# Estimation of Subgraph Densities in Noisy Networks

Jinyuan Chang, Eric D. Kolaczyk & Qiwei Yao

Taylor & Francis
Taylor & Francis Group

Check for updates

# Estimation of Subgraph Densities in Noisy Networks

Jinyuan Chang[a], Eric D. Kolaczyk[b], and Qiwei Yao[c]

[a]School of Statistics, Southwestern University of Finance and Economics, Chengdu, China; [b]Department of Mathematics and Statistics, Boston University, Boston, MA; [c]Department of Statistics, London School of Economics and Political Science, London, UK

## ABSTRACT

While it is common practice in applied network analysis to report various standard network summary statistics, these numbers are rarely accompanied by uncertainty quantification. Yet any error inherent in the measurements underlying the construction of the network, or in the network construction procedure itself, necessarily must propagate to any summary statistics reported. Here we study the problem of estimating the density of an arbitrary subgraph, given a noisy version of some underlying network as data. Under a simple model of network error, we show that consistent estimation of such densities is impossible when the rates of error are unknown and only a single network is observed. Accordingly, we develop method-of-moment estimators of network subgraph densities and error rates for the case where a minimal number of network replicates are available. These estimators are shown to be asymptotically normal as the number of vertices increases to infinity. We also provide confidence intervals for quantifying the uncertainty in these estimates based on the asymptotic normality. To construct the confidence intervals, a new and nonstandard bootstrap method is proposed to compute asymptotic variances, which is infeasible otherwise. We illustrate the proposed methods in the context of gene coexpression networks. Supplementary materials for this article are available online.

## 1. Introduction

An applied analysis in network science typically includes the following three steps: (i) gather basic measurements relevant to the interactions among elements in a system of interest, (ii) construct a network-based representation of that system, with nodes serving as elements and links indicating interactions between pairs of elements, and (iii) report various numerical summaries of network structure (e.g., density, centralities, etc.). Necessarily, uncertainty at the level of the basic measurements in the first step will propagate to the network constructed in the second step and thus to the summaries reported in the third step.

The potential for measurement error arises in nearly every network analysis application. Here, by "measurement error" we will specifically mean true edges being observed as nonedges, and vice versa—there are, of course, other notions of error that might be considered. Such edge noise occurs in online social networks (e.g., Facebook), which are often based on the extraction and merging of lists of "friends" from millions of individual accounts, where uniqueness of names is not assured. Similarly, it can be found in biological networks (e.g., of gene regulatory relationships), which are often based on notions of association (e.g., correlation, partial correlation, etc.) among experimental measurements of gene activity levels that are determined by some form of statistical inference. Finally, maps of the logical internet traditionally have been synthesized from the results of surveys in which paths along which information flows are learned experimentally through a large set of packet probes (e.g., via traceroute). See Chapter 3.5 of Kolaczyk (2009) for several

detailed examples of applied network analyses associated with such data.

That there is measurement error associated with these and other common types of network constructions is typically well-understood by practitioners. And in many settings the general issue has received substantial attention, such as, for example, in the context of protein-protein interaction networks (e.g., Hart, Ramani, and Marcotte 2006) or social networks (e.g., Almquist 2012). But, to our best knowledge, there has been little attention to date given toward formal development of statistical methods accounting for propagation of network error. Exceptions include statistical methodology for predicting network topology or attributes with models that explicitly include a component for network noise (e.g., Jiang, Gold, and Kolaczyk 2011; Jiang and Kolaczyk 2012), the "denoising" of noisy networks (e.g., Chatterjee 2015), and the adaptation of methods for vertex classification using networks observed with errors (Priebe et al. 2015).

Motivating our own work is that of Balachandran, Kolaczyk, and Viles (2017). Working with the analogue of a "signal plus noise" model for networks, these authors characterize the asymptotic distribution of the empirical edge density (i.e., formally, the density of observed edges) in noisy networks, in the context of what they call low-rate measurement error. Gan and Kolaczyk (2018) offered a refinement. The edge density is an important prototype, as it is a fundamental characteristic of networks. Its calculation generally is one of the first steps in an applied network analysis, analogous to computing a sample mean in analyzing traditional data. Additionally, the edge density is understood to be a key driver of various other

network characteristics—for example, placing limits on the frequency of higher-order subgraphs (e.g., Turán 1941). We note that the work in these two papers is entirely probabilistic in nature, focused on approximation error using Stein's method. Here our focus is statistical in nature.

In particular, here we study the problem of estimating subgraph densities, with the edge density serving as a critical initial case. We adopt a simple model for noisy networks that, conditional on some true underlying network, assumes we observe a version of that network corrupted by an independent random noise that effectively flips the status of (non)edges. If it is known the rates at which edges are instead observed as nonedges, and nonedges as edges, then it is straightforward to construct a moment-based estimator of the density of a given subgraph of interest from a single noisy network. However, in the more realistic setting in which one or both of these error rates are unknown and must themselves be estimated, the problem of identifiability arises. The problem in this case is analogous to estimation under a two-component mixture model. We show that consistent estimation of any subgraph density is in fact impossible under this setting.

The primary contribution in this article is our development of method-of-moments estimators for network subgraph densities and the underlying rates of error when replicates of the observed network are available. Beginning with the fundamental case of edge density, we provide estimators that are asymptotically normal (as the number of vertices increases to infinity) when one or both of the error rates are unknown, using a minimum of two or three replicates, respectively. The asymptotic normality in turn facilitates interval estimation for network edge density. We then extend the method-of-moments estimator to the context of an arbitrary higher-order subgraph density, and illustrate with the cases of two-star and triangle densities, as well as the clustering coefficient (or transitivity). To construct their confidence intervals, a new and nonstandard bootstrap method is proposed in order to compute asymptotic variances, which is infeasible otherwise. Numerical simulation suggests that high accuracy is possible for networks of even modest size. We illustrate the practical use of our estimators in the context of gene coexpression networks, where a small number of replicates of the basic underlying measurements (e.g., microarray expression) are frequently available.

It is difficult to overstate how ubiquitous is the use of subgraph densities in empirical network analysis. As a result, our work here is relevant to a broad and diverse cross-section of literature in humanities, social, and natural sciences, as touched by the applied network analysis literature. Certain subgraph densities (i.e., the edge density and the two-star and triangle densities, through the clustering coefficient) are reported as commonly in network analysis as one reports, say, the mean, median, and standard deviation in standard data analysis. In fact, they feature in what at least one author has termed "the network analysis 'five-number summary'" (Luke 2015). Prolific use of subgraph densities is also found in the so-called "triad census" that is standard in social network analysis (e.g., Wasserman and Faust 1994) and in the context of "motif analysis" (Milo et al. 2002), the latter being fundamental to both computational biology (e.g., Stone, Simberloff, and Artzy-Randrup 2019) and computational neuroscience (e.g., Sporns and Betzel 2016).

To date researchers doing empirical network analysis have necessarily had to report these and other types of subgraph densities simply as descriptive summaries, lacking a statistically principled framework for assessing and correcting for bias and for quantifying uncertainty due to network noise. Our work here not only provides such a framework but also demonstrates, in the context of a typical exercise in computational biology, that the nature and impact of network noise on the standard practice of reporting subgraph densities is almost surely more nuanced and pronounced than the general practitioner likely imagines. In addition, of independent interest specifically to statisticians within our work are (i) the impossibility theorem described in Theorem 1, and (ii) the nonstandard bootstrap algorithm following Theorem 4.

The rest of the article is organized as follows. Section 2 introduces the problem to be tackled. Section 3 deals with the estimation of error rates and the inference of edge density. It also reveals the innate difficulty associated with estimation when the error rates are unknown. Section 4 addresses the inference of subgraph densities in general. Numerical illustration is reported in Section 5. Some further discussion of our work is stated in Section 6. All technical proofs are relegated to supplementary materials.

## 2. Notation, Assumptions, and Problem Statement

### 2.1. Noisy Networks

Let $G = (V, E)$ be a graph, with vertices $V = \{1, \ldots, p\}$ and edges $E \subseteq V^2$. We observe a noisy version of $G$, say, $G^{\mathrm{obs}} = (V, E^{\mathrm{obs}})$, where we implicitly assume that the vertex set $V$ is known. Denote the $p \times p$ adjacency matrix of $G$ by $\mathbf{A} = (A_{i,j})_{p \times p}$, and that of $G^{\mathrm{obs}}$ by $\mathbf{Y} = (Y_{i,j})_{p \times p}$. Hence $A_{i,j} = 1$ if there is a true edge between the $i$th vertex and the $j$th vertex, and 0 otherwise, while $Y_{i,j} = 1$ if an edge is observed between the $i$th vertex and the $j$th vertex, and 0 otherwise. We assume throughout that $G$ and $G^{\mathrm{obs}}$ are simple, that is, that they possess neither multi-edges nor self-loops. An implication of the latter is that $A_{i,i} = Y_{i,i} \equiv 0$. Note that for the sake of exposition, we assume $G$ to be undirected. Then $A_{i,j} = A_{j,i}$ and $Y_{i,j} = Y_{j,i}$ for any $i \neq j$. Extension to directed graphs is straightforward and discussed briefly in Section 6.

Following Balachandran, Kolaczyk, and Viles (2017), we specify the errors in the noisy network $G^{\mathrm{obs}}$ as follows:

$$\mathbb{P}(Y_{i,j} = 1 \,|\, A_{i,j} = 0) = \alpha \quad \text{and} \quad \mathbb{P}(Y_{i,j} = 0 \,|\, A_{i,j} = 1) = \beta \tag{1}$$

for any $1 \leq i < j \leq p$. Note that $\alpha$ and $\beta$ may be interpreted, respectively, as Type I and II error rates. We assume that both $\alpha$ and $\beta$ remain constant over different edges. For some applications, $\alpha$ is known as, for example, the nominal significance level of statistical tests for the null hypothesis that there is no edge between one vertex and another. If one applies the same test method over different vertex pairs, and assumes (approximately) equal strength of "signal" across the network, then the power of the test $1 - \beta$, though unknown, also remains (approximately) the same.

Inspired by the conventional treatment of regression analysis in which inference is conditionally on regressors (i.e., treating

them as constants) and with additive noise, we treat $A_{i,j}$ as constants and assume

$$Y_{i,j} = A_{i,j}I(\varepsilon_{i,j} = 0) + I(\varepsilon_{i,j} = 1) \qquad (2)$$

for any $1 \le i < j \le p$, where $I(\cdot)$ denotes the indicator function, and $\{\varepsilon_{i,j}\}$ are specified in Assumption 1.

*Assumption 1.* The $\varepsilon_{i,j}$, for all $1 \le i < j \le p$, are independent random variables with $\mathbb{P}(\varepsilon_{i,j} = 1) = \alpha$, $\mathbb{P}(\varepsilon_{i,j} = 0) = 1 - \alpha - \beta$, and $\mathbb{P}(\varepsilon_{i,j} = -1) = \beta$, where $\alpha, \beta \ge 0$ and $\alpha + \beta < 1$.

Now (1) follows from (2) and Assumption 1 immediately. The independence condition in Assumption 1 is not strictly necessary. See Remark 1 in Section 3.2.1.

## 2.2. Subgraph Density

A standard quantity of general interest in practice is the density of certain subgraphs in $G$. Subgraphs of common interest include (i) edges, (ii) two-stars (also called triples) and other higher-order $k$-stars, (iii) triangles and other higher-order cliques, (iv) chains, and (v) cycles. Subgraph density is simply the total number of times a given subgraph, say $H$, is found in $G$ (where, note, overlap among copies of $H$ is allowed), divided by the maximum number of copies possible in a graph of the same number of vertices as $G$. There are different ways to express this notion formally. Intuitively, for example, the count $f_H(G)$ of the number of distinct copies of a subgraph $H$ in $G$ is represented as

$$f_H(G) = \frac{1}{|\text{Iso}(H)|} \sum_{H' \subseteq K_p, H' \cong H} I(H' \subseteq G), \qquad (3)$$

where $K_p$ is the complete graph on $p$ vertices and $H \subseteq G$ indicates that $H$ is a subgraph of $G$ (i.e., $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$). The value $|\text{Iso}(H)|$ is a normalization factor for the number of isomorphisms of $H$. Normalizing $f_H(G)$, in turn, by the total number of copies of $H$ possible in the complete graph $K_p$ then yields the density of subgraph $H$ in $G$.

For our purposes, it is more convenient to adopt an alternative expression for subgraph density—albeit one that is notationally more cumbersome. Consider an arbitrary subgraph $H = (V_H, E_H)$ of interest, of order $|V_H| \ge 2$. We characterize such subgraphs in terms of an index set $\mathcal{V} = \mathcal{V}_H$ of the following generic form

$$\mathcal{V} = \{(i_1, i'_1, \ldots, i_k, i'_k) :$$
$$i_\ell \ne i'_\ell \text{ for each } \ell = 1, \ldots, k, \ |\{i_{\ell_1}, i'_{\ell_1}\} \cap \{i_{\ell_2}, i'_{\ell_2}\}| \le 1$$
$$\text{for any } \ell_1 \ne \ell_2, \text{ and } i_1, i'_1, \ldots, i_k, i'_k \text{ also}$$
$$\text{satisfying other restrictions imposed by } H\},$$
$$(4)$$

and $k$ prescribed values $\tau_1, \ldots, \tau_k \in \{0, 1\}$. We then represent the subgraph density for any subgraph $H$ in $G$ as

$$C_{\mathcal{V}}(\tau_1, \ldots, \tau_k) = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v} = (i_1, i'_1, \ldots, i_k, i'_k) \in \mathcal{V}}$$
$$A_{i_1, i'_1}^{\tau_1} (1 - A_{i_1, i'_1})^{1-\tau_1} \cdots A_{i_k, i'_k}^{\tau_k} (1 - A_{i_k, i'_k})^{1-\tau_k}. \qquad (5)$$

Here, we adopt the convention $0^0 = 1$.

The quantity $C_{\mathcal{V}}(\tau_1, \ldots, \tau_k)$ defined in (5) is quite general. For example, if we let $k = 1$ and $\tau_1 = 1$, it reduces to the edge density defined in (6) below, which is arguably the most important single-number summary for networks. If we select $\tau_1 = \cdots = \tau_k = 1$ and $\mathcal{V} = \{(i_1, i'_1, \ldots, i_k, i'_k) : i'_\ell = i_{\ell+1} \text{ for each } \ell = 1, \ldots, k-1, i_1 \ne i_2 \ne \cdots \ne i_k \ne i'_k\}$, then

$$C_{\mathcal{V}}(\tau_1, \ldots, \tau_k) = \frac{1}{p \cdots (p - k)} \sum_{i_1 \ne \cdots \ne i_{k+1}} A_{i_1, i_2} \cdots A_{i_k, i_{k+1}},$$

which is the density of $k$ connected edges in $G$ passing through $k + 1$ different nodes—that is, the density of paths of length $k$. If in addition we impose the constraint that the path must start and end with the same vertex, we select $\tau_1 = \cdots = \tau_k = 1$ and $\mathcal{V} = \{(i_1, i'_1, \ldots, i_k, i'_k) : i'_\ell = i_{\ell+1} \text{ for each } \ell = 1, \ldots, k-1, i'_k = i_1, i_1 \ne i_2 \ne \cdots \ne i_k\}$, yielding

$$C_{\mathcal{V}}(\tau_1, \ldots, \tau_k) = \frac{1}{p \cdots (p - k + 1)} \sum_{i_1 \ne \cdots \ne i_k} A_{i_1, i_2} \cdots A_{i_k, i_1},$$

which is the density of cycles of length $k$ in $G$. An important special case of the latter is when $k = 3$, which yields the density of closed triples in $G$ (generally interpreted as three times the density of triangles). Similarly, if the summands $A_{i_1,i_2}A_{i_2,i_3}A_{i_3,i_1}$ associated with the triangle density are instead replaced by $A_{i_1,i_2}A_{i_2,i_3}(1 - A_{i_3,i_1})$, we obtain the density of (open) connected triples or two-stars. In turn, the ratio of the first of these two quantities to its sum with the second defines the clustering coefficient (also called the transitivity) of $G$—arguably the second most important summary statistic in practice after the edge density.

In practice, given a noisy network, researchers currently report the empirical subgraph densities (i.e., $C_{\mathcal{V}}$ applied to $G^{\text{obs}}$ with adjacency matrix $\mathbf{Y}$) and assume that they are reflective of the corresponding true subgraph densities (i.e., $C_{\mathcal{V}}$ applied to $G$ with adjacency matrix $\mathbf{A}$). The work of Balachandran, Kolaczyk, and Viles (2017) shows that, under conditions similar to those assumed here, there is in general no reason to expect that these empirical (or "plug in") estimates are even consistent. Our goal in this article is to produce principled and accurate estimates of subgraph densities. In what follows, we treat the estimation of edge density as a special base case, which helps inform the exposition of our results for general subgraph density estimation.

## 3. Inference for Edge Density

In this section, we consider inference of the edge density with unknown error rates $\alpha$ and $\beta$. The edge is the simplest subgraph. The count of the number of edges or, upon normalization, the so-called edge density (aka network density) is defined as follows:

$$\delta = \frac{2}{p(p - 1)} \sum_{i<j} A_{i,j}. \qquad (6)$$

It is both useful, from the perspective of our mathematical development, and fundamental, from the perspective of network theory and applications, to focus first on the edge density $\delta$ as the estimand of interest. It reveals the innate difficulty associated with estimation under unknown error rates. See Section 3.1. The inference for general subgraphs will be presented in Section 4.
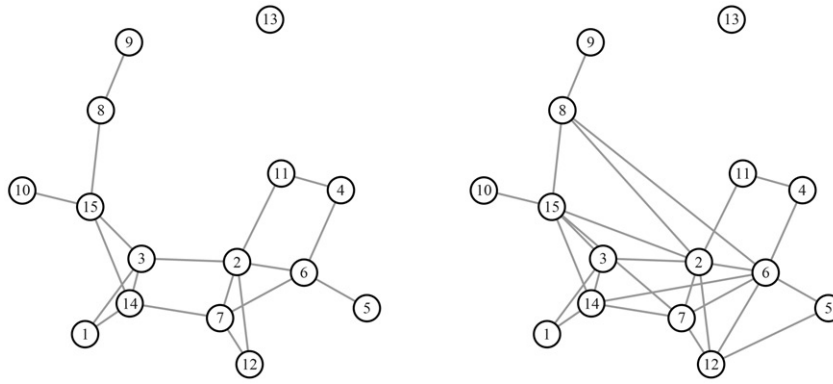
**Figure 1.** Left panel: True network $G$, with $p = 15$ nodes and density $\delta = 0.181$. Right panel: Noisy network $G^{\text{obs}}$, with $\alpha = 0.05$, $\beta = 0.15$, and $\bar{Y} = 0.229$. The goal is to estimate $\delta$ based on the noisy network.

### 3.1. Difficulty of Estimating Subgraph Densities

Consider estimation of the network edge density $\delta$ in (6). Figure 1 presents a simple visual illustration of our task. The network on the left with $p = 15$ nodes is defined by a deterministic adjacency matrix $\mathbf{A}$ with 19 edges, and hence the network density $\delta = 2 \times 19/(15 \times 14) = 0.181$. The noisy network on the right defined by the adjacency matrix $\mathbf{Y}$ was observed with 24 edges, where $\mathbf{Y} = (Y_{i,j})_{15 \times 15}$ is generated from $\mathbf{A}$ by (2) with $\alpha = 5\%$ and $\beta = 15\%$. Our task is to estimate $\delta$ based on $\mathbf{Y}$.

A natural estimator for $\delta$ is given by

$$\bar{Y} = \frac{2}{p(p-1)} \sum_{i<j} Y_{i,j}.$$

In the illustration of Figure 1, this value is $\bar{Y} = 0.229$, in comparison to the true value $\delta = 0.181$. Let $\mathcal{S} = \{(i,j) : A_{i,j} = 1, \, i < j\}$ and $\mathcal{S}^c = \{(i,j) : A_{i,j} = 0, \, i < j\}$. From (6), we know $\bar{Y}$ is a biased estimator for $\delta$. More specifically, we have

$$\mathbb{E}(\bar{Y}) = \frac{2}{p(p-1)} \sum_{(i,j) \in \mathcal{S}} \mathbb{E}(Y_{i,j}) + \frac{2}{p(p-1)} \sum_{(i,j) \in \mathcal{S}^c} \mathbb{E}(Y_{i,j}) \tag{7}$$

$$= \delta(1-\beta) + (1-\delta)\alpha.$$

But if $\alpha$ and $\beta$ are known, (7) suggests estimating $\delta$ instead by

$$\widetilde{Y} = \frac{\bar{Y} - \alpha}{1 - \alpha - \beta}. \tag{8}$$

Equation (8) defines a consistent estimator for $\delta$.

In practice, however, values for $\alpha$ and $\beta$ typically are not readily obtainable, and one or both must be estimated. This makes the problem of estimating $\delta$ decidedly more difficult. In fact, it is essentially impossible to estimate any subgraph count $f_H(G)$ from a single noisy observation $G^{\text{obs}}$.

Formally, let $\mathcal{M} = \{(\alpha, \beta, \mathbf{A}) : 0 \le \alpha \le 1, 0 \le \beta \le 1, A_{i,j} = 0 \text{ or } 1, A_{i,j} = A_{j,i}\}$ be the class of all models defined under (2) and Assumption 1. For any model $M = (\alpha, \beta, \mathbf{A}) \in \mathcal{M}$, we define its dual model as $M^* = (1 - \beta, 1 - \alpha, \mathbf{A}^*)$, where $\mathbf{A}^* = (A_{i,j}^*)_{p \times p}$ satisfies $A_{i,j}^* = 1 - A_{i,j}$ for any $i \ne j$. Denote by $F_M$ and $F_{M^*}$ the joint distributions of $\mathbf{Y}$ when $\mathbf{Y}$ follows models $M$ and $M^*$, respectively. Finally, for a given subgraph density $f$ of interest, define

$$d_f = \sup_{M \in \mathcal{M}} |f(M) - f(M^*)|,$$

where $f(M)$ and $f(M^*)$ are the associated subgraph densities based on model $M$ and its dual model $M^*$, respectively. We then have the following result. Note that Theorem 1 holds for any subgraph density $f$. When $f$ is the edge density, $d_f = 1$.

*Theorem 1.* Write $\mathcal{E}$ for the class of all measurable functionals of the data $\mathbf{Y}$. Let Assumption 1 hold. If $d_f > 0$, then it holds that

$$\inf_{\hat{f} \in \mathcal{E}} \sup_{\mathcal{M}} \mathbb{P}\left(|\hat{f} - f| \ge \frac{d_f}{2}\right) \ge \frac{1}{2}.$$

Theorem 1 indicates that it is in general impossible to produce a consistent estimate of a subgraph density $f$ based on only one noisy version of the adjacency matrix $\mathbf{A}$.

To build intuition for the difficulty of this problem, consider again Equation (7), which indicates that $\bar{Y}$ is an unbiased estimate of

$$u_1 \equiv (1-\delta)\alpha + \delta(1-\beta),$$

rather than of $\delta$. This observation suggests use of the (asymptotically) unbiased estimating equation

$$\hat{u}_1 = (1-\delta)\alpha + \delta(1-\beta), \tag{9}$$

where $\hat{u}_1 = \bar{Y}$. It is obvious that $\alpha, \beta$, and $\delta$ cannot all be uniquely identified from this single equation.

Fortunately, in certain key areas of application we may observe more than one noisy version of the target network $G$. For example, in computational biology, the common use of replicates at the most basic level of measurement (e.g., microarray expression) often allows for the construction of replicate networks (e.g., coexpression networks), as we demonstrate in Section 5. Similarly, in the context of computational neuroscience, it has become common now to obtain imaging measurements (e.g., fMRI) on multiple individuals within a given subpopulation (e.g., healthy females of a given age) and to create networks (e.g., functional connectivity networks) for each individual. In the remainder of this section, we demonstrate how to estimate the edge density of the adjacency matrix $\mathbf{A}$ consistently using just two or three replicates. We then develop generalizations of these results for the case of arbitrary subgraphs in Section 4.

### 3.2. Estimation of Unknown Error Rates

#### 3.2.1. One of $\alpha$ or $\beta$ Known

In some settings, one of either $\alpha$ or $\beta$ may be known. For example, if the edges in **Y** are inferred through formal hypothesis testing, then $\alpha$ would be the user-specified rate of Type I error. In this case, there are only two unknown parameters that need to be estimated, and we demonstrate how two replicates are sufficient to do so.

Suppose that **Y** is defined as above, and that $\mathbf{Y}_* = (Y_{i,j,*})_{p \times p}$ is an independent and identically distributed replicate of **Y**. Both are then noisy versions of the same adjacency matrix **A**, observed with the same error rates $\alpha$ and $\beta$. It follows from (2) that for $(i,j)$ with $A_{i,j} = 1$,

$$
Y_{i,j,*} - Y_{i,j} = \begin{cases} -1, & \text{with probability } \beta(1 - \beta), \\ 0, & \text{with probability } 1 - 2\beta(1 - \beta), \\ 1, & \text{with probability } \beta(1 - \beta), \end{cases}
$$

and for $(i,j)$ with $A_{i,j} = 0$,

$$
Y_{i,j,*} - Y_{i,j} = \begin{cases} -1, & \text{with probability } \alpha(1 - \alpha), \\ 0, & \text{with probability } 1 - 2\alpha(1 - \alpha), \\ 1, & \text{with probability } \alpha(1 - \alpha). \end{cases}
$$

Similar to (7), we have

$$
\mathbb{E}\left\{ \frac{2}{p(p-1)} \sum_{i<j} |Y_{i,j,*} - Y_{i,j}| \right\}
$$
$$
= 2\{(1 - \delta)\alpha(1 - \alpha) + \delta\beta(1 - \beta)\}.
$$

Let

$$
u_2 \equiv (1 - \delta)\alpha(1 - \alpha) + \delta\beta(1 - \beta), \tag{10}
$$

for which the method of moment estimate is

$$
\hat{u}_2 = \frac{1}{p(p-1)} \sum_{i<j} |Y_{i,j,*} - Y_{i,j}|.
$$

Therefore, we have a second estimating equation:

$$
\hat{u}_2 = (1 - \delta)\alpha(1 - \alpha) + \delta\beta(1 - \beta). \tag{11}
$$

Combining (9) and (11), when $\alpha$ is known, the estimators for $\beta$ and $\delta$ are

$$
\begin{cases} \hat{\beta} = \dfrac{\hat{u}_2 - \alpha + \hat{u}_1 \alpha}{\hat{u}_1 - \alpha}, \\[2ex] \hat{\delta} = \dfrac{(\hat{u}_1 - \alpha)^2}{\hat{u}_1 - \hat{u}_2 - 2\hat{u}_1 \alpha + \alpha^2}, \end{cases} \tag{12}
$$

and when $\beta$ is known, the estimators for $\alpha$ and $\delta$ are

$$
\begin{cases} \hat{\alpha} = \dfrac{\hat{u}_1 \beta - \hat{u}_2}{\hat{u}_1 + \beta - 1}, \\[2ex] \hat{\delta} = \dfrac{\hat{u}_1^2 - \hat{u}_1 + \hat{u}_2}{\hat{u}_1 + \hat{u}_2 - 2\hat{u}_1 \beta - (1 - \beta)^2}. \end{cases} \tag{13}
$$

The following proposition gives the convergence rates for the proposed estimators.

*Proposition 1.* Let $N = p(p-1)/2$. Under Assumption 1, if $N_1 = p(p-1)\delta \to \infty$ and $N_2 = p(p-1)(1-\delta) \to \infty$, it holds that (i) $\hat{\beta} = \beta + O_p(N^{-1/2})$ and $\hat{\delta} = \delta + O_p(N^{-1/2})$, provided that $\alpha$ is known and $\delta(1 - \alpha - \beta)^2 \geq c$ for some positive constant $c$, (ii) $\hat{\alpha} = \alpha + O_p(N^{-1/2})$ and $\hat{\delta} = \delta + O_p(N^{-1/2})$, provided that $\beta$ is known and $(1 - \delta)(1 - \alpha - \beta)^2 \geq c$ for some positive constant $c$.

*Remark 1.* Since our estimation of the unknown parameters is based on moment estimation, the independent noise dictated by Assumption 1 is not strictly necessary. As is shown in the proof of Proposition 1, the convergence rate for the moment estimation of the unknown parameters is determined by the convergence rates of $\hat{u}_1 - u_1$ and $\hat{u}_2 - u_2$. For any $i < j$, let $e_{i,j} = I(\varepsilon_{i,j} = 0, 1) - (1 - \beta)$ for $(i,j) \in \mathcal{S}$ and $e_{i,j} = I(\varepsilon_{i,j} = 1) - \alpha$ for $(i,j) \in \mathcal{S}^c$. Recall $\mathbb{P}(\varepsilon_{i,j} = 1) = \alpha$, $\mathbb{P}(\varepsilon_{i,j} = 0) = 1 - \alpha - \beta$ and $\mathbb{P}(\varepsilon_{i,j} = -1) = \beta$. Then $\mathbb{E}(e_{i,j}) = 0$ for any $i < j$. If $\text{var}(N^{-1/2} \sum_{i<j} e_{i,j}) \leq C$ for some positive constant $C$, then $\hat{u}_1 = u_1 + O_p(N^{-1/2})$ without the independence assumption. When Assumption 1 is satisfied, $\text{var}(N^{-1/2} \sum_{i<j} e_{i,j}) = \delta\beta(1-\beta) + (1-\delta)\alpha(1-\alpha)$. Analogously, $\hat{u}_2 = u_2 + O_p(N^{-1/2})$ still holds when some dependency among $\varepsilon_{i,j}$ $(i < j)$ is present. Hence, the results of Proposition 1 still hold when there is some dependency among $\varepsilon_{i,j}$ $(i < j)$.

*Remark 2.* It is not strictly necessary that $\mathbf{Y}_*$ derive from the same underlying adjacency matrix **A** as **Y**. More specifically, let $\mathbf{A}_* = (A_{i,j,*})_{p \times p}$ be the adjacency matrix underlying the observation $\mathbf{Y}_*$, and let $\mathcal{B}_1 = \{(i,j) : A_{i,j} = A_{i,j,*}, i < j\}$. The average of $|Y_{i,j,*} - Y_{i,j}|$ over $\mathcal{B}_1$ provides an unbiased estimator for the parameter $u_2$ defined in (10), while the original estimator $\hat{u}_2$ defined in (11) is no longer unbiased if $|\mathcal{B}_1| < p(p-1)/2$. As long as $\theta_1 = 2|\mathcal{B}_1|/\{p(p-1)\}$ is sufficiently close to 1, e.g. $|1 - \theta_1| = o(p^{-1})$, the bias term in $\hat{u}_2$ will be asymptotically negligible, which means the estimators (12) and (13) will still be consistent.

*Theorem 2.* Let $N = p(p-1)/2$. Under Assumption 1, if $N_1 = p(p-1)\delta \to \infty$ and $N_2 = p(p-1)(1-\delta) \to \infty$, it holds that (i) $\sqrt{N}(\hat{\beta} - \beta, \hat{\delta} - \delta)^{\mathsf{T}} \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{1,\alpha})$ with $\boldsymbol{\Sigma}_{1,\alpha}$ defined as (A.2) in the Appendix, provided that $\alpha$ is known and $\delta(1 - \alpha - \beta)^2 \geq c$ for some positive constant $c$, (ii) $\sqrt{N}(\hat{\alpha} - \alpha, \hat{\delta} - \delta)^{\mathsf{T}} \to_d \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{1,\beta})$ with $\boldsymbol{\Sigma}_{1,\beta}$ defined as (A.3) in the Appendix, provided that $\beta$ is known and $(1 - \delta)(1 - \alpha - \beta)^2 \geq c$ for some positive constant $c$.

We can construct approximate confidence intervals for $\delta$ based on the asymptotic normality stated in Theorem 2. Let $\sigma^2$ denote the asymptotic variance of $\sqrt{N}(\hat{\delta} - \delta)$. Then $\sigma$ depends on unknown parameters $\delta$ and $\beta$ or $\alpha$. Replacing those unknown parameters by their estimates, we obtain an estimated asymptotic variance denoted by $\hat{\sigma}^2$. Then an approximate 95% confidence interval for $\delta$ is

$$
\left( \hat{\delta} - 1.96\hat{\sigma}N^{-1/2}, \ \hat{\delta} + 1.96\hat{\sigma}N^{-1/2} \right). \tag{14}
$$

#### 3.2.2. Both $\alpha$ and $\beta$ Unknown

When both $\alpha$ and $\beta$ are unknown, together with $\delta$ there are three unknown parameters to be estimated. We show that three

replicates are sufficient for asymptotically consistent estimation in this setting.

Let $\mathbf{Y}$, $\mathbf{Y}_*$, and $\mathbf{Y}_{**}$ be independent and identically distributed replicates from (2). Hence, for $(i, j)$ with $A_{i,j} = 1$,

$$
Y_{i,j,**} - 2Y_{i,j,*} + Y_{i,j} = \begin{cases} -2, & \text{with probability } \beta^2(1-\beta), \\ -1, & \text{with probability } 2\beta(1-\beta)^2, \\ 0, & \text{with probability } \beta^3 + (1-\beta)^3, \\ 1, & \text{with probability } 2\beta^2(1-\beta), \\ 2, & \text{with probability } \beta(1-\beta)^2, \end{cases}
$$

and for $(i, j)$ with $A_{i,j} = 0$,

$$
Y_{i,j,**} - 2Y_{i,j,*} + Y_{i,j} = \begin{cases} -2, & \text{with probability } \alpha(1-\alpha)^2, \\ -1, & \text{with probability } 2\alpha^2(1-\alpha), \\ 0, & \text{with probability } \alpha^3 + (1-\alpha)^3, \\ 1, & \text{with probability } 2\alpha(1-\alpha)^2, \\ 2, & \text{with probability } \alpha^2(1-\alpha). \end{cases}
$$

Arguing in a fashion analogous to that used in producing the parameters $u_1$ and $u_2$, we emerge with the parameter

$$
u_3 \equiv (1-\delta)\alpha(1-\alpha)^2 + \delta\beta^2(1-\beta), \tag{15}
$$

with corresponding method of moment estimator

$$
\hat{u}_3 = \frac{2}{3p(p-1)} \sum_{i<j} I(Y_{i,j,**} - 2Y_{i,j,*} + Y_{i,j} = 1 \text{ or } -2),
$$

from which we obtain a third estimating equation:

$$
\hat{u}_3 = (1-\delta)\alpha(1-\alpha)^2 + \delta\beta^2(1-\beta). \tag{16}
$$

Combining (9), (11), and (16), we have a nonlinear system of three equations with three unknowns. This nonlinear system can be solved by some simple numerical iterations. For example, it follows from (15) that

$$
\hat{\alpha} = \frac{\hat{u}_3 - \delta\beta^2(1-\beta)}{(1-\delta)(1-\alpha)^2}. \tag{17}
$$

Starting with an initial value $\alpha_0$, we compute the estimates for $\beta$, $\delta$, and $\alpha$ recursively using (12) and (17) until the absolute difference between two successive values for $\hat{\alpha}$ is smaller than a prescribed small number. Analogous to Proposition 1 and Theorem 2, we have the following result.

*Theorem 3.* Let $N = p(p-1)/2$. Under Assumption 1, if $N_1 = p(p-1)\delta \to \infty$ and $N_2 = p(p-1)(1-\delta) \to \infty$, it holds that $\hat{\alpha} = \alpha + O_p(N^{-1/2})$, $\hat{\beta} = \beta + O_p(N^{-1/2})$, and $\hat{\delta} = \delta + O_p(N^{-1/2})$, provided that $\delta(1-\delta)(1-\alpha-\beta)^4 \geq c$ for some positive constant $c$. More specifically, we have $\sqrt{N}(\hat{\alpha} - \alpha, \hat{\beta} - \beta, \hat{\delta} - \delta)^{\mathrm{T}} \to_d \mathcal{N}(\mathbf{0}, \Sigma_2)$ with $\Sigma_2$ defined as (A.4) in the Appendix.

## 4. Inference for Higher-Order Subgraph Densities

Now we address the inference of higher-order subgraph densities $C_{\mathcal{V}}(\tau_1, \ldots, \tau_k)$ defined in (5) with $k \geq 2$. We continue to use method of moments estimation, but with the error rates $\alpha$ and/or $\beta$ replaced by their estimators obtained in Section 3.2. The resulting estimators admit a uniform representation; see (25). However, interval estimation for $C_{\mathcal{V}}(\tau_1, \ldots, \tau_k)$ requires the evaluation of an asymptotic variance that is a function of the individual (unknown) network edges $A_{i,j}$. Accordingly, we propose a new and nonstandard bootstrap method to overcome this obstacle. To highlight the key ideas, we first proceed in Section 4.1 with both $\alpha$ and $\beta$ assumed to be known. The development with unknown $\alpha$ and $\beta$ is then presented in Section 4.2.

### 4.1. Inference for Subgraph Densities With Known Error Rates

In this subsection, we assume that both $\alpha$ and $\beta$ are known. All inference will be based on one observed network $\mathbf{Y} = (Y_{i,j})_{p \times p}$ only. It follows from (2) and Assumption 1 that

$$
A_{i,j} = \frac{\mathbb{E}(Y_{i,j} - \alpha)}{1 - \alpha - \beta} \quad \text{and} \quad 1 - A_{i,j} = \frac{\mathbb{E}(1 - \beta - Y_{i,j})}{1 - \alpha - \beta}.
$$

Hence (5) admits a more compact representation

$$
C_{\mathcal{V}} =: C_{\mathcal{V}}(\tau_1, \ldots, \tau_k)
$$
$$
= \frac{1}{(1 - \alpha - \beta)^k} \cdot \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \prod_{\ell=1}^{k} \mathbb{E}\{\varphi_\ell(Y_{i_\ell, i_\ell'})\},
$$

where

$$
\varphi_\ell(x) = (x - \alpha)^{\tau_\ell}(1 - \beta - x)^{1 - \tau_\ell}. \tag{18}
$$

Note that $|\{i_{\ell_1}, i_{\ell_1}'\} \cap \{i_{\ell_2}, i_{\ell_2}'\}| \leq 1$ for any $\ell_1 \neq \ell_2$, Assumption 1 implies that the $\{Y_{i_\ell, i_\ell'}\}_{\ell=1}^k$ are independent of each other. Therefore, a natural method of moments estimator for $C_{\mathcal{V}}$ can be defined as

$$
\widetilde{C}_{\mathcal{V}} = \frac{\widetilde{T}_{\mathcal{V}}}{(1 - \alpha - \beta)^k}, \tag{19}
$$

where

$$
\widetilde{T}_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \prod_{\ell=1}^{k} \varphi_\ell(Y_{i_\ell, i_\ell'}).
$$

To state the asymptotic properties of $\widetilde{C}_{\mathcal{V}}$, we need to introduce some notation. For any $\mathbf{v} = (i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}$ with $\mathcal{V}$ given in (4) and $1 \leq \ell_1 < \cdots < \ell_s \leq k$ with $1 \leq s \leq k-1$, we define

$$
\mathcal{G}_{\ell_1, \ldots, \ell_s}(\mathbf{v}) = \{(\theta_1, \theta_1', \ldots, \theta_s, \theta_s') : (i_1, i_1', \ldots, i_{\ell_1-1}, i_{\ell_1-1}', \theta_1, \theta_1',
$$
$$
i_{\ell_1+1}, i_{\ell_1+1}', \ldots, i_{\ell_2-1}, i_{\ell_2-1}', \theta_2, \theta_2',
$$
$$
\ldots, i_{\ell_s-1}, i_{\ell_s-1}', \theta_s, \theta_s', i_{\ell_s+1}, i_{\ell_s+1}', \ldots, i_k, i_k') \in \mathcal{V}\}.
$$

In turn, we define the quantity

$$
\aleph_{\mathcal{V}}(s) = \max_{\mathbf{v} \in \mathcal{V}} \max_{1 \leq \ell_1 < \cdots < \ell_s \leq k} |\mathcal{G}_{\ell_1, \ldots, \ell_s}(\mathbf{v})| \tag{20}
$$

and

$$
\aleph_{\mathcal{V}} = \max_{1 \leq s \leq k-1} \aleph_{\mathcal{V}}(s).
$$

*Proposition 2.* Under Assumption 1, if $|1 - \alpha - \beta| \geq c$ for some positive constant $c$, it holds that $|\widetilde{C}_\mathcal{V} - C_\mathcal{V}| = O_p(\sqrt{\aleph_\mathcal{V}/|\mathcal{V}|})$ as $p \to \infty$.

Notice that $\aleph_\mathcal{V}(1) \leq \cdots \leq \aleph_\mathcal{V}(k-1)$, so that $\aleph_\mathcal{V} = \aleph_\mathcal{V}(k-1)$. If we select $\mathcal{V} = \{(i_1, i_1', \ldots, i_k, i_k') : i_\ell' = i_{\ell+1} \text{ for each } \ell = 1, \ldots, k-1, i_1 \neq i_2 \neq \cdots \neq i_k \neq i_k'\}$, which corresponds to counting paths of length $k \geq 2$, then $|\mathcal{V}| = p \cdots (p - k)$ and $\aleph_\mathcal{V}(s) = (p - k + s - 1) \cdots (p - k)$ for any $1 \leq s \leq k - 1$. Alternately, if we select $\mathcal{V} = \{(i_1, i_1', \ldots, i_k, i_k') : i_\ell' = i_{\ell+1} \text{ for each } \ell = 1, \ldots, k-1, i_k' = i_1, i_1 \neq i_2 \neq \cdots \neq i_k\}$, which corresponds to counting cycles of length $k \geq 3$, then $|\mathcal{V}| = p \cdots (p - k + 1)$, $\aleph_\mathcal{V}(1) = 1$ and $\aleph_\mathcal{V}(s) = (p - k + s - 1) \cdots (p - k + 1)$ for any $2 \leq s \leq k - 1$. As a result, in the case of counting paths or cycles of length $k$, $\aleph_\mathcal{V}/|\mathcal{V}| = \{p(p-1)\}^{-1}$. Letting $N = p(p-1)/2$, we then have $|\widetilde{C}_\mathcal{V} - C_\mathcal{V}| = O_p(N^{-1/2})$.

To investigate the asymptotic distribution of $\widetilde{C}_\mathcal{V} - C_\mathcal{V}$, we require the following mild assumption.

*Assumption 2.* (i) $\aleph_\mathcal{V}(s)/\aleph_\mathcal{V} \to 0$ for any $1 \leq s \leq k - 2$, and (ii)

$$\max_{\mathbf{v} \in \mathcal{V}} \max_{1 \leq \ell_1 < \cdots < \ell_{k-1} \leq k} |\mathcal{G}_{\ell_1, \ldots, \ell_{k-1}}(\mathbf{v})|$$
$$\asymp \min_{\mathbf{v} \in \mathcal{V}} \min_{1 \leq \ell_1 < \cdots < \ell_{k-1} \leq k} |\mathcal{G}_{\ell_1, \ldots, \ell_{k-1}}(\mathbf{v})|.$$

Let

$$S_\mathcal{V} = \frac{\sqrt{N}}{(1 - \alpha - \beta)^k}$$
$$\sum_{j=1}^k \frac{(-1)^{1-\tau_j}}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \left[ \mathring{Y}_{i_j, i_j'} \prod_{\ell \neq j} \mathbb{E}\{\varphi_\ell(Y_{i_\ell, i_\ell'})\} \right],$$

(21)

where $\mathring{Y}_{i_j, i_j'} = Y_{i_j, i_j'} - \mathbb{E}(Y_{i_j, i_j'})$ and $\varphi_\ell(\cdot)$ is defined as (18).

*Proposition 3.* Let $N = p(p-1)/2$, $\aleph_\mathcal{V}/|\mathcal{V}| \asymp N^{-1}$ and $|1 - \alpha - \beta| \geq c$ for some positive constant $c$. Under Assumptions 1 and 2, it holds that $\sqrt{N}(\widetilde{C}_\mathcal{V} - C_\mathcal{V}) = S_\mathcal{V} + o_p(1)$ for $S_\mathcal{V}$ defined as (21).

Recall $Y_{i,j} = Y_{j,i}$ for any $i \neq j$, and

$$\max_{\mathbf{v} \in \mathcal{V}} \max_{1 \leq \ell_1 < \cdots < \ell_{k-1} \leq k} |\mathcal{G}_{\ell_1, \ldots, \ell_{k-1}}(\mathbf{v})|$$
$$\asymp \min_{\mathbf{v} \in \mathcal{V}} \min_{1 \leq \ell_1 < \cdots < \ell_{k-1} \leq k} |\mathcal{G}_{\ell_1, \ldots, \ell_{k-1}}(\mathbf{v})|.$$

Notice that $\aleph_\mathcal{V}/|\mathcal{V}| \asymp N^{-1}$. Then it holds that

$$S_\mathcal{V} = \frac{1}{\sqrt{N}} \sum_{i<j} \mathring{Y}_{i,j} K_{i,j}$$

for some constants $K_{i,j}$. Since $\{Y_{i,j}\}_{i<j}$ are independent, it follows from the central limit theorem that

$$\sqrt{N}(\widetilde{C}_\mathcal{V} - C_\mathcal{V}) \xrightarrow{d} \mathcal{N}(0, \sigma_\mathcal{V}^2)$$

as $p \to \infty$, where the asymptotic variance $\sigma_\mathcal{V}^2$ satisfies

$$\sigma_\mathcal{V}^2 = \lim_{p \to \infty} \frac{1}{N} \sum_{i<j} K_{i,j}^2 \text{var}(Y_{i,j}).$$

(22)

It is easy to see from (2) that $\text{var}(Y_{i,j}) = A_{i,j}(1 - \alpha - \beta)(\beta - \alpha) + \alpha(1 - \alpha)$. As we do not know $A_{i,j}$, it is impossible to compute $\sigma_\mathcal{V}^2$ based on (22) (except for some simple special cases such as when $K_{i,j}^2$ does not vary with respect to $i$ and $j$). To overcome this difficulty, we propose a nonstandard bootstrap procedure as follows: we draw bootstrap samples $Y^\dagger$ according to

$$Y_{i,j}^\dagger \equiv Y_{j,i}^\dagger = Y_{i,j} I(\eta_{i,j} = 0) + I(\eta_{i,j} = 1) \quad \text{for } i < j, \quad (23)$$

where $\{\eta_{i,j}\}$ are independent random variables, $\mathbb{P}(\eta_{i,j} = 0) = \gamma_1$, $\mathbb{P}(\eta_{i,j} = 1) = \gamma_2$ and $\mathbb{P}(\eta_{i,j} = -1) = 1 - \gamma_1 - \gamma_2$, with $\gamma_1 > 0$, $\gamma_2 > 0$ and $\gamma_1 + \gamma_2 < 1$ satisfying

$$\begin{cases} \gamma_1(1 - \gamma_1 - 2\gamma_2) = \beta - \alpha, \\ \gamma_2(1 - \gamma_2) = \alpha(1 - \beta). \end{cases} \quad (24)$$

Now let

$$S_\mathcal{V}^\dagger = \frac{\sqrt{N}}{(1 - \alpha - \beta)^k}$$
$$\sum_{j=1}^k \frac{(-1)^{1-\tau_j}}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \left\{ \mathring{Y}_{i_j, i_j'}^\dagger \prod_{\ell \neq j} \varphi_\ell(Y_{i_\ell, i_\ell'}) \right\}$$

with $\mathring{Y}_{i_j, i_j'}^\dagger = Y_{i_j, i_j'}^\dagger - Y_{i_j, i_j'}\gamma_1 - \gamma_2$ and $\varphi_\ell(\cdot)$ defined as in (18). Theorem 4 shows that the distribution of $\sqrt{N}(\widetilde{C}_\mathcal{V} - C_\mathcal{V})$ can be approximated by the conditional distribution of $S_\mathcal{V}^\dagger$ given $\mathbf{Y} = (Y_{i,j})_{p \times p}$.

Note that (24) may admit more than one legitimate solution for $(\gamma_1, \gamma_2)$; any one of them can serve for our purpose. Furthermore, the bootstrap sample $(Y_{i,j}^\dagger)_{p \times p}$ does not necessarily resemble the full behavior of the original sample $(Y_{i,j})_{p \times p}$. What matters here is the fact that it has the correct (conditional expected) variance:

$$\mathbb{E}\{\text{var}(Y_{i,j}^\dagger \mid \mathbf{Y})\} = \text{var}(Y_{i,j}).$$

Note that $\text{var}(Y_{i,j}^\dagger \mid \mathbf{Y}) = Y_{i,j}(\beta - \alpha) + \alpha(1 - \beta)$, which is guaranteed by (24).

*Theorem 4.* Under the conditions of Proposition 3, it holds that

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\{\sqrt{N}(\widetilde{C}_\mathcal{V} - C_\mathcal{V}) > z\} - \mathbb{P}(S_\mathcal{V}^\dagger > z \mid \mathbf{Y}) \right| \to 0$$

as $p \to \infty$.

Theorem 4 can be extended to multiple cases easily, which is required for constructing the joint confidence regions for several subgraph densities, or their functions such as the clustering coefficient. For given $(\mathcal{V}_1, \tau_{1,1}, \ldots, \tau_{1,k_1}), \ldots, (\mathcal{V}_m, \tau_{m,1}, \ldots, \tau_{m,k_m})$, we approximate the joint distribution of $\sqrt{N}(\widetilde{C}_{\mathcal{V}_1} - C_{\mathcal{V}_1}, \ldots, \widetilde{C}_{\mathcal{V}_m} - C_{\mathcal{V}_m})^\mathrm{T}$ by the following parametric bootstrap procedure:

1: **repeat**
2:     given data $\mathbf{Y} = (Y_{i,j})_{p \times p}$ draw bootstrap samples $\mathbf{Y}^\dagger = (Y_{i,j}^\dagger)_{p \times p}$ as in (23)

3:    calculate bootstrap estimate $\boldsymbol{\vartheta}^{\dagger} = (\vartheta_1^{\dagger}, \ldots, \vartheta_m^{\dagger})^{\mathrm{T}}$, where

$$\vartheta_q^{\dagger} = \frac{\sqrt{N}}{(1-\alpha-\beta)^{k_q}} \sum_{j=1}^{k_q} \frac{(-1)^{1-\tau_{q,j}}}{|\mathcal{V}_q|}$$
$$\sum_{\mathbf{v}=(i_1, i_1', \ldots, i_{k_q}, i_{k_q}') \in \mathcal{V}_q} \left\{ \mathring{Y}_{i_j, i_j'}^{\dagger} \prod_{\ell \neq j} \varphi_{q,\ell}(Y_{i_\ell, i_\ell'}) \right\}$$

for each $q = 1, \ldots, m$ with $\mathring{Y}_{i_j, i_j'}^{\dagger} = Y_{i_j, i_j'}^{\dagger} - Y_{i_j, i_j'} \gamma_1 - \gamma_2$ and $\varphi_{q,\ell}(x) = (x-\alpha)^{\tau_{q,\ell}} (1-\beta-x)^{1-\tau_{q,\ell}}$ for any $x \in \{0, 1\}$

4:  **until** $B$ replicates obtained, for a large integer $B$
5:  **approximate** the joint distribution by the empirical distribution function of $\{\boldsymbol{\vartheta}_1^{\dagger}, \ldots, \boldsymbol{\vartheta}_B^{\dagger}\}$

*Remark 3.* For estimating two-star density, we let $k = 3$, $\tau_1 = \tau_2 = 1$, $\tau_3 = 0$ and $\mathcal{V} = \{(i_1, i_1', i_2, i_2', i_3, i_3') : i_1' = i_2, i_2' = i_3, i_3' = i_1, i_1 \neq i_2 \neq i_3\}$. Then

$$S_{\mathcal{V}} = \frac{(1-\alpha-\beta)^{-1}\sqrt{N}}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} \{\mathring{Y}_{i_1, i_2} A_{i_2, i_3}(1 - A_{i_3, i_1})$$
$$+ \mathring{Y}_{i_2, i_3} A_{i_1, i_2}(1 - A_{i_3, i_1}) - \mathring{Y}_{i_3, i_1} A_{i_1, i_2} A_{i_2, i_3}\}$$

and

$$S_{\mathcal{V}}^{\dagger} = \frac{(1-\alpha-\beta)^{-3}\sqrt{N}}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} \{\mathring{Y}_{i_1, i_2}^{\dagger} (Y_{i_2, i_3} - \alpha)(1 - \beta - Y_{i_3, i_1})$$
$$+ \mathring{Y}_{i_2, i_3}^{\dagger} (Y_{i_1, i_2} - \alpha)(1 - \beta - Y_{i_3, i_1}) - \mathring{Y}_{i_3, i_1}^{\dagger} (Y_{i_1, i_2} - \alpha)(Y_{i_2, i_3} - \alpha)\}.$$

*Remark 4.* For estimating triangle density, we let $k = 3$, $\tau_1 = \tau_2 = \tau_3 = 1$ and $\mathcal{V} = \{(i_1, i_1', i_2, i_2', i_3, i_3') : i_1' = i_2, i_2' = i_3, i_3' = i_1, i_1 \neq i_2 \neq i_3\}$. Then

$$S_{\mathcal{V}} = \frac{(1-\alpha-\beta)^{-1}\sqrt{N}}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} (\mathring{Y}_{i_1, i_2} A_{i_2, i_3} A_{i_3, i_1}$$
$$+ \mathring{Y}_{i_2, i_3} A_{i_1, i_2} A_{i_3, i_1} + \mathring{Y}_{i_3, i_1} A_{i_1, i_2} A_{i_2, i_3})$$

and

$$S_{\mathcal{V}}^{\dagger} = \frac{(1-\alpha-\beta)^{-3}\sqrt{N}}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} \{\mathring{Y}_{i_1, i_2}^{\dagger} (Y_{i_2, i_3} - \alpha)(Y_{i_3, i_1} - \alpha)$$
$$+ \mathring{Y}_{i_2, i_3}^{\dagger} (Y_{i_1, i_2} - \alpha)(Y_{i_3, i_1} - \alpha)$$
$$+ \mathring{Y}_{i_3, i_1}^{\dagger} (Y_{i_1, i_2} - \alpha)(Y_{i_2, i_3} - \alpha)\}.$$

## 4.2. Estimation of Subgraph Densities With Unknown Error Rates

When the error rates $\alpha$ and $\beta$ are unknown, we simply use the estimator $\widetilde{C}_{\mathcal{V}}$ defined in (19) with $\alpha$ and $\beta$ replaced by their estimators derived in Section 3.2. Then its asymptotic properties are more complex, and, consequently, the construction of confidence sets is more involved. Note that we need at most three samples $\mathbf{Y}, \mathbf{Y}_*, \mathbf{Y}_{**}$ for estimating $\alpha$ and $\beta$ in Section 3.2. Obviously an improvement to the approach outlined below can be entertained by combining the three estimators obtained from computing (19), using one of the three available samples each time. For simplicity, we do not pursue this idea further here.

Given estimators $(\tilde{\alpha}, \tilde{\beta})$ for $(\alpha, \beta)$, we define

$$\widehat{C}_{\mathcal{V}} = \frac{\widehat{T}_{\mathcal{V}}}{(1-\tilde{\alpha}-\tilde{\beta})^k} \tag{25}$$

as an estimator for $C_{\mathcal{V}}$, where

$$\widehat{T}_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}}$$
$$\prod_{\ell=1}^{k} (Y_{i_\ell, i_\ell'} - \tilde{\alpha})^{\tau_\ell} (1 - \tilde{\beta} - Y_{i_\ell, i_\ell'})^{1-\tau_\ell}.$$

See also (19). Here we let $(\tilde{\alpha}, \tilde{\beta}) = (\alpha, \hat{\beta})$ for $\hat{\beta}$ defined in (12) if $\alpha$ is known, $(\tilde{\alpha}, \tilde{\beta}) = (\hat{\alpha}, \beta)$ for $\hat{\alpha}$ defined in (13) if $\beta$ is known, and $(\tilde{\alpha}, \tilde{\beta}) = (\hat{\alpha}, \hat{\beta})$ for $(\hat{\alpha}, \hat{\beta})$ defined in Section 3.2.2 if both $\alpha$ and $\beta$ are unknown. Let

$$\Delta_{\alpha, \mathcal{V}} = \frac{kC_{\mathcal{V}}}{1-\alpha-\beta} - \frac{1}{(1-\alpha-\beta)^k}$$
$$\sum_{j:\tau_j=1} \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \prod_{\ell \neq j} \mathbb{E}\{\varphi_\ell(Y_{i_\ell, i_\ell'})\} \tag{26}$$

and

$$\Delta_{\beta, \mathcal{V}} = \frac{kC_{\mathcal{V}}}{1-\alpha-\beta} - \frac{1}{(1-\alpha-\beta)^k}$$
$$\sum_{j:\tau_j=0} \frac{1}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1, i_1', \ldots, i_k, i_k') \in \mathcal{V}} \prod_{\ell \neq j} \mathbb{E}\{\varphi_\ell(Y_{i_\ell, i_\ell'})\} \tag{27}$$

with $\varphi_\ell(\cdot)$ defined as in (18).

*Proposition 4.* Let $N = p(p-1)/2$, $\aleph_{\mathcal{V}}/|\mathcal{V}| \asymp N^{-1}$, $\max\{|\tilde{\alpha} - \alpha|, |\tilde{\beta} - \beta|\} = O_p(N^{-1/2})$ and $|1 - \alpha - \beta| \geq c$ for some positive constant $c$. Under Assumption 1, it holds that $|\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}}| = O_p(N^{-1/2})$. Furthermore, if Assumption 2 also holds, then $\sqrt{N}(\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}}) = S_{\mathcal{V}} + \Delta_{\alpha, \mathcal{V}}\sqrt{N}(\tilde{\alpha} - \alpha) + \Delta_{\beta, \mathcal{V}}\sqrt{N}(\tilde{\beta} - \beta) + o_p(1)$, where $S_{\mathcal{V}}$ is defined as (21).

In comparison to Proposition 3, the leading term of $\sqrt{N}(\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}})$ with unknown $\alpha$ or/and $\beta$ has an additional part

$$\Xi_{\mathcal{V}} := \Delta_{\alpha, \mathcal{V}}\sqrt{N}(\tilde{\alpha} - \alpha) + \Delta_{\beta, \mathcal{V}}\sqrt{N}(\tilde{\beta} - \beta), \tag{28}$$

which is a linear combination of $\sqrt{N}(\tilde{\alpha} - \alpha)$ and $\sqrt{N}(\tilde{\beta} - \beta)$. Since $S_{\mathcal{V}}$ and $\Xi_{\mathcal{V}}$ both converge to normal distributions, $\sqrt{N}(\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}})$ is also asymptotically normal. Let $\kappa_1 = \alpha(1-\alpha)$, $\kappa_2 = \beta(1-\beta)$ and $\kappa_3 = 1 - \alpha - \beta$. Define

$$\mathbf{G} = \begin{pmatrix} g_{\alpha,1} & g_{\alpha,2} & g_{\alpha,3} \\ g_{\beta,1} & g_{\beta,2} & g_{\beta,3} \end{pmatrix}, \tag{29}$$

where $(g_{\alpha,1}, g_{\alpha,2}, g_{\alpha,3}, g_{\beta,1}, g_{\beta,2}, g_{\beta,3})$ are specified as follows.

- If only $\alpha$ is known, $g_{\alpha,1} = g_{\alpha,2} = g_{\alpha,3} = 0$, $g_{\beta,1} = \frac{\kappa_1 - \kappa_2}{\delta\kappa_3^2}$, $g_{\beta,2} = \frac{1}{\delta\kappa_3}$, and $g_{\beta,3} = 0$.
- If only $\beta$ is known, $g_{\alpha,1} = \frac{\kappa_1 - \kappa_2}{(1-\delta)\kappa_3^2}$, $g_{\alpha,2} = \frac{1}{(1-\delta)\kappa_3}$, $g_{\alpha,3} = 0$, and $g_{\beta,1} = g_{\beta,2} = g_{\beta,3} = 0$.

- If both $\alpha$ and $\beta$ are unknown, $g_{\alpha,1} = \frac{(1-2\beta)\alpha+\beta^2}{(1-\delta)\kappa_3^2}$, $g_{\alpha,2} = \frac{\alpha-2\beta}{(1-\delta)\kappa_3^2}$, $g_{\alpha,3} = \frac{1}{(1-\delta)\kappa_3^2}$, $g_{\beta,1} = -\frac{(1-2\alpha)\beta+\alpha^2}{\delta\kappa_3^2}$, $g_{\beta,2} = \frac{\beta-2\alpha+1}{\delta\kappa_3^2}$, and $g_{\beta,3} = -\frac{1}{\delta\kappa_3^2}$.

Let $\kappa_4 = \beta - \alpha$. We define a three-dimensional vector $\mathbf{h}_{\mathcal{V}}$ such that

$$\mathbf{h}_{\mathcal{V}}^{\mathrm{T}} = \left[ 6\kappa_4, 3(\kappa_4^2 - \kappa_1 - \kappa_2), 2\{\kappa_4(-6\alpha\beta + 3\kappa_3^2 - 4\kappa_3) \right.$$
$$\left. + (1-\alpha)(\beta - 2\alpha)\} \right] \times \frac{1}{3} \sum_{j=1}^{k} (-1)^{1-\tau_j}$$
$$C_{\mathcal{V}}(\tau_1, \ldots, \tau_{j-1}, 1, \tau_{j+1}, \ldots, \tau_k)$$
$$+ \left\{ 6\kappa_1, 3\kappa_1(1-2\alpha), 2\kappa_1(1-\alpha)(1-3\alpha) \right\}$$
$$\times \frac{1}{3\kappa_3^k} \sum_{j=1}^{k} \frac{(-1)^{1-\tau_j}}{|\mathcal{V}|} \sum_{\mathbf{v}=(i_1,i_1',\ldots,i_k,i_k')\in\mathcal{V}} \prod_{\ell\neq j} \mathbb{E}\{\varphi_\ell(Y_{i_\ell,i_\ell'})\}.$$
(30)

Now we can state the following theorem.

*Theorem 5.* Let $N = p(p-1)/2$, $\aleph_{\mathcal{V}}/|\mathcal{V}| \asymp N^{-1}$, $|1-\alpha-\beta| \geq c$ for some positive constant $c$, $N_1 = p(p-1)\delta \to \infty$ and $N_2 = p(p-1)(1-\delta) \to \infty$. Under Assumptions 1 and 2, it holds that $\sqrt{N}(\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}}) \to_d \mathcal{N}(0, \phi_{\mathcal{V}}^2)$ with $\phi_{\mathcal{V}}^2 = \sigma_{\mathcal{V}}^2 + (\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}^{\mathrm{T}}(\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})^{\mathrm{T}} + \mathbf{h}_{\mathcal{V}}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}(\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})^{\mathrm{T}}$, where $\sigma_{\mathcal{V}}^2$ and $\boldsymbol{\Sigma}$ are defined as (22) and (A.1) in the Appendix, respectively, provided that one of the following three conditions holds: (i) $\delta(1-\alpha-\beta)^2 \geq c$ for some positive constant $c$ when only $\alpha$ is known, (ii) $(1-\delta)(1-\alpha-\beta)^2 \geq c$ for some positive constant $c$ when only $\beta$ is known, or (iii) $\delta(1-\delta)(1-\alpha-\beta)^4 \geq c$ for some positive constant $c$ when both of $\alpha$ and $\beta$ are unknown.

Recall that $\sqrt{N}(\widehat{C}_{\mathcal{V}} - C_{\mathcal{V}}) = S_{\mathcal{V}} + \Xi_{\mathcal{V}} + o_p(1)$. The asymptotic variance $\phi_{\mathcal{V}}^2$ stated in Theorem 5 actually can be divided into three parts. The first term $\sigma_{\mathcal{V}}^2$ is the asymptotic variance of $S_{\mathcal{V}}$. The second term $(\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}^{\mathrm{T}}(\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})^{\mathrm{T}}$ is the asymptotic variance of $\Xi_{\mathcal{V}}$. The third term $\mathbf{h}_{\mathcal{V}}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}(\Delta_{\alpha,\mathcal{V}}, \Delta_{\beta,\mathcal{V}})^{\mathrm{T}}$ is two times the asymptotic covariance between $S_{\mathcal{V}}$ and $\Xi_{\mathcal{V}}$.

*Remark 5.* By way of comparison with Theorem 5 here, based on Theorem 10 of Balachandran, Kolaczyk, and Viles (2017) and the discussion immediately following that theorem, we can conclude that in general the empirically observed subgraph counts will not even be consistent estimates of $C_{\mathcal{V}}$.

### 4.3. Joint Inference of Subgraph Densities With Unknown Error Rates

Theorem 5 can be extended to the case of multiple subgraph densities, which is required for constructing the joint confidence regions for several subgraph densities or a smooth function thereof. Given $(\mathcal{V}_1, \tau_{1,1}, \ldots, \tau_{1,k_1}), \ldots, (\mathcal{V}_m, \tau_{m,1}, \ldots, \tau_{m,k_m})$, it holds that the random vector $\sqrt{N}(\widehat{C}_{\mathcal{V}_1} - C_{\mathcal{V}_1}, \ldots, \widehat{C}_{\mathcal{V}_m} - C_{\mathcal{V}_m})^{\mathrm{T}}$ converges to a multivariate normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{V})$. Let

$$\boldsymbol{\vartheta} = (S_{\mathcal{V}_1}, \ldots, S_{\mathcal{V}_m})^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{\theta} = (\Xi_{\mathcal{V}_1}, \ldots, \Xi_{\mathcal{V}_m})^{\mathrm{T}},$$

where $S_{\mathcal{V}_q} = S_{\mathcal{V}_q}(\tau_{q,1}, \ldots, \tau_{q,k_q})$ and $\Xi_{\mathcal{V}_q} = \Xi_{\mathcal{V}_q}(\tau_{q,1}, \ldots, \tau_{q,k_q})$ are defined in the same manner as (21) and (28), respectively, but

in which $(\mathcal{V}, \tau_1, \ldots, \tau_k)$ is replaced by $(\mathcal{V}_q, \tau_{q,1}, \ldots, \tau_{q,k_q})$ now. It follows from Proposition 4 that $\mathbf{V} = \lim_{p\to\infty} \mathbf{V}_p$ with

$$\mathbf{V}_p = \underbrace{\mathrm{var}(\boldsymbol{\vartheta})}_{\mathbf{V}_{1,p}} + \underbrace{\mathrm{var}(\boldsymbol{\theta})}_{\mathbf{V}_{2,p}} + \underbrace{\mathrm{cov}(\boldsymbol{\vartheta}, \boldsymbol{\theta}) + \mathrm{cov}(\boldsymbol{\theta}, \boldsymbol{\vartheta})}_{\mathbf{V}_{3,p}}. \quad (31)$$

The first term $\mathbf{V}_{1,p}$ can be consistently estimated by the bootstrap procedure presented in Section 4.1 with $(\alpha, \beta)$ replaced by $(\tilde{\alpha}, \tilde{\beta})$. To evaluate $\mathbf{V}_{2,p}$ and $\mathbf{V}_{3,p}$, we put

$$\boldsymbol{\Delta} = \begin{pmatrix} \Delta_{\alpha,\mathcal{V}_1} & \Delta_{\beta,\mathcal{V}_1} \\ \vdots & \vdots \\ \Delta_{\alpha,\mathcal{V}_m} & \Delta_{\beta,\mathcal{V}_m} \end{pmatrix} \quad \text{and} \quad \mathbf{H} = \begin{pmatrix} \mathbf{h}_{\mathcal{V}_1}^{\mathrm{T}} \\ \vdots \\ \mathbf{h}_{\mathcal{V}_m}^{\mathrm{T}} \end{pmatrix},$$

where $\Delta_{\alpha,\mathcal{V}_q}$, $\Delta_{\beta,\mathcal{V}_q}$ and $\mathbf{h}_{\mathcal{V}_q}^{\mathrm{T}}$ are defined in the same manner as (26), (27), and (30), respectively, with $(\mathcal{V}, \tau_1, \ldots, \tau_k)$ replaced by $(\mathcal{V}_q, \tau_{q,1}, \ldots, \tau_{q,k_q})$ now. Then it holds that

$$\mathbf{V}_{2,p} = \boldsymbol{\Delta}\mathbf{G}\boldsymbol{\Sigma}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}} + o(1) \quad \text{and}$$
$$\mathbf{V}_{3,p} = \frac{1}{2}\left(\mathbf{H}\mathbf{G}^{\mathrm{T}}\boldsymbol{\Delta}^{\mathrm{T}} + \boldsymbol{\Delta}\mathbf{G}\mathbf{H}^{\mathrm{T}}\right) + o(1), \quad (32)$$

where $\mathbf{G}$ and $\boldsymbol{\Sigma}$ are defined as (29) and (A.1) in the Appendix, respectively.

For given $q = 1, \ldots, m$, $\tau_{q,1}, \ldots, \tau_{q,k_q} \in \{0, 1\}$ and $(\tilde{\alpha}, \tilde{\beta})$, define $\tilde{\varphi}_{q,\ell}(x) = (x - \tilde{\alpha})^{\tau_{q,\ell}}(1 - \tilde{\beta} - x)^{1-\tau_{q,\ell}}$ for $x \in \{0, 1\}$. Let $\tilde{\kappa}_1 = \tilde{\alpha}(1 - \tilde{\alpha})$, $\tilde{\kappa}_2 = \tilde{\beta}(1 - \tilde{\beta})$ and $\tilde{\kappa}_3 = 1 - \tilde{\alpha} - \tilde{\beta}$. Since

$$\frac{1}{|\mathcal{V}_q|} \sum_{\mathbf{v}=(i_1,i_1',\ldots,i_{k_q},i_{k_q}')\in\mathcal{V}_q} \prod_{\ell\neq j} \mathbb{E}\{\varphi_{q,\ell}(Y_{i_\ell,i_\ell'})\}$$

can be consistently estimated by

$$\frac{1}{|\mathcal{V}_q|} \sum_{\mathbf{v}=(i_1,i_1',\ldots,i_{k_q},i_{k_q}')\in\mathcal{V}_q} \prod_{\ell\neq j} \tilde{\varphi}_{q,\ell}(Y_{i_\ell,i_\ell'}),$$

then

$$\widehat{\Delta}_{\alpha,\mathcal{V}_q} = \frac{k_q}{\tilde{\kappa}_3}\widehat{C}_{\mathcal{V}_q} - \frac{1}{\tilde{\kappa}_3^{k_q}}$$
$$\sum_{j:\tau_{q,j}=1} \frac{1}{|\mathcal{V}_q|} \sum_{\mathbf{v}=(i_1,i_1',\ldots,i_{k_q},i_{k_q}')\in\mathcal{V}_q} \prod_{\ell\neq j} \tilde{\varphi}_{q,\ell}(Y_{i_\ell,i_\ell'})$$

and

$$\widehat{\Delta}_{\beta,\mathcal{V}_q} = \frac{k_q}{\tilde{\kappa}_3}\widehat{C}_{\mathcal{V}_q} - \frac{1}{\tilde{\kappa}_3^{k_q}}$$
$$\sum_{j:\tau_{q,j}=0} \frac{1}{|\mathcal{V}_q|} \sum_{\mathbf{v}=(i_1,i_1',\ldots,i_{k_q},i_{k_q}')\in\mathcal{V}_q} \prod_{\ell\neq j} \tilde{\varphi}_{q,\ell}(Y_{i_\ell,i_\ell'})$$

are consistent estimates for $\Delta_{\alpha,\mathcal{V}_q}$ and $\Delta_{\beta,\mathcal{V}_q}$, respectively. Replacing $\Delta_{\alpha,\mathcal{V}_q}$, $\Delta_{\beta,\mathcal{V}_q}$ and $(\alpha, \beta)$ by $\widehat{\Delta}_{\alpha,\mathcal{V}_q}$, $\widehat{\Delta}_{\beta,\mathcal{V}_q}$ and $(\tilde{\alpha}, \tilde{\beta})$, respectively, we can obtain consistent estimates of $\boldsymbol{\Delta}$, $\mathbf{G}$, $\mathbf{H}$, and $\boldsymbol{\Sigma}$, and, consequently, consistent estimates of $\mathbf{V}_{2,p}$ and $\mathbf{V}_{3,p}$. For $i = 1, 2, 3$, denote by $\widehat{\mathbf{V}}_{i,p}$ the consistent estimate of $\mathbf{V}_{i,p}$. Then the joint distribution of $\sqrt{N}(\widehat{C}_{\mathcal{V}_1} - C_{\mathcal{V}_1}, \ldots, \widehat{C}_{\mathcal{V}_m} - C_{\mathcal{V}_m})^{\mathrm{T}}$ can be approximated by $\mathcal{N}(\mathbf{0}, \widehat{\mathbf{V}}_p)$ with $\widehat{\mathbf{V}}_p = \widehat{\mathbf{V}}_{1,p} + \widehat{\mathbf{V}}_{2,p} + \widehat{\mathbf{V}}_{3,p}$.

*Remark 6.* For estimating two-star density, we let $k = 3$, $\tau_1 = \tau_2 = 1$, $\tau_3 = 0$ and $\mathcal{V} = \{(i_1, i'_1, i_2, i'_2, i_3, i'_3) : i'_1 = i_2, i'_2 = i_3, i'_3 = i_1, i_1 \neq i_2 \neq i_3\}$. Then $\widehat{C}_\mathcal{V} = \tilde{\kappa}_3^{-3}\{p(p-1)(p-2)\}^{-1} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha})(1 - \tilde{\beta} - Y_{i_3,i_1})$, $\widehat{\Delta}_{\alpha,\mathcal{V}} = 3\tilde{\kappa}_3^{-1}\widehat{C}_\mathcal{V} - 2\tilde{\kappa}_3^{-3}\{p(p-1)(p-2)\}^{-1} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(1 - \tilde{\beta} - Y_{i_3,i_1})$, $\widehat{\Delta}_{\beta,\mathcal{V}} = 3\tilde{\kappa}_3^{-1}\widehat{C}_\mathcal{V} - \tilde{\kappa}_3^{-3}\{p(p-1)(p-2)\}^{-1} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha})$ and

$$
\begin{aligned}
\widehat{\mathbf{h}}_\mathcal{V}^\mathsf{T} = \frac{1}{3} &\left\{ 2\tilde{C}_\mathcal{V} - \frac{\tilde{\kappa}_3^{-3}}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha}) \right. \\
&\left. (Y_{i_3,i_1} - \tilde{\alpha}) \right\} \times \left[ 6\tilde{\kappa}_4, 3\left(\tilde{\kappa}_4^2 - \tilde{\kappa}_1 - \tilde{\kappa}_2\right), 2\left\{ \tilde{\kappa}_4 \left( -6\tilde{\alpha}\tilde{\beta} + 3\tilde{\kappa}_3^2 - 4\tilde{\kappa}_3 \right) \right.\right. \\
&\left.\left. + (1 - \tilde{\alpha})(\tilde{\beta} - 2\tilde{\alpha}) \right\} \right] \right] + \frac{\tilde{\kappa}_3^{-3}}{3p(p-1)(p-2)} \\
&\sum_{i_1 \neq i_2 \neq i_3} \left\{ 2(Y_{i_1,i_2} - \tilde{\alpha})(1 - \tilde{\beta} - Y_{i_3,i_1}) - (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha}) \right\} \\
&\times \{6\tilde{\kappa}_1, 3\tilde{\kappa}_1(1 - 2\tilde{\alpha}), 2\tilde{\kappa}_1(1 - \tilde{\alpha})(1 - 3\tilde{\alpha})\}.
\end{aligned}
$$

*Remark 7.* For estimating triangle density, we let $k = 3$, $\tau_1 = \tau_2 = \tau_3 = 1$ and $\mathcal{V} = \{(i_1, i'_1, i_2, i'_2, i_3, i'_3) : i'_1 = i_2, i'_2 = i_3, i'_3 = i_1, i_1 \neq i_2 \neq i_3\}$. Then $\widehat{C}_\mathcal{V} = \tilde{\kappa}_3^{-3}\{p(p-1)(p-2)\}^{-1} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha})(Y_{i_3,i_1} - \tilde{\alpha})$, $\widehat{\Delta}_{\alpha,\mathcal{V}} = 3\tilde{\kappa}_3^{-1}\widehat{C}_\mathcal{V} - 3\tilde{\kappa}_3^{-3}\{p(p-1)(p-2)\}^{-1} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha})$, $\widehat{\Delta}_{\beta,\mathcal{V}} = 3\tilde{\kappa}_3^{-1}\widehat{C}_\mathcal{V}$ and

$$
\begin{aligned}
\widehat{\mathbf{h}}_\mathcal{V}^\mathsf{T} = &\widehat{C}_\mathcal{V}\left[ 6\tilde{\kappa}_4, 3(\tilde{\kappa}_4^2 - \tilde{\kappa}_1 - \tilde{\kappa}_2), 2\{\tilde{\kappa}_4(-6\tilde{\alpha}\tilde{\beta} \right.\\
&\left. + 3\tilde{\kappa}_3^2 - 4\tilde{\kappa}_3) + (1 - \tilde{\alpha})(\tilde{\beta} - 2\tilde{\alpha})\}\right] \\
&+ \frac{1}{\tilde{\kappa}_3^3}\{6\tilde{\kappa}_1, 3\tilde{\kappa}_1(1 - 2\tilde{\alpha}), 2\tilde{\kappa}_1(1 - \tilde{\alpha})(1 - 3\tilde{\alpha})\} \\
&\times \frac{1}{p(p-1)(p-2)} \sum_{i_1 \neq i_2 \neq i_3} (Y_{i_1,i_2} - \tilde{\alpha})(Y_{i_2,i_3} - \tilde{\alpha}).
\end{aligned}
$$

## 5. Numerical Illustration

### 5.1. Simulations

We conduct some simulations to illustrate the finite sample properties of the proposed estimation methods. For given $\delta \in (0, 1)$ and integers $p$, $N_{2*}$, and $N_\triangle$, we specify a $p \times p$ deterministic adjacency matrix $\mathbf{A}$ with $\lfloor \delta p(p-1)/2 \rfloor$ edges randomly allocated among vertex pairs subject to the condition that there are exactly $N_{2*}$ two-stars (also called triplets), and $N_\triangle$ triangles. Hence the clustering coefficient of the corresponding network is

$$\gamma = 3N_\triangle/N_{2*}. \tag{33}$$

Generating such $\mathbf{A}$ is accomplished by an adaptation of the rewiring ideas of Mahadevan et al. (2006), which, to our best knowledge, is new. Note that $\delta$ is the edge density, $2N_{2*}/\{p(p-1)(p-2)\}$ and $6N_\triangle/\{p(p-1)(p-2)\}$ are, respectively, the two-star density and the triangle density. We set $\alpha = 0.05$, $\beta = 0.05$ or $0.20$, $p = 30, 50, 100$, and $200$. We assume that both $\alpha$ and $\beta$ are unknown. Therefore, we need 3 noisy observations $\mathbf{Y}, \mathbf{Y}_*, \mathbf{Y}_{**}$ to facilitate the estimation, which are generated according to (2).

We evaluate the point estimates for $\delta, \alpha$, and $\beta$ iteratively using (12) and (17). More precisely we set an initial value $\alpha_0 = 0.2$, and obtain $\hat{\beta}$ and $\hat{\delta}$ from (12). Plugging $(\alpha, \hat{\beta}, \hat{\delta})$ into the right-hand side of (17), we obtain $\hat{\alpha}$. We repeat this exercise by setting $\alpha = \hat{\alpha}$, and terminate the recursion when the absolute difference of two successive values of $\alpha$ is smaller than $10^{-4}$. We also calculate the approximate confidence intervals for $\delta$ based on the asymptotic normality stated in Theorem 3. More precisely, the confidence interval is in the same form as (14) with the asymptotic variance determined by (A.4) in the Appendix in which $\alpha, \beta, \delta$ are replaced by their respective estimates.

Having obtained estimates $\hat{\alpha}$ and $\hat{\beta}$, the point estimates for the densities of two-star edges and triangles are $\widehat{C}_\mathcal{V}$ defined in (25); see also Remarks 6 and 7. Then a plug-in estimate for clustering coefficient is obtained based on (33). To compute their confidence intervals is more involved, and is based on the procedure described in Section 4.3. More precisely, we calculate the joint asymptotic distribution of the normalized estimators for two-star edge density and triangle density, which is a two-dimensional normal distribution with zero mean and variance-covariance matrix $\mathbf{V}_p = \mathbf{V}_{1,p} + \mathbf{V}_{2,p} + \mathbf{V}_{3,p}$, as given in the form (31). Note that $\mathbf{V}_{2,p}$, and $\mathbf{V}_{3,p}$ can be calculated directly; see (32) and also Remarks 6 and 7. To calculate $\mathbf{V}_{1,p}$, we have to apply the bootstrap algorithm presented in Section 4.1 with $\alpha = \hat{\alpha}$ and $\beta = \hat{\beta}$; see also Remarks 3 and 4. We replicate bootstrap sampling 500 times. Then a 95% confidence interval is $\widehat{C}_\mathcal{V} \pm 1.96s$, where $s$ is the square-root of, respectively, the (1,1)-element or the (2,2)-element of $2\mathbf{V}_p/\{p(p-1)\}$ for two-star density or triangle density. Consequently, a confidence interval for clustering coefficient is deduced based on (33).

To assess the performance of the estimation procedure, we replicate the simulation 500 times for each setting. The results are reported in Tables 1 and 2. As the densities for two-stars and triangles are very small (i.e., smaller than $10^{-2}$), we report the estimates for the counts $N_{2*}$ and $N_\triangle$ instead. The mean absolute errors (MAE) for the point estimates for the error rates $\alpha, \beta$, the edge density $\delta$, the two-star count $N_{2*}$, the triangle count $N_\triangle$, and the clustering coefficient $\gamma$ are reported in Table 1. For example, $\mathrm{MAE}(\hat{\delta}) = \frac{1}{500}\sum_{i=1}^{500}|\delta_i - \delta|$, where $\delta_1, \ldots, \delta_{500}$ denote the estimated values in the 500 replications of simulation, and $\delta$ denotes the true value. When $p$ increases, the estimation errors for $\alpha, \beta, \delta$ and $\gamma$ decrease. Furthermore the errors with $\beta = 0.2$ are always greater than those with $\beta = 0.05$. This is due to greater (Type II) errors occurring in the observations $Y_{i,j}$. The estimation for the edge density $\delta$ is very accurate, and is more accurate than that for the clustering coefficient $\gamma$ which is a higher-order quantity, though $\gamma$ can be estimated accurately too especially when $p \geq 100$. Also noticeable are greater errors in estimating $\beta$ than those in estimating $\alpha$. For sparser networks (such as $\delta = 0.1$ or $0.2$), there are a comparatively smaller number of $A_{i,j}$ taking value 1, and, hence, the information on $\beta$ is less. Note that the estimation for $\beta$ improves when $\delta$ increases from 0.1 to 0.2. The MAE for the two-star count and the triangle count depend on the magnitudes of the counts themselves. Note that the relative MAE (i.e., $\mathrm{MAE}(\widehat{N}_{2*})/N_{2*}$ or $\mathrm{MAE}(\widehat{N}_\triangle)/N_\triangle$) are small or very small. Indeed they decrease too when $p$ increases.

The estimated 95% confidence intervals for $\delta, N_{2*}, N_\triangle$, and $\gamma$ are reported in Table 2. The estimated coverage probabilities

**Table 1.** Mean absolute errors (MAE) of the point estimates for error rates $\alpha$, $\beta$, edge density $\delta$, two-star count $N_{2*}$, triangle count $N_\triangle$, and clustering coefficient $\gamma$ in the simulation with 500 replications for noisy network with $p$ nodes, and $\alpha = 0.05$.

| $p$ | $\beta$ | $\delta$ | $N_{2*}$ | $N_\triangle$ | $\gamma$ | MAE($\hat{\alpha}$) | MAE($\hat{\beta}$) | MAE($\hat{\delta}$) | MAE($\widehat{N}_{2*}$) | MAE($\widehat{N}_\triangle$) | MAE($\hat{\gamma}$) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 0.05 | 0.1 | 100 | 15 | 0.4500 | 0.0057 | 0.0369 | 0.0103 | 24.61 | 3.450 | 0.1079 |
|  | 0.20 |  |  |  |  | 0.0064 | 0.0622 | 0.0150 | 34.66 | 5.781 | 0.1897 |
| 30 | 0.05 | 0.2 | 430 | 40 | 0.2791 | 0.0058 | 0.0228 | 0.0103 | 48.54 | 6.248 | 0.0279 |
|  | 0.20 |  |  |  |  | 0.0072 | 0.0385 | 0.0162 | 74.44 | 11.25 | 0.0538 |
| 50 | 0.05 | 0.1 | 1260 | 50 | 0.1190 | 0.0034 | 0.0243 | 0.0058 | 105.0 | 11.45 | 0.0204 |
|  | 0.20 |  |  |  |  | 0.0037 | 0.0397 | 0.0086 | 170.3 | 17.57 | 0.0334 |
| 50 | 0.05 | 0.2 | 2300 | 140 | 0.1826 | 0.0037 | 0.0138 | 0.0061 | 145.5 | 16.54 | 0.0132 |
|  | 0.20 |  |  |  |  | 0.0048 | 0.0275 | 0.0111 | 255.4 | 27.98 | 0.0230 |
| 100 | 0.05 | 0.1 | 5000 | 150 | 0.0900 | 0.0017 | 0.0125 | 0.0030 | 299.1 | 22.84 | 0.0107 |
|  | 0.20 |  |  |  |  | 0.0020 | 0.0237 | 0.0048 | 481.3 | 35.45 | 0.0170 |
| 100 | 0.05 | 0.2 | 22,000 | 1800 | 0.2455 | 0.0019 | 0.0071 | 0.0031 | 630.7 | 82.42 | 0.0054 |
|  | 0.20 |  |  |  |  | 0.0024 | 0.0157 | 0.0058 | 1199 | 154.2 | 0.0096 |
| 200 | 0.05 | 0.1 | 40,000 | 1500 | 0.1125 | 0.0008 | 0.0065 | 0.0016 | 1235 | 82.09 | 0.0039 |
|  | 0.20 |  |  |  |  | 0.0010 | 0.0126 | 0.0027 | 2179 | 137.1 | 0.0063 |
| 200 | 0.05 | 0.2 | 155,000 | 10,000 | 0.1935 | 0.0008 | 0.0036 | 0.0016 | 2444 | 258.1 | 0.0023 |
|  | 0.20 |  |  |  |  | 0.0012 | 0.0078 | 0.0027 | 4249 | 431.4 | 0.0036 |

**Table 2.** The 95% confidence intervals for edge density $\delta$, two-star count $N_{2*}$, triangle count $N_\triangle$, and clustering coefficient $\gamma$ in the simulation with 500 replications for noisy networks with $p$ nodes, and $\alpha = 0.05$.

| | | | True value | | | $\delta$ | | $N_{2*}$ | | $N_\triangle$ | | $\gamma$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $p$ | $\beta$ | $\delta$ | $N_{2*}$ | $N_\triangle$ | $\gamma$ | RF | Length | RF | Length | RF | Length | RF | Length |
| 30 | 0.05 | 0.1 | 100 | 15 | 0.4500 | 0.950 | 0.0520 | 0.950 | 130.5 | 0.978 | 20.92 | 0.982 | 0.6316 |
|  | 0.20 |  |  |  |  | 0.938 | 0.0602 | 0.899 | 146.0 | 0.939 | 26.10 | 0.986 | 0.9709 |
| 30 | 0.05 | 0.2 | 430 | 40 | 0.2791 | 0.954 | 0.0496 | 0.950 | 239.1 | 0.960 | 33.60 | 0.982 | 0.1633 |
|  | 0.20 |  |  |  |  | 0.929 | 0.0747 | 0.920 | 349.9 | 0.941 | 52.57 | 0.990 | 0.2582 |
| 50 | 0.05 | 0.1 | 1260 | 50 | 0.1190 | 0.952 | 0.0301 | 0.956 | 544.1 | 0.964 | 62.04 | 0.966 | 0.1144 |
|  | 0.20 |  |  |  |  | 0.950 | 0.0396 | 0.946 | 765.7 | 0.947 | 82.04 | 0.990 | 0.1519 |
| 50 | 0.05 | 0.2 | 2300 | 140 | 0.1826 | 0.942 | 0.0295 | 0.946 | 705.8 | 0.966 | 88.06 | 0.976 | 0.0770 |
|  | 0.20 |  |  |  |  | 0.950 | 0.0530 | 0.940 | 1256 | 0.955 | 152.5 | 0.991 | 0.1313 |
| 100 | 0.05 | 0.1 | 5000 | 150 | 0.0900 | 0.960 | 0.0150 | 0.966 | 1521 | 0.970 | 129.1 | 0.972 | 0.0637 |
|  | 0.20 |  |  |  |  | 0.954 | 0.0253 | 0.954 | 2571 | 0.990 | 216.2 | 0.999 | 0.1070 |
| 100 | 0.05 | 0.2 | 22,000 | 1800 | 0.2455 | 0.954 | 0.0145 | 0.954 | 3011 | 0.956 | 404.1 | 0.968 | 0.0281 |
|  | 0.20 |  |  |  |  | 0.948 | 0.0288 | 0.950 | 6081 | 0.956 | 808.0 | 0.978 | 0.0541 |
| 200 | 0.05 | 0.1 | 40,000 | 1500 | 0.1125 | 0.948 | 0.0074 | 0.948 | 6014 | 0.958 | 435.1 | 0.968 | 0.0228 |
|  | 0.20 |  |  |  |  | 0.944 | 0.0131 | 0.940 | 10,559 | 0.960 | 768.7 | 0.986 | 0.0399 |
| 200 | 0.05 | 0.2 | 155,000 | 10,000 | 0.1935 | 0.942 | 0.0072 | 0.940 | 11,399 | 0.938 | 1197 | 0.964 | 0.0111 |
|  | 0.20 |  |  |  |  | 0.970 | 0.0142 | 0.972 | 22,323 | 0.966 | 2329 | 0.970 | 0.0211 |

NOTE: Reported in the table are the relative frequencies (RF) of the event that a confidence interval covers the corresponding true value, and also the average length of the intervals.

are indeed around 95%. The interval estimation for the edge density $\delta$ is accurate as the average interval lengths are small, varying from 0.0602 when $p = 30$ to 0.0072 when $p = 200$. Note that the true value of $\delta$ is either 0.1 or 0.2. The confidence intervals for the clustering coefficient $\gamma$ tend to be conservative with the coverage probabilities ranging from 96.4% to 99.9%. Nevertheless, the average interval lengths are also small, especially for large $p$. For example, when $p = 200$ and $\gamma = 0.1935$, the average interval length is 0.0111 when $\beta = 0.05$, or 0.0211 when $\beta = 0.2$.

### 5.2. Application: Gene Expression Networks

It is a standard exercise in computational biology to construct and analyze networks from gene expression data. For the purpose of illustration, we consider the data and network construction described in Section 7.3.1 of Kolaczyk and Csárdi (2014). These data, originally published by Faith et al. (2007), contain (log) gene expression levels in the bacteria *Escherichia coli* (*E. coli*), measured for 153 genes under each of 40 different experimental conditions, with three replicates of each condition. For each set of replicates, we constructed a network among the 153

genes by applying a threshold to the Fisher transformation of the Pearson correlation coefficients calculated for the expression levels between all pairs of genes. A Bonferonni correction was used to adjust for multiple testing, with the family-wise error rate controlled at the 0.05 level. While there are numerous other approaches to construction of gene coexpression networks, this simple method is both immediately amenable to our illustration and not uncommon in practice.

The empirical edge density in each of the three resulting networks is quite stable, that is, approximately 0.073, 0.075, and 0.074, respectively. With $153 \times 152/2 = 11{,}628$ hypothesis tests, the nominal value of $\alpha$ in this analysis is at most $4.3 \times 10^{-6}$. Taking this value as known, and calculating the estimates in (12) for two of the networks, we obtain $\hat{\beta} = 0.456$ and $\hat{\delta} = 0.135$. The corresponding approximate 95% confidence interval for $\delta$ is $(0.131, 0.139)$. Similar results are obtained for the other possible pairings of the three networks. These numbers suggest that the true edge density $\delta$ differs substantially from those observed empirically. However, it is well known that the nominal Type I error rates in this setting can be quite inaccurate (e.g., Cosgrove, Gardner, and Kolaczyk 2010). If we instead treat $\alpha$ as unknown, the estimators defined by (9), (11), and (16) yield estimates $\hat{\alpha} = 0.024$, $\hat{\beta} = 0.232$, and $\hat{\delta} = 0.067$. These numbers suggest

that the Type I error rate is orders of magnitude higher than nominally expected, and furthermore that the Type II error rate is nearly one in four. On the other hand, the resulting method-of-moments estimate of the edge density $\delta$ suggests that the empirical edge densities observed in our networks over-estimate only slightly.

However, consider now estimation of higher-order quantities—specifically, of the number of two-stars $N_{2*}$, the number of triangles $N_\triangle$, and the clustering coefficient $\gamma$. For the three networks, the empirical values of these numbers are, respectively, 19,112, 3373, and 0.53 for the first network, 22,952, 4814, and 0.63 for the second network, and 21,820, 4349, and 0.60 for the third network. Thus, we see substantially more variability in these numbers across networks than we did for the empirical edge density. Applying our proposed method-of-moments estimators to these data, we obtain estimates of approximately $\widehat{N}_{2*} = 25{,}248$, $\widehat{N}_\triangle = 7243$, and $\hat{\gamma} = 0.86$. These are all substantially higher than their empirical counterparts, indicating a nontrivial upward adjustment for network noise, presumably driven in large part by the high estimated rate of Type II error.

Finally, applying our bootstrap-based methodology for construction of asymptotic confidence intervals, we obtain an approximate 95% confidence interval for $\delta$ of $(0.06, 0.074)$, which further reinforces the evidence that the true network edge density is less than that observed empirically. At the same time, the corresponding confidence interval for the clustering coefficient $\gamma$ is $(0.81, 0.91)$, suggesting that the true network clustering coefficient is roughly 1/3 larger than observed empirically. Furthermore, the confidence intervals for $N_{2*}$ and $N_\triangle$ are $(21{,}580, 28{,}915)$ and $(5879, 8607)$, respectively, by which we see that the triangle count appears to be more adversely affected by noise than the two-star count.

Ultimately, we see that the ability to account for network noise appropriately in reporting these basic summary statistics can lead to distinctly different numbers and conclusions. From a biological perspective, the fact that the empirically observed edge density is inferred to be fairly accurate, while the clustering coefficient is inferred to be noticeably larger than observed empirically, is suggestive. Specifically, increasing clustering coefficient has been found to trend with increasing modularity in a variety of biological networks (Ravasz et al. 2002; Pavlopoulos et al. 2011). Modules (i.e., groups of highly connected nodes) in gene co-expression networks are understood to be reflective of groups of genes that cooperate in common biological functions. Our results suggest that the presence of modularity in gene co-expression—and, hence, the level of functional cooperation among genes—may well be even more pronounced than currently believed.

## 6. Discussion

Here we have developed a general framework for estimation and uncertainty quantification of arbitrary subgraph densities in contexts wherein one has observations of noisy networks. Our approach requires as few as two or three replicates of network observations, and employs method-of-moments techniques to derive estimators and establish their asymptotic consistency and

normality. Simulations demonstrate that substantial inferential accuracy is possible in networks of even modest size when nontrivial noise is present. And our application to coexpression networks in the context of computational biology shows that the gains offered by our approach over presenting traditional empirical network summaries can be substantial.

The approach we develop here is relevant and broadly applicable to numerous contexts wherein it is possible to obtain some notion of a handful of network replicates. For example, multiple observations of networks are encountered in genetics (e.g., Bartlett, Olhede, and Zaikin 2014), computational neuroscience (e.g., Biswal, Menness, and Zuo 2010), online social media (e.g., Mukherjee, Sarkar, and Lin 2017), and in the study of psychiatric disorders (e.g., Nelson et al. 2017). Similarly, we note that most papers on dynamic networks assume that the networks observed over different times are (conditionally) independent of each other as the connection probabilities evolve over time. As a result, for connection probabilities that do not evolve too quickly, our results are directly applicable within small windows of time (i.e., in light of Remark 2, following Theorem 1). See Pensky (2019) and Zhao, Chen, and Lin (2019), and the references therein, for a variety of examples of relevant dynamic networks.

Our development here is general and supported by formal theoretical arguments. In practice, other approaches have been utilized to date for uncertainty quantification in certain specific contexts, albeit—to our best knowledge—without the formal justification developed here. For example, in the context of gene expression measurements (as in the application described in Section 5.2), investigators will sometimes use bootstrapping of the original experiments to resample many pseudo-data sets and construct many networks, from which in turn they generate bootstrap distributions of network summaries of interest (e.g., Xulvi-Brunet and Li 2009).

We have pursued a frequentist approach to the problem of uncertainty quantification for network summary statistics. If the replicates necessary for our approach are unavailable in a given setting, a Bayesian approach is a natural alternative. For example, posterior-predictive checks for goodness of fit based on examination of a handful of network summary measures is common practice (e.g., Bloem-Reddy and Orbanz 2018, sec 5.3). Note, however, that the Bayesian approach requires careful modeling of the generative process underlying $G$ and typically does not distinguish between signal and noise components. Our analysis is conditional on $G$, and hence does not require that $G$ be modeled. It is effectively a "signal plus noise" model, with the signal taken to be fixed but unknown. Related and more formal work has been done in the context of graphon modeling, with the goal of estimating network motif frequencies (e.g., Latouche and Robin 2016). However, again, one typically does not distinguish between signal and noise components in this setting. Additionally, we note that the problem of practical graphon estimation itself is still a developing area of research.

Our work here sets the stage for extensions of various levels of difficulty. For example, while we have focused here on the case of undirected graphs, the extension to directed graphs is straightforward. For directed graphs, $A_{i,j} \neq A_{j,i}$ and $Y_{i,j} \neq Y_{j,i}$. The representation (2) relies on independent $\varepsilon_{i,j}$ for $1 \leq i \neq j \leq p$. The statistics used for estimation should be changed accordingly

too. For example in (6) the sum should be taken for all $i \neq j$ instead of $i < j$, and the sum should be divided by $p(p-1)$ instead of $p(p-1)/2$. Then the current technical proofs for undirected graphs are applicable identically to directed graphs. On the other hand, whereas we have focused on estimation solely in the case of subgraph densities, which rests on the behavior of counting statistics, we anticipate that the estimation of non-counting network summaries (e.g., summaries based on shortest path lengths) from noisy network data is likely nontrivial, due to the fact that the latter are based on extremes rather than counts.

## Appendix

Here, we derive expressions for the covariance matrices in Theorem 2. Let $\kappa_1 = \alpha(1-\alpha)$, $\kappa_2 = \beta(1-\beta)$ and $\kappa_3 = 1 - \alpha - \beta$. Let

$$\mathbf{W}_\alpha = \begin{pmatrix} \frac{\kappa_2-\kappa_1}{\delta\kappa_3^2} & -\frac{1}{\delta\kappa_3} \\ \frac{2\beta-1}{\kappa_3^2} & -\frac{1}{\kappa_3^2} \end{pmatrix}, \quad \mathbf{W}_\beta = \begin{pmatrix} \frac{\kappa_2-\kappa_1}{(1-\delta)\kappa_3^2} & -\frac{1}{(1-\delta)\kappa_3} \\ \frac{2\alpha-1}{\kappa_3^2} & \frac{1}{\kappa_3^2} \end{pmatrix}$$

and

$$\mathbf{W} = \begin{pmatrix} \frac{(1-2\beta)\alpha+\beta^2}{(1-\delta)\kappa_3^2} & \frac{\alpha-2\beta}{(1-\delta)\kappa_3^2} & \frac{1}{(1-\delta)\kappa_3^2} \\ -\frac{(1-2\alpha)\beta+\alpha^2}{\delta\kappa_3^2} & \frac{\beta-2\alpha+1}{\delta\kappa_3^2} & -\frac{1}{\delta\kappa_3^2} \\ \frac{3\kappa_3+6\alpha\beta-2}{\kappa_3^3} & \frac{3\kappa_3+6\beta-2}{\kappa_3^3} & -\frac{2}{\kappa_3^3} \end{pmatrix}.$$

Define a matrix

$$\mathbf{\Sigma} = (\sigma_{ij})_{3\times 3} \tag{A.1}$$

with $\sigma_{11} = \delta\kappa_2 + (1-\delta)\kappa_1$, $\sigma_{22} = \delta\kappa_2(1/2 - \kappa_2) + (1-\delta)\kappa_1(1/2 - \kappa_1)$, $\sigma_{33} = \delta\beta\kappa_2(1/3 - \beta\kappa_2) + (1-\delta)\kappa_1(1-\alpha)\{1/3 - \kappa_1(1-\alpha)\}$, $\sigma_{12} = \sigma_{21} = \delta\kappa_2(\beta - 1/2) + (1-\delta)\kappa_1(1/2 - \alpha)$, $\sigma_{13} = \sigma_{31} = \delta\kappa_2(\beta^2/3 - 2\kappa_2/3) + (1-\delta)\kappa_1\{(1-\alpha)^2/3 - 2\kappa_1/3\}$, and $\sigma_{23} = \sigma_{32} = \delta\beta\kappa_2(1/3 - \kappa_2) + (1-\delta)(1-\alpha)\kappa_1(1/3 - \kappa_1)$. Denote by $\mathbf{\Sigma}_1 = (\sigma_{ij})_{2\times 2}$ the $2 \times 2$ submatrix of $\mathbf{\Sigma}$. Based on such defined $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_1$, let

$$\mathbf{\Sigma}_{1,\alpha} = \mathbf{W}_\alpha \mathbf{\Sigma}_1 \mathbf{W}_\alpha^{\mathrm{T}}, \tag{A.2}$$

$$\mathbf{\Sigma}_{1,\beta} = \mathbf{W}_\beta \mathbf{\Sigma}_1 \mathbf{W}_\beta^{\mathrm{T}}, \tag{A.3}$$

and

$$\mathbf{\Sigma}_2 = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^{\mathrm{T}}. \tag{A.4}$$

## Supplementary Materials

The on-line supplementary material contains all the technical proofs for the theoretical results in this paper.

## Acknowledgments

The authors are grateful to the editor, an associate editor, and two referees for their helpful suggestions.

## Funding

## References

Almquist, Z. (2012), "Random Errors in Egocentric Networks," *Social Networks*, 34, 493–505. [361]

Balachandran, P., Kolaczyk, E., and Viles, W. (2017), "On the Propagation of Low-Rate Measurement Error to Subgraph Counts in Large Networks," *Journal of Machine Learning Research*, 18, 1–33. [361,362,363,369]

Bartlett, E. T., Olhede, S. C., and Zaikin, A. (2014), "A DNA Methylation Network Interaction Measure and Detection of Network Oncomarkers," *PLoS One*, 9, e84573. [372]

Biswal, B. B., Menness, M., and Zuo, X.-N. (2010), "Toward Discovering Science of Human Brain Function," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 4734–4739. [372]

Bloem-Reddy, B., and Orbanz, P. (2018), "Random Walk Models of Network Formation and Sequential Monte Carlo Methods for Graphs," *Journal of the Royal Statistical Society*, Series B, 80, 871–898. [372]

Chatterjee, S. (2015), "Matrix Estimation by Universal Singular Value Thresholding," *The Annals of Statistics*, 43, 177–214. [361]

Cosgrove, E. J., Gardner, T. S., and Kolaczyk, E. (2010), "On the Choice and Number of Microarrays for Transcriptional Regulatory Network Inference," *BMC Bioinformatics*, 11, 454. [371]

Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J., and Gardner, T. (2007), "Large-Scale Mapping and Validation of *Escherichia coli* Transcriptional Regulation From a Compendium of Expression Profiles," *PLoS Biology*, 5, e8. [371]

Gan, H. L., and Kolaczyk, E. (2018), "Approximation of the Difference of Two Poisson-Like Counts by Skellam," *Journal of Applied Probability*, 55, 416–430. [361]

Hart, G., Ramani, A., and Marcotte, E. (2006), "How Complete Are Current Yeast and Human Protein-Interaction Networks?," *Genome Biology*, 7, 120. [361]

Jiang, X., Gold, D., and Kolaczyk, E. (2011), "Network-Based Auto-Probit Modeling for Protein Function Prediction," *Biometrics*, 67, 958–966. [361]

Jiang, X., and Kolaczyk, E. (2012), "A Latent Eigenprobit Model With Link Uncertainty for Prediction of Protein-Protein Interactions," *Statistics in Biosciences*, 4, 84–104. [361]

Luke, D. A. (2015), *A User's Guide to Network Analysis in R*, London: Springer. [362]

Kolaczyk, E. (2009), *Statistical Analysis of Network Data: Methods and Models*, New York: Springer-Verlag. [361]

Kolaczyk, E., and Csárdi, G. (2014), *Statistical Analysis of Network Data With R*, New York: Springer. [371]

Latouche, P., and Robin, S. (2016), "Variational Bayes Model Averaging for Graphon Functions and Motif Frequencies Inference in W-Graph Models," *Statistics and Computing*, 26, 1173–1185. [372]

Mahadevan, P., Krioukov, C., Fall, K., and Vahdat, A. (2006), "Systematic Topology Analysis and Generation Using Degree Correlations," in *SIGCOMM'06 Proceedings of the 2006 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications*, pp. 135–146. [370]

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. [362]

Mukherjee, S. S., Sarkar, P., and Lin, L. (2017), "On Clustering Network-Valued Data," in *Advances in Neural Information Processing Systems (NIPS)* (Vol. 30), pp. 7071–7081. [372]

Nelson, N. G., Bassett, J., CamBass, D. S., Bullmore, E. T., and Lim, K. O. (2017), "Comparison of Largescale Human Brain Functional and Anatomical Networks in Schizophrenia," *NeuroImage:Clinical*, 15, 439–448. [372]

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R., and Bagos, P. G. (2011), "Using Graph Theory to Analyze Biological Networks," *BioData Mining*, 4, 10. [372]

Pensky, M. (2019), "Dynamic Network Models and Graphon Estimation," *The Annals of Statistics*, 47, 2378–2403. [372]

Priebe, C., Sussman, D., Tang, M., and Vogelstein, J. (2015), "Statistical Inference on Errorfully Observed Graphs," *Journal of Computational and Graphical Statistics*, 24, 930–953. [361]

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A. L. (2002), "Hierarchical Organization of Modularity in Metabolic Networks," *Science*, 297, 1551–1555. [372]

Sporns, O., and Betzel, R. F. (2016), "Modular Brain Networks," *Annual Review of Psychology*, 67, 613–640. [362]

Stone, L., Simberloff, D., and Artzy-Randrup, Y. (2019), "Network Motifs and Their Origins," *PLoS Computational Biology*, 15, e1006749. [362]

Turán, P. (1941), "On an Extremal Problem in Graph Theory," *Matematikai és Fizikai Lapok*, 48, 436–452. [362]

Wasserman, S., and Faust, K. (1994), *Social Network Analysis: Methods and Applications*, Cambridge: Cambridge University Press. [362]

Xulvi-Brunet, R., and Li, H. (2009), "Co-Expression Networks: Graph Properties and Topological Comparisons," *Bioinformatics*, 26, 205–214. [372]

Zhao, Z., Chen, L., and Lin, L. (2019), "Change-Point Detection in Dynamic Networks via Graphon Estimation," arXiv no. 1908.01823. [372]