HONORS THESIS

Hidden Erdös-Rényi Graphs in Stochastic Block Model Networks

Wenhao Wang

Department of Applied Mathematics

University of Colorado - Boulder

Defended on April 1 2022

Committee Members

Thesis Advisor: Manuel E. Lladser, Department of Applied Mathematics (APPM)

Anne Dougherty, College of Arts and Sciences Honors Program & APPM

Katharina Kann, Department of Computer Science (CSCI)

Hidden Erdös-Rényi Graphs in Stochastic Block Model Networks

Wenhao Wang

Abstract

In this thesis, we use a new probabilistic and statistical methodology to estimate the fraction of spurious edges in a Stochastic Block model (SBM). By spurious we mean unstructured connections between nodes in the network, which we attribute as coming from an Erdös-Rényi random graph model. We define the "hidden weight" of a SBM as the largest expected fraction of placed or missing edges that can be attributed to an Erdös-Rényi model. We characterize explicitly and asymptotically this weight in terms of the parameters defining the SBM, and test our results with simulations.

Table of Contents

1 Introduction

Detecting clusters in a graph is a fundamental problem in physics, biology and computer science because they can reveal underlying structures in complex networks [1]. Cluster detection aims to group similar objects (in a network) from objects (in the same network) with different characteristics. The grand challenge in cluster detection is to infer those groups without knowing the characteristics that define them.

Cluster detection can be used in some real networks, such as social networks [2], improving recommendation systems [3] and biological networks [4]. It is also crucial in image analysis [5] and bioinformatics [6]. In this thesis, we use the Erdös-Rényi model (ER) and Stochastic Block model (SBM) to better understand the underlying structure of random networks.

The Erdös-Rényi (ER) model was first introduced by the Hungarian mathematicians Paul Erdős and Alfréd Rényi in 1959 [7]. The ER model is a basic probability model for network analysis where all edges have the same probability of occurrence, independently from one another [8].

Stochastic Block models were introduced by Holland et al in 1983 [9]. They are popular and effective for modeling hidden clusters in large and complex networks. In general, a Stochastic Block model $\text{SBM}_{p,q}(n)$ indicates that each vertex has probability p to be connected to any node within the same cluster, but has probability q to be connected with any node in a different cluster. In the past few years, various works have developed algorithms and methods to detect or recover clusters in networks under different type of models [10]. In 2011, Karrer and Newman worked with undirected multi-graphs and first termed *Degree-corrected blockmodel* [11]. In 2017, Pexioto introduced the *Microcanonical stochastic block model* and its nested version [12]. In most models, each node is attributed to a single community; however, in real-world networks, a node may belong to two or more groups. *Mixed membership stochastic block models* (MMSBM) by Airoldi et al. [13] extended blockmodels to allow nodes to belong to multiple clusters.

One of the chief motivations of this thesis is to estimate the expected fraction of spurious edges in a SBM. For this, in Section 2, we introduce the notion of "hidden weight" of a SBM with respect to an Erdös-Rényi model. This notion may be regarded a specialization of the concept of latent weights [14, 15] to random graphs. In Section 3, we discuss the approximate distribution of hidden ER-weights under different configurations of the SBM. The key tools in this section are the Delta method [16] and order statistics [17]. Conclusions and future work are discussed in Section 4.

2 Hidden ER-weight of a Stochastic Block Model

2.1 Basic Models

All random network models in this manuscript are *undirected.*

The general Stochastic Block model, SBM(n,W), is a random graph defined as follows:

- n is the number of vertices;
- W is a symmetric matrix of dimensions $k \times k$. Here, k is the number of clusters in the model, which we label with elements in the set $C := \{1, \dots, k\}$. We denote the set of vertices in cluster i as V_i; in particular, $\sum_{i \in C} |V_i| = n$. Here and in what follows, $|\cdot|$ denotes the *size* (i.e. *cardinality*) of the set within the parentheses.
- Each entry w_{ij} in W denotes the probability of connection between a node in cluster i and a node in cluster j.

In some applications, only the size of each cluster is specified by a probability vector α of dimension k; in particular, $\alpha_i \geq 0$ for each $i \in C$, and $\sum_{i \in C} \alpha_i = 1$. In this case, $|V_i| = \alpha_i \cdot n$. Other applications, however, interpret α as a probability distribution that assigns each node to V_i with probability α_i. In this case, V_i and its size is random. In fact, $|V_i| \sim \text{Binomial}(n, \alpha_i)$.

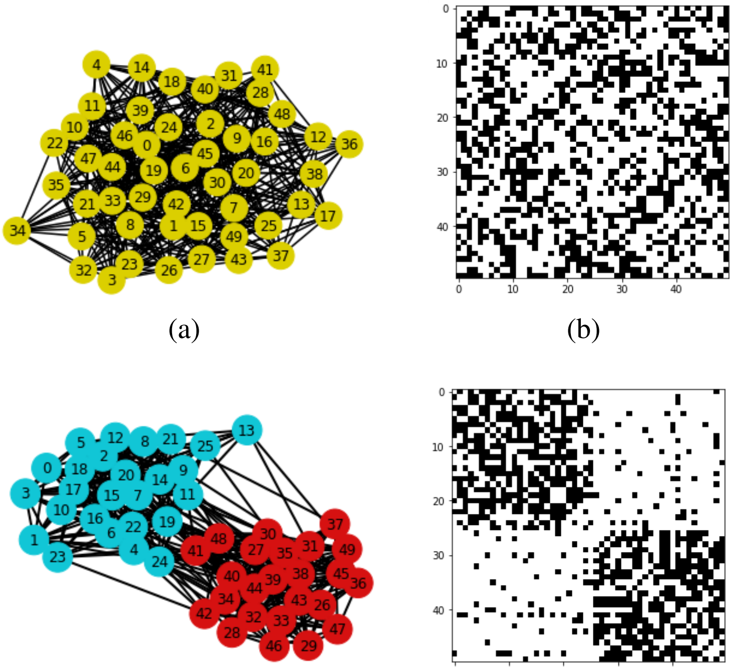
The Erdös-Rényi model, ER(n, p) is defined as follows:

- n is the number of vertices in the graph.

- p denotes the probability of connection between each pairs of different vertices.

In the literature, ER(n, p) is usually denoted $G(n, p)$. We note that an Erdös-Rényi model is equivalent to a Stochastic Block model with a single community.

Samples of both models are displayed in Figure 1.



(a)

(b)

(c)

(d)

Figure 1: **Samples from an ER and SBM models.** (a) Realization of ER$(n = 50, p = 0.4)$. (b) Adjacency matrix of the graph on the left. (c) Realization of SBM$(n = 50, W)$ with $k = 2$, $w_{11} = 0.6$, $w_{12} = w_{21} = 0.1$, and $w_{22} = 0.5$. Each community, identified by node colors, has size 25. (d) Adjacency matrix of the graph on the left.

2.2 Hidden ER-weight

As stated in the introduction, the chief motivation of this thesis is to estimate the expected fraction of spurious edges in a Stochastic Block model. Following the recent framework in [14, 15], this requires abstracting a Stochastic Block model as a mixture of two models as follows:

$$\text{SBM}(n, W) = \lambda \cdot \text{ER}(n, p) + (1 - \lambda) \cdot \text{SBM}(n, W'), \tag{1}$$

for certain $0 \leq \lambda, p \leq 1$, and symmetric $k \times k$ matrix W' with entries between 0 and 1, associated with an alternative Stochastic Block model.

Namely, we interpret the placed as well as missing connections in a realization of SBM(n, W) as produced by an Erdös-Rényi model with probability λ, and by an alternative Stochastic Block model with probability $(1 - \lambda)$. In this context, it is natural to refer to the (mixture) component ER(n, p) as the *hidden ER-model* of SBM(n, W). Edges placed by this model would be interpreted as *spurious* because Erdös-Rényi models have no clusters. *The expected fraction of spurious edges is therefore* $\lambda \cdot p$.

There is however a key issue with the decomposition in (1). In order for it to be unique, the component $\text{SBM}(n, W')$ cannot contain itself a hidden-ER model. Otherwise, if say

$$\text{SBM}(n, W') = \lambda' \cdot \text{ER}(n, p') + (1 - \lambda') \cdot \text{SBM}(n, W''),$$

for some $\lambda' > 0$, then

$$\text{SBM}(n, W) = \lambda \cdot \text{ER}(n, p) + (1 - \lambda)\lambda' \cdot \text{ER}(n, p') + (1 - \lambda)(1 - \lambda') \cdot \text{SBM}(n, W'').$$

But, because a mixture of Erdös-Rényi models is another Erdös-Rényi model, and the same can be said about a mixture of Stochastic Block models, the above identity is equivalent to

$$\text{SBM}(n, W) = \big(\lambda + \lambda' - \lambda\lambda'\big) \cdot \text{ER}(n, p'') + \big(1 - \lambda - \lambda' + \lambda\lambda'\big) \cdot \text{SBM}(n, W'''), \tag{2}$$

where—due to *Bayes' formula*:

$$p'' := \frac{\lambda p + (1 - \lambda)\lambda' p'}{\lambda + (1 - \lambda)\lambda'};$$
$$W''' := \frac{(1 - \lambda)W' + (1 - \lambda)(1 - \lambda')W''}{(1 - \lambda) + (1 - \lambda)(1 - \lambda')}.$$

But note that $(\lambda + \lambda' - \lambda\lambda') = \lambda + (1 - \lambda)\lambda' > \lambda$. (There can be equality only if $\lambda = 1$, but then the original SBM would be an ER-model to start with—so the mixture decomposition would be purposeless.) In particular, the decomposition in (2) would carry a hidden ER-model with a strictly larger weight than in (1).

As discussed in [14, 15], the uniqueness issue can be avoided altogether if we request the weight λ in (1) to be as large as possible. We refer to that maximal λ as the *hidden ER-weight* of $\text{SBM}(n, W)$. From the perspective of equation (1), λ may be interpreted as *the largest expected fraction of placed or missing edges in a Stochastic Block model realization that can be attributed to an Erdös-Rényi model.* Likewise, $\lambda \cdot p$ can be interpreted as the *the largest expected fraction of spurious edges in a Stochastic Block model realization.*

2.3 Hidden ER-weight Calculation

Following the methodology in [14], to find the largest weight λ for which the decomposition in (1) is possible, we need to maximize the λ for which there is some $0 \le p \le 1$ such that, for each possible edge e:

$$\begin{array}{rcl} \mathbb{P}(e \in \text{SBM}(n, W)) & \geq & \lambda \cdot \mathbb{P}(e \in \text{ER}(n, p)) = \lambda \cdot p \\ \mathbb{P}(e \notin \text{SBM}(n, W)) & \geq & \lambda \cdot \mathbb{P}(e \notin \text{ER}(n, p)) = \lambda \cdot (1 - p) \end{array}$$

Equivalently, since the probabilities on the left hand-side above depend only on the communities of the vertices that each edge e connects, we need:

$$(\forall i, j \in C)\ (\exists p \in [0, 1]) : 0 \le \lambda \le \frac{w_{ij}}{p} \ \text{ and } \ 0 \le \lambda \le \frac{1 - w_{ij}}{1 - p}.$$

In other words, if $\min(W)$ denotes the minimum entry in W, and $\max(W)$ denotes the maximum entry in W, the above is equivalent to requiring that

$$\exists p \in [0, 1] : 0 \le \lambda \le \frac{\min(W)}{p} \ \text{ and } \ 0 \le \lambda \le \frac{1 - \max(W)}{1 - p}.$$

But note that the function $p \longrightarrow \frac{\min(W)}{p}$ is monotone decreasing, and $p \longrightarrow \frac{1-\max(W)}{1-p}$ is monotone increasing when $p \in [0,1]$. In particular, the maximum value of λ is associated with the intersection of the graphs of these functions—if it exists at all. By solving the equation:

$$\frac{\min(W)}{p} = \frac{1-\max(W)}{1-p},$$

we find that

$$p = \frac{\min(W)}{1-\max(W)+\min(W)}.$$

Since $0 \le p \le 1$, it follows that the hidden-ER weight of SBM(n, W) is given by the formula:

$$\lambda = 1 - \max(W) + \min(W). \tag{3}$$

Since $0 \le \min(W) \le \max(W) \le 1$, we find that $0 \le \lambda = 1 - \big(\max(W) - \min(W)\big) \le 1$; which is required for a probabilistic interpretation of this weight.

Observe that $\lambda \cdot p = \min(W)$. In particular, $\min(W)$ *is the largest expected fraction of spurious edges in a realization of SBM*(n, W).

2.3.1 2-Clusters Example

Consider the Stochastic Block model with $k = 2$ clusters and connection probability matrix:

$$W = \begin{bmatrix} 0.5 & 0.3 \\ 0.3 & 0.1 \end{bmatrix}.$$

For this model, $\max(W) = 1/2$ and $\min(W) = 1/10$. In particular, $\lambda = 3/5$ and the hidden-ER model has parameter $p = 1/6$. In fact, in accordance with the identity in equation (1), we can rewrite the connection probability as

$$W = \frac{3}{5} \cdot \begin{bmatrix} 1/6 & 1/6 \\ 1/6 & 1/6 \end{bmatrix} + \frac{2}{5} \cdot \begin{bmatrix} 1 & 1/2 \\ 1/2 & 0 \end{bmatrix}.$$

Thus, we expect the majority of the edges ($3/5$ of them) to be placed or missing with no relation whatsoever to its clusters because they are governed by an Erdös-Rényi model. The remaining placed or missing edges are instead governed by the model:

$$\text{SBM}\left(n, \begin{bmatrix} 1 & 1/2 \\ 1/2 & 0 \end{bmatrix}\right).$$

Remarkably, this model allows all connections between nodes in the first cluster, but none in the second one. Instead, connections between the two clusters are totally random.

2.3.2 3-Clusters Example

Consider now the Stochastic Block model with $k = 3$ clusters and connection probability matrix:

$$W = \begin{bmatrix} 0.5 & 0.8 & 0.2 \\ 0.8 & 0.3 & 0.7 \\ 0.2 & 0.7 & 0.1 \end{bmatrix}.$$

In this example, $\max(W) = 0.8$ and $\min(W) = 0.1$. Hence the hidden-ER weight is $\lambda = 3/10$, which is associated to a hidden-ER model with $p = 1/3$. In fact:

$$W = \frac{3}{10} \cdot \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} + \frac{7}{10} \cdot \begin{bmatrix} 4/7 & 1 & 1/7 \\ 1 & 2/7 & 6/7 \\ 1/6 & 6/7 & 0 \end{bmatrix}.$$

Unlike the previous example, in this case we expect the majority of the placed or missing edges ($7/10$ of them) to be attributed to the Stochastic Block model with connection probabilities matrix:

$$W' := \begin{bmatrix} 4/7 & 1 & 1/7 \\ 1 & 2/7 & 6/7 \\ 1/7 & 6/7 & 0 \end{bmatrix}.$$

3 Asymptotic Distribution of Hidden ER-weights

In what remains of this thesis we focus on the problem of estimating hidden-ER weights in Stochastic Block models.

In most practical situations the number k of clusters in a Stochastic Block model is assumed known, but the remaining parameters, i.e, cluster membership and matrix W, are estimated from a single realization of the model. The basic idea behind the estimation methods is to group vertices with similar degrees into the same cluster; which then can be used to estimate the entries in the matrix W.

With an estimate of W, formula (3) provides a natural point estimate for the hidden-ER weight λ of the SBM. Associated with any estimate, however, one would like to quantify the uncertainty associated with it. This motivates to study the asymptotic distribution of the hidden ER-weight when the network size n is large; in particular, *we assume in what follows that the clusters and connection probability matrix have been estimated accurately.*

To accomplish the above, let V_i denote a set of vertices in the cluster $i \in C$. Define $|V_i| = a_i \cdot n$. Namely, a_i is the overall fraction of nodes in cluster i in the network. Accordingly, $a_i \geq 0$, for each $i \in C$, and $\sum_{i \in C} a_i = 1$.

Let N_{ij} denote the number of edges between clusters V_i and V_j. This is a random quantity that depends on the realization of the Stochastic Block model. In fact, from the definition of the model, we have that

$$N_{ij} \sim \text{Binomial}(w_{ij}, n_{ij})$$

where

$$n_{ij} \sim \begin{cases} \binom{a_i n}{2}, & i = j; \\ a_i a_j n^2, & i \neq j. \end{cases}$$

The expected value and variance of the random variable N_{ij} are given by $\mathbb{E}(N_{ij}) = n_{ij} w_{ij}$ and $\mathbb{V}(N_{ij}) = n_{ij} w_{ij} (1 - w_{ij})$, respectively. The natural estimator of w_{ij} is

$$\hat{w}_{ij} := \frac{N_{ij}}{n_{ij}}.$$

As $n \to \infty$, the *Central Limit Theorem* asserts the following *convergence in distribution*:

$$\frac{\hat{w}_{ij} - w_{ij}}{\sqrt{\frac{w_{ij}(1-w_{ij})}{n_{ij}}}} \xrightarrow{d} \text{Normal}(0, 1).$$

Since $n_{ii} \sim \frac{a_i^2 n^2}{2}$, the above is equivalent to:

$$n(\hat{w}_{ij} - w_{ij}) \xrightarrow{d} \text{Normal}\left(0, \frac{\rho_{ij} \cdot w_{ij} \cdot (1 - w_{ii})}{a_i \cdot a_j}\right), \tag{4}$$

where

$$\rho_{ij} := \begin{cases} 2, & i = j; \\ 1, & i \neq j. \end{cases}$$

3.1 No-Ties Asymptotics

In what follows, since the matrix W is symmetric of dimensions $k \times k$, we think of it as a (column) vector of dimension $k(k+1)/2$. This is because we may arrange the entries along and above the diagonal of W along a vector without loosing any information about this matrix. With this convention, the above findings imply that W has an asymptotic multivariate normal distribution of the form:

$$n(\hat{W} - W) \xrightarrow{d} \text{Normal}_{\frac{k(k+1)}{2}}(0, \Sigma),$$

where the variance-covariance matrix Σ is diagonal.

If $g : \mathbb{R}^{\frac{k(k+1)}{2}} \longrightarrow \mathbb{R}$ is *differentiable* at the parameter W of the Stochastic Block model then the so-called *multivariate Delta method* [16] implies that

$$n\left(g(\hat{W}) - g(W)\right) \xrightarrow{d} \text{Normal}\left(0, \nabla g'(W) \cdot \Sigma \cdot \nabla g(W)\right), \tag{5}$$

as $n \to \infty$. Above, ∇ denotes the gradient operator, and the prime the transpose operator.

Due to formula (3), the natural choice for the function g to determine the asymptotic distribution of the hidden-ER weight is

$$g(x) := 1 - \max(x) + \min(x), \text{ for all } x \in \mathbb{R}^{\frac{k(k+1)}{2}}.$$

In this case, the natural *estimator of the hidden-weight* is

$$\hat{\lambda} := g(\hat{W}) = 1 - \max(\hat{W}) + \min(\hat{W}).$$

The above function g is differentiable only at those vectors $x \in \mathbb{R}^{\frac{k(k+1)}{2}}$ such that the $\max(x)$ and $\min(x)$ are achieved at a single entry of x. So, *if* $\max(W)$ *and* $\min(W)$ *are attained at a single entry of (the vector) W then the asymptotic formula in (5) holds.* In this case—so long n is large enough—we may replace $g(x)$ by an affine linear transformation of x involving only two of its entries; in particular, $\nabla g(x)$ is constant. In fact:

$$\nabla g(x) = \left(\tau_1, \tau_2, ..., \tau_{\frac{k(k+1}{2}}\right)', \text{ for all } x \in \mathbb{R}^{\frac{k(k+1)}{2}},$$

where

$$\tau_i := \begin{cases} -1, & W_i = \min(W); \\ 1, & W_i = \max(W); \\ 0, & \text{otherwise.} \end{cases}$$

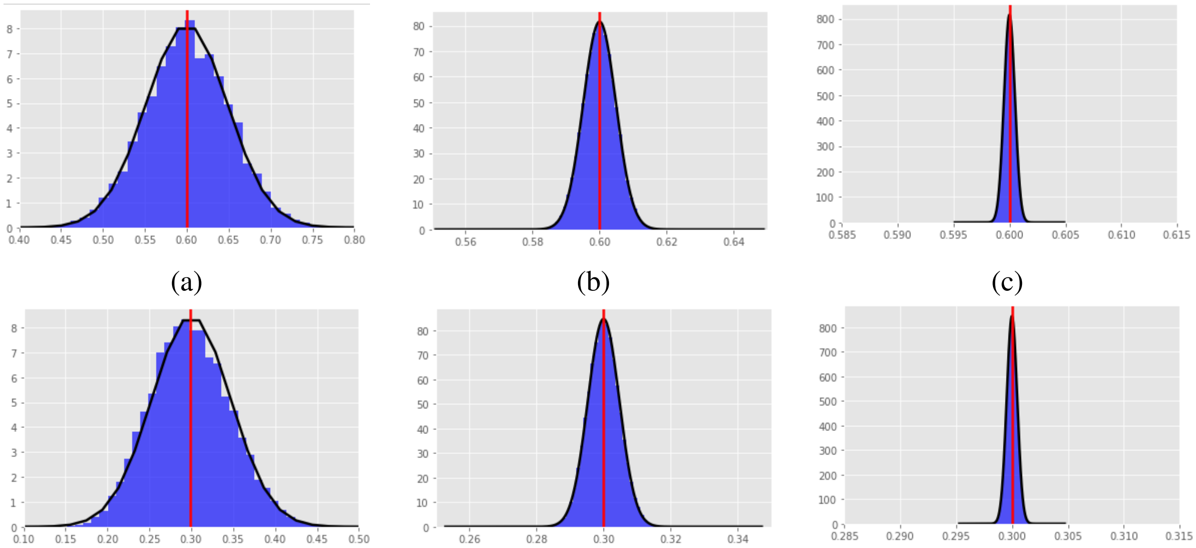
Taking this into account, equation (5) may be rewritten as

$$n(\hat{\lambda} - \lambda) \xrightarrow{d} \text{Normal}\left(0, \sum_{i=1}^{\frac{k(k+1)}{2}} \tau_i^2 \cdot \Sigma_{ii}\right).$$

In other words, *when n is large and there are no ties for the maximum and minimum of the entries in W:*

$$\hat{\lambda} \overset{d}{\approx} \text{Normal}\left(\lambda, \frac{1}{n^2} \sum_{i=1}^{\frac{k(k+1)}{2}} \tau_i^2 \cdot \Sigma_{ii}\right). \tag{6}$$

As seen in Figure 2, the above approximation performs well on numerical experiments associated with the 2- and 3-clusters examples in Section 2.3.1 and 2.3.2, respectively. The simulations were carried out mostly using *Python 3*.



(a) (b) (c)

(d) (e) (f)

Figure 2: **Performance of the Delta method for no-ties Stochastic Block models.** Vertical red lines denote the true hidden-ER weight for each model. Blue histograms (each based on 500,000 samples) summarize estimates of this weight based on simulations. Black curve is the asymptotic distribution approximation of hidden-ER estimators. Top row: plots associated with the example in Section 2.3.1 with $k = 2$ and $\alpha = (0.3, 0.7)$ for (a) n =50, (b) n =500, (c) n =5,000 nodes. Bottom row: plots associated with the example in Section 2.3.2 with $k = 3$ and $\alpha = (0.3, 0.5, 0.2)$ for (d) n =50, (e) n =500, (f) n =5,000 nodes.

3.1.1 Tied Case Issue with Delta Method

Consider now the Stochastic Block model with $k = 2$ clusters and connection probability matrix:

$$W = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.3 \end{bmatrix}.$$

In this case, $\max(W) = 0.3$ and $\min(W) = 0$; in particular, $\lambda = 0.7$.

In this case, the maximum entry in W is achieved with ties at w_{11} and w_{22}. There are no ties for the minimum. (Because W is symmetric, the determination of ties only considers entries along or above its diagonal.)

As seen in Figure 3, the approximation given by equation (6) is not longer valid. In fact, the distribution of the estimated hidden-ER weight looks skewed to the left—*suggesting that there is no asymptotic normality when there are ties in the maximum or maximum entry in W.*

We may explain this noticing that

$$\hat{W} = \begin{bmatrix} \hat{w}_{11} & 0 \\ 0 & \hat{w}_{22} \end{bmatrix}.$$

Hence, $g(\hat{W}) = 1 - \hat{w}_{11}$ when $\hat{w}_{11} > \hat{w}_{22}$; however, $g(\hat{W}) = 1 - \hat{w}_{22}$ when $\hat{w}_{11} < \hat{w}_{22}$. In particular, the choice of function g to apply the Delta method depends on the sample—which is troublesome. Nevertheless, this suggests that the asymptotic distribution of $\hat{\lambda}$ relates to the conditional distribution of $1 - \hat{w}_{11}$ when $\hat{w}_{11} > \hat{w}_{22}$, and the conditional distribution of $1 - \hat{w}_{22}$ when $\hat{w}_{11} < \hat{w}_{22}$.

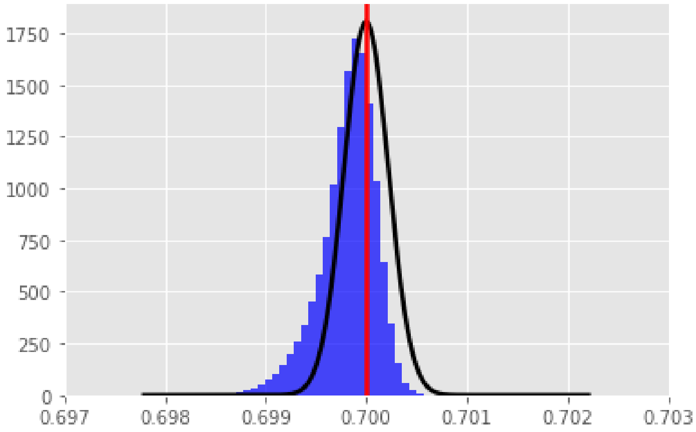


Figure 3: **Flawed Delta method approximation for tied Stochastic Block models.** Vertical red line denotes the true hidden-ER weight of the model in Section 3.1.1. Blue histogram (based on 500,000 samples) summarizes estimates $\hat{\lambda}$ of λ based on simulations. Black curve is the asymptotic distribution approximation of hidden-ER estimators.

3.2 Tied Case Asymptotics

As discussed in Section 3.1.1, when the $\max(W)$ or $\min(W)$ are achieved at more than one entry along or above the diagonal of W, the asymptotic distribution of the estimator $\hat{\lambda}$ of the hidden-ER weight λ associated with a SBM(n, W) needs not to be normally distributed.

In this section *we show via examples how to determine the asymptotic distribution of $\hat{\lambda}$ when there may be ties in $\max(W)$ or $\min(W)$.* In doing so, we require first some general observations about maxima and minima of independent normal random variables.

3.2.1 Distribution of Normal Extremes

In the sections that follow, *when Z is a continuous random variable, $\mathbb{P}(Z = z)$ denotes its p.d.f.* Instead, *if Z is asymptotically continuous (w.r.t. some parameter tending to infinity) then $\mathbb{P}(Z = z)$ denotes its approximate p.d.f.*

In the remainder, $\varphi(\cdot)$ denotes the *probability density function* (p.d.f.) and $\Phi(\cdot)$ the *cumulative probability function* (c.d.f.) of the *standard normal distribution*, respectively.

If $X \sim \text{Normal}(\mu_1, \sigma_1^2)$ and $Y \sim \text{Normal}(\mu_2, \sigma_2^2)$ are *independent* then the p.d.f. of $\max(X, Y)$ is given by

$$\begin{aligned}\mathbb{P}\big(\max(X,Y) = z\big) &= \mathbb{P}\big(\max(X,Y) = z, X \geq Y\big) + \mathbb{P}\big(\max(X,Y) = z, X \leq Y\big)\\ &= \mathbb{P}(X = z, Y \leq z) + \mathbb{P}(Y = z, X \leq z)\\ &= \mathbb{P}(X = z) \cdot \mathbb{P}(Y \leq z) + \mathbb{P}(Y = z) \cdot \mathbb{P}(X \leq z)\\ &= \varphi\big(\frac{z - \mu_1}{\sigma_1}\big) \cdot \Phi\big(\frac{z - \mu_2}{\sigma_2}\big) + \varphi\big(\frac{z - \mu_2}{\sigma_2}\big) \cdot \Phi\big(\frac{z - \mu_1}{\sigma_1}\big),\end{aligned} \tag{7}$$

where for the last identity we have used that $(X - \mu_1)/\sigma_1$ and $(Y - \mu_2)/\sigma_2$ are standard normal random variables.

Similarly:

$$\mathbb{P}\big(\min(X,Y) = z\big) = \varphi\big(\frac{z - \mu_2}{\sigma_2}\big) \cdot \Big(1 - \Phi\big(\frac{z - \mu_1}{\sigma_1}\big)\Big) + \varphi\big(\frac{z - \mu_1}{\sigma_1}\big) \cdot \Big(1 - \Phi\big(\frac{z - \mu_2}{\sigma_2}\big)\Big). \tag{8}$$

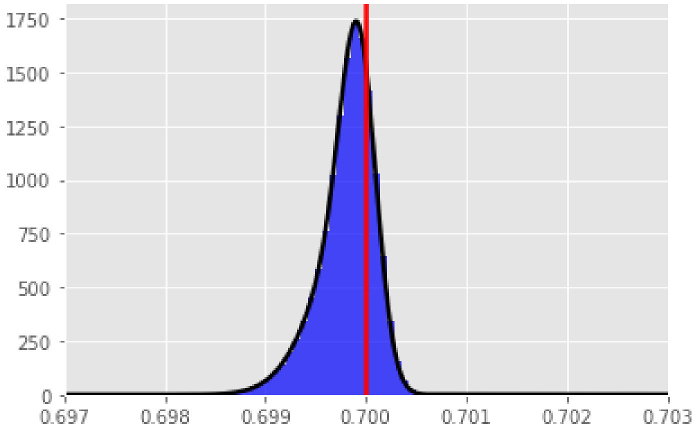


Figure 4: **Corrected approximation for a tied Stochastic block model.** In this model n =5,000, k = 2, and a tie occurs at the maximum of W. Vertical red line denotes the true hidden-ER weight λ of the model in Section 3.1.1. Blue histogram (based on 500,000 samples) summarizes estimates $\hat{\lambda}$ of λ based on simulations. Black curve is the asymptotic distribution approximation of the hidden-ER estimator based on equation (7).

3.2.2 Revisiting the Model with a Flawed Delta Method Approximation

In Section 3.1.1, we saw that the delta method was not adequate to estimating the asymptotic distribution of $\hat{\lambda}$ when the connection probability matrix is of the form

$$W = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}.$$

(See Figure 3.) In this case, regardless of the instance of the SBM(n, W) model, the estimated connection probability matrix will be of the form

$$\hat{W} = \begin{bmatrix} \hat{w}_{11} & 0 \\ 0 & \hat{w}_{22} \end{bmatrix}.$$

In particular, $\min(\hat{W}) = 0$, hence $\hat{\lambda} = 1 - \max(\hat{w}_{11}, \hat{w}_{22})$. Note that $\hat{w}_{11}$ and $\hat{w}_{22}$ are independent and asymptotically normal. In fact, due to equation (4), asymptotically:

$$\hat{w}_{11} \overset{d}{\approx} \text{Normal}\left(0.3, \frac{0.42}{|V_1|^2}\right); \tag{9}$$

$$\hat{w}_{22} \overset{d}{\approx} \text{Normal}\left(0.3, \frac{0.42}{|V_2|^2}\right). \tag{10}$$

Hence, the asymptotic distribution of $\max(\hat{w}_{11}, \hat{w}_{22})$ should be well-approximated by the p.d.f. in equation (7) with $\mu_1 = 0.3$, $\sigma_1^2 = 0.42/|V_1|^2$, $\mu_2 = 0.3$, and $\sigma_2^2 = 0.42/|V_2|^2$. As a matter of fact, as seen in Figure 4, this approximation works extremely well.

3.2.3 Another Single Tie Model

Consider now the Stochastic Block model with $k = 2$ clusters and connection probability matrix:

$$W = \begin{bmatrix} 0.3 & 1 \\ 1 & 0.3 \end{bmatrix}.$$

In this case, $\max(\hat{W}) = 1$ regardless of the realization of the Stochastic block model associated with this matrix. In particular, $\hat{\lambda} = \min(\hat{w}_{11}, \hat{w}_{22})$. In this case, the approximations in equations (9)-(10) still apply, and the p.d.f. of $\hat{\lambda}$ should be approximately given by equation (8) with $\mu_1 = 0.3$, $\sigma_1^2 = 0.42/|V_1|^2$, $\mu_2 = 0.3$, and $\sigma_2^2 = 0.42/|V_2|^2$.

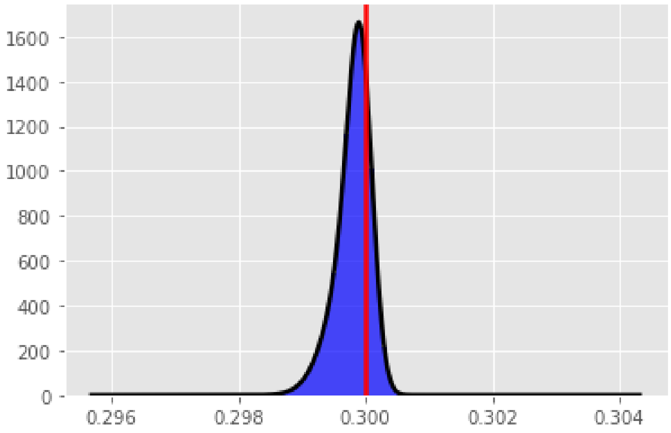


Figure 5: **Approximation for another single tie Stochastic block model.** In this model, n =5,000, $k = 2$, and the tie occurs at the minimum rather than the maximum of W. Vertical red line denotes the true hidden-ER weight λ of the model in Section 3.2.3. Blue histogram (based on 500,000 samples) summarizes estimates $\hat{\lambda}$ of λ based on simulations. Black curve is the asymptotic distribution approximation of the hidden-ER estimator based on equation (8).

3.2.4 A Heuristic for a Model with Two Ties

So far we have only analyzed examples on which there is a tie at either maximum or minimum of W, while the other extreme of $\hat{W}$ is constant.

In this section, we consider models where both $\max(W)$ *and* $\min(W)$ *are achieved with ties but* $\max(\hat{W})$ *and* $\min(\hat{W})$ *are random.* In this setting, the distribution of $\hat{\lambda} := 1 - \max(\hat{W}) + \min(\hat{W})$ is more difficult to characterize because $\max(\hat{W})$ and $\min(\hat{W})$ are *dependent* random variables. In particular, the marginal distributions of $\max(\hat{W})$ and $\min(\hat{W})$ are not enough to characterize their *joint distribution*.

Nevertheless, when there is a significant gap between $\max(W)$ and $\min(W)$, we expect that $\max(\hat{W})$ and $\min(\hat{W})$ are approximately independent. This is because—with high probability—the entries in $\hat{W}$ that usually produce the $\max(\hat{W})$ will be different from those that usually produce the $\min(\hat{W})$. So, since the entries (along or above the diagonal) of $\hat{W}$ are independent, $\max(\hat{W})$ *and* $\min(\hat{W})$ *should be approximately independent when* $\max(W)$ *and* $\min(W)$ *are very different*. In this case, the distribution of $\hat{\lambda}$ should be well approximated by the (shifted) convolution of the distributions of $-\max(\hat{W})$ and $\min(\hat{W})$. In other words, $\hat{\lambda}$ should have the approximate p.d.f.:

$$\mathbb{P}(\hat{\lambda} = z) \approx \int_{-\infty}^{\infty} \mathbb{P}(\max(\hat{W}) = t) \cdot \mathbb{P}(\min(\hat{W}) = z + t - 1)\, dt; \tag{11}$$

$$= \int_{-\infty}^{\infty} \mathbb{P}(\max(\hat{W}) = 1 + t - z) \cdot \mathbb{P}(\min(\hat{W}) = t)\, dt. \tag{12}$$

To fix ideas consider the three-communities Stochastic block model with connection probability matrix

$$W = \begin{bmatrix} 0.3 & 0.6 & 0.9 \\ 0.6 & 0.5 & 0.9 \\ 0.9 & 0.9 & 0.3 \end{bmatrix}.$$

In this case, $\max(W) = 0.9 = w_{13} = w_{23}$ and $\min(W) = 0.3 = w_{11} = w_{33}$; in particular, $\lambda = 0.4$. Since both extremes in W are achieved with ties, the asymptotic distribution of $\max(\hat{W})$ and $\min(\hat{W})$ should each be well-approximated using the methods in Sections 3.2.2-3.2.3. As seen in Figure 6, the approximation that then follows form equation (11) or (12) agrees very well with the simulations.

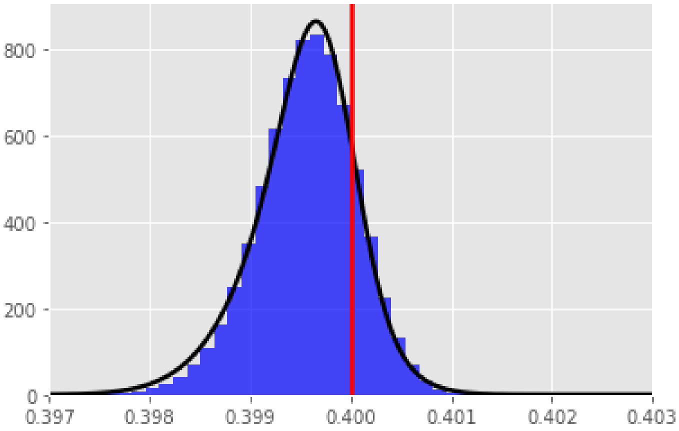


Figure 6: **Approximation for a Stochastic block model with ties at both extremes.** In this model n =5,000, k = 3, and both extremes of W occur with ties. Vertical red line denotes the true hidden-ER weight of the model in Section 3.2.4. Blue histogram (based on 500,000 samples) summarizes estimates $\hat{\lambda}$ of λ based on simulations. Black curve is the asymptotic distribution approximation of the hidden-ER estimator based on the approximation in equations (11) or (12).

3.3 Hidden-ER Weights in Connection with Order Statistics

We complete this chapter relating our previous discussion to the more general setting of *order statistics* [17].

To start observe that $\hat{\lambda}$ is a function of $(\max(W), \min(W))$; in particular, to approximate its distribution it should suffice to approximate the joint distribution of $\max(\hat{W})$ and $\min(\hat{W})$. For this note that the *joint c.d.f.* of these random variables is

$$\begin{aligned} F(x,y) &:= \mathbb{P}(\min(\hat{W}) \le x, \max(\hat{W}) \le y) \\ &= \mathbb{P}(\max(\hat{W}) \le y) - \mathbb{P}(x < \min(\hat{W}), \max(\hat{W}) \le y) \\ &= \mathbb{P}(\forall i \le j : \hat{w}_{ij} \le y) - \mathbb{P}(\forall i \le j : x < \hat{w}_{ij} \le y) \\ &= \prod_{i \le j} \Phi_{ij}(y) - \prod_{i \le j} \big(\Phi_{ij}(y) - \Phi_{ij}(x)\big), \text{ for all } x \le y; \end{aligned}$$

where Φ_{ij} denotes the c.d.f. of $\hat{w}_{ij}$. But, due to equation (4), when n is large enough, the entries are independent (though not necessarily identically distributed) normal random variables. As a result

$$F(x,y) \approx \prod_{i \le j} \Phi\left(\frac{(y - w_{ij})\sqrt{|V_i||V_j|}}{\sqrt{\rho_{ij} w_{ij}(1 - w_{ij})}}\right) - \prod_{i \le j} \left\{\Phi\left(\frac{(y - w_{ij})\sqrt{|V_i||V_j|}}{\sqrt{\rho_{ij} w_{ij}(1 - w_{ij})}}\right) - \Phi\left(\frac{(x - w_{ij})\sqrt{|V_i||V_j|}}{\sqrt{\rho_{ij} w_{ij}(1 - w_{ij})}}\right)\right\}.$$

The above approximation works in any case—regardless if there are or not ties in the extremes of W. Furthermore, at least theoretically, we may approximate the joint p.d.f. $f(x,y)$ of $\max(\hat{W})$ and $\min(\hat{W})$ differentiating the right hand-side above w.r.t. x and then w.r.t. y.

$$\mathbb{P}(\hat{\lambda} = z) \approx \int_{-\infty}^{\infty} \frac{\partial^2 F}{\partial y \partial x}(t, z + t - 1)\, dt.$$

In practice, however, the above approach is easier said than done due to the appearance of

$$\frac{k(k+1)}{2} \cdot \left(\frac{k(k+1)}{2} - 1\right) \sim \frac{k^4}{4}$$

integrals, each of which has an integrand composed by k factors. This makes the numerical approximation of the above integral impractical, even for moderately small values of k.

4 Conclusions and Future Work

In Section 2, we introduced the notion of hidden-ER weight to assess how much a given Stochastic block model resembles an Erdös-Rényi one. For a given blockmodel SBM(n, W), we defined its hidden-ER weight as $\lambda := 1 - \max(W) + \min(W)$. This weight measures the largest expected fraction of placed or missing edges in the blockmodel that may be attributed to an Erdös-Rényi model. Edges placed by this later model may be interpreted as “spurious” because ER-models have no communities by definition. In particular, $\min(W)$ is the largest fraction of spurious edges we expect in a realization of SBM(n, W).

In Section 3, we studied how to estimate the hidden-ER weight when the matrix W is estimated from various realizations of the blockmodel. Given an estimate $\hat{W}$ of W, we estimated the hidden-ER weight by $\hat{\lambda} := 1 - \max(\hat{W}) + \min(\hat{W})$. We determined or approximated the asymptotic distribution of this estimator under different configurations of the blockmodel. Overall, the variance of $\hat{\lambda}$ is small, seemingly $O(1/n)$. However, when $\max(W)$ or $\min(W)$ is attained at more than one entry along or above the diagonal of W, $\hat{\lambda}$ has a negative bias.

The methods in Section 3 require various estimates of W to approximate the distribution of $\hat{\lambda}$. This is very undesirable for real-world networks because, often, one counts with a single realization of the network. There are methods to estimate k and W when there are no preconceived or natural communities in a network [10]. Nevertheless, apart from a point estimate of λ, it is not possible to weight the uncertainty associated with the estimate with a single realization. There may be ways around this, however. Indeed, if we select nodes at random (i.e, without reference to the communities) in a Stochastic block model, the resulting graph is a realization of the blockmodel with the same matrix W, just with a reduced number of nodes. We may exploit this strategy to generate various realizations of the same model from a single instance of it.

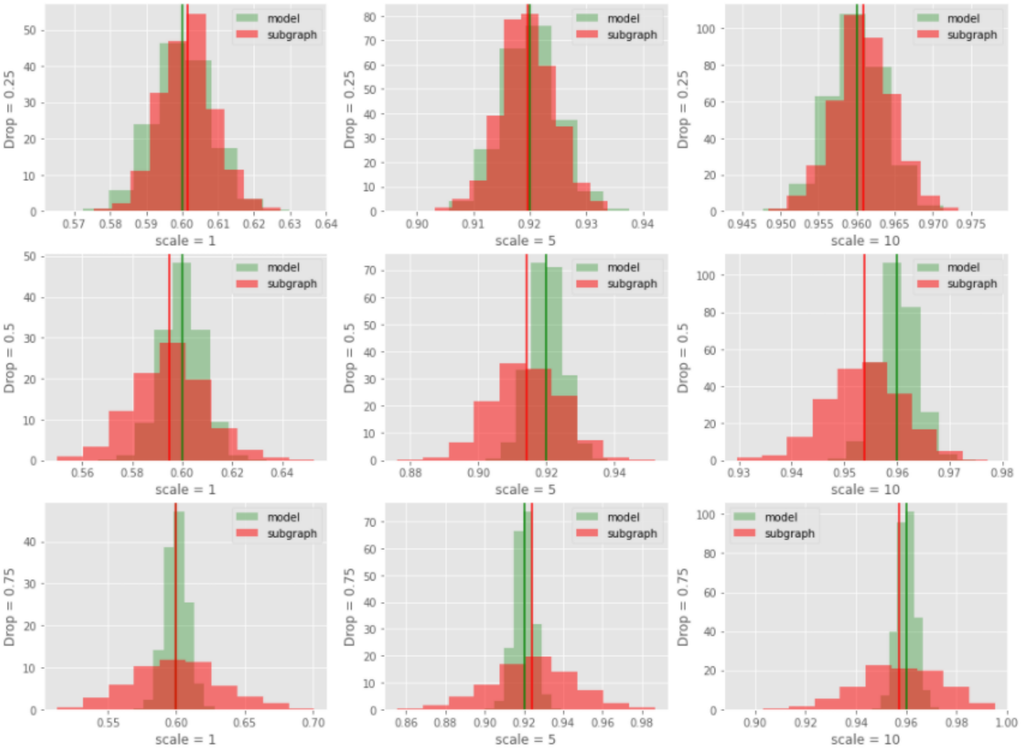


Figure 7: **Testing single blockmodel realization to estimate hidden-ER weights.** Experiments associated with blockmodels of the form SBM$(n = 200, W/s)$, with $s = 1, 5, 10$, and W as in equation (13). Plots in the same row have the same drop percentages (25%, 50%, or 75% from top to bottom, respectively). Each histogram is based on 1,000 samples. The green vertical line shows the mean of green histogram, and the red vertical line shows the mean of red histogram.

To fix ideas, consider the blockmodel with $n = 200$ and connection probability matrix

$$W = \begin{bmatrix} 0.5 & 0.3 \\ 0.3 & 0.7 \end{bmatrix}. \tag{13}$$

In particular, $\lambda = 0.6$. We simulated this blockmodel 1,000 times, using two communities of equal size. The green histograms in the first column of Figure 7 display the estimated hidden-ER weights associated with each of the 1,000 estimates of W. In addition, for each of these realizations, we estimated the matrix W using a random sub-graph of the original network according to different drop percentages. The red histograms in the fist column of Figure 7 display the estimated hidden-ER weights using these sub-graphs. Each column in Figure 7 is associated with a different "scale" parameter that controls the sparsity of the network. Instead, each row is associated with a different "drop percentage." See the figure caption for details on how these parameters vary across the columns and rows. In all the plots, the red vertical line shows the mean of red histogram, whereas the green vertical line shows the mean of green histogram.

In Figure 7, the plots with 50% or 75% drop percentage have histograms that do not resemble well each other. Nevertheless, regardless of the drop percentage and scale parameter, the mean of the red histogram is always within the support green histogram—provably because of the large amount of sub-graphs generated to recover the matrix W. Further, with a 25% drop percentage, we obtain the best and reasonably accurate estimates of the hidden-ER weight λ.

Future work could focus on the best estimation based on a specific drop percentage or sparsity regarding the estimation of hidden-ER weights when the communities of a blockmodel are known but the matrix W is not. Also, regarding the use of order statistics to deal with the general case of ties in W, it would be valuable to find a more feasible approximation of the joint p.d.f. of $\max(\hat{W})$ and $\min(\hat{W})$. Finally, it remains to explore if hidden-ER weights could be helpful for blockmodel selection in real-world networks without preconceived communities.

References

[1] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, p. 163–166, 2016.

[2] P. Li, H. Dau, G. Puleo, and O. Milenkovic, "Motif clustering and overlapping clustering for social network analysis," *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 2017.

[3] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, p. 76–80, 2003.

[4] P. Sah, L. O. Singh, A. Clauset, and S. Bansal, "Exploring community structure in biological networks with random graphs," *BMC Bioinformatics*, vol. 15, no. 1, 2014.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, p. 888–905, 2000.

[6] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein–protein interaction network," *Bioinformatics*, vol. 22, no. 18, p. 2283–2290, 2006.

[7] P. Erdős and A. Rényi, "On random graphs. i," *Publicationes Mathematicae*, p. 290–297, 1959.

[8] S. E. Fienberg, "A brief history of statistical models for network analysis and open challenges," *Journal of Computational and Graphical Statistics*, vol. 21, no. 4, p. 825–839, 2012.

[9] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: First steps," *Social Networks*, vol. 5, no. 2, p. 109–137, 1983.

[10] C. Lee and D. J. Wilkinson, "A review of stochastic block models and extensions for graph clustering," *Applied Network Science*, vol. 4, no. 1, 2019.

[11] B. Karrer and M. E. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, no. 1, 2011.

[12] T. P. Peixoto, "Nonparametric bayesian inference of the microcanonical stochastic block model," *Physical Review E*, vol. 95, no. 1, 2017.

[13] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, “Mixed membership stochastic blockmodels,” *Journal of Machine Learning Research*, vol. 9, no. 65, pp. 1981–2014, 2008.

[14] A. Pearson and M. E. Lladser, “On Contamination of Symbolic Datasets.” (Submitted: arXiv.2002.05592).

[15] A. Pearson and M. E. Lladser, “Hidden Independence in Unstructured Probabilistic Models,” in *31st International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms (AofA 2020)*, vol. 159 of *Leibniz International Proceedings in Informatics (LIPIcs)*, (Dagstuhl, Germany), pp. 23:1–23:13, Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020.

[16] J. M. Ver Hoef, “Who invented the delta method?,” *The American Statistician*, vol. 66, no. 2, p. 124–127, 2012.

[17] H. A. David and H. N. Nagaraja, *Expected Values and Moments*, p. 159–239. Wiley-Interscience, 2003.