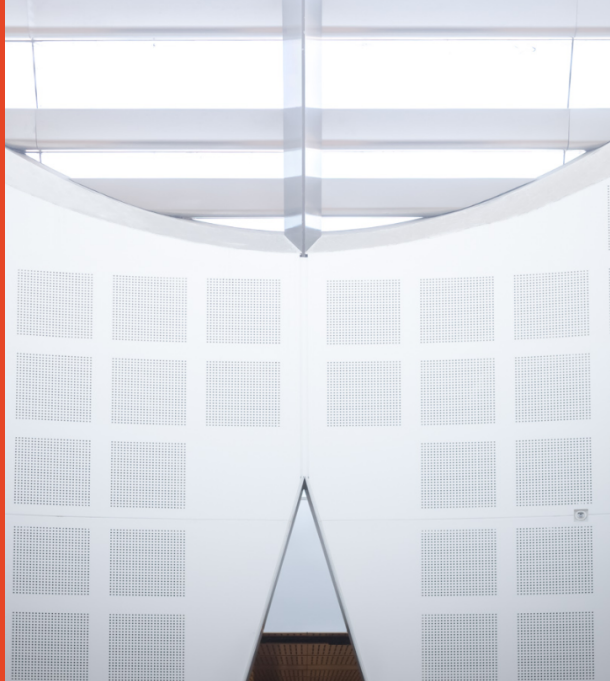
COMP5313/COMP4313 - Large Scale Networks

Week 1: Introduction

Lijun Chang

Feburary 27, 2025

THE UNIVERSITY OF SYDNEY

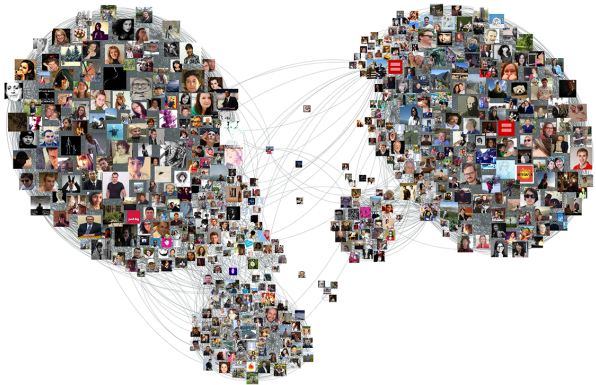


Outline

What is a Network?

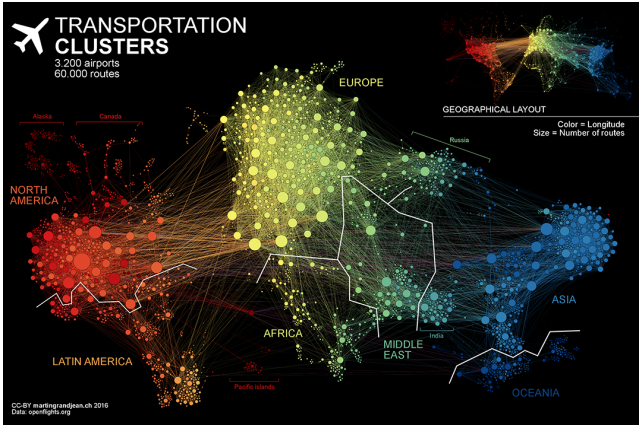
- A network is any collection of objects (nodes) in which some pairs of these objects are connected by links.
  - Flexible, many different forms of relationships or connections can be used to define links.
  - Network is the backbone of complex systems.

Online Social Networks



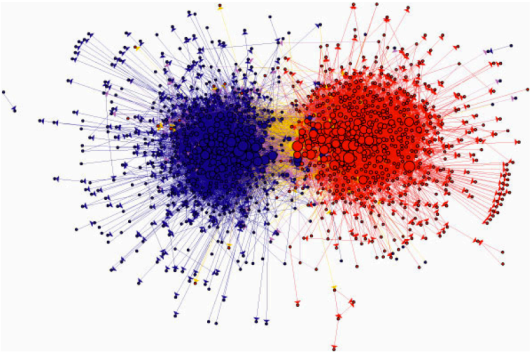
Source: http://kateto.net/

Global Air Traffic Network



Source: http://www.martingrandjean.ch/connected-world-air-traffic-network/

Blog Web Pages



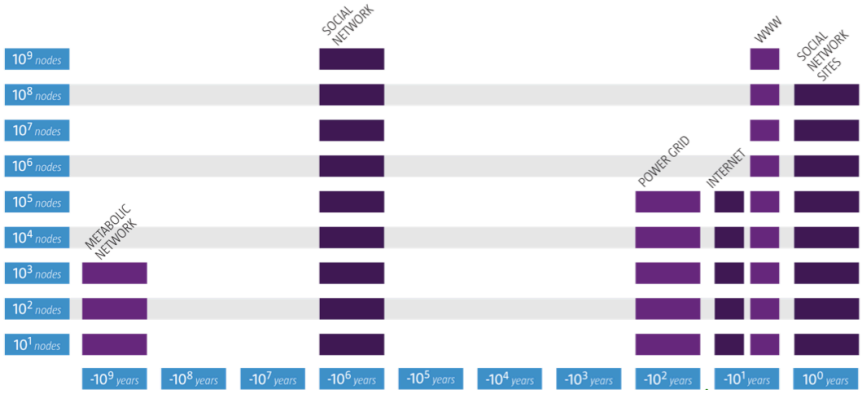
Community structure of political blogs [1]

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 us election: divided they blog," in Proceedings of the 3rd international workshop on Link discovery, pp. 36–43, ACM, 2005.

Why You Should Care About Networks?

- Universal language for describing complex data.
  - Networks from science, nature, and technology are more similar than one would expect
- Shared vocabulary between fields
  - Computer science, social science, physics, economics, statistics, biology
- Data availability (/computational challenges)
  - Web/mobile, bio, health, and medical

Why Now?



Source: http://web.stanford.edu/class/cs224w/ - By Jure Leskovec

Outline

Personnel: Myself

- **Lijun Chang** (Coordinator and Lecturer)
  - Office: Room 440, SCS Building
  - Phone: (02) 903 69756
  - Email: Lijun.Chang@sydney.edu.au
  - Web: https://sydney.edu.au/engineering/people/lijun.chang.php
  - Office hours: via email appointment

- **Research area**
  - Design efficient and scalable algorithms to process large-scale networks with billions of edges

Timetable

- **Title:** COMP5313/COMP4313 Large-Scale Networks
- **Credit points:** 6cp
- **Lectures:** Thursday 18:00-20:00
  Location: G04.270.Wilkinson Building.Wilkinson Lecture Theatre 270
- **Tutorials:** Please go to the one that is on your timetable
  Locations: Thursday 20:00-21:00, weeks 2–12
  or Friday 18:00–19:00, weeks 2–7, 9–12
  Tutor: Yuefan Shang and Mouyi Xu

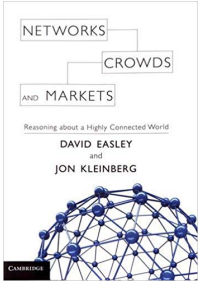
Canvas

- Course information: https://canvas.sydney.edu.au/

- Slides will be uploaded right before the lecture
- Video recording should be automatically posted on Canvas (usually the next day)
- All assessments & notifications are communicated through the Canvas website
- Questions about content & clarifications: post on Ed Discussion (access via Canvas)
  - You all are encouraged to participate in the discussions & share interesting material
  - I will reply to common questions

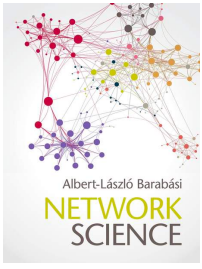
Textbook (Required)

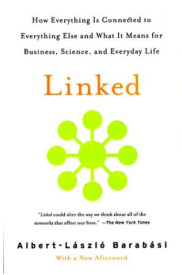
- Networks, Crowds and Markets. Easley and Kleinberg

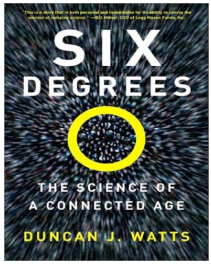


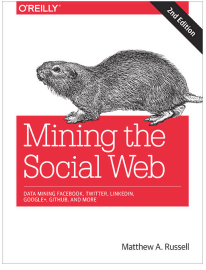
- Required textbook
- Freely available online at
  http://www.cs.cornell.edu/home/kleinber/networks-book/

Other Books (Optional)









- Albert-László Barabási. Network Science. 2016
- Albert-László Barabási. Linked: The New Science of Networks. 2002.
- Duncan Watts. Six Degrees: The Science of a Connected Age. 2003.
- Matthew A. Russell: Mining the Social Web. 2013.

Outline

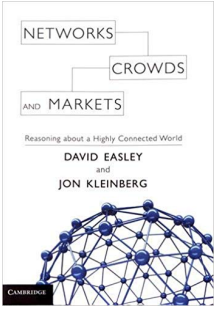
Course Description

The growing connectedness of modern society translates into simplifying global communication and accelerating spread of news, information and epidemics. The focus of this unit is on the key concepts to address the challenges induced by the recent scale shift of complex networks. In particular, the course will present how scalable solutions exploiting graph theory, sociology, game theory and probability tackle the problems of communicating (routing, diffusing, aggregating) in dynamic and social networks.

- Full details can be found at https://sydney.edu.au/units/COMP5313

Topics Covered

- We cover:
  - Graph theory
  - Information networks
  - Network dynamics
  - Basic graph machine learning
- We do not cover:
  - Game theory
  - Economics
  - Auctions & Markets



Assessments

- **Assignment 1 [15%]**
  - Solve problems
  - Available from week 3 and due on week 6
- **Midterm quiz (closed book, in-person) [10%]**
  - Answer multiple-choice questions
  - Time: during week 8's Thursday lab time (20:00–21:00, April 17, 2025)
- **Assignment 2 - group project [25%]**
  - Normally, teams of two students
  - Teams of one or three students need my approval
  - Project report (4–6 pages) and recorded presentation are due on week 11
  - More details next week
- **Final exam [50]%**
  - During exam period

Assessments

- It is a policy of the School of Computer Science that in order to pass this unit,
  1. A student must achieve **at least 40% in the final examination.**
  2. A student must also achieve **an overall mark of 50 or more.**

- Any student not meeting these requirements may be given a maximum final mark of no more than 45.

- ChatGPT (and other AI-assisted tools) are not allowed to be used for directly generating the assignment report and the presentation recording.
  - They can be used for polishing the writing.

Special Consideration

- If your performance on assessments is affected by illness or misadventure
- Follow proper bureaucratic procedures
  - Have professional practitioner sign special USyd form
  - Submit application for special consideration online, upload scans
  - Note you have to apply within three days of the assessment
  - http://sydney.edu.au/current_students/special_consideration/
- Also, notify coordinator by email as soon as anything begins to go wrong
- There is a similar process if you need special arrangements, eg., for religious observance, military service, representative sports

Late Submissions

- Penalties for lateness when special consideration is not granted: 5% per calendar day according to the Clause 7A of the assessment procedures.
  - *E.g. An assignment that would normally get 9/10 and is 2 days late loses 10% of the full 10 marks, i.e. new mark = 8/10*
  - *E.g. An assignment that would normally get 5/10 and is 5 days late loses 25% of the full 10 marks, i.e. new mark = 2.5/10*
  - Assignments more than 10 days late get 0.

- Late attendance of midterm quiz is not allowed.

Outline

Tentative Schedule

- **Social networks**
  - Week 2 - Graph ties
  - Week 3 - Structural balance and network evolution
  - Week 4 - Graph community detection and partitioning
- **Information networks**
  - Week 5 - Hubs & Authorities
  - Week 6 - Google's PageRank algorithm
- **Machine learning on graphs**
  - Week 7 - Machine learning on graphs (I)
  - Week 8 - Machine learning on graphs (II)
- **Network dynamics**
  - Week 9 - Information cascades and Power laws
  - Week 10 - Structural models for decentralized search
  - Week 11 - Peer-to-Peer networks

1) Social Networks: Ties

- Concepts that were first studied by social scientists on small-scale networks and then verified by computer scientists on large-scale networks.
  - Small world phenomenon
  - Strength of weak ties
  - local bridges
  - strong triadic closure

- Strong ties, representing close and frequent social contacts, tend to be embedded in tightly-linked regions of the network
- Weak ties, representing more casual and distinct social contacts, tend to cross between these regions
- At a global scale, it suggests some of the ways in which weak ties can act as short-cuts that link together distant parts of the world, resulting in the phenomenon colloquially known as the six degrees of separation.

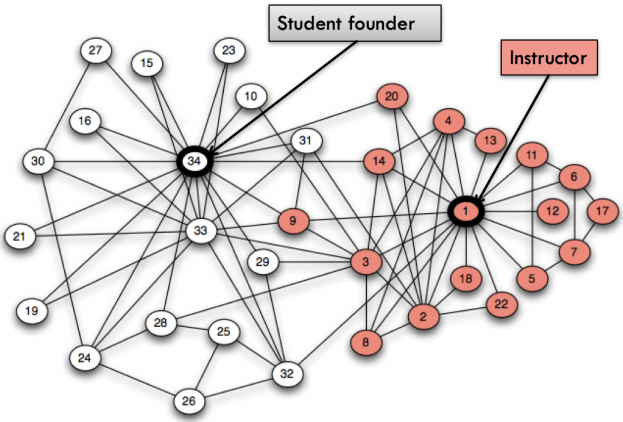
Social Networks: Structural Balance and Network Evolution

- Structural balance
  - The relationship can be positive (e.g., friends) and negative (e.g., enemies)
  - How positive and negative relationships affect the structure of a network

- Network Evolution
  - Homophily is one of the most basic notions governing the structure of social networks
    - Your friends are generally similar to you in terms of ethnicity, age, and other mutable characteristics (e.g., occupations, interests, beliefs)
    - Homophily can divide a social network into densely connected homogeneous parts that are weakly connected to each other
  - We will also look at how links are formed in social networks.

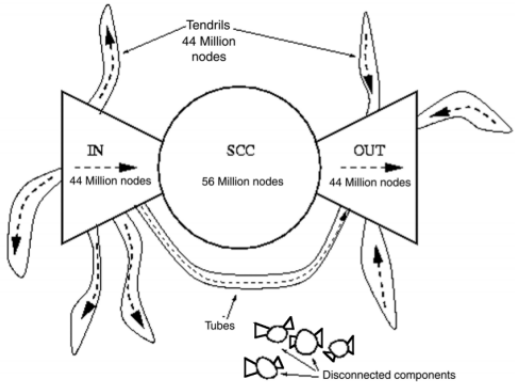
Social Networks: Community Detection

- Tightly-knit groups are usually regarded as communities.



- How to identify the tightly-knit groups?

2) Information Networks

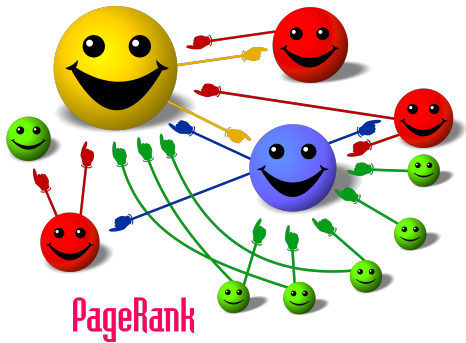


Bow-tie structure of the web [2]

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the web," Computer networks, vol. 33, no. 1-6, pp. 309–320, 2000.

Information Networks: Ranking Webpages

- Web search engines such as Google make extensive use of network structure in evaluating the quality and relevance of Web pages.

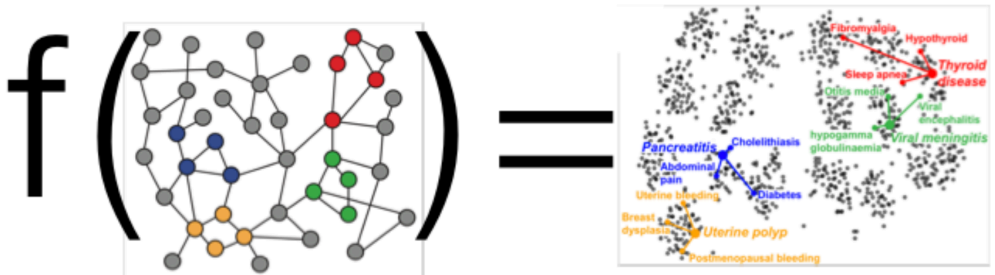


Information Networks: PageRank and Personalized PageRank

- PageRank is used as a network centrality measure
  - Yields the importance of each node in light of the entire graph structure
  - At each time step, the random surfer has two options
    - With probability α, follow a link at random
    - With probability $1 - \alpha$, jump to a random page among all pages

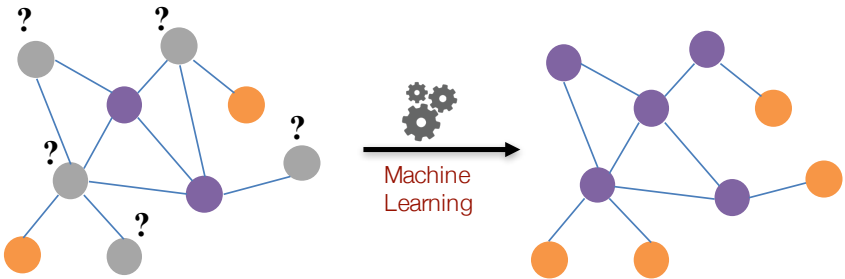
- Personalized PageRank is used to illuminate a region of a large graph around a target set S of interest.
  - At each time step, the random surfer has two options
    - With probability α, follow a link at random
    - With probability $1 - \alpha$, jump to a random page among a set S of pre-selected pages

3) Machine Learning on Graphs



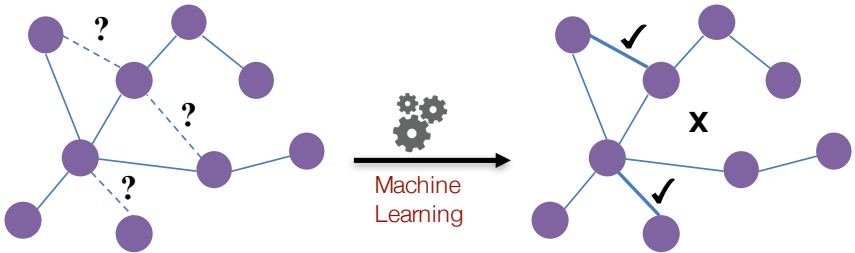
Source: Jure Leskovec Graph Representation Learning CS224W

Machine Learning on Graphs: Node Classification



Source: Jure Leskovec Graph Representation Learning CS224W

Machine Learning on Graphs: Link Prediction



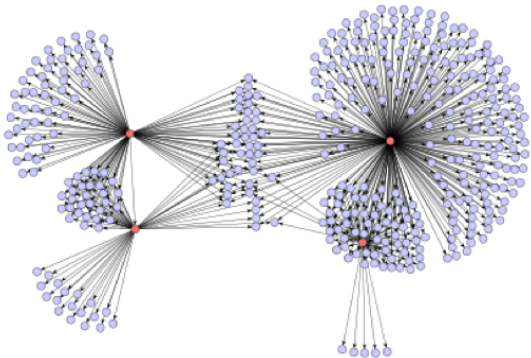
Source: Jure Leskovec Graph Representation Learning CS224W

4) Network Dynamics: Information Cascade

- A cascading behaviour spreads from one person to another
  - People influence each other's behaviour: if you see more and more people doing something, you generally become more likely to do it, too
  - Like a biological epidemics
  - Called "social contagion"

- Cascading effects may arise when individuals have incentives to adopt the behaviour of their neighbours in the network
  - A new behaviour starts with only a small set of initial adopters
  - Then can spread radially outward through the network

Network Dynamics: Social Contagion



E-mail recommendations for a particular Japanese graphic novel spread outward from four initial purchasers [3]

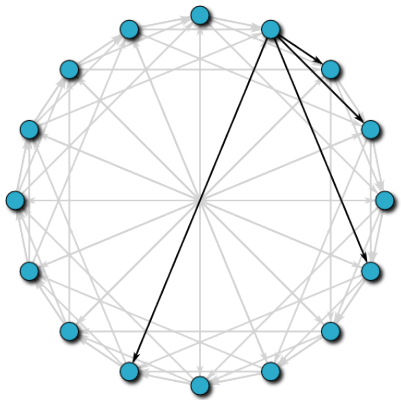
[3] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," ACM Transactions on the Web (TWEB), vol. 1, no. 1, p. 5, 2007.

Applications

- When understood, the properties of networks can be applied elsewhere
- News spreading: Six degrees of separation is useful to disseminate information rapidly
- File lookup: Selecting shortcuts adequately can help navigating rapidly to a destination

Applications

The overlay networks use shortcuts to effectively retrieve files in a peer-to-peer file sharing system



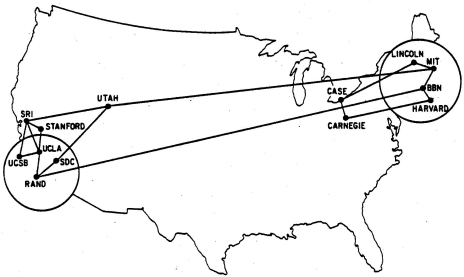
Outline

Networks and Graphs

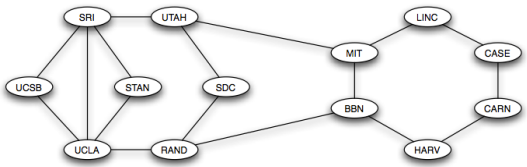
- A network often denotes a real system.
  - With nodes connected by links
- A graph is a mathematical representation
  - With vertices connected by edges
- These terms are often used interchangeably
- A graph $G = (V, E)$ has a set of vertices V and a set of edges E.

Mathematical Representation

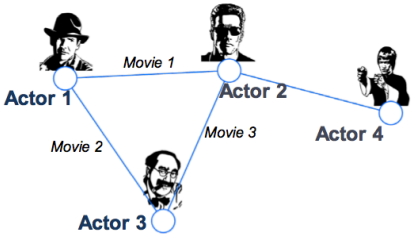
- ARPANET (former Internet) in 1970

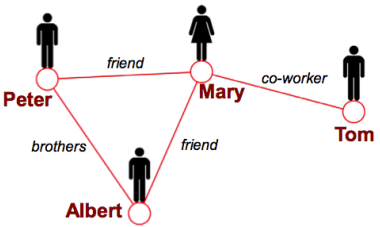


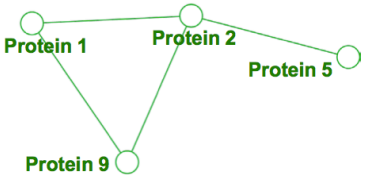
- A simplified graph representation of ARPANET

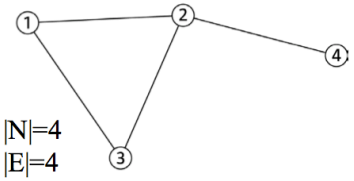


Graph Examples



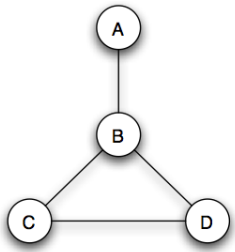




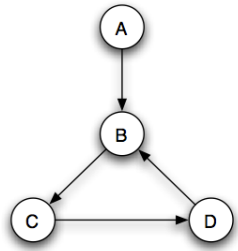


Undirected vs. Directed Graphs

- Undirected graphs
  - Ex: Facebook friendship

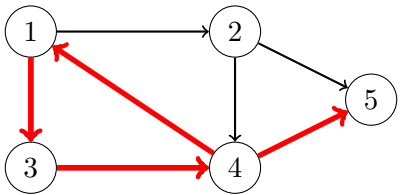


- Directed graphs
  - Ex: The world wide web



Path and Cycle

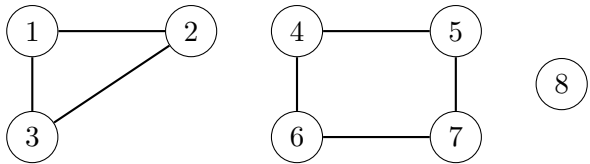
- A path leads from vertex u to vertex v through edges of the graph.
  - The length of a path is the number of edges in it.



- A path is a cycle if the first vertex and last vertex are the same.
- A path is simple if each node appears at most once in the path.

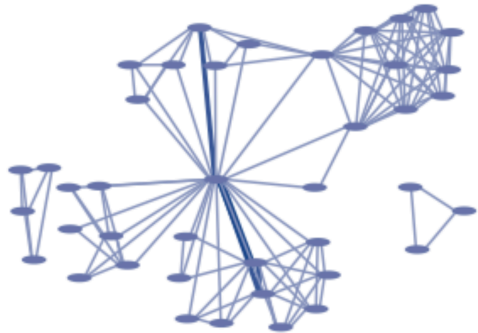
Connected Components

- An undirected graph is connected if there is a path between any two vertices.
- The maximal connected subgraphs of an undirected graph are called its connected components.



Connected Components

- The collaboration graph of the biological research center Structural Genomics of Pathogenic Protozoa (SGPP).
- How many connected components does it have?



Giant (Connected) Component

Qualitative think about connected components of large scale networks

- Consider the social network world-wide
- Is it connected?
  - Presumably not
  - A person with no living friends will be a one-node component
- But, it probably has a giant component: a connected component with a large fraction of all nodes
  - You probably have friends in a foreign country
  - You are then in the same component as them
  - Their friends and descendants are also in the same component

Giant Component

When a network contains a giant component it contains generally only one

- Try to imagine that there were two giant components, each with hundreds of millions of people
- All it would take is a single edge from someone in the first of these components to someone in the second to become one giant component
- It is generally unconceivable that such edge would not form!

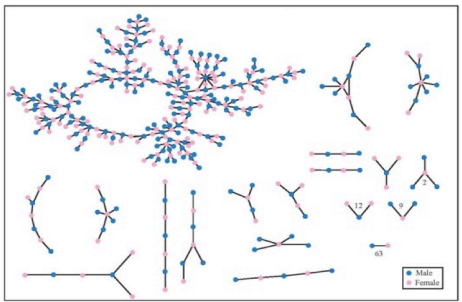
Merge of Giant Components

The merge of giant components may have dramatic consequences

- Think about the colonisation of Australia (1788)
  - One could see British and Australians as two giant connected components that co-existed for a long time
  - Human diseases evolved independently
  - Upon colonization the two components merged
- Series of European diseases such as
  - Measles
  - Smallpox
  - Tuberculosis
- A smallpox epidemic in 1789 is estimated to have killed up to 90% of the Darug people

Components

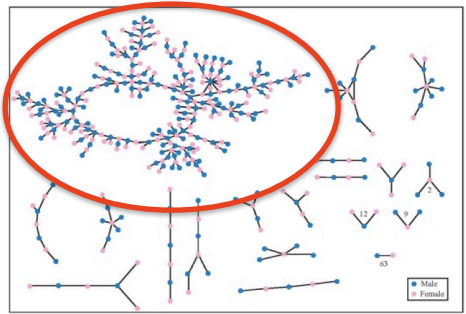
- A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted



- What do you observe?

Components

- A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted



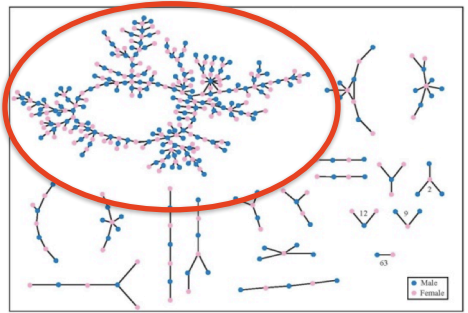
- What do you observe?

Components

The large component plays an important role in the spread of sexually transmitted diseases

- Take a student who had a single partner during the study
- Without knowing it, he is likely part of the large component

$\Longrightarrow$ So, he may be part of many paths of potential transmissions



Distance

- Besides connectivity through a path, the length of a path is also important
  - The length of a path is the number of edges the path contains
- The distance between two vertices in a graph is the length of the shortest path between them
  - If the distance from u to v is 3, then there is no path between them with length < 3
- The diameter of a graph is the longest distance among all pairs of nodes
- Determining the distance may seem easy on small networks, but requires a systematic way for large-scale networks

Distance

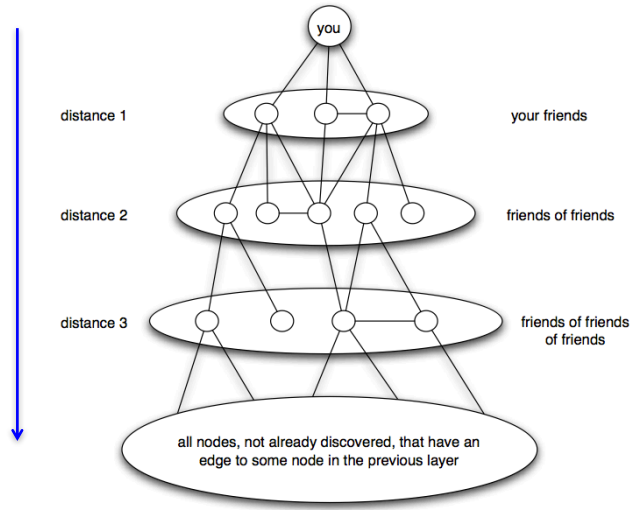
Breadth-First Search: searching network from starting node, closest nodes first

1. You first declare all of your actual friends to be at distance 1.
2. You then find all of their friends (not counting people who are already friends of yours), and declare these to be at distance 2.
3. Then you find all of their friends (again, not counting people who you've already found at distances 1 and 2) and declare these to be at distance 3

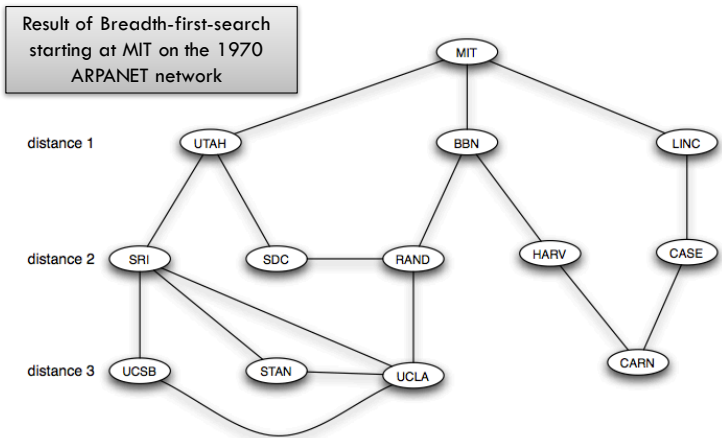
Continuing in this way, you search in successive layers, each representing the next distance out. Each new layer is built from all those nodes that *(i) have not already been discovered in earlier layers*, and that *(ii) have an edge to some node in the previous layer.*

Distance

Breadth-First Search discovers layers one at a time from the starting node downward



Distance


Result of Breadth-first-search starting at MIT on the 1970 ARPANET network
distance 1
distance 2
distance 3
MIT
UTAH
BBN
LINC
SRI
SDC
RAND
HARV
CASE
UCSB
STAN
UCLA
CARN

Conclusion

- A network is a graph with vertices/nodes and edges
- Large-scale networks are hard to analyze
  - Manually is impossible
  - Sometimes too large for computers as well
- Important properties of networks
  - Path, length, distance
  - connected component
- Intuition tells us that large social networks tend to have:
  - One giant connected component (if not yet, it is likely that merging will occur)
  - Other connected components tend to be much smaller

Reading

- Reading for this week
  - Chapters 1 and 2 of the textbook
- Reading for next week
  - Chapter 3 of the textbook, excluding the advanced material
- NO TUTORIAL THIS WEEK