Model Based Clustering of **Nonparametric Weighted Networks**

Amal Agarwal

Joint with Lingzhou Xue

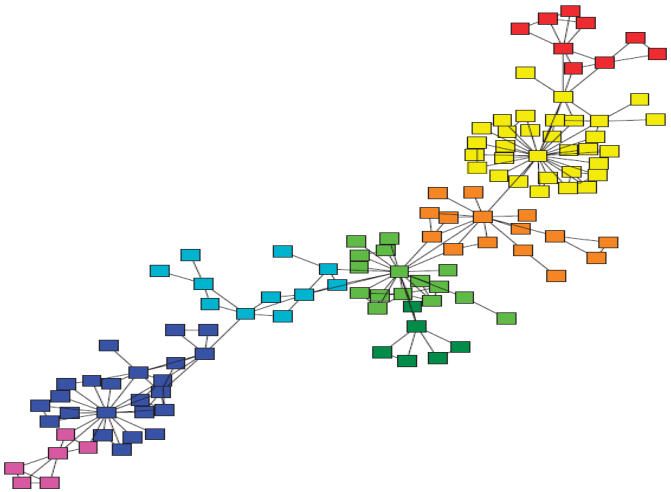
PENNSTATE

1855

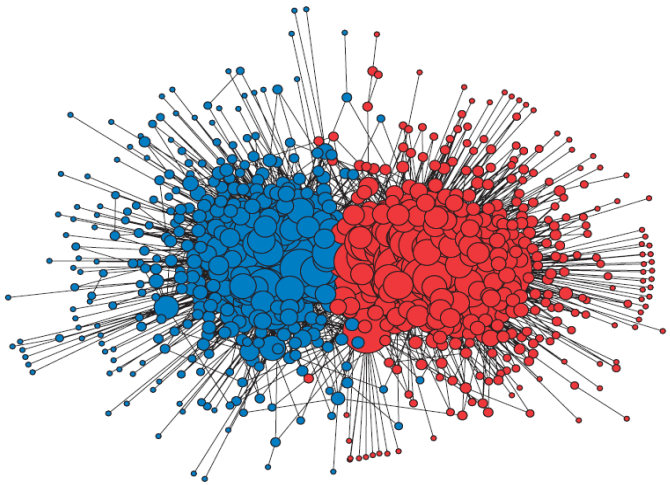
July 30, 2018

Clustering networks is synonymous to finding groups of nodes exhibiting similar behaviour.

- Clustering based on algorithm optimizing community criterion
- Clustering based on probability model
  - Stochastic Block Model (SBM)
  - Exponential-family Random Graph Model (ERGM)



Network of collaborations among scientists *(Newman 2011)*



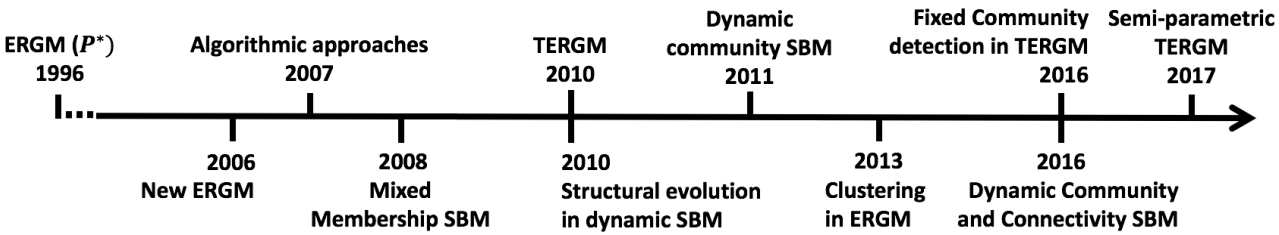
Network of links between US political blogs *(Adamic and Glance 2005)*

Exponential-family Random Graph Model (ERGM)

- **ERGM can be written in the following form:** *(Wasserman and Pattison 1996)*

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \exp\{\boldsymbol{\theta}' \boldsymbol{g}(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\}$$

- $\mathbf{Y} = (Y_{ij})_{1 \le i,j \le n}$ is a random network edge indicator adjacency matrix.
- $\boldsymbol{\theta}$ are network parameters, $\boldsymbol{g}(\mathbf{y})$ is sufficient network statistics and $\psi(\boldsymbol{\theta})$ is the normalization constant.



- **Weighted Random Networks:** *(Ambroise and Matias 2012)*

$$P_\beta(\boldsymbol{Y}_{ij} = \boldsymbol{y}_{ij} | \mathbf{Z}_{ik}\mathbf{Z}_{jl} = 1) = p_{kl} f(\cdot, \beta_{kl}) + (1 - p_{kl})\delta_0(\cdot)$$

- $\boldsymbol{Y} = (Y_{ij})_{1 \le i,j \le N}$ is a weighted network with $y_{ij} \in \mathbb{R}$ as weights.
- $Z_{ik} = 1$ iff node i lies in cluster k.

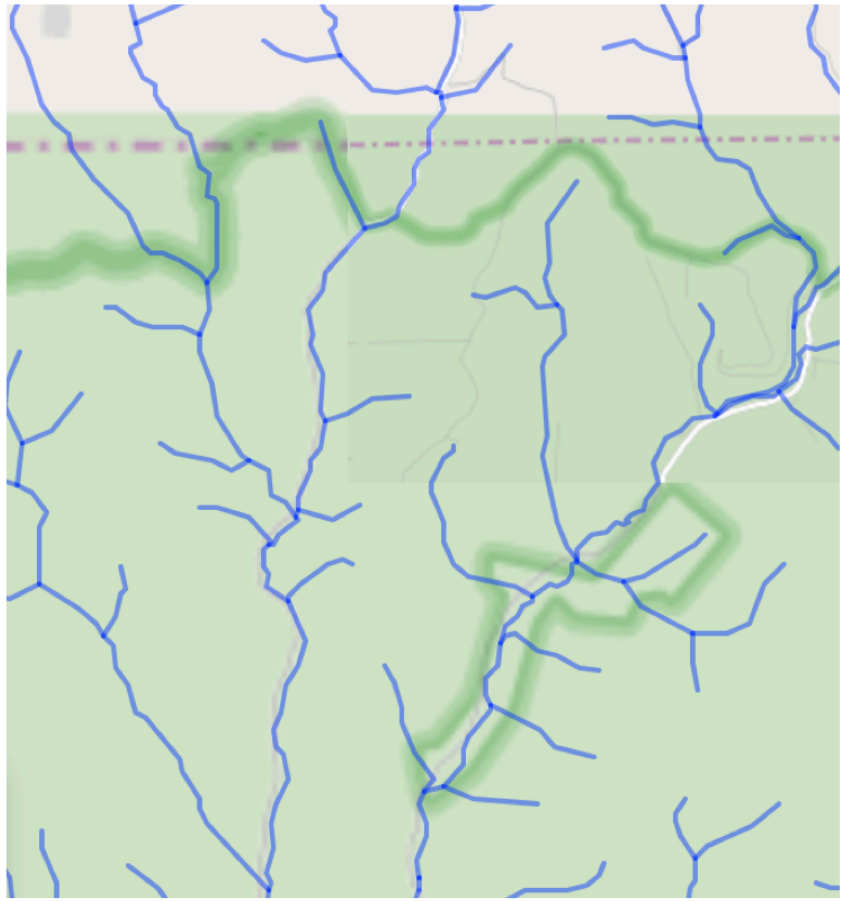
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

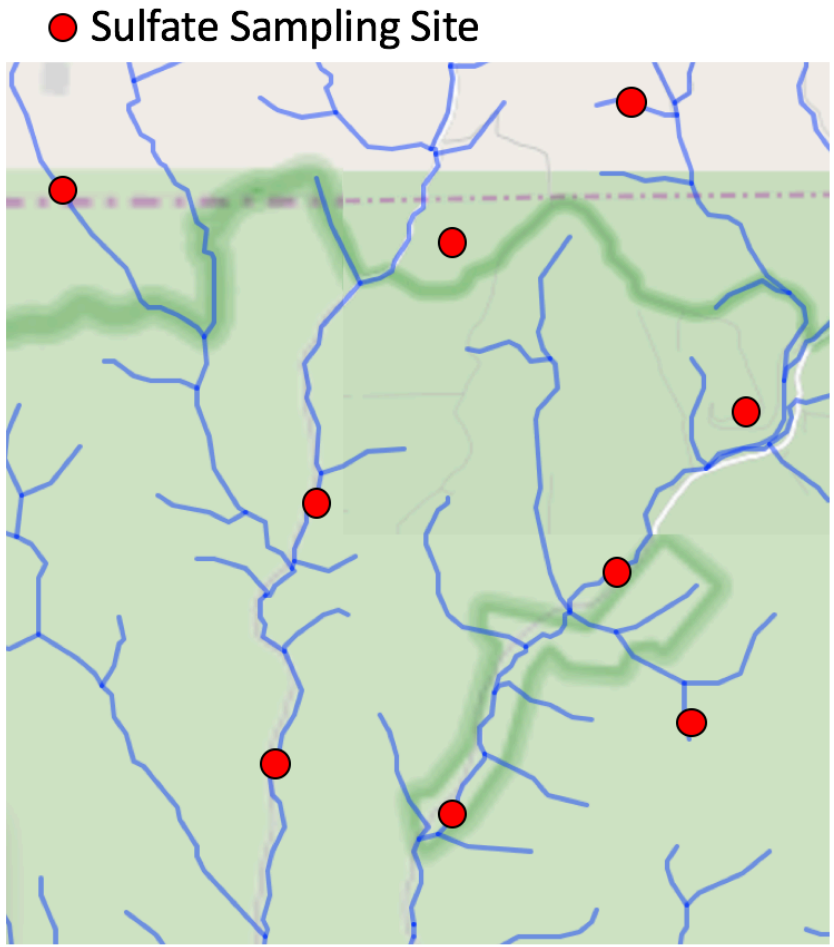
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

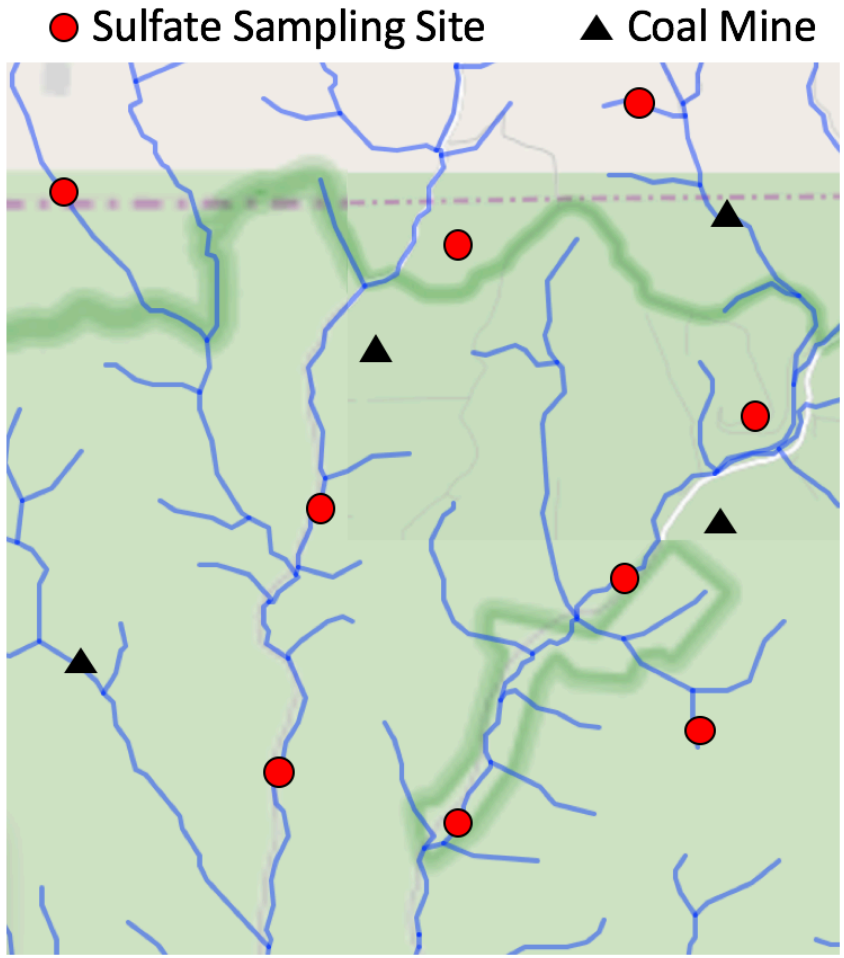
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

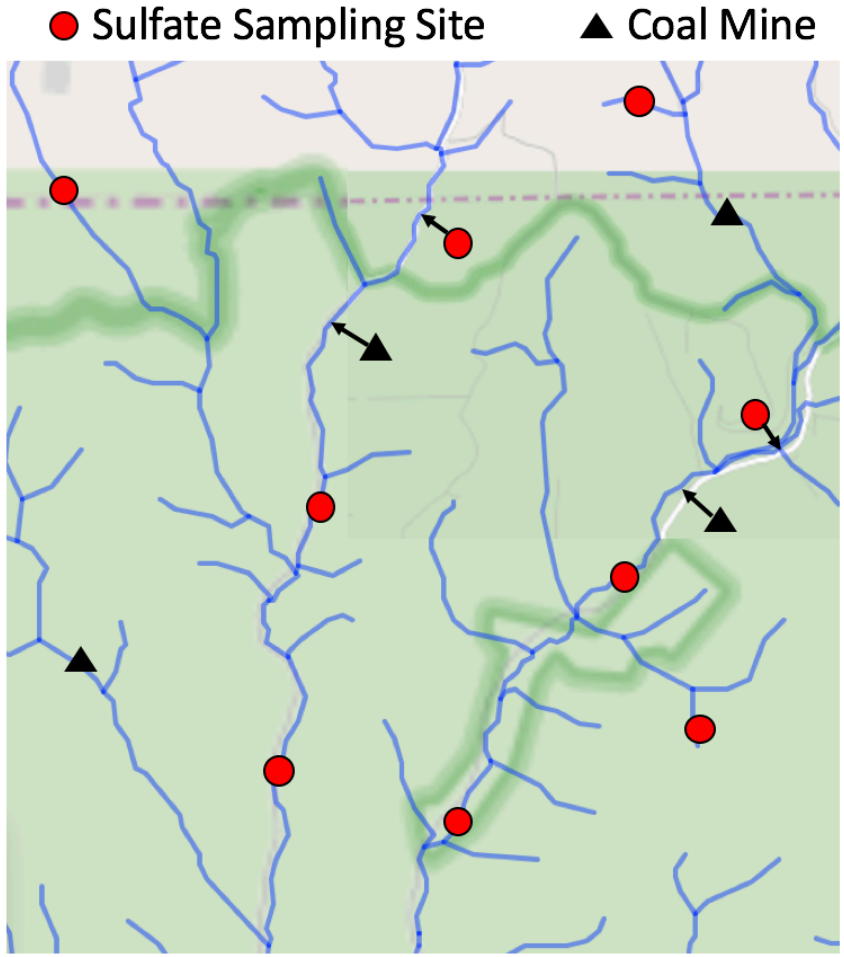
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

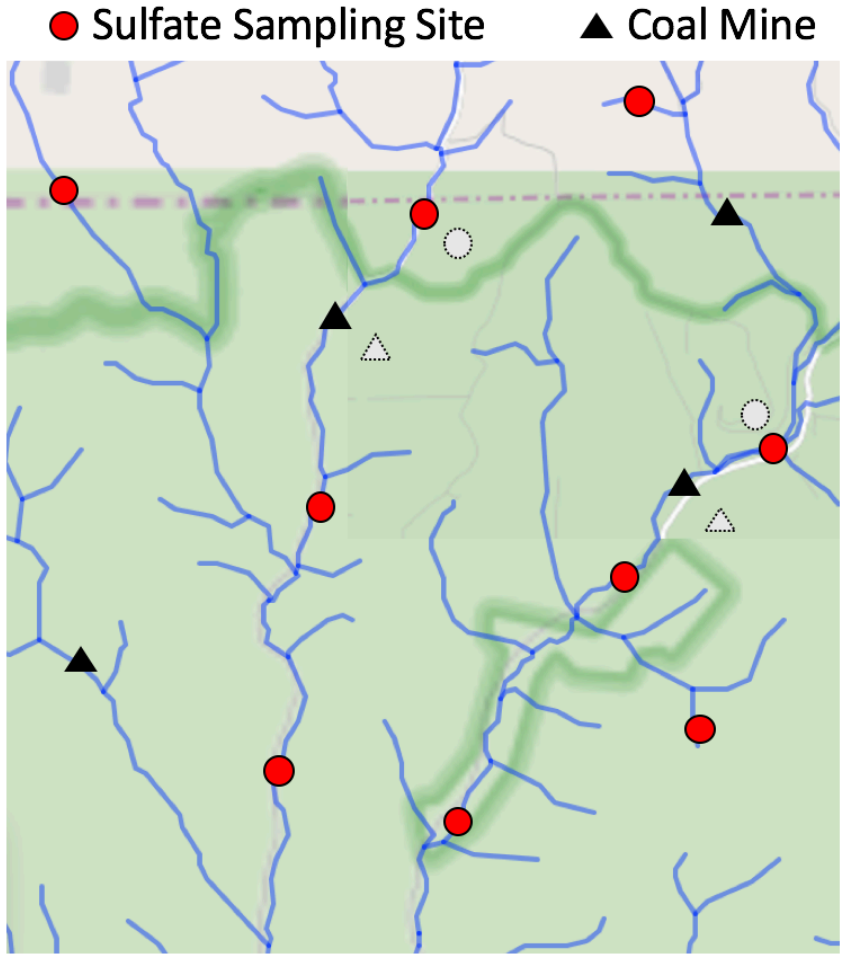
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

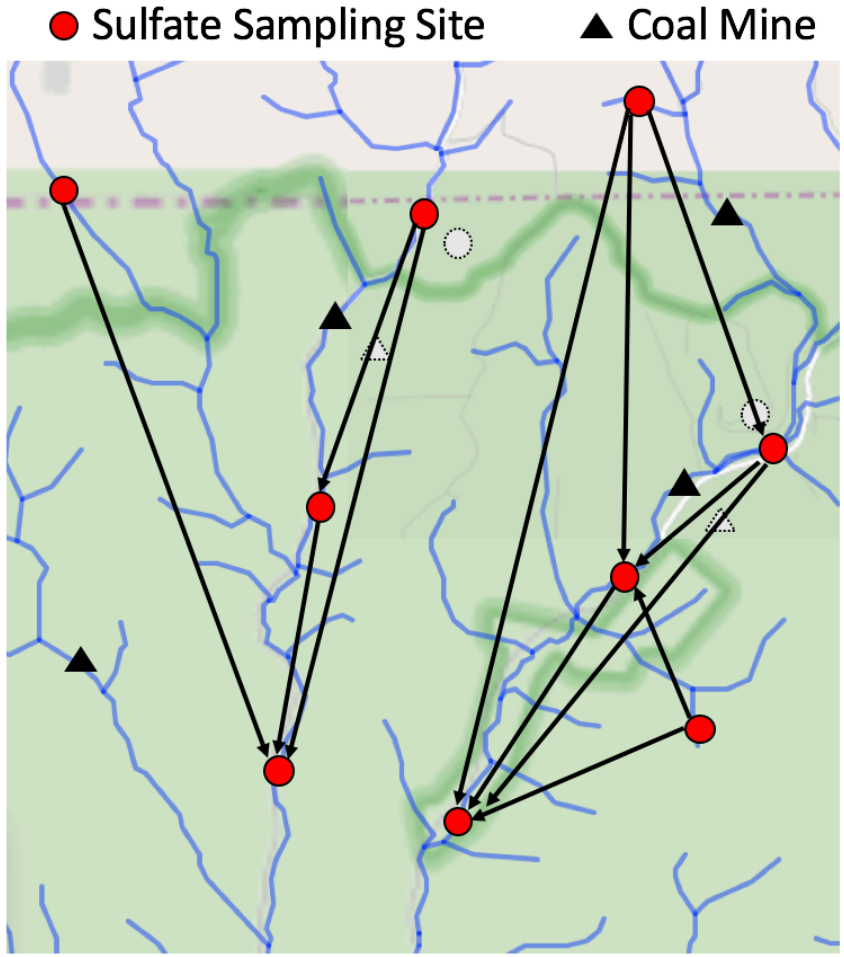
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

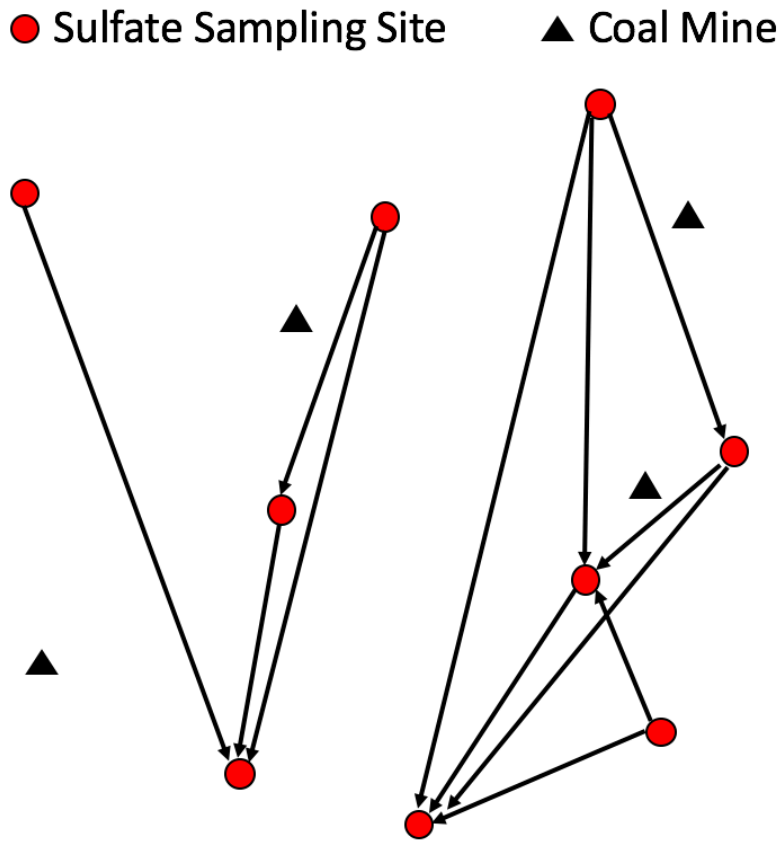
Water pollution is caused by high dissolved sulfate in river streams of Ohio Watershed, PA.

Integrating three topologies:

- Base river flow network
- $\sim$ **900** Sulfate sampling sites
- $\sim$ **100** Coal mines

Preprocessing:

- Mapping the sampling sites and coal mines on the base river network (measurement error).
- Directed weighted network over sampling sites based on river flow.



Different stages in preprocessing sulfate network

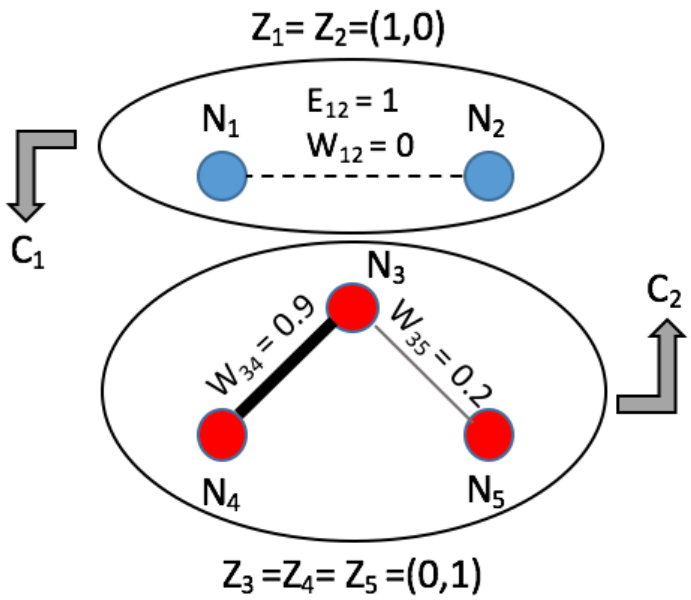
Major challenges for clustering weighted networks

1. How to incorporate weights in the network model?
2. How to estimate block densities in weighted networks?
3. How to choose an appropriate number of clusters in a weighted network model?
4. How to deal with computationally intractable log-likelihood?

Outline

Notations and Model Setup

- n: Number of nodes.
- K: Number of clusters.
- $\mathbf{Y} = (\mathbf{E}, \mathbf{W})$: Weighted Network
  - $\mathbf{E} = (E_{ij})_{1 \le i,j \le n}$: Edge indicator adjacency matrix.
  - $\mathbf{W} = (W_{ij})_{1 \le i,j \le N}$: Weight matrix.
- $\boldsymbol{Z} = (Z_{i,1 \le i \le n})$: Cluster memberships
- θ: Network parameters
- f_{kl}: Nonparametric density functions.
- π: Mixture proportions



An example of a clustered weighted network

Nonparametric weighted network model and local likelihood estimation for the degenerate $K = 1$ case

Log-likelihood of the nonparametric weighted network

$$\log(P_f(\mathbf{Y} = \mathbf{y})) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}_{e_{ij} \neq 0} \left[\log(f(w_{ij})) - \left(\int_{\mathcal{X}} f(u)du - 1 \right) \right] \text{ (Loader 1996)}$$

Local log-likelihood estimation through orthogonal basis approximation

$$\ell(\beta, w; \mathbf{Y}) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{1}_{e_{ij} \neq 0} \left[K_h(w_{ij} - w)\Phi^p(w_{ij} - w) - \left(\int_{\mathcal{X}} K_h(u - w) \exp(\Phi^p(u - w))du - 1 \right) \right]$$

Clustering nonparametric weighted network

Log-likelihood of the observed weighted network

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

Conditional dyadic independence *(Vu et al. 2013)*

$$P_{\theta,\boldsymbol{f}}(\mathbf{Y} = \mathbf{y} | \mathbf{Z} = \mathbf{z}) = \prod_{1 \le i < j \le n} P_{\theta_{z_i z_j}}(E_{ij} = e_{ij} \mid \mathbf{Z} = \mathbf{z}) f_{z_i z_j}(W_{ij} = w_{ij} \mid e_{ij} = 1, \mathbf{Z} = \mathbf{z})$$

For any given pair of nodes (i, j),

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

Local likelihood block density estimation in a clustered nonparametric weighted network

Conditional log-likelihood of the clustered weighted network

$$\log(P_{\theta,f}(\mathbf{Y}=\mathbf{y}\mid\mathbf{Z}=\mathbf{z}))=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\mathbb{1}_{e_{ij}\neq 0}\log(p_{z_iz_j})+\mathbb{1}_{e_{ij}=0}\log(1-p_{z_iz_j})\right]$$

$$+\mathbb{1}_{e_{ij}\neq 0}\left[\log(f_{z_iz_j}(w_{ij}))-\left(\int_{\mathcal{X}}f_{z_iz_j}(u)du-1\right)\right]$$

Local conditional log-likelihood estimation through orthogonal basis approximation

$$\ell(\theta,\beta,w;\boldsymbol{Y}\mid\boldsymbol{Z})=\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left[\mathbb{1}_{e_{ij}\neq 0}\log(p_{z_iz_j})+\mathbb{1}_{e_{ij}=0}\log(1-p_{z_iz_j})\right]+\mathbb{1}_{e_{ij}\neq 0}\times$$

$$\left[K_h(w_{ij}-w)\Phi^{p}_{z_iz_j}(w_{ij}-w)-\left(\int_{\mathcal{X}}K_h(u-w)\exp(\Phi^{p}_{z_iz_j}(u-w))du-1\right)\right]$$

Outline

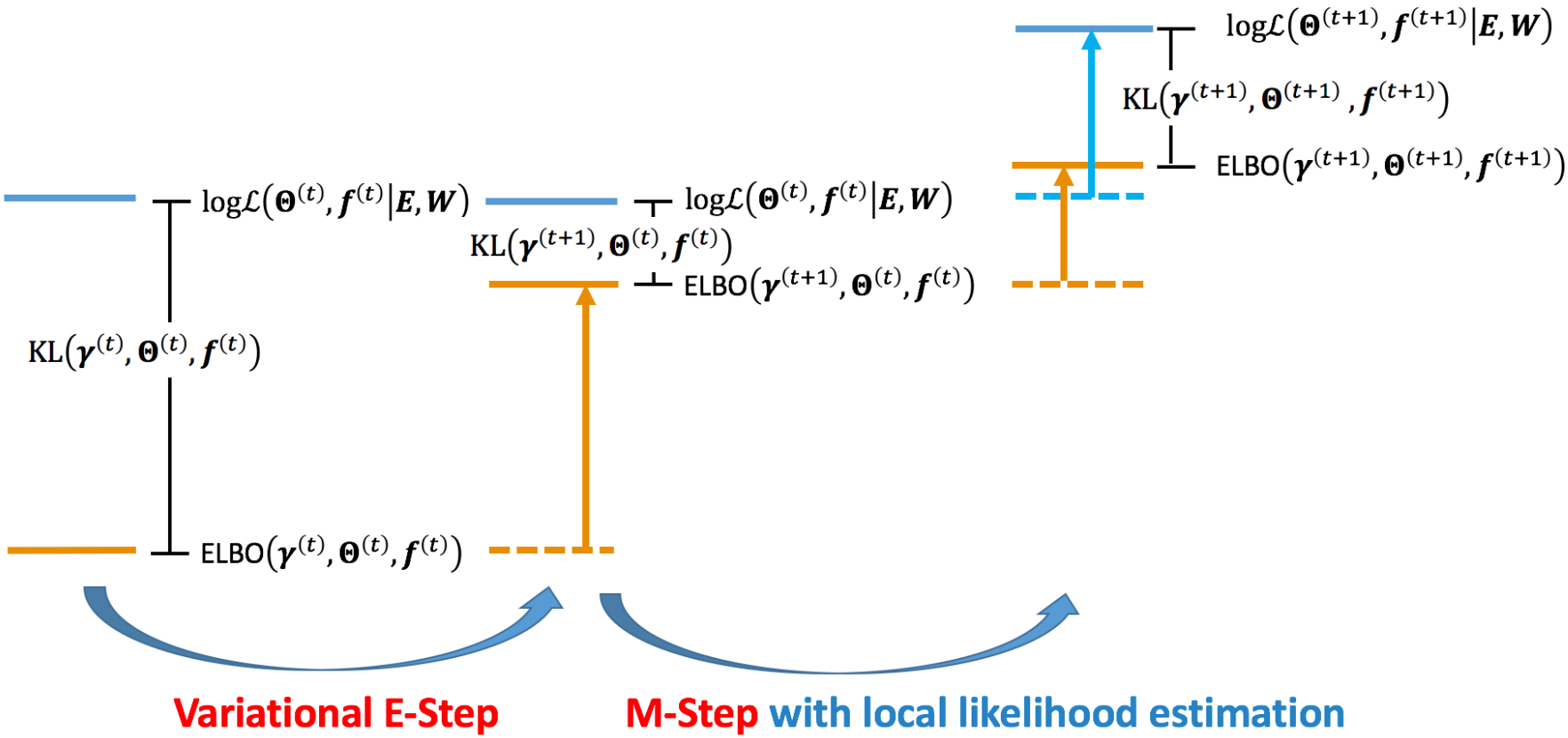
Variational EM Algorithm

- Why Variational EM (VEM) instead of EM?

1. VEM uses $A(\boldsymbol{z})$ to approximate $\mathbb{P}_\theta(\boldsymbol{Z}|\boldsymbol{Y})$ (e.g. mean field approximation)
2. Construct a tractable lower bound of the intractable log-likelihood (by using Jensen's inequality)
3. Maximize the lower bound, yielding approximate ML estimates. *(Wainwright and Jordan 2008, Hunter and Lange, 2004)*

- Variational inference: A review for statisticians *(Blei et al. 2017)*
- Theoretical and Computational Guarantees *(Zhang and Zhou, 2017)*

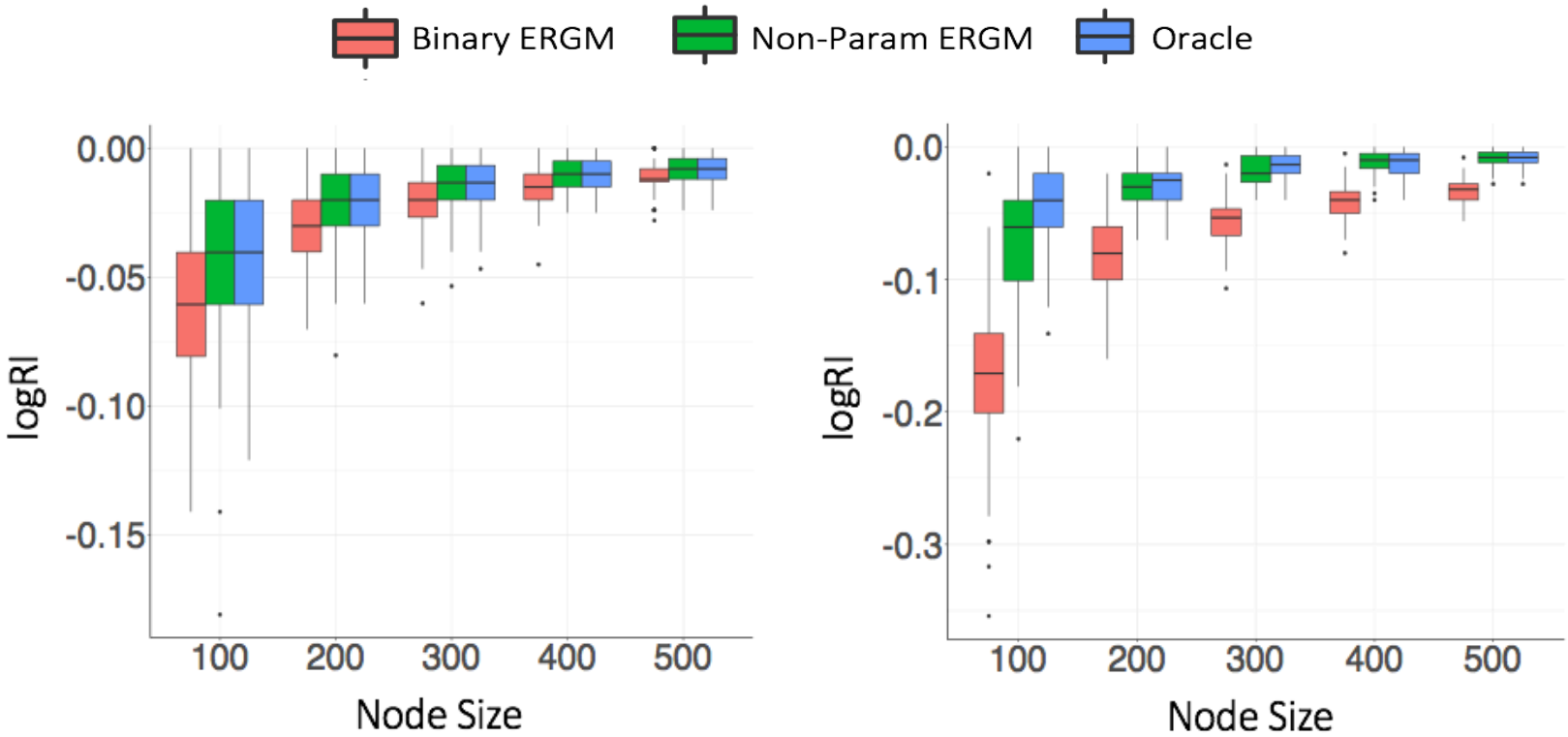
Variational EM algorithm follows ascent property!



Approximate MLE and ascent property

Outline

Clustering performance for our proposed nonparametric ERGM is comparable to Oracle!

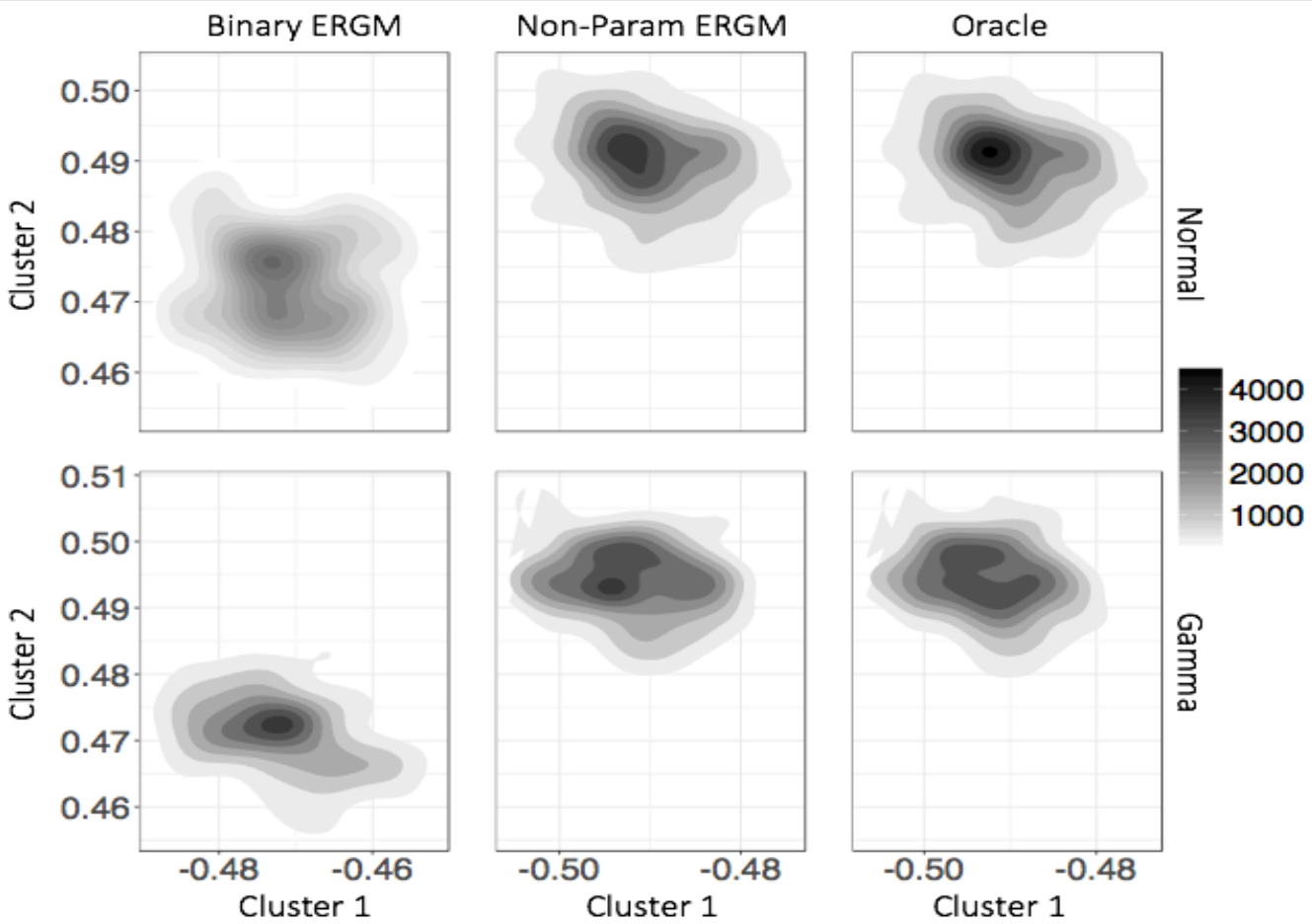


θ_{s_1} with Normal distributions

θ_{s_2} with Gamma distributions

Clustering Performance measured using logRI against different node sizes comparing the three models for different simulation settings

Estimation performance for our proposed nonparametric ERGM is comparable to Oracle!



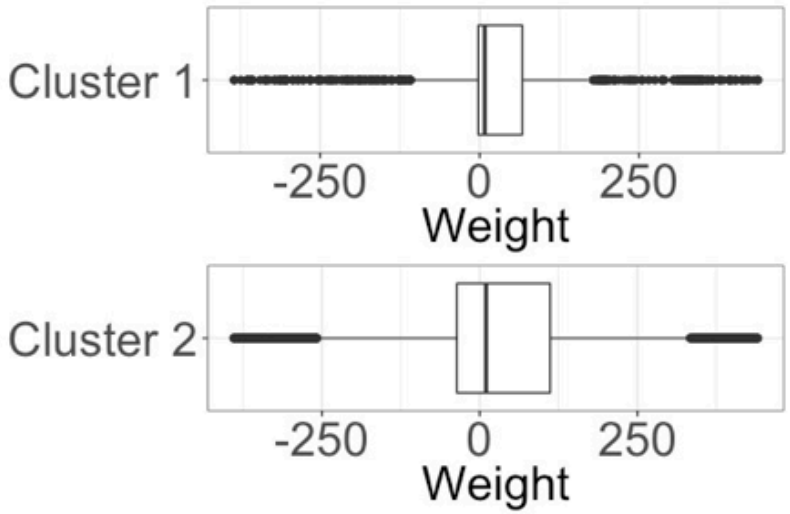
Plots of empirical joint distributions of network parameters θ_{S_2} over 100 simulations with 500 nodes, comparing the three models for different block distributions

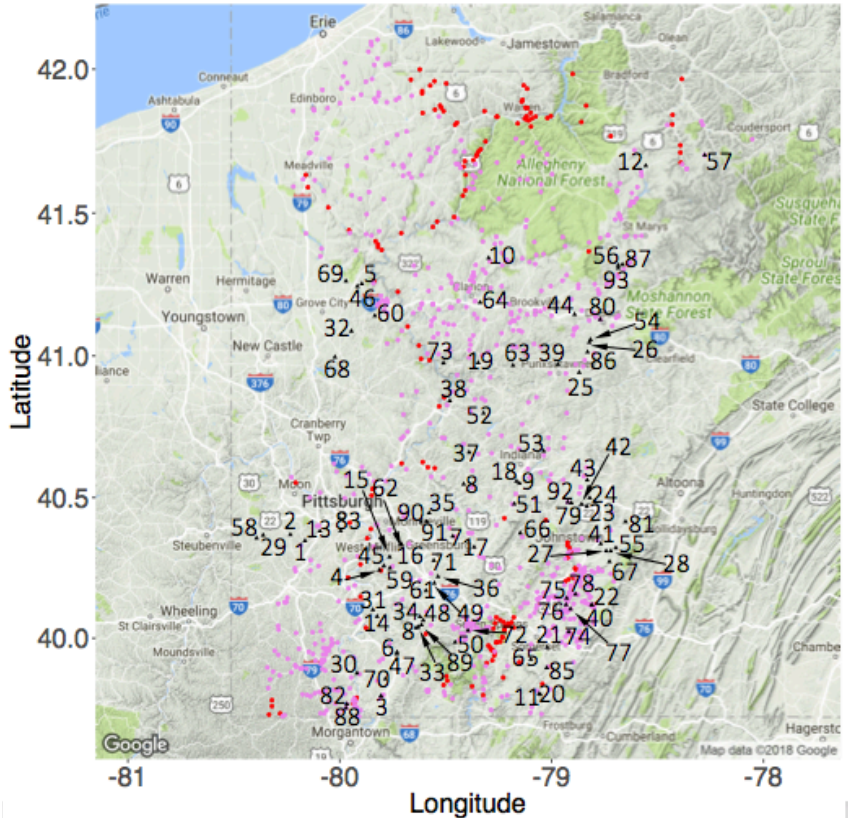
Outline

We detect **two** clusters in sulfate pollution network based on our proposed nonparametric ERGM

Summary Statistic	C_1	C_2
Number of nodes	147	718
Average Degree	18.72	5.21

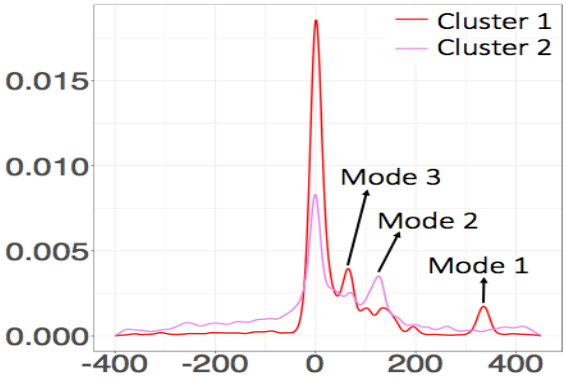
Descriptive statistics of the two clusters



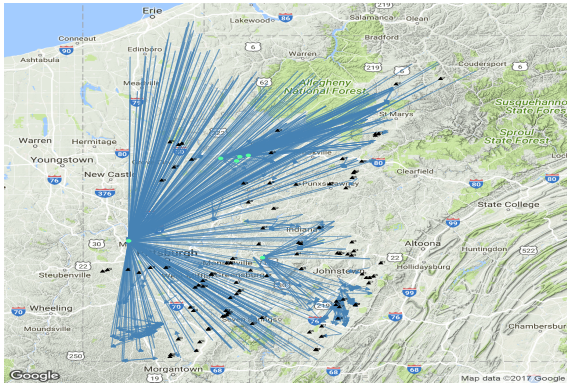


Clustered sulfate sampling sites with C_1 in red & C_2 in pink and coal mining sites in black (1-93)

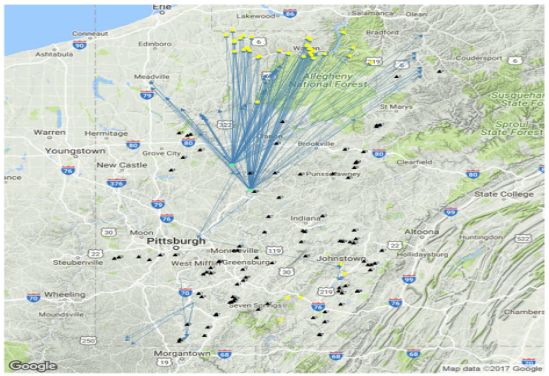
Nonparametrically estimated block densities and associated sub-networks



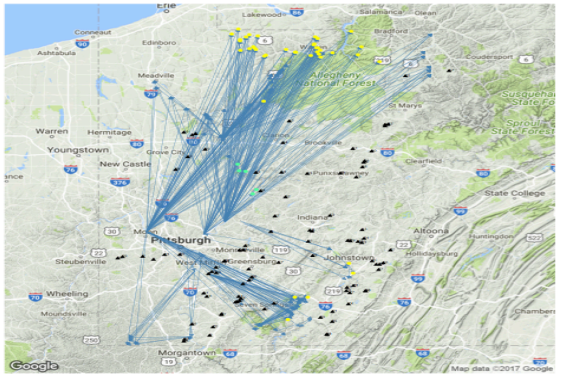
Non-parametrically estimated densities



Sub-network for Mode 2 in C_2



Sub-network for Mode 1 in C_1

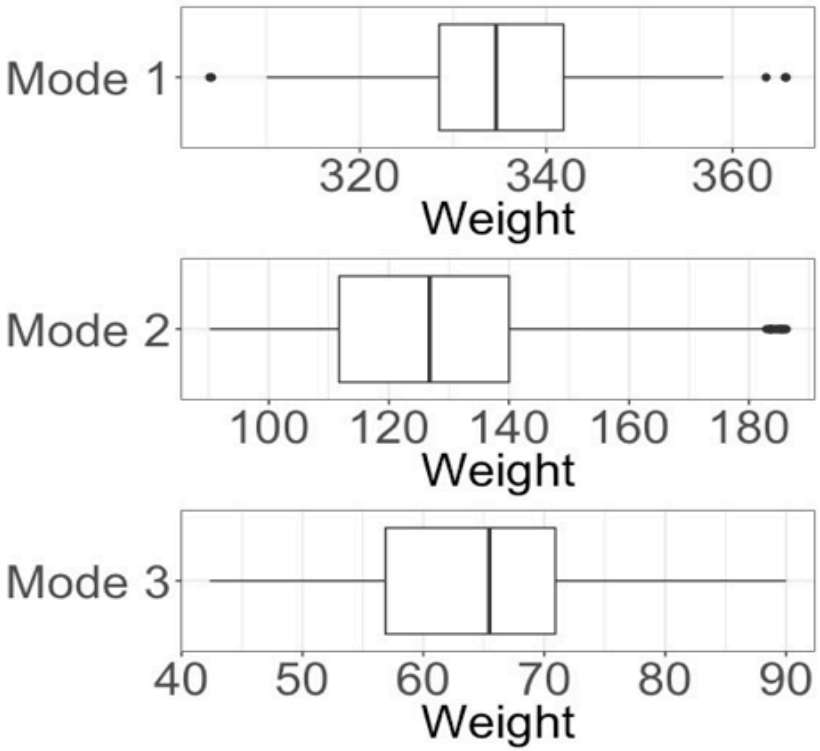


Sub-network for Mode 3 in C_1

Non-parametrically estimated densities for ties within clusters and subnetworks for different modes

Coal mines around high degree and high mode sub-networks are responsible for significant increase in river pollution!

- **Mode 1:** high degrees and higher differences of sulfate concentrations among adjacent nodes;
- **Mode 2:** sparse sub-regions with low degrees and moderate differences of sulfate concentrations;
- **Mode 3:** dense sub-regions with high degrees and low differences of sulfate concentrations.



Contributions

1. How to incorporate weights in the network model?
   - **Proposed a novel nonparametric ERGM for modeling weighted networks.**
2. How to estimate block densities in weighted networks?
   - **Introduced the notion of local log likelihood estimation without any distributional assumptions.**
3. How to choose an appropriate number of communities in the nonparametric setup?
   - **Proposed ICL to choose number of clusters.**
4. How to deal with computationally intractable log-likelihood?
   - **Designed an efficient VEM algorithm.**

References and Acknowledgements

- Link to **arXiv preprint**: https://arxiv.org/pdf/1712.07800.pdf
- **GeoNet**- A web application for detecting polluters in river networks: http://shiny.science.psu.edu/lxx6/geonet/
- Special thanks to **Susan Brantley** and her research group at Department of Geosciences, Pennstate.