HEART: Statistics and Data Science With Networks

Joshua Agterberg

Johns Hopkins University

Fall 2021

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
- Mixed-Membership SBM

Random Graph Models

Common models we discussed:

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

Random Graph Models

Common models we discussed:

- Erdos-Renyi Graph
- Stochastic Blockmodel
- Mixed-Membership SBM
- Degree-Corrected SBM
- More General Models (RDPGs, Graphons)

Random Graph Models

Network Models are *models*, so do not cover real-world problems. But real-world graphs are:

- Sparse

Random Graph Models

Network Models are *models*, so do not cover real-world problems. But real-world graphs are:

- Sparse
- Low-Rank

Random Graph Models

Network Models are *models*, so do not cover real-world problems. But real-world graphs are:

- Sparse
- Low-Rank
- Have triangles

Random Graph Models

Network Models are *models*, so do not cover real-world problems. But real-world graphs are:

- Sparse
- Low-Rank
- Have triangles

Need to choose a model to make things work, but you also need it to work well on real data!

Graph Clustering (and more) “Pipeline”

- First, choose the dimension *d* by finding an “elbow” from plotting the eigenvalues (scree plot)

Graph Clustering (and more) “Pipeline”

- First, choose the dimension d by finding an “elbow” from plotting the eigenvalues (scree plot)
- Obtain an $n \times d$ matrix called a“graph embedding” or “spectral embedding” by looking at the corresponding eigenvectors (maybe multiplying by eigenvalues)

Graph Clustering (and more) "Pipeline"

- First, choose the dimension d by finding an "elbow" from plotting the eigenvalues (scree plot)
- Obtain an $n \times d$ matrix called a"graph embedding" or "spectral embedding" by looking at the corresponding eigenvectors (maybe multiplying by eigenvalues)
- Cluster the *rows* of the graph embedding

Graph Clustering (and more) “Pipeline”

- First, choose the dimension d by finding an “elbow” from plotting the eigenvalues (scree plot)
- Obtain an $n \times d$ matrix called a“graph embedding” or “spectral embedding” by looking at the corresponding eigenvectors (maybe multiplying by eigenvalues)
- Cluster the *rows* of the graph embedding
- Can also treat the rows as data itself

Multiple Graphs

- We saw some different graph models

Multiple Graphs

- We saw some different graph models
- Frontier of network literature!

Multiple Graphs

- We saw some different graph models
- Frontier of network literature!
- Idea: get elbow as before, only now with a matrix created with all the adjacency matrices

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
- Bootstrapping networks:
  - Levin and Levina
  - Subgraph Count I
  - Subgraph Counts II
- Hypothesis Testing

Open Problems

- Covariates?
- Edge weights?
- Edge Covariates?
- Dependence
- Bootstrapping networks:
  - Levin and Levina
  - Subgraph Count I
  - Subgraph Counts II
- Hypothesis Testing
- Hypergraphs, multiple graphs, more...

Reconciling Theory and Practice

- Triangles

Reconciling Theory and Practice

- Triangles
- Removing the low-rank assumption

Reconciling Theory and Practice

- Triangles
- Removing the low-rank assumption
- Connecting these problems to other areas of data science...

Outline

1 What We've Learned

2 Open Problems in Statistical Network Analysis

3 Overall Perspective of Data Science

Deep Graph Learning

- Deep learning is modern, but we want things to be *principled*

Deep Graph Learning

- Deep learning is modern, but we want things to be *principled*
- Lots of recent work trying to understand deep learning (e.g. here)

Deep Graph Learning

- Deep learning is modern, but we want things to be *principled*
- Lots of recent work trying to understand deep learning (e.g. here)
- Can learn from deep learning and vice versa???

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)