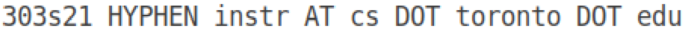
Social and Information Networks

University of Toronto CSC303
Winter/Spring 2021

Week 1: January 11-15 (2021)

Course Organization

Course Instructor: Ian Berlot-Attwell



- Email: 303s21 HYPHEN instr AT cs DOT toronto DOT edu

Teaching Assistants: Ruijian An, Soroush Ebadian, Julian Posada, Fengwei Sun

Communications

Communication:

1. Course Web page: source of first resort
`https://www.cs.toronto.edu/~ianberlot/303s21/`
2. Announcements will also be sent via Quercus, and information that shouldn't be accessible to the public (e.g. Zoom link) will also be on Quercus
3. Discussion board: **Discourse** for questions of general interest
`https://bb-2021-01.teach.cs.toronto.edu/c/csc303`
Instructor and TA will monitor and respond as appropriate. I encourage questions in-class which leads to less confusion especially with regard to technical questions.
4. Office hours: TBA

Course Materials

Course Materials: CSC303 is based on the text by Easley and Kleinberg, previous parts of (the now discontinued) CSC200 by Borodin and Craig Boutilier, and the current course developed by Ashton Anderson at UTSC.

1. Text: D. Easley, J. Kleinberg. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, 2010. Online version available at `http://www.cs.cornell.edu/home/kleinber/networks-book/` We will supplement with some topics and material not in the text.
2. Additional materials will be linked on course web page.

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - Lectures or Tutorials Monday, Wednesday and Friday. We will usually have the tutorials on Wednesdays, and lectures Mondays and Fridays. **Zoom Links on Quercus**
  - However, if necessary, we will sometimes rearrange the schedule between tutorial time and lecture times. You should be available M,W,F 15:00-16:00 each week whether it is a lecture or a tutorial.
- More generally
  - Readings posted on web site usually posted in advance.
  - The readings often (but not always!) cover all or most of the lecture material – I suggest doing them in advance if possible
  - Lecture slides (some detailed, some less so) will usually be posted one or two days *after* the class. **You are responsible (i.e., can be tested) for information that occurs in lectures and tutorials.**

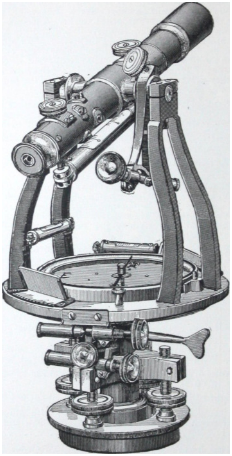
Lecture/Tutorial/Course Structure

- Lectures & Tutorials will be recorded
  - Links to recordings will be made available on the course website (only accessible with a UTORid)
  - **Let me know ASAP if you have objections to being recorded so that an arrangement can be made**
  - Although recordings are available, I *strongly* suggest attending if you can – the opportunity to interact with myself, the TAs, and your classmates is invaluable for your learning

- Feedback, suggestions, & ideas for improving the course are welcome via
  - Email
  - Anonymously via `https://forms.gle/8dXBGyRoD6tWLYXr5`

Survey on Office Hours & Delivery Method

- See Quercus
- 5 minutes
- Closes Friday Jan 15th



Preparation, reading scheme and schedule

You should be comfortable with basic probability and discrete math concepts as would be covered in the prerequisites. I have posted a probability primer on the course web page.

Grading Scheme

1. Participation: 5% – Quercus quizzes
2. Assignments: Two, each worth 15% = 30%
   Tentative due dates: February 12 and March 26
3. Critical review of a current article (groups 3-4): Worth 10%
   Tentative due date: March 26
4. Term Test (take-home): Worth 20%
   Tentative date: March 12-14, should take you 2-4 hours
5. Final Exam (take-home): Worth 35%
   Tentative date: TBD, 48 hour window, should take you 3-5 hours

Be careful! Feb 12 is sooner than you'd think, and a lot of material is due in the last few weeks.

Policies

1. No late submissions accepted beyond 12 2-hour grace tokens for assignments. But I do make an individual alternative grading scheme to accommodate medical and other legitimate issues.
2. All requests for remarking must be submitted on Markus within one week of work being graded. The only exception is for any calculation errors in adding up grades which I can correct immediately..
3. Collaboration and Plagiarism: In general, we encourage discussion of course materials. However, any work submitted must be your own! Advice: do not take away written notes from discussions about any work you will be submitting. Any material you obtain from a published source must be properly cited.
4. The "20%" rule: For any question or subquestion on any quiz, test, assignment or the final exam, you will receive 20% of the assigned question credit if you state "I do not know how to answer this question". That is, it is important to know what you do not know. If you have partial ideas then provide them; but no credit will be given for answers that do not show any understanding of the question.

What’s in a name? Graphs or Networks?

Networks are graphs with (for some people) different terminology where graphs have vertices connected by edges, and networks have nodes connected by links. I do not worry about this “convention”, to the extent it is really a vague convention without any real significance.

Here is one explanation for the different terminology: We use networks for settings where we think of links transmitting or transporting “things” (e.g. information, physical objects, friendship).

Many different types of networks

- Social networks
- Information networks
- Transportation networks
- Communication networks
- Biological networks (e.g., protein interactions)
- Neural networks

Visualizing Networks

- nodes: entities (people, countries, companies, organizations, . . .)
- links (may be directed or weighted): relationship between entities
  - friendship, classmates, did business together, viewed the same web pages, . . .
  - membership in a club, class, political party, . . .

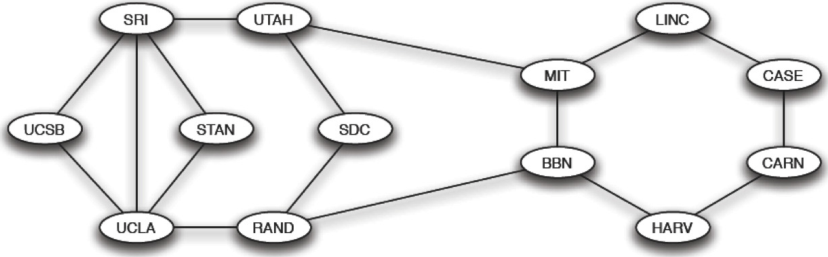
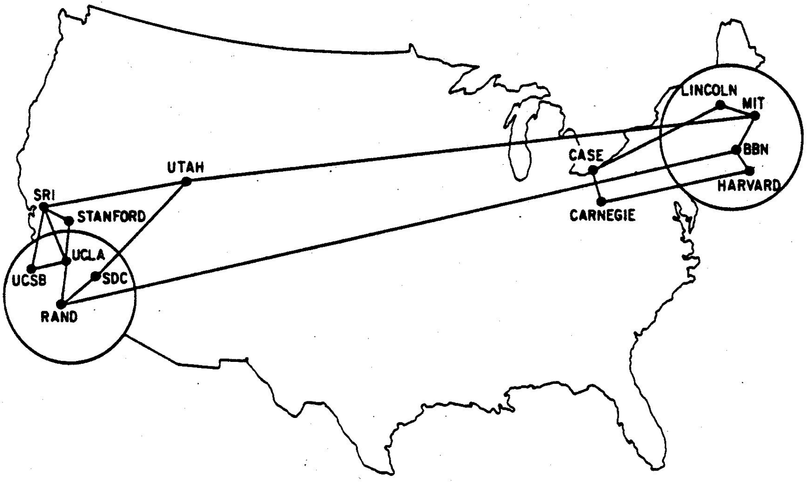


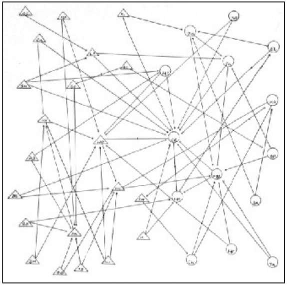
Figure: Initial internet: Dec. 1970 [E&K, Ch.2]

December 1970 internet visualized geographically [Heart et al 1978]

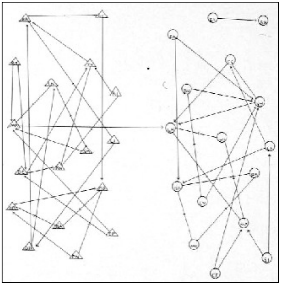


The first social network analysis

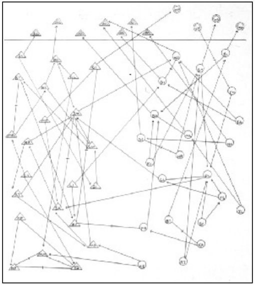
In his **1934** book *Who Shall Survive: A New Approach to the Problem of Human Interrelations*, Jacob Moreno (Romanian-US psychiatrist) introduced *sociograms* and used these graphs/networks to understand relationships. In one study (that was repeated to test changes) he asked each child in various elementary grades at a public school to choose two children to sit next to in class. He used this to study inter-gender relationships (and other relationships). Here boys are depicted by triangles and girls by circles.



1st grade



4th grade



8th grade

A closer look at grade 1 in Moreno sociogram

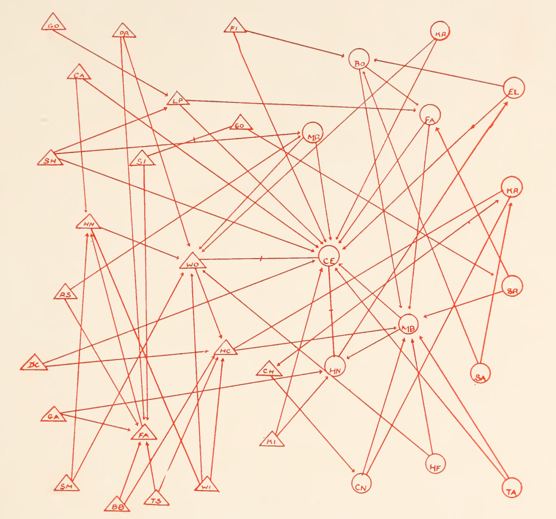


Figure: 21 boys, 14 girls. Directed graph. Most nodes have out-degree 2. 18 are not chosen, thus having in-degree 0. Note also that there are some “stars” with high in-degree.

A closer look at grade 4 in Moreno sociogram

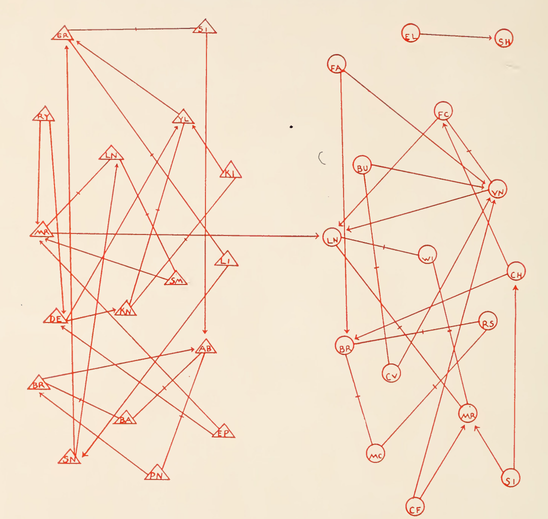


Figure: 17 boys, 16 girls. Directed graph with 6 unchosen having in-degree 0. Moreno depicted his graphs to emphasize inter-gender relations. Note only one edge from a boy to a girl.

A closer look at grade 8 in Moreno sociogram

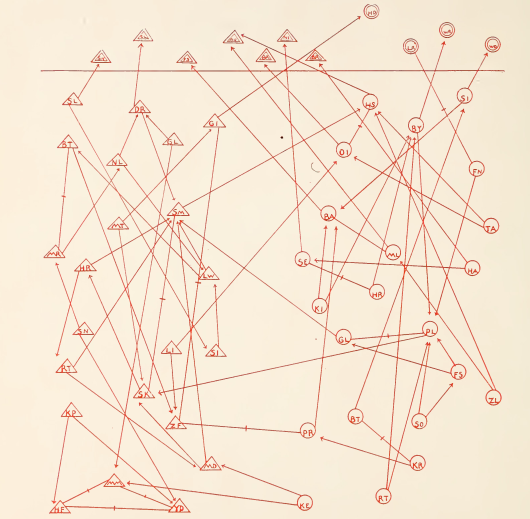


Figure: 22 boys, 22 girls. Directed graph with 12 unchosen having in-degree 0. Some increase in inter-gender relations. Double triangles and circles above line indicate individuals outside of the class.

Romantic Relationships [Bearman et al, 2004]

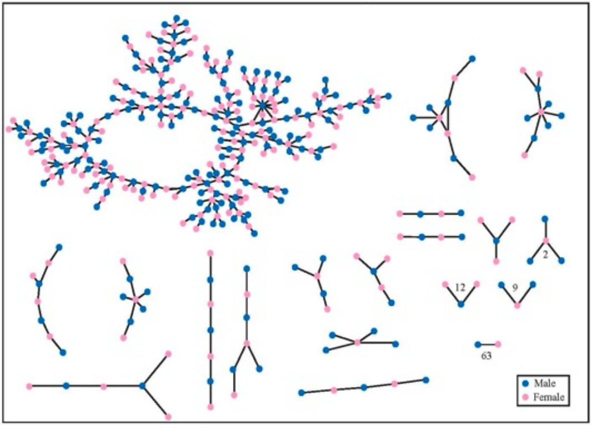


Figure: Dating network in US high school over 18 months.

- Illustrates common structural properties of many networks
- What is the benefit of understanding this network structure?

Kidney Exchange: Swap Chains

- Waiting list for kidney donation: approximately 100K in US and growing (i.e., new patients added but many deaths while waiting). The wait for a deceased donor could be 5 years and longer.
- Live kidney donations becoming somewhat more common in N.A. to get around waiting list problems: requires donor-recipient pairs
- Exchange: supports willing pairs who are incompatible
  1. allows multiway-exchange
  2. supported by sophisticated algorithms to find matches

Kidney Exchange: Swap Chains

- Waiting list for kidney donation: approximately 100K in US and growing (i.e., new patients added but many deaths while waiting). The wait for a deceased donor could be 5 years and longer.
- Live kidney donations becoming somewhat more common in N.A. to get around waiting list problems: requires donor-recipient pairs
- Exchange: supports willing pairs who are incompatible
  1. allows multiway-exchange
  2. supported by sophisticated algorithms to find matches
- But what if someone renegs? ⇒ Cycles require simultaneous transplantation; Paths require an altruistic donor!

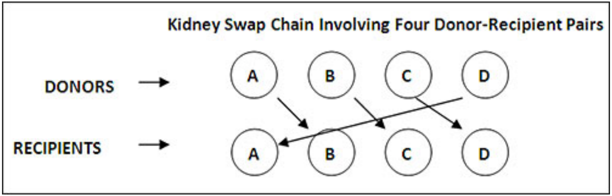
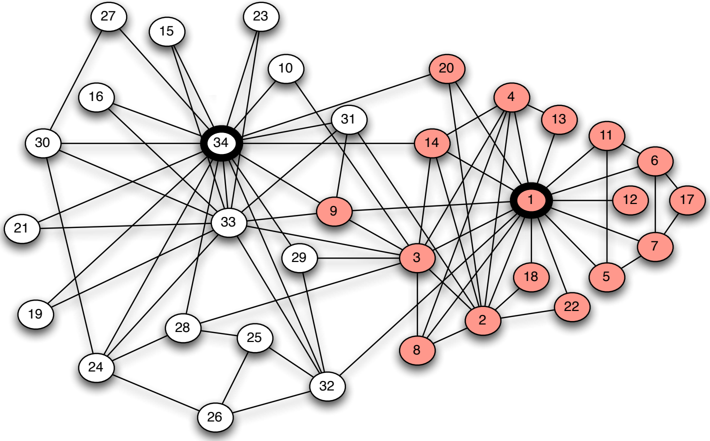


Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

Communities: Karate club division



Karate Club social network, Zachary 1977

Figure: Karate club splis into two clubs

Communities: 2004 Political blogsphere

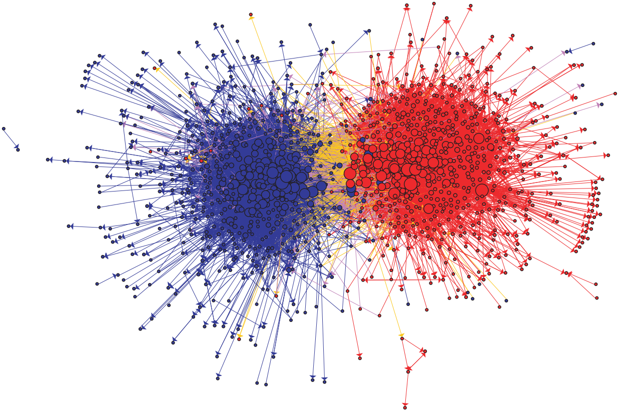


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones from conservative to liberal. The size of each blog reflects the number of other blogs that link to it.

Figure: [E&K, Fig 1.4]

Communities: 2017 Twitter online discourse regarding Black Lives Matter

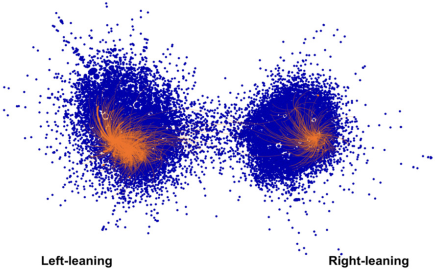


Fig. 1. Retweet Network Graph: RU-IRA Agents in #BlackLivesMatter Discourse. The graph (originally published [3]) shows accounts active in Twitter conversations about #BlackLivesMatter and shooting events in 2016. Each node is an account. Accounts are closer together when one account retweeted another account. The structural graph shows two distinct communities (pro-BlackLivesMatter on the left; anti-BlackLivesMatter on the right).

Accounts colored orange were determined by Twitter to have been operated by Russia's Internet Research Agency. Orange lines represent retweets of those account, showing how their content echoed across the different communities.
The graph shows IRA agents active in both "sides" of that discourse.

Figure: From Starbird et al [2017, 2019]

Communities and hierarchical structure: Email communication

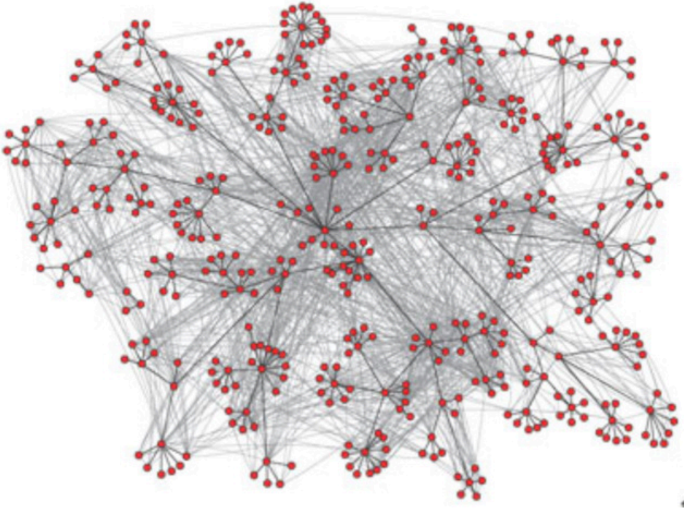
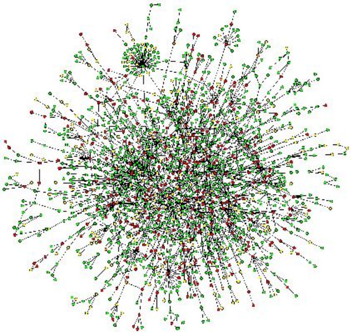


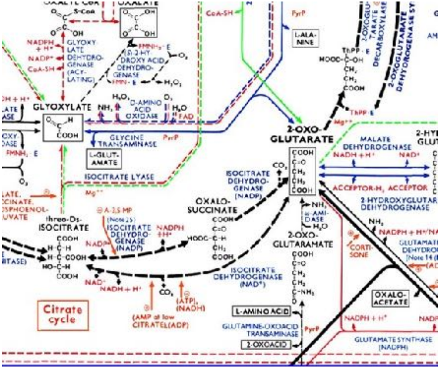
Figure: Email communication among 436 employees of Hewlett Packard Research Lab, superimposed on the organizational hierarchy [Fig 1.2, EK textbook]

Protein-protein interaction network



Protein-Protein Interaction Networks
Nodes: Proteins
Edges: 'physical' interactions

Metabolic network



Metabolic networks

Nodes: Metabolites and enzymes

Edges: Chemical reactions

The web as a directed graph of hyperlinks

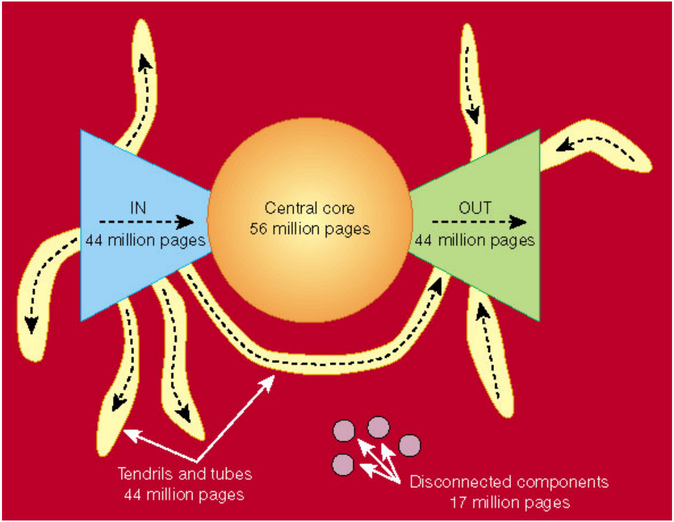


Figure: A schematic picture of the bow tie structure of the 1999 Web. Although the numbers are outdated, the structure has persisted. [Fig 13.7, EK textbook]

The current interest in networks

- Clearly there are complex systems and networks that we are in contact with daily.
- The population of the world can be thought of as social network of approximately 7.8 billion people. As of the second quarter of 2020, the people on Facebook are a *subnetwork* of approximately 2.7 billion active monthly users.
- The language of networks and graph analysis provides a common language and framework to study systems in diverse disciplines. Moreover, networks relating to diverse disciplines may sometimes share common features and analysis.
- The current impact of social and information networks will almost surely continue to escalate (even if Facebook and other social networks are under increasing pressure to protect privacy and eliminate "bad actors").

What can one accomplish by studying networks

We use networks as **a model** of real systems. As such, we always have to keep in mind the goals of any model which necessarily simplifies things to make analysis possible.

In studying social and information networks we can hopefully

- Discover interesting phenomena and statistical properties of the network and the system it attempts to model.
- Formulate hypotheses as to say how networks form and evolve over time
- Predict behaviour for the system being modeled.

And how do we accomplish stated goals

Much of what people do in this field is empirical analysis. We formulate our network model, hypotheses and predictions and then compare against real world (or sometimes synthetically generated) data.

Sometimes we can theoretically analyze properties of a network and then again compare to real or synthetic data.

What are the challenges?

- Real world data is sometimes hard to obtain. For example, search engine companies treat much of what they do as proprietary.
- Many graph theory problems are known to be computationally difficult (i.e., NP hard) and given the size of many networks, results can often only be approximated and even then this may require a significant amount of specialized heuristics and approaches to help overcome (to some extent) computational limitations.
- And we are always faced with the difficulty of bridging the simplification of a model with that of the many real world details that are lost in the abstraction.

Network concepts used in this course

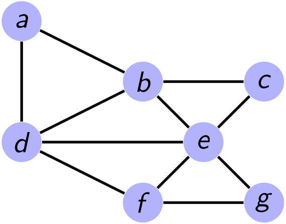
- Two main mathematical subjects of primary relevance to this course:
  1. graph theoretic concepts
  2. probability

- In motivating the course, we have already seen a number of examples of networks and hinted at some basic graph-theoretic concepts. We will now continue that discussion (i.e. material from Chapter 2 of the text) and for part of the next lecture before moving on to Chapter 3.

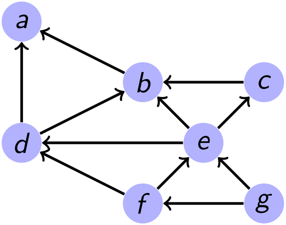
- We use the previous examples and some new ones to illustrate the basic graph concepts and terminology we will be using.

Graphs: come in two varieties

1. undirected graphs (graph usually means an undirected graph.)



2. directed graphs (often called di-graphs).



Visualizing Networks as Graphs

- nodes: entities (people, countries, companies, organizations, . . .)
- links (may be directed or weighted): relationship between entities
  - friendship, classmates, did business together, viewed the same web pages, . . .
  - membership in a club, class, political party, . . .

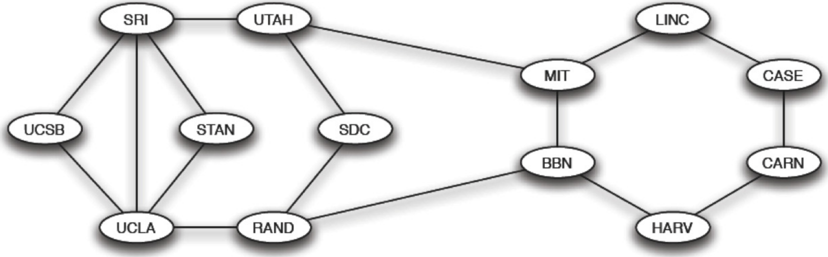


Figure: Internet: Dec. 1970 [E&K, Ch.2]

Adjacency matrix for graph induced by eastern sites in alphabetical order) in 1970 internet graph: another way to represent a graph

$$A(G) = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

- This node induced subgraph is a 6 node regular graph of degree 2. It is a simple graph in that there are no self-loops or multiple edges (two or more edges between the same two nodes).
- Note that the adjacency matrix of an (undirected) simple graph is a symmetric matrix (i.e. $A_{i,j} = A_{j,i}$) with $\{0,1\}$ entries.
- To specify distances, we would need to give weights to the edges to represent the distances.

Directed Graph Example: Kidney Exchange

- Live kidney donation common in North America to get around waiting list problems: donor-recipient pairs are nodes and links are directed.
- Exchange: supports willing pairs who are incompatible
  1. allows multiway-exchange
  2. supported by sophisticated algorithms to find matches
- But what if someone reneges? ⇒ require simultaneous transplantation! Non-cyclic paths can be started by an altruistic donor!

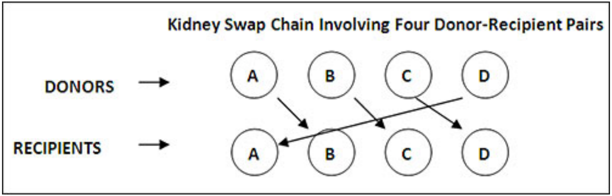


Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

More definitions and terminology

- In order to refer to the nodes and edges of a graph, we define graph $G = (V, E)$, where
  - V is the set of nodes (often called vertices)
  - E is the set of edges (sometimes called links or arcs)

More definitions and terminology

- In order to refer to the nodes and edges of a graph, we define graph $G = (V, E)$, where
    - V is the set of nodes (often called vertices)
    - E is the set of edges (sometimes called links or arcs)

- Undirected graph: an edge (u, v) is an unordered pair of nodes.

More definitions and terminology

- In order to refer to the nodes and edges of a graph, we define graph $G = (V, E)$, where
  - V is the set of nodes (often called vertices)
  - E is the set of edges (sometimes called links or arcs)

- Undirected graph: an edge (u, v) is an unordered pair of nodes.

- Directed graph: a directed edge (u, v) is an ordered pair of nodes $\langle u, v \rangle$.
  - However, we usually know when we have a directed graph and just write (u, v).

Basic definitions continued

- First start with undirected graphs G = (V,E).

Basic definitions continued

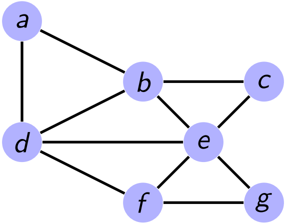
- First start with undirected graphs G = (V,E).
- A path between two nodes, say u and v is a sequence of nodes, say $u_1, u_2, \ldots, u_k$, where for every $1 \leq i \leq k-1$,
  - the pair (u_i, u_{i+1}) is an edge in E,
  - $u = u_1$ and $v = u_k$

Basic definitions continued

- First start with undirected graphs G = (V,E).
- A path between two nodes, say u and v is a sequence of nodes, say $u_1, u_2, \ldots, u_k$, where for every $1 \leq i \leq k-1$,
    - the pair (u_i, u_{i+1}) is an edge in E,
    - $u = u_1$ and $v = u_k$
- The length of a path is the number of edges on that path.

Basic definitions continued

- First start with undirected graphs G = (V,E).
- A path between two nodes, say u and v is a sequence of nodes, say $u_1, u_2, \ldots, u_k$, where for every $1 \leq i \leq k-1$,
    - the pair (u_i, u_{i+1}) is an edge in E,
    - $u = u_1$ and $v = u_k$
- The length of a path is the number of edges on that path.
- A graph is a connected if there is a path between every pair of nodes. For example, the following graph is connected.



Romantic Relationships [Bearman et al, 2004]

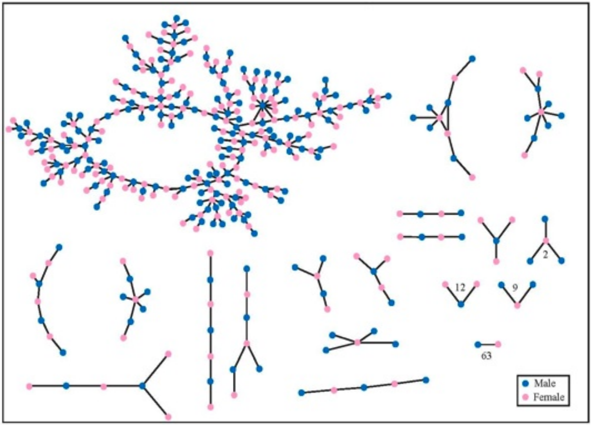
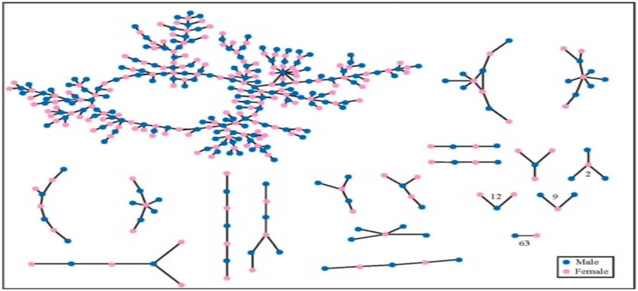


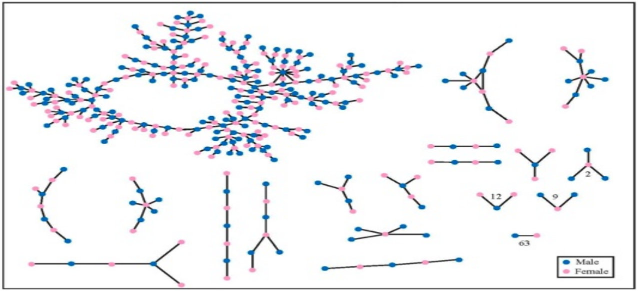
Figure: Dating network in US high school over 18 months.

- Illustrates common structural properties of many networks
- What predictions could you use this for?

More basic definitions



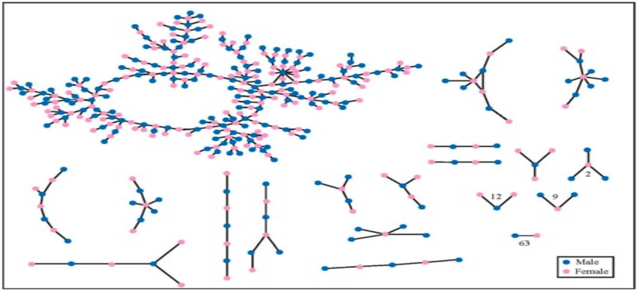
More basic definitions



Observation

Many connected components including one "giant component"

More basic definitions



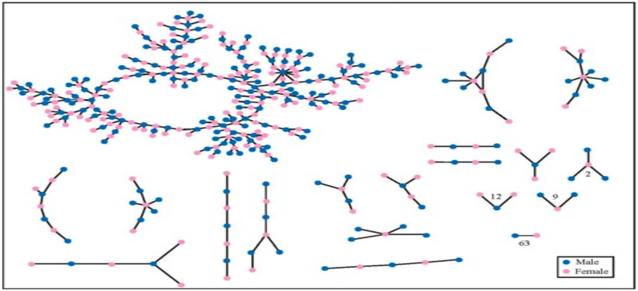
Observation

Many connected components including one "giant component"

- We will use this same graph to illustrate some other basic concepts.
- A cycle is path $u_1, u_2, \ldots, u_k$ such that $u_1 = u_k$; that is, the path starts and ends at the same node.

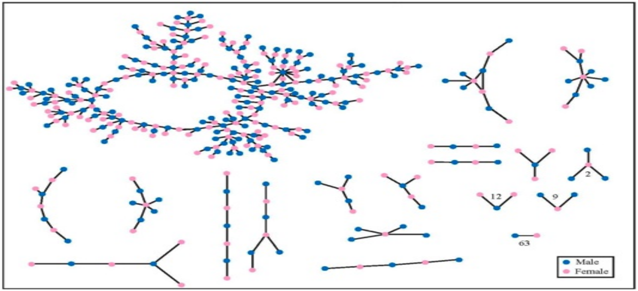
Simple paths and simple cycles

- Usually only consider simple paths and simple cycles: no repeated nodes (other than the start and end nodes in a simple cycle.)



Simple paths and simple cycles

- Usually only consider simple paths and simple cycles: no repeated nodes (other than the start and end nodes in a simple cycle.)



Observation

- There is one big simple cycle and (as far as I can see) three small simple cycles in the "giant component".
- Only one other connected component has a cycle: a triangle having three nodes. Note: this graph is "almost" bipartite and "almost" acyclic.

Example of an acyclic bipartite graph

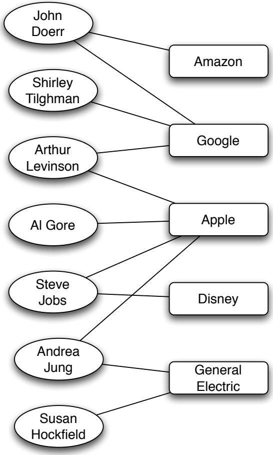


Figure: [E&K, Fig 4.4] One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors. A very small portion of this network (as of mid-2009) is shown here.

Florentine marriages and “centrality”

- Medici connected to more families, but not by much
- More importantly: lie between most pairs of families
  - shortest paths between two families: coordination, communication
  - Medici lie on 52% of all shortest paths; Guadagni 25%; Strozzi 10%

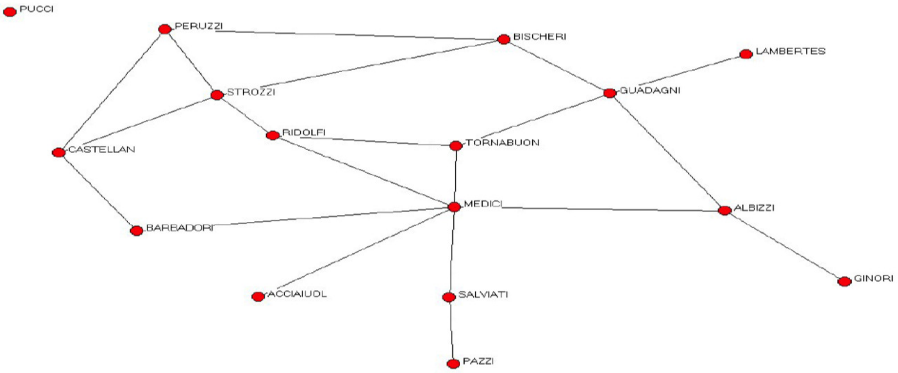


Figure: see [Jackson, Ch 1]

Breadth first search and path lengths [E&K, Fig 2.8]

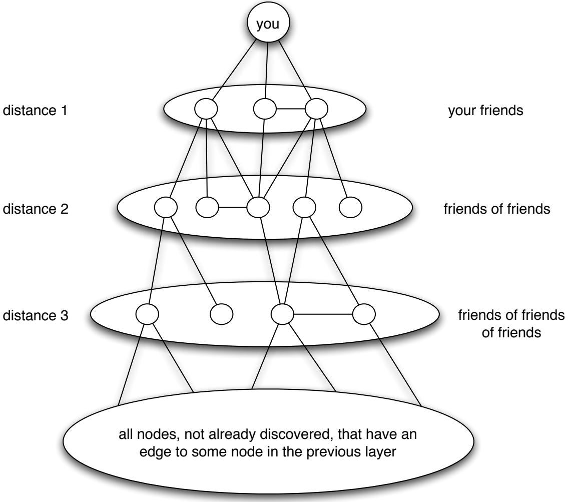

you
distance 1
your friends
distance 2
friends of friends
distance 3
friends of friends of friends
all nodes, not already discovered, that have an edge to some node in the previous layer

Figure: Breadth-first search discovers distances to nodes one layer at a time. Each layer is built of nodes adjacent to at least one node in the previous layer.