BA990/ECE590: Statistical inference on graphs

Spring 2020
Administrivia

- **Schedule:** Wed and Fri, 3:05PM - 4:20PM, Seminar Room G, Fuqua
- **Instructor:** Prof. Jiaming Xu [jiaming.xu868@duke.edu](mailto:jiaming.xu868@duke.edu), Fuqua W313
  - Office hours: by appointment
- **Course website:** [https://faculty.fuqua.duke.edu/~jx77/Course_graphs.html](https://faculty.fuqua.duke.edu/~jx77/Course_graphs.html)
1 Course prerequisites:

- Maturity with probability theory and linear algebra
- Familiarity with statistical theory, optimization, and algorithms
Course prerequisites:

- Maturity with probability theory and linear algebra
Course prerequisites:
- Maturity with probability theory and linear algebra
- Familiarity with statistical theory, optimization, and algorithms
Course prerequisites:

- Maturity with probability theory and linear algebra
- Familiarity with statistical theory, optimization, and algorithms

Participation (30%):

- Attending classes on time
- Raising good questions or providing good answers to questions
- Proofreading lecture notes and pointing out errors

Homework (30%): three to four problem sets

Final project (40%):

- Team of 1-2 students
- Either presenting paper(s) or a standalone research project
- Guidelines and project ideas will be announced before March 1
- Mid-term project report due March 18 (after spring break)
- Project presentation (15 mins each team) on April 10 and 15
- Final project report due April 27
Administrivia

1. Course prerequisites:
   - Maturity with probability theory and linear algebra
   - Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   - Attending classes on time :-)

3. Homework (30%): three to four problem sets

4. Final project (40%):
   - Team of 1-2 students
   - Either presenting paper(s) or a standalone research project
   - Guidelines and project ideas will be announced before March 1
   - Mid-term project report due March 18 (after spring break)
   - Project presentation (15 mins each team) on April 10 and 15
   - Final project report due April 27
Administrivia

1 Course prerequisites:
   - Maturity with probability theory and linear algebra
   - Familiarity with statistical theory, optimization, and algorithms

2 Participation (30%):
   - Attending classes on time :-)
   - Raising good questions or providing good answers to questions

3 Homework (30%): three to four problem sets

4 Final project (40%)
   - Team of 1-2 students
   - Either presenting paper(s) or a standalone research project
   - Guidelines and project ideas will be announced before March 1
   - Mid-term project report due March 18 (after spring break)
   - Project presentation (15 mins each team) on April 10 and 15
   - Final project report due April 27
Administrivia

1. Course prerequisites:
   - Maturity with probability theory and linear algebra
   - Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   - Attending classes on time :-)  
   - Raising good questions or providing good answers to questions
   - Proofreading lecture notes and pointing out errors

3. Homework (30%): three to four problem sets

4. Final project (40%)
   - Team of 1-2 students
   - Either presenting paper(s) or a standalone research project
   - Guidelines and project ideas will be announced before March 1
   - Mid-term project report due March 18 (after spring break)
   - Project presentation (15 mins each team) on April 10 and 15
   - Final project report due April 27
Course prerequisites:
  ▶ Maturity with probability theory and linear algebra
  ▶ Familiarity with statistical theory, optimization, and algorithms

Participation (30%):
  ▶ Attending classes on time :-)  
  ▶ Raising good questions or providing good answers to questions  
  ▶ Proofreading lecture notes and pointing out errors

Homework (30%): three to four problem sets
Administrivia

1. Course prerequisites:
   - Maturity with probability theory and linear algebra
   - Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   - Attending classes on time :-)
   - Raising good questions or providing good answers to questions
   - Proofreading lecture notes and pointing out errors

3. Homework (30%): three to four problem sets

4. Final project (40%)
Course prerequisites:
  ▶ Maturity with probability theory and linear algebra
  ▶ Familiarity with statistical theory, optimization, and algorithms

Participation (30%):
  ▶ Attending classes on time :-) 
  ▶ Raising good questions or providing good answers to questions
  ▶ Proofreading lecture notes and pointing out errors

Homework (30%): three to four problem sets

Final project (40%)
  ▶ Team of 1-2 students
Administrivia

1. Course prerequisites:
   ▶ Maturity with probability theory and linear algebra
   ▶ Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   ▶ Attending classes on time :-)  
   ▶ Raising good questions or providing good answers to questions  
   ▶ Proofreading lecture notes and pointing out errors

3. Homework (30%): three to four problem sets

4. Final project (40%)  
   ▶ Team of 1-2 students  
   ▶ Either presenting paper(s) or a standalone research project
1 Course prerequisites:
   ▶ Maturity with probability theory and linear algebra
   ▶ Familiarity with statistical theory, optimization, and algorithms

2 Participation (30%):
   ▶ Attending classes on time :-)
   ▶ Raising good questions or providing good answers to questions
   ▶ Proofreading lecture notes and pointing out errors

3 Homework (30%): three to four problem sets

4 Final project (40%)
   ▶ Team of 1-2 students
   ▶ Either presenting paper(s) or a standalone research project
   ▶ Guidelines and project ideas will be announced before March 1
Course prerequisites:
- Maturity with probability theory and linear algebra
- Familiarity with statistical theory, optimization, and algorithms

Participation (30%):
- Attending classes on time :-)
- Raising good questions or providing good answers to questions
- Proofreading lecture notes and pointing out errors

Homework (30%): three to four problem sets

Final project (40%)
- Team of 1-2 students
- Either presenting paper(s) or a standalone research project
- Guidelines and project ideas will be announced before March 1
- Mid-term project report due March 18 (after spring break)
Administrivia

1. Course prerequisites:
   ▶ Maturity with probability theory and linear algebra
   ▶ Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   ▶ Attending classes on time :-)
   ▶ Raising good questions or providing good answers to questions
   ▶ Proofreading lecture notes and pointing out errors

3. Homework (30%): three to four problem sets

4. Final project (40%)
   ▶ Team of 1-2 students
   ▶ Either presenting paper(s) or a standalone research project
   ▶ Guidelines and project ideas will be announced before March 1
   ▶ Mid-term project report due March 18 (after spring break)
   ▶ Project presentation (15 mins each team) on April 10 and 15
Administrivia

1. Course prerequisites:
   - Maturity with probability theory and linear algebra
   - Familiarity with statistical theory, optimization, and algorithms

2. Participation (30%):
   - Attending classes on time :-)  
   - Raising good questions or providing good answers to questions
   - Proofreading lecture notes and pointing out errors

3. Homework (30%): three to four problem sets

4. Final project (40%)
   - Team of 1-2 students
   - Either presenting paper(s) or a standalone research project
   - Guidelines and project ideas will be announced before March 1
   - Mid-term project report due March 18 (after spring break)
   - Project presentation (15 mins each team) on April 10 and 15
   - Final project report due April 27
Materials: Lecture notes and additional reading materials will be posted on the course website.
Administ rivia

1. Materials: Lecture notes and additional reading materials will be posted on the course website.

2. March 4 and 6 classes will be in RAND classroom just below seminar G.
Materials: Lecture notes and additional reading materials will be posted on the course website

March 4 and 6 classes will be in RAND classroom just below seminar G

Classes will be video recorded using Duke’s Panopto system
Statistical problems

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

\[ \theta \in \Theta \mapsto X \mapsto \hat{\theta} \]

- Understanding the fundamental limits:
  - Characterize statistical (information-theoretic) limit: What is possible/impossible?
  - 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 \( \to \infty \))
Statistical problems

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

\[ \theta \in \Theta \mapsto X \mapsto \hat{\theta} \]

- parameter
- data
- estimate

• Understanding the fundamental limits:
Statistical problems

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

\[ \theta \in \Theta \mapsto X \mapsto \hat{\theta} \]

parameter \quad data \quad 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)

\[ \theta \in \Theta \quad \mapsto \quad X \quad \mapsto \quad \hat{\theta} \]

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

\[ \theta \in \Theta \mapsto X \mapsto \hat{\theta} \]

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

\[ \theta \in \Theta \mapsto X \mapsto \hat{\theta} \]

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

  \[ \theta \in \Theta \quad \mapsto \quad X \quad \mapsto \quad \hat{\theta} \]

  - parameter \quad data \quad estimate

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

\[ \theta \in \Theta \rightarrow X \rightarrow \hat{\theta} \]

• Understanding the fundamental limits:
  1. Characterize statistical (information-theoretic) limit: What is possible/impossible?
  2. 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 $\rightarrow \infty$)
Basic definitions of graphs

A graph \( G = (V, E) \) consists of

- A **vertex set** \( V = [n] \equiv \{1, \ldots, n\} \) for some positive integer \( n \).
- An **edge set** \( E \subset \binom{V}{2} \). Each element of \( E \) is an edge \( e = (i, j) \) (unordered pair). We say \( i \) and \( j \) are connected and write \( i \sim j \) if \( (i, j) \in E \).

We mostly focus on graphs that are **undirected** and **simple**.

**Adjacency matrix representation:** \( A = (A_{ij})_{i,j \in [n]} \) is an \( n \times n \) symmetric binary matrix with zero diagonal and

\[
A_{ij} = 1\{i \sim j\} = \begin{cases} 
1 & (i, j) \in E \\
0 & \text{o.w.}
\end{cases}
\]

\[
A = \begin{bmatrix}
0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0
\end{bmatrix}
\]
Basic definitions of graphs

- **The neighborhood** of a given vertex $v \in V$:
  \[
  N(v) = \{u \in V : u \sim v\}
  \]
- **The degree** of $v$:
  \[
  d_v = |N(v)|
  \]
- **Induced subgraph**: For any $S \subset V$, the subgraph induced by $S$ is $G[S] = (S, E_S)$, where
  \[
  E_S \triangleq \{(u, v) \in E : u, v \in S\}
  \]
- **A clique** is a **complete** subgraph. A graph is complete iff all pairs of vertices in the graph are connected.
Planted clique – graph view

A set $S$ of $k$ vertices is chosen to form a clique. For every other pair of vertices, add an edge w.p. 1.
A set $S$ of $k$ vertices is chosen to form a clique.
A set $S$ of $k$ vertices is chosen to form a clique
Planted clique – graph view

1. A set $S$ of $k$ vertices is chosen to form a clique
2. For every other pair of vertices, add an edge w.p. $\frac{1}{2}$
Planted clique – graph view

1. A set $S$ of $k$ vertices is chosen to form a clique
2. For every other pair of vertices, add an edge w.p. $\frac{1}{2}$
Planted clique – adjacency matrix view
Planted clique – adjacency matrix view
Planted clique – adjacency matrix view
Planted clique recovery

Planted clique $S \longrightarrow$ graph $G \longrightarrow$ Estimated clique $\hat{S}$

- **Minimax framework:** Find an estimator $\hat{S} = \hat{S}(G)$ that performs well in worst-case
  $$\min_{S \in \binom{[n]}{k}} \mathbb{P} \left[ \hat{S}(G) = S \right] \approx 1$$

- **Bayesian framework:** Find an estimator $\hat{S} = \hat{S}(G)$ that performs well on average
  $$\mathbb{E}_{S \sim \text{Unif}(\binom{[n]}{k})} \mathbb{P} \left[ \hat{S}(G) = S \right] \approx 1$$

- The two formulations are equivalent by the permutation invariance of the model:
  $$\sup_{\hat{S}} \min_{S \in \binom{[n]}{k}} \mathbb{P} \left[ \hat{S}(G) = S \right] = \sup_{\hat{S}} \mathbb{E}_{S \sim \text{Unif}(\binom{[n]}{k})} \mathbb{P} \left[ \hat{S}(G) = S \right].$$
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]

Figure 1: Community structure of political blogs (expanded set), shown using utilizing the GUESS visualization tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Because of bloggers' ability to identify and frame breaking news, many mainstream media sources keep a close eye on the best known political blogs. A number of mainstream news sources have started to discuss and even to host blogs. In an online survey asking editors, reporters, columnists and publishers to each list the “top 3” blogs they read, Drezner and Farrell [4] identified a short list of dominant “A-list” blogs. Just 10 of the most popular blogs accounted for over half the blogs on the journalists’ lists. They also found that, besides capturing most of the attention of the mainstream media, the most popular political blogs also get a disproportionate number of links from other blogs. Shirky [12] observed the same effect for blogs in general and Hindman et al. [7] found it to hold for political websites focusing on various issues.

While these previous studies focused on the inequality of citation links for political blogs overall, there has been comparatively little study of subcommunities of political blogs. In the context of political websites, Hindman et al. [7] noted that, for example, those dealing with the issue of abortion, gun control, and the death penalties, contain subcommunities of opposing views. In the case of the pro-choice and pro-life web communities, an earlier study [1] found pro-life websites to be more densely linked than pro-choice ones. In a study of a sample of the blogosphere, Herring et al.[6] discovered densely interlinked (non-political) blog communities focusing on the topics of Catholicism and homeschooling, as well as a core network of A-list blogs, some of them political.

Recently, Butts and Cross [3] studied the response in the structure of networks of political blogs to polling data and election campaign events. In another political blog study, Welsch [15] gathered a single-day snapshot of the network neighborhoods of Atrios, a popular liberal blog, and Instapundit, a popular conservative blog. He found the Instapundit neighborhood to include many more blogs than the Atrios one, and observed no overlap in the URLs cited between the two neighborhoods. The lack of overlap in liberal and conservative interests has previously been observed in purchases of political books on Amazon.com [8]. This brings about the question of whether we are witnessing a cyberbalkanization [11, 13] of the Internet, where the proliferation of specialized online news sources allows people with different political leanings to be exposed only to information in agreement with their previously held views. Yale law professor Jack Balkin provides a counter-argument by pointing out that such segregation is unlikely in the blogosphere because bloggers systematically comment on each other, even if only to voice disagreement.

In this paper we address both hypotheses by examining in a systematic way the linking patterns and discussion topics of political bloggers. In doing so, we not only measure the degree of interaction between liberal and conservative blogs, but also uncover differences in the structure of the two communities. Our data set includes the posts of 40 A-list blogs over the period of two months preceding the U.S. Presidential Election of 2004. We also study a large network of over 1,000 political blogs based on a single day snapshot that includes blogrolls (the list of links to other blogs frequently found in sidebars), and so presents a more static picture of a broader blogosphere.

From both samples we find that liberal and conservative blogs did indeed have different lists of favorite news sources.
Stochastic block model – graph view

- $n$ nodes are partitioned into $2$ equal-sized communities.
- For every pair of nodes in the same community, add an edge with probability $p$.
- For every pair of nodes in different communities, add an edge with probability $q$. 
1 $n$ nodes are partitioned into 2 equal-sized communities
Stochastic block model – graph view

1. $n$ nodes are 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 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 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

nz = 7962
Stochastic block model – estimation

Planted community $\sigma \in \{\pm 1\}^n \rightarrow$ graph $G \rightarrow$ Estimated community $\hat{\sigma}$

$$\mathbb{P}[(i, j) \in E] = \begin{cases} p & \sigma_i = \sigma_j \\ q & \sigma_i \neq \sigma_j \end{cases},$$

- If $p$ and $q$ are known, estimate the planted community $\sigma$
- If $p$ and $q$ are unknown, jointly estimate $p, q$ and $\sigma$

When $p = q$, the SBM reduces to the Erdős-Rényi random graph $G(n, p)$
Asymptotic Notation

For two sequences of numbers $a_n$ and $b_n$, where $b_n > 0$ for all sufficiently large $n$. Then

- $a_n = O(b_n)$ or $a_n \lesssim b_n$, if there exist constants $C$ and $n_0$ such that $|a_n| \leq Cb_n$ for all $n \geq n_0$.
- $a_n = \Omega(b_n)$ or $a_n \gtrsim b_n$, if there exist constants $c > 0$ and $n_0$ such that $a_n \geq cb_n$ for all $n \geq n_0$.
- $a_n = \Theta(b_n)$ or $a_n \asymp b_n$, if $a_n = O(b_n)$ and $a_n = \Omega(b_n)$.
- $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$.
- $a_n = o(b_n)$, if $a_n/b_n \to 0$ as $n \to \infty$.
- $a_n = \omega(b_n)$ if $a_n/b_n \to \infty$ as $n \to \infty$.
- $a_n \ll b_n$ if $a_n \geq 0$ and $a_n = o(b_n)$.

Also, we say that a sequence of events $\mathcal{E}_n$ holds with high probability (w.h.p.), if $\mathbb{P}\{\mathcal{E}_n\} \to 1$ as $n \to \infty$. 