TECHNOMETRICS
2020, VOL. 62, NO. 2, 161–172
https://doi.org/10.1080/00401706.2019.1623076

Taylor & Francis
Taylor & Francis Group

Model-Based Clustering of Nonparametric Weighted Networks With Application to Water Pollution Analysis

Amal Agarwal and Lingzhou Xue

Department of Statistics, Pennsylvania State University, University Park, PA

ABSTRACT

Water pollution is a major global environmental problem, and it poses a great environmental risk to public health and biological diversity. This work is motivated by assessing the potential environmental threat of coal mining through increased sulfate concentrations in river networks, which do not belong to any simple parametric distribution. However, existing network models mainly focus on binary or discrete networks and weighted networks with known parametric weight distributions. We propose a principled nonparametric weighted network model based on exponential-family random graph models and local likelihood estimation, and study its model-based clustering with application to large-scale water pollution network analysis. We do not require any parametric distribution assumption on network weights. The proposed method greatly extends the methodology and applicability of statistical network models. Furthermore, it is scalable to large and complex networks in large-scale environmental studies. The power of our proposed methods is demonstrated in simulation studies and a real application to sulfate pollution network analysis in Ohio watershed located in Pennsylvania, United States.

ARTICLE HISTORY
Received February 2018
Accepted May 2019

KEYWORDS
Environmental studies; Exponential-family random graphical model; Local likelihood; Variational inference

1. Introduction

Water pollution is the leading cause of deaths and diseases, and it is a major global problem. It is known that nearly 80% of the world's population lives in areas exposed to high levels of threat to water security (Vörösmarty et al. 2010). The recent national report on water quality by EPA (2017) pointed out that 46% of rivers, 21% of lakes, 18% of coastal waters, and 32% of wetlands in the United States are in poor biological condition or rated poor based on a water quality index. The major pollutant sources include agriculture, atmospheric deposition, construction, industrial production, municipal sewage, resource extraction, spills, and urban runoff. They pose severe health hazards like cancer, cardiovascular, respiratory, neurologic, and developmental damage. This work is motivated by assessing the potential environmental threat of coal mining in Ohio watershed of Pennsylvania through increased sulfate concentrations in the surface water, which is an important scientific problem in geoscience. Bernhardt et al. (2012) mapped surface coal mining of southern West Virginia and linked these maps with water quality and biological data of 223 streams. When these coal mines occupy $>5.4\%$ of their contributing watershed area, the sulfate concentrations within catchments could exceed 50 mg/L (Niu et al. 2018). Residential proximity to heavy coal production is associated with higher risk for cardiopulmonary disease, chronic lung disease, hypertension, and kidney disease (Hendryx and Ahern 2008). The study of water-quality risks will help the whole society manage them now and in future.

With advances in data collection, there are more and more modern statistical research on environmental studies using nonparametric regression, causal inference, mixture model, network analysis, and variable selection (e.g., Ebenstein 2012; Liang et al. 2015; Li et al. 2017; Lin et al. 2017; Wen et al. 2019, among many others). Especially, network analysis becomes increasingly important in large-scale environmental studies and geoscientific research to assess environmental impacts and risks for water pollution (Smith, Alexander, and Wolman 1987; Lienert, Schnetzer, and Ingold 2013; Ruzol et al. 2017). For example, Gianessi and Peskin (1981) proposed a water network model to explore the impact of cropland sediment controls on improved water quality, and Montgomery (1972) and Anastasiadis et al. (2016) studied weighted pollution networks where the weights measure the pollution diminishing transition. However, none of aforementioned network models took into account the spatial heterogeneity and the hub structure of river networks. Without exploring spatial heterogeneity, these models could fail to differentiate polluted regions from less polluted but well connected regions in river networks. The hubs in river networks usually determine the flow of pollutants, and they may help identify polluted and well connected regions.

In this work, we introduce a principled model-based clustering of networks to effectively deal with the spatial heterogeneity of river pollution and efficiently identify the hub structure in river networks. Model-based clustering of networks based on stochastic block models (SBMs) and exponential-family random graph models (ERGMs) have received considerable attention in recent literature, including Snijders and Nowicki (1997), Nowicki and Snijders (2001), Girvan and Newman (2002), Airoldi et al. (2008), Karrer and Newman (2011), Zhao,

CONTACT Lingzhou Xue *lzxue@psu.edu* Department of Statistics, Pennsylvania State University, University Park, PA 16802.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/TECH*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TECH*.

Levina, and Zhu (2012), Vu, Hunter, and Schweinberger (2013), Saldana, Yu, and Feng (2017), Wang and Bickel (2017), Lee, Xue, and Hunter (2017), among many others. It is worth pointing out that existing research mainly focuses on the model-based clustering of networks with binary or discrete edges. In a recent paper by Ambroise and Matias (2012), parametric distributions are incorporated into a stochastic block model to model continuous network weights. Alternatively, Bayesian variational methods were proposed by Aicher, Jacobs, and Clauset (2014) to approximate posterior distribution of weights over latent block structures. However, from our motivating example, sulfate concentrations in the surface river network do not belong to any simple parametric distribution, which will be illustrated in Section 5. Thus, we need to relax the parametric assumption in network models to account for the unknown distribution of continuous network weights. To address this issue, we propose a new nonparametric weighted network model based on ERGMs and local likelihood estimation, and study its model-based clustering with application to large-scale water pollution network analysis. The proposed method greatly extends the methodology and applicability of statistical network models. Furthermore, it is scalable to large and complex networks in real-world applications.

The rest of this article is organized as follows. Section 2 presents the methodology of our proposed nonparametric weighted network model. In Section 3, we introduce a novel variational expectation-maximization (EM) algorithm to solve the approximate maximum likelihood estimation. Section 4 demonstrates the numerical performances of our proposed methods and algorithms in simulation studies. In Section 5, we apply the proposed method to analyze the large-scale water pollution network of sulfate concentrations in the Ohio watershed of Pennsylvania.

2. Methodology

We define some necessary notation before presenting our proposed method. Let n be the number of nodes in the observed network. Let $\mathbf{Y} = (Y_{ij})_{1\le i,j\le n}$ be the corresponding weighted network such that $Y_{ij} = (E_{ij}, W_{ij})$, where E_{ij} is a binary indicator denoting the existence of an edge in the network and W_{ij} is the corresponding weight when $E_{ij} = 1$. The weight matrix $W = (W_{ij})_{1\le i,j\le n}$ consists of continuous weights in the network. Further we assume that the edge distribution of each network belongs to an exponential family (Besag 1974; Frank and Strauss 1986). We write the distribution of edge indicator matrix $\boldsymbol{E}$ as

$$P_{\boldsymbol{\theta}}(\boldsymbol{E} = \boldsymbol{e}) = \exp\{\boldsymbol{\theta}'\boldsymbol{g}(\boldsymbol{e}) - \psi(\boldsymbol{\theta})\}, \tag{1}$$

where $\psi(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{e}\in\mathcal{E}} \exp\left[\boldsymbol{\theta}'\boldsymbol{g}(\boldsymbol{e})\right]$ is the log of the normalizing constant, $\boldsymbol{\theta} \in \mathbb{R}^p$ are canonical network parameters of interest, and $\boldsymbol{g} : \mathcal{E} \to \mathbb{R}^p$ is the sufficient statistic. Here $\mathcal{E}$ is the space for $\boldsymbol{E}$ consisting of 2^n possible binary edge structures.

One of the major limitations of this binary network model is that it cannot deal with large number of nodes due to large computational time for evaluating the likelihood function. For undirected networks, this computing time scales with node size as $\exp((n(n-1)\log 2)/2)$. Many estimation algorithms have been developed (Snijders 2002; Hunter and Handcock 2006; Møller et al. 2006; Koskinen, Robins, and Pattison 2010; Caimo and Friel 2011), however, most of them are time-consuming and therefore unrealistic for fitting large networks. This issue of non-scalability can be resolved by the assumption of dyadic independence which assumes that all dyads are independent of each other. Note that dyad is a general term applicable for both directed and undirected edges. In the undirected weighted networks, it would imply the ties and weights are independent of each other and for all pairs of nodes, that is,

$$P_{\boldsymbol{\theta}}(\boldsymbol{Y} = \boldsymbol{y}) = \prod_{1\le i<j\le n} \exp\{\boldsymbol{\theta}'\boldsymbol{g}(e_{ij}) - \psi(\boldsymbol{\theta})\} P_{\boldsymbol{\theta}}(W_{ij} = w_{ij}). \tag{2}$$

This assumption facilitates both estimation and simulation of large networks as well as solves the issue of degeneracy (Strauss 1986; Handcock et al. 2003; Schweinberger 2011; Krivitsky 2012). However, dyadic independence is too restrictive and most models following this assumption are either very trivial, failing to capture relational dependencies (Gilbert 1959; Erdős and Rényi 1959) or non-parsimonious, with a large number of parameters (Holland and Leinhardt 1981).

We consider a model-based clustering framework to relax the dyadic independence. More specifically, we introduce the finite K-component mixture form together with a much less restrictive assumption of *conditional dyadic independence* (CDI) (Snijders and Nowicki 1997; Nowicki and Snijders 2001; Girvan and Newman 2002; Vu, Hunter, and Schweinberger 2013). Under this assumption, we propose the nonparametric weighted network as

$$P_{\boldsymbol{\theta}\boldsymbol{f}}(\mathbf{Y} = \mathbf{y}|\mathbf{Z} = \mathbf{z}) = \prod_{1\le i<j\le n} P_{\boldsymbol{\theta}_{z_iz_j} f_{z_iz_j}}(Y_{ij} = y_{ij}|\boldsymbol{Z} = \boldsymbol{z}), \tag{3}$$

where $\mathbf{Z} = (\boldsymbol{Z}_i)_{1\le i\le n}$ denote the latent cluster memberships of nodes. Here, $\boldsymbol{Z}_i$ is a $K \times 1$ vector such that $z_{ik} = 1$ if and only if node i lies in cluster k, otherwise $z_{ik} = 0$. The conditional dyadic independence strikes in a nice balance between model complexity and parsimony for model-based clustering. The number of parameters are reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(K^2)$, thus enabling simple inter and intra community interpretations. The CDI assumption induces a block structure similar to SBMs. SBMs have been thoroughly studied in the context of social networks (Holland, Laskey, and Leinhardt 1983; Airoldi et al. 2008) and have a long history in multiple scientific communities (Bui et al. 1987; Dyer and Frieze 1989; Bollobás, Janson, and Riordan 2007). We omit discussion of SBMs here except to point out a major difference from our current setup of ERGM. ERGMs can allow several kinds of dynamic network statistics like density, stability, transitivity (Hanneke, Fu, and Xing 2010), thus effectively generalizing the simple density case in SBM methodology which makes them appealing in practice. To effectively model continuous weights, for any given pair of nodes (i,j), we have

$$P_{\boldsymbol{\theta}\boldsymbol{f}}(Y_{ij} = y_{ij}|\boldsymbol{Z} = \boldsymbol{z}) = \left(p_{z_iz_j} f_{z_iz_j}(w_{ij})\right)^{\mathbb{1}_{e_{ij}\neq 0}} \left(1 - p_{z_iz_j}\right)^{\mathbb{1}_{e_{ij}=0}}. \tag{4}$$

Here, $(p_{kl})_{1\le k,l\le K} = (P_{\boldsymbol{\theta}_{z_iz_j}}(E_{ij} = e_{ij}|\mathbf{Z} = \mathbf{z}))_{1\le k,l\le K}$ take the parametric specification of exponential-family distributions of

the network statistics as explained in (1) and network weights $(w_{ij})_{1\le i,j\le n:z_i=k,z_j=l}$ are assumed to be an iid sample observed from a population following an univariate nonparametric density function f_{kl}. Note that $(f_{kl})_{1\le k,l\le K}$ do not necessarily have any parametric form. We implicitly assume that conditioning on the full $Z = z$ is same as conditioning on just z_i and z_j, that is, E_{ij} and W_{ij} depend on Z only via z_i and z_j. For ease of presentation, we assume the additive structure of p_{kl} to be parameterized by network sparsity parameters $\boldsymbol{\theta}$ as $p_{kl} = \text{logit}^{-1}(\theta_k + \theta_l)$. It is worth pointing out that the clusters $z_1, \ldots, z_n$ are determined by two sources of information in network models. On the one hand, the random block structure and also the additive structure of $(p_{kl})_{1\le k,l\le K}$ contribute to the exploration of different degrees among the clusters. On the other hand, the nonparametric network weights modulate the separation between clusters with different weight distributions.

Combining (3) and (4), the corresponding log-likelihood function given the cluster memberships can be written as

$$\begin{aligned}&\log(P_{\boldsymbol{\theta},\boldsymbol{f}}(\mathbf{Y}=\mathbf{y}\mid\mathbf{Z}=\mathbf{z}))\\&=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Big\{\big[\mathbb{1}_{e_{ij}\neq 0}\log(p_{z_iz_j})+\mathbb{1}_{e_{ij}=0}\log(1-p_{z_iz_j})\big]\\&+\mathbb{1}_{e_{ij}\neq 0}\Big[\log(f_{z_iz_j}(w_{ij}))-\Big(\int_{\mathcal{X}}f_{z_iz_j}(u)du-1\Big)\Big]\Big\},\end{aligned}\tag{5}$$

where we also introduce the penalty term $\left(\int_{\mathcal{X}} f_{kl}(u)du - 1\right)$. Thus, (5) can be treated as a likelihood for any nonnegative function f_{kl} without imposing the additional constraint $\int_{\mathcal{X}} f_{kl}(u)du = 1$. This specification follows the similar spirit of Loader (1996).

Now we derive the localized version of the conditional log-likelihood evaluated at an arbirary grid point w as

$$\begin{aligned}&\ell(\boldsymbol{\theta},\boldsymbol{f},w;\boldsymbol{Y}\mid\boldsymbol{Z})\\&=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Big\{\big[\mathbb{1}_{e_{ij}\neq 0}\log(p_{z_iz_j})+\mathbb{1}_{e_{ij}=0}\log(1-p_{z_iz_j})\big]\\&+\mathbb{1}_{e_{ij}\neq 0}\times\\&\Big[K_h(w_{ij}-w)\log(f_{z_iz_j}(w_{ij}))\\&-\Big(\int_{\mathcal{X}}K_h(u-w)f_{z_iz_j}(u)du-1\Big)\Big]\Big\},\end{aligned}\tag{6}$$

where K_h is the rescaled kernel function with a positive bandwidth h. We approximate $\log(f_{kl}(u))$ by Φ^p_{kl}, a linear combination of orthogonal basis functions $(\phi_m)_{1\le m\le p}$, namely, $\Phi^p_{kl}(u-w) = \sum_{m=0}^{p}\beta_m^{(kl)}\phi_m(u-w)$. With this approximation, the local conditional log-likelihood becomes

$$\begin{aligned}&\ell(\boldsymbol{\theta},\boldsymbol{\beta},w;\boldsymbol{Y}\mid\boldsymbol{Z})\\&=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\Big\{\big[\mathbb{1}_{e_{ij}\neq 0}\log(p_{z_iz_j})+\mathbb{1}_{e_{ij}=0}\log(1-p_{z_iz_j})\big]\\&+\mathbb{1}_{e_{ij}\neq 0}\times\\&\Big[K_h(w_{ij}-w)\Phi^p_{z_iz_j}(w_{ij}-w)-\Big(\int_{\mathcal{X}}K_h(u-w)\exp\\&(\Phi^p_{z_iz_j}(u-w))du-1\Big)\Big]\Big\}.\end{aligned}\tag{7}$$

We assume that membership indicators $\mathbf{Z} = (Z_i)_{1\le i\le n}$ follow a multinomial distribution with a single trial and mixture proportions as $\pi = (\pi_k)_{1\le k\le K}$. The log-likelihood of the observed weighted network can be written as

$$\ell(\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{f}) = \log\Big(\sum_{\mathbf{z}\in\{1,\ldots,K\}^n} P_{\boldsymbol{\theta},\boldsymbol{f}}(\mathbf{Y}=\mathbf{y}\mid\mathbf{Z}=\mathbf{z})P_{\boldsymbol{\pi}}(\mathbf{Z}=\mathbf{z})\Big).\tag{8}$$

In view of (5) and (7), we maximize the log-likelihood function (8) of the observed network described to estimate model parameters $\boldsymbol{\theta}$ together with block densities $\boldsymbol{f}$.

Remark 1. Allman, Matias, and Rhodes (2011) proved the identifiability of parameters for a broad class of parametric or nonparametric weighted network models. As shown in Section 4 of Allman, Matias, and Rhodes (2011), we can uniquely identify the parameters under mild conditions for parametric weighted networks while the identifiability result for nonparametric weighted networks depend on binning the values of the edge variables into a finite set. Please see Theorem 15 of Allman, Matias, and Rhodes (2011) for more details about the identifiability.

Remark 2. The proposed nonparametric weighted network model can be further extended to discrete temporally evolving weighted networks. Given the dynamic network series $\mathbf{Y} = (\mathbf{Y}_t)_{1\le t\le T} = (\mathbf{E}_t, \mathbf{W}_t)_{1\le t\le T}$, the dynamic nonparametric weighted network model can be derived by assuming a discrete-time Markov structure over the time (Hanneke, Fu, and Xing 2010; Krivitsky and Handcock 2014; Kim et al. 2018), that is,

$$P_{\boldsymbol{\theta},\boldsymbol{f}}(\mathbf{Y}=\mathbf{y}|\mathbf{Z}=\mathbf{z}) = \prod_{2\le t\le T} P_{\boldsymbol{\theta},\boldsymbol{f}}(\mathbf{Y}_t=\mathbf{y}_t|\mathbf{Y}_{t-1}=\mathbf{y}_{t-1},\mathbf{Z}=\mathbf{z}),$$

where $P_{\boldsymbol{\theta},\boldsymbol{f}}(\mathbf{Y}_t = \mathbf{y}_t|\mathbf{Y}_{t-1} = \mathbf{y}_{t-1},\mathbf{Z} = \mathbf{z})$ follows the similar specification of (3) and (4). We can incorporate dynamic network statistics such as stability and transitivity in the exponential-family specification of $P_{\boldsymbol{\theta}_{z_iz_j}}(E_{t,ij} = e_{t,ij} \mid E_{t-1,ij} = e_{t-1,ij},\mathbf{Z} = \mathbf{z})$. Hence, the interpretation of clusters will reflect the impact from dynamic network statistics. For example, the use of stability network statistic will contribute to the exploration of different levels of stability among the clusters. The dynamic nonparametric weighted network is beyond our current scope and we will study it in the future.

3. Computation

This section proposes a variational EM algorithm to approximately solve the maximum likelihood estimation. It is infeasible to directly maximize the log-likelihood function (8) due to two key challenges: (i) exponential-family form of $P(\boldsymbol{Y} \mid \boldsymbol{Z})$ is not scalable for large networks; (ii) the sum is over every possible assignment to $\boldsymbol{Z}$, where for each $1 \le i \le n$, $Z_i = z_i$ can take one of K possible values.

To resolve the first challenge, the CDI assumption (3) plays a crucial role. Typically parameters in a mixture model are estimated using the classical EM algorithm (Dempster, Laird, and Rubin 1977). The E-step proceeds by writing the complete

data log-likelihood (9), assuming the network is observed while node membership indicators $\boldsymbol{Z}$ are unobserved.

$$\begin{aligned}&\log(P_{\boldsymbol{\theta},\boldsymbol{\pi},f}(\mathbf{Y}=\mathbf{y},\mathbf{Z}=\mathbf{z}))\\&=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}z_{ik}z_{jl}\log P_\theta(Y_{ij}=y_{ij}\mid\mathbf{Z}=\mathbf{z})\\&\quad+\sum_{i=1}^{n}\sum_{k=1}^{K}z_{ik}\log(\pi_k).\end{aligned}\tag{9}$$

Next we take expectation of this complete log-likelihood with the distribution $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Z}\mid\boldsymbol{Y})$. Clearly the distribution $\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Z}\mid\boldsymbol{Y})=\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{Z}_{i,1\le i\le n}\mid\boldsymbol{Y})$ cannot be further factored over nodes since for $i\neq j$, Z_i is not independent of Z_j given the observed network $\boldsymbol{Y}$. This poses a huge computational challenge. The intractability of the complete log-likelihood motivates the use of variational approximation. Variational methods are well studied in literature (Blei, Kucukelbir, and McAuliffe 2017). The basic idea is to posit a tractable auxiliary distribution $A_\gamma(\boldsymbol{z})\equiv P(\boldsymbol{Z}=\boldsymbol{z})$ for the latent variables $\boldsymbol{Z}$ and find the optimal setting for variational parameters γ that minimizes the Kullback–Liebler divergence between the approximation and true distribution. We use this auxiliary distribution to construct a tractable lower bound of the log-likelihood using Jensen's inequality and then maximize this lower bound, yielding approximate maximum likelihood estimates.

$$\begin{aligned}\ell(\boldsymbol{\theta},\boldsymbol{\pi},\boldsymbol{f})&=\log\left(\sum_{\mathbf{z}\in\{1,\dots,K\}^n}\frac{P_{\boldsymbol{\theta},f}(\mathbf{Y}=\mathbf{y}|\mathbf{z})P_{\boldsymbol{\pi}}(\mathbf{Z}=\mathbf{z})}{A_{\boldsymbol{\gamma}}(\mathbf{z})}A_{\boldsymbol{\gamma}}(\mathbf{z})\right)\\&\ge\sum_{\mathbf{z}\in\{1,\dots,K\}^n}\log\left(\frac{P_{\boldsymbol{\theta},f}(\mathbf{Y}=\mathbf{y}\mid\mathbf{z})P_{\boldsymbol{\pi}}(\mathbf{Z}=\mathbf{z})}{A_{\boldsymbol{\gamma}}(\mathbf{z})}\right)A_{\boldsymbol{\gamma}}(\mathbf{z})\\&=\text{ELBO}\,(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{f}).\end{aligned}\tag{10}$$

The derivation in (10) uses an auxiliary distribution and also Jensen's inequality. We choose the variational distribution $A(\mathbf{Z})$ from the mean-field family as,

$$A_{\boldsymbol{\gamma}}(\mathbf{Z})=\prod_{i=1}^{n}P_{\boldsymbol{\gamma}_i}(\mathbf{Z}_i)=\prod_{i=1}^{n}\prod_{k=1}^{K}\gamma_{ik}^{z_{ik}},\tag{11}$$

where $\forall\, i\in\{1,\dots,n\}$, $k\in\{1,\dots,K\}$, we have $0\le\gamma_{ik}\le 1$ with the constraint $\sum_{k=1}^{K}\gamma_{ik}=1$ and $z_{ik}=\mathbb{1}_{z_i=k}$. This class of probability distributions $A_{\boldsymbol{\gamma}}$ considers independent laws through different nodal memberships. With the definition of $A(\mathbf{Z})$ in (11), we derive the following effective lower bound (ELBO).

$$\begin{aligned}&\text{ELBO}\,(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{f})\\&=\sum_{i=1}^{N}\sum_{j=i+1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big\{\gamma_{ik}\gamma_{jl}\big[\mathbb{1}_{e_{ij}\neq0}\log(p_{kl})+\mathbb{1}_{e_{ij}=0}\log(1-p_{kl})\big]\\&\quad+\mathbb{1}_{e_{ij}\neq0}\Big[\log(f_{kl}(w_{ij}))-\Big(\int_{\mathcal{X}}f_{kl}(u)du-1\Big)\Big]\Big\}\\&\quad+\sum_{i=1}^{N}\sum_{k=1}^{K}\gamma_{ik}\left[\log(\pi_k)-\log(\gamma_{ik})\right].\end{aligned}\tag{12}$$

Variational E-step: We maximize ELBO $(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\pi},\boldsymbol{f})$ in (12) to obtain $\boldsymbol{\gamma}^{(t)}$

$$\boldsymbol{\gamma}^{(t)}=\underset{\boldsymbol{\gamma}}{\arg\max}\ \text{ELBO}\,(\boldsymbol{\theta}^{(t-1)},\boldsymbol{\gamma},\boldsymbol{\pi}^{(t-1)},\boldsymbol{f}^{(t-1)}).\tag{13}$$

The direct maximization of ELBO in (13) is difficult, since the lower bound depends on the products $\gamma_{ik}\gamma_{jl}$ and, therefore, the fixed-point updates of γ_{ik} depend on all other γ_{jl} (Daudin, Picard, and Robin 2008). To separate the parameters in this maximization problem, we adopt an MM algorithm that involves constructing a surrogate (minorizing) function and optimizing it iteratively (Hunter and Lange 2004). The surrogate function Q must satisfy the following properties to qualify as a valid minorizing function

$$Q(\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\pi}^{(t)},\boldsymbol{f}^{(t)};\boldsymbol{\gamma})\le\text{ELBO}(\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma},\boldsymbol{\pi}^{(t)},\boldsymbol{f}^{(t)})\quad,\forall\boldsymbol{\gamma},\tag{14}$$

$$Q(\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\pi}^{(t)},\boldsymbol{f}^{(t)};\boldsymbol{\gamma}^{(t)})=\text{ELBO}(\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\pi}^{(t)},\boldsymbol{f}^{(t)}).\tag{15}$$

First we note that for all $(\theta_{kl})_{1\le k\le l\le K}$ we have $\log(p_{kl})<0$ and $\log(1-p_{kl})<0$ which gives rise to following inequalities using the arithmetic geometric mean inequality

$$\gamma_{ik}\gamma_{jl}\log(p_{kl})\ge\left(\gamma_{ik}^2\frac{\hat\gamma_{jl}}{2\hat\gamma_{ik}}+\gamma_{jl}^2\frac{\hat\gamma_{ik}}{2\hat\gamma_{jl}}\right)\log(p_{kl}),\tag{16}$$

$$\gamma_{ik}\gamma_{jl}\log(1-p_{kl})\ge\left(\gamma_{ik}^2\frac{\hat\gamma_{jl}}{2\hat\gamma_{ik}}+\gamma_{jl}^2\frac{\hat\gamma_{ik}}{2\hat\gamma_{jl}}\right)\log(1-p_{kl}),\tag{17}$$

with equality if $\gamma_{ik}=\hat\gamma_{ik}$ and $\gamma_{jl}=\hat\gamma_{jl}$. Also the concavity of the logarithm function gives rise to following inequality (Vu, Hunter, and Schweinberger 2013)

$$-\log(\gamma_{ik})\ge-\log(\hat\gamma_{ik})-\frac{\gamma_{ik}}{\hat\gamma_{ik}}+1.\tag{18}$$

We construct the surrogate function that satisfies (14) and (15) using the inequalities (16), (17), and (18), thus guaranteeing the ascent property of ELBO.

$$\begin{aligned}&\text{Q}\,(\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t-1)},\boldsymbol{\pi}^{(t)},\boldsymbol{f}^{(t)};\boldsymbol{\gamma})\\&=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\Bigg\{\left(\gamma_{ik}^2\frac{\gamma_{jl}^{(t-1)}}{2\gamma_{ik}^{(t-1)}}+\gamma_{jl}^2\frac{\gamma_{ik}^{(t-1)}}{2\gamma_{jl}^{(t-1)}}\right)\\&\quad\Big[\mathbb{1}_{e_{ij}\neq0}\log(p_{kl}^{(t)})+\\&\quad\mathbb{1}_{e_{ij}=0}\log(1-p_{kl}^{(t)})\Big]+\mathbb{1}_{e_{ij}\neq0}\Big[\log(f_{kl}^{(t)}(w_{ij}))\\&\quad-\Big(\int_{\mathcal{X}}f_{kl}^{(t)}(u)du-1\Big)\Big]\Bigg\}\\&\quad+\sum_{i=1}^{n}\sum_{k=1}^{K}\gamma_{ik}\left[\log(\pi_k^{(t)})-\log(\gamma_{ik}^{(t-1)})-\frac{\gamma_{ik}}{\gamma_{ik}^{(t-1)}}+1\right].\end{aligned}\tag{19}$$

To maximize (19), we solve n separate quadratic programming problems of K variables $\boldsymbol{\gamma}_i$ under the constraints $\gamma_{ik}\ge0,\forall\,k\in\{1,\dots,K\}$ together with $\sum_{k=1}^{K}\gamma_{ik}=1$.

M-step: We first maximize the (12) with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$. We have a closed-form update for $\boldsymbol{\pi}$ as $\pi_k^{(t)}=\sum_{i=1}^{n}\gamma_{ik}^{(t)}/n$.

To update $\boldsymbol{\theta}$, we maximize (12) using the modified Newton–Raphson method with line search to guarantee the ascent property (Dennis and Schnabel 1996). Now, it remains to update block densities $\boldsymbol{f}$. To this end, we use (6) and the approximation of $\log(f_{kl}(.))$. For simplicity, we use a local polynomial approximation such that $\log(f_{kl}(.))$ can be approximated by a low-degree polynomial ζ_{kl} in a neighborhood of the fitting point w (Loader 1996): $\log(f_{kl}(u)) \approx \zeta_{kl}(u-w) = \sum_{m=0}^{p} \beta_m^{(kl)}(u-w)^m$. With this approximation, we rewrite (7) as

$$
\begin{aligned}
&\ell(\boldsymbol{\theta}, \boldsymbol{\beta}, w; \boldsymbol{Y} \mid \boldsymbol{Z}) \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \Big\{ \left[\mathbb{1}_{e_{ij}\neq 0} \log(p_{z_i z_j}) + \mathbb{1}_{e_{ij}=0} \log(1-p_{z_i z_j}) \right] \\
&\quad + \mathbb{1}_{e_{ij}\neq 0} \Big[K_h(w_{ij}-w) \\
&\quad \zeta_{z_i z_j}(w_{ij}-w) - \Big(\int_{\mathcal{X}} K_h(u-w) \exp(\zeta_{z_i z_j}(u-w)) du - 1 \Big) \Big] \Big\},
\end{aligned}
$$

and then derive the corresponding ELBO as

$$
\begin{aligned}
&\mathrm{ELBO}(w; \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \Big\{ \gamma_{ik}\gamma_{jl} \Big(\left[\mathbb{1}_{e_{ij}\neq 0} \log(p_{kl}) + \mathbb{1}_{e_{ij}=0} \log(1-p_{kl}) \right] \\
&\quad + \mathbb{1}_{e_{ij}\neq 0} \\
&\quad \times \Big[K_h(w_{ij}-w)\zeta_{kl}(w_{ij}-w) \\
&\quad - \Big(\int_{\mathcal{X}} K_h(u-w) \exp(\zeta_{kl}(u-w)) du - 1 \Big) \Big] \Big) \Big\} \\
&\quad + \sum_{i=1}^{N} \sum_{k=1}^{K} \gamma_{ik} \Big[\log(\pi_k) - \log(\gamma_{ik}) \Big].
\end{aligned}
\tag{20}
$$

We maximize (20) over a sequence of grid points to estimate the block densities f_{kl}.

Remark 3. We note that, the estimated densities in the M-step directly affect the optimization over variational parameters γ, which can seen by the ELBO constructed after the variational approximation. Given that cluster memberships are usually estimated by the hard clustering over these parameters, weights also affect the clusters.

4. Simulation Studies

In this section, we conduct simulation studies to examine our proposed nonparametric methods. The general procedure that we adopt to simulate entails the following steps:

1. First we simulate the membership indicators for all nodes from multinomial distribution with parameter vector π corresponding to uniform mixture proportions.
2. We simulate the binary adjacency matrix by simulating dyads in the static network given the cluster membership indicators of nodes. While simulating these dyads we use the network parameters with two settings $\theta_{s_1} = (-1, 1)$ and $\theta_{s_2} = (-0.5, 0.5)$. The first setting corresponds to well separated clusters on the basis of density of edges while second setting considers the more extreme case when the clusters are relatively close.
3. For each node pair with an edge, we simulate the weight on that edge using true distribution with block parameter that depends on their cluster memberships.

We consider two distributions separately: Normal and Gamma. For space consideration, we include all results about Gamma distributions in the supplementary materials.

We compare three different model-based clustering methods in each simulation, which are based on binary ERGM, proposed nonparametric weighted ERGM and "oracle" parametric weighted ERGM (Desmarais and Cranmer 2012) with the correct specification of weight distributions. We consider different node sizes from 100 to 500 and 100 repetitions. Before proceeding, we introduce several average metrics to measure clustering and model parameters estimation performance for different simulation settings over 100 replications. First, to assess the clustering performance, we calculate the log of Rand Index (logRI). The measure $\mathrm{RI}(\boldsymbol{z}, \hat{\boldsymbol{z}})$ calculates the proportion of pairs whose estimated labels correspond to the true labels in terms of being assigned to the same or different groups (Rand 1971). We calculate logRI as,

$$
\mathrm{logRI} = \log\left(\frac{1}{\binom{n}{2}} \sum_{i<j} (I\{z_i = z_j\} I\{\hat{z}_i = \hat{z}_j\} + I\{z_i \neq z_j\} I\{\hat{z}_i \neq \hat{z}_j\}) \right).
$$

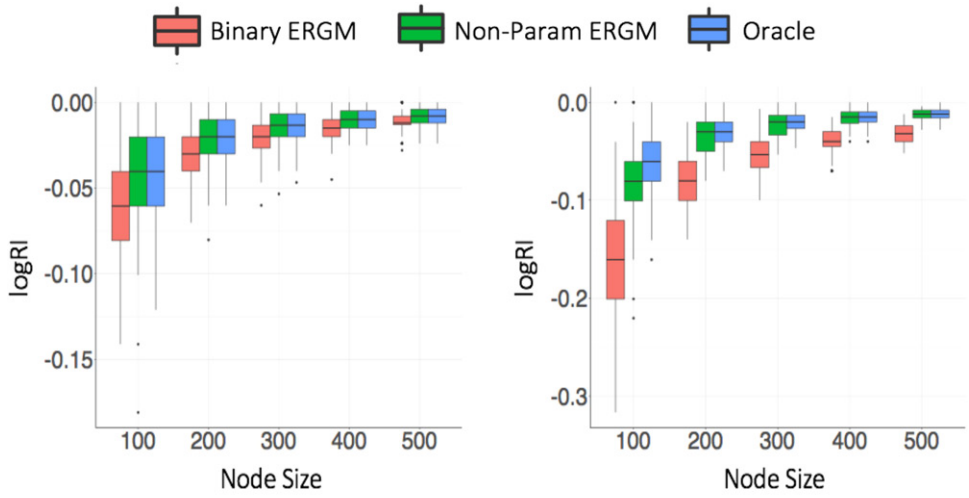
Next, to assess the performance of the estimators of network parameters $\boldsymbol{\theta}$, we consider log of square root of the average squared error (logRASE),

$$
\mathrm{logRASE}_{\boldsymbol{\theta}} = \frac{1}{2} \log\left(\frac{1}{K} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_k)^2 \right).
$$

To assess the performance of the density estimation $\boldsymbol{f}$, we consider the Kolmogorov–Smirnov (KS) statistic,

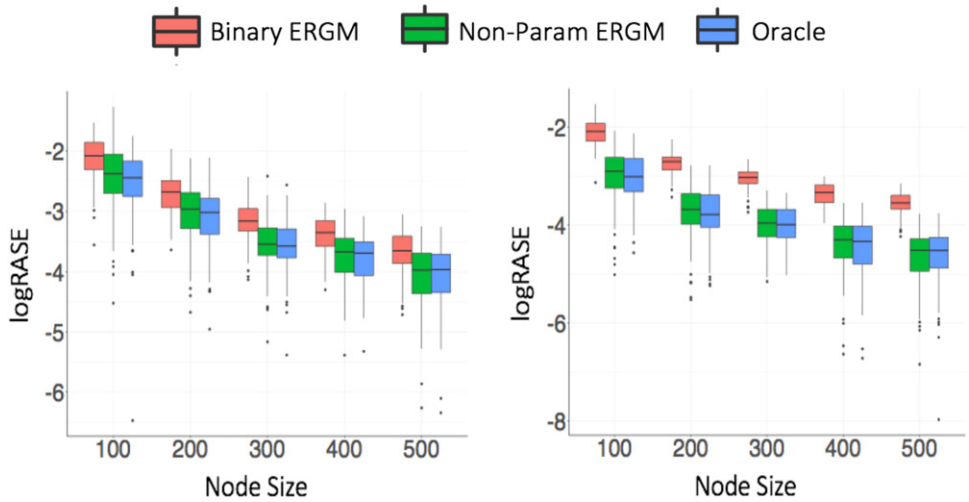
$$
\mathrm{KS}_f = \sup_w \mid \hat{\boldsymbol{f}}(w) - \boldsymbol{f}(w) \mid .
$$

Based on the metrics defined here, Figures 1 and 2 show clustering and θ estimation performance for different sparsity parameter settings averaged over 100 simulations of graphs for normal weight distributions. Corresponding figures for Gamma weight distributions look similar and have been moved to Appendix A. The differences in logRI and logRASE for θ_{s_1} and θ_{s_2} evidently confirms the expected fact that separating two very close clusters is difficult compared to well separated clusters. It appears that the both the distributions allow a reasonable recovery of the cluster membership indicators, when the graphs considered have more than 100 nodes. As expected, the node size improves the recovery of latent structure and estimation of network parameters θ in all cases. It can be observed that our proposed nonparametric ERGM outperforms the binary ERGM by a large difference and performs competitively with the oracle method (parametric ERGM with true weight distributions) for all simulation settings. We note that our proposed strategy is best suited for real world applications when the true distributions for block pairs are unknown.



(a) θ_{s_1} with Normal distributions (b) θ_{s_2} with Normal distributions

Figure 1. Clustering performance measured using logRI against different node sizes comparing the three models for different sparsity parameter settings under Normal weight distributions.



(a) θ_{s_1} with Normal distributions (b) θ_{s_2} with Normal distributions

Figure 2. θ Estimation Performance measured using logRASE against different node sizes comparing the three models for different sparsity parameter settings under Normal weight distributions.

Figures 3 and 4 show the empirical distributions of network parameters θ for normal weight distributions. Corresponding figures for Gamma weight distributions have been moved to Appendix A. The proposed nonparametric model estimation again outperforms the binary ERGM uniformly for all settings. We also note that the contour plots for the proposed model seem really close to Oracle model, thus demonstrating the power of our approach.

Figure 5 shows the estimated block densities within 2.5 and 97.5 percentiles for node size of 500 for normal weight distributions. Corresponding figure for Gamma weight distributions have been moved to Appendix A. Comparing θ_{s_1} and θ_{s_2}, it is evident that within cluster 1 density estimation improves substantially for θ_{s_2}. This is because cluster 1 is more sparse for θ_{s_1} compared to θ_{s_2}. Comparing the normal and gamma distributions, we observe that asymmetry of gamma distribution leads to underestimation of within cluster 1 estimated density. Cluster 1 is again most affected since it is most sparse. We point out here that for larger node sizes, these estimated densities will converge to true densities (Loader 1996).

Table 1 gives the summary of KS statistic for various simulation settings. We note that for sparse cluster 1, comparing θ_{s_1} and θ_{s_2}, there is a huge improvement when the true distribution is Normal. However the improvement is only minor in case of Gamma due to asymmetry. The differences are much less substantial for other blocks, however, Normal uniformly outperforms Gamma for all settings.

5. Application to Water Pollution Analysis

In this section, we demonstrate the power of our methodology in an environmental application to study sulfate concentrations in river networks. The dataset consists of three main parts. The

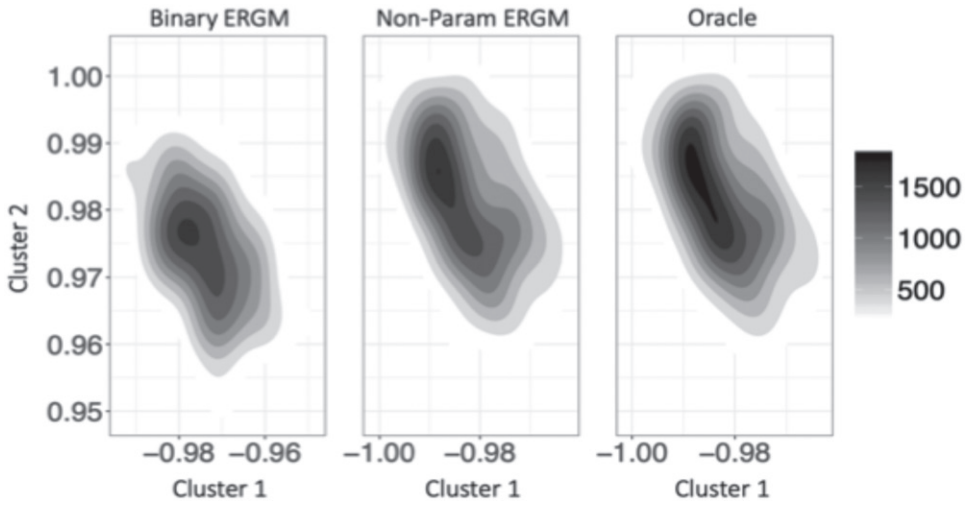


Figure 3. Plots of empirical joint distributions of network parameters θ_{s_1} for Normal weight distributions over 100 simulations with 500 nodes, comparing the three models for different block distributions.

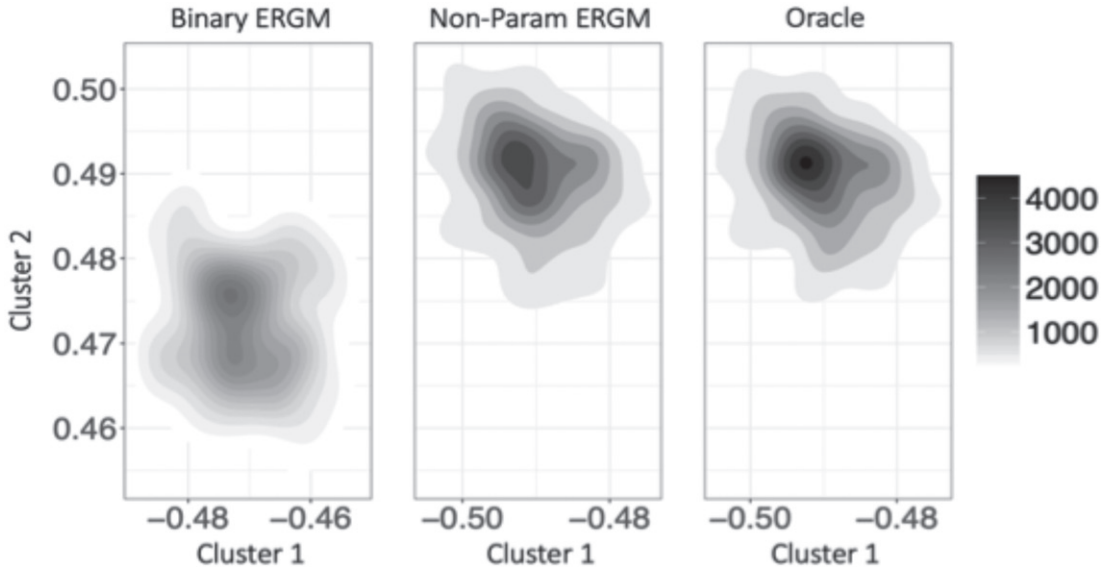


Figure 4. Plots of empirical joint distributions of network parameters θ_{s_2} for Normal weight distributions over 100 simulations with 500 nodes, comparing the three models for different block distributions.

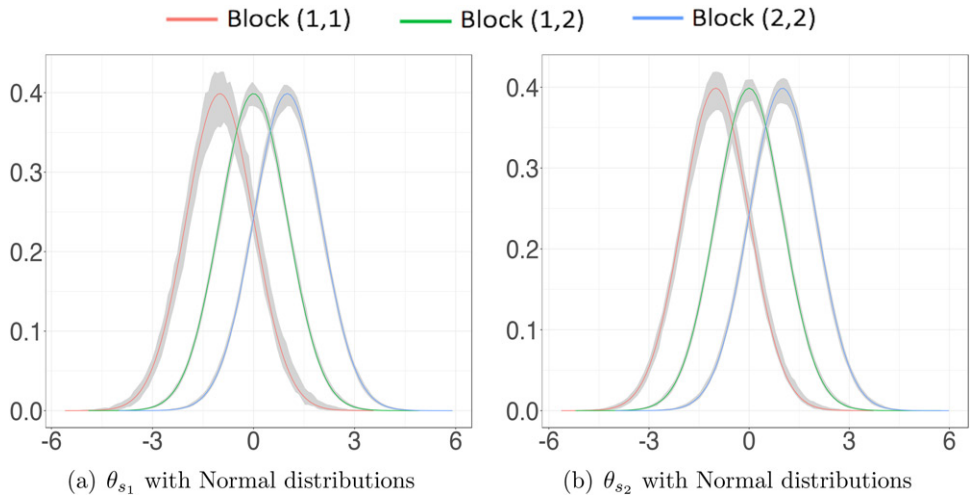


Figure 5. Estimated block densities for normal weight distributions.

Table 1. Summary of KS statistic ($\times 10^2$) for the three block densities under different simulation settings for the proposed model, computed over 100 simulations of graphs with 500 nodes.

		θ_{s_1}		θ_{s_2}	
	Summary statistic	Normal	Gamma	Normal	Gamma
Block (1,1)	Median	3.94	4.51	2.77	4.26
	Mean	4.03	4.70	2.84	4.36
Block (1,2)	Median	1.40	1.52	1.46	1.57
	Mean	1.41	1.56	1.49	1.61
Block (2,2)	Median	1.51	1.53	1.63	1.76
	Mean	1.53	1.60	1.66	1.76

first part consists of approximately 865 sulfate samples measured as concentrations in the units of parts per million (ppm) over several creeks in the Ohio watershed in Pennsylvania. The source include the following online databases: the USGS National Water Information System, the Susquehanna River Basin Commission database, the EPA STORET Data Warehouse, and the Shale Network database (*doi:10.4211/his-data-shalenetwork*). The second part consists of the directed geographical river network in the form of locations of all streams and creeks in Pennsylvania. The directions in the network correspond to the actual river flow. The third part consists of 93 coal mine locations which are suspected to be potential polluters of the river streams posing an environmental risk. Both the latter parts are publically available at Pennsylvania Spatial Data Access (PASDA). We map the latitude and longitude of the sampling sites onto the geographical river network. The 865 sulfate sampling sites become the nodes. Thus the nodes in the network are defined through first part of the data after mapping to nearest streams in the second part. We define the edges according to the path of river flow; for sampling sites A and B, the edge is present when river can flow either from A to B or from B to A directly or through one of its sub-tributary or some intermediary stream. The path information used to define the edges entirely comes from the second part. The spatial weights are constructed as proportional to the strength of influence of the upstream site on the downstream site, measured by difference in concentrations between the sampling locations (Peterson, Theobald, and ver Hoef 2007). For example, if the river is flowing from A to B, with measured sulfate concentrations C_A and C_B, then weight on the edge between A and B would be defined as $w_{AB} = C_B - C_A$. Note in this definition, the weights could be negative if $C_B < C_A$. In this context, we have transformed the geographically defined "river network" to a weighted network defined above for the purpose of our analysis.

It is an important question to study the water pollution in river networks. Most of the existing spatial clustering methods rely on some "neighborhood" metric that cluster the data points based on their spatial proximity. However, these approaches fail in a river network setup when the two points are very close spatially but still not connected by the river flow or vice-versa. These methods usually rely on some criteria to choose the number of clusters that is heuristic and not rigorously founded on model likelihood. There have been several parametric approaches to study spatial concentration gradients (see, e.g., Lawson and Denison 2002). However most methods fail in water pollution applications where the gradients tend to be asymmetric, heavy tailed and multimodal, with an unknown

Table 2. Distributional properties of sulfate concentration gradients.

	Skewness	Kurtosis	Hartigan's dip test p-value
Whole network	0.050	3.885	$<2.2 \times 10^{-16}$
Block (1,1)	0.762	6.817	$<2.2 \times 10^{-16}$
Block (1,2)	−0.011	3.447	9.37×10^{-6}
Block (2,2)	0.026	3.672	$<2.2 \times 10^{-16}$

number of modes. We adopt skewness, kurtosis and Hartigan's dip test of unimodality, respectively, to infer these properties over the whole sulfate network and individual clusters obtained using our model. The results, summarized in Table 2, clearly indicate that the gradients follow an asymmetric, leptokurtic and multimodal distribution for all block pairs. This calls for a generalized framework of clustering the weighted network while modeling the concentration gradients without making any distributional assumptions.

The proposed nonparametric weighted network models address all the aforementioned challenges. Now, we apply the proposed method to analyze the sulfate concentration network. In practice, we need to effectively choose the number of clusters (i.e., K). Since the likelihood is intractable (Biernacki, Celeux, and Govaert 2000), we follow (Daudin, Picard, and Robin 2008) to introduce a modified integrated classification likelihood (ICL) criterion

$$\mathrm{ICL}_K = \log P(\boldsymbol{Y}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{f}}) - (K-1)\log n - K \log\left(\frac{n(n-1)}{2}\right), \tag{21}$$

where the complete log-likelihood with estimated membership and densities becomes

$$\begin{aligned}\log P(\boldsymbol{Y}, \hat{\boldsymbol{Z}}, \hat{\boldsymbol{f}}) = \sum_{i<j}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big\{\hat{z}_{ik}\hat{z}_{jl}\Big(&\log P_{\boldsymbol{\theta}_{z_i z_j}}(E_{ij} = e_{ij} \mid \mathbf{Z} = \mathbf{z}) \\ &+ \log \hat{f}_{z_i z_j}(W_{ij} = w_{ij} \mid e_{ij} = 1, \mathbf{Z} = \mathbf{z})\Big)\Big\} \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}\log\hat{\pi}_k.\end{aligned}$$

The ICL follows the philosophy of Bayesian model selection criterion. The second term in (21) penalizes for the $K-1$ free parameters in the mixture proportions $\boldsymbol{\pi}$. The third term accounts for the penalization of the network parameters (Matias and Miele 2017) given the additive structure of the specified ERGM.

We choose the optimal number of clusters by maximizing the modified ICL criterion, which suggests $K = 2$. The estimated network sparsity parameters corresponding to these two clusters labelled C_1 and C_2 are −0.521 and −2.084, respectively, indicating that C_1 has higher degree in terms of the edges $(e_{ij})_{i,j\in C_1}$ on average compared to C_2. We plot the sampling sites belonging to the two clusters overlaying the potential polluter locations in Figure 6. It can be seen that coal mines 4, 5, 15, 20, 22, 38, 40, 42, 45, 46 73, 75–78, 82, and 88 lie directly either upstream or downstream of nodes in C_1 suggesting they may significantly affect the sulfate concentrations. As show in Table 3, C_1 consists of relatively more hubs with higher degree on average while C_2 consists more nodes with lower degree on average.

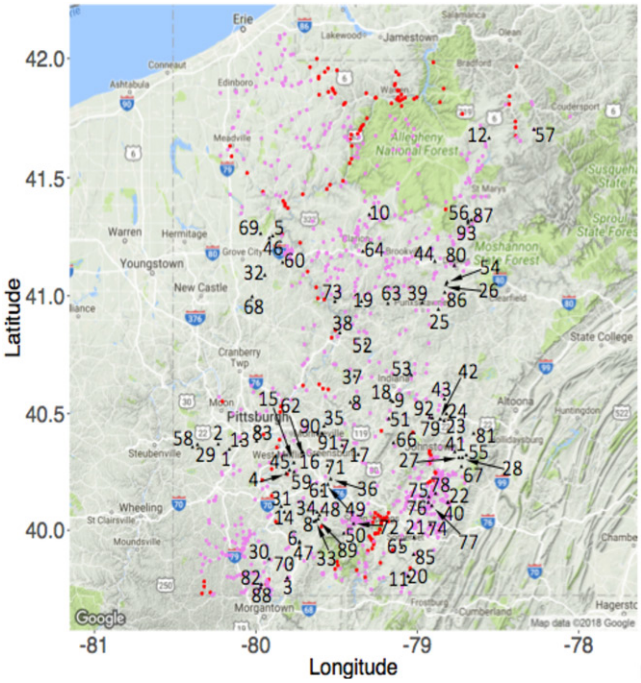


Figure 6. Clustered sulfate sampling sites with C_1 in red and C_2 in pink and coal mining sites in black (numbered 1–93).

Table 3. Summary of nodes, degree and edge weights of the two clusters.

Summary	C_1	C_2
Number of nodes	147	718
Average degree	18.72	5.21
Minimum	−385.50	−389.70
1st quantile	−2.00	−35.99
Median	8.081	10.24
Mean	40.41	23.66
3rd quantile	66.96	111.80
Maximum	436.40	441.00

In Figure 7(a), we plot densities estimated for these clusters. From this density plot, it is clear that C_1 has two modes (Mode 1 and Mode 3) and C_2 has one mode (Mode 2) on the positive sulfate concentrations. Based on these modes and estimated network parameters, we identify subregions of interest corresponding to 3 modes:

- Mode 1 consists of subregions with high degrees and higher differences of sulfate concentrations among adjacent nodes;
- Mode 2 consists of sparse subregions with low degrees and moderate differences of sulfate concentrations among adjacent nodes;
- Mode 3 consists of dense subregions with high degrees and low differences of sulfate concentrations among adjacent nodes.

The modes on positive weights indicate that the pollutant's concentration increases downstream, thus pointing toward polluters making significant impact on the environment. We analyze three modes by plotting their corresponding subnetworks in Figure 7 and summarizing the descriptive statistics of edge weights in Table 4.

Figures 7(c) and (d) give sub-networks for the positive modes in C_1. We identify coal mine ids 10, 19, 64, and 73 directly lying

Table 4. Descriptive statistics of edge weights of the three modes.

Summary	Mode 1	Mode 2	Mode 3
Minimum	303.9	90.18	42.33
1st quantile	328.5	111.70	56.86
Median	334.60	126.80	65.46
Mean	334.80	128.30	65.42
3rd quantile	341.9	140.00	70.92
Maximum	365.8	186.20	90.02

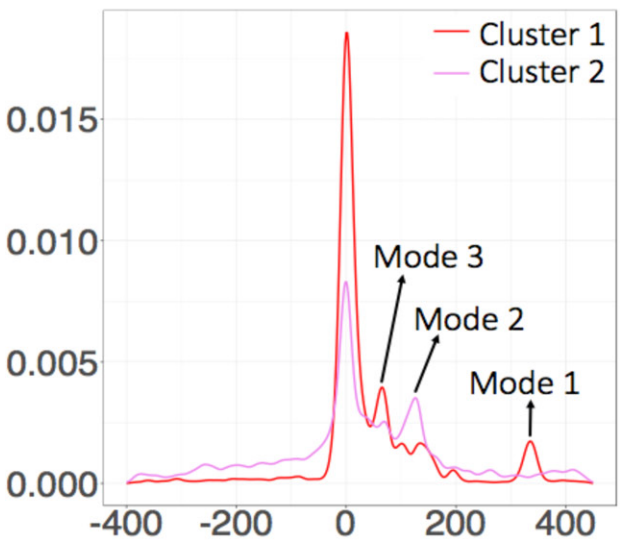
on highly weighted edges around Mode 2 and emanating from the hubs. We define the hubs as nodes with an outlier degree. The outlier degree is defined in the usual sense of outliers, that is, any degree which is greater than $Q_3 + 1.5 \times \text{IQR}$, where Q_3 is the upper quartile of the degree distribution and IQR is the interquartile range. These nodes correspond to sampling sites usually located at junctions of several streams. The coal mines and hubs are marked black and green, respectively. Since these coal mines belong to Mode 1, they are most likely to cause serious water pollution and must be investigated in a prioritized manner. Next we identify coal mine ids 2, 4, 5, 15, 32, 36, 38, 45, 49, 59, and 68 in Mode 3. These mines emanate from nodes that are connected densely to other nodes and so the affected region is larger, even though the pollution is relatively low. These mines in Mode 3 cause moderate impact over a large section of river network and must be monitored accordingly. Figure 7(b) shows the sub-network for the only positive mode in C_2. This corresponds to Mode 2 with relatively moderate to low local impacts.

We also compare the binary ERGM model with weighted model and observe that binary ERGMs cluster 1 consists of 108 nodes, a strict subset of 147 nodes of weighted model's C_1. Figures 7(c) and (d) show this difference of nodes by yellow. These indicate that several edges would have been absent in C_1 had we used binary model, thus missing significant coal mines that lie in Mode 1. Clearly taking weights into account while clustering helps to differentiate mines between Mode 1 and Mode 3 and hence could uncover important potential polluters.

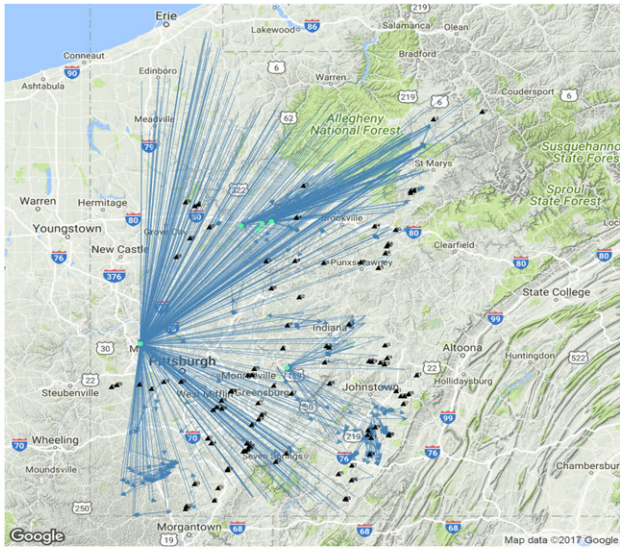
6. Discussion

We introduce a new nonparametric model-based approach for clustering large-scale weighted networks. The ERGM specification allows the flexibility to incorporate interesting network statistics and the nonparametric density function provides the robustness to study the network weights. We illustrate the power of our proposed method in a real application to study water pollution networks.

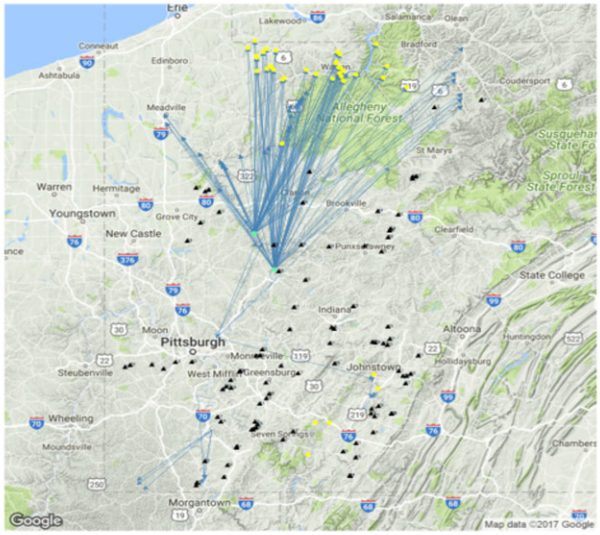
In general, our proposed method does not require a parametric specification of network weights and thus it is robust to the model mis-specification, and it can be extended to incorporate the nonparametric mixture functions or parametric constraints (DeSarbo, Chen, and Blank 2017; Lee and Xue 2018). Like most nonparametric methods, when the sample size is limited, our proposed method may not perform well. The semiparametric extension such as Xue and Zou (2012, 2014) and Fan, Xue, and Zou (2016) seems a promising alternative to the proposed nonparametric method. Moreover, the proposed method could be computationally intensive when



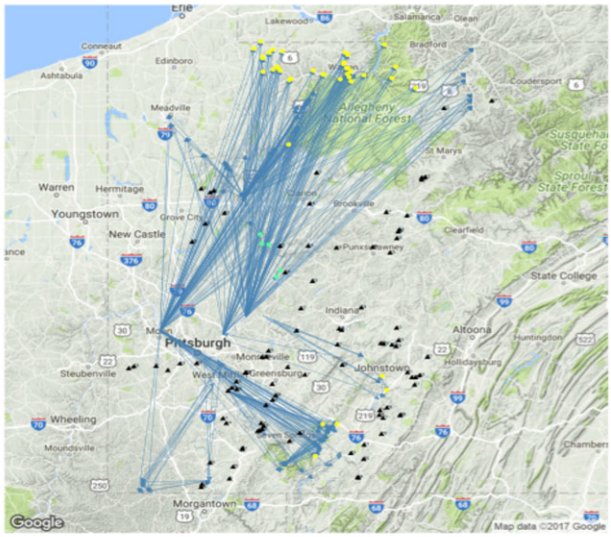
(a) Non-parametrically estimated densities



(b) Sub-network for Mode 2 in C_2



(c) Sub-network for Mode 1 in C_1



(d) Sub-network for Mode 3 in C_1

Figure 7. Non parametrically estimated densities for ties within clusters and subnetworks for different modes.

the number of nodes is huge. To make the proposed methods scalable, we shall follow the stochastic variational methods (Hoffman et al. 2013) to employ the minibatch sampling scheme.

Supplementary Materials

We include the additional figures from simulation studies in the supplementary materials.

Acknowledgments

The authors thank the editor, an associate editor, and two referees for their constructive comments and suggestions. Amal Agarwal and Lingzhou Xue have been partially supported by the National Institute on Drug Abuse grant P50DA039838 and the National Science Foundation grants DMS-1505256 and DMS-1811552.

Funding

Amal Agarwal and Lingzhou Xue have been partially supported by the National Institute on Drug Abuse grant P50DA039838 and the National Science Foundation grants DMS-1505256 and DMS-1811552.

References

Aicher, C., Jacobs, A. Z., and Clauset, A. (2014), "Learning Latent Block Structure in Weighted Networks," *Journal of Complex Networks*, 3, 221–248. [162]

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008), "Mixed Membership Stochastic Blockmodels," *Journal of Machine Learning Research*, 9, 1981–2014. [161,162]

Allman, E. S., Matias, C., and Rhodes, J. A. (2011), "Parameter Identifiability in a Class of Random Graph Mixture Models," *Journal of Statistical Planning and Inference*, 141, 1719–1736. [163]

Ambroise, C., and Matias, C. (2012), "New Consistent and Asymptotically Normal Parameter Estimates for Random-Graph Mixture

Models," *Journal of the Royal Statistical Society*, Series B, 74, 3–35. [162]

Anastasiadis, E., Deng, X., Krysta, P., Li, M., Qiao, H., and Zhang, J. (2016), "Network Pollution Games," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, International Foundation for Autonomous Agents and Multiagent Systems, pp. 23–31. [161]

Bernhardt, E. S., Lutz, B. D., King, R. S., Fay, J. P., Carter, C. E., Helton, A. M., Campagna, D., and Amos, J. (2012), "How Many Mountains can We Mine? Assessing the Regional Degradation of Central Appalachian Rivers by Surface Coal Mining," *Environmental Science & Technology*, 46, 8115–8122. [161]

Besag, J. (1974), "Spatial Interaction and the Statistical Analysis of Lattice Systems," *Journal of the Royal Statistical Society*, Series B, 36, 192–236. [162]

Biernacki, C., Celeux, G., and Govaert, G. (2000), "Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 719–725. [168]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. [164]

Bollobás, B., Janson, S., and Riordan, O. (2007), "The Phase Transition in Inhomogeneous Random Graphs," *Random Structures & Algorithms*, 31, 3–122. [162]

Bui, T. N., Chaudhuri, S., Leighton, F. T., and Sipser, M. (1987), "Graph Bisection Algorithms With Good Average Case Behavior," *Combinatorica*, 7, 171–191. [162]

Caimo, A., and Friel, N. (2011), "Bayesian Inference for Exponential Random Graph Models," *Social Networks*, 33, 41–55. [162]

Daudin, J.-J., Picard, F., and Robin, S. (2008), "A Mixture Model for Random Graphs," *Statistics and Computing*, 18, 173–183. [164,168]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [163]

Dennis, J. E., Jr, and Schnabel, R. B. (1996), *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Philadelphia, PA: SIAM. [165]

DeSarbo, W. S., Chen, Q., and Blank, A. S. (2017), "A Parametric Constrained Segmentation Methodology for Application in Sport Marketing," *Customer Needs and Solutions*, 4, 37–55. [169]

Desmarais, B. A., and Cranmer, S. J. (2012), "Statistical Inference for Valued-Edge Networks: The Generalized Exponential Random Graph Model," *PLoS One*, 7, e30136. [165]

Dyer, M. E., and Frieze, A. M. (1989), "The Solution of Some Random NP-Hard Problems in Polynomial Expected Time," *Journal of Algorithms*, 10, 451–489. [162]

Ebenstein, A. (2012), "The Consequences of Industrialization: Evidence From Water Pollution and Digestive Cancers in China," *Review of Economics and Statistics*, 94, 186–201. [161]

EPA (2017), "2017 National Water Quality Inventory Report to Congress," Technical Report, Washington, D.C., United States. [161]

Erdős, P., and Rényi, A. (1959), "On Random Graphs I," *Publicationes Mathematicae Debrecen*, 6, 290–297. [162]

Fan, J., Xue, L., and Zou, H. (2016), "Multitask Quantile Regression Under the Transnormal Model," *Journal of the American Statistical Association*, 111, 1726–1735. [169]

Frank, O., and Strauss, D. (1986), "Markov Graphs," *Journal of the American Statistical Association*, 81, 832–842. [162]

Gianessi, L. P., and Peskin, H. M. (1981), "Analysis of National Water Pollution Control Policies: 2. Agricultural Sediment Control," *Water Resources Research*, 17, 803–821. [161]

Gilbert, E. N. (1959), "Random Graphs," *The Annals of Mathematical Statistics*, 30, 1141–1144. [162]

Girvan, M., and Newman, M. E. (2002), "Community Structure in Social and Biological Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 7821–7826. [161,162]

Handcock, M. S., Robins, G., Snijders, T. A., Moody, J., and Besag, J. (2003), "Assessing Degeneracy in Statistical Models of Social Networks," Technical Report, Citeseer. [162]

Hanneke, S., Fu, W., and Xing, E. P. (2010), "Discrete Temporal Models of Social Networks," *Electronic Journal of Statistics*, 4, 585–605. [162,163]

Hendryx, M., and Ahern, M. M. (2008), "Relations Between Health Indicators and Residential Proximity to Coal Mining in West Virginia," *American Journal of Public Health*, 98, 669–671. [161]

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *The Journal of Machine Learning Research*, 14, 1303–1347. [170]

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983), "Stochastic Blockmodels: First Steps," *Social Networks*, 5, 109–137. [162]

Holland, P. W., and Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs," *Journal of the American Statistical Association*, 76, 33–50. [162]

Hunter, D. R., and Handcock, M. S. (2006), "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics*, 15, 565–583. [162]

Hunter, D. R., and Lange, K. (2004), "A Tutorial on MM Algorithms," *The American Statistician*, 58, 30–37. [164]

Karrer, B., and Newman, M. E. J. (2011), "Stochastic Blockmodels and Community Structure in Networks," *Physical Review E*, 83, 016107. [161]

Kim, B., Lee, K. H., Xue, L., and Niu, X. (2018), "A Review of Dynamic Network Models With Latent Variables," *Statistics Surveys*, 12, 105–135. [163]

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010), "Analysing Exponential Random Graph (p-Star) Models With Missing Data Using Bayesian Data Augmentation," *Statistical Methodology*, 7, 366–384. [162]

Krivitsky, P. N. (2012), "Exponential-Family Random Graph Models for Valued Networks," *Electronic Journal of Statistics*, 6, 1100. [162]

Krivitsky, P. N., and Handcock, M. S. (2014), "A Separable Model for Dynamic Networks," *Journal of the Royal Statistical Society*, Series B, 76, 29–46. [163]

Lawson, A. B., and Denison, D. G. (2002), *Spatial Cluster Modelling*, Boca Raton, FL: CRC Press. [168]

Lee, K. H., and Xue, L. (2018), "Nonparametric Finite Mixture of Gaussian Graphical Models," *Technometrics*, 60, 511–521. [169]

Lee, K. H., Xue, L., and Hunter, D. R. (2017), "Model-Based Clustering of Time-Evolving Networks Through Temporal Exponential-Family Random Graph Models," arXiv no. 1712.07325. [162]

Li, Z., Zheng, G., Agarwal, A., Xue, L., and Lauvaux, T. (2017), "Discovery of Causal Time Intervals," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, SIAM, pp. 804–812. [161]

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015), "Assessing Beijing's $PM_{2.5}$ Pollution: Severity, Weather Impact, APEC and Winter Heating," *Proceedings of the Royal Society A*, 471, 20150257. [161]

Lienert, J., Schnetzer, F., and Ingold, K. (2013), "Stakeholder Analysis Combined With Social Network Analysis Provides Fine-Grained Insights Into Water Infrastructure Planning Processes," *Journal of Environmental Management*, 125, 134–148. [161]

Lin, N., Jing, R., Wang, Y., Yonekura, E., Fan, J., and Xue, L. (2017), "A Statistical Investigation of the Dependence of Tropical Cyclone Intensity Change on the Surrounding Environment," *Monthly Weather Review*, 145, 2813–2831. [161]

Loader, C. R. (1996), "Local Likelihood Density Estimation," *The Annals of Statistics*, 24, 1602–1618. [163,165,166]

Matias, C., and Miele, V. (2017), "Statistical Clustering of Temporal Networks Through a Dynamic Stochastic Block Model," *Journal of the Royal Statistical Society*, Series B, 79, 1119–1141. [168]

Møller, J., Pettitt, A. N., Reeves, R., and Berthelsen, K. K. (2006), "An Efficient Markov Chain Monte Carlo Method for Distributions With Intractable Normalising Constants," *Biometrika*, 93, 451–458. [162]

Montgomery, W. D. (1972), "Markets in Licenses and Efficient Pollution Control Programs," *Journal of Economic Theory*, 5, 395–418. [161]

Niu, X., Wendt, A., Li, Z., Agarwal, A., Xue, L., Gonzales, M., and Brantley, S. L. (2018), "Detecting the Effects of Coal Mining, Acid Rain, and Natural Gas Extraction in Appalachian Basin Streams in

Pennsylvania (USA) Through Analysis of Barium and Sulfate Concentrations," *Environmental Geochemistry and Health*, 40, 865–885. [161]

Nowicki, K., and Snijders, T. A. (2001), "Estimation and Prediction for Stochastic Blockstructures," *Journal of the American Statistical Association*, 96, 1077–1087. [161,162]

Peterson, E. E., Theobald, D. M., and ver Hoef, J. M. (2007), "Geostatistical Modelling on Stream Networks: Developing Valid Covariance Matrices Based on Hydrologic Distance and Stream Flow," *Freshwater Biology*, 52, 267–279. [168]

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850. [165]

Ruzol, C., Banzon-Cabanilla, D., Ancog, R., and Peralta, E. (2017), "Understanding Water Pollution Management: Evidence and Insights From Incorporating Cultural Theory in Social Network Analysis," *Global Environmental Change*, 45, 183–193. [161]

Saldana, D. F., Yu, Y., and Feng, Y. (2017), "How Many Communities Are There?," *Journal of Computational and Graphical Statistics*, 26, 171–181. [162]

Schweinberger, M. (2011), "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families," *Journal of the American Statistical Association*, 106, 1361–1370. [162]

Smith, R. A., Alexander, R. B., and Wolman, M. G. (1987), "Water-Quality Trends in the Nation's Rivers," *Science*, 235, 1607–1616. [161]

Snijders, T. A. (2002), "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure*, 3, 1–40. [162]

Snijders, T. A., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Blockmodels for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [161,162]

Strauss, D. (1986), "On a General Class of Models for Interaction," *SIAM Review*, 28, 513–527. [162]

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., and Liermann, C. R. (2010), "Global Threats to Human Water Security and River Biodiversity," *Nature*, 467, 555–561. [161]

Vu, D. Q., Hunter, D. R., and Schweinberger, M. (2013), "Model-Based Clustering of Large Networks," *The Annals of Applied Statistics*, 7, 1010. [162,164]

Wang, Y. R., and Bickel, P. J. (2017), "Likelihood-Based Model Selection for Stochastic Block Models," *The Annals of Statistics*, 45, 500–528. [162]

Wen, T., Agarwal, A., Xue, L., Chen, A., Herman, A., Li, Z., and Brantley, S. L. (2019), "Assessing Changes in Groundwater Chemistry in Landscapes With More Than 100 Years of Oil and Gas Development," *Environmental Science: Processes & Impacts*, 21, 384–396. [161]

Xue, L., and Zou, H. (2012), "Regularized Rank-Based Estimation of High-Dimensional Nonparanormal Graphical Models," *The Annals of Statistics*, 24, 2541–2571. [169]

——— (2014), "Rank-Based Tapering Estimation of Bandable Correlation Matrices," *Statistica Sinica*, 40, 83–100 [169]

Zhao, Y., Levina, E., and Zhu, J. (2012), "Consistency of Community Detection in Networks Under Degree-Corrected Stochastic Block Models," *The Annals of Statistics*, 40, 2266–2292. [162]