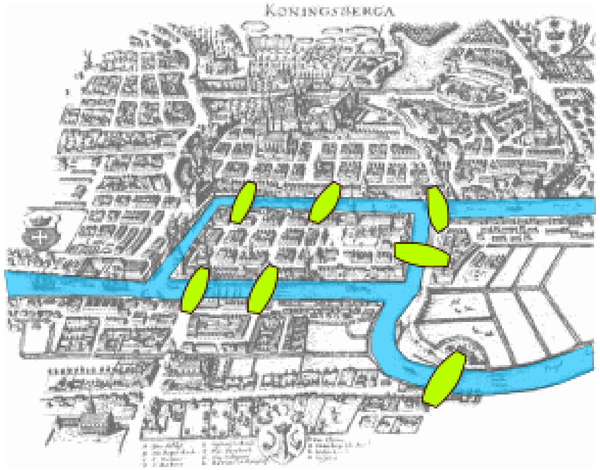
CS4973

network analysis

Plan for the day

- **Graph Theory - Origin Story**
- **Why analyze networks?**
  - **Case studies**
- **Social media mining – what and why?**
  - **Data characteristics**
  - **Challenges in social media mining**

Seven Bridges of Königsberg
Leonhard Euler, 1736

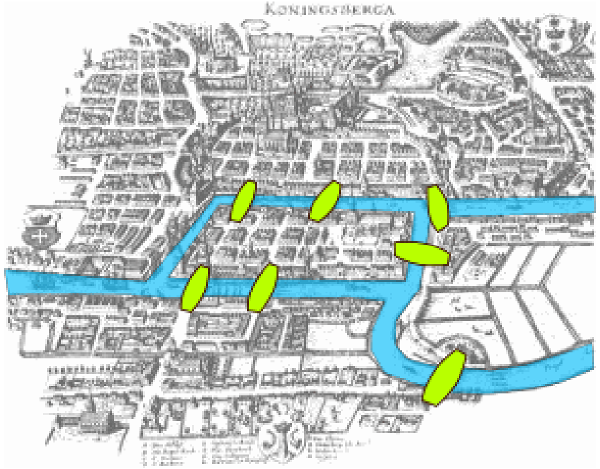


- The city of Königsberg (now Kaliningrad, Russia) was divided by the Pregel River, with two islands connected to the mainland by seven bridges.

- The challenge was to determine whether it was possible to devise a walk through the city that crossed each bridge exactly once.

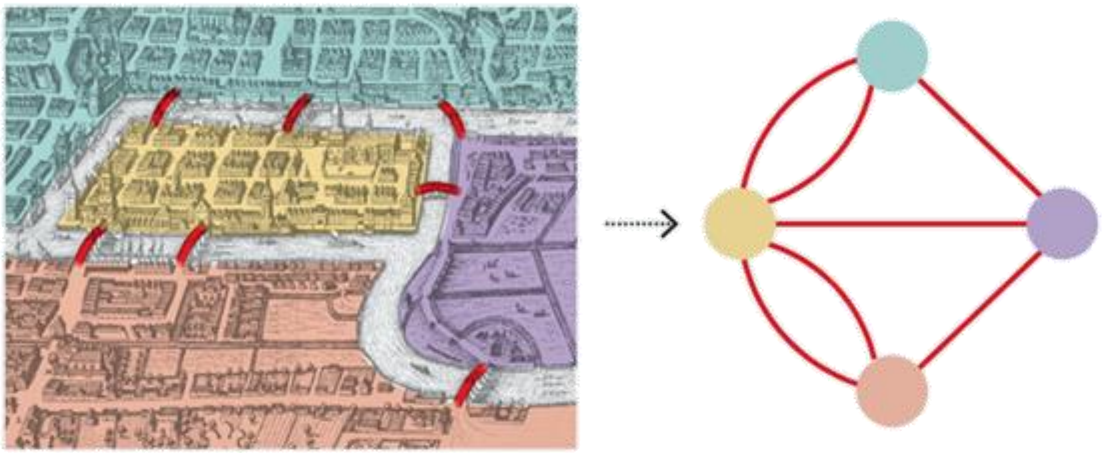
Seven Bridges of Königsberg

Leonhard Euler, 1736



How would you solve this problem?

Seven Bridges of Königsberg
Leonhard Euler, 1736



- Abstraction – remove extraneous information
- What we’re left with, is a ***graph***.
- *Can you solve it now?*

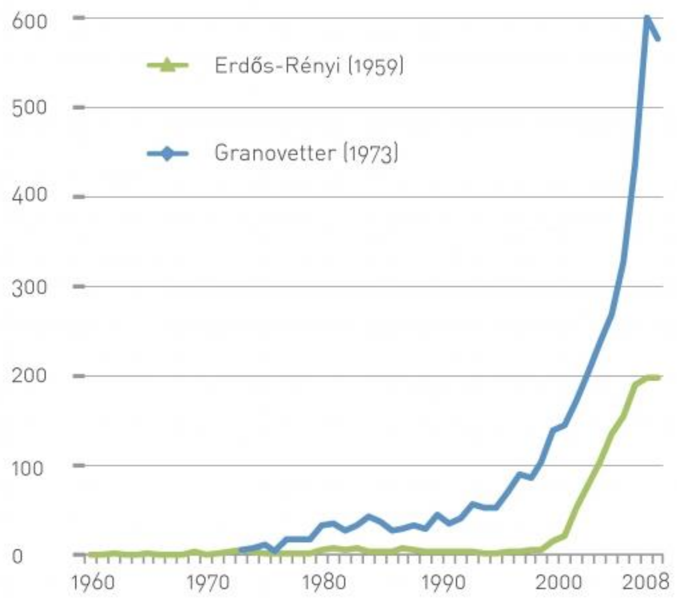
Other Graph Problems

- Route planning (*path finding*)
- Playing games (*also path finding)*
- Constraint satisfaction/scheduling (*min cover*)

Fun (sad) fact: graph theory helps me track down academic integrity violations…

Network Science

- Relatively new discipline (21st century)
- Interest skyrockets around 2000
- Internet helps centralize information
- Hardware advances improve problem scaling capabilities



Citation counts for two seminal graph theory papers

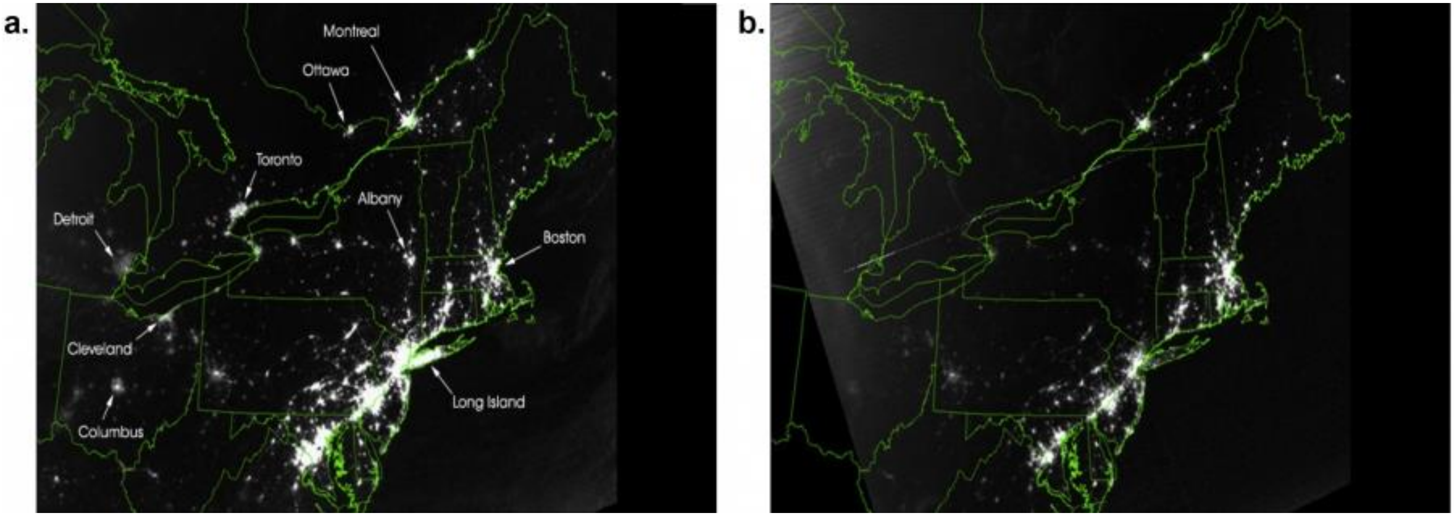
Data Availability

- 1990s – central databases for chemical reactions within cells
- Imdb – Hollywood networks (six degrees of Kevin Bacon)
- Google Scholar/ArXiv networks – co-authorship tracking
- … and many more

Interdisciplinarity

- Network Science gives researchers a common framework
- Cell biologists, psychologists, and computer scientists often solve variations on the same problem – mapping interconnectivity
- Empirical & data-driven, computational in nature
- Societal Impact
  - Drug discovery (network biology)
  - Web search
  - Security (fighting terrorism)
  - Understanding epidemics/pandemics
  - Management (uncovering internal structure)

Case Study - 1



a) Satellite image on Northeast United States on August 13th, 2003,at 9:29pm (EDT), 20 hours before the 2003 blackout

b) 5 hours after the blackout

The 2003 Northeast America Blackout

~45 million in USA and 10 million in Ontario lost power

Case Study - 1

- **This is an example of a *cascading* failure.**
- When one node in the grid fails, its load is distributed over its neighbors.
- If extra load is negligible, the network will absorb it.
- If extra load is too much, redistribution may occur until failure.

Case Study - 1

- How would you use networks to prevent future blackouts?
  - What are the entities and relationships in your model?
  - What are you studying?
  - What would you change/modify/focus on?
- Discuss with the person closest to you and write down 2-3 ideas.

Case Study - 1

- Your model may include
  - **Entities**: Power stations, substations, and distribution points.
  - **Relationships**: Transmission lines connecting nodes.
  - **Metadata**: Capacity of transmission lines (e.g., power flow limits).

- Critical Component Identification
  - Identify critical transmission **lines or power stations** that handle the majority of the power flow.
  - Highlight nodes with many connections, which could be high-risk hubs.

Case Study - 1

- Cascading Failure Simulation
  - Simulate failures on critical lines or power stations and propagate the effects using the graph structure.
  - This helps identify potential cascading paths and zones prone to blackout.
- Load Redistribution
  - Partition the grid into smaller, independently stable subnetworks.
  - Balancing power loads across the graph.
- Robustness Analysis
  - Evaluate **graph resilience** using measures like connectivity
  - Identify ways to strengthen weak areas through redundancy

Case Study - 1

- Similar failures happen within computer networks & the internet
- CPU/GPU cores load balancing
- Financial institutions, economy
- Group projects!

Network Connectedness

- Complex and interconnected systems form the backbone of modern society.

- Blessing and a curse!

    - Allows trade/exchange and lowers consumer costs
    - Allows rapid exchange of information

    - Failures have far-reaching impact
    - Allows rapid exchange of misinformation/bad information

Case Study - 2

- Facebook-Cambridge Analytica Scandal (2018)
- Data from millions of users were harvested via a personality quiz app
    - This Is Your Digital Life, Aleksandr Kogan
    - **270,000 users** installed the app, gave permissions to access Facebook data
    - Facebook policies allowed app to collect friends data without consent
    - Kogan gained knowledge about **87 million Facebook users** without their consent

Case Study - 2

- Data sold to Cambridge Analytica, against FB policies
  - Data used to build detailed psychological profiles of individuals
  - Includes political leanings, personality traits
  - Data used for Ted Cruz and Donald Trump's presidential campaigns
  - Targeted advertising, political messaging

- Scandal revealed by whistleblower, CA shuts down in 2018

- **$5 billion fine** from the Federal Trade Commission (FTC)

Case Study - 2

- How would you use network science to tackle similar problems?
  - What are the entities and relationships in your model?
  - What are you studying?
  - What would you change/modify/focus on?
- Discuss with the person closest to you and write down 2-3 ideas.

Case Study - 2

- Limiting access based on network structure
  - With proper consent, allow access only to close-knit groups, such as families
  - App access restricted to user's *subgraph*, and not unrelated areas - track this
  - Automate such partitioning using community detection algorithms

- Early Detection of Suspicious App Behavior
  - Monitor the rate of new edges created
  - Anomalies in edge creation rate could indicate abuse

- Define privacy metrics based on graph structure, policy enforcement

Social Media Mining

- Key aspects:
  - Representation of social media data
  - Pattern extraction
  - Deriving meaningful insights
- Vast amount of user-generated content
- Rich source of human behavior data
- Allows us to study trends, public opinion, business insights…

Data Characteristics

- Large scale of participation
- Openness under the guise of anonymity
- Explicit community structure
- Very rich data

Challenges

- Big Data Paradox
  - Large overall volume, but sparse individual data
  - Social media data is notoriously noisy (self-reported nature)

- Sampling Issues
  - Ensuring representative samples often a challenge
  - API limits, scraping bans
  - *I once got my WashU lab's compute nodes blocked from Google search API for 24 hours*

- Noise removal is hard
  - Extensive preprocessing vital to many data mining approaches
  - Garbage in, garbage out
  - Removing noise can worsen problem (limited data to begin with)

- Evaluation Dilemma
  - Absence of ground truth
  - Unsupervised/semi-supervised approaches

Some open questions…

- How does behavior of individuals change across sites?
  - What behaviors remain consistent and what behaviors likely change?
  - What are possible reasons behind these differences?

- Do you believe social media algorithms are enhancing or limiting your life experiences?
  - How might they be shaping our worldviews?
  - Relationships - are online friendships different from in-person ones?
  - Are we aware of these influences?

- How do you think social media will evolve in the next decade?
  - What new challenges or opportunities might arise?