Network Science Analytics

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

February 1, 2021
Introductions

Networks - A birds-eye view

Class description and contents
Who we are, where to find me, lecture times

- **Gonzalo Mateos**
- Associate Professor, Dept. of Electrical and Computer Engineering
- CSB 726, gmateosb@ece.rochester.edu
- [http://www.ece.rochester.edu/~gmateosb](http://www.ece.rochester.edu/~gmateosb)
- **Where?** We meet online via Zoom
  - Meeting ID: 771 885 0098, passcode sent via email
- **When?** Mondays and Wednesdays 3:25 pm to 4:40 pm
- My office hours, **Tuesdays at 2 pm** via Zoom (771 885 0098)
  - Anytime, as long as you have something interesting to tell me
- **Class website**
  - [http://www.ece.rochester.edu/~gmateosb/ECE442.html](http://www.ece.rochester.edu/~gmateosb/ECE442.html)
Teaching assistants

Two great TAs to help you with your homework and project

Narges Mohammadi
- Email: nmohamm4@ur.rochester.edu
- Her office hours, Fridays at 1 pm
- Zoom: 381 188 3230

Shadi Sartipi
- Email: ssartipi@ur.rochester.edu
- Her office hours, Thursdays at 11 am
- Zoom: 527 274 2525
Prerequisites

(I) Graph theory and statistical inference
- Graphs are mathematical abstractions of networks
- Statistical inference useful to “learn” from network data
- Basic knowledge expected. Will review in first four lectures

(II) Probability theory and linear algebra
- Random variables, distributions, expectations, Markov processes
- Vector/matrix notation, systems of linear equations, eigenvalues

(III) Programming
- Will use e.g., Matlab for homework and your project
- You can use the language/network analysis package your prefer
- Plenty of libraries in Python and R
Homework, project and grading

(I) Homework sets (3 in 14 weeks) worth 20%
   ▶ Mix of analytical problems and programming assignments
   ▶ Collaboration accepted, welcomed, and encouraged

(II) Research project on a topic of your choice, worth 80%
   ▶ Important and demanding part of this class. Three deliverables:
     1) Proposal by the end of week 6, worth 15%
     2) Progress report by the end of week 10, worth 15%
     3) Final report and in-class presentation, worth 50%

   ▶ This is a special topics, research-oriented graduate level class
     ⇒ Focus should be on thinking, reading, asking, implementing
     ⇒ Goal is for everyone to earn an A
We will use lecture slides to cover the material.

Research papers, tutorials also posted in the class website.

Basic book I will follow is: Eric D. Kolaczyk, “Statistical Analysis of Network Data: Methods and Models,” Springer

Available online from http://www.library.rochester.edu/


Be nice

- I work hard for this course, expect you to do the same

✓ Come to class, be on time, pay attention, ask
✓ Check out the additional suggested readings
✓ Play with network analysis software
✓ Search for datasets
✓ Do all of your homework
× Do not hand in as yours the solution of others

- Let me know of your interests. I can adjust topics accordingly

- Come and learn. Useful down the road. More on impact next
Stop the spread

WEAR A MASK
Cover your nose and mouth with a face covering. Refrain from touching your face.

DR. CHAT BOT
Complete the Dr. Chat Bot screening online every day at uofr.us/chatbot.

WASH HANDS
Wash your hands often. Follow the CDC guidelines posted in restrooms.

STAY APART
Stay six feet apart. Adhere to physical distancing recommendations.

Learn more at rochester.edu/coronavirus-update
Introductions

Networks - A birds-eye view

Class description and contents
As per the dictionary: *A collection of inter-connected things*

- Ok. There are multiple things, they are connected. Two extremes

1) A real (complex) system of inter-connected components
2) A graph representing the system

- Understand complex systems ⇔ Understand networks behind them
Historical background

- Network-based analysis in the sciences has a long history
- Mathematical foundations of graph theory (L. Euler, 1735)
  - The seven bridges of Königsberg
- Laws of electrical circuitry (G. Kirchoff, 1845)
- Molecular structure in chemistry (A. Cayley, 1874)
- Network representation of social interactions (J. Moreno, 1930)
- Power grids (1910), telecommunications and the Internet (1960)
Why networks? Why now?

- Understand complex systems ⇔ Understand networks behind them

- Relatively small field of study up until ∼ the mid-90s

- Epidemic-like explosion of interest recently. A few reasons:
  - Systems-level perspective in science, away from reductionism
  - Ubiquitous high-throughput data collection, computational power
  - Globalization, the Internet, connectedness of modern societies
Network Science

- Study of complex systems through their network representations

- Universal language for describing complex systems and data
  - Striking similarities in networks across science, nature, technology

- Shared vocabulary across fields, cross-fertilization
  - From biology to physics, economics to statistics, CS to sociology

- Impact: social networking, drug design, smart infrastructure, . . .
Economic impact

- **Google**
  Market cap: $1.24 trillion

- **Facebook**
  Market cap: $736 billion

- **Cisco**
  Market cap: $188 billion

- **Apple**
  Market cap: $2.22 billion
Healthcare impact

▶ Prediction of epidemics, e.g. the 2009 H1N1 pandemic

Real

Predicted

▶ Human Connectome Project to map-out brain circuitry
Homeland security impact

- Social network analysis key to capturing S. Hussein
What are the goals of Network Science?

- **Reveal** patterns and statistical properties of network data
- **Understand** the underpinnings of network behavior and structure
- **Engineer** more resource-efficient, robust, socially-intelligent networks

**Characteristics:** interdisciplinary, empirical, quantitative, computational

**Empirical** study of graph-valued data to find patterns and principles
- Collection, measurement, summarization, visualization?

**Mathematical models.** Graph theory meets statistical inference
- Understand, predict, discern nominal vs anomalous behavior?

**Algorithms** for graph analytics
- Computational challenges, scalability, tractability vs optimality?
Examples of networks

- Network analysis spans the sciences, humanities and arts

- Let’s see a few examples from four general areas
  - Technological
  - Biological
  - Social
  - Informational

- Standard taxonomy, by no means the only one
  ⇒ “Soft” classification, networks may fall in multiple categories
Technological networks

- **Ex:** communication, transportation, energy, sensor networks

Q1: What does the Internet look like today? How big is it?
Q2: How will the traffic from New York to Chicago look tomorrow?
Q3: How can we unveil anomalous traffic patterns?
Biological networks

- **Ex:** neurons, gene regulatory, protein interaction, metabolic paths, predator-prey, ecological networks

**Q1:** Are certain gene interactions more common than expected?

**Q2:** Which parts of the brain “communicate” during a given task?

**Q3:** Can we predict biological function of proteins from interactions?
Social networks

- **Ex:** friendship, corporate, email exchange, international relations, financial networks

- **Q1:** What are the mechanisms underpinning friendship formation?
- **Q2:** Which actors are central to the network and which peripheral?
- **Q3:** Can we identify overlapping communities?
Informational networks

Ex: WWW, Twitter, co-citation between academic journals, blogosphere, paper co-authorship, peer-to-peer networks

Q1: How does the size and structure of the WWW change in time?
Q2: How can we use network analysis for authorship attribution?
Q3: Can we track information cascades in online social media?
Class contents

Introductions

Networks - A birds-eye view

Class description and contents
What is this class about?

- **Our focus:** Statistical analysis of network data
- Measurements of or from a system conceptualized as a network

- **Unique challenges**
  - Relational aspect of the data
  - Complex statistical dependencies
  - High-dimensional and often massive in quantity

- Will examine how these challenges arise in relation to
  - Visualization
  - Summarization and description
  - Sampling and inference
  - Modeling
Q: How does one go about ‘mapping’ the ‘landscape’ of ‘Science’?

Statistical challenges

- Defining the population of interest
- Representativeness of our data
- Appropriate notions of units (vertices and edges)
- How to visualize it effectively?
Q: How to describe/summarize the complex interactions during a seizure?

Statistical challenges
- Criterion for defining ‘brain networks’
- Choice of network summary statistics
- Assessing significance of changes/differences
Q: Can we monitor characteristics of massive social media networks?

Statistical challenges

- Computer protocols correspond to what sampling designs?
- What sort of biases are inherent to the sampling?
- Can we compensate for those biases?
Q: Can we leverage protein-protein interactions to infer function?

Statistical challenges
- To what extent do interacting proteins share common function?
- How do we incorporate a network as an explanatory variable?
- Can we account for uncertainty in the training data and/or network?
Four thematic blocks in this class

(I) Graph theory, probability and statistical inference review (∼ 4 lectures)
   ▶ Vertices and edges, degrees, subgraphs, families of graphs, connectivity, . . .
   ▶ Algebraic graph theory, adjacency and Laplacian matrices, spectrum, . . .
   ▶ Estimation, prediction and hypothesis testing. Case studies

   ▶ Will follow a statistical taxonomy: descriptive an inferential techniques
      ⇒ Issues on data collection, data management and computing

(II) Descriptive analysis and properties of large networks (∼ 7 lectures)
(III) Sampling, modeling and inference of networks (∼ 9 lectures)
(IV) Processes evolving over network graphs (∼ 8 lectures)
Descriptive analysis and properties of networks

- The WWW and other large directed graphs exhibit a “bowtie” structure

- Power-law degree distributions are ubiquitous in real-world networks

- Of interest: network graph construction and visualization, centrality measures, community detection, network sampling, small-world

- Applications: Google’s PageRank, marketing, epilepsy, transportation
Watts-Strogatz model captures small-world structure in real graphs

- Highly structured locally (like social groups); and
- “Small” globally (like purely random graphs)

Of interest: random graph models, network topology inference, growth models for evolving networks, preferential attachment

Applications: detecting motifs, inferring gene-regulatory interactions, mapping the Internet, predicting popularity in Twitter
Processes evolving over network graphs

- Tracking of end-to-end delay in the Internet
  - Only 30 out of 62 paths sampled, routing induces spatial correlations
  - “Ground-truth” delays compared to real-time estimates

- Of interest: Markov random fields, kernel regression on graphs, epidemic modeling, network flow models, traffic matrix estimation

- Applications: computer network health monitoring, electric load data cleansing, information cascades in social media, viral marketing