S&DS 684: Statistical inference on graphs

Fall 2018

Administrivia

- Schedule: Wed 2:30-5pm, 24 Hillhouse Rm 107
- Instructor: Prof. Yihong Wu yihongwu@illinois.edu, Rm 235 Dunham Lab (10 Hillhouse)
  - Office hours: by appointment
- Website: http://stat.yale.edu/~yw562/684.html

Administrivia

1. Course prerequisites:

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, $\LaTeX$ template online

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, LaTeX template online
3. Homeworks (30%): three to four problem sets

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, LaTeX template online
3. Homeworks (30%): three to four problem sets
4. Final project (40%)

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, LaTeX template online
3. Homeworks (30%): three to four problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone reseach project.

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, LaTeX template online
3. Homeworks (30%): three to four problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone reseach project.
   - topics announced around week 6

Administrivia

1. Course prerequisites:
   - Maturity with probability theory, familiarity with mathematical statistics.
   - Linear algebra
2. Participation (30%):
   - attendance :-)
   - scribe: take turns, due in one week from the lecture, LaTeX template online
3. Homeworks (30%): three to four problem sets
4. Final project (40%)
   - either presenting paper(s) or a standalone reseach project.
   - topics announced around week 6
5. Materials: Lecture notes and additional reading materials will be posted online.

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - Q2 Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - Q2 Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?
- In this course:

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - Q2 Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?
- In this course:
  - Data = graph

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - Q2 Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?
- In this course:
  - Data = graph
  - Parameter = hidden (latent, or planted) structure

Statistical problems

- Statistical tasks: using data to make informed decisions (hypotheses testing, estimation, etc)

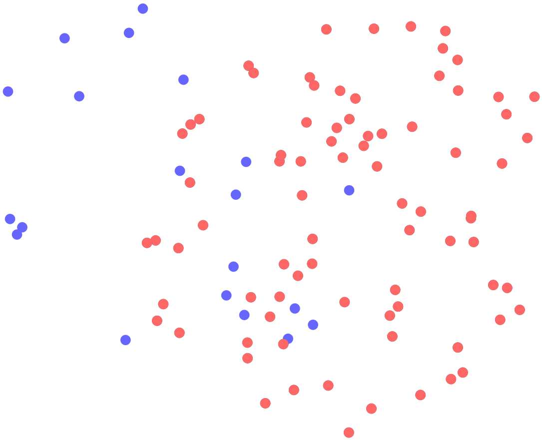
$$\underbrace{\theta \in \Theta}_{\text{parameter}} \mapsto \underbrace{X}_{\text{data}} \mapsto \underbrace{\hat{\theta}}_{\text{estimate}}$$

- Understanding the fundamental limits:
  - Q1 Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - Q2 Can statistical limits be attained computationally efficiently, e.g., in polynomial time? If yes, how? If not, why?
- In this course:
  - Data = graph
  - Parameter = hidden (latent, or planted) structure
  - Focus on large-graph limit (number of vertices $\to \infty$)

Planted clique – graph view

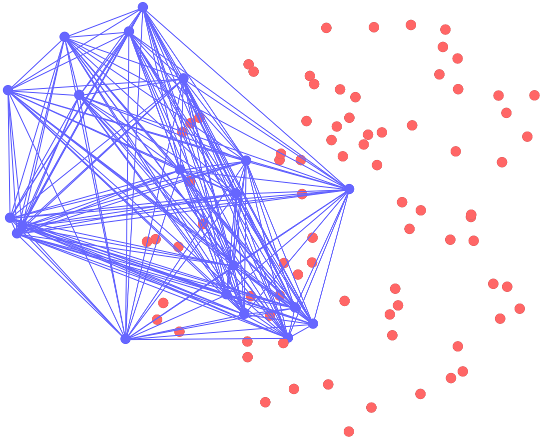
Planted clique – graph view

1. A clique of k vertices are chosen randomly to form a clique



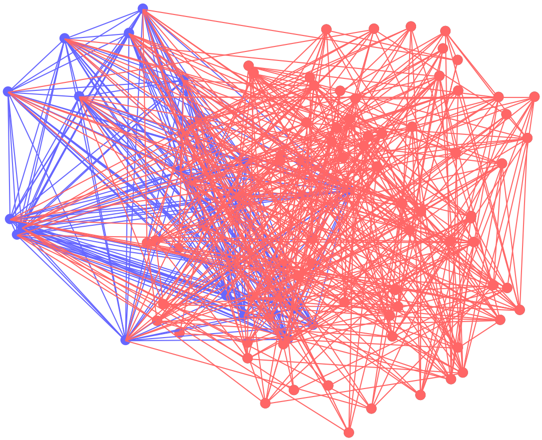
Planted clique – graph view

1. A clique of k vertices are chosen randomly to form a clique



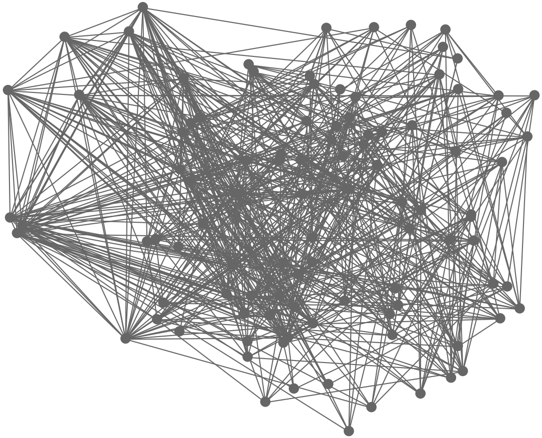
Planted clique – graph view

1. A clique of k vertices are chosen randomly to form a clique
2. For every other pair of nodes, add an edge w.p. $\frac{1}{2}$

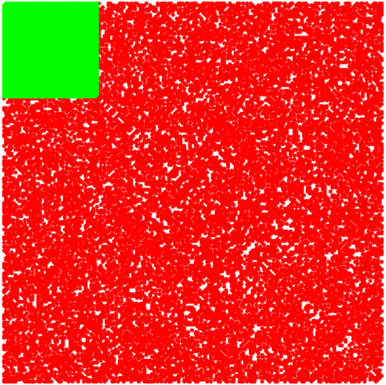


Planted clique – graph view

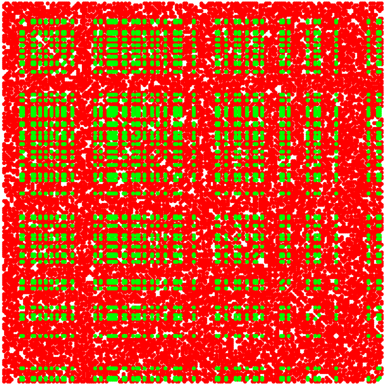
1. A clique of k vertices are chosen randomly to form a clique
2. For every other pair of nodes, add an edge w.p. $\frac{1}{2}$



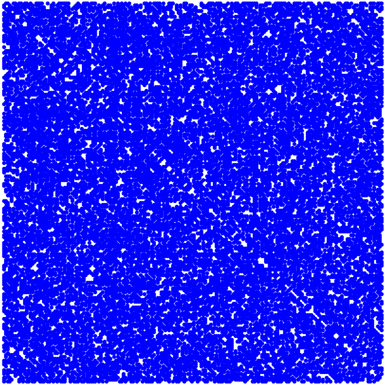
Planted clique – adjacency matrix view



Planted clique – adjacency matrix view



Planted clique – adjacency matrix view



Community detection in networks

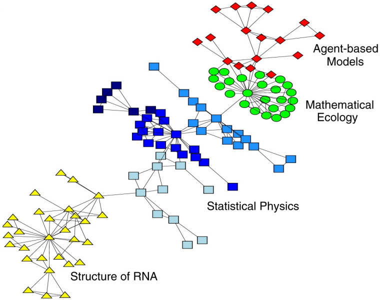
- Networks with community structures arise in many applications

Community detection in networks

- Networks with community structures arise in many applications
- Task: Discover underlying communities based on the network topology alone

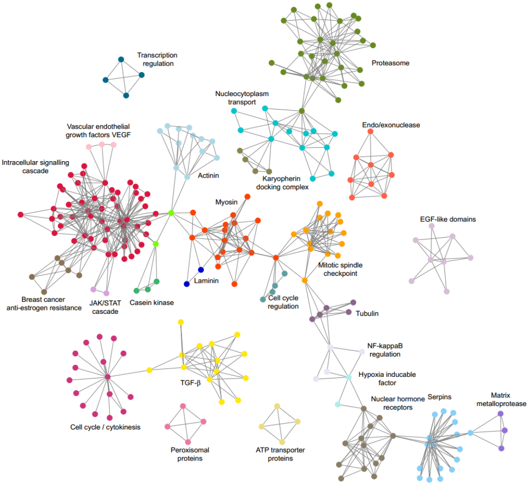
Example 1

Santa Fe Institute Collaboration network [Girvan-Newman '02]



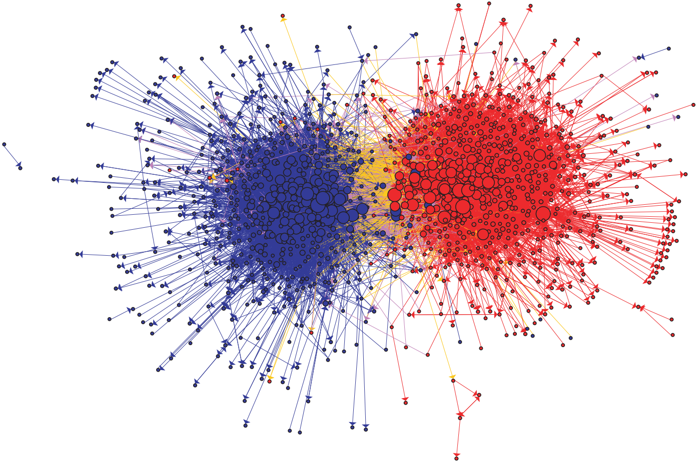
Example 2

Protein-protein interaction networks [Jonsson et al. 06']

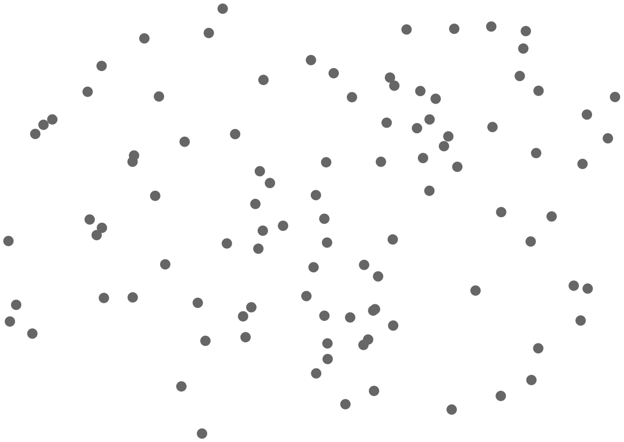


Example 3

Political blogosphere and the 2004 U.S. election [Adamic-Glance '05]

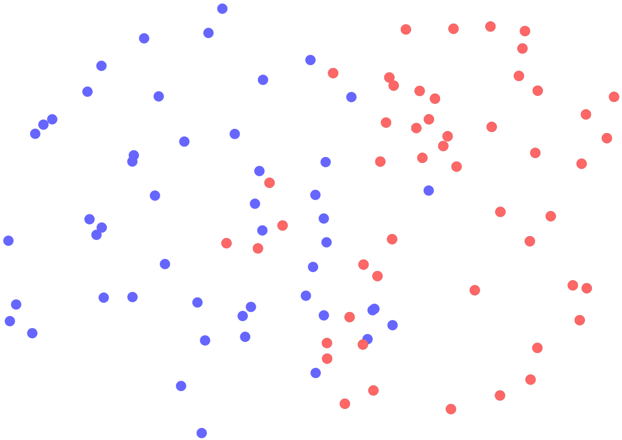


Stochastic block model – graph view



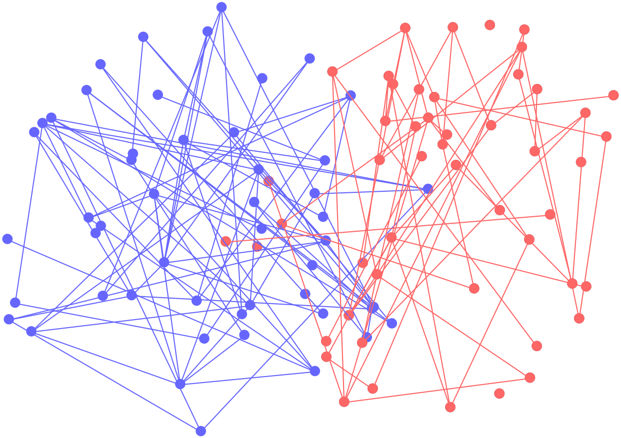
Stochastic block model – graph view

❶ n nodes are randomly partitioned into 2 equal-sized communities



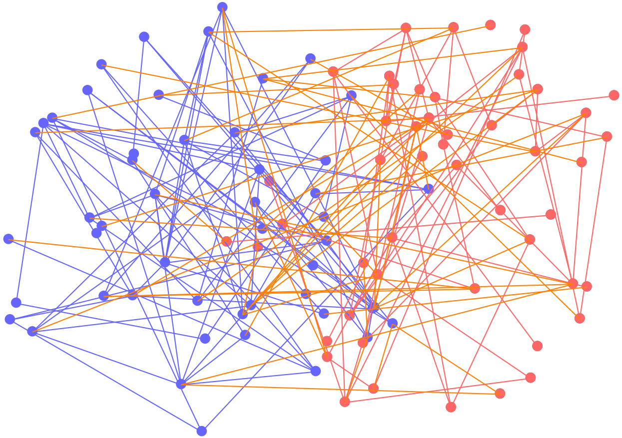
Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p



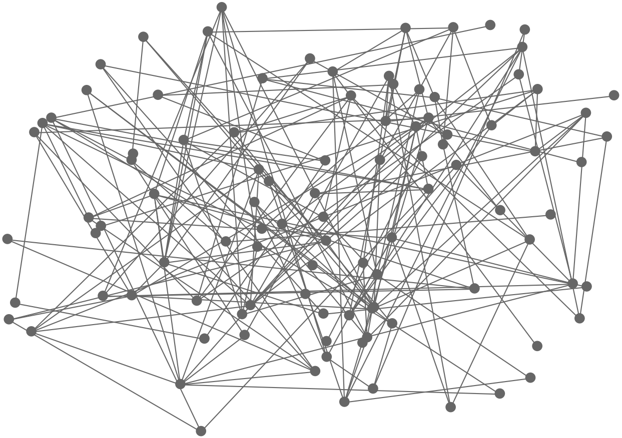
Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q

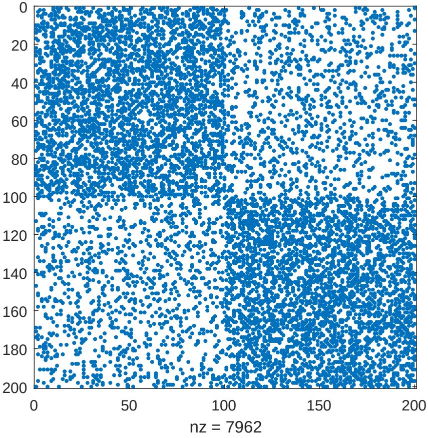


Stochastic block model – graph view

1. n nodes are randomly partitioned into 2 equal-sized communities
2. For every pair of nodes in same community, add an edge w.p. p
3. For every pair of nodes in diff. community, add an edge w.p. q



Stochastic block model – adjacency matrix view



Stochastic block model – adjacency matrix view