Statistical Analysis of Network Data

James Boyle
supervised by George Bolt

September 4, 2020

Network $G = (V, E)$

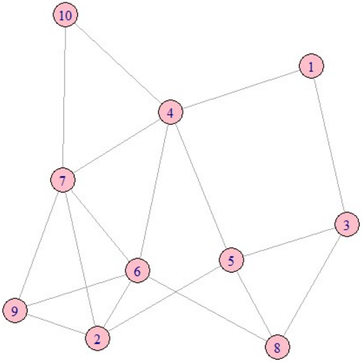


Figure: A graph with $N_V = 10$ vertices

Network $G = (V, E)$

- Vertices $V = \{1, \ldots, N_V\}$

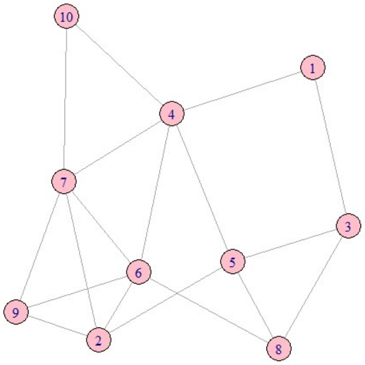


Figure: A graph with $N_V = 10$ vertices

Network Basics

Network $G = (V, E)$

- Vertices $V = \{1, \ldots, N_V\}$
- Edges $\{i, j\}$ joining vertices

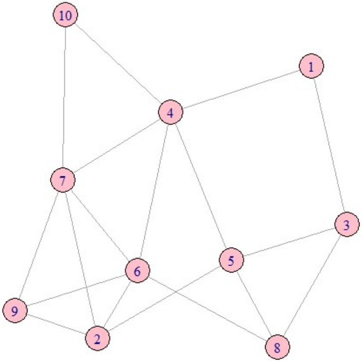
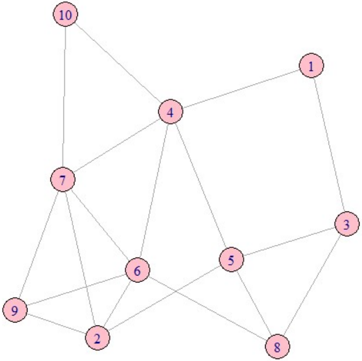


Figure: A graph with $N_V = 10$ vertices

Network $G = (V, E)$



Adjacency matrix $A \in \mathbb{R}^{N_V \times N_V}$

$$a_{ij} = \begin{cases} 1 & \text{if } \{i, j\} \in E, \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

Network Characteristics

Vertex *centrality*, measures of how “important” a vertex is:

Vertex *centrality*, measures of how “important” a vertex is:

- *degree* - number of edges incident to a vertex

Network Characteristics

Vertex *centrality*, measures of how "important" a vertex is:

- *degree* - number of edges incident to a vertex
- *closeness centrality* - $c_{cl}(v) = \frac{1}{\sum_{u \in V} \mathrm{d}(v,u)}$

Vertex *centrality*, measures of how “important” a vertex is:

- *degree* - number of edges incident to a vertex
- *closeness centrality* - $c_{cl}(v) = \frac{1}{\sum_{u \in V} \mathrm{d}(v,u)}$
- *betweenness centrality* - proportion of shortest paths between pairs of vertices passing through v

Random Graphs

Stochastic Block Model

Idea: Split the vertices into groups, and consider all vertices in a given group stochastically equivalent.

Stochastic Block Model

Idea: Split the vertices into groups, and consider all vertices in a given group stochastically equivalent.

Parameters: $N_V, K \in \mathbb{N}$, $B \in \mathbb{R}^{K \times K}$

Stochastic Block Model

Idea: Split the vertices into groups, and consider all vertices in a given group stochastically equivalent.

Parameters: $N_V, K \in \mathbb{N}$, $B \in \mathbb{R}^{K \times K}$

- Split the N_V vertices into K classes (*a priori* or at random). Class memberships $c = (c_1, \ldots, c_{N_V})$

Stochastic Block Model

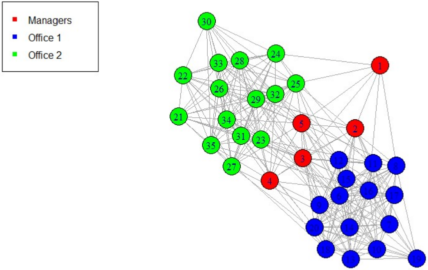
Idea: Split the vertices into groups, and consider all vertices in a given group stochastically equivalent.

Parameters: $N_V, K \in \mathbb{N}$, $B \in \mathbb{R}^{K \times K}$

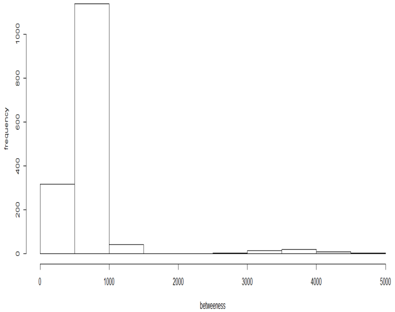
- Split the N_V vertices into K classes (*a priori* or at random). Class memberships $c = (c_1, \ldots, c_{N_V})$
- $\mathbb{P}(\text{edge between vertices } i \text{ and } j) = b_{c_i c_j}$

Stochastic Block Model - Example

- $N_V = 35$
- $K = 3$
- $B = \begin{pmatrix} 0.9 & 0.1 & 0.1 \\ 0.1 & 0.3 & 0.05 \\ 0.1 & 0.05 & 0.3 \end{pmatrix}$



Measures of Vertex Centrality



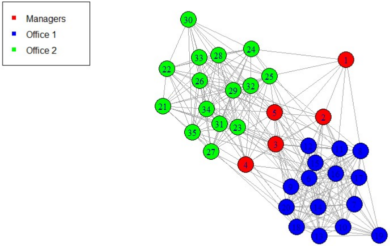


Figure: Betweenness Centrality

Measures of Vertex Centrality

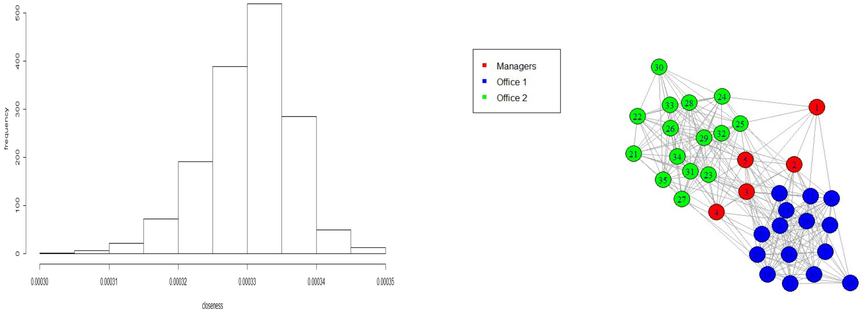


Figure: Closeness Centrality

Multiple Network Models

Single network models are in general not suitable for modelling multiple network observations, e.g. brain scans.

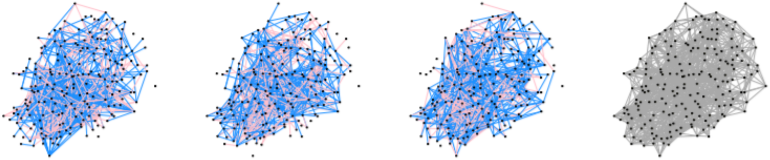


Figure: Brain Networks[4]

Multiple Network Models

Aim: Model multiple noisy realisations of a single "true" network, i.e. observations of the form

$$\text{True Network} + \text{Noise}$$

Multiple Network Models

Aim: Model multiple noisy realisations of a single “true” network, i.e. observations of the form

$$\text{True Network} + \text{Noise}$$

For a binary network, noise can only manifest itself in the form of false positive and false negative observations

The Measurement Error Model

Model Assumptions[3][1]:

- True network $A \sim \mathrm{StochasticBlockModel}(N_V, K, B)$

[1] C. M. Le, K. Levin, E. Levina, et al. Estimating a network from multiple noisy realizations.Electronic Journal of Statistics, 12(2):4697–4740, 2018

The Measurement Error Model

Model Assumptions[3][1]:

- True network $A \sim \mathrm{StochasticBlockModel}(N_V, K, B)$
- Observation noise $A^{(1)}, \ldots, A^{(n)}$ respects the block structure

[1] C. M. Le, K. Levin, E. Levina, et al. Estimating a network from multiple noisy realizations.Electronic Journal of Statistics, 12(2):4697–4740, 2018

The Measurement Error Model

Model Assumptions[3][1]:

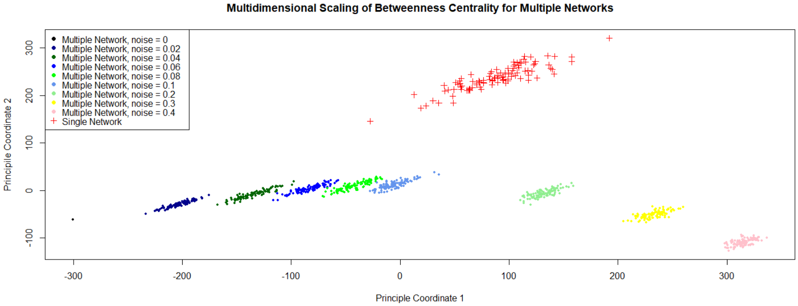
- True network $A \sim \mathrm{StochasticBlockModel}(N_V, K, B)$
- Observation noise $A^{(1)}, \dots, A^{(n)}$ respects the block structure

Concretely, letting $P, Q \in \mathbb{R}^{K \times K}$ be the matrices of false positive and false negative rates respectively, we suppose that

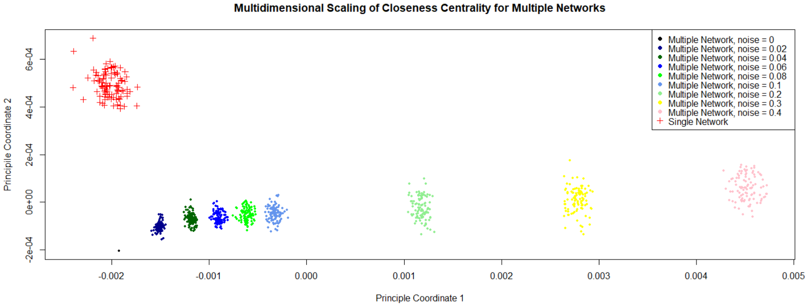
$$A_{ij}^{(m)} \sim \begin{cases} \mathrm{Bernoulli}(P_{c_i c_j}) & \text{if } A_{ij} = 0 \\ \mathrm{Bernoulli}(1 - Q_{c_i c_j}) & \text{if } A_{ij} = 1 \end{cases}$$

[1] C. M. Le, K. Levin, E. Levina, et al. Estimating a network from multiple noisy realizations.Electronic Journal of Statistics, 12(2):4697–4740, 2018

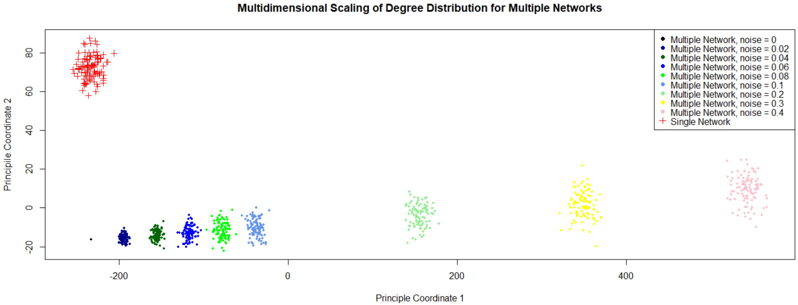
The Measurement Error Model



The Measurement Error Model



The Measurement Error Model



Metric Based Models

Idea[4][2]: Assign probabilities to networks based on their distance from a central, "true", network.

[2] Lunagomez S., Olhed, S. C., and Wolfe P. J. (2020). Modeling network populations via graph distances. Journal of the American Statistical Association (just-accepted):1–59

Metric Based Models

Idea[4][2]: Assign probabilities to networks based on their distance from a central, "true", network.

e.g. For a true network G^{true}, the **Spherical Network Model** assigns

$$\mathbb{P}(G; G^{true}, \gamma) \propto \exp(-\gamma d(G, G^{true}))$$

[2] Lunagomez S., Olhed, S. C., and Wolfe P. J. (2020). Modeling network populations via graph distances.Journal of the American Statistical Association (just-accepted):1–59

On Things not Covered

Network Path data - Each data point is a path through a network

e.g. vertex $\leftrightarrow$ webpage

edge $\leftrightarrow$ navigation by user between webpages

On Things not Covered

Network Path data - Each data point is a path through a network
e.g. vertex $\leftrightarrow$ webpage
edge $\leftrightarrow$ navigation by user between webpages

Inference - Very complicated, so approximate methods such as MCMCMLE or EMA must be used

On Things not Covered

Network Path data - Each data point is a path through a network
e.g. vertex $\leftrightarrow$ webpage
edge $\leftrightarrow$ navigation by user between webpages

Inference - Very complicated, so approximate methods such as MCMCMLE or EMA must be used

Single Network Models

On Things not Covered

Network Path data - Each data point is a path through a network
e.g. vertex $\leftrightarrow$ webpage
edge $\leftrightarrow$ navigation by user between webpages

Inference - Very complicated, so approximate methods such as MCMCMLE or EMA must be used

Single Network Models

Dynamic networks

Bibliography

B. Kim, K. H. Lee, L. Xue, and X. Niu.
A review of dynamic network models with latent variables.
Statistics surveys, 12:105, 2018.

E. D. Kolaczyk and G. Csárdi.
Statistical analysis of network data with R, volume 65.
Springer, 2014.

C. M. Le, K. Levin, E. Levina, et al.
Estimating a network from multiple noisy realizations.
Electronic Journal of Statistics, 12(2):4697–4740, 2018.

S. Lunagómez, S. C. Olhede, and P. J. Wolfe.
Modeling network populations via graph distances.
Journal of the American Statistical Association, (just-accepted):1–59, 2020.

M. Salter-Townshend, A. White, I. Gollini, and T. Murphy.
Review of statistical network analysis: Models, algorithms and software supplementary material.
2012.