

# MODEL SELECTION FOR NETWORK DATA BASED ON SPECTRAL INFORMATION

Jairo Ivan Peña Hidalgo AND Jonathan R. Stewart

*Florida State University*

*Abstract:* We introduce a new methodology for model selection in the context of modeling network data. The statistical network analysis literature has developed many different classes of network data models, with notable model classes including stochastic block models, latent position models, and exponential families of random graph models. A persistent question in the statistical network analysis literature lies in understanding how to compare different models for the purpose of model selection and evaluating goodness-of-fit, especially when models have different mathematical foundations. In this work, we develop a novel non-parametric method for model selection in network data settings which exploits the information contained in the spectrum of the graph Laplacian in order to obtain a measure of goodness-of-fit for a defined set of network data models. We explore the performance of our proposed methodology to popular classes of network data models through numerous simulation studies, demonstrating the practical utility of our method through two applications.

*Key words and phrases:* Statistical network analysis, network data, model selection, social network analysis.

## 1. Introduction

Network data have witnessed a surge of interest across a variety of fields and disciplines in recent decades, including the study of social networks (Lusher et al., 2013), network epidemiology (involving the spread of disease through networks of contacts) (Morris, 2004), covert networks of criminal activity and terrorism (Coutinho et al., 2020), brain networks (Obando and de Vico Fallani, 2017), financial markets (Finger and Lux, 2017), and more. Network data, as a data structure, are typically represented as a graph (Kolaczyk, 2009), consisting of a set of nodes representing the elements of a population of interest (e.g., researchers in a collaboration network) and a set of pairwise observations or measurements between nodes represented as edges between nodes (e.g., co-authorship on a paper).

Many classes of models have been proposed and developed to study and model network data. A non-exhaustive review of some of the more prominent examples include exponential families of

random graph models (ERGMs) (e.g., Lusher et al., 2013; Schweinberger et al., 2020), stochastic block models (SBMs) (e.g., Holland et al., 1983; Anderson et al., 1992; Wang and Bickel, 2017), latent position models (LPMs) (e.g., Hoff et al., 2002; Sewell and Chen, 2015; Tang et al., 2013; Athreya et al., 2017), and more. Each class offers a unique mathematical platform for constructing models of networks from observed network data, with respective strengths and weaknesses. The exponential family class provides a flexible parametric platform for building models of networks with dependent edges. In contrast, stochastic block models can capture network structure and clustering of nodes through a discrete latent space, whereas latent position models capture network structure and edge dependence through latent node positions in (e.g.) a latent Euclidean space.

A persistent challenge in statistical network analysis applications is how to compare different models and select models for specific network data sets. At present, the literature has primarily focused on model selection problems within each class of models, tailoring methods to specific classes of models (SBMs: Wang and Bickel (2017); Latouche et al. (2014); ERGMs: Hunter et al. (2008); LSMs: Ryan et al. (2017)). As a result, there is a gap in the literature which explores methods for comparing model fit or performing model selection across models from different mathematical platforms, e.g., comparing an ERGM to an SBM to an LPM. In this work, we introduce a novel non-parametric methodology for model selection in network data settings that can be applied to a broad class of models under weak assumptions, capable of facilitating comparison of models with different mathematical foundations. Our method utilizes information in the spectrum of the graph Laplacian in order to select a best fitting model for an observed network, and essentially only requires the ability to simulate adjacency matrices from candidate models and compute eigenvalues of the graph Laplacian derived from the adjacency matrices.

The rest of the paper is organized as follows. Section 2 reviews spectral properties of the graph Laplacian for networks and motivates the use of spectral information in the model selection problem for network data. Our proposed methodology is introduced in Section 3. We present experimental studies and simulations in Section 4, and two applications of our methodology in Section 5.

## 2. Spectral properties of the graph Laplacian

We consider simple undirected networks defined on a set of  $N$  nodes with corresponding adjacency matrix  $\mathbf{X} \in \{0, 1\}^{N \times N}$ , where  $X_{i,j} = 1$  corresponds to the event that there is an edge between nodes  $i$  and  $j$  and  $X_{i,j} = 0$  otherwise. We adopt the standard conventions that  $X_{i,j} = X_{j,i}$  and  $X_{i,i} = 0$ . Extensions of our methodology to directed networks is discussed in Section 4. Extensions to networks with valued edges is possible, but beyond the scope of this work. Let  $\mathbf{d} = \text{deg}(\mathbf{X}) = (\sum_{j=1}^N X_{i,j} : i = 1, \dots, N) \in \mathbb{R}^N$  be the vector of node degrees of the network. The Laplacian matrix, also called the graph Laplacian, is defined as  $\mathbf{L}(\mathbf{X}) := \text{diag}(\mathbf{d}) - \mathbf{X}$ , where  $\text{diag}(\mathbf{d})$  is the  $N \times N$  diagonal matrix with diagonal  $\mathbf{d}$ . Since  $\mathbf{L}(\mathbf{X})$  is symmetric and positive semi-definite (Brouwer

and Haemers, 2012), the eigenvalues of  $L(\mathbf{X})$  will all be real and non-negative. Throughout, we will let  $\boldsymbol{\lambda} \in \mathbb{R}^N$  denote the vector of ordered eigenvalues (from smallest to largest) of the Laplacian matrix  $L(\mathbf{X})$ . The vector  $\boldsymbol{\lambda}$  will depend on the adjacency matrix  $\mathbf{X}$  through  $L(\mathbf{X})$ , however, for ease of presentation, we do not make this dependence explicit notationally, as it will be clear contextually.

Eigenvalues of Laplacian matrices encode many well known properties of a network. For example, the multiplicity of the eigenvalue 0 corresponds to the number of connected components in the network (Brouwer and Haemers, 2012). The second smallest eigenvalue (possibly 0) is known as the algebraic connectivity (Fiedler, 1973), and measures the overall connectivity of a graph (de Abreu, 2007). It is also used in establishing Cheeger inequalities (Donetti et al., 2006), which have applications in image segmentation (Shi and Malik, 2000), graph clustering (Kwok et al., 2013) and expander graphs (Hoory et al., 2006). The subsequent eigenvalues of the Laplacian matrix are related to the minimal cuts (weighted edge deleting) required to partition a network (Bollobás and Nikiforov, 2004).

Two undirected graphs with adjacency matrices  $\mathbf{A}$  and  $\mathbf{B}$  are isomorphic if there exists a permutation matrix  $\mathbf{P}$  such that  $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^t$ , which requires that the adjacency matrices be similar  $\mathbf{A} = \mathbf{P}\mathbf{B}\mathbf{P}^{-1}$ , noting that a permutation matrix  $\mathbf{P}$  satisfies  $\mathbf{P}^t = \mathbf{P}^{-1}$ . In such cases, the corresponding graph Laplacian matrices will be similar as well:

$$L(\mathbf{A}) = \text{deg}(\mathbf{P}\mathbf{B}\mathbf{P}^t) - \mathbf{P}\mathbf{B}\mathbf{P}^t = \mathbf{P}\text{deg}(\mathbf{B})\mathbf{P}^t - \mathbf{P}\mathbf{B}\mathbf{P}^t = \mathbf{P}L(\mathbf{B})\mathbf{P}^t.$$

Consequently, since  $L(\mathbf{B})$  is Hermitian, there exists an eigen decomposition  $L(\mathbf{B}) = \mathbf{U}\mathbf{D}\mathbf{U}^t$ . Hence,  $L(\mathbf{A}) = \mathbf{P}(\mathbf{U}\mathbf{D}\mathbf{U}^t)\mathbf{P}^t = (\mathbf{P}\mathbf{U})\mathbf{D}(\mathbf{P}\mathbf{U})^t$ . As a result, if  $\boldsymbol{\lambda}$  is a vector of eigenvalues of  $L(\mathbf{B})$ , it is also a vector of eigenvalues of  $L(\mathbf{A})$ . In our context, this implies one can always differentiate two non-isomorphic networks if their eigenvalues are different. The reverse result is not true in general. There are graphs possessing the same eigenvalue decomposition (referred to as cospectral or isospectral) which are not isomorphic (Cvetković et al., 1980). However, numerical evidence suggests that the fraction of (non-isomorphic) cospectral graphs tends to zero as the number of nodes in a graph grows (Brouwer and Haemers, 2012).

Several applications of spectral decomposition of the Laplacian matrix have been proposed in the network analysis literature. For example, spectral clustering (Von Luxburg, 2007) is a well known clustering algorithm based on the leading eigenvectors of the Laplacian of a similarity matrix. Lei and Rinaldo (2015) established, under mild conditions, the consistency of the spectral clustering method for stochastic block models. Another example is in Newman (2006), where a family of community detection algorithms were proposed for networks based on the spectral decomposition of the graph Laplacian. Lastly, Shore and Lubin (2015) proposed a statistic for evaluating goodness-of-fit for network models reminiscent of the  $R^2$  statistic in regression settings,

which compares eigenvalues of the graph Laplacian generated from a model fit to the eigenvalues of the graph Laplacian from a pre-specified *null* model (typically taken to be a Bernoulli random graph model, referred to as a density only model).

In light of these results, it is natural to regard the vector of eigenvalues  $\lambda$  as a signature of a network, containing important topographical and structural information which can be exploited for the purposes of model selection. Our proposed methodology compares the empirical distribution of the spectrum of the graph Laplacian of candidate models to that of an observed network. Our methodology is motivated by the following considerations regarding properties of the graph Laplacian.

First, if the true data generating process is in the list of candidate models, the observed eigenvalues derived from an observed network are expected to fall within the spectral distribution of the data generating process. If, in practice, none of the proposed models are the true generating process, candidate models can still be assessed by their ability to capture the spectrum of the observed graph Laplacian, providing a means for developing a methods for model selection. Second, we can obtain a relative measure of fit among competing models depending on how well the spectrum of the observed graph Laplacian is captured by candidate models, providing a means to not only select a best fitting model, but also to compare the fit of the best fitting model to unselected alternatives. Third, our methodology requires no parametric assumptions on the data generating process and is able to compare models across different mathematical platforms, including models which do not have a well-defined likelihood function or which are constructed through a stochastic process, an example of which are agent-based models (e.g., [Snijders et al., 2010](#); [Jackson and Watts, 2002](#)) or generative algorithms based on preferential attachment models (e.g., [Barabasi and Albert, 1999](#); [Zeng et al., 2013](#)).

### 3. Methodology

We outline a methodology for model selection in network data settings which exploits the spectral properties of the graph Laplacian, motivated by considerations in the previous section. We assume throughout that the network is completely observed, denoted by  $\mathbf{X}_{\text{obs}}$ . The corresponding observed vector of eigenvalues of the observed graph Laplacian  $L(\mathbf{X}_{\text{obs}})$  is denoted by  $\lambda_{\text{obs}}$ . Our fundamental inferential goal is to select a best fitting model for the observed network  $\mathbf{X}_{\text{obs}}$  from a set of candidate models  $\{\mathcal{M}_1, \dots, \mathcal{M}_M\}$  ( $M \geq 2$ ). We frame the problem as a classification problem and aim to construct a classifier  $\mathcal{P} : \mathbb{R}^N \mapsto \{1, \dots, M\}$  trained on the spectrum of the graph Laplacian for each of the candidate models in order to predict a class  $m^* \in \{1, \dots, M\}$  for a given vector of eigenvalues, namely  $\lambda_{\text{obs}}$ . We present our model selection method algorithm in [Table 1](#).

**Model selection procedure:**

1. Simulate  $K$  networks  $\mathbf{X}^{(m,1)}, \dots, \mathbf{X}^{(m,K)}$  from each of the candidate models  $\mathcal{M}_m \in \{\mathcal{M}_1, \dots, \mathcal{M}_M\}$ .
2. For each  $\mathbf{X}^{(m,k)}$ , compute the Laplacian matrix  $\mathbf{L}(\mathbf{X}^{(m,k)})$  and the corresponding vector of eigenvalues  $\boldsymbol{\lambda}^{(m,k)} \in \mathbb{R}^N$ .
3. Construct a design matrix  $\mathbf{D} \in \mathbb{R}^{(KM) \times N}$  by stacking the  $KM$  vectors of eigenvalues  $\boldsymbol{\lambda}^{(m,k)}$  to form the rows of  $\mathbf{D}$ .
4. Train a classifier  $\mathcal{P} : \mathbb{R}^N \mapsto \{1, \dots, M\}$  to predict a model  $m^* \in \{1, \dots, M\}$  using the  $K$  simulated vectors of eigenvectors  $\boldsymbol{\lambda}^{(m,k)}$  for each class  $m \in \{1, \dots, M\}$  contained in the design matrix  $\mathbf{D}$ . Feature engineering is advised at this stage.
5. Compute the Laplacian matrix  $\mathbf{L}(\mathbf{X}_{\text{obs}})$  for the observed network  $\mathbf{X}_{\text{obs}}$  and the corresponding vector of eigenvalues  $\boldsymbol{\lambda}_{\text{obs}}$ .
6. Predict a class  $m^* = \mathcal{P}(\boldsymbol{\lambda}_{\text{obs}})$  for the observed network using the trained classifier from Step 4 and set  $\mathcal{M}^* = \mathcal{M}_{m^*}$ .

Table 1: Description of the model selection algorithm.

### 3.1 Selection of classifier

Real life networks can possess hundreds, thousands or even millions of nodes. As the dimension of the vector of eigenvalues of the graph Laplacian matrices is equal to the number of nodes in the network, classification methods based on eigenvalues of the Laplacian matrix will be prone to the usual pitfalls of high dimensional classification problems. The literature for classification methods is quite extensive, which makes the choice of classifier a critical step in our methodology, although we show in Section 4 that the effect of the choice of classifier may not have a significant effect on the results of our methodology under certain circumstances. In light of these results, we consider practical concerns of the implementation of the choice of classifier.

Linear discriminant analysis, which requires the computation of the inverse of a covariance matrix, has been shown in practice to suffer a decay in performance as the number of variables increases and the sample size is fixed (Bickel and Levina, 2004). Alternative methods include support vector machines, neural networks, random forests, and boosting algorithms, which generally perform well in high-dimensional settings (Hastie et al., 2011). Within this class is the eXtreme Gradient Boosting (XGBoost) method, which offers both scalability and state-of-the-art performance (Chen and Guestrin, 2016). In the rest of this paper we use exclusively XGBoost, with the notable exception being Simulation study 5 in Section 4, in which we compare the performance of different classifiers to establish the claim made earlier in this section.

## 3.2 Relative measure of goodness-of-fit

Many classification algorithms return more than just a predicted class, often returning a vector of propensity scores  $\mathbf{s} = (s_1, \dots, s_M)$  with the property that  $\|\mathbf{s}\|_1 = 1$ . If several models were considered, the propensity scores for many of the models can shrink simply because of the larger number of classes being considered, meaning that the interpretation of propensity scores  $s_1, \dots, s_M$  can depend on  $M$ . To overcome this issue and facilitate the comparison of fit between models, we propose to normalize the propensity scores to obtain a measure of goodness-of-fit which is independent of the number of candidate models  $M$ . To this end, we define

$$\tilde{s}_i = \frac{s_i}{\|\mathbf{s}\|_\infty}, \quad i = 1, \dots, M,$$

to be the normalized score, which is equal to 1 for the highest scoring model. For all remaining models, the normalized score is a measure of the fit of the model relative to the highest scoring model. By rescaling all propensity scores in this manner, the number of models  $M$  which is considered in the candidate set of models has no effect on the interpretation of the (relative) propensity scores.

## 4. Simulation studies

We conduct a number of simulation studies to demonstrate the potential of our proposed methodology. Specifically, we aim to examine the extent to which the signature of a network is contained within the spectrum of the graph Laplacian. Simulation studies permit knowledge of the true data-generating model, which facilitates empirical studies which aim to clarify the conditions under which our proposed methodology is able to successfully differentiate different network models and structural properties of networks.

### 4.1 Simulation study 1: curved exponential families

We study the performance of our methodology on curved exponential families, which have gained popularity in the social network analysis community (e.g., [Snijders et al., 2006](#); [Hunter and Handcock, 2006](#)), as well as other applications (e.g., [Obando and de Vico Fallani, 2017](#); [Schweinberger et al., 2020](#); [Stivala and Lomi, 2021](#)). The prominence of curved exponential family parameterizations for random graph models emerged out of a desire to solve challenges related to degeneracy and fitting of early and ill-posed model specifications ([Snijders et al., 2006](#)). Additionally, curved exponential family parameterizations are able to parsimoniously model complex sequences of graph statistics, such as degree sequences and shared partner sequences, without sacrificing interpretabil-

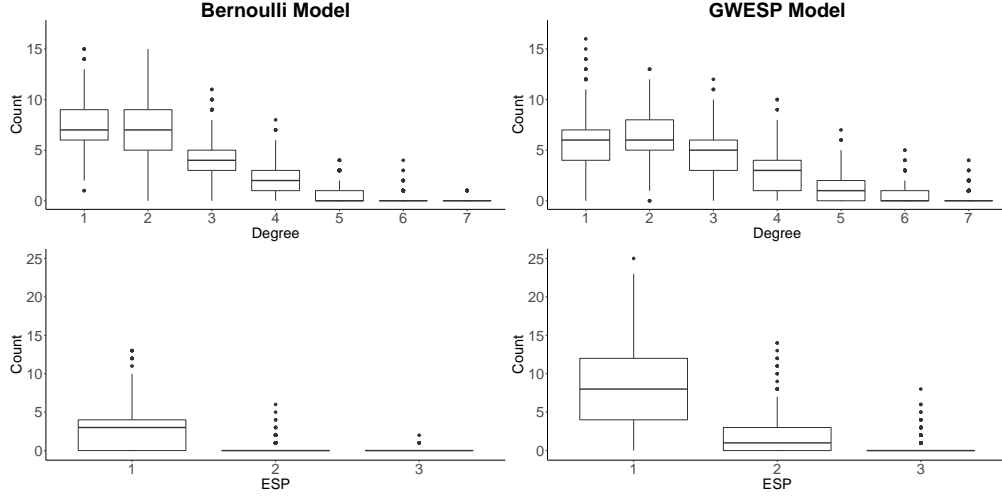


Figure 1: We visualize the degree and ESP distributions of the Bernoulli and GWESP model by simulating 1000 networks from (1) with data-generating parameter vector  $(\theta_1, \theta_2, \theta_3) = (-2.5, 0, 1)$  (Bernoulli) and  $(\theta_1, \theta_2, \theta_3) = (-2.5, .3, 1)$  (GWESP). Each column corresponds to each model and we evidence the rightward shift in the degree and ESP distribution of the GWESP model, relative to the Bernoulli model.

ity (Hunter, 2007; Stewart et al., 2019). A prototypical example used in the social network analysis literature is the geometrically-weighted edgewise shared partner model, which models transitivity through the shared partner sequence (Snijders et al., 2006; Hunter, 2007; Stewart et al., 2019).

We simulate networks according to the following model:

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) \propto \exp\left(\theta_1 \sum_{i < j} x_{i,j} + \sum_{t=1}^{N-2} \sum_{i < j} \eta_t(\theta_2, \theta_3) SP_t(\mathbf{x})\right), \quad (1)$$

where  $\theta_1 \in \mathbb{R}$  controls the baseline propensity for edge formation, and

$$\eta_t(\theta_2, \theta_3) = \theta_2 \exp(\theta_3) [1 - (1 - \exp(-\theta_3 t))], \quad t \in \{1, \dots, N-2\},$$

parameterizes the sequence of shared partner statistics

$$SP_t(\mathbf{x}) = \sum_{i < j} x_{i,j} \mathbb{1}\left(\sum_{h \neq i,j} x_{i,h} x_{h,j} = t\right), \quad t \in \{1, \dots, N-2\}.$$

In words,  $SP_t(\mathbf{x})$  counts the number of edges in the network between nodes which have exactly  $t$  mutual connections, commonly called shared partners in the social network analysis literature. While  $\theta_2 \in \mathbb{R}$ , in typical applications  $\theta_2 \geq 0$  and  $\theta_3 \in (0, \infty)$ , as values of  $\theta_3 < -\log 2$  correspond

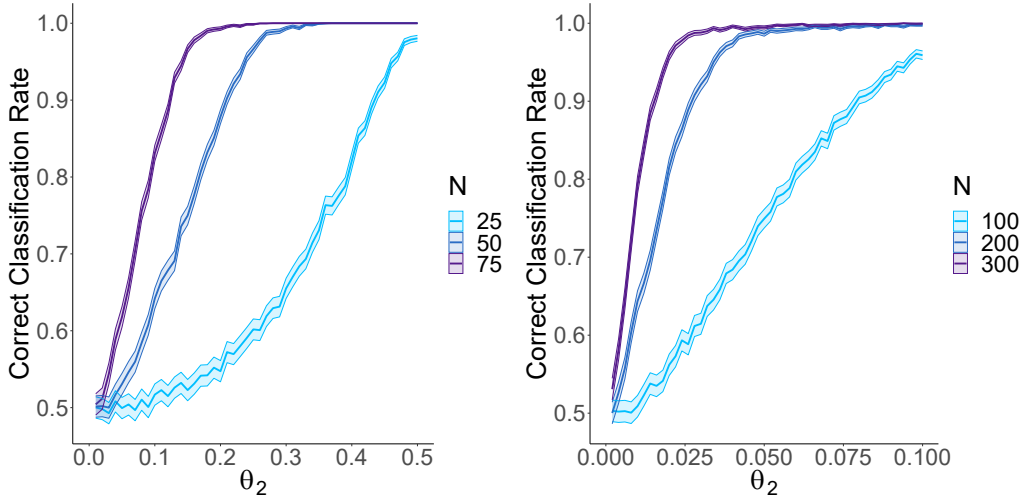


Figure 2: Results of Simulation study 1. (left) Estimate of the correct classification rate with 95% confidence band for networks of sizes  $N = 25, 50, 75$ . (right) Estimate of the correct classification rate with 95% confidence band for networks of sizes  $N = 100, 200, 300$ .

to models which are unstable in the sense of [Schweinberger \(2011\)](#), and empirical evidence suggests that  $\theta_3 \in (0, \infty)$  in many applications ([Schweinberger, 2011](#); [Stewart et al., 2019](#)). The effect that the GWESP model specified by (1) has on the degree and shared partner distributions of networks is visualized in [Figure 1](#), where positive values of  $\theta_2$  stochastically encourage network formations with more transitive edges, i.e., edges between nodes with at least one shared partner, relative to the Bernoulli random graph model with  $\theta_2 = 0$ . This is evidenced by the rightward shift in the ESP distribution of the GWESP model, relative to the Bernoulli model.

We take the true data-generating model  $\mathcal{M}^*$  to be the curved exponential family specified by (1) with parameter vector  $\boldsymbol{\theta}^* = (-2.5, \theta_2, 1)$ , with  $\theta_2$  on a grid covering the interval  $[0, 0.5]$ . Note that when  $\theta_2 = 0$ , the model reduces to a Bernoulli random graph model with edge probability  $p = [1 + \exp(-2.5)]^{-1}$ . We consider the problem of selecting between two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , where  $\mathcal{M}^* = \mathcal{M}_1$  and  $\mathcal{M}_2$  is the Bernoulli random graph model with edge probability  $p = [1 + \exp(2.5)]^{-1}$ . By varying  $\theta_2$  we are able to study the threshold of effect size ( $\theta_2$ ) for which we are able to correctly detect the presence of transitivity in the network, as modeled by the geometrically-weighted edgewise shared partner model in (1).

We vary the network size  $N = 25, 50, 75, 100, 200, 300$ , performing 5000 replicates for each network size. The results of this simulation study are summarized in [Figure 2](#). When  $\theta_2$  is close to 0, the point at which  $\mathcal{M}_1 = \mathcal{M}_2$ , as discussed above, our methodology tends to select  $\mathcal{M}_1$  and  $\mathcal{M}_2$  with equal probability. However, once  $\theta_2$  is sufficiently large (relative to the network size  $N$ ), our methodology correctly selects  $\mathcal{M}_1$  in almost every replicate. The effect of the size of the network is seen as we vary  $N$  from 25 to 300. When the network size is larger ( $N = 100, 200, 300$ ), we are able



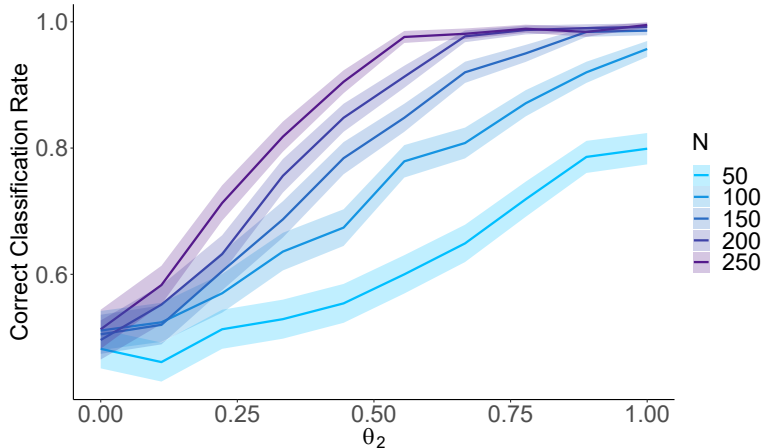


Figure 3: Results of Simulation study 2. Estimates of the correct classification rate with 95% confidence band for various network sizes  $N$ .

to correctly find the data-generating model  $\mathcal{M}_1$  with high probability for smaller values of  $\theta_2$ . In contrast, we require  $\theta_2 \geq .25$  before we are able to have a high confidence in correctly selecting the data-generating model in networks of size  $N = 75$ , requiring  $\theta_2 \geq .5$  for networks of size  $N = 25$ .

## 4.2 Simulation study 2: reciprocity in directed networks

When the adjacency matrix  $\mathbf{X}$  is undirected, the corresponding Laplacian matrix  $\mathbf{L}(\mathbf{X})$  will be positive semidefinite (Brouwer and Haemers, 2012), resulting in a real-valued vector of eigenvalues  $\boldsymbol{\lambda} \in \mathbb{R}^N$ . However, when  $\mathbf{X}$  is the adjacency matrix of a directed network, the graph Laplacian, as defined for undirected networks, may not be positive semidefinite, and may involve complex valued eigenvalues. A common adaptation for directed networks in the literature is to consider the incidence matrix  $\mathbf{B} \in \{0, 1, -1\}^{N \times |E|}$ , where  $|E|$  is the total number of edges in the network. On each column of the incidence matrix exactly one element will be  $-1$ , indicating the node where an edge begins, and exactly one element will be  $1$ , indicating the node where said edge ends. Every other entry is zero. In this manner, a directed network is completely specified by listing all existing edges as columns that indicate which nodes are connected and an orientation between them. We can adapt our proposed methodology to directed networks by considering the symmetric graph Laplacian defined by  $\mathbf{L} := \mathbf{B}^t \mathbf{B}$  (Brouwer and Haemers, 2012).

We simulate directed networks from the probability mass function

$$\mathbb{P}(\mathbf{X} = \mathbf{x}) \propto \prod_{i < j}^N \exp \left( \theta_1 (x_{i,j} + x_{j,i}) + \frac{\theta_2}{2} x_{i,j} x_{j,i} \right), \quad (2)$$

We apply our methodology taking  $\mathcal{M}_1$  to be the density only model with fixed  $\theta_2 = 0$  in (2). We take  $\mathcal{M}^* = \mathcal{M}_2$  to be the general model specified via (2) with unrestricted parameters. We conduct a simulation study by taking  $\theta_1 = -2.5$  in both  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , taking  $\theta_2 = 0$  in  $\mathcal{M}_1$ , and varying  $\theta_2$  on a uniform grid of 100 values in  $[0, 1]$  for  $\mathcal{M}_2$ . The simulation results in Figure 3 are based on 1000 replications in each case, reconfirming findings in the previous simulation study which suggested that the ability of our methodology to detect the true data-generating model depends on how far  $\theta_2$  is from 0, the point at which  $\mathcal{M}_1 = \mathcal{M}_2$ , and the size of the network. Moreover, this study uniquely demonstrates that our methodology can be applied successfully to directed networks.

### 4.3 Simulation study 3: latent position models

Latent variable models for networks, especially latent position models, have witnessed increased popularity and attention since the seminal work of Hoff et al. (2002). In this class of models, nodes are given a latent position  $\mathbf{z}_i \in \mathbf{Z}$  ( $i = 1, \dots, N$ ) in a latent space, typically taken to be the Euclidean space (i.e.,  $\mathbf{Z} = \mathbb{R}^k$ ), although alternative spaces and geometries have been proposed as well, as is the case of ultrametric spaces (Schweinberger and Snijders, 2003), dot product similarity resulting in bilinear forms (Hoff et al., 2002; Athreya et al., 2017), as well as hyperbolic (Krioukov et al., 2010) and elliptic geometries (Smith et al., 2019). Edges in the network are assumed to be conditionally independent given the latent positions of nodes. Following Hoff et al. (2002), we simulate networks in this study accordingly:

$$\log \frac{\mathbb{P}(X_{i,j} = 1 | \mathbf{z}_i, \mathbf{z}_j)}{\mathbb{P}(X_{i,j} = 0 | \mathbf{z}_i, \mathbf{z}_j)} = \theta - \|\mathbf{z}_i - \mathbf{z}_j\|_2, \quad (3)$$

where  $\theta \in \mathbb{R}$  and  $\mathbf{z}_i, \mathbf{z}_j \in \mathbb{R}^k$ . Under this specification, the odds of two nodes forming an edge decreases in the Euclidean distance  $\|\mathbf{z}_i - \mathbf{z}_j\|_2$  between the positions of the two nodes in the latent metric space.

We explore the ability of our methodology to detect the true dimension of a latent space by generating networks from the latent Euclidean model described above, varying the dimension of the latent metric space  $k \in \{1, 2, 3, 4, 5\}$ . Latent positions of nodes are randomly generated from a multivariate normal distribution in dimension  $k \in \{1, 2, 3, 4, 5\}$  with zero mean vector and identity covariance matrix. The candidate competing models are generated in the same fashion across dimensions  $1, \dots, 5$ . We set  $\theta = -2.5$  to ensure a low baseline probability of edge formation, reflecting the sparsity of many real-world networks, and vary the network size  $N \in \{50, 100, 150, 200, 250\}$ . We apply our model selection methodology in each case and compute the percentage of times our methodology selects each of the candidate latent space models.

We summarize the results of the simulation study in Figure 4, which demonstrates that our methodology is able to correctly identify the true dimension of the data-generating latent space

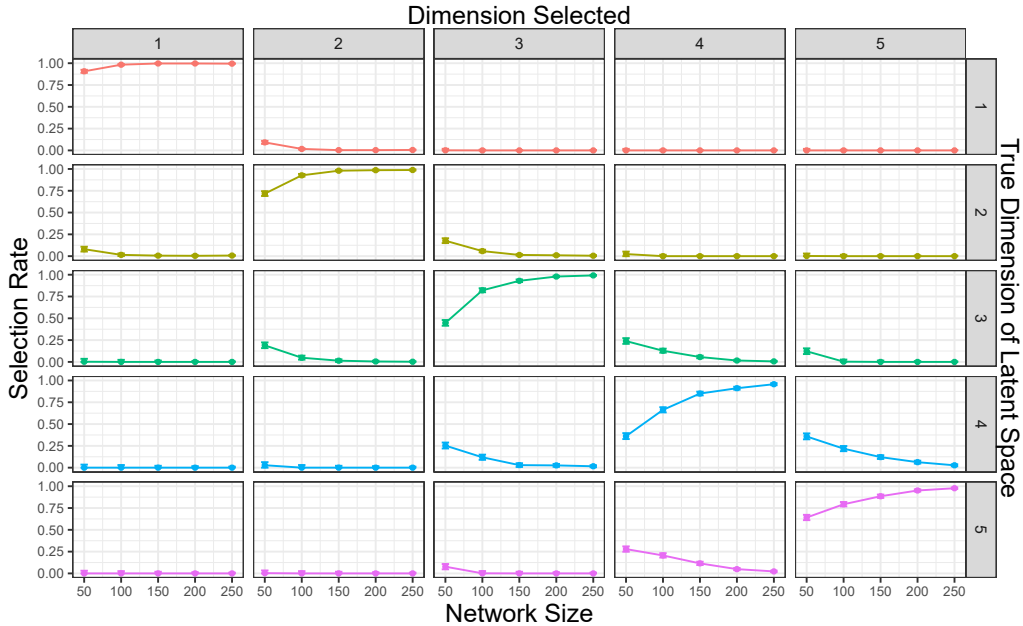


Figure 4: Results of Simulation study 3. Estimates of the correct classification rate with 95% confidence intervals for selected network sizes and across various latent space dimensions. The diagonal panels correspond to correct classification where the selection rate is desired to be highest.

model provided the network size is sufficiently large. The diagonal panels in Figure 4 correspond to correct selection of the dimension of the latent space. Of particular note, the problem becomes more challenging as the dimension of the latent space grows, but this effect is mitigated as the network size increases, with most correct selection rates in this study close to 1 for networks of size  $N = 250$ .

#### 4.4 Simulation study 4: comparing different latent mechanisms

We next study whether our proposed methodology is capable of distinguishing different latent mechanisms for edge formation in a latent position model. The first one is the same latent space model specified in (3), while the second one replaces the Euclidean distance term  $-\|z_i - z_j\|_2$  with the dot product  $z_i^t z_k$ , commonly referred to as a bilinear form. A related class of latent position models which utilize bilinear forms of latent node positions are random dot product graphs (Athreya et al., 2017). As in the previous simulation study, latent positions of nodes are randomly generated from a multivariate normal distribution with zero mean vector but this time with covariance matrix  $\sigma^2 \mathbf{I}$ , with  $\mathbf{I}$  being the identity matrix (of appropriate dimension) and  $\sigma^2 \in \{0.1, 0.2, \dots, 1.0\}$  a

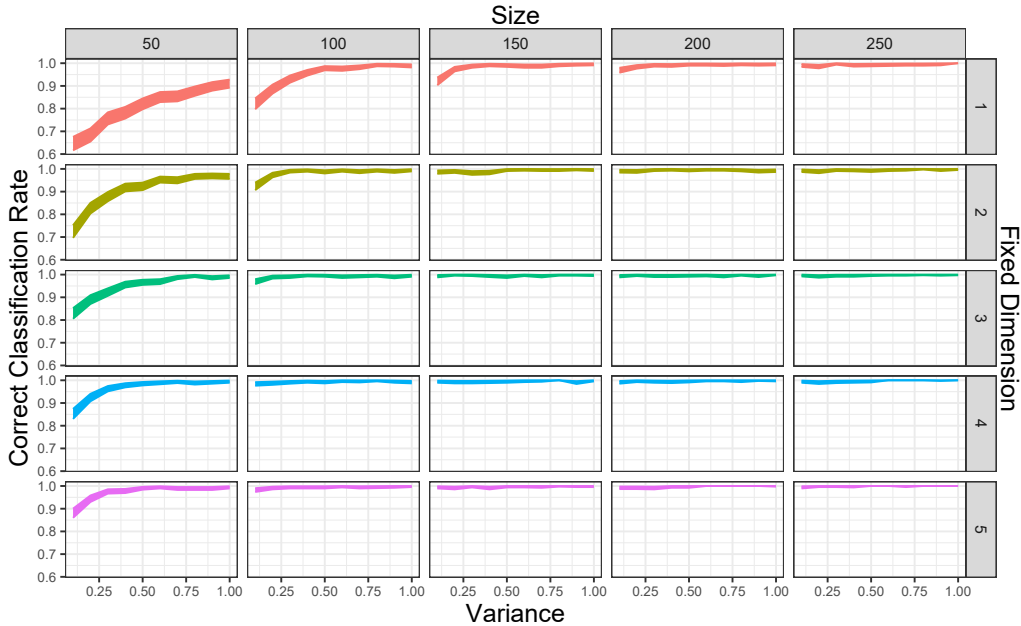


Figure 5: Results of Simulation study 4 comparing a distance based model (true model) to a similarity based model. Estimates of the correct classification rate with 95% confidence band for different networks at different sizes and across different dimensions of latent spaces.

scale factor. As the scale factor tends to zero, both models converge to a density only model so detecting the true generating process becomes more difficult. We summarize the results of the simulation study in Figure 5, which demonstrates that our methodology is able to correctly identify the true model (distance based) when compared to a bilinear (similarity based) model. Of particular note, performance improves as the dimension of the latent space increases and as the network size increases, as in the previous studies conducted.

#### 4.5 Simulation study 5: effect of the choice of classifier

In this study, we repeat Simulation study 1 using three different classifiers, XGBoost (Chen and Guestrin, 2016), Random Forest (Ho, 1995; Liaw and Wiener, 2002) and Naive Bayes (Hand and Yu, 2001; Majka, 2019). Doing so allows us to examine the effect that the choice of classifier has on the results of this simulation study, as well as to explore the relative effectiveness of each classifier in this simulation study. Figure 6 shows a similar performance for all classifiers in this simulation study, with the notable exception being the naive Bayes classifier when networks are size 25, suggesting that the choice of classifier has a weak effect on the performance of our proposed methodology, provided the network is sufficiently large. In line with conclusions in the previous simulation studies, larger network sizes result in more pronounced model signatures. In light of

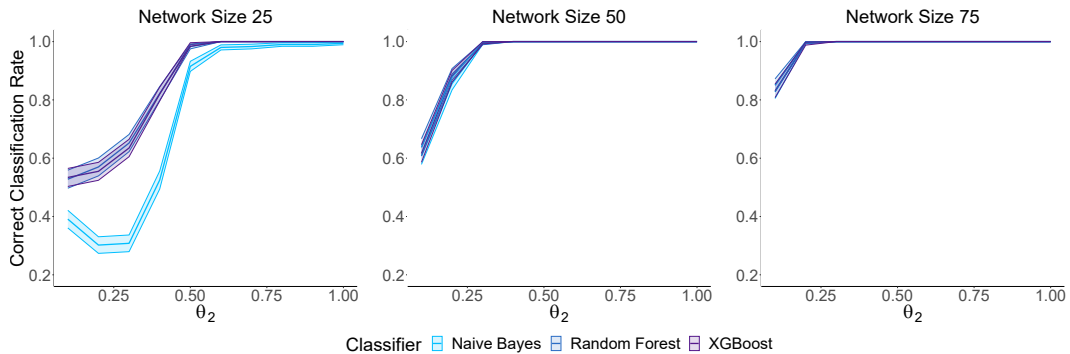


Figure 6: Results of Simulation study 5. Estimates of the correct classification rate with 95% confidence band for different classifiers.

these results, the effect of the choice of classifier appears to diminish if the model signal is sufficiently strong.

## 5. Applications

In order to study the performance of our proposed model selection methodology in applications to real-world network data, we study two network data sets which have previously been studied in the literature, in order to have a baseline for evaluating whether our methodology confirms existing results and knowledge about these networks.

### 5.1 Application 1: Sampson’s monastery network

We apply our model selection methodology to the Sampson’s monastery network data on social relationships (likeness) among 18 monk novices in a New England monastery in 1968 (Sampson, 1968). Based on the existing literature studying this network, we propose different model structures for this network which are well-designed to capture the community structure known to be a critical component of the network. In order to model this structure, stochastic block models have been applied to the network (Airoldi et al., 2008), as well as latent position models with a hierarchical group-based prior distribution structure on the latent positions (Handcock et al., 2007). We consider the following models:

- SBM:  $\mathcal{M}_1$ – $\mathcal{M}_4$  correspond to stochastic block models with  $K = 1, 2, 3, 4$  blocks ( $\mathcal{M}_1$  being equivalent to a density only model).
- LPM:  $\mathcal{M}_5$ – $\mathcal{M}_8$  correspond to latent position models with model terms for density and reciprocity and latent space dimensions  $K = 1, 2, 3, 4$ .

Model	$s_i$	$\tilde{s}_i$	Model	$s_i$	$\tilde{s}_i$
$\mathcal{M}_1$ (SBM, $K = 1$ )	0.002	0.004	$\mathcal{M}_2$ (SBM, $K = 2$ )	0.003	0.007
$\mathcal{M}_3$ (SBM, $K = 3$ )	0.032	0.077	$\mathcal{M}_4$ (SBM, $K = 4$ )	<b>0.410</b>	<b>1</b>
$\mathcal{M}_5$ (LPM, $d = 1$ )	0.028	0.068	$\mathcal{M}_6$ (LPM, $d = 2$ )	0.028	0.069
$\mathcal{M}_7$ (LPM, $d = 3$ )	0.005	0.013	$\mathcal{M}_8$ (LPM, $d = 4$ )	0.003	0.008
$\mathcal{M}_9$ (GLPM, $K = 2, d = 1$ )	0.023	0.055	$\mathcal{M}_{10}$ (GLPM, $K = 3, d = 1$ )	0.044	0.108
$\mathcal{M}_{11}$ (GLPM, $K = 4, d = 1$ )	0.043	0.104	$\mathcal{M}_{12}$ (GLPM, $K = 2, d = 2$ )	0.060	0.147
$\mathcal{M}_{13}$ (GLPM, $K = 3, d = 2$ )	0.083	0.202	$\mathcal{M}_{14}$ (GLPM, $K = 4, d = 2$ )	0.036	0.089
$\mathcal{M}_{15}$ (GLPM, $K = 2, d = 3$ )	0.020	0.050	$\mathcal{M}_{16}$ (GLPM, $K = 3, d = 3$ )	0.041	0.101
$\mathcal{M}_{17}$ (GLPM, $K = 4, d = 3$ )	0.061	0.148	$\mathcal{M}_{18}$ (GLPM, $K = 2, d = 4$ )	0.012	0.029
$\mathcal{M}_{19}$ (GLPM, $K = 3, d = 4$ )	0.030	0.074	$\mathcal{M}_{20}$ (GLPM, $K = 4, d = 4$ )	0.035	0.085

Table 2: Propensity scores  $s_i$  and normalized propensity scores  $\tilde{s}_i$  for models  $\mathcal{M}_1$ – $\mathcal{M}_{20}$  for the Sampson’s monastery network.

- GLPM:  $\mathcal{M}_9$ – $\mathcal{M}_{20}$  combine the two previous specifications by utilizing the hierarchical group-based prior distribution structure of [Handcock et al. \(2007\)](#), considering all combinations of group number  $K = 2, 3, 4$  and latent space dimension  $d = 1, 2, 3, 4$ .

Each model was fit and our model selection methodology was applied to choose a best fitting model. The latent space models were fit with [Krivitsky and Handcock \(2014\)](#) and the stochastic block models were fit with [Leger \(2016\)](#). Table 2 presents the results. The model with the highest propensity score is  $\mathcal{M}_4$ , the stochastic block model with  $K = 4$  blocks.

It has been well-established in the literature that the Sampson’s monastery network features strong community structure ([Handcock et al., 2007](#); [Airoldi et al., 2008](#)), featuring three labeled groups. However, statistical analyses have revealed the presence of a potential fourth group, evidenced in analysis which employ mixed membership stochastic block models ([Airoldi et al., 2008](#)), as well as evidence in studies which employ latent position models which suggests certain nodes may have strong connections to two or more labeled groups ([Handcock et al., 2007](#)). Within the context of the models we considered here, the choice of a stochastic block model with  $K = 4$  blocks appears to be sufficient to capture the mixing patterns of the communities as well as the reciprocity

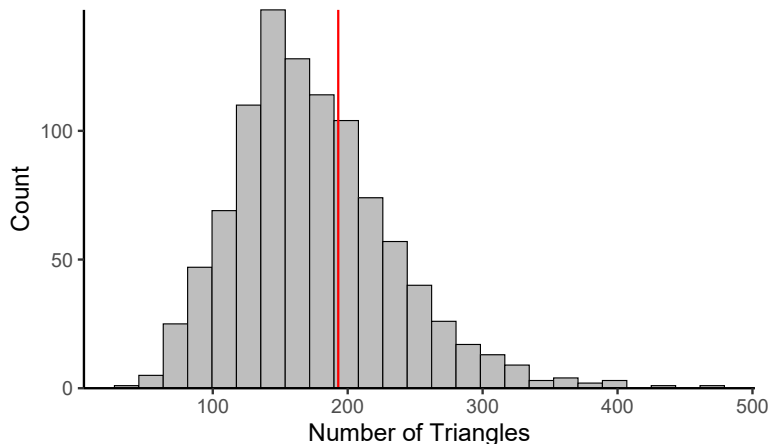


Figure 7: Fit of the observed number of triangles in the Sampson network relative to the distribution of triangles from simulated networks from  $\mathcal{M}_4$ . The observed number of triangles is indicated in red.

from the inclusion of a reciprocity term. We hold the opinion that the expression of transitivity is not sufficiently strong in this network, otherwise the latent position model with  $K = 4$  groups would potentially serve as a better model, as latent position models are able to capture network transitivity through the latent metric space. Figure 7 supports this claim by simulating networks from  $\mathcal{M}_4$  and comparing the empirical triangle count distribution of these simulated networks to the observed number of triangles in the network, demonstrating good model fit in this regard.

## 5.2 Application 2: multilevel school network

We end the section with an application to a multilevel network consisting of 6,607 third grade students over 306 classes across 176 primary schools in Poland in the 2010/2011 academic year (Maluchnik and Modzelewski, 2014). Our interest in this data set lies in the fact that it has already been extensively studied in Stewart et al. (2019), which provides the closest we can get to a data-generating model. The network contains 306 classes, but features a significant portion of non-response resulting in a large percentage of missing edge data in the network. The issues of missing data require careful consideration and are beyond the scope of this work. As such, we restrict our study in this work to the 44 classes within the multilevel network that did not feature any missing edge data. The data set employed is a directed network of 906 nodes corresponding to the individual students within the 44 classes without missing edge data, where a directed edge  $i \rightarrow j$  implies that person  $i$  stated they were friends person  $j$ . Part of the data collected included the sex of each student (recorded as male or female). This multilevel network data set naturally fits into the local dependence framework of Schweinberger and Handcock (2015), for which class based sampling is justified under the local dependence assumption (Proposition 2 & Theorem 2,

Model Term	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
Edges	✓	✓	✓	✓
Mutual	✓	✓	✓	✓
Out-degrees (1–6)	✓	✓	✓	✓
Out-degree (Female)	✓	✓	✓	✓
In-degree (Female)	✓	✓	✓	✓
Sex-match	✓	✓	✓	✓
GWESP (decay parameter fixed at 0)		✓		
GWESP (decay parameter fixed at .25)			✓	
GWESP (decay parameter estimated)				✓

Table 3: Descriptions of Models 1–4 found in [Stewart et al. \(2019\)](#).

[Schweinberger and Stewart, 2020](#)); additional details of the data set can be found in [Stewart et al. \(2019\)](#).

In this application, we study whether our proposed methodology for model selection coincides with published findings for this network by studying Models 1–4 published in [Stewart et al. \(2019\)](#), which we summarize in Table 3. The first three model terms (edges, mutual, and out-degree terms) control for structural effects within the network, including density, reciprocity, and fitting the degree distribution. The next three model terms adjust for different sex-based edge effects and homophily. The last three model terms correspond to the geometrically-weighted shared partner (GWESP) term specified in (1) that was studied in Simulation study 1. The inclusion of this model term is aimed at capturing a stochastic tendency towards network transitivity and triad formations based on values of the base parameter ( $\theta_2$  in (1)) and the decay parameter ( $\theta_3$  in (1)). Model 1 includes no GWESP term, whereas Model 2 and Model 3 fix the decay parameter at specific values found in the literature, reducing the curved exponential family to a canonical exponential family (see discussions in [Hunter \(2007\)](#) and [Stewart et al. \(2019\)](#)). Model 4 estimates the decay parameter.

We fit each of the four models  $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$  and apply our model selection methodology, which selects model  $\mathcal{M}_4$  (propensity = 0.997) above all other candidate models. This coincides with the findings of [Stewart et al. \(2019\)](#), who explored the fit of various models to the data set with respect to common-place heuristic measures ([Hunter et al., 2008](#)), as well as out-of-sample measures and through the Bayesian Information Criterion (BIC). Figure 8 demonstrates the model fit to important network features.

## References

Airoldi, E., D. Blei, S. Fienberg, and E. Xing (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research* 9, 1981–2014.



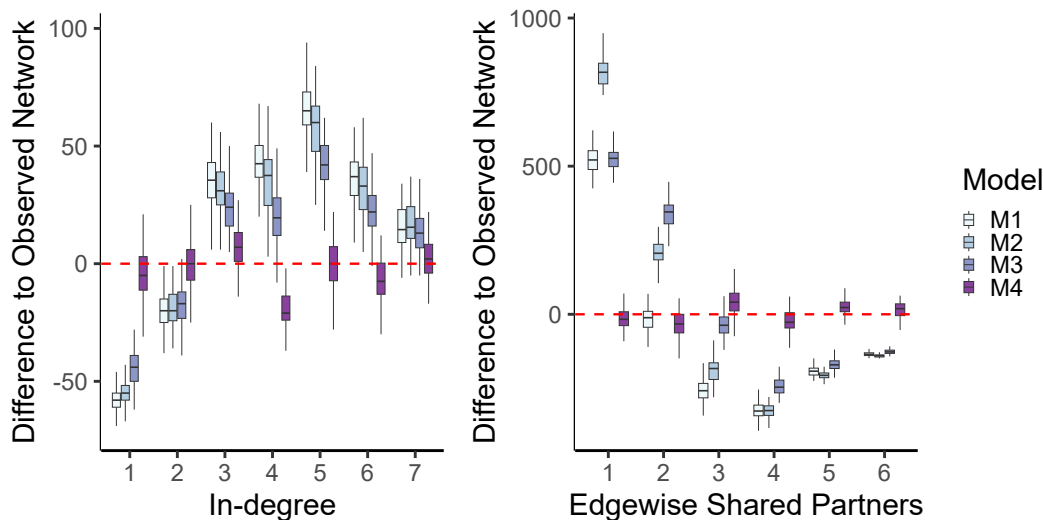


Figure 8: Difference in selected statistics between fitted models and Polish school network.

Anderson, C. J., S. Wasserman, and K. Faust (1992). Building stochastic blockmodels. *Social Networks* 14, 137–161.

Athreya, A., D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* 18(1), 8393–8484.

Barabasi, A.-L. and R. Albert (1999). Emergence of scaling in random networks. *Science* 286(5439), 509–512.

Bickel, P. J. and E. Levina (2004). Some theory for fisher’s linear discriminant function, ‘naive bayes’, and some alternatives when there are many more variables than observations. *Bernoulli* 10(6), 989–1010.

Bollobás, B. and V. Nikiforov (2004). Graphs and hermitian matrices: eigenvalue interlacing. *Discrete mathematics* 289(1-3), 119–127.

Brouwer, A. E. and W. H. Haemers (2012). *Spectra of Graphs*. Academic Press.

Chen, T. and C. Guestrin (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, New York, NY, USA, pp. 785–794. ACM.

Coutinho, J. A., T. Diviák, D. Bright, and J. Koskinen (2020). Multilevel determinants of collaboration between organised criminal groups. *Social Networks* 63, 56–69.

- Cvetković, D. M., M. Doob, and H. Sachs (1980). *Spectra of Graphs: Theory and Application*. New York: Academic Press.
- de Abreu, N. M. M. (2007). Old and new results on algebraic connectivity of graphs. *Linear Algebra and its Applications* 423(1), 53–73. Special Issue devoted to papers presented at the Aveiro Workshop on Graph Spectra.
- Donetti, L., F. Neri, and M. A. Muñoz (2006, Aug). Optimal network topologies: expanders, cages, ramanujan graphs, entangled networks and all that. *Journal of Statistical Mechanics: Theory and Experiment* 2006(08), P08007–P08007.
- Fiedler, M. (1973). Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* 23(2), 298–305.
- Finger, K. and T. Lux (2017). Network formation in the interbank money market: An application of the actor-oriented model. *Social Networks* 48, 237–249.
- Hand, D. J. and K. Yu (2001, December). Idiot’s Bayes—Not So Stupid After All? *International Statistical Review* 69(3), 385–398.
- Handcock, M. S., A. E. Raftery, and J. M. Tantrum (2007). Model-based clustering for social networks. *Journal of the Royal Statistical Society, Series A (with discussion)* 170, 301–354.
- Hastie, T., R. Tibshirani, and J. Friedman (2011). *The Elements of Statistical Learning* (2 ed.). New York: Springer-Verlag.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, M, pp. 278–282. IEEE Computer Society.
- Hoff, P. D., A. E. Raftery, and M. S. Handcock (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* 97(460), 1090–1098.
- Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). Stochastic block models: some first steps. *Social Networks* 5, 109–137.
- Hoory, S., N. Linial, and A. Wigderson (2006). Expander graphs and their applications. *Bull. Amer. Math. Soc.* 43(04), 439–562.
- Hunter, D. R. (2007). Curved exponential family models for social networks. *Social Networks* 29, 216–230.

- Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit of social network models. *Journal of the American Statistical Association* 103, 248–258.
- Hunter, D. R. and M. S. Handcock (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics* 15, 565–583.
- Jackson, M. O. and A. Watts (2002). The evolution of social and economic networks. *Journal of Economic Theory* 106(2), 265–295.
- Kolaczyk, E. D. (2009). *Statistical Analysis of Network Data: Methods and Models*. New York: Springer-Verlag.
- Krioukov, D., F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguñá (2010, Sep). Hyperbolic geometry of complex networks. *Phys. Rev. E* 82, 036106.
- Krivitsky, P. N. and M. S. Handcock (2014). *latentnet: Latent position and cluster models for statistical networks*. The Comprehensive R Archive Network. R package version 2.5.1.
- Kwok, T. C., L. C. Lau, Y. T. Lee, S. Oveis Gharan, and L. Trevisan (2013). Improved cheeger’s inequality: Analysis of spectral partitioning algorithms through higher order spectral gap. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, STOC ’13*, New York, NY, USA, pp. 11–20. Association for Computing Machinery.
- Latouche, P., E. Birmelé, and C. Ambroise (2014). Model selection in overlapping stochastic block models. *Electronic journal of statistics* 8(1), 762–794.
- Leger, J.-B. (2016). Blockmodels: A r-package for estimating in latent block model and stochastic block model, with various probability functions, with or without covariates.
- Lei, J. and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* 43(1), 215–237.
- Liaw, A. and M. Wiener (2002). Classification and regression by randomforest. *R News* 2(3), 18–22.
- Lusher, D., J. Koskinen, and G. Robins (2013). *Exponential Random Graph Models for Social Networks*. Cambridge, UK: Cambridge University Press.
- Majka, M. (2019). *naivebayes: High Performance Implementation of the Naive Bayes Algorithm in R*. The Comprehensive R Archive Network. R package version 0.9.7.

- Maluchnik, M. and M. Modzelewski (2014). Próba badawcza i proces zbierania danych. In R. Dolata (Ed.), *Czy szkoła ma znaczenie? Zróżnicowanie wyników nauczania po pierwszym etapie edukacyjnym oraz jego pozaszkolne i szkolne uwarunkowania*, Volume 1. Warsaw: Instytut Badań Edukacyjnych.
- Morris, M. (2004). *Network epidemiology: A handbook for survey design and data collection*. Oxford University Press on Demand.
- Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical review E* 74(3), 036104.
- Obando, C. and F. de Vico Fallani (2017). A statistical model for brain networks inferred from large-scale electrophysiological signals. *Journal of The Royal Society Interface* 14(128), 20160940.
- Ryan, C., J. Wyse, and N. Friel (2017). Bayesian model selection for the latent position cluster model for social networks. *Network Science* 5(1), 70–91.
- Sampson, S. (1968). *A novitiate in a period of change: An experimental and case study of relationships*. Ph. D. thesis, Department of Sociology, Cornell University.
- Schweinberger, M. (2011). Instability, sensitivity, and degeneracy of discrete exponential families. *Journal of the American Statistical Association* 106(496), 1361–1370.
- Schweinberger, M. and M. S. Handcock (2015). Local dependence in random graph models: characterization, properties and statistical inference. *Journal of the Royal Statistical Society, Series B* 77, 647–676.
- Schweinberger, M., P. N. Krivitsky, C. T. Butts, and J. Stewart (2020). Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios. *Statistical Science* 35, 627–662.
- Schweinberger, M. and T. A. B. Snijders (2003). Settings in social networks: a measurement model. In R. M. Stolzenberg (Ed.), *Sociological Methodology*, Volume 33, Chapter 10, pp. 307–341. Boston & Oxford: Basil Blackwell.
- Schweinberger, M. and J. Stewart (2020). Concentration and consistency results for canonical and curved exponential-family models of random graphs. *The Annals of Statistics* 48, 374–396.
- Sewell, D. K. and Y. Chen (2015). Latent space models for dynamic networks. *Journal of the American Statistical Association* 110, 1646–1657.
- Shi, J. and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(8), 888–905.

- Shore, J. and B. Lubin (2015). Spectral goodness of fit for network models. *Social Networks* 43, 16–27.
- Smith, A. L., D. M. Asta, and C. A. Calder (2019). The geometry of continuous latent space models for network data. *Statistical Science* 34(3), 428 – 453.
- Snijders, T. A., G. G. van de Bunt, and C. E. Steglich (2010). Introduction to stochastic actor-based models for network dynamics. *Social Networks* 32(1), 44–60.
- Snijders, T. A. B., P. E. Pattison, G. L. Robins, and M. S. Handcock (2006). New specifications for exponential random graph models. *Sociological Methodology* 36, 99–153.
- Stewart, J., M. Schweinberger, M. Bojanowski, and M. Morris (2019). Multilevel network data facilitate statistical inference for curved ERGMs with geometrically weighted terms. *Social Networks* 59, 98–119.
- Stivala, A. and A. Lomi (2021). Testing biological network motif significance with exponential random graph models. *Applied Network Science* 6(1), 1–27.
- Tang, M., D. L. Sussman, and C. E. Priebe (2013). Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics* 41, 1406–1430.
- Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing* 17(4), 395–416.
- Wang, Y. X. R. and P. J. Bickel (2017). Likelihood-based model selection for stochastic block models. *The Annals of Statistics* 45(2), 500–528.
- Zeng, R., Q. Z. Sheng, L. Yao, T. Xu, and D. Xie (2013). A practical simulation method for social networks. In *Proceedings of the First Australasian Web Conference-Volume 144*, pp. 27–34.