HEART: Statistics and Data Science With Networks

Joshua Agterberg

Johns Hopkins University

Fall 2021

Outline

Outline

Random and Real World Graphs

Real World Graphs:

- Power-Laws
- Triangles
- Community Structure

Random and Real World Graphs

Real World Graphs:

- Power-Laws
- Triangles
- Community Structure

Random Graph Models:

- Sparsity (not scale-free networks)
- Community Structure (SBMs)
- Low-rank property (My work)

Low-Rank Random Graphs

- A common assumption to make is that $\mathbb{E}A$ is a low-rank matrix

Low-Rank Random Graphs

- A common assumption to make is that $\mathbb{E}A$ is a low-rank matrix
- Special cases:
  - Erdos-Renyi ($A_{ij} \sim \mathrm{Bernoulli}(p)$)
  - Stochastic blockmodel and variants

Low-Rank Random Graphs

- A common assumption to make is that $\mathbb{E}A$ is a low-rank matrix
- Special cases:
  - Erdos-Renyi ($A_{ij} \sim \mathrm{Bernoulli}(p)$)
  - Stochastic blockmodel and variants
- Allows us to use dimensionality reduction techniques based on *spectral embeddings*

Outline

Stochastic Blockmodels

- Each edge probability depends only on community membership

Stochastic Blockmodels

- Each edge probability depends only on community membership
- Allows for more generality than E-R

Stochastic Blockmodels

- Each edge probability depends only on community membership
- Allows for more generality than E-R
- Important theoretical model to understand performance

Stochastic Blockmodels

- Each edge probability depends only on community membership
- Allows for more generality than E-R
- Important theoretical model to understand performance
- Doesn’t capture within-vertex heterogeneity, since two vertices in the same community are “equivalent”

Degree-Corrected Stochastic Blockmodels

- If vertex i and j belong to communities k and l, the DCSBM is defined by

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \mathbf{B}_{kl}$$

Degree-Corrected Stochastic Blockmodels

- If vertex i and j belong to communities k and l, the DCSBM is defined by

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \mathbf{B}_{kl}$$

- Allows for vertex-specific heterogeneity by a degree-correction factor θ_i associated to vertex i

Degree-Corrected Stochastic Blockmodels

- If vertex i and j belong to communities k and l, the DCSBM is defined by

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \mathbf{B}_{kl}$$

- Allows for vertex-specific heterogeneity by a degree-correction factor θ_i associated to vertex i
- Special Case: $\theta_i = 1$ for all i

Degree-Corrected Stochastic Blockmodels

- If vertex i and j belong to communities k and l, the DCSBM is defined by

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \mathbf{B}_{kl}$$

- Allows for vertex-specific heterogeneity by a degree-correction factor θ_i associated to vertex i
- Special Case: $\theta_i = 1$ for all i
- Satisfies $\mathbb{E}A = \Theta Z B Z^\top \Theta$ (low-rank property)

Degree-Corrected Stochastic Blockmodels

- If vertex i and j belong to communities k and l, the DCSBM is defined by

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j \mathbf{B}_{kl}$$

- Allows for vertex-specific heterogeneity by a degree-correction factor θ_i associated to vertex i
- Special Case: $\theta_i = 1$ for all i
- Satisfies $\mathbb{E}A = \Theta Z B Z^\top \Theta$ (low-rank property)
- Empirical Work

Mixed-Membership Stochastic Blockmodels

- Allows each vertex to belong to a weighted combination of communities

Mixed-Membership Stochastic Blockmodels

- Allows each vertex to belong to a weighted combination of communities
- If vertex i belongs to community k with weight w_k, and vertex j belongs to community l with weight w_l, then

$$\mathbb{P}(A_{ij} = 1) = \sum_{k,l} w_k w_l B_{kl}$$

Mixed-Membership Stochastic Blockmodels

- Allows each vertex to belong to a weighted combination of communities
- If vertex i belongs to community k with weight w_k, and vertex j belongs to community l with weight w_l, then

$$\mathbb{P}(A_{ij} = 1) = \sum_{k,l} w_k w_l B_{kl}$$

- Require $\sum_k w_k = 1$

Mixed-Membership Stochastic Blockmodels

- Allows each vertex to belong to a weighted combination of communities
- If vertex i belongs to community k with weight w_k, and vertex j belongs to community l with weight w_l, then

$$\mathbb{P}(A_{ij} = 1) = \sum_{k,l} w_k w_l B_{kl}$$

- Require $\sum_k w_k = 1$
- Special case: $w_k = 1$ for some k and $w_{k'} = 0$ for $k' \neq k$

Mixed-Membership Stochastic Blockmodels

- Allows each vertex to belong to a weighted combination of communities
- If vertex i belongs to community k with weight w_k, and vertex j belongs to community l with weight w_l, then

$$\mathbb{P}(A_{ij} = 1) = \sum_{k,l} w_k w_l B_{kl}$$

- Require $\sum_k w_k = 1$
- Special case: $w_k = 1$ for some k and $w_{k'} = 0$ for $k' \neq k$
- Satisfies $\mathbb{E}A = WBW^\top$

Outline

Random Dot Product Graphs

- Motivated by idea that each vertex has a vector in a *latent space* associated to it

Random Dot Product Graphs

- Motivated by idea that each vertex has a vector in a *latent space* associated to it
- If vertex i has vector X_i and vertex j has vector X_j, then

$$\mathbb{P}(A_{ij} = 1) = \langle X_i, X_j \rangle.$$

Random Dot Product Graphs

- Motivated by idea that each vertex has a vector in a *latent space* associated to it
- If vertex i has vector X_i and vertex j has vector X_j, then

$$\mathbb{P}(A_{ij} = 1) = \langle X_i, X_j \rangle.$$

- Need the above to be in $[0, 1]$

Random Dot Product Graphs

- Motivated by idea that each vertex has a vector in a *latent space* associated to it
- If vertex i has vector X_i and vertex j has vector X_j, then

$$\mathbb{P}(A_{ij} = 1) = \langle X_i, X_j \rangle.$$

- Need the above to be in $[0, 1]$
- Special Case: X_i depends only on community membership

Random Dot Product Graphs

- Motivated by idea that each vertex has a vector in a *latent space* associated to it
- If vertex i has vector X_i and vertex j has vector X_j, then

$$\mathbb{P}(A_{ij} = 1) = \langle X_i, X_j \rangle.$$

- Need the above to be in $[0, 1]$
- Special Case: X_i depends only on community membership
- Satisfies $\mathbb{E}A = XX^\top$ (psd matrix)

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$
- Very general model

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$
- Very general model
- Special cases:
  - $\mathbb{E}A$ is psd

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$
- Very general model
- Special cases:
  - $\mathbb{E}A$ is psd
  - $\mathbb{E}A = WBW^\top$

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$
- Very general model
- Special cases:
  - $\mathbb{E}A$ is psd
  - $\mathbb{E}A = WBW^\top$
  - $\mathbb{E}A = ZBZ^\top$

Low-Rank Random Graphs

- Only require that $\mathbb{E}A$ is low-rank, and edges are independent for $i \leq j$
- Very general model
- Special cases:
  - $\mathbb{E}A$ is psd
  - $\mathbb{E}A = WBW^\top$
  - $\mathbb{E}A = ZBZ^\top$
  - $\mathbb{E}A = \Theta ZBZ^\top \theta$

General Latent-Space Models

- Require that $\mathbb{P}(A_{ij} = 1) = f(X_i, X_j)$ for some function f called the *link function*

General Latent-Space Models

- Require that $\mathbb{P}(A_{ij} = 1) = f(X_i, X_j)$ for some function f called the *link function*
- Much broader set of models

General Latent-Space Models

- Require that $\mathbb{P}(A_{ij} = 1) = f(X_i, X_j)$ for some function f called the *link function*
- Much broader set of models
- Problem: before we had a known link function, now we need to estimate it...

General Latent-Space Models

- Require that $\mathbb{P}(A_{ij} = 1) = f(X_i, X_j)$ for some function f called the *link function*
- Much broader set of models
- Problem: before we had a known link function, now we need to estimate it...
- Allows for lots of variability in probabilities, but hard to analyze...

Outline

Problems with statistical models

- As with all statistical models, more freedom in design makes analysis difficult

Problems with statistical models

- As with all statistical models, more freedom in design makes analysis difficult
- The problem with low-rank graphs and triangles

Problems with statistical models

- As with all statistical models, more freedom in design makes analysis difficult
- The problem with low-rank graphs and triangles
- Dense graphs cannot have power laws (average expected degree grows proportional to $\log(n)$, but power laws it grows constant in n)

Problems with statistical models

- As with all statistical models, more freedom in design makes analysis difficult
- The problem with low-rank graphs and triangles
- Dense graphs cannot have power laws (average expected degree grows proportional to $\log(n)$, but power laws it grows constant in n)
- Important math problem: Fundamental limits of recovery

Problems with statistical models

- As with all statistical models, more freedom in design makes analysis difficult
- The problem with low-rank graphs and triangles
- Dense graphs cannot have power laws (average expected degree grows proportional to $\log(n)$, but power laws it grows constant in n)
- Important math problem: Fundamental limits of recovery
- Important practical problem: for low-rank graphs we still need to choose the rank!

Choosing the rank

- Scree plot: plot the eigenvalues of A in descending order
- Look for an "elbow" (Zhu and Ghodsi)

Choosing the rank

- Scree plot: plot the eigenvalues of A in descending order
- Look for an “elbow” (Zhu and Ghodsi)
- Universal Singular Value Thresholding (USVT)

Choosing the rank

- Scree plot: plot the eigenvalues of A in descending order
- Look for an "elbow" (Zhu and Ghodsi)
- Universal Singular Value Thresholding (USVT)
- More complicated methods?
- Equation (27)