HEART: Statistics and Data Science With Networks

Joshua Agterberg

Johns Hopkins University

Fall 2022

Outline

1. What We've Learned
2. Open Problems in Statistical Network Analysis
3. Overall Perspective of Data Science

Outline

1. What We've Learned
2. Open Problems in Statistical Network Analysis
3. Overall Perspective of Data Science

Basic Stuff: Probability and Linear Algebra

We discussed:

- Probability: necessary to understand Bernoulli random variables (e.g. edges of a graph)

Basic Stuff: Probability and Linear Algebra

We discussed:

- Probability: necessary to understand Bernoulli random variables (e.g. edges of a graph)
- Linear algebra: eigenvectors and eigenvalues (to find graph embeddings)

Basic Stuff: Probability and Linear Algebra

We discussed:

- Probability: necessary to understand Bernoulli random variables (e.g. edges of a graph)
- Linear algebra: eigenvectors and eigenvalues (to find graph embeddings)
- Also discussed notions from graph theory (Adjacency matrix, Laplacian matrix, etc.)

Random Graph Models

Common models we discussed:

- Erdos-Renyi Graph

Random Graph Models

Common models we discussed:

- Erdos-Renyi Graph
- Stochastic Blockmodel

Random Graph Models

Common models we discussed:

- Erdos-Renyi Graph
- Stochastic Blockmodel
- Degree-Corrected SBM

Network Analysis Pipeline

- Step 0: get the adjacency matrix

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by):

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
- Step 2:

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an “elbow” in the scree plot.
- Step 2: compute those eigenvectors!

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an “elbow” in the scree plot.
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
- Step 3: Cluster!

Network Analysis Pipeline

- Step 0: get the adjacency matrix
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
- Step 3: Cluster!

Knobs to Tweak

- Step 0: get the adjacency matrix

Knobs to Tweak

- Step 0: get the adjacency matrix
  - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.

Knobs to Tweak

- Step 0: get the adjacency matrix
    - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an “elbow” in the scree plot.
    - Could use automated methods to find an elbow!
    - Could use “eyeball” method!

Knobs to Tweak

- Step 0: get the adjacency matrix
  - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an “elbow” in the scree plot.
  - Could use automated methods to find an elbow!
  - Could use “eyeball” method!
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length

Knobs to Tweak

- Step 0: get the adjacency matrix
  - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
  - Could use automated methods to find an elbow!
  - Could use "eyeball" method!
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
  - Could not do this at all!
  - Could normalize in a different way!

Knobs to Tweak

- Step 0: get the adjacency matrix
  - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
  - Could use automated methods to find an elbow!
  - Could use "eyeball" method!
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
  - Could not do this at all!
  - Could normalize in a different way!
- Step 3: Cluster!

Knobs to Tweak

- Step 0: get the adjacency matrix
    - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
    - Could use automated methods to find an elbow!
    - Could use "eyeball" method!
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
    - Could not do this at all!
    - Could normalize in a different way!
- Step 3: Cluster!
    - Could choose a different algorithm for clustering!

Knobs to Tweak

- Step 0: get the adjacency matrix
  - Could use another matrix here!
- Step 1: choose the number of clusters (embedding dimension by): finding an "elbow" in the scree plot.
  - Could use automated methods to find an elbow!
  - Could use "eyeball" method!
- Step 2: compute those eigenvectors!
- Step 2.5: Normalize the eigenvectors by their length
  - Could not do this at all!
  - Could normalize in a different way!
- Step 3: Cluster!
  - Could choose a different algorithm for clustering!
- Step 4: do something besides cluster!

Multiple Graphs

- Frontier of network literature!

Multiple Graphs

- Frontier of network literature!
- Ways to aggregate networks:
  - Averaging!
  - Averaging the squared adjacency matrices!
  - Running the whole "pipeline" on one network and then using the singular vectors of the concatenated eigenvectors!

Outline

1. What We've Learned
2. Open Problems in Statistical Network Analysis
3. Overall Perspective of Data Science

Open Problems

- Covariates?

Open Problems

- Covariates?
- Edge weights?

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence
- Hypothesis Testing

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence
- Hypothesis Testing
- Hypergraphs

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence
- Hypothesis Testing
- Hypergraphs
- Graph models that model real world networks

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence
- Hypothesis Testing
- Hypergraphs
- Graph models that model real world networks
- Multiple graph models that model real-world networks

Reconciling Theory and Practice

- Triangles

Reconciling Theory and Practice

- Triangles
- Sparse networks

Reconciling Theory and Practice

- Triangles
- Sparse networks
- “Heterophily” versus “homophily”

Reconciling Theory and Practice

- Triangles
- Sparse networks
- "Heterophily" versus "homophily"
- Is a community model the "right" model?

Outline

1. What We've Learned
2. Open Problems in Statistical Network Analysis
3. Overall Perspective of Data Science

Data Science

This class focused on:

- Unsupervised Learning

Data Science

This class focused on:

- Unsupervised Learning
- Spectral methods (e.g. here)

Data Science

This class focused on:

- Unsupervised Learning
- Spectral methods (e.g. here)

Can also study

- Supervised Learning

Data Science

This class focused on:

- Unsupervised Learning
- Spectral methods (e.g. here)

Can also study

- Supervised Learning
- Manifold Learning (e.g. here and here)
- Deep learning...