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/

January 11, 2022

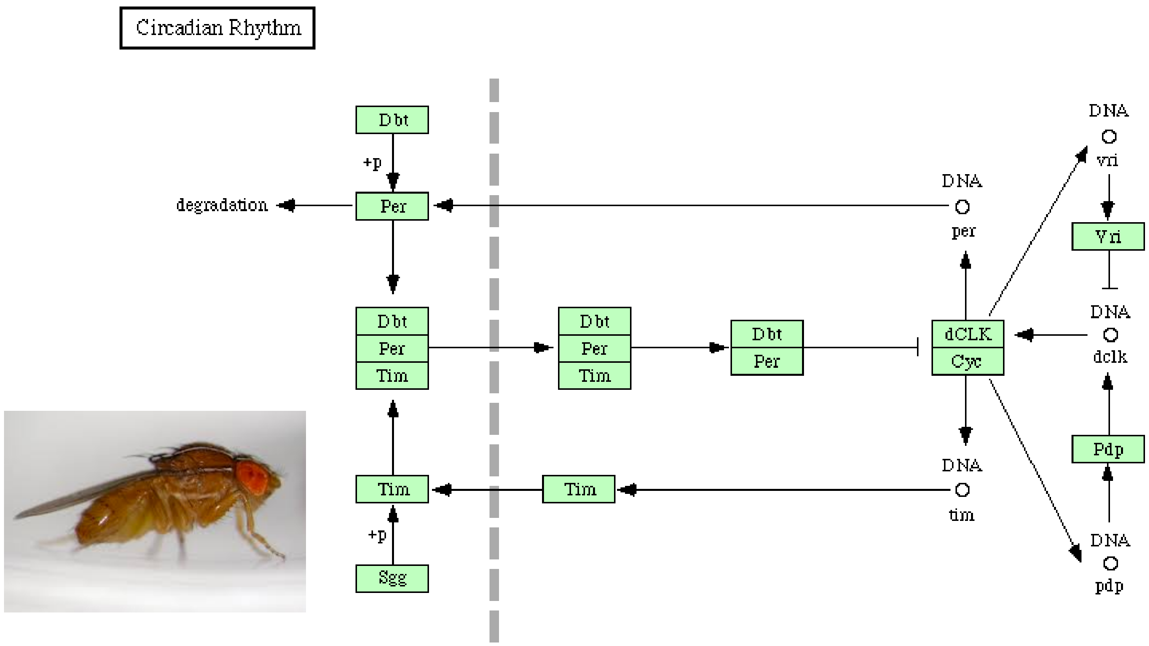
Networks

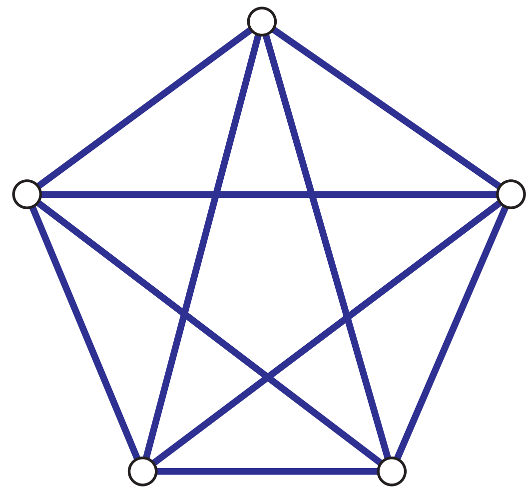
Introductions

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



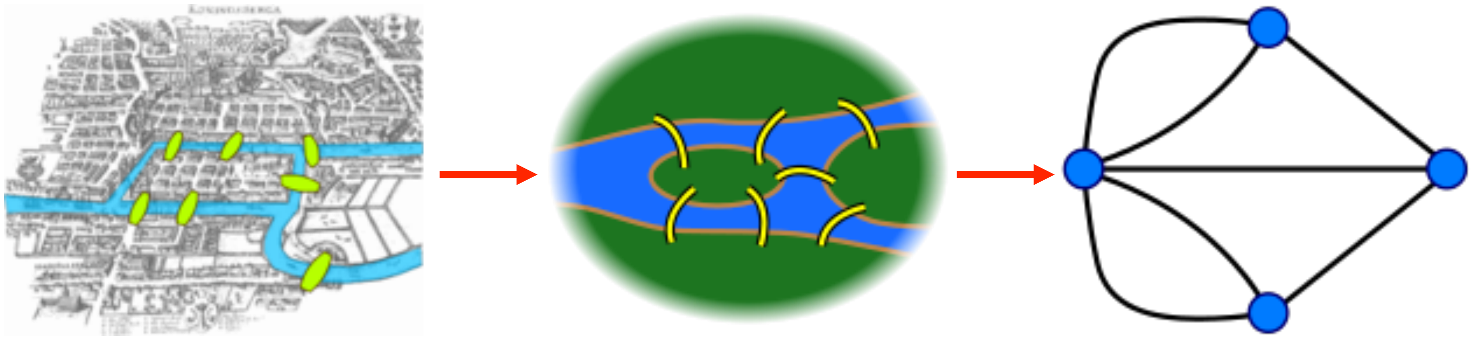


1) A real (complex) system of inter-connected components
2) A graph representing the system

- Understand complex systems $\Leftrightarrow$ Understand networks behind them

Historical background

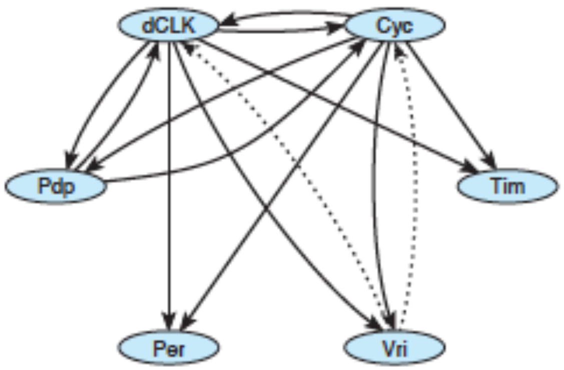
- Network-based analysis in the sciences has a long history
- Mathematical foundations of graph theory (L. Euler, 1735)

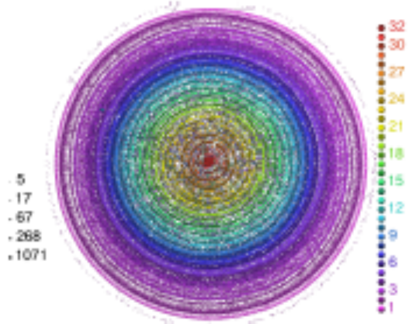


  - The seven bridges of Königsberg
- Laws of electrical circuitry (G. Kirchoff, 1845)
- Molecular structure in chemistry (A. Cayley, 1874)
- Network representation of social interactions (J. Moreno, 1930)
- Power grids (1910), telecommunications and the Internet (1960)
- Google (1997), Facebook (2004), Twitter (2006), ...

Why networks? Why now?

- Understand complex systems ⇔ Understand networks behind them

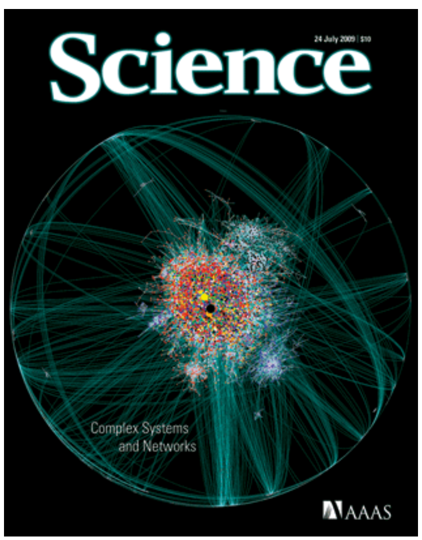


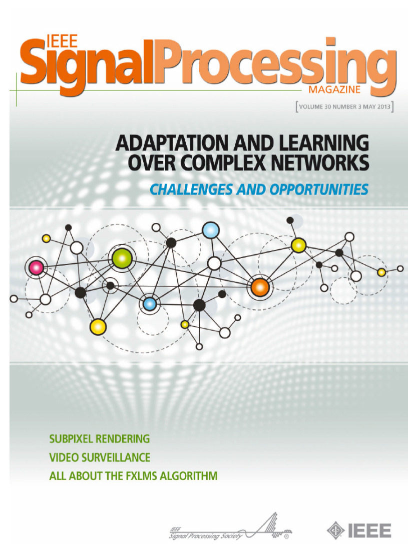


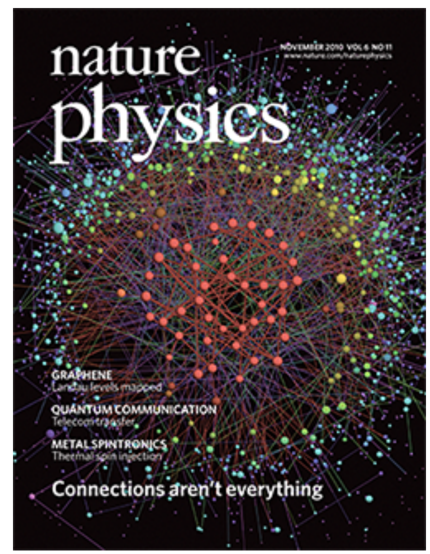
- Relatively small field of study up until ∼ the mid-90s
- Epidemic-like explosion of interest recently. A few reasons:
  - Systems-level perspective in science, away from reductionism
  - Ubiquitous high-throughput data collection, computational power
  - Globalization, the Internet, connectedness of modern societies

Network Science

- Study of complex systems through their network representations
  Ex: economy, metabolism, brain, society, Web, . . .
- 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

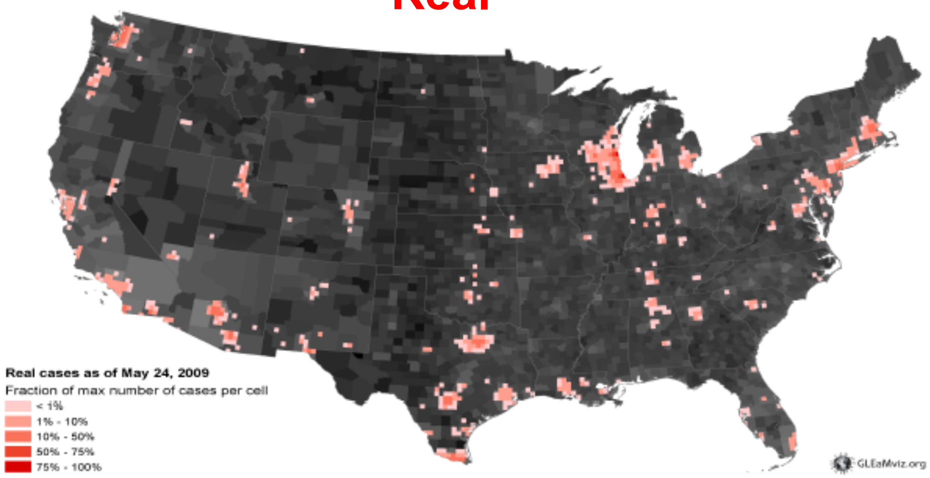
- Facebook
  Market cap:
  $736 billion

- Cisco
  Market cap:
  $188 billion

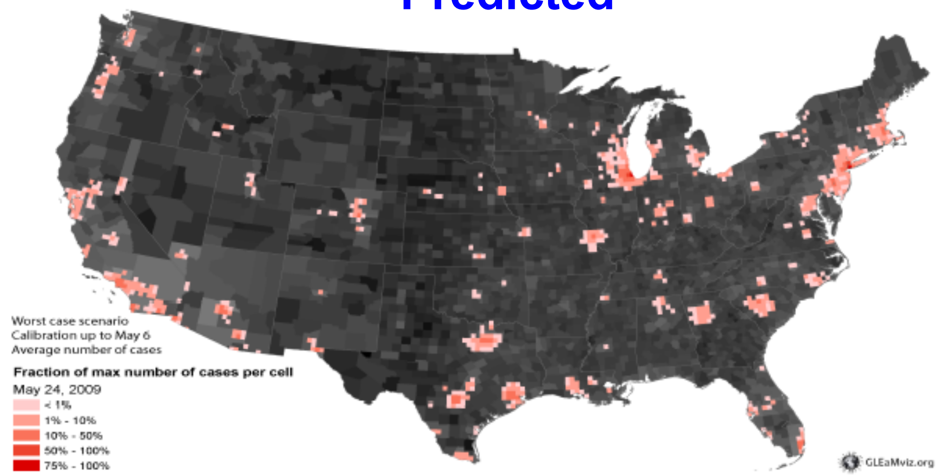
- Apple
  Market cap:
  $2.22 billion

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

Real

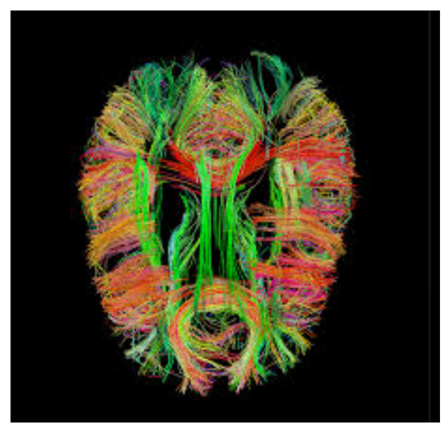


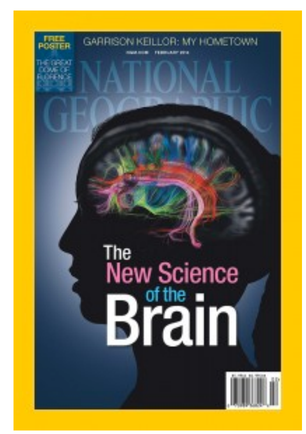
Predicted



- Human Connectome Project to map-out brain circuitry

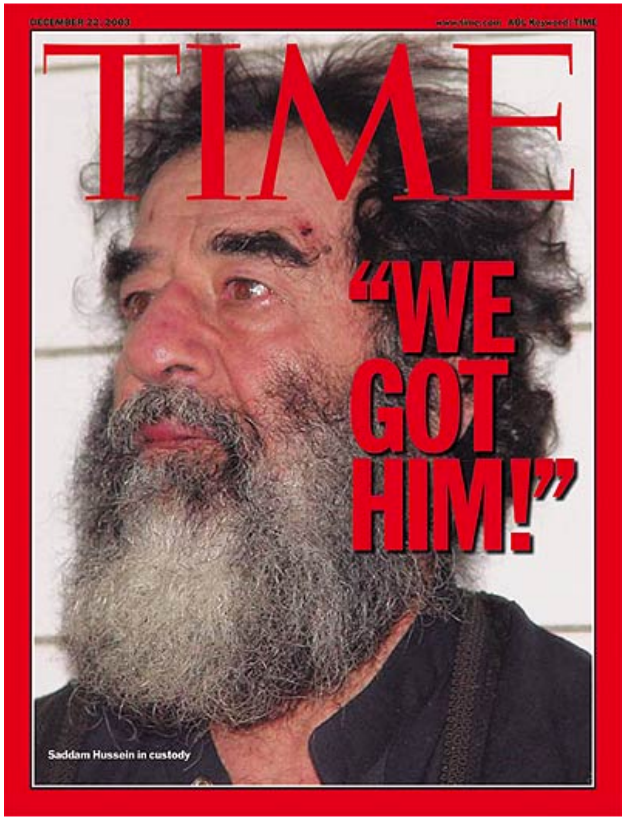
HUMAN Connectome PROJECT

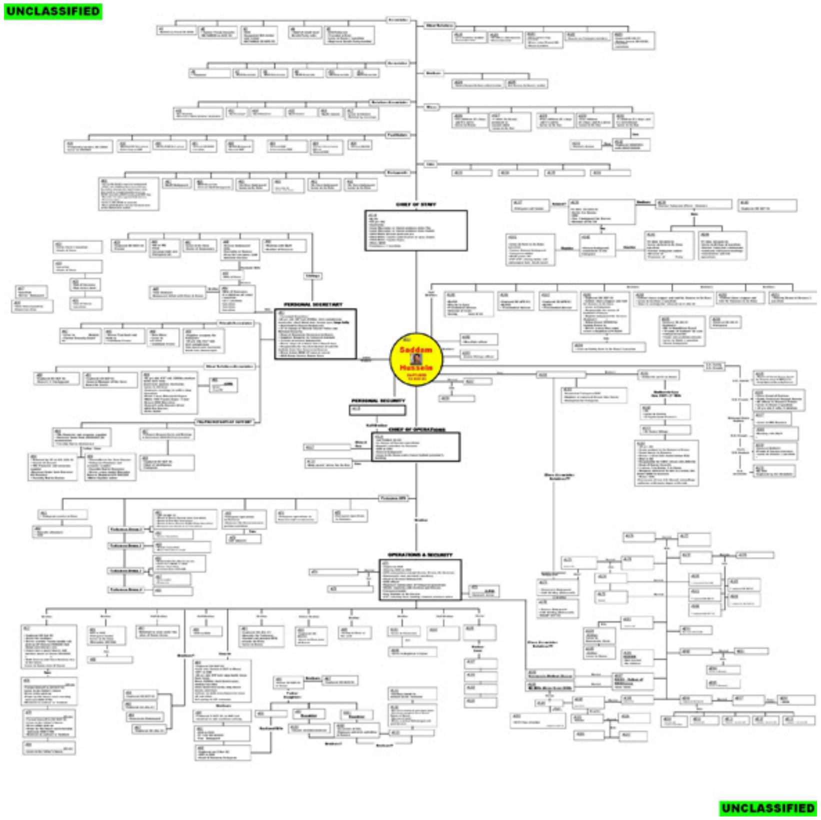




Homeland security impact

- Social network analysis key to capturing S. Hussein





Desiderata and Network Science characteristics

- 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?

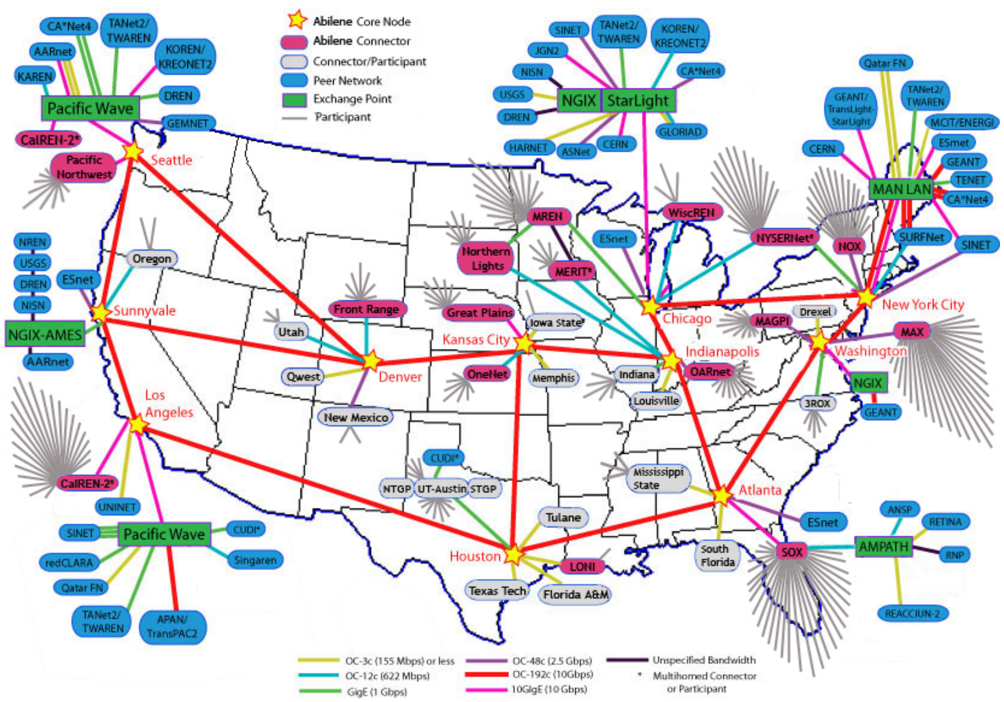
Examples of networks

- 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

  $\Rightarrow$ "Soft" classification, networks may fall in multiple categories

Technological networks

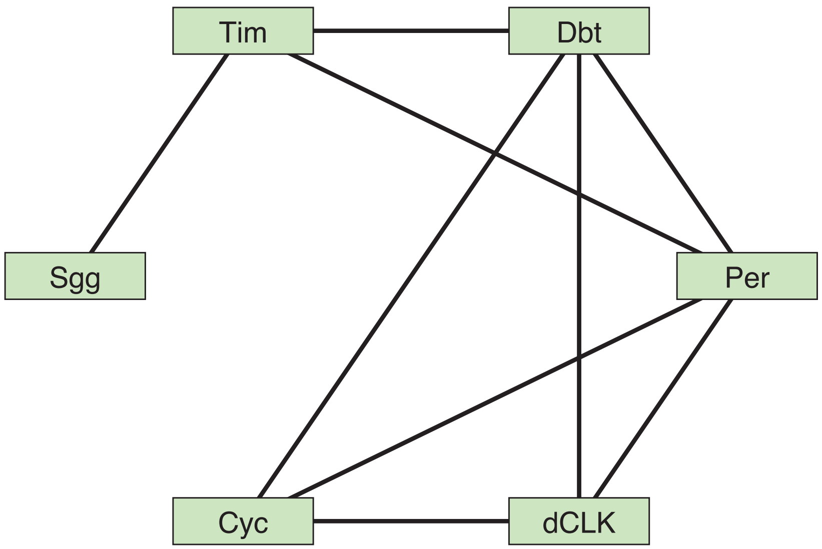
- Ex: communication, transportation, energy, sensor networks

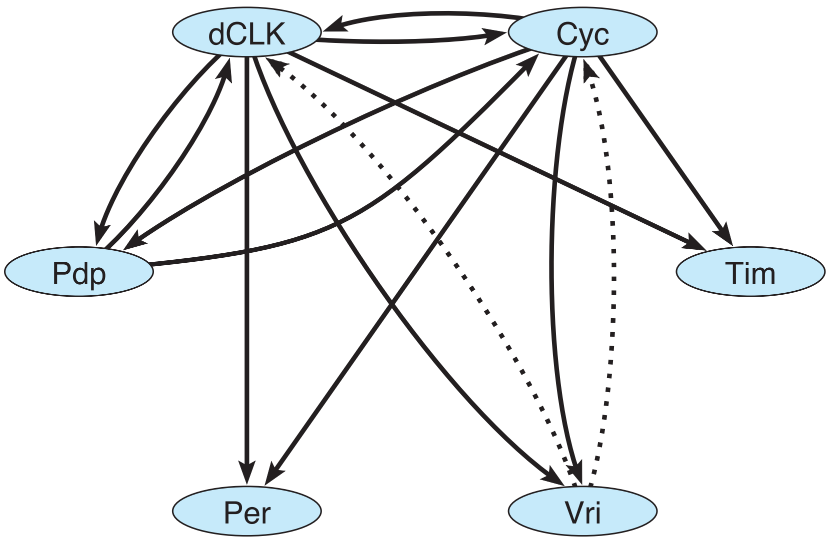


- Q1: What does the Internet look like today? How big is it?
- Q2: How will the traffic from New York to Chicago look tomorrow?
- Q3: How can we unveil anomalous traffic patterns?

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?

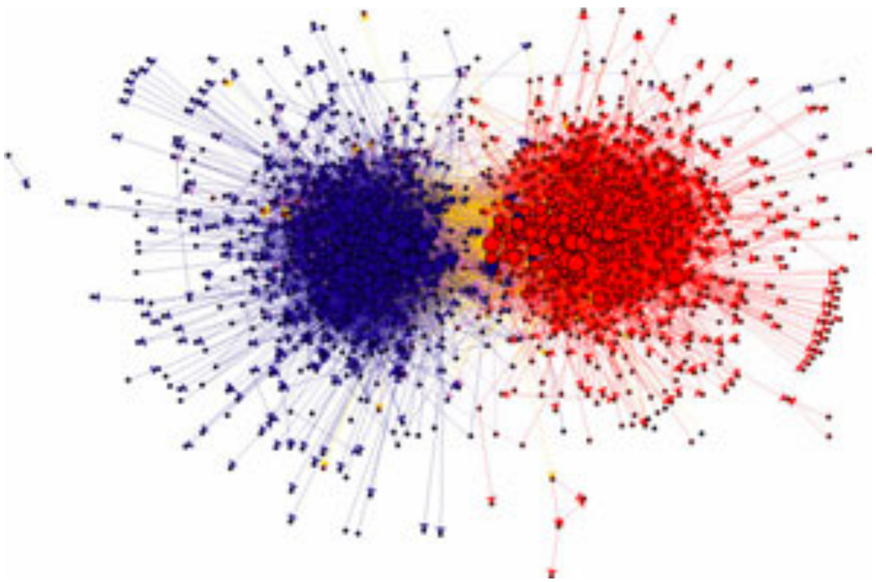
Social networks

- 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

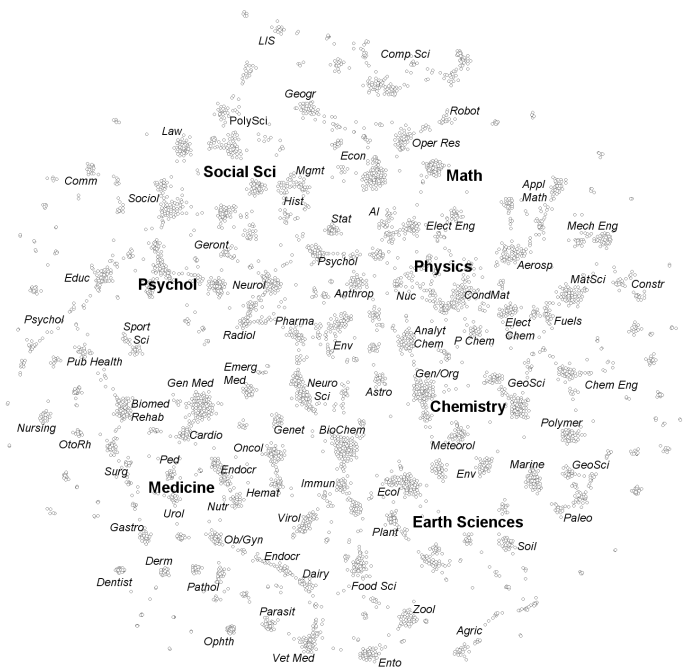
Class description and contents

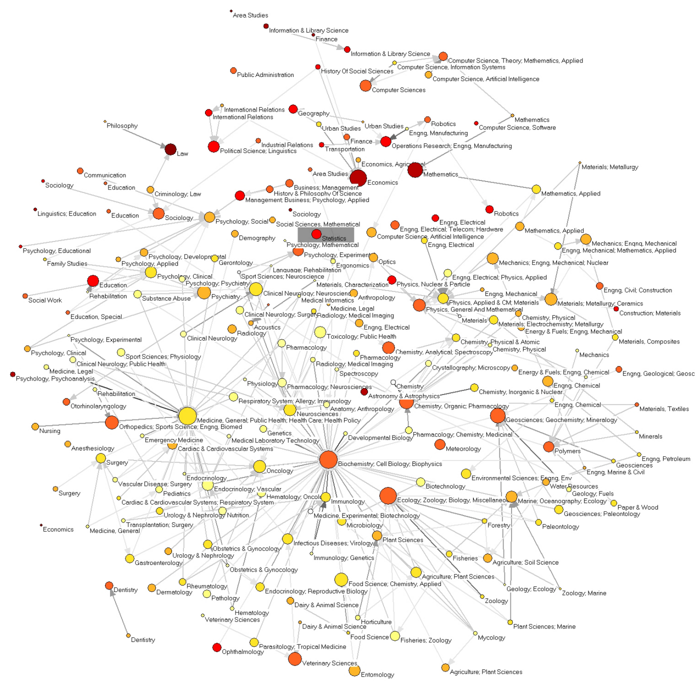
What is this class about?

- 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

Mapping Science

- Q: How does one go about ‘mapping’ the ‘landscape’ of ‘Science’?

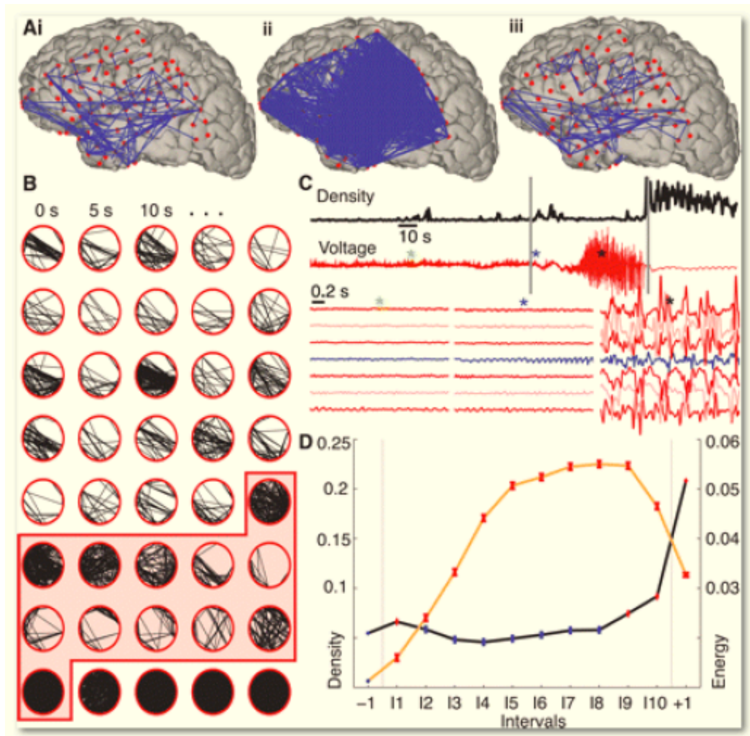




- Statistical challenges
  - Defining the population of interest
  - Representativeness of our data
  - Appropriate notions of units (vertices and edges)
  - How to visualize it effectively?

UNIVERSITY of ROCHESTER

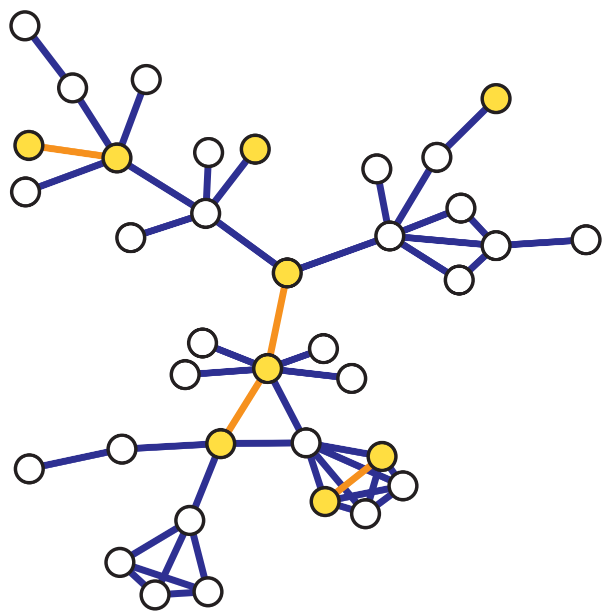
- Q: How to describe/summarize the complex interactions during a seizure?



- Statistical challenges
  - Criterion for defining ‘brain networks’
  - Choice of network summary statistics
  - Assessing significance of changes/differences

Monitoring social media

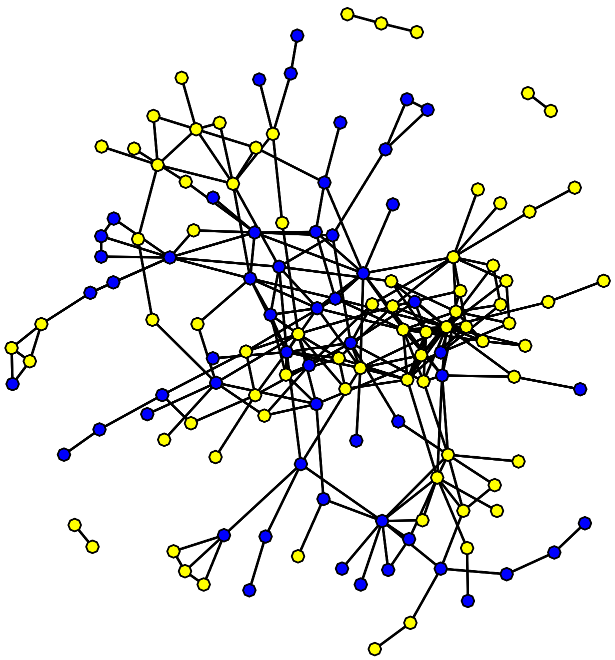
- Q: Can we monitor characteristics of massive social media networks?



- Statistical challenges
  - Computer protocols correspond to what sampling designs?
  - What sort of biases are inherent to the sampling?
  - Can we compensate for those biases?

Predicting protein function

- 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 ($\sim$ 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

  $\Rightarrow$ Issues on data collection, data management and computing

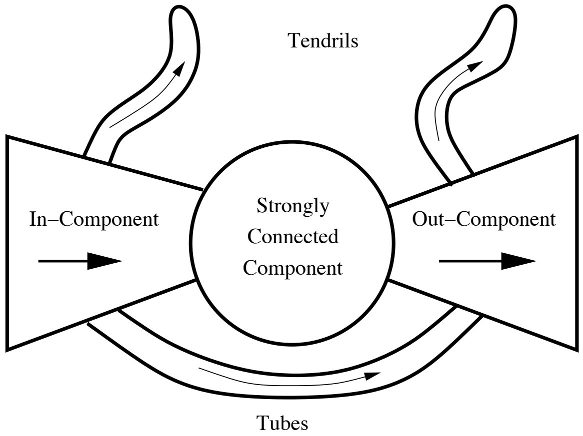
(II) Descriptive analysis and properties of large networks ($\sim$ 7 lectures)

(III) Sampling, modeling and inference of networks ($\sim$ 9 lectures)

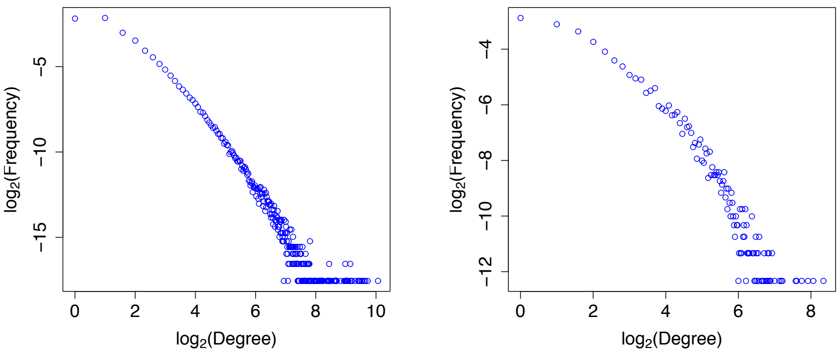
(IV) Processes evolving over network graphs ($\sim$ 8 lectures)

Descriptive analysis and properties of networks

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

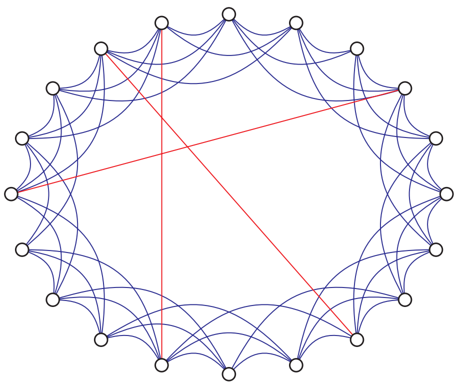


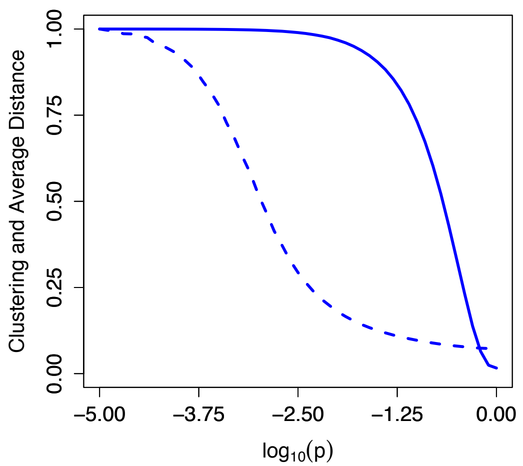
- Power-law degree distributions are ubiquitous in real-world networks



- 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)

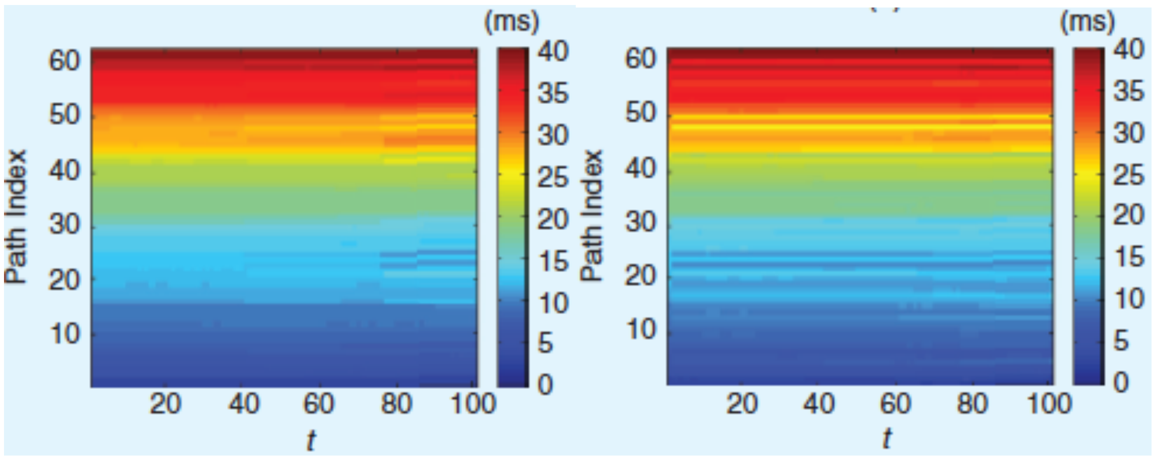




- Of interest: random graph models, network topology inference, growth models for evolving networks, preferential attachment
- Applications: detecting motifs, inferring gene-regulatory interactions, mapping the Internet, predicting popularity in Twitter

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-thruth" 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