown. The key idea is based on adaptation of the "blocking" argument, developed for bootstrapping of time series and re-tiling of spatial data, to random networks. We sample network blocks (patches) and bootstrap the data within these patches. To select an optimal patch size, we develop a new computationally efficient and data-driven cross-validation algorithm. The proposed fast patchwork bootstrap (FPB) methodology further extends the ideas developed by [33] for a case of network mean degree, to inference on a degree distribution. In addition, the FPB is substantially less computationally expensive, requires less information on a graph, and is free from nuisance parameters. In our simulation study, we show that the new bootstrap method outperforms competing approaches by providing sharper and better calibrated confidence intervals for functions of a network degree distribution than other available approaches. We illustrate the FPB in application to a study of the Erdös collaboration network.

CCS Concepts

•**Mathematics of computing** → **Nonparametric statistics;** *Random graphs; Stochastic processes;*

Keywords

computational efficiency; large networks; network degree distribution; sampling; resampling

MLG2016'16 August 14, 2016, San Francisco, California, USA

ACM ISBN 123-4567-24-567/08/06…$15.00

DOI: 10.1145/1235

1. INTRODUCTION

As the world continues to embrace the wealth of information provided by modern social media, from devising viral marketing strategies, to predicting fashion trends, to analysing public health perception and preventing terrorist attacks, there is a flare of interest in development of new statistical methodology for analysis of large network structures. Indeed, probabilistic models have been dominating the area of network inference, whereas development of statistical inference, particularly for large networks, was noticeably delayed, and statistical network models are yet relatively scant and poorly investigated (see [16, 30, 22, 19] and references therein). Motivated by a plethora of modern large network applications and rapid advances in computing technologies, the area of statistical network modeling is undergoing a vigorous developmental boom, spreading over numerous disciplines, from statistics to engineering to social and health sciences.

Challenges of parametric model specification and validation inspire a search for more data-driven and flexible nonparametric (at least, semiparametric) approaches for network inference. As [14] state, "*statistical modeling of networks cries for nonparametric estimation, because of the inaccuracy often resulting from fallacious parametric assumptions*". In spite of that, the scope and availability of nonparametric procedures for random network inference still remains very limited and scarce (for some recent results and overview see [27, 3, 22, 6, 33] and references therein). In this light, it is appealing and promising to follow a nonparametric bootstrap path for statistical inference on random networks that can potentially allow us to avoid many restrictive conditions on network degree distribution and model specification. To our knowledge, the pioneers in this area are Snijders and Borgatti [32] who proposed to employ an induced graph sampling for estimation of standard errors in network density estimation and comparison of two networks. The procedure is, however, limited to very small networks, assumes availability of the entire network data upfront as well as requires resampling of the entire data set.

Despite all the recent interest in nonparametric network analysis, bootstrap methodology for inference on random networks still remains virtually unexplored in statistical literature. And, whereas some recent results target quantification of estimation accuracy for subgraph patterns [2, 6], issues with reliable evaluation of estimation errors for a degree distribution are largely unaddressed [31]. Recently, [33] propose a nonparametric resampling-based patchwork bootstrap, with a focus on a network mean degree. In this paper,

we further advance the patchwork of [33] and develop a fast and information greedy bootstrap for quantification of estimation uncertainties in functions of degree distribution.

The degree distribution is one of the primary interests in analysis of graph-structured data and there exist numerous methods for obtaining the degree distribution directly from a graph (for overview see, for instance, [26, 1, 22, 31, 35] and references therein). However, what *missing is quantification of estimation uncertainty*, that is, how reliable the obtained estimates of the degree distribution are. Clearly, if the *entire and complete* graph-structured data that constitute the study interest are available upfront, then we can simply obtain the degree sequence (for instance, we might be interested in friendship of prison inmates or sexual contacts on an island that are perfectly recorded). In practice, however, such an approach might be not only computationally inefficient but also infeasible — indeed, we typically observe only partial information, e.g., a subset of Facebook data, and the goal is to make a reliable inference on how people interact online. To our knowledge, our approach is the first attempt to quantify estimation uncertainty in degree distribution using nonparametric bootstrap.

Our idea behind the bootstrap path is intuitive: as the classical bootstrap of [11] was originally suggested for independent and identically distributed data and then adapted to time series and spatial processes [17, 8, 23, 28], we borrow the "blocking" argument developed for resampling of space and time dependent processes and adjust it to networks. In this sense, a random graph can be viewed as a mathematical object representing a hybrid of time and space dependent processes, with a natural metric induced by a shortest path between two vertices. Similar to the "blocking" argument, we select local vicinities, or patches, around randomly selected vertices, and then resample vertices within each patch. Since patches are allowed to overlap, our procedure can be said to follow the "Künsch rule" [23]. In contrast to the classical "blocking" argument in time series, we do not aim to reconstruct the network data generating process (DGP). Although such DGP reconstruction would certainly be desirable, we believe that this ambitious goal cannot be attained with the patchwork bootstrap or any other bootstrap technique on networks without imposing very restrictive (thus, impractical) conditions on the network structure.

In this paper, we apply the new fast patchwork bootstrap (FPB) to estimate network degree distribution and quantify its estimation uncertainty, i.e., develop a confidence interval, under the assumption that both network degree distribution and network order are unknown but the network distribution is involution invariant. The property of involution invariance can be viewed as a network analogue of stationarity of a stochastic process [27]. Stationarity is typically an essential condition for consistency of block bootstrap for space and time dependent data, thus, again linking our bootstrap procedure with the "blocking" argument.

In addition, similarly to the block bootstrap for space and time dependent data [18], we found that the new information-greedy bootstrap procedure is sensitive to the size of the patch. We address this issue by developing a data-driven and computationally efficient optimal patch selection algorithm based on a cross-validation argument.

The main contributions of our study are as follows:

- To our knowledge, this is the first approach to develop bootstrap inference and bootstrap confidence intervals for network degree distribution. In fact, while there exists a vast literature on graph sampling for estimating network properties, very little is known on how to *reliably evaluate associated errors of estimation* (outside of extensive, information costly and typically impractical simple random sampling).
- We introduce a novel nonparametric bootstrap method for evaluating uncertainty in functions of a population network degree distribution, under no prior information on network degree distribution and network order. Note that this is very different than developing a point estimator of a quantity of interest, as our new method enables to assess the error of estimation and construct reliable confidence intervals in a fully data-driven way. Moreover, in contrast to other methods, the network can be sparse and can be only partially observable.
- We develop a new computationally efficient and data-driven cross-validation algorithm for selecting an optimal patch size.
- We validate the new bootstrap procedure by extensive simulations and show that the new method outperforms the competing approaches by providing sharper and better calibrated confidence intervals for functions of a network degree distribution. We illustrate utility of the FPB in applications to the Erdös collaboration networks.
- Our method allows to draw statistical inference about the "true" (population) unobserved network, using only a small portion of observed graph.

The paper is organized as follows. Section 2 provides some preliminary notations on random graphs and presents the new FPB procedure. In Section 3, we discuss a cross-validation algorithm for optimal patch selection. The new bootstrap algorithm is then evaluated by extensive numerical studies in Section 4. In Section 5, we illustrate applications of new fast patchwork bootstrap procedure to analysis of the Erdös collaboration network. The paper is concluded by discussion in Section 6.

2. BACKGROUND AND APPROACH

2.1 Assumptions

Consider an undirected random graph $G = (V, E)$ with a set of vertices, $V(G)$, and a set of edges, $E(G)$. The order and size of G are defined as the number of vertices and edges in G, i.e., $|V(G)|$ and $|E(G)|$, respectively. We assume that G has no self-loops, i.e., $u \neq v$ for any edge $e_{uv} \in E$. The degree of a vertex v is the number of edges incident to v. We denote the probability that a randomly selected node has a degree k by $f(k)$, the degree distribution of G by $F = \{f(k), k \geq 0\}$, and the mean degree of G by $\mu(G)$. We assume that G is involution invariant [24, 27], that is from the vantage point of any randomly selected vertex, the rest of the connected network is probabilistically the same.

Graph G represents some hypothetical "true" random graph of interest that is never fully observed, and its degree distribution F with finite mean and its order are unknown. Instead, we observe a random graph G_n with a degree distribution $F_n = \{f_n(k), k \geq 0\}$. Let $N_k^{(n)}$ be the number of vertices with a degree k in G_n. Observed graph G_n is a

realization of G in a sense that as $n \to \infty$, $N_k^{(n)}/n \to f(k)$ in probability (empirical distribution F_n converges in probability to F) and joint degree distribution of G_n approaches that of G (see [7, 34] and references therein).

2.2 Fast patchwork bootstrap (FPB)

We develop a new nonparametric bootstrap-based inference for an unknown population degree distribution F of G using the observed realization G_n. Let $\eta(G)$ be the statistical parameter of interest based on F (e.g., $\eta(G)$ can be a probability of observing a vertex of degree k, network mean degree, variance or tail indexes), and let $\hat{\eta}(G_n)$ be an empirical estimator of $\eta(G)$ obtained from an observed realization G_n. Our goal is to assess estimation uncertainty of the population parameter $\eta(G)$ using a bootstrap distribution of the sample statistic $\hat{\eta}(G_n)$.

Our patchwork algorithm consists of two main steps: *sampling*, or creation of patches that aim to "mirror" G_n, and *resampling*, or bootstrap, within the patches that aims to quantify estimation uncertainty of the parameter of interest $\eta(G)$. This new method significantly extends and simplifies the approach of [33], particularly, excludes any nuisance parameters from constructing confidence intervals and does not assume independence of patches.

Sampling-resampling procedure is summarized in Algorithm 1. To generate patches we employ a modified version of snowball sampling, namely the Labeled Snowball with Multiple Inclusions (LSMI, Figure 1) of [33]. Unlike snowball sampling, LSMI incorporates new information conditionally on the links that have been already recorded, thus, does not trace the same edge multiple times and hence minimizes bias in degree estimation. LSMI may be viewed as a fusion of classical snowball sampling, induced subgraph sampling and star sampling [21, 13].

We apply a modified bootstrap-based Horvitz–Thompson method to estimate a degree distribution [33]:

$$\hat{f}^*(k) = \frac{|\{v_s^*(k)\}| + (1-\hat{p}_0^*)|\{v_{ns}^*(k)\}|}{|\{v_s^*\}| + |\{v_{ns}^*\}|}, \qquad (1)$$

where $k > 0$, $v_s^*(k)$ and $v_{ns}^*(k)$ are bootstrap seeds and non-seeds with degree k, respectively, $|\cdot|$ denotes cardinality of a set, and $\hat{p}_0^*$ is the proportion of zeros in the set of bootstrapped seeds $\{v_s^*\}$, $\hat{f}^*(0) = \hat{p}_0^*$. The corresponding bootstrap-based mean degree estimator is:

$$\hat{\mu}(G_n)^* = \sum_{k \geq 0} k \hat{f}^*(k). \qquad (2)$$

The intuitive idea behind (1) is that its numerator represents an estimate of the number of all nodes with a degree k, with the first term delivering information from seeds and the second term delivering information from non-seeds. Denominator in (1) is an estimator of a network order that is, similarly, based on seeds and non-seeds.

For each seed-wave combination j, we construct the Efron $100(1-\alpha)\%$ bootstrap confidence interval

$$BCI_j = \left(\hat{\eta}^{j*}_{[B\alpha/2]}, \hat{\eta}^{j*}_{[B(1-\alpha/2)]} \right), \qquad (3)$$

where $j = 1, \ldots, J$, $J = kd$, d is the number of waves, $m_1, \ldots, m_k$ are different sample sizes for the seeds, $\hat{\eta}^{j*}_{[B\alpha/2]}$ and $\hat{\eta}^{j*}_{[B(1-\alpha/2)]}$ are the empirical quantiles from the bootstrap distribution based on B bootstrap replications (see Section 3 on a data-driven choice of the optimal seed-wave combination).

Algorithm 1: Labeled snowball with multiple inclusions (LSMI) sampling and patchwork bootstrap [33].

input : network G_n; number of seeds m ($m \ll n$); number of waves d; number of bootstrap samples B.
output: a sample of m seeds $\{v_s\}$ with up to d waves around each seed $\{v_{ns}\}$, and corresponding bootstrap samples $\{v_s^*\}_b$ and $\{v_{ns}^*\}_b$, $b = 1, \ldots, B$.

```
{v_s} = sample randomly without replacement m seeds;
for i = 1, ..., m do
    start with original network G_n (with all edges);
    included_0 = {v_s}_i;
    for j = 1, ..., d do
        let wave_j be all immediate neighbours of the
        vertices from the set included_{j-1};
        included_j = included_{j-1} ∪ wave_j;
        eliminate all edges that were used to locate
        wave_j;
    end
    {v_ns}_i = {wave_j}_1^d;          /* Multiset Union */
end
for b = 1, ..., B do
    {v_s^*}_b = sample with replacement from {v_s};
    {v_ns^*}_b = sample with replacement from {v_ns} with
    weights proportional to inverse of their degrees.
end
```

What do we gain by combining seeds and non-seeds into a joint estimator? While many estimators of graph totals based solely on seeds are unbiased [12], variance of such seed-based estimator might be high if the number of seeds is low. At the same time, sampling more seeds might be prohibitively expensive (see overview by [20] and references therein). Adding information from non-seeds into the degree estimator increases bias but reduces variance. For example, Figure 2 shows the effect of adding waves of non-seeds into the mean degree estimator (2). Hence, a choice on number of seeds and waves of non-seeds in LSMI leads to a classical bias vs. variance trade-off, and we propose to address it using a cross-validation procedure (Section 3).

3. SELECTING AN OPTIMAL SEED-WAVE COMBINATION

Similar to findings of [17, 8, 23, 18] for block bootstrap for space and time dependent processes, performance of the new FPB procedure strongly depends on the size of patches defined by the number of seeds and the number of waves in a patch. We propose to select an optimal seed-wave combination by a data-driven cross-validation procedure (Algorithm 2). Note that in contrast to the earlier method of [33] which requires multiple LSMI (≈ 25), the new cross-validation Algorithm 2 requires substantially less data and is based on 1 LSMI, which makes it particularly attractive for streaming applications.

4. SIMULATION STUDY

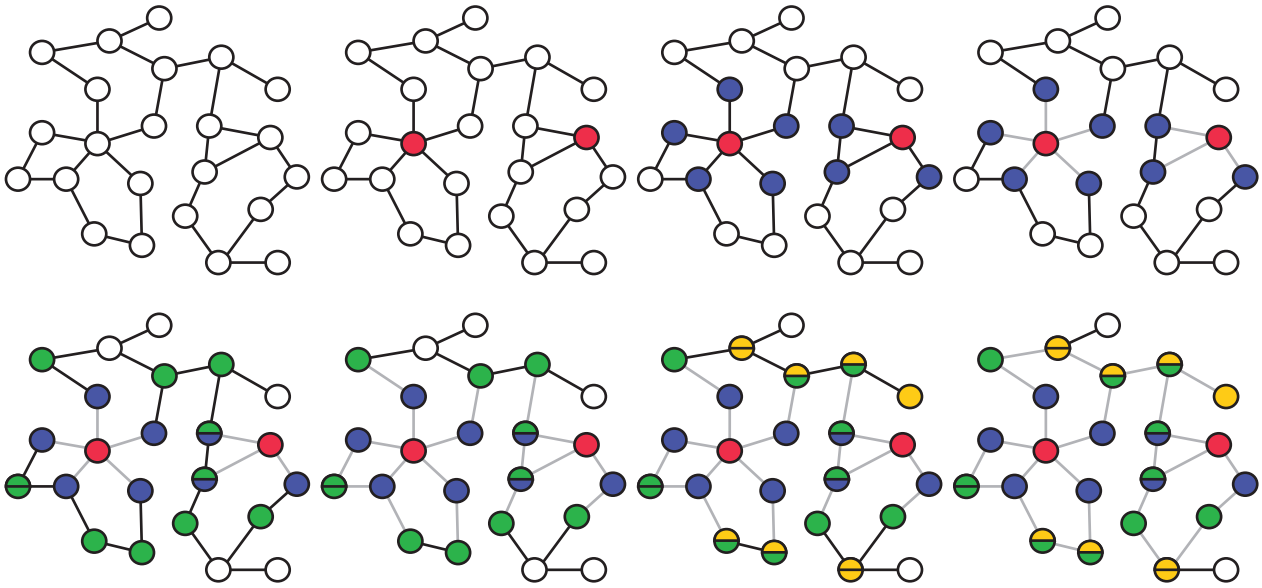

Figure 1: Steps of the LSMI algorithm with $m = 2$ seeds and $d = 3$ waves applied to a network of order $n = 23$.

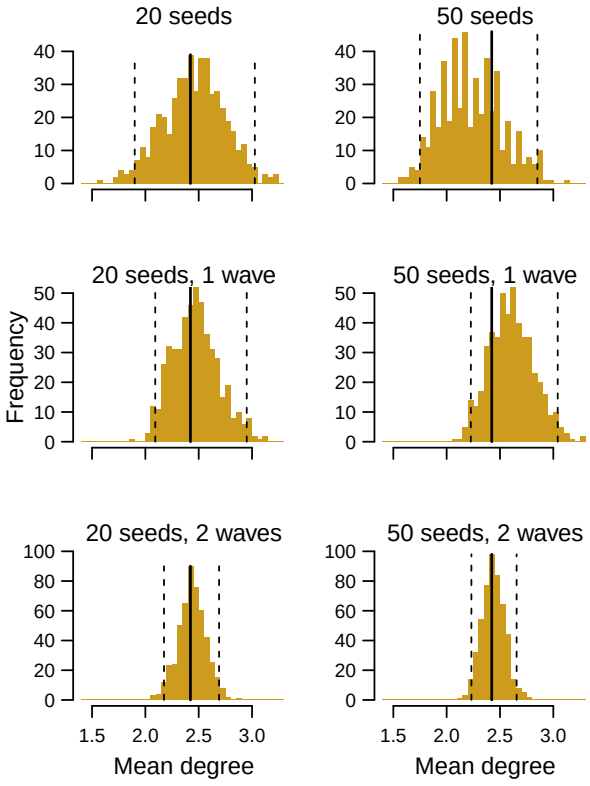

Figure 2: Histograms of bootstrap mean degrees $\hat{\mu}(G_n)^*$ for a simulated network of order 10,000 with polylogarithmic(0.1,2) degree distribution. The 95% confidence intervals (dashed vertical lines) are for $\mu(G) = 2.42$ (solid vertical line).

Algorithm 2: Cross-validation algorithm to select an optimal seed-wave combination.

input : network G_n; IDs of seeds that were used in the patch, U; bootstrap confidence intervals BCI_j for J seed-wave combinations, $j = 1, \ldots, J$; proxy sample size h; N number of times for obtaining proxy; confidence level α.

output: an optimal seed-wave combination j_{opt} selected from j, and corresponding bootstrap interval $BCI_{j_{opt}}$.

for $i = 1, \ldots, N$ **do**
  sample h nodes from U;
  estimate $\hat{\eta}_i^{proxy}$ from the h sampled nodes;
  for $j = 1, \ldots, J$ **do**

$$count_{i,j} = \begin{cases} 1 & \text{if} \quad \hat{\eta}_i^{proxy} \in BCI_j \\ 0 & \text{otherwise} \end{cases}$$

  end
end
$j_{opt} = \arg\min_{j=1,\ldots,J} \left| N^{-1} \sum_{i=1}^{N} count_{i,j} - (1 - \alpha) \right|$;
$BCI_{j_{opt}}$.

In this section, we examine finite sample properties of the new fast patchwork bootstrap and cross-validation procedure, by extensive Monte Carlo experiments.

Validation Metrics We use two standard statistical metrics to validate the proposed bootstrap method: coverage probability and sharpness. Coverage probability for a $100(1-\alpha)\%$-confidence interval is defined by a relative proportion of times when the confidence interval contains the estimated parameter. Coverage probability is a measure of *calibration*. Average width of the developed confidence intervals provides assessment of *sharpness*. Calibrated confidence inter-

vals with shorter widths are preferred. Conservative confidence intervals (over-estimating coverage) are preferred over liberal confidence intervals (under-estimating coverage).

Using the Chung-Lu algorithm [9], we simulate 10,000 networks for three different distributions, namely, zero-truncated Poisson and two different polylogarithmic distributions [26, 33], and for varying network orders (1,000, 3,000, 5,000 and 10,000 vertices). Among the considered degree distributions, polylogarithmic distribution with parameters (2,3) exhibits the lightest tail, whereas the longest tail belongs to polylogarithmic distribution with parameters (0.1,2) (Figure 3). We consider patches with 20, 30, 40 and 50 seeds and 1 to 5 waves around each seed ($J = 20$ different seed-wave combinations in each network realization).

We validate our patchwork bootstrap procedure for quantifying estimation of a population mean degree against two competing procedures. The first competing approach is a 95% parametric confidence interval (CI) based on normal distribution. That is, using simple random sampling (SRS) without replacement, we select M nodes and estimate proportion of nodes with degree k, i.e., $\hat{f}(k)$. Then, normal-based confidence interval (NCI) is given by $NCI^{\{M\}} = \hat{f}(k) \pm 1.96\hat{\sigma}_{\hat{f}(k)}$ where $\hat{\sigma}_{\hat{f}(k)}$ is a sample standard deviation based on the M sampled nodes. The second competing approach is a nonparametric quantile-based bootstrapped confidence interval based on the M nodes from SRS. In particular, we resample with replacement the degrees of M previously selected nodes, calculate the respective proportions of nodes with degree k and repeat the resampling procedure B times. The respective the Efron bootstrap confidence interval is given by

$$QCI^{\{M*\}} = \left(\hat{f}^{\{M*\}}_{[B\alpha/2]}(k),\ \hat{f}^{\{M*\}}_{[B(1-\alpha/2)]}(k) \right), \qquad (4)$$

where $\hat{f}^{\{M*\}}_{[B\alpha/2]}(k)$ and $\hat{f}^{\{M*\}}_{[B(1-\alpha/2)]}(k)$ are the empirical quantiles estimated solely from the M nodes from SRS. (Throughout the paper, nominal significance level α is 0.05.)

We now evaluate performance of the FPB in quantifying estimation uncertainty of the network degree probabilities $f(k)$.

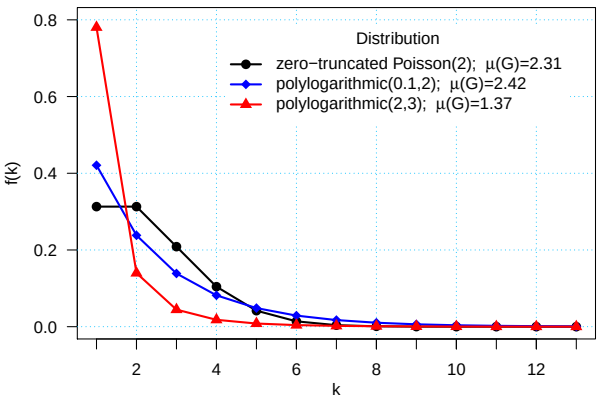

Figure 3: Theoretical degree distributions.

Quantifying estimation uncertainty for probabilities $f(k)$ of observing a node of degree k. We now apply the FPB to quantify uncertainty in estimating theoretical probabilities $f(k)$, $k \in Z^+$.

Table 1 presents the results of the new fast patchwork bootstrap procedure along with the competing NCI and QCI. The FPB provides the most calibrated and sharp confidence intervals (CIs) for all considered degree distributions and network orders.

In particular, for the zero-truncated Poisson distribution and polylogarithmic distribution with parameters $(0.1, 2)$, coverage of the FPB fluctuates around the declared 95% confidence level (coverage is between 92% and 98%), while both NCI and QCI, despite consistently yielding around 40% wider intervals than FPB, noticeably underestimate the nominal coverage probability, especially for $f(4)$.

Moreover, difference in performance among the FPB, NCI and QCI is particularly striking for a sparse network (i.e., polylogarithmic distribution with parameters $(2, 3)$). Here, the FPB delivers well calibrated intervals, closely resembling the declared 95% confidence level; however, despite producing noticeably wider intervals, NCI and QCI severely underestimated coverage, yielding only 60% vs. the declared 95% for $f(4)$. (We also explored applicability of FPB to Poisson distribution and our findings are similar.)

Thus, the FPB can be viewed as a preferred procedure for fast and reliable inference in sparse networks, under limited prior information. Moreover, the FPB method is both computationally efficient and information-greedy (i.e., it minimizes information that is collected from the network). Hence, the FPB approach can be of particular importance in analysis of complex social networks, for example, for quantifying estimation uncertainty and hypothesis testing for number of friends, collaborators, and sexual partners, including hard-to-reach populations.

5. ERDÖS NETWORKS

In this section, we revisit collaboration networks of two groups of mathematical scientists, which are a part of the Erdös collaboration network [26, 4, 25, 10]. We define the networks as in [33]: based on the Erdös number, which is the shortest path length between an author and Erdös, where the latter has the Erdös number of zero. The authors with an Erdös number from 1 to 4 represent "senior" researchers, whereas the authors with an Erdös number from 5 to 7 represent "junior" researchers. [33] show that the group of junior researchers has significantly lower mean degree than their senior colleagues, and the observed $\hat{\mu}(G_n)$ is 2.44 for junior (n = 80,607) vs. 5.53 for senior researchers (n = 94,766). Here we further investigate collaboration patterns in the two groups of researchers and, in particular, probabilities of collaborating with one or more co-authors.

Table 2 reports 95% confidence intervals for $f_k(G)$, $k = 1, \ldots, 5$, along with observed frequencies $\hat{f}_k(G_n)$. The main difference between two subnetworks is in the proportion of nodes with degree 1: it is almost twice higher for the collaborators with the Erdös number 5 to 7. Remarkably, the FPB 95% confidence intervals for $f_1(G)$ do not overlap for the two networks, confirming the significance of the difference, whereas the NCI and QCI intervals are much wider and overlap. Given the results of our simulation study that suggests higher reliability of FPB, we tend to adopt the conclusion of a statistically significant difference among networks, delivered by FPB. This phenomenon is likely due to the fact that the group of researchers with the Erdös number 5 to 7 is dominated by "junior" researchers, and the latter tend to have more collaboration solely with his or her supervisor.

Table 1: Coverage of theoretical probabilities $f_k(G)$ of observing a node of degree k, $k = 2, 4$, by 95% confidence intervals for varying network orders. Average interval width is given in parentheses. Network degree distributions are zero-truncated Poisson(2) (ztP(2), $\mu(G) = 2.31$), polylogarithmic(0.1,2) (pl(0.1,2), $\mu(G) = 2.42$), and polylogarithmic(2,3) (pl(2,3), $\mu(G) = 1.37$). Methods of obtaining confidence intervals are fast patchwork bootstrap (FPB), normal interval based on estimated proportions and their variance using 50 random nodes (NCI$^{\{50\}}$), bootstrap of 50 random nodes (QCI$^{\{50*\}}$). Number of bootstrap resamples is 500. Number of Monte Carlo simulations is 1,000.

Distribution	k	Method	Network order n 2,000	3,000	5,000	10,000
ztP(2)	2	FPB	92.4 (0.15)	93.3 (0.15)	93.7 (0.16)	94.7 (0.15)
		NCI$^{\{50\}}$	93.0 (0.25)	92.6 (0.25)	92.0 (0.26)	93.2 (0.26)
		QCI$^{\{50*\}}$	93.5 (0.25)	94.5 (0.25)	92.9 (0.25)	94.4 (0.25)
	4	FPB	96.4 (0.10)	97.3 (0.10)	97.9 (0.10)	97.7 (0.10)
		NCI$^{\{50\}}$	89.5 (0.17)	88.7 (0.17)	89.8 (0.17)	89.8 (0.17)
		QCI$^{\{50*\}}$	90.0 (0.16)	89.6 (0.16)	90.0 (0.16)	89.1 (0.16)
pl(0.1,2)	2	FPB	92.2 (0.13)	92.5 (0.13)	92.3 (0.14)	94.0 (0.13)
		NCI$^{\{50\}}$	91.5 (0.23)	90.8 (0.23)	91.6 (0.23)	91.8 (0.24)
		QCI$^{\{50*\}}$	93.8 (0.23)	93.6 (0.23)	94.5 (0.23)	93.5 (0.23)
	4	FPB	93.9 (0.082)	96.5 (0.08)	96.7 (0.09)	98.2 (0.09)
		NCI$^{\{50\}}$	90.0 (0.14)	91.4 (0.15)	90.9 (0.15)	93.4 (0.15)
		QCI$^{\{50*\}}$	89.9 (0.14)	91.4 (0.14)	90.9 (0.14)	93.3 (0.14)
pl(2,3)	2	FPB	96.0 (0.13)	95.1 (0.13)	95.8 (0.13)	96.7 (0.14)
		NCI$^{\{50\}}$	89.9 (0.19)	92.7 (0.19)	92.0 (0.19)	90.7 (0.19)
		QCI$^{\{50*\}}$	90.6 (0.18)	93.3 (0.19)	93.0 (0.18)	92.7 (0.18)
	4	FPB	96.8 (0.05)	95.8 (0.05)	95.6 (0.05)	96.1 (0.05)
		NCI$^{\{50\}}$	59.3 (0.06)	58.7 (0.06)	59.4 (0.06)	60.6 (0.06)
		QCI$^{\{50*\}}$	58.2 (0.05)	57.1 (0.05)	58.2 (0.05)	59.5 (0.05)

Confidence intervals for $f_k(G)$, $k = 2, \ldots, 5$ unveil no further differences between the two degree distributions. Note the FPB intervals in these two networks are about 1.7 times narrower than the NCI and QCI intervals.

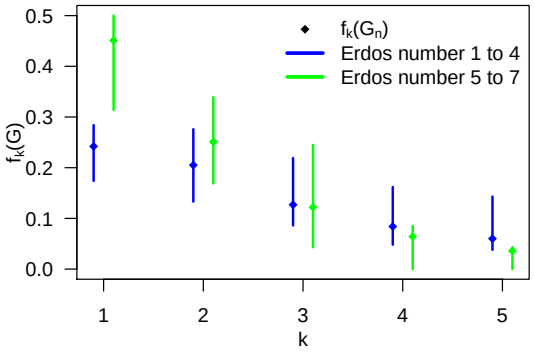

Figure 4: Observed frequencies $f_k(G_n)$ (points) and FPB 95% intervals (lines) for $f_k(G)$, for the two subnetworks of researchers, based on their Erdös number.

6. CONCLUSIONS

In this paper we propose a novel data-driven and computationally efficient method for quantifying uncertainty in network degree distribution using nonparametric bootstrap. We primarily focus on developing confidence intervals for functions of a network degree distribution of some "true" underlying network and perceive the collected network data as a single realization of this "true" unobserved network. The proposed patchwork idea is intrinsically linked to block bootstrap and re-tiling in space-time processes where patches, or analogues of blocks and tiles, are grown around randomly selected seeds, and then both seeds and their neighbors are resampled. Similarly to resampling procedures for weakly dependent space-time processes, finite sample performance of the new FPB depends on number of seeds and waves around them, and we address this challenge by developing a new data-driven cross-validation procedure. We show that the FPB provides well-calibrated and sharp confidence intervals for network mean degree and probabilities of observing a node of a prespecified degree and outperforms its parametric and nonparametric competitors in terms of accuracy, computational costs and required network information. The new bootstrap method can be further extended to quantification of estimation uncertainty in point centrality and centralization measures, network heterogeneity and similarity measures for multiple network comparisons based on a degree distribution. In the future we plan to explore combination of the FPB approach with other type of degree estimators, particularly, the respondent-driven sampling (RDS) framework for hard to reach populations [15]. Another interesting direction is application of bootstrap for goodness-of-fit testing on networks and optimal parameter selection, for instance, in conjunction with parameterization of the shortest-path dis-

Table 2: The 95% confidence intervals for the population probabilities $f_k(G)$ of two Erdös subnetworks. Methods of obtaining confidence intervals are fast patchwork bootstrap (FPB), normal interval based on estimated proportions and their variance using 50 random nodes ($NCI^{\{50\}}$), bootstrap of 50 random nodes ($QCI^{\{50*\}}$). In FPB, 12 seed-wave combinations were considered: waves from 1 to 3, seeds 20, 30, 40, and 50. Cross-validation is based on a random selection of 100 seeds 13 times. Number of bootstrap resamples is 500.

		FPB		$NCI^{\{50\}}$		$QCI^{\{50*\}}$	
k	$\hat{f}_k(G_n)$	lower	upper	lower	upper	lower	upper
Subnetwork with Erdös number 1 to 4							
1	0.242	0.174	0.284	0.282	0.558	0.280	0.540
2	0.205	0.133	0.276	0.043	0.237	0.060	0.240
3	0.127	0.086	0.219	0.043	0.237	0.040	0.240
4	0.084	0.048	0.162	0.000	0.126	0.000	0.120
5	0.060	0.038	0.143	0.000	0.126	0.000	0.140
Subnetwork with Erdös number 5 to 7							
1	0.451	0.314	0.500	0.207	0.473	0.210	0.480
2	0.251	0.169	0.339	0.088	0.312	0.100	0.320
3	0.122	0.043	0.245	0.088	0.312	0.100	0.320
4	0.064	0.000	0.085	0.029	0.211	0.040	0.220
5	0.036	0.000	0.043	0.004	0.156	0.020	0.160

tance distribution of networks using the generalized Gamma distribution [5]. The current version of the code is available from R package *snowboot* [29].

7. REFERENCES

[1] D. Achlioptas, A. Clauset, D. Kempe, and C. Moore. On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In *Proceedings of the 37th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2005.

[2] N. K. Ahmed, N. Duffield, J. Neville, and R. Kompella. Graph sample and hold: A framework for big-graph analytics. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1446–1455. ACM, 2014.

[3] E. M. Airoldi, T. B. Costa, and S. H. Chan. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems*, pages 692–700, 2013.

[4] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590–614, 2002.

[5] C. Bauckhage, K. Kersting, and F. Hadiji. Parameterizing the distance distribution of undirected networks. In *Proceedings of the 31th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 121–130. AUAI, 2015.

[6] S. Bhattacharyya and P. J. Bickel. Subsampling bootstrap of count features of networks. *Annals of Statistics*, 43(6):2384–2411, 2015.

[7] T. Britton, M. Deijfen, and A. Martin-Löf. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 6(124):1377–1397, 2006.

[8] E. Carlstein. The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *The Annals of Statistics*, 14(3):1171–1179, 09 1986.

[9] F. Chung and L. Lu. Connected components in random graphs with given degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

[10] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

[11] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 01 1979.

[12] O. Frank. Estimation of graph totals. *Scandinavian Journal of Statistics*, 4(2):81–89, 1977.

[13] O. Frank. Survey sampling in networks. *The SAGE handbook of social network analysis. Sage, London*, pages 389–403, 2011.

[14] A. Freno, M. Keller, G. C. Garriga, and M. Tommasi. Spectral estimation of conditional random graph models for large-scale network data. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 121–130. AUAI, 2012.

[15] K. J. Gile. Improved inference for respondent-driven sampling data with application to hiv prevalence estimation. *Journal of the American Statistical Association*, 106(493):135–146, 2011.

[16] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2(2):129–233, 2010.

[17] P. Hall. Resampling a coverage pattern. *Stochastic Processes and their Applications*, 20(2):231–246, 1985.

[18] P. Hall, J. L. Horowitz, and B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82(3):561–574, 1995.

[19] T. Hellmann and M. Staudigl. Evolution of social networks. *European Journal of Operational Research*, 234(3):583–596, 2014.

[20] J. Illenberger and G. Flötteröd. Estimating network properties from snowball sampled data. *Social Networks*, 34(4):701–711, 2012.

[21] E. D. Kolaczyk. *Statistical analysis of network data: methods and models*. Springer, New York, 2009.

[22] E. D. Kolaczyk and G. Csárdi. *Statistical analysis of network data with R*, volume 65. Springer, 2014.

[23] H. R. Künsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 09 1989.

[24] L. Lovász. *Large networks and graph limits*, volume 60. Colloquium Publications. American Mathematical Society, 2012.

[25] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2):026113, 2004.

[26] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*,

64(2):026118, 2001.

[27] P. Orbanz and D. M. Roy. Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2):437–461, 2015.

[28] D. Politis and J. P. Romano. A circular block-resampling procedure for stationary data. In R. LePage and L. Billard, editors, *Exploring the Limits of Bootstrap*, pages 263–270. John Wiley, New York, 1992.

[29] L. L. Ramirez Ramirez, K. Nezafati, V. Lyubchich, and Y. R. Gel. *snowboot: Bootstrap Methods for Network Inference*, 2016. R package version 0.5.1.

[30] J. Scott. Social network analysis, overview of. In *Computational Complexity*, pages 2898–2911. Springer, 2012.

[31] O. Simpson, C. Seshadhri, and A. McGregor. Catching the head, tail, and everything in between: a streaming algorithm for the degree distribution. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 979–984. IEEE, 2015.

[32] T. A. B. Snijders and S. P. Borgatti. Non-parametric standard errors and tests for network statistics. *Connections*, 22(2):161–170, 1999.

[33] M. E. Thompson, L. L. Ramirez Ramirez, V. Lyubchich, and Y. R. Gel. Using bootstrap for statistical inference on random graphs. *Canadian Journal of Statistics*, 10.1002/cjs.11271, 2015.

[34] R. van der Hofstad. *Random Graphs and Complex Networks*. http://www.win.tue.nl/ rhofstad/NotesRGCN.pdf, 2014.

[35] Y. Zhang, E. Kolaczyk, and B. Spencer. Estimating network degree distributions under sampling: An inverse problem, with applications to monitoring social media networks. *Annals of Applied Statistics*, 9(1):166–199, 2015.