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

Amal Agarwal & Lingzhou Xue

To cite this article: Amal Agarwal & Lingzhou Xue (2019): Model-Based Clustering of Nonparametric Weighted Networks With Application to Water Pollution Analysis, Technometrics, DOI: 10.1080/00401706.2019.1623076

To link to this article: https://doi.org/10.1080/00401706.2019.1623076
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.

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 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 ecological 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; Ruzo 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,
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 \( Y = (Y_{ij})_{1 \leq i < j \leq n} \) 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 \( W = (W_{ij})_{1 \leq i < j \leq n} \) 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 \( E \) as

\[
P_{\theta}(E = e) = \exp\left(\theta' g(e) - \psi(\theta)\right),
\]

where \( \psi(\theta) = \log \sum_{e \in E} \exp \left(\theta' g(e)\right) \) is the log of the normalizing constant, \( \theta \in \mathbb{R}^p \) are canonical network parameters of interest, and \( g : E \to \mathbb{R}^p \) is the sufficient statistic. Here \( E \) is the space for \( E \) 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,

\[
P_{\theta}(Y = y) = \prod_{1 \leq i < j \leq n} \exp(\theta' g(e_{ij}) - \psi(\theta))P_{\theta}(W_{ij} = w_{ij}).
\]  

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

\[
P_{\theta,f}(Y = y|Z = z) = \prod_{1 \leq i < j \leq n} P_{\theta_{z_{ij}f_{z_{ij}}}}(Y_{ij} = y_{ij}|Z = z),
\]

where \( Z = (Z_{i})_{1 \leq i \leq n} \) denote the latent cluster memberships of nodes. Here, \( Z_i \) 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 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 of continuous network weights, for any given pair of nodes \((i,j)\), we have

\[
P_{\theta,f}(Y_{ij} = y_{ij}|Z = z) = \left(p_{z_{ij}f_{z_{ij}}}(w_{ij})\right)^{y_{ij} \neq 0} \left(1 - p_{z_{ij}z_{ij}}\right)^{y_{ij} = 0}.
\]  

Here, \( (p_{kl})_{1 \leq k,l \leq K} = (P_{\theta_{z_{ij}f_{z_{ij}}}}(E_{ij} = e_{ij}|Z = z))_{1 \leq k,l \leq K} \) take the parametric specification of exponential-family distributions of
the network statistics as explained in (1) and network weights
\((w_{ij})_{1 \leq i,j \leq n, z_i=z_j = 1}\) are assumed to be an iid sample observed
from a population following an univariate nonparametric den-
sity function \(f_{kl}\). Note that \((f_{kl})_{1 \leq k \leq K}\) 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 \(\theta\) as \(p_{kl} = \log^{-1}(\theta_k + \theta_l)\). 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 \((p_{kl})_{1 \leq k \leq K}\) 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
\[
\ell(n-1) \sum_{i=1}^{n} \sum_{l=1}^{n} \left\{ \log(f_{kl}(u_{ij})) - \left( \int_{\mathcal{X}} f_{kl}(u) du - 1 \right) \right\},
\]
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 arbitrary grid point \(w\) as
\[
\ell(\theta, f, w; Y | Z)
= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ 1_{e_{ij} \neq 0} \log(p_{z_i z_j}) + 1_{e_{ij} = 0} \log(1 - p_{z_i z_j}) \right\}
+ 1_{e_{ij} = 0} \times \left[ K_h(w_{ij} - w) \log(f_{z_i z_j}(w_{ij})) - \left( \int_{\mathcal{X}} K_h(u - w) f_{z_i z_j}(u) du - 1 \right) \right],
\]
where \(K_h\) is the rescaled kernel function with a positive band-
width \(h\). We approximate \(\log(f_{z_i z_j}(u))\) by \(\Phi^p_{kl}\), a linear combina-
tion of orthogonal basis functions \((\phi_m)_{1 \leq m \leq p}\), namely, \(\Phi^p_{kl}(u - w) = \sum_{m=0}^{p} \rho_m^{(kl)} \phi_m(u - w)\). With this approximation, the local conditional log-likelihood becomes
\[
\ell(\theta, \beta, w; Y | Z)
= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ 1_{e_{ij} \neq 0} \log(p_{z_i z_j}) + 1_{e_{ij} = 0} \log(1 - p_{z_i z_j}) \right\}
+ 1_{e_{ij} = 0} \times \left[ K_h(w_{ij} - w) \Phi^p_{z_i z_j}(w_{ij} - w) - \left( \int_{\mathcal{X}} K_h(u - w) \exp(\Phi^p_{z_i z_j}(u - w)) du - 1 \right) \right].
\]

We assume that membership indicators \(Z = (Z_i)_{1 \leq i \leq n}\) follow
a multinomial distribution with a single trial and mixture propor-
tions as \(\pi = (\pi_k)_{1 \leq k \leq K}\). The log-likelihood of the observed
weighted network can be written as
\[
\ell(\theta, \pi, f) = \log \left( \sum_{z \in [1, \ldots, K]} P_{\theta f}(Y = y | Z = z)P_{\pi}(Z = z) \right).
\]

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

**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
nonparametric 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 \(Y = (Y_t)_{1 \leq t \leq T}\) = \((E_t, W_t)_{1 \leq t \leq T}\), 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,
\[
P_{\theta f}(Y = y | Z = z) = \prod_{2 \leq t \leq T} P_{\theta f}(Y_t = y_t | Y_{t-1} = y_{t-1}, Z = z),
\]
where \(P_{\theta f}(Y_t = y_t | Y_{t-1} = y_{t-1}, Z = z)\) 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 \(P_{\theta f}(E_{ij} = e_t | Y_{t-1} = e_{t-1}, Z = z)\).
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 approxi-
mately solve the maximum likelihood estimation. It is infe-
sible to directly maximize the log-likelihood function (8)
due to two key challenges: (i) exponential-family form of
\(P(Y | Z)\) is not scalable for large networks; (ii) the sum
is over every possible assignment to \(Z\), where for each
\(1 \leq l \leq i \leq n, Z_i = z_l\) 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 \( Z \) are unobserved.

\[
\log(P_{\theta,Y}(Y = y, Z = z)) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{K} z_{ik} z_{ji} \log P_{\theta}(Y_{ij} = y_{ij} | Z = z) \]

\[
\geq \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k).
\]

Next we take expectation of this complete log-likelihood with the distribution \( P_{\theta}(Z | Y) \). Clearly the distribution \( P_{\theta}(Z | Y) = \mathbb{P}_{\theta}(Z_{i1 \leq i \leq n} | Y) \) cannot be further factored over nodes since for \( i \neq j, Z_i \) is not independent of \( Z_j \) given the observed network \( Y \). 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 \( A_Y(z) = \mathbb{P}(Z = z) \) for the latent variables \( Z \) and find the optimal setting for variational parameters \( \gamma \) 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.

\[
\ell(\theta, \pi, \gamma, f) = \log \left( \sum_{z \in \{1, \ldots, K\}^n} P_{\theta,Y}(Y = y | z) P_{\theta}(Z = z) / A_Y(z) \right) \geq \ell(\theta, \pi, \gamma, f).
\]

(10)

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

\[
A_Y(Z) = \prod_{i=1}^{n} \mathbb{P}_{\gamma_i}(Z_i) = \prod_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} Z_i.
\]

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

\[
\text{ELBO}(\theta, \gamma, \pi, f) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[ \log(\pi_k) - \log(\gamma_{ik}) + \log(f_{ik}(w_{ij})) - \int_{\chi} f_{ij}(u) \log(u) \, du - 1 \right] + \sum_{k=1}^{K} \log(\gamma_{ik}) - \log(\gamma_{ik}) + \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[ \log(\pi_k) - \log(\gamma_{ik}) \right].
\]

(12)

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

\[
\gamma(t) = \arg \max_{\gamma} \text{ELBO}(\theta^{(t-1)}, \gamma, \pi^{(t-1)}, f^{(t-1)}).
\]

(13)

The direct maximization of ELBO in (13) is difficult, since the lower bound depends on the products \( \gamma_{ik} f_{ij} \) and, therefore, the fixed-point updates of \( \gamma_{ik} \) depend on all other \( \gamma_{ij} \) (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

\[
Q(\theta(t), \gamma(t), \pi(t), f(t); \gamma) \leq \text{ELBO}(\theta(t), \gamma(t), \pi(t), f(t)) - \text{ELBO}(\theta^{(t-1)}, \gamma^{(t-1)}, \pi^{(t-1)}, f^{(t-1)}). \forall \gamma(t), \quad Q(\theta(t), \gamma(t), \pi(t), f(t); \gamma(t)) = \text{ELBO}(\theta(t), \gamma(t), \pi(t), f(t)).
\]

(14)

First we note that for all \( (\theta_{ij})_{1 \leq k \leq K} \) we have \( \log(\gamma_{ik}) < 0 \) and \( \log(1 - \gamma_{ik}) < 0 \) which gives rise to following inequalities using the arithmetic geometric mean inequality

\[
\gamma_{ik} f_{ij} \log(\gamma_{ik}) \geq \left( \frac{\gamma_{ik} f_{ij}}{2} \right)^2 \log(\gamma_{ik}),
\]

(16)

\[
\gamma_{ik} f_{ij} \log(1 - \gamma_{ik}) \geq \left( \frac{\gamma_{ik} f_{ij}}{2} \right)^2 \log(1 - \gamma_{ik}).
\]

(17)

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

\[
-\log(\gamma_{ik}) \geq -\log(\hat{\gamma}_{ik}) + \frac{\gamma_{ik}}{\hat{\gamma}_{ik}} + 1.
\]

(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.

\[
Q(\theta(t), \gamma^{(t-1)}, \pi^{(t-1)}, f^{(t-1)}, \gamma(t)) = \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=1}^{K} \left( \frac{\hat{\gamma}_{ik} f_{ij} \gamma_{ij}^{(t-1)}}{2} + \frac{\hat{\gamma}_{ik} f_{ij} \gamma_{ij}^{(t-1)}}{2} \right) \left\{ \delta_{\epsilon_{ij} \neq 0} \log(\hat{\gamma}_{ik} f_{ij}) + \delta_{\epsilon_{ij} = 0} \left[ \log(\hat{\gamma}_{ik} f_{ij})(w_{ij}) - \int_{\chi} f_{ij}(u) \log(u) \, du - 1 \right] \right\}.
\]

(19)

To maximize (19), we solve \( n \) separate quadratic programming problems of \( K \) variables \( \gamma \) under the constraints \( \gamma_{ik} \geq 0, \forall k \in \{1, \ldots, K\} \) together with \( \sum_{k=1}^{K} \gamma_{ik} = 1 \).

\textbf{M-step:} We first maximize the (12) with respect to \( \pi \) and \( \theta \).

We have a closed-form update for \( \pi_k^{(t)} = \frac{\sum_{i=1}^{n} \gamma_{ik}^{(t)}}{n} \).
To update $\theta$, 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 $f$. To this end, we use (6) and the approximation of $\log(f_{kl}(u))$. For simplicity, we use a local polynomial approximation such that $\log(f_{kl}(u))$ can be approximated by a low-degree polynomial $\xi_{kl}$ in a neighborhood of the fitting point $w$ (Loader 1996): $\log(f_{kl}(u)) \approx \xi_{kl}(u - w) = \sum_{m=0}^{p} p_{m}(u - w)^{m}$. With this approximation, we rewrite (7) as

$$\ell(\theta, \beta, w; Y | Z) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \left[ 1_{e_{ij} \neq 0} \log(p_{z_{i}z_{j}}) + 1_{e_{ij} = 0} \log(1 - p_{z_{i}z_{j}}) \right] + \sum_{i=1}^{n} \left[ K_{h}(w_{ij} - w) \right. \\
\left. \xi_{z_{i}z_{j}}(w_{ij} - w) - \left( \int_{X} K_{h}(u - w) \exp(\xi_{z_{i}z_{j}}(u - w)) du - 1 \right) \right]$$

and then derive the corresponding ELBO as

$$\text{ELBO}(w; \theta, \beta, \gamma) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{K} \sum_{l=1}^{K} \gamma_{kl} \left[ 1_{e_{ij} \neq 0} \log(p_{kl}) + 1_{e_{ij} = 0} \log(1 - p_{kl}) \right] + \sum_{i=1}^{n} \left[ K_{h}(w_{ij} - w) \xi_{kl}(w_{ij} - w) \right. \\
\left. - \left( \int_{X} K_{h}(u - w) \exp(\xi_{kl}(u - w)) du - 1 \right) \right] + \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma_{ik} \left[ \log(\sigma_{ik}) - \log(\gamma_{ik}) \right]$$

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 $\gamma$, which can be 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 $\pi$ 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 $\theta_1 = (-1, 1)$ and $\theta_2 = (-0.5, 0.5)$. 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 $RI(z, \hat{z})$ 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,

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

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

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

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

$$\text{KS}_{f} = \sup_{w} \left| \hat{f}(w) - f(w) \right|$$

Based on the metrics defined here, Figures 1 and 2 show clustering and $\theta$ 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 $\theta_1$ and $\theta_2$ 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 $\theta$ 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.
Figure 1. Clustering performance measured using logRI against different node sizes comparing the three models for different sparsity parameter settings under Normal weight distributions.

Figure 2. $\theta$ 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 $\theta$ 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 $\theta_1$ and $\theta_2$, it is evident that within cluster 1 density estimation improves substantially for $\theta_2$. This is because cluster 1 is more sparse for $\theta_1$ compared to $\theta_2$. 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 $\theta_1$ and $\theta_2$, 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 $\theta_1$ 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 $\theta_2$ 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.

<table>
<thead>
<tr>
<th>Summary statistic</th>
<th>Block (1,1)</th>
<th>Block (1,2)</th>
<th>Block (2,2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Median</td>
<td>Mean</td>
</tr>
<tr>
<td>$\theta_1$</td>
<td>3.94</td>
<td>2.77</td>
<td>4.26</td>
</tr>
<tr>
<td>$\theta_2$</td>
<td>4.03</td>
<td>2.84</td>
<td>4.36</td>
</tr>
</tbody>
</table>

Table 2. Distributional properties of sulfate concentration gradients.

<table>
<thead>
<tr>
<th></th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Hartigan's dip test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole network</td>
<td>0.050</td>
<td>3.885</td>
<td>$&lt;2.2 \times 10^{-16}$</td>
</tr>
<tr>
<td>Block (1,1)</td>
<td>0.762</td>
<td>6.817</td>
<td>$&lt;2.2 \times 10^{-16}$</td>
</tr>
<tr>
<td>Block (1,2)</td>
<td>-0.011</td>
<td>3.447</td>
<td>$9.37 \times 10^{-6}$</td>
</tr>
<tr>
<td>Block (2,2)</td>
<td>0.026</td>
<td>3.672</td>
<td>$&lt;2.2 \times 10^{-16}$</td>
</tr>
</tbody>
</table>

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 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

\[
\text{ICL}_K = \log P(Y, \hat{Z}, \hat{f}) - (K - 1) \log n - K \log \left( \frac{n(n - 1)}{2} \right),
\]

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

\[
\log P(Y, \hat{Z}, \hat{f}) = \sum_{i<j} \sum_{k=1}^{K} \left\{ \hat{z}_{ik} \hat{z}_{jk} \left( \log P_{\hat{f}_{yz}}(E_{ij} = e_{ij} | Z = z) + \log \frac{1}{Z} W_{ij} = w_{ij} | e_{ij} = 1, Z = z) \right) \right\}
+ \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \hat{\pi}_k.
\]

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 $\pi$. 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.

<table>
<thead>
<tr>
<th>Summary</th>
<th>$C_1$</th>
<th>$C_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of nodes</td>
<td>147</td>
<td>718</td>
</tr>
<tr>
<td>Average degree</td>
<td>18.72</td>
<td>5.21</td>
</tr>
<tr>
<td>Minimum</td>
<td>−385.50</td>
<td>−389.70</td>
</tr>
<tr>
<td>1st quantile</td>
<td>−2.00</td>
<td>−35.99</td>
</tr>
<tr>
<td>Median</td>
<td>8.081</td>
<td>10.24</td>
</tr>
<tr>
<td>Mean</td>
<td>40.41</td>
<td>23.66</td>
</tr>
<tr>
<td>3rd quantile</td>
<td>66.96</td>
<td>111.80</td>
</tr>
<tr>
<td>Maximum</td>
<td>436.40</td>
<td>186.20</td>
</tr>
</tbody>
</table>

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 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.

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

<table>
<thead>
<tr>
<th>Summary</th>
<th>Mode 1</th>
<th>Mode 2</th>
<th>Mode 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>303.9</td>
<td>90.18</td>
<td>42.33</td>
</tr>
<tr>
<td>1st quantile</td>
<td>328.5</td>
<td>111.70</td>
<td>56.86</td>
</tr>
<tr>
<td>Median</td>
<td>334.60</td>
<td>128.60</td>
<td>65.46</td>
</tr>
<tr>
<td>Mean</td>
<td>334.80</td>
<td>128.30</td>
<td>65.42</td>
</tr>
<tr>
<td>3rd quantile</td>
<td>341.9</td>
<td>140.00</td>
<td>70.92</td>
</tr>
<tr>
<td>Maximum</td>
<td>365.8</td>
<td>186.20</td>
<td>90.02</td>
</tr>
</tbody>
</table>

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...
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


Pennsylvania (USA) Through Analysis of Barium and Sulfate Concentrations," *Environmental Geochemistry and Health*, 40, 865–885. [1]


