Social and Information Networks

University of Toronto CSC303
Winter/Spring 2024

Week 1: Jan 8-12

Welcome! Behind the mask, I am smiling ;)

- Call me Ian :)
- He/Him/His pronouns

This week's high-level learning goals

By the end of the week you'll be able to:

- Recall or locate the course organization
  - People
  - Resources & communication
  - Course structure
  - Pre-Term Survey
  - Key policies
- Define basic graph terms & concepts and apply them to example graphs
- Provide examples of what can graphs represent
- Recognize strengths and weaknesses of network modelling
  - Assess how these may apply to a new problem
- Contrast embeddedness and dispersion in the example application of detecting romantic relationships
  - Define & calculate the embeddedness of an edge
  - Define & calculate the dispersion of an edge
  - Conceptually explain what they measure

Course Organization

Course Instructor: Ian Berlot-Attwell

- Email: 303s24 HYPHEN instr AT cs DOT toronto DOT edu

Teaching Assistants: TBA

Communications

1. Course Web page: source of first resort
   `https://www.cs.toronto.edu/~ianberlot/303s24/`
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: **Piazza** for questions of general interest
   `https://www.cs.toronto.edu/~ianberlot/303s24/`
   - I encourage questions in-class which leads to less confusion especially with regard to technical questions.
   - Questions about extensions to the material, or possible examples are also welcome – they often lead to some of the most interesting discussion!
4. Office hours: regular schedule TBA, also by appointment

Course Materials

Key Resources:

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 mostly follow the textbook, supplemented with papers and some new topics and material
2. Additional materials will be linked on course web page.

Aside: 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.

Lecture/Tutorial/Course Structure

- Course (both lectures & tutorials): in-person
- Midterm: in-person
- Final Exam: in-person

- Tentatively, lectures will be recorded and accessible via Zoom – however, there are no guarantees
- Tentatively, some tutorial sections will be recorded – however, there are no guarantees
  - OCCS lecture & tutorial recordings on Quercus

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - Lectures: Mondays & Wednesdays
  - Tutorials on Fridays; starting Fri Jan 19 in BA2165, BA2175, and BA2195
    - ★ You should be enrolled in a tutorial section on ACORN
    - ★ In general, you are free to attend whichever room you prefer
    - ★ BA2165 will not be recorded
  - **This Fri Jan 12, is a lecture in WB116**
  - Zoom Links for lecture are Quercus
  - When necessary, we will rearrange the schedule – therefore you should be available M,W,F 16:00-17:00 each week whether it is a lecture or a tutorial.

Lecture/Tutorial/Course Structure

Week	Dates	Lecture Topics	Tutorial Topic	Suggested Readings
1	Jan 8-12	Networks, graph concepts	N/A	Ch 1,2
2	Jan 15-19	Strong and weak ties	Mini Lecture	Ch 3
3	Jan 22-26	Homophily and Influence	Mini Lecture	Ch 4
4	Jan 29-Feb 2	Structural balance	Paper Discussion	Ch 5
5	Feb 5-9	Small worlds	Problem Set	Ch 20
6	Feb 12-16	Power laws, Web link analysis	Mini Lecture	Ch 18,14
7	Feb 26-Mar 1	Rumour spread, influence maximization	A1 Solutions	Ch 19
8	Mar 4-8	Influence models, disease spread	Midterm	Ch 19,21
9	Mar 11-15	Mitochondrial Eve, Bargaining power	Problem Set	Ch 21,12
10	Mar 18-22	Stable marriage, Network traffic	Paper Discussion	Ch 8
11	Mar 25-29	Braess' paradox, kidney exchange.	Univ. closed	Ch 8
12	Apr 1-5	Additional topics and course review.	A2 Solutions	

- Core readings are in the table above.
  - Readings often (but not always!) cover all or most of the lecture material
  - Optional but encouraged – some students find them invaluable for deeper understanding
- Additional materials will be posted on the course website at the end of the Course Contents page
- 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 (hopefully) be recorded
  - I will do my best, but make no guarantees as to the quality, consistency, or completeness
  - Depending on how things go, I may have to abandon recordings
  - 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 I hope to make recordings available, I *strongly* suggest attending if you can – as the opportunity to interact with myself, the TAs, and your classmates can be invaluable for your learning
    - ⋆ Using recordings to memorize may be counterproductive, instead try expanding material to new scenarios (Joordens, 2009)

Lecture/Tutorial/Course Structure

- Lectures will (hopefully) be broadcast via Zoom (chat only)
  - This purely a best-efforts offering on my part (as that's outside of the proper scope of the course)
  - I reserve the right to abandon the idea if I find it to be impractical or too time consuming
    - ⋆ Note that it would be only lectures, as I couldn't think of a practical way to make tutorials hybrid
  - Zoom can be used for automatic closed captioning of the lecture
- Laptops welcome, and I encourage students to look at the papers we cover.
  - All I ask is that *if* you work on anything non CSC303 related, that you try to sit towards the back of the lecture hall to avoid distracting others.

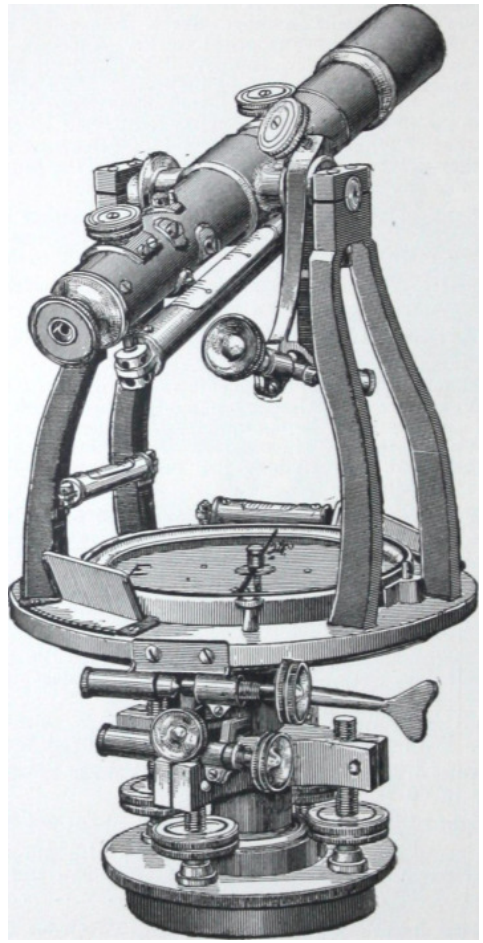
Lecture/Tutorial/Course Structure

- The midterm will be in-tutorial: So make sure you have no ongoing scheduling conflicts!

- Feedback, suggestions, & ideas for improving the course are all welcome!
  - Email
  - Anonymously via `https://forms.gle/vARbnB6vqaNEQz6R6`

Pre-Term Survey (Office Hours, Zoom, Interests)

- Found under the Quercus Quizzes tab
- `https://q.utoronto.ca/courses/337249/quizzes/363998`
- 5 minutes
- Closes Sat Jan 13



Preparation

To learn the material you'll need to be comfortable with basic probability and linear algebra concepts as covered in the prerequisites.

Assignment 0 will help you identify any areas to review.

Probability & linear algebra review material is on the course web page.

`https://www.cs.toronto.edu/~ianberlot/303s24/material.html`

Grading scheme and schedule

Grading Scheme

1. Assignment 0: 3%
   Due: Thu Jan 25
2. Assignment 1: 14%
   Due: Thu Feb 15
3. Assignment 2: 14%
   Due: Thu Mar 21
4. Critical review of a current article (groups 3-4): Worth 14%
   - Draft is due Fri Mar 22, final submission due Thu Apr 4
5. Term Test (closed book with aid sheet, in tutorial rooms): Worth 20%
   Fri Mar 8
6. Final Exam (closed book with aid sheet, TBD): Worth 35%
   Date: During Final Exam Period, date TBA, intended 3 hours duration

Be careful! Thu Feb 15 is sooner than you'd think, and a lot of material is due in the last few weeks.

Submitting work late – this slide is important!

- For A0, A1, A2, and the critical review project, all students have automatic extension (see the syllabus for details & the respective lengths); this means you can submit the assignment late, without penalty
  - These automatic extensions are intended to cover mistakes, illness, unexpected situations, etc...
  - *However*: these automatic extensions are *not* a new deadline, and the expectation is that very few students will need or use the extra time
  - You should aim for the deadline, and we'll start collecting and grading work at the deadline
  - Furthermore, as the additional time *is an extension intended to cover extraordinary circumstances*, additional extensions **will not be provided**
  - *Only* in the case of unforeseeable, unavoidable, and significantly impacting events, will I make accommodations if someone misses both the deadline & extension – and the accommodation will not be to extend the deadline further

Make-up midterm – this slide is important!

- This year, I'm experimenting with allowing any student to take the make-up midterm, for any reason
  - The make-up midterm is exactly one week later during tutorial
    - ★ Those writing the make-up midterm will miss that week's tutorial
  - The policy is meant to cover unexpected & significantly impacting situations, and immovable prior commitments etc...
  - The make-up and the main midterm will be completely different tests
    - ★ If one turned out to be meaningfully harder, then I will adjust the marks on the make-up midterm
  - THIS IS NOT AN EXTRA WEEK TO STUDY!!!!
    - ★ It will force you to miss a tutorial
    - ★ It will take away time from other assignments
    - ★ There is NO makeup makeup midterm!
  - You *cannot* write both tests. I'll discard the makeup midterm test mark if you try it

Make-up midterm – this slide is important!

- Good reasons to write the make-up midterm: unexpected & significantly impacting, or unavoidable
  - You're sick
  - The TTC broke down and you arrived 20 minutes late
  - You forgot your aid sheet
  - Unexpected personal tragedy
  - Religious obligations
  - Unavoidable travel

- Bad reasons to write the make-up midterm: anything else
  - Hoping to game the system
    - ★ Different tests, and any difference in difficulty will be adjusted for
  - Test anxiety (unless there's solid reason why it'll be better in a week)
    - ★ Try contacting accessibility services, now
  - Seeking "perfect" testing conditions
    - ★ Ain't possible, the random noise will average out with time, and seeking perfection isn't worth the price

Policies

1. All requests for remarking must be submitted on Markus within one week of work being graded.
2. Please don't discuss solutions until work has been graded, solutions have been released, or a week has passed since the original deadline.
3. Collaboration and Plagiarism:
   - We encourage discussion of course materials, but any work submitted must be your own!
   - Any material from a source outside the course must be properly cited.
   - You are only allowed to discuss assignment problems in broad strokes.
   - Do not share your solutions!
   - Advice: do not take away written notes from discussions about any work you will be submitting.

Policies

1. 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".
   - It is important to know what you do not know.
   - It is optional, but strongly encouraged to write out *why* you don't know
   - If you have partial ideas then provide them
   - Additional marks may be given for partial ideas, or explanations as to why you can't answer the question, that show understanding of the material.
2. There are many other guidelines and policies; I strongly encourage everyone to read through the course handbook for details, to help you succeed in the course :)

   `https://www.cs.toronto.edu/~ianberlot/303s24/syllabus.pdf`

Recap

- Course organization
  - People
  - Resources & communication
  - Course structure
  - Pre-Term Survey
  - Key policies

What's in a name? Graphs or Networks?

Networks are graphs. Typically, terminology varies:

- graphs have vertices connected by edges
- 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"

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, . . .

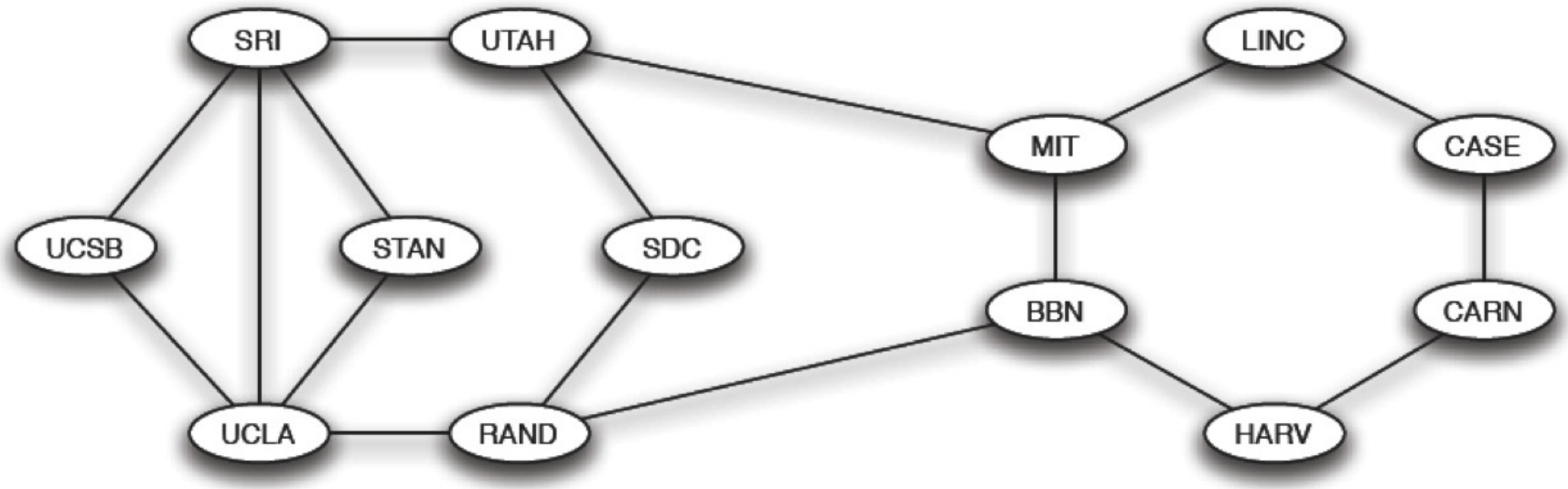
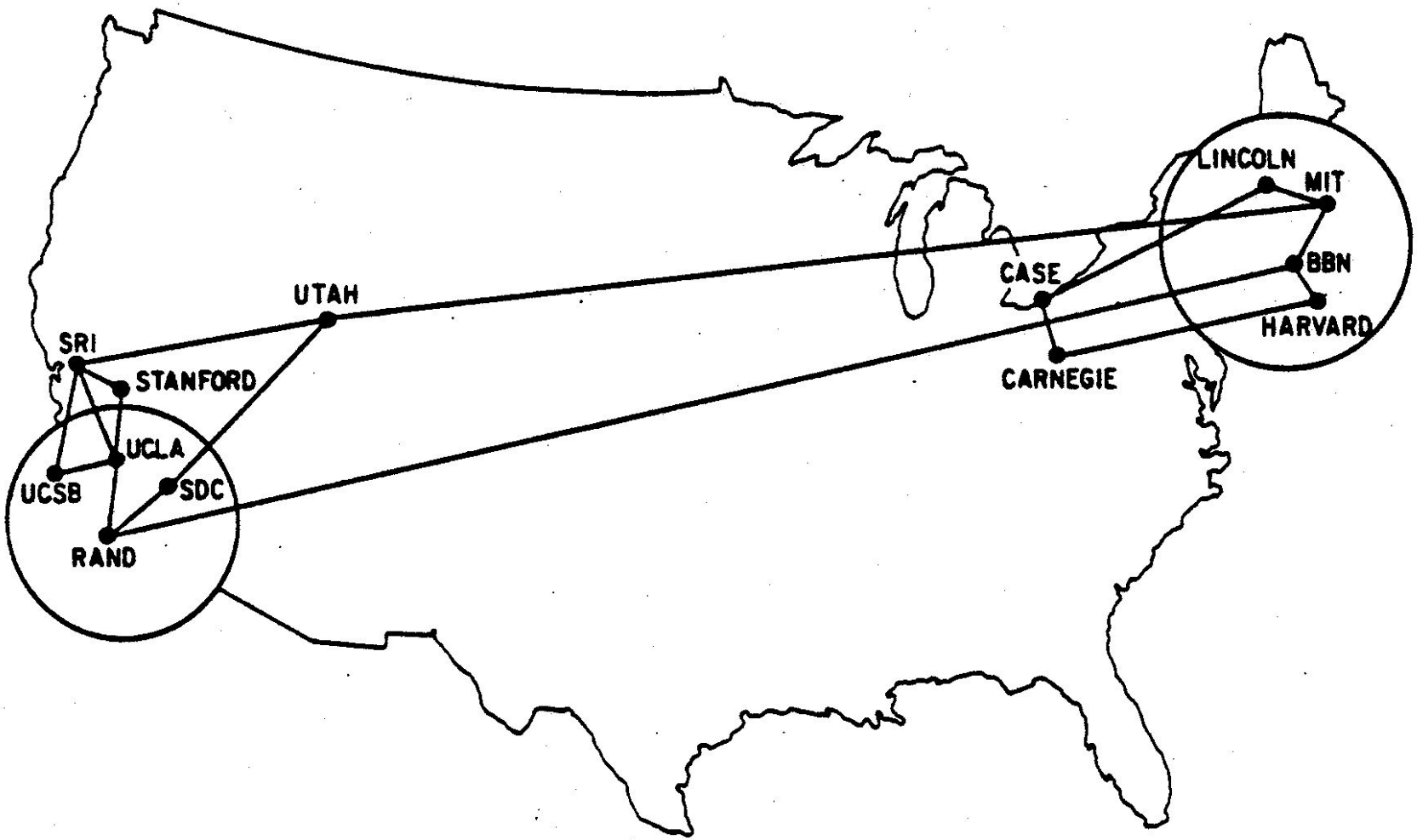


Figure: Initial internet: Dec. 1970 [E&K, Ch.2]

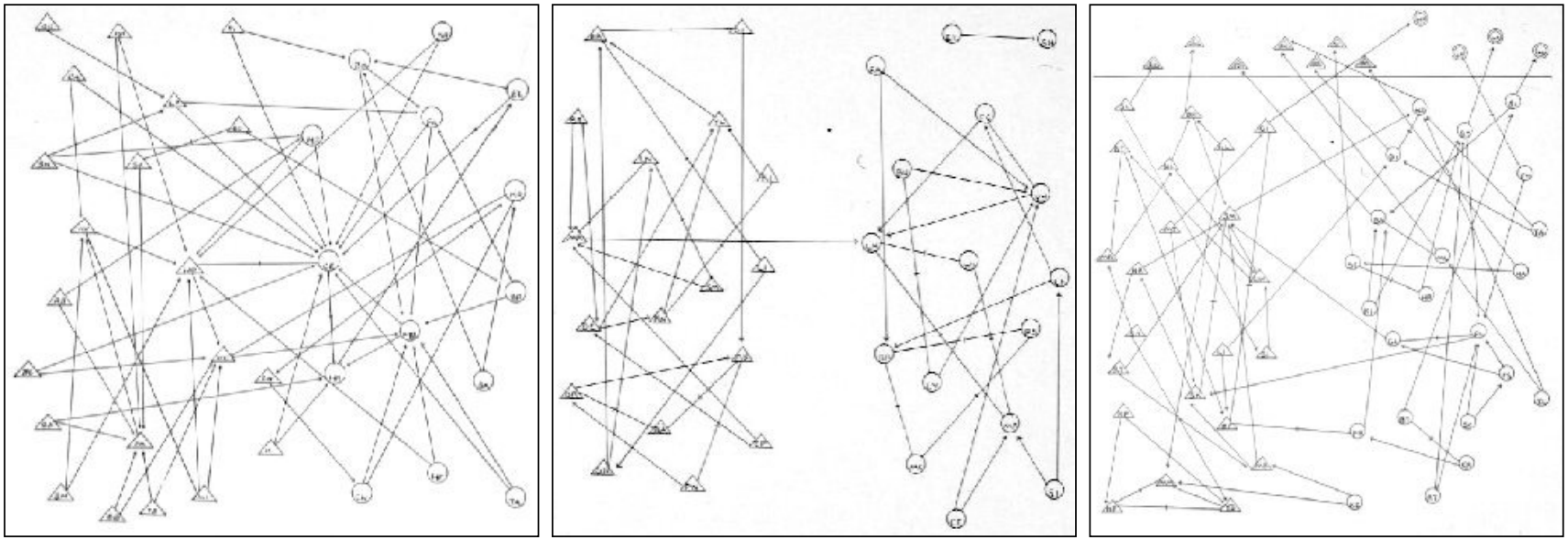
December 1970 internet visualized geographically [Heart et al 1978]



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 he went to various elementary classrooms at a public school and he asked each child 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

Moreno's sociograms, 1924

A closer look at grade 1 in Moreno sociogram

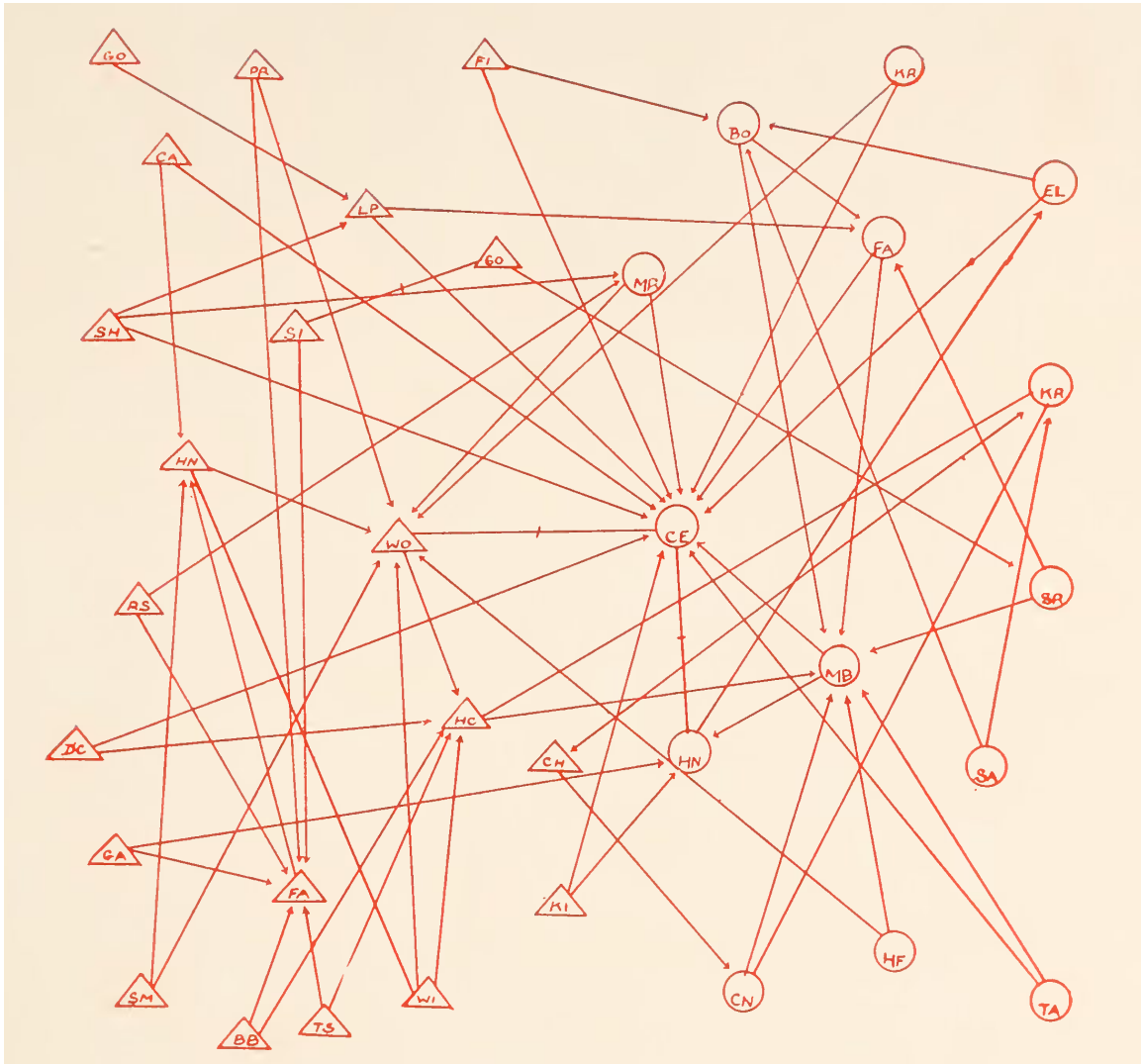


Figure: 21 boys, 14 girls. Directed graph. Most nodes have out-degree 2. 18 are not chosen, thus having in-degree 0. Note also that there are some "stars" with high in-degree.

A closer look at grade 4 in Moreno sociogram

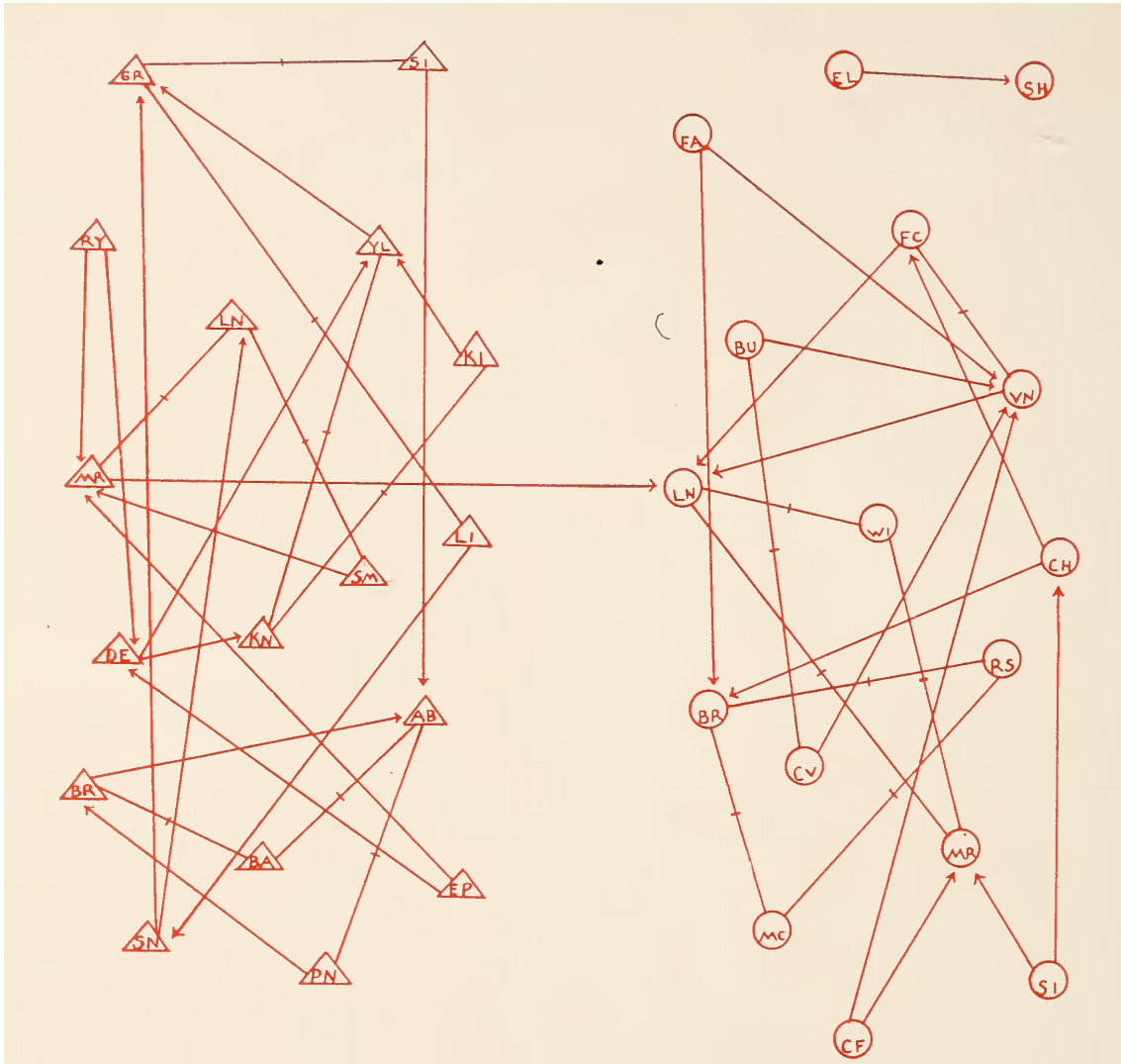


Figure: 17 boys, 16 girls. Directed graph with 6 unchosen having in-degree 0. Moreno depicted his graphs to emphasize inter-gender relations. Note only one edge from a boy to a girl.

A closer look at grade 8 in Moreno sociogram

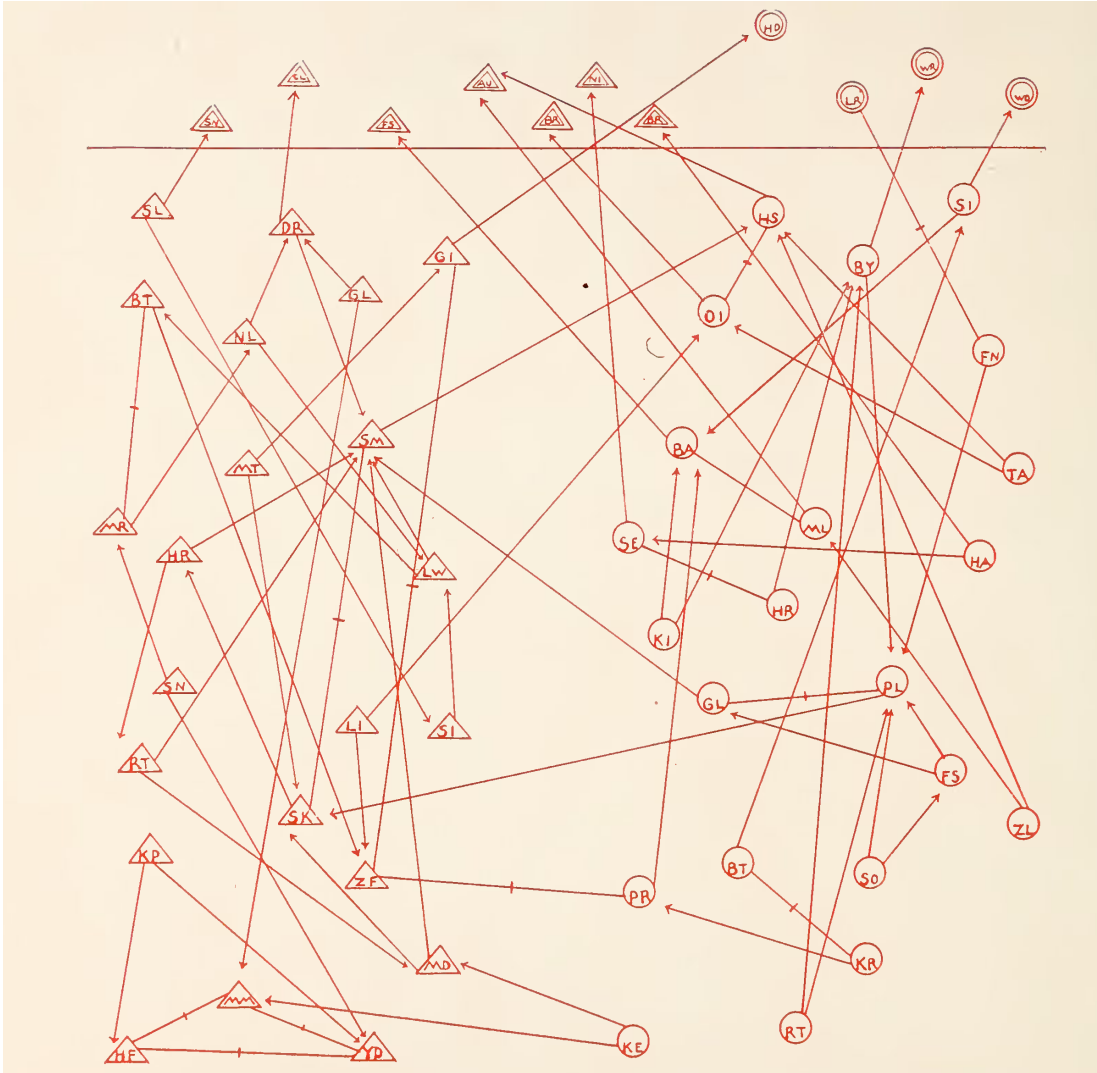


Figure: 22 boys, 22 girls. Directed graph with 12 unchosen having in-degree 0. Some increase in inter-gender relations. Double triangles and circles above line indicate individuals outside of the class.

Romantic Relationships [Bearman et al, 2004]

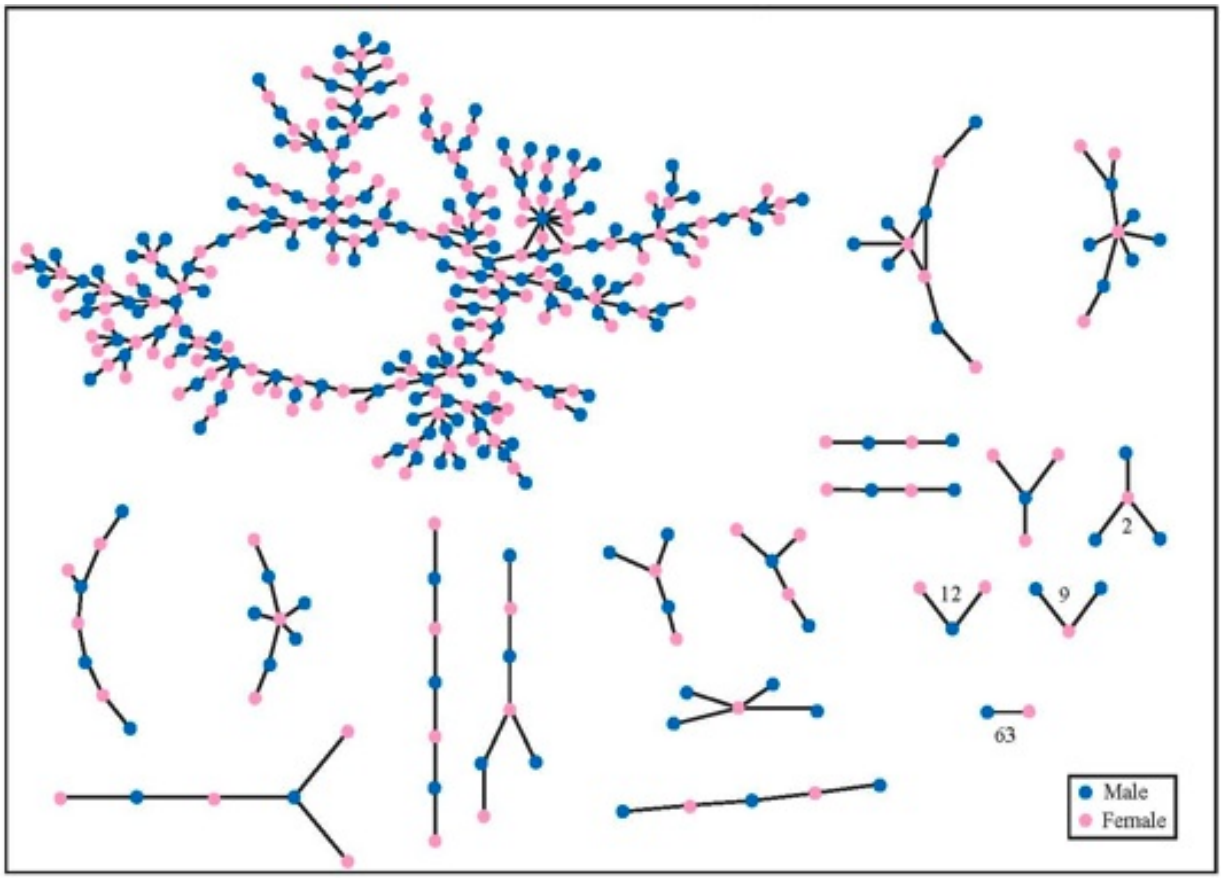
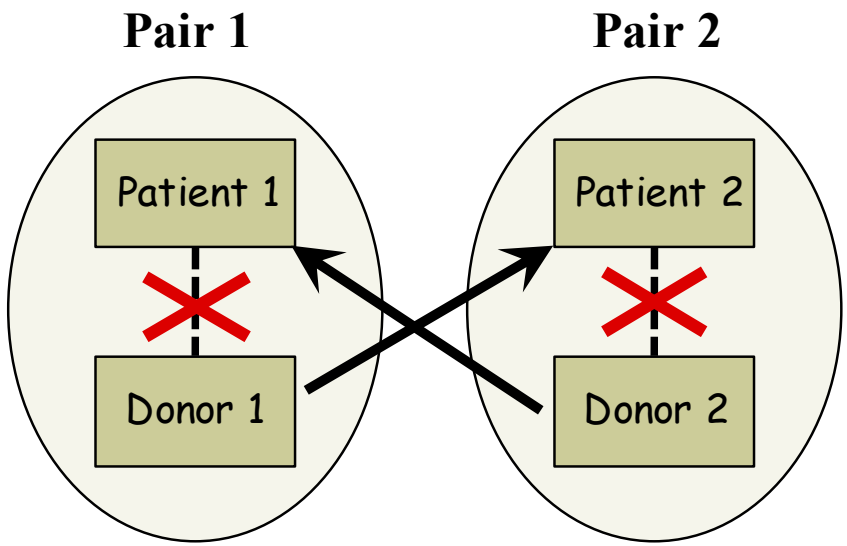


Figure: Dating network in US high school over 18 months.

- 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 get around waiting list problems: requires donor-recipient pairs. What if they are incompatible?



- Exchange: supports willing pairs who are incompatible
  1. allows multiway-exchange
  2. supported by sophisticated algorithms to find matches

Kidney Exchange: Swap Chains

- But what if someone renegs? $\Rightarrow$ Cycles require simultaneous transplantation; Paths require an altruistic donor!

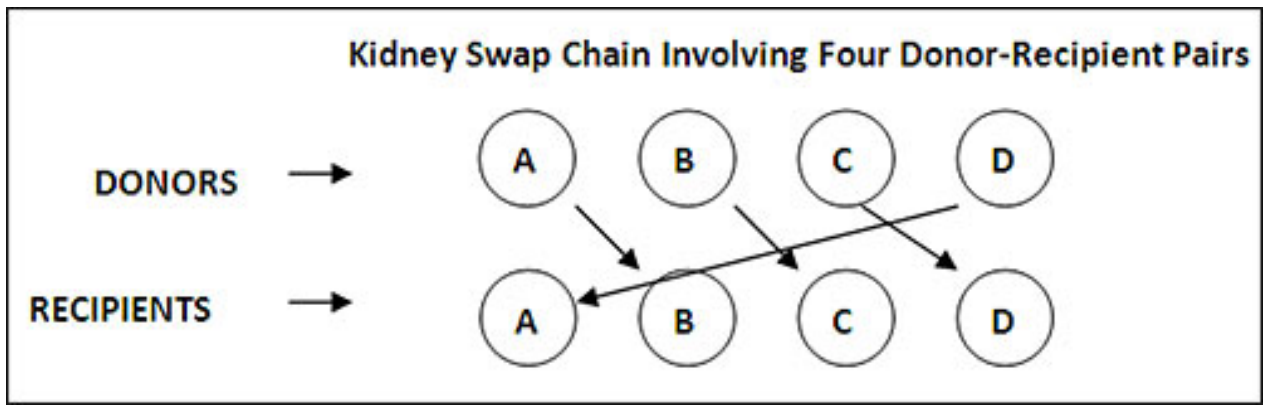
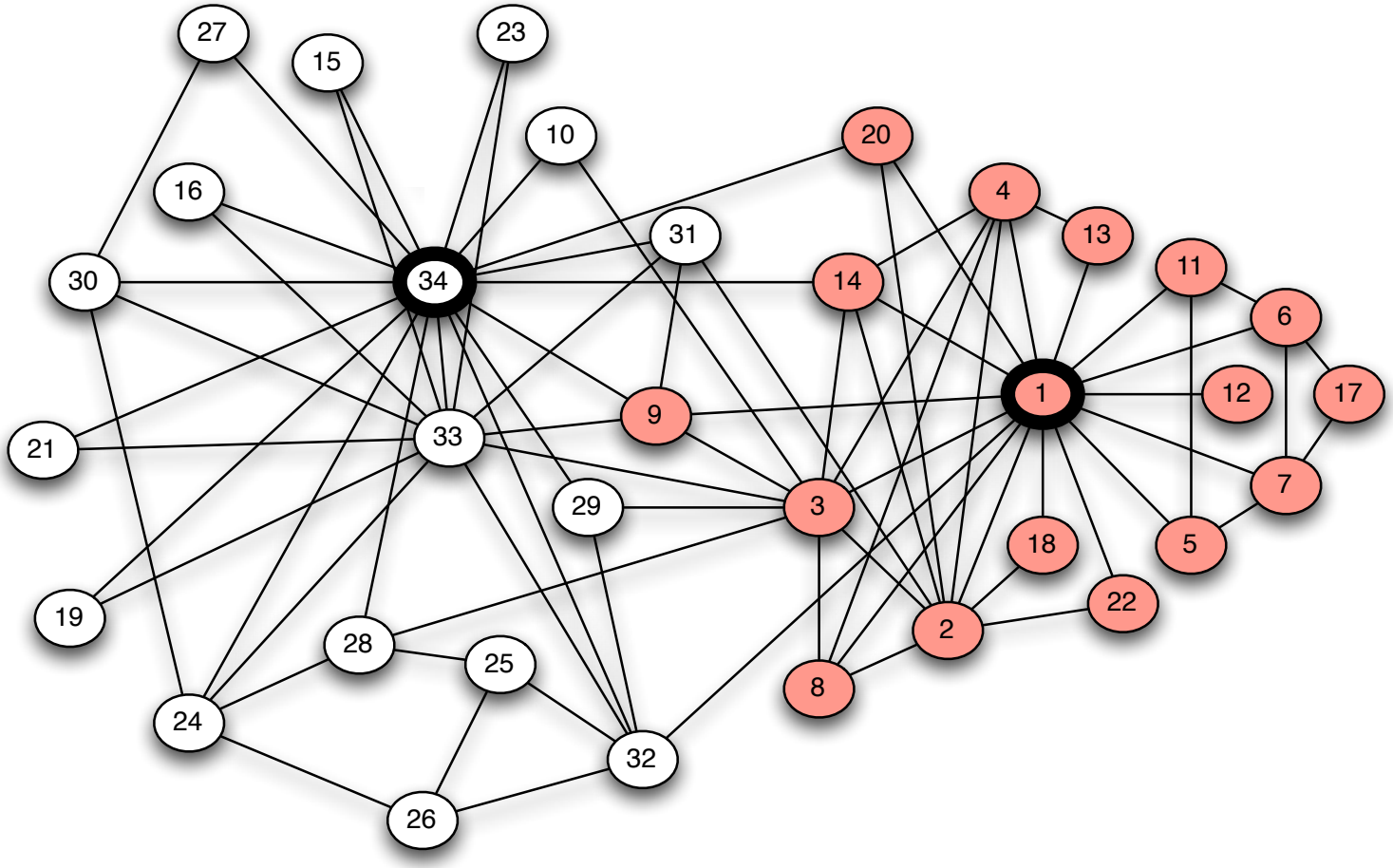


Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

Communities: Karate club division



Karate Club social network, Zachary 1977

Figure: Karate club splis into two clubs

Communities: 2004 Political blogsphere

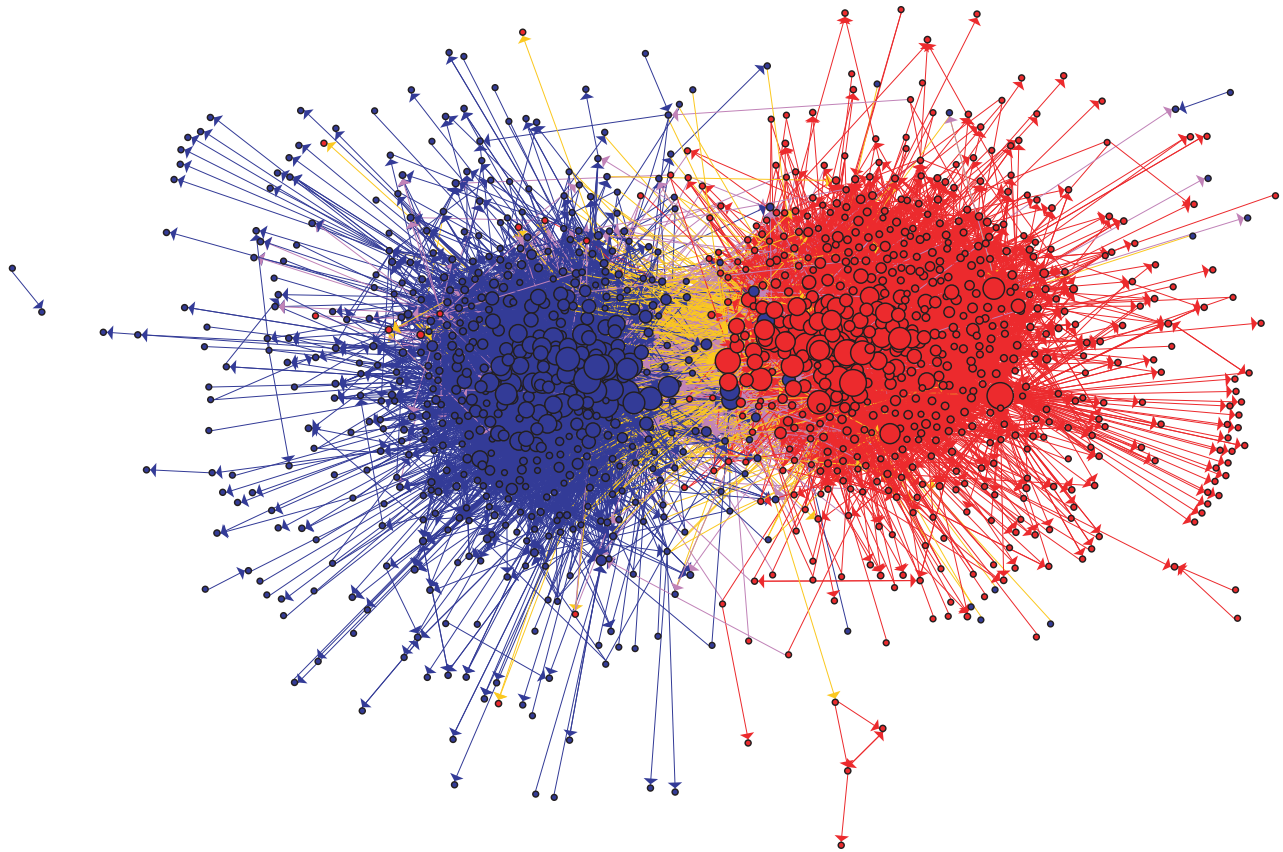


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones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

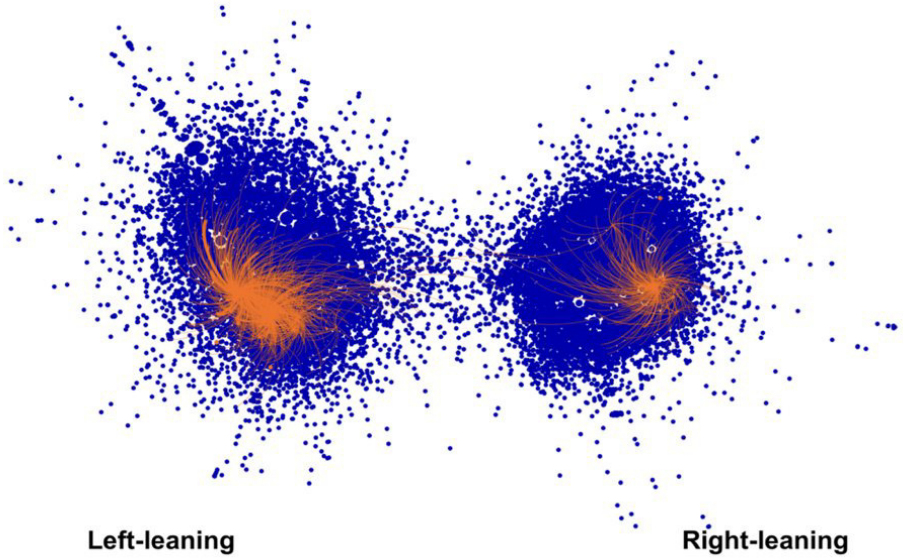


Fig. 1. Retweet Network Graph: RU-IRA Agents in #BlackLivesMatter Discourse. The graph (originally published [3]) shows accounts active in Twitter conversations about #BlackLivesMatter and shooting events in 2016. Each node is an account. Accounts are closer together when one account retweeted another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

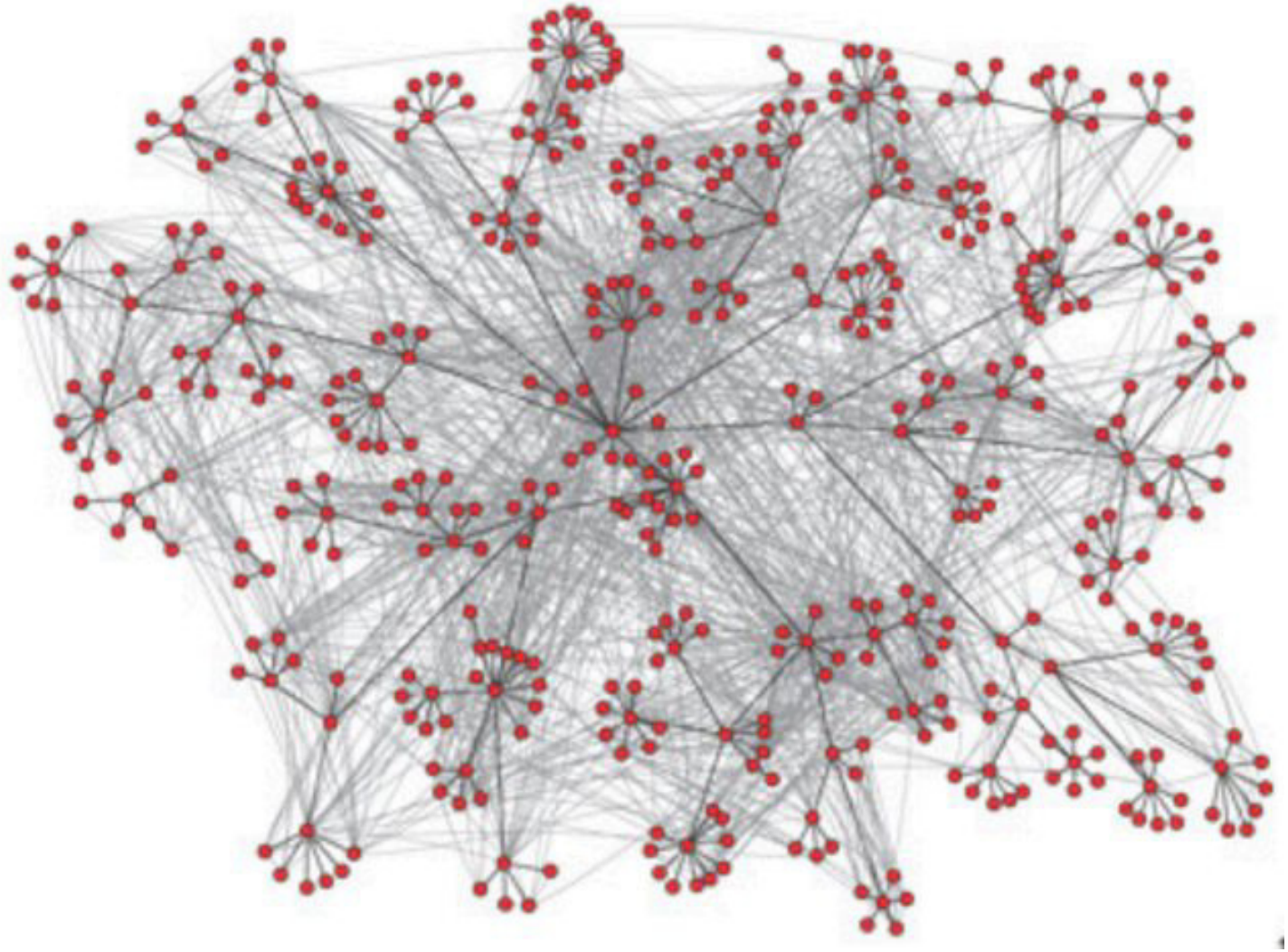
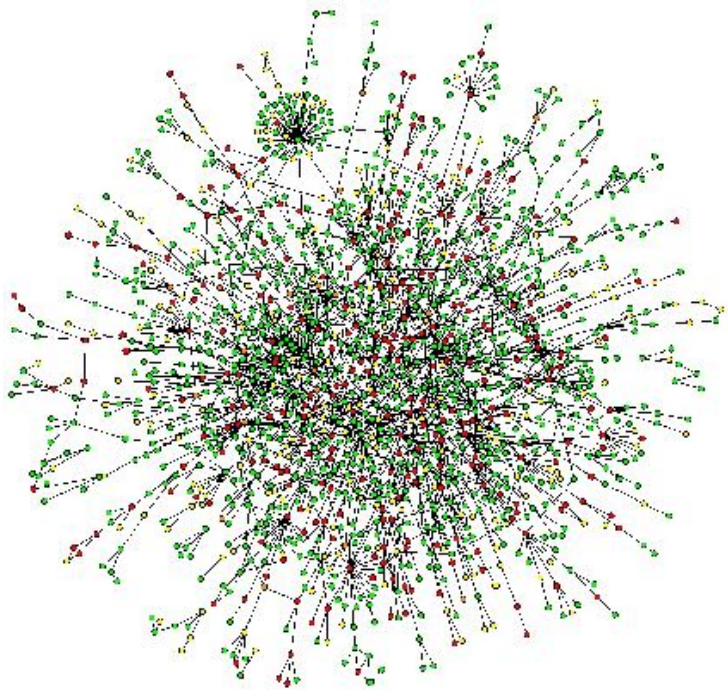


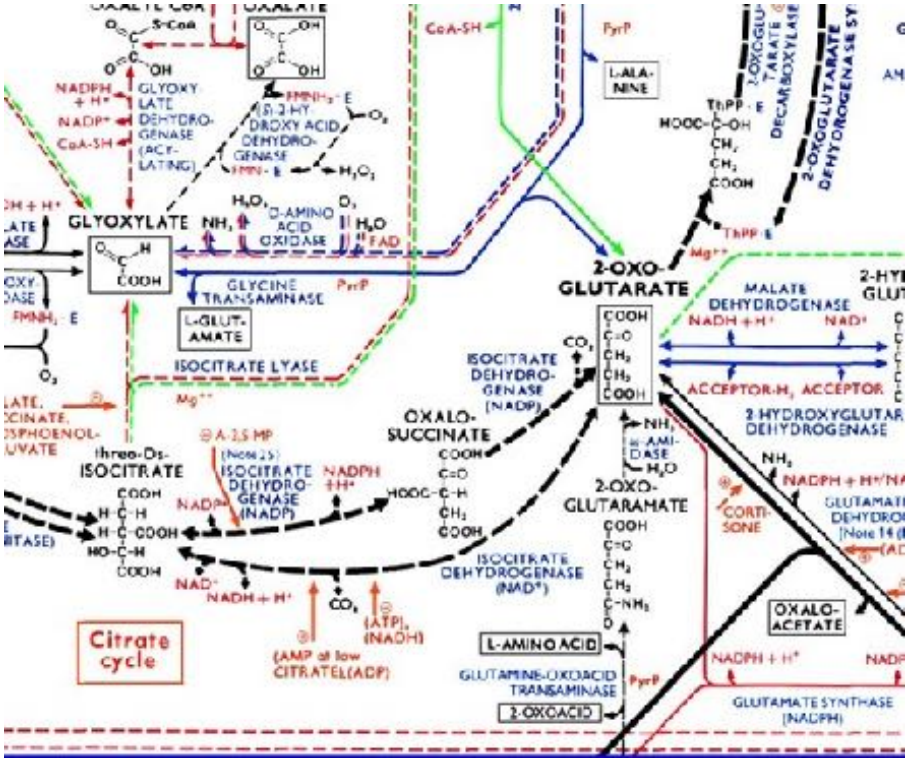
Figure: Email communication among 436 employees of Hewlett Packard Research Lab, superimposed on the organizational hierarchy [Fig 1.2, EK textbook]

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

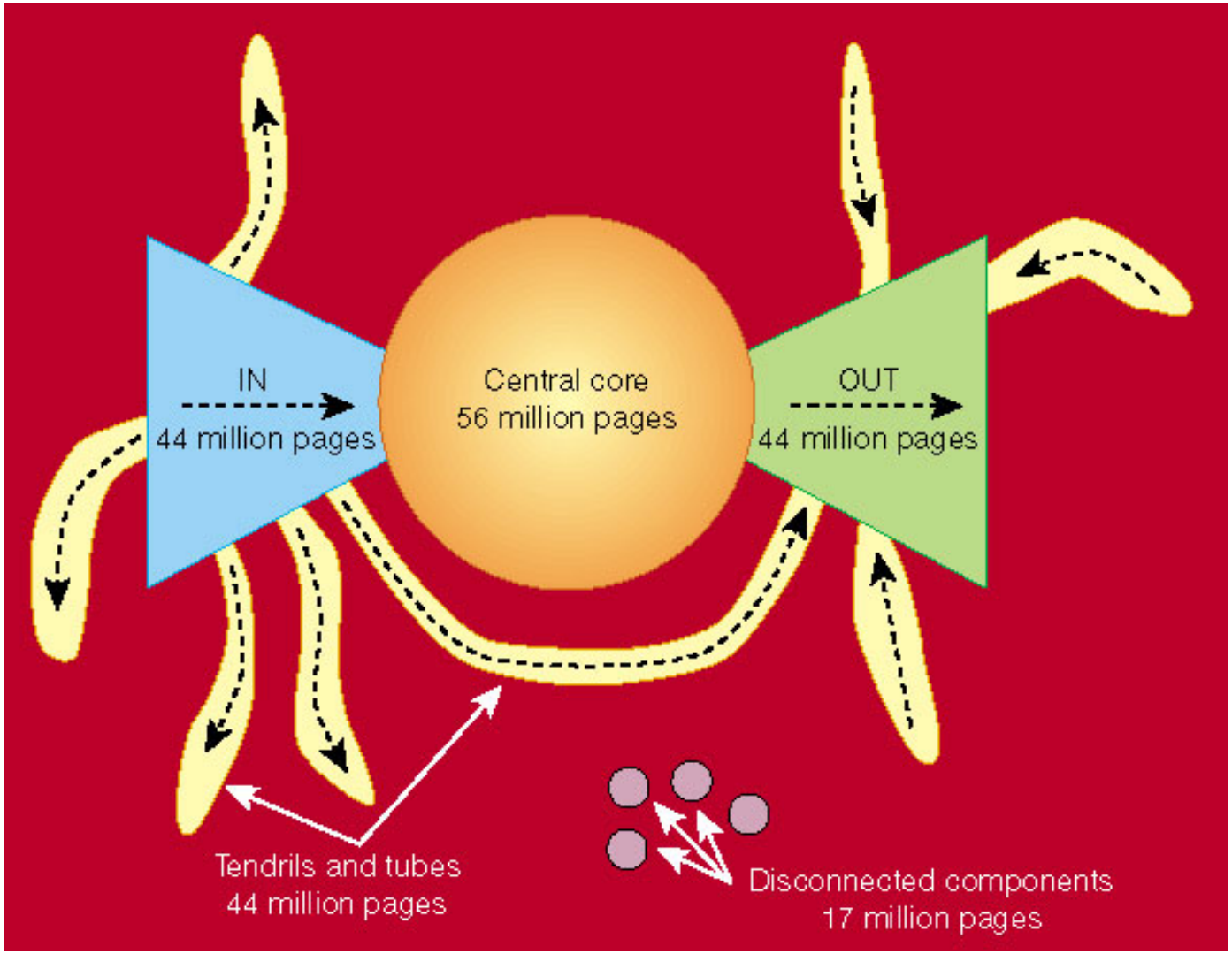


Figure: A schematic picture of the bow tie structure of the 1999 Web [Fig 13.7, EK textbook]

The web as a directed graph of hyperlinks

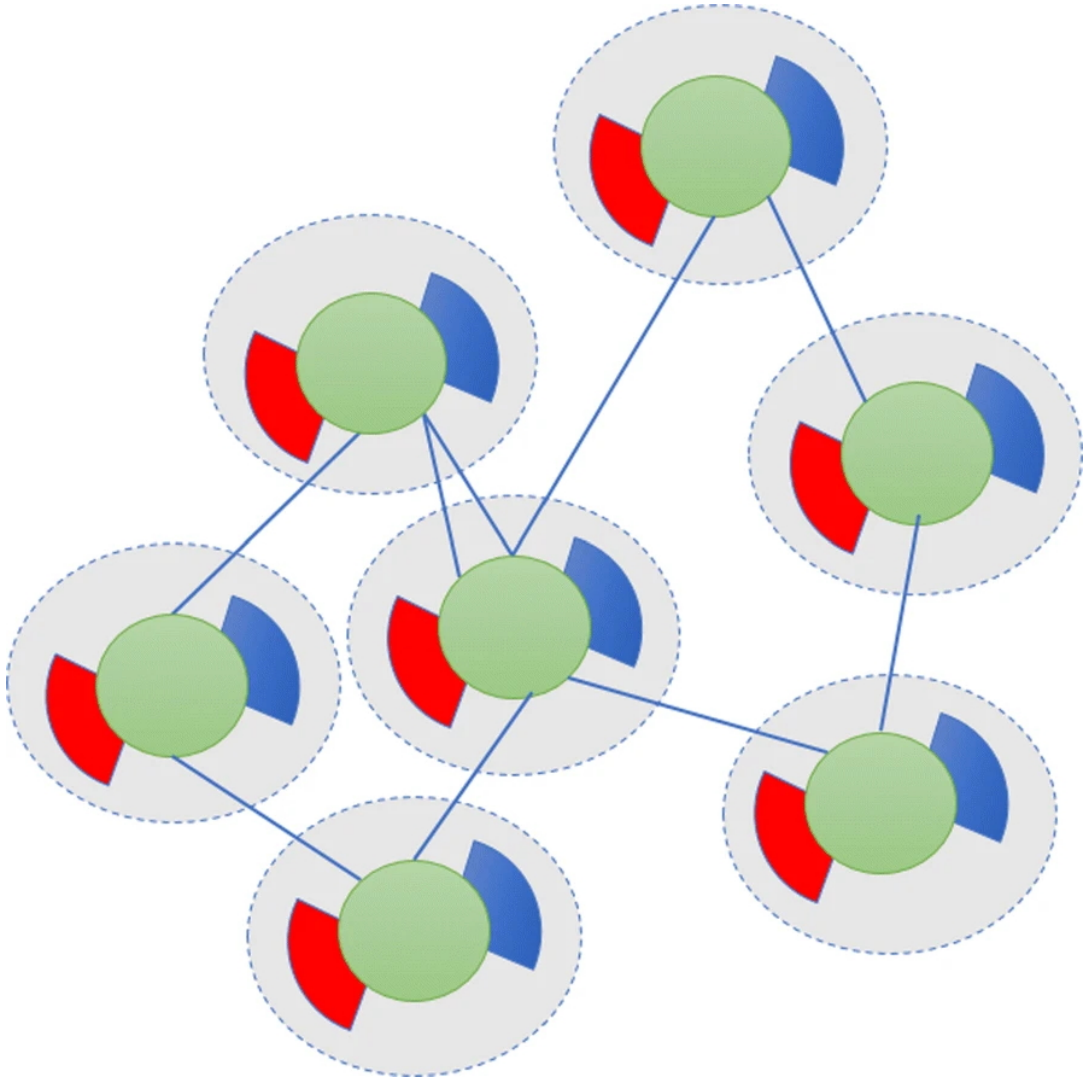


Figure: A schematic picture of the local bow tie structure of the modern web Fujita et al. 2019]

The current interest in networks

- Clearly there are complex systems and networks that we are in contact with daily.
- World population can be thought of as social network of approximately 8 billion people.
  - Circa 2020, Facebook is 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.
  - Importantly, 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* analyse properties of a network, and then 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.

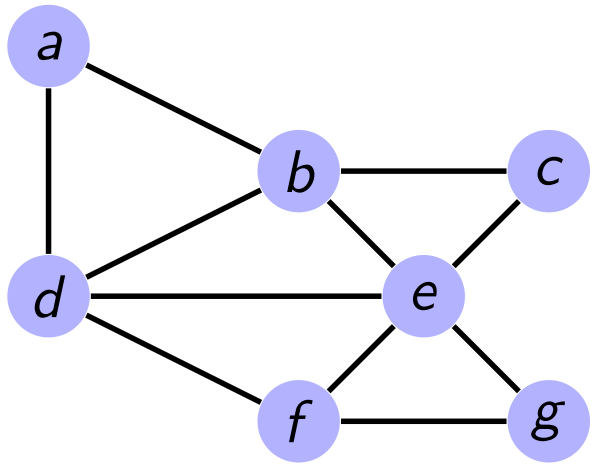
- We use the previous examples and some new ones to illustrate the basic graph concepts and terminology we will be using.

Wed. Jan 10th: Announcements & Corrections

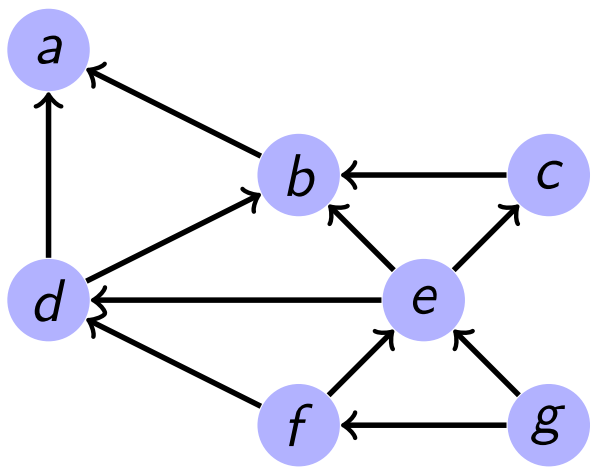
- Assignment 0 has been released!
  - Covers prerequisites, due in about 2 weeks.
  - Thanks for bearing with the double announcement – my announcement script had a hiccup!
- Monday's recording is available on Quercus
  - See the OCCS Student App tab for streaming
- If you're looking to make a CSC303 study group, why not join or lead a registered study group which'll go on you co-curricular record?
  - `https://sidneysmithcommons.artsci.utoronto.ca/recognized-study-groups/`

Graphs: come in two varieties

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



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



Simple Graphs

- A simple graph has:
  1. No self-loops (i.e., an edge from a node to itself).
  2. No multiple edges (several identical edges between the same pair of nodes).
- Unless otherwise stated, we will always be dealing with simple graphs (both directed & undirected).

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, . . .

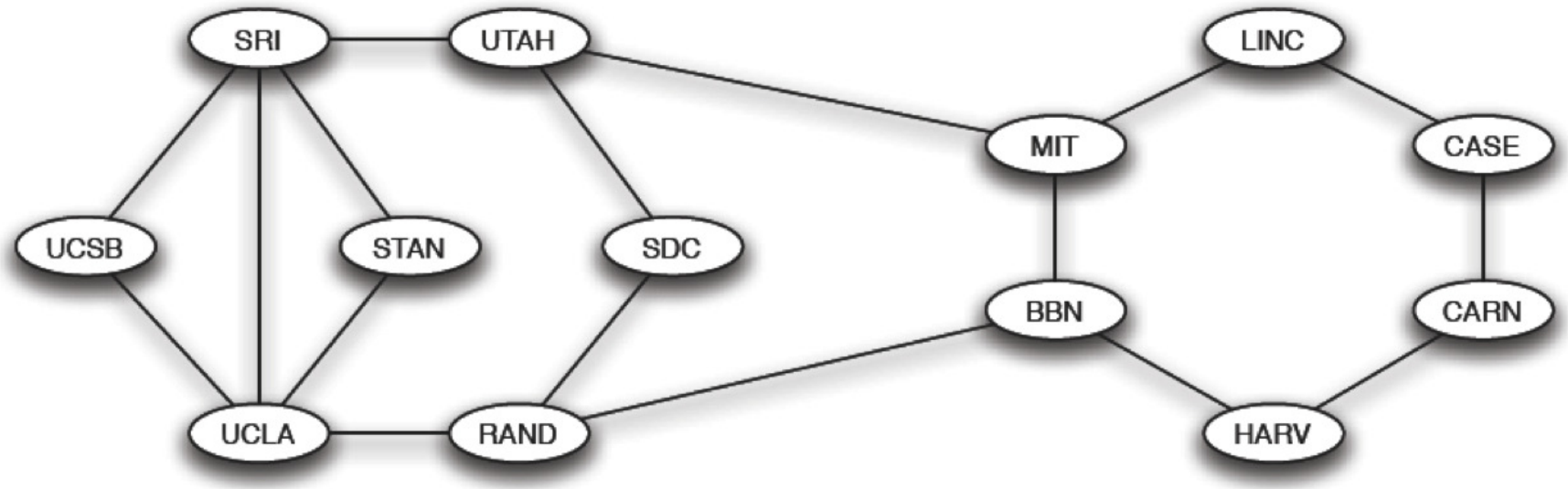


Figure: Internet: Dec. 1970 [E&K, Ch.2]

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 also a simple graph.
- 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.

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

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

- This node induced subgraph is a 6 node regular graph of degree 2. It is also a simple graph.
- 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? $\Rightarrow$ require simultaneous transplantation! Non-cyclic paths can be started by an altruistic donor!

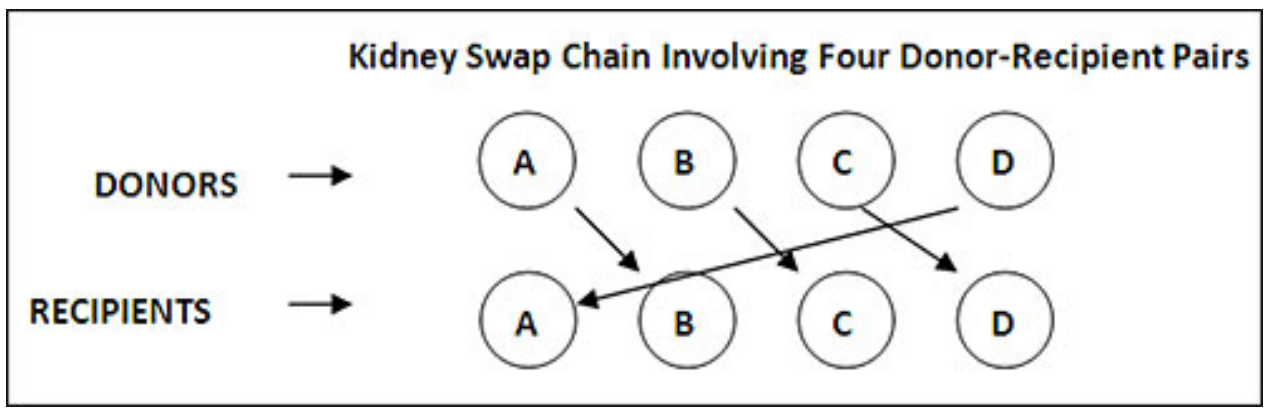


Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

More definitions and terminology

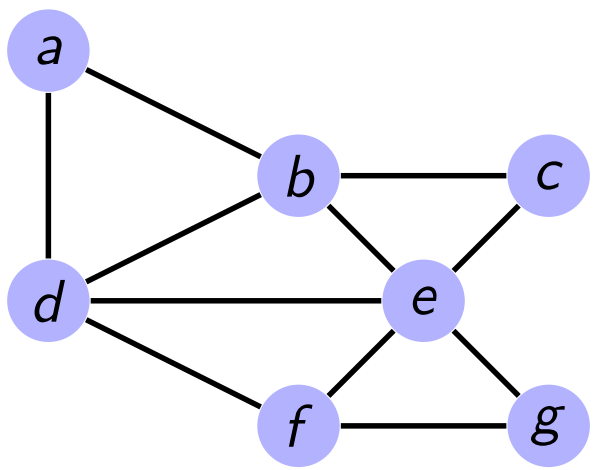
- In order to refer to the nodes and edges of a graph, we define graph $G = (V, E)$, where
  - V is the set of nodes (often called vertices)
  - E is the set of edges (sometimes called links or arcs)

- 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).

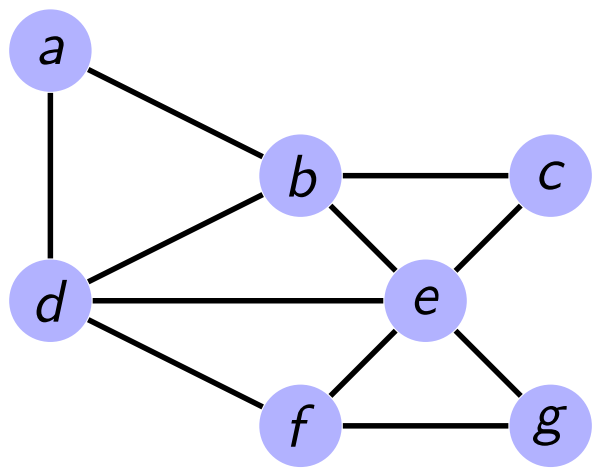
Basic definitions continued

- First start with undirected graphs G = (V,E).
- The degree of a node is the number of edges connected to it.
- 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 or subgraph is a connected if there is a path between every pair of nodes. For example, the following graph is connected.
- A connected component is a connected subgraph, that isn't contained by a larger connected subgraph



Basic definitions continued

- A subset of vertices $V' \subseteq V$ is a clique if there is an edge between every pair of nodes
  - i.e., $\forall u, v \in V' : u \neq v \implies (u, v) \in E$
  - Note that, unlike a connected component, a clique need not be maximal
    - ★ Question: Is a single node a clique?



Romantic Relationships [Bearman et al, 2004]

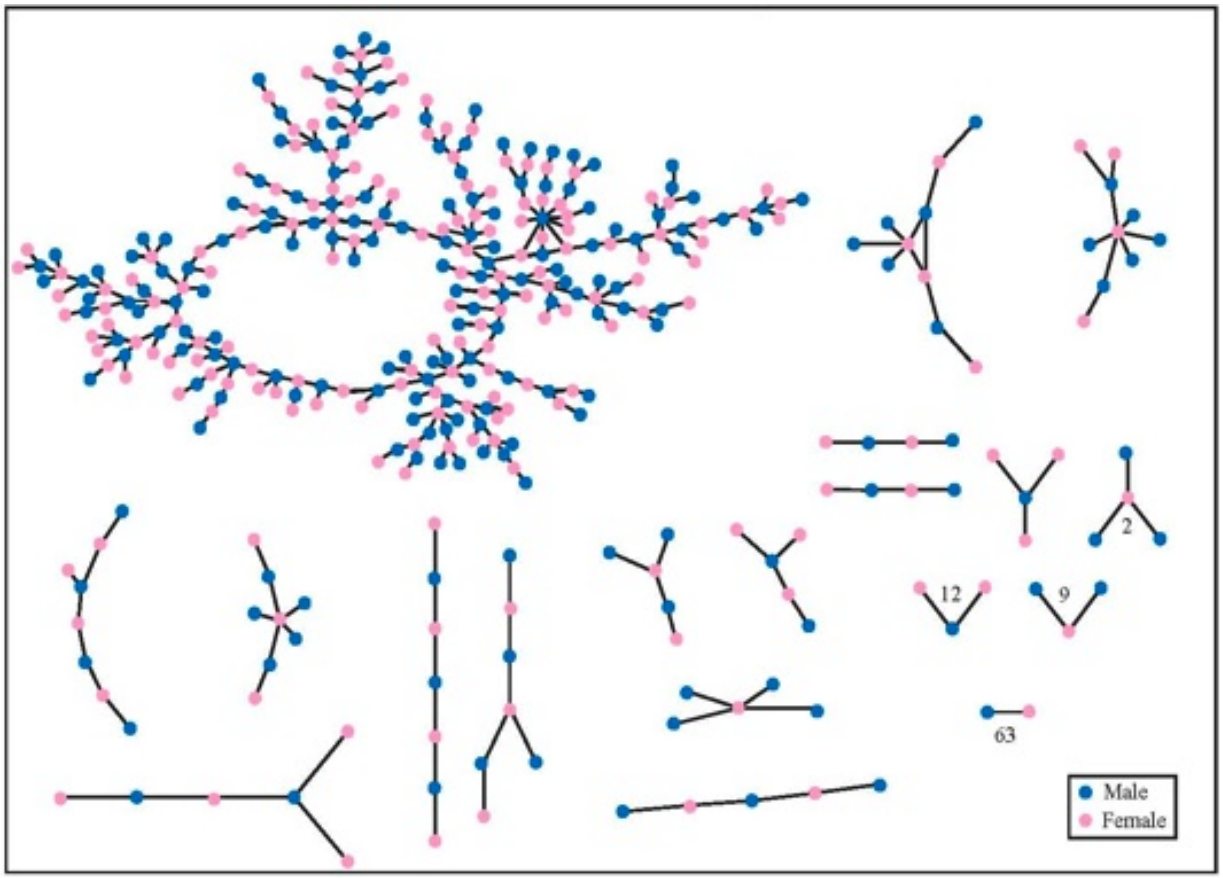
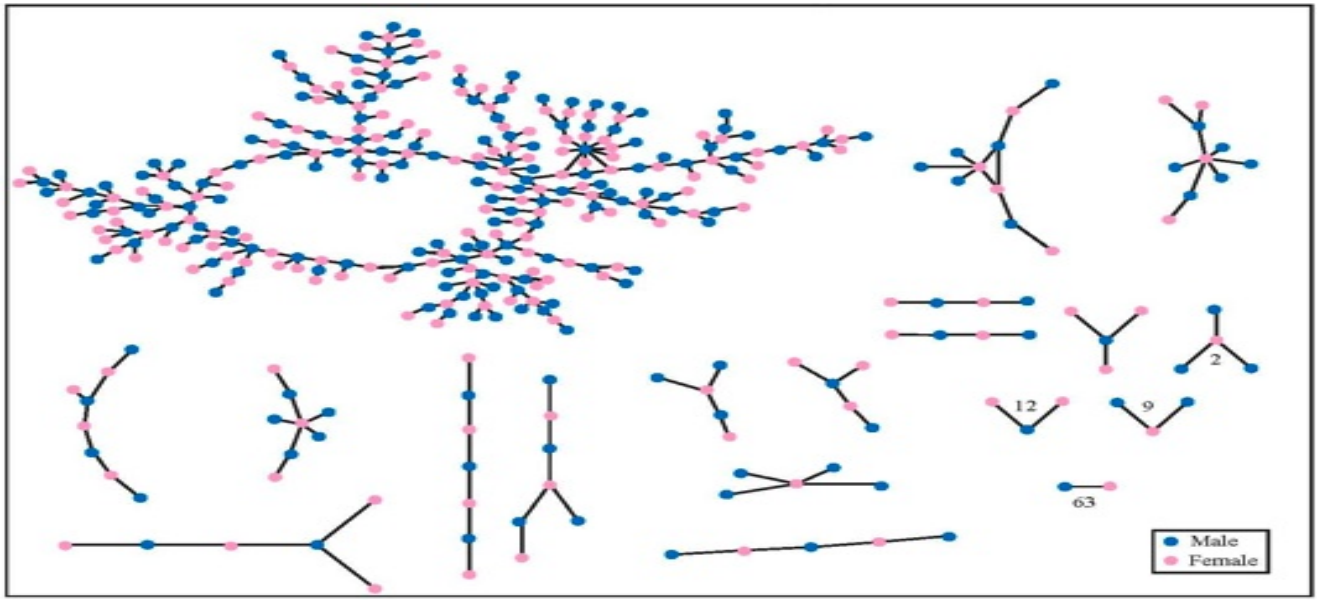


Figure: Dating network in US high school over 18 months.

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

More basic definitions



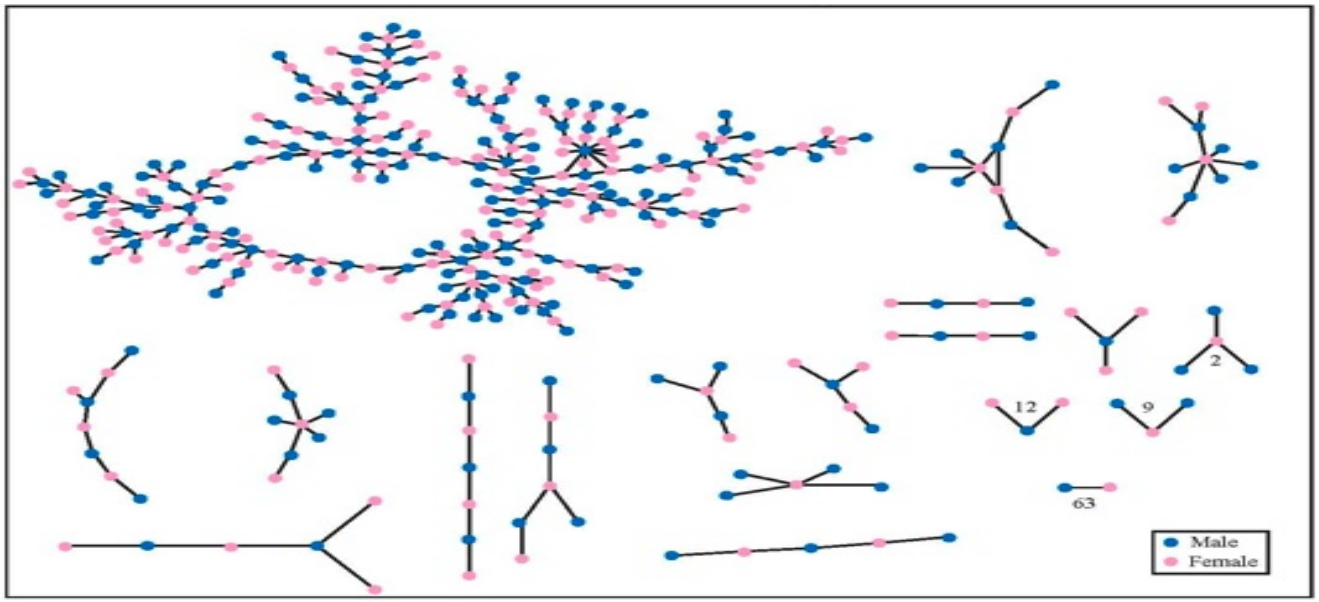
Observation

Many connected components including one "giant component"

Definition (Giant Component)

A connected component that contains a "significant" fraction of all the nodes

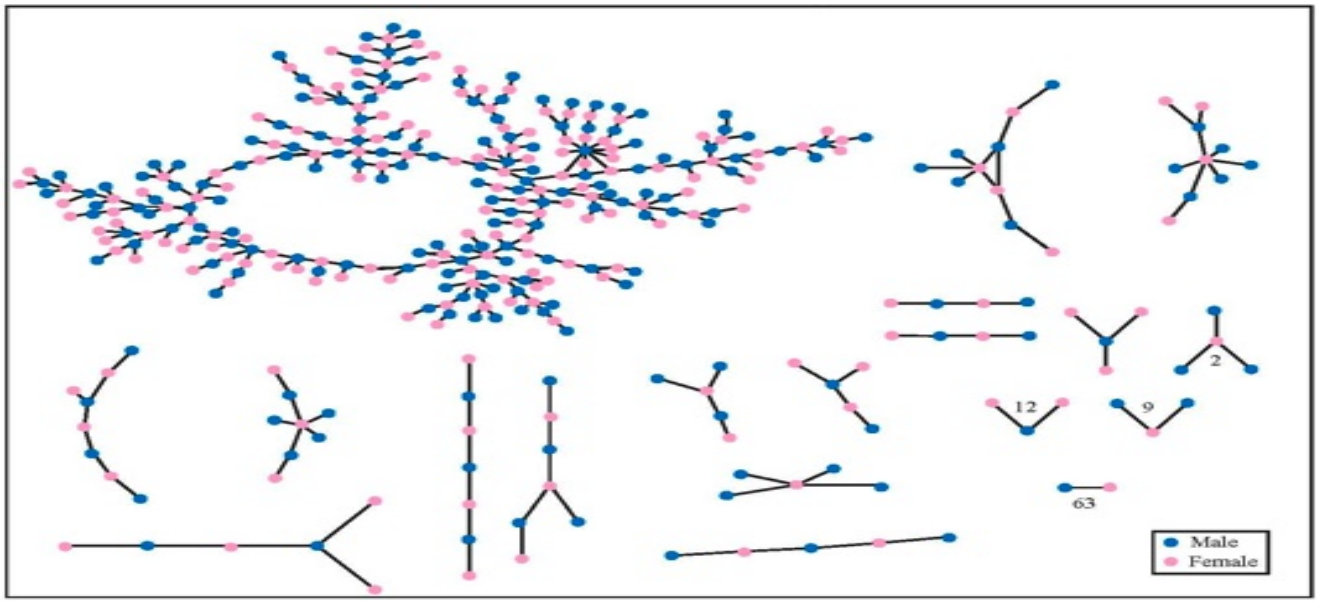
More basic definitions



- 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.)

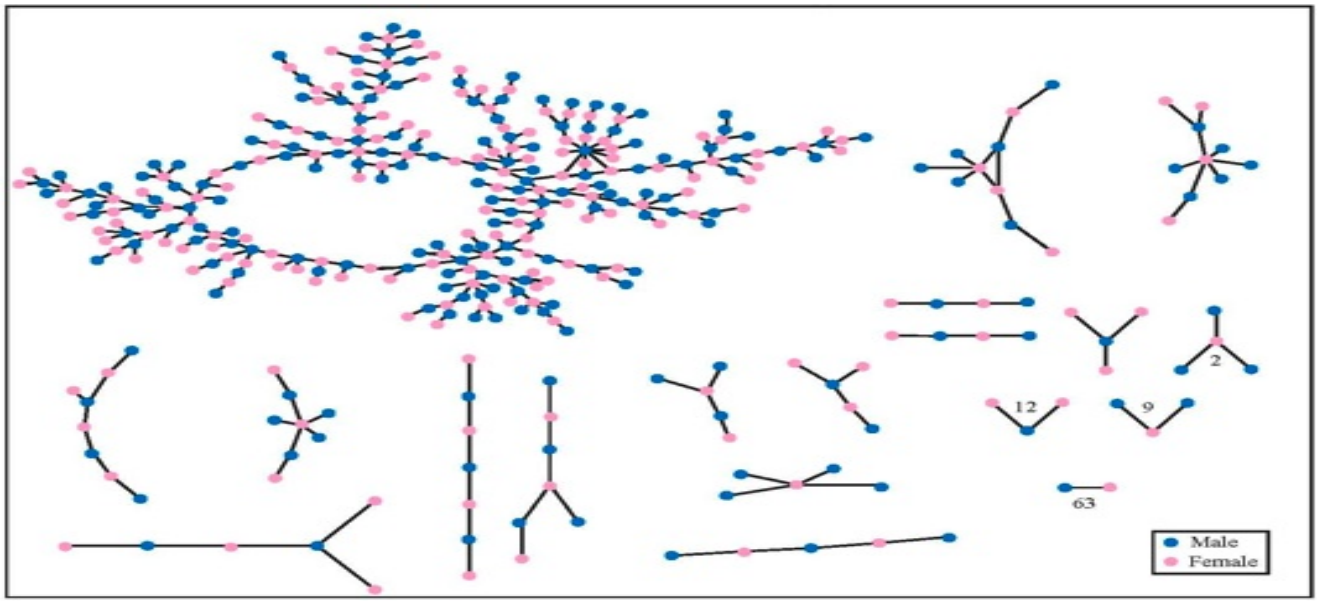


Observation

- 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" acyclic.

Bipartite graphs

- This graph is "almost" bipartite



- An undirected graph is bipartite if the nodes can be partitioned into two groups, such that all edges are between the groups
  - i.e., $\exists X, Y \subseteq V : X \cup Y = V, X \cap Y = \varnothing, \forall x, x' \in X : (x, x') \notin E, \forall y, y' \in Y : (y, y') \notin E$
  - Homework: Convince yourself that a graph is bipartite iff it contains no cycles of odd length

Example of an acyclic bipartite graph

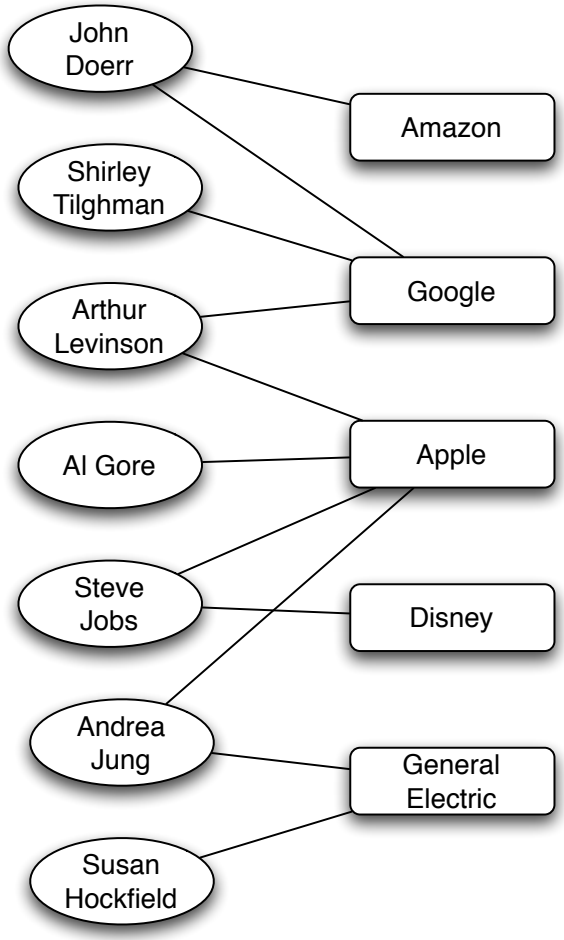


Figure: [E&K, Fig 4.4] One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors. A very small portion of this network (as of mid-2009) is shown here.

Florentine marriages and "centrality"

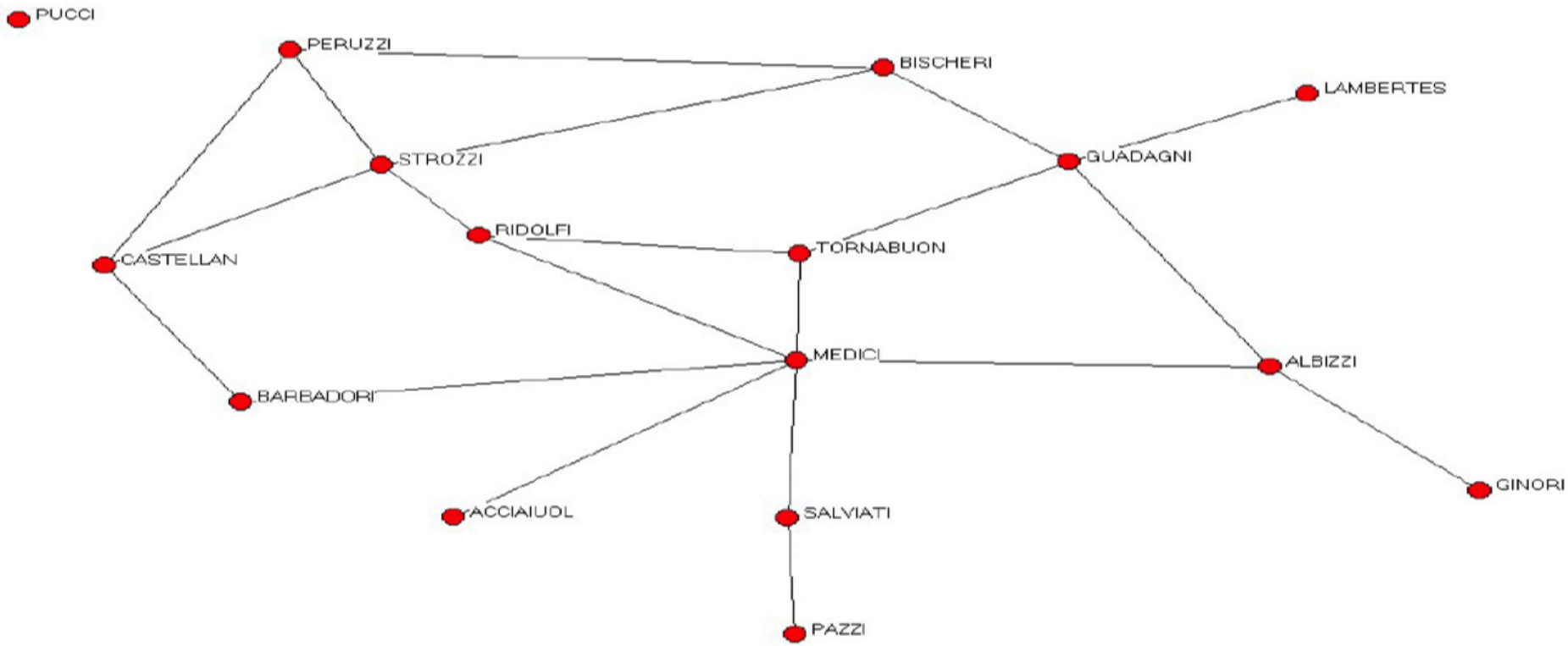


Figure: see [Jackson, Ch 1]

- 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%

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

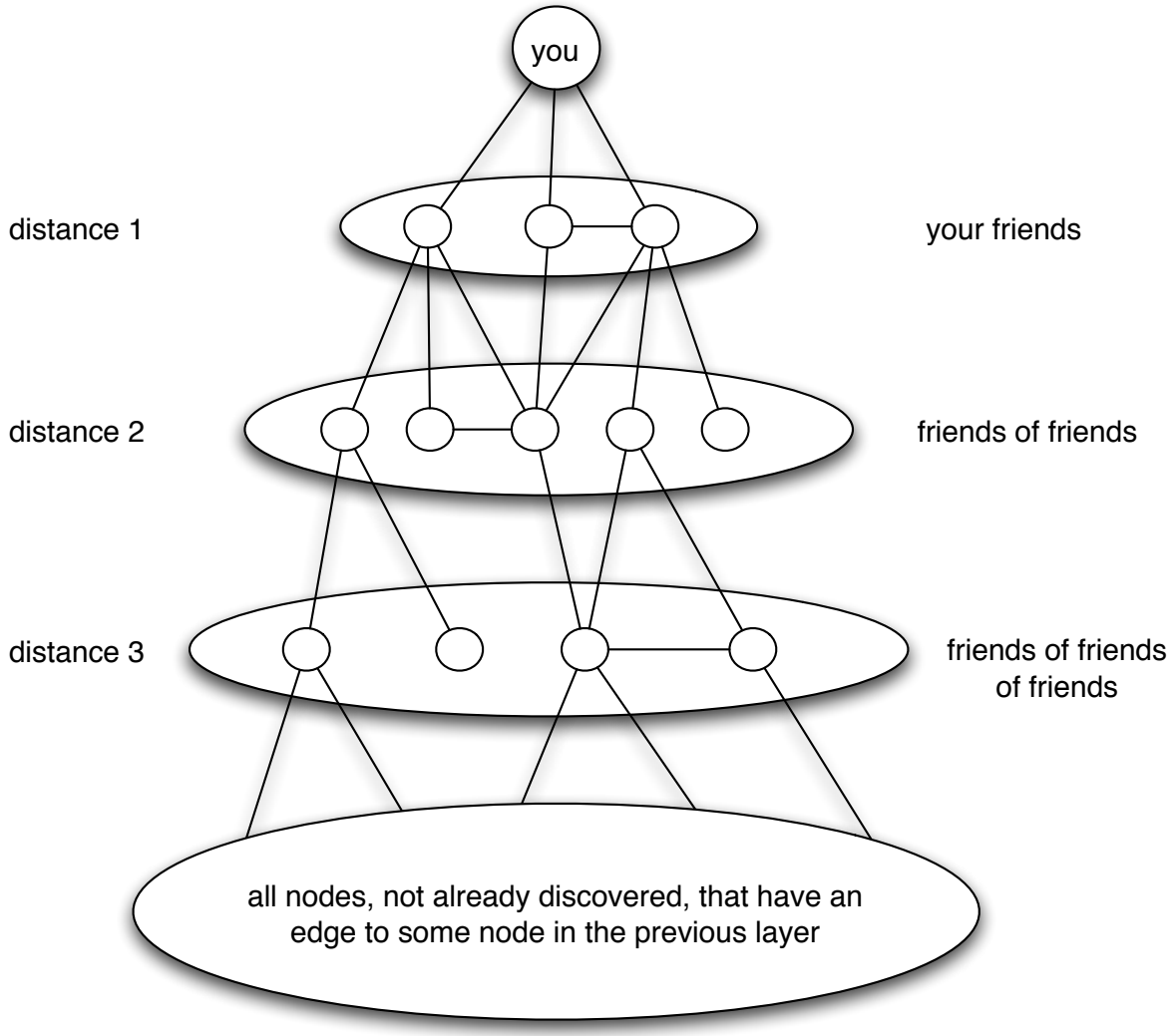


Figure: Breadth-first search discovers distances to nodes one "layer" at a time. Each layer is built of nodes adjacent to at least one node in the previous layer.

The Small World Phenomena

The small world phenomena suggests that in a connected social network any two individuals are likely to be connected (i.e. know each other indirectly) by a short path.

Later in the course we will study 1967 Milgram's small world experiment where he asked random people in Omaha Nebraska to forward a letter to a specified individual in a suburb of Boston which became the origin of the idea of six degrees of separation.

Small Collaboration Worlds

- For now, let us just consider collaboration networks
- For mathematicians (or more generally say scientists) we can form the collaboration network in which an edge represents co-authorship on a published paper
  - For mathematicians, we considers their Erdos number: the length of the shortest path to Paul Erdos.
- For actors, the edges represents actors performing in the same movie.
  - For actors, a popular notion is ones Bacon number, the shortest path to Kevin Bacon.

Erdos collaboration graph drawn by Ron Graham [http:/www.oakland.edu/enp/cgraph.jpg]

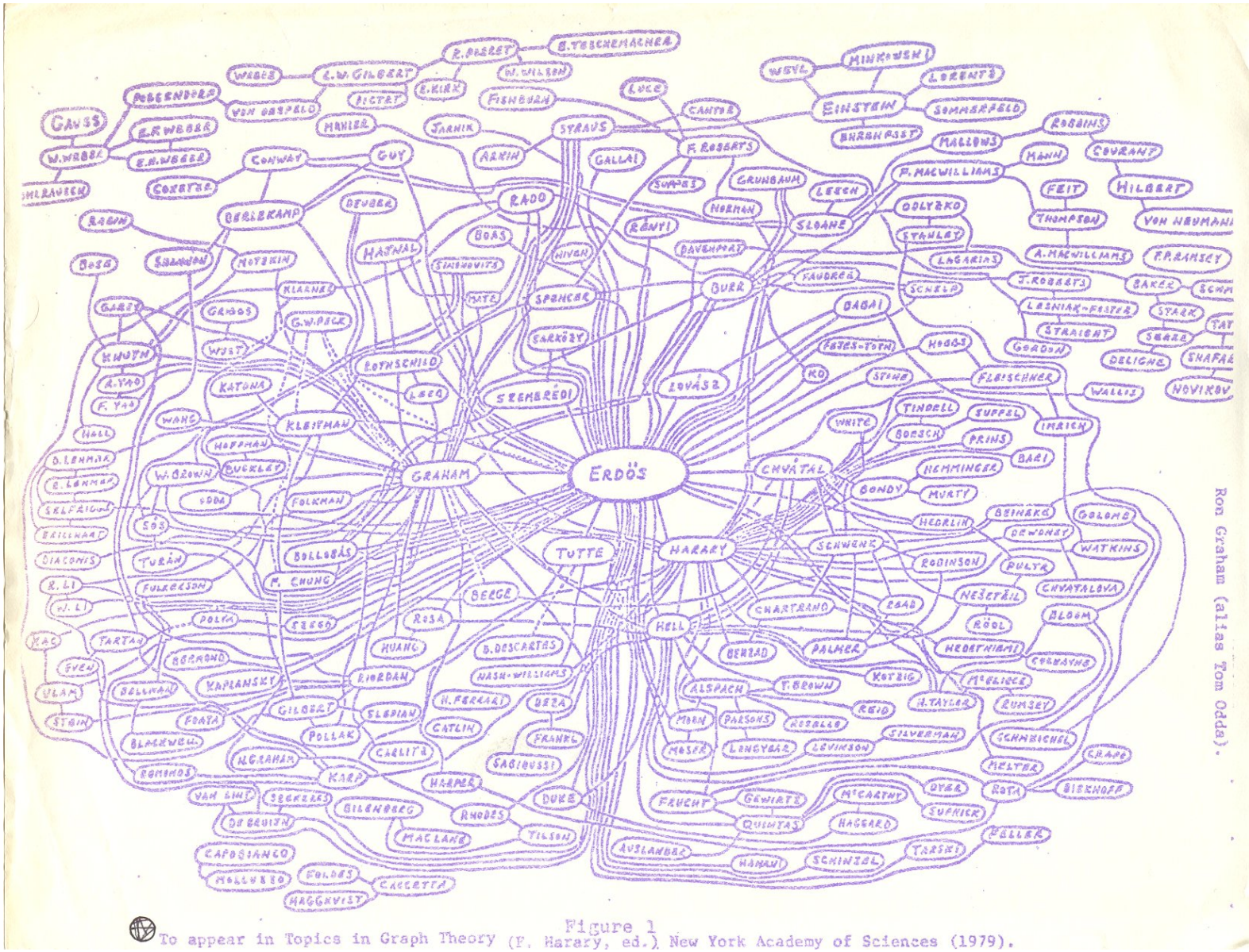


Figure 1

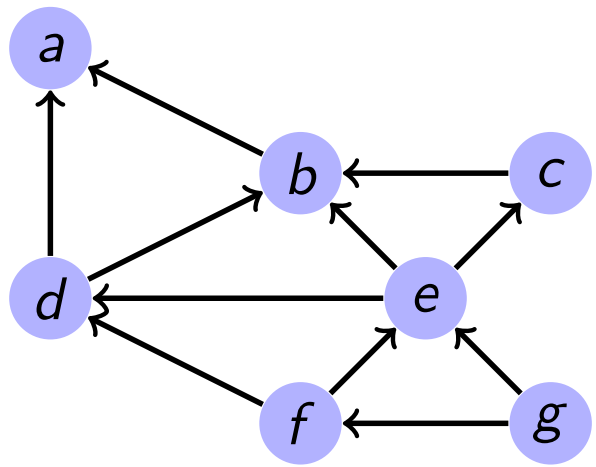
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

Analogous concepts for directed graphs

- We use the same notation for directed graphs, i.e. denoting a di-graph as $G = (V, E)$, where now the edges in E are directed.
- Formally, an edge $\langle u, v \rangle \in E$ is now an ordered pair in contrast to an undirected edge (u, v) which is unordered pair.
  - However, it is usually clear from context if we are discussing undirected or directed graphs and in both cases most people just write (u, v).

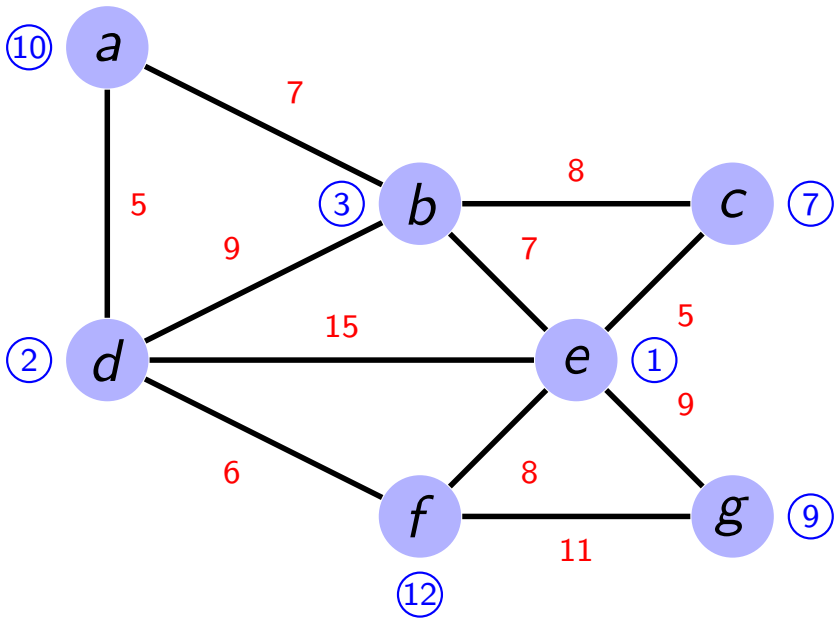
Analogous concepts for directed graphs

- Directed graphs instead have in-degree, out-degree, directed paths & directed cycles.
- Instead of connected components, we have strongly connected components.
  - A graph is strongly connected if there is a directed path between all ordered pairs of nodes
  - A strongly connected component is a maximal strongly connected subgraph



Weighted graphs

- We will often consider weighted graphs. Lets consider a (directed or undirected) graph $G = (V, E)$. Example:



- red numbers: edge weights
- blue numbers: vertex weights

- We can have a weight $w(v)$ for each node $v \in V$ and/or a weight $w(e)$ for each edge $e \in E$.
- For example, in a social network whose nodes represent people, the weight $w(v)$ of node v might indicate the importance of this person.
- The weight $w(e)$ of edge e might reflect the strength of a friendship.

Edge weighted graphs

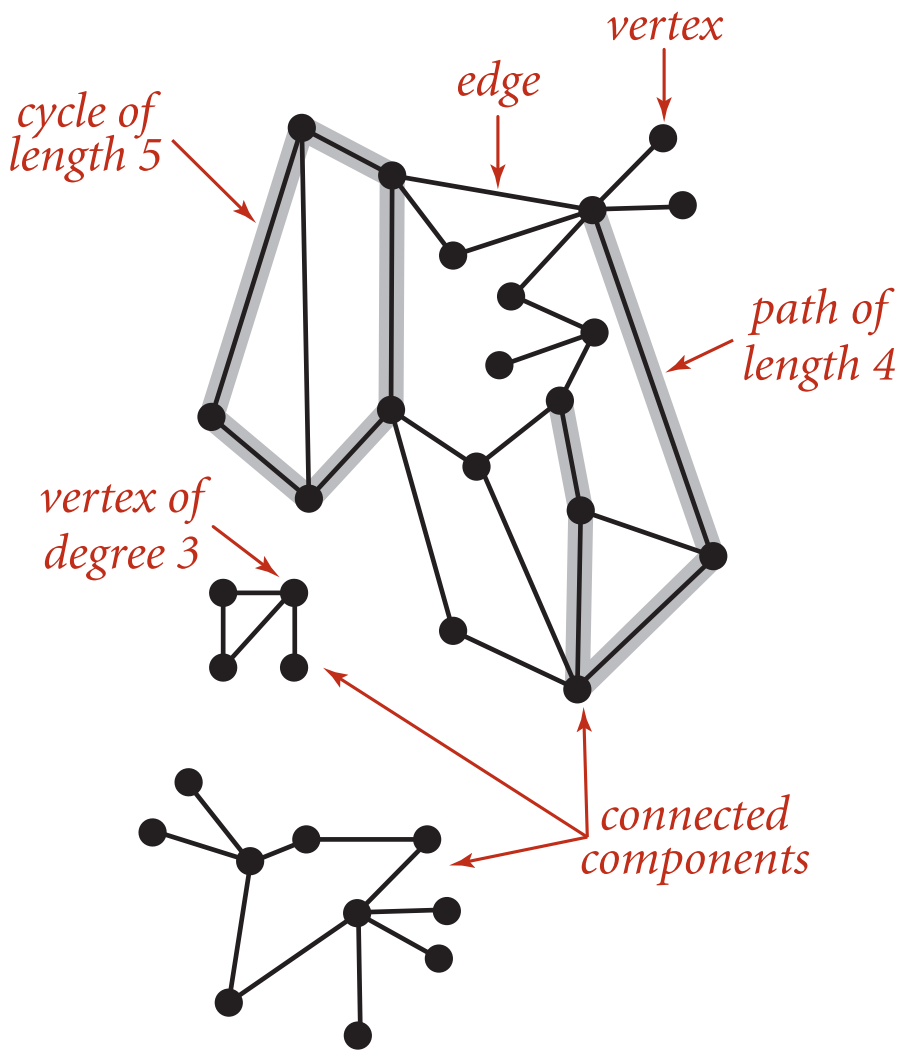
- When considering edge weighted graphs, we often have edge weights $w(e) = w(u, v)$ which are non negative (with $w(e) = 0$ or $w(e) = \infty$ meaning no edge depending on the context).

- In some cases, weights can be either positive or negative. A positive (resp. negative) weight reflects the intensity of connection (resp. repulsion) between two nodes (with $w(e) = 0$ being a neutral relation).

- Analogous to shortest paths in an unweighted graph, we often wish to compute least cost paths, where the cost of a path is the sum of weights of edges in the path.

- Sometimes (as in Chapter 3) we will only have a qualitative (rather than quantitative) weight, to reflect a strong or weak relation (tie).

Undirected graph anatomy: summary thus far

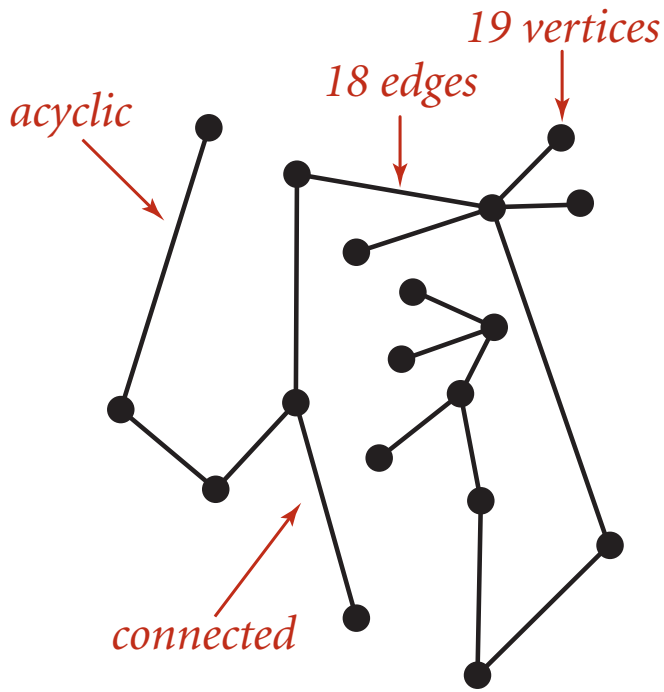


[from Algorithms, 4th Edition by Sedgewick and Wayne]

Acyclic graphs (forests)

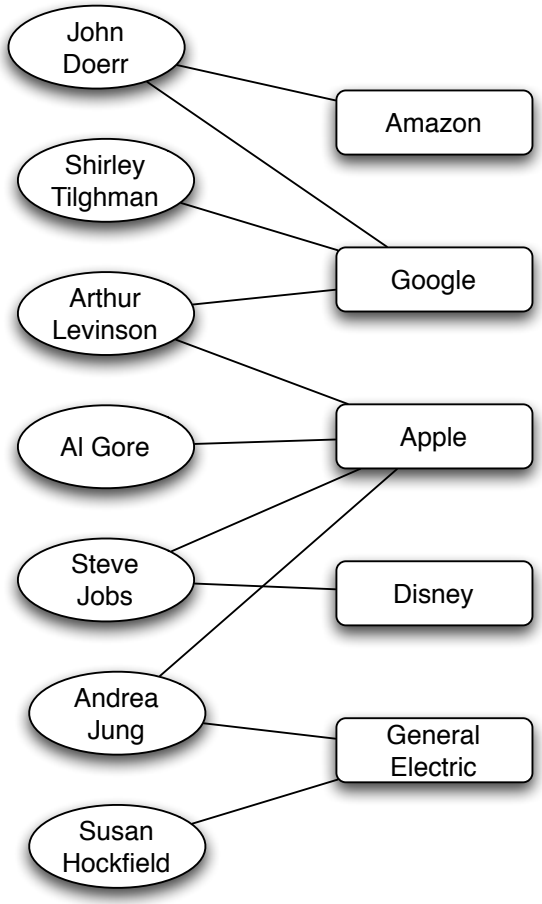
- A graph that has no cycles is called a forest.
- Each connected component of a forest is a tree.

  - A tree is a connected acyclic graph.
  - Question: Why are such graphs called trees?
  - Fact: There are always $n-1$ edges in an n node tree.



- Thus, a forest is simply a collection of trees.

Another tree [E&K Figure 4.4]



- The bipartite graph from before (depicting membership on corporate boards) is also an example of a tree.
- In general, bipartite graphs can have cycles.
- Question: is an acyclic graph always bipartite?

Facts

- It is computationally easy to decide if a graph is acyclic or bipartite.
- However, we (in CS) strongly "believe" it is not easy to determine if a graph is tripartite (i.e. 3-colourable).

Analogous concepts for directed graphs

- We now have directed paths and directed cycles.
- Instead of the degree of a node, we have the in-degree and out-degree of a node.

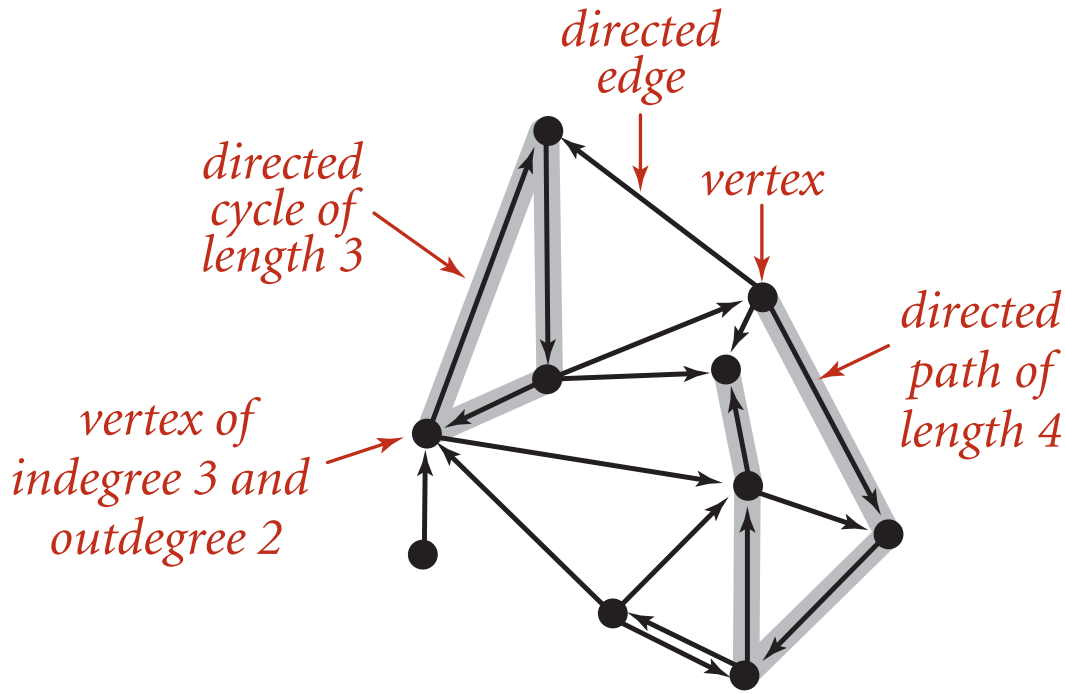
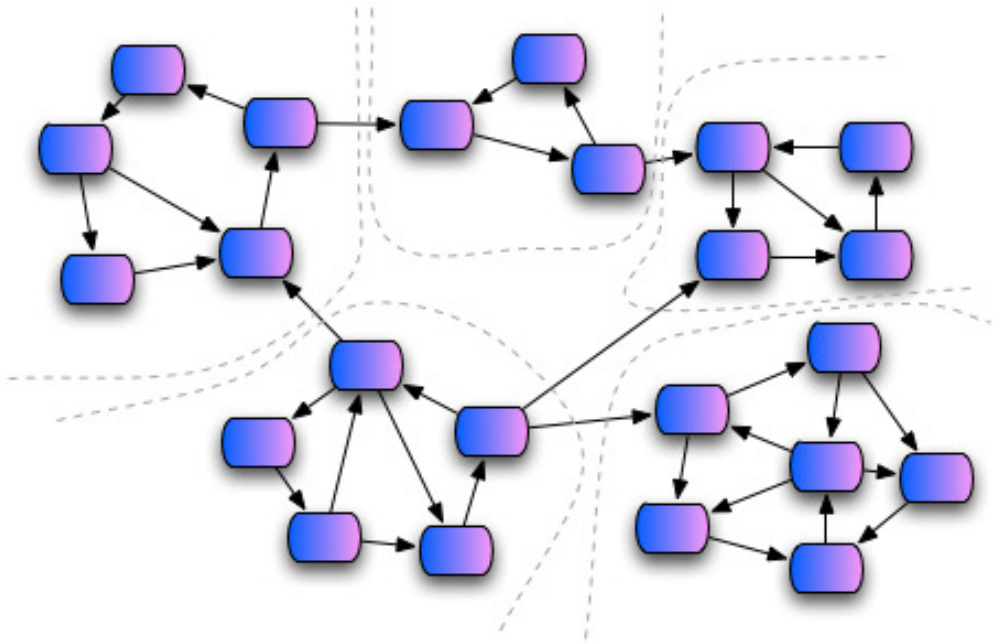


Figure: Directed graph antonomy [from Sedgewick and Wayne]

More analogous concepts for directed graphs

- Acyclic mean no directed cycles.
- Instead of connected components, we have strongly connected components.

[from `http://scientopia.org/blogs/goodmath/`]



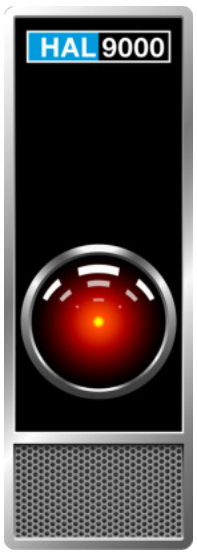
- Instead of trees, we have directed (i.e. rooted) trees which have a unique root node with in-degree 0 and having a unique path from the root to every other node.
- Question: What is a natural example of a rooted tree?

Recap

- Basic graph terminology
- What can graphs represent?
- Why study networks?
  - What are the challenges?

Fri. Jan 12th: Announcements & Corrections

- Happy Friday! You've survived your first week back ;)
- The Pre-Term survey closes tomorrow; a big thank you to those who've finished it :)
- It's international talk like Hal day! Happy birthday Hal!



Fri. Jan 12th: Announcements & Corrections

- I got some questions on Wednesday – Excellent! :D
- Clarifying "maximal" when we say that a strongly connected component is a maximal strongly connected subgraph.
  - Subgraph of $G = (V, E)$: Some $G' = (V', E')$ such that $V' \subseteq V$, and $E' = \{(u, v) \in E | u, v \in V'\}$
  - Strongly connected subgraph of the directed graph $G = (V, E)$: Any subgraph $G' = (V', E')$ such that there is a directed path between every ordered pair of vertices in V'
  - Strongly connected component of the directed graph $G = (V, E)$: Any strongly connected subgraph $G' = (V', E')$, such that there is no strongly connected subgraph $\tilde{G} = (\tilde{V}, \tilde{E})$ where $V' \subset \tilde{V}$
- Question: Is the claim "there is no larger strongly connected subgraph $\tilde{G}$ that completely contains our strongly connected subgraph G'", equivalent to saying "there is no individual node that we can add to our subgraph such that the result is still strongly connected"?
  - Also answer this question for connected subgraphs! (i.e., the undirected case)
- Tic-tac-toe game tree: A rooted tree or not?

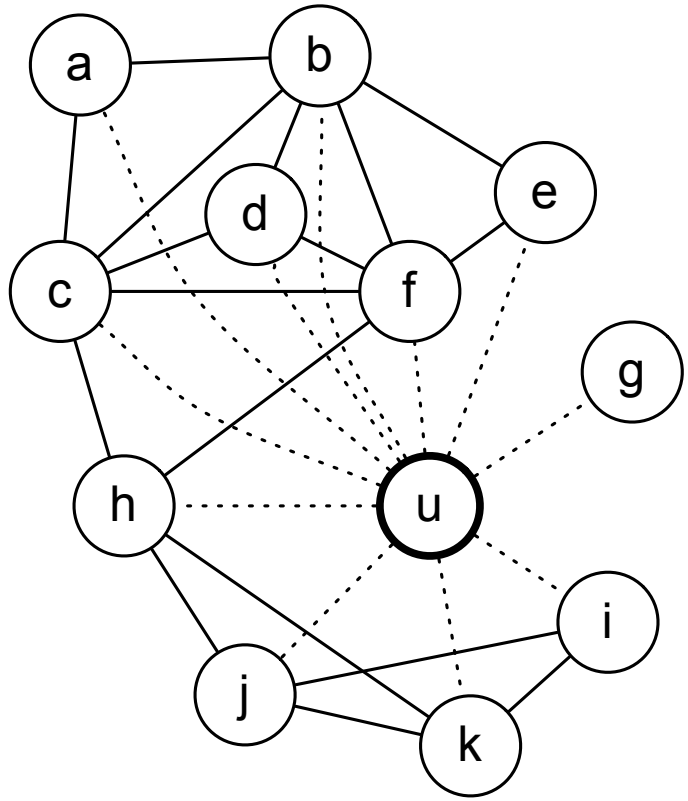
Detecting the romantic relation in Facebook: Course motivation and a lead in to Chapter 3 of text

- There is an interesting paper by Backstrom and Kleinberg (http://arxiv.org/abs/1310.6753) on detecting "the" romantic relation in a subgraph of Facebook users who specify that they are in such a relationship.
- Backstrom and Kleinberg construct two datasets of randomly sampled Facebook users:
  - (i) an extended data set consisting of 1.3 million users declaring a spouse or relationship partner, each with between 50 and 2000 friends
  - (ii) a smaller data set extracted from neighbourhoods of the above data set (used for the more computationally demanding experimental studies)
- The main experimental results are nearly identical for both data sets.
- Question: How would you go about identifying someone's spouse given their Facebook profile & feed?

Detecting the romantic relation (continued)

- They consider various "interaction features" including
  1. the number of photos in which both A and B appear.
  2. the number of profile views within the last 90 days.
- Their focus was various graph structural features of edges, including
  1. the *embeddedness* of an edge
     - ★ Intuitively: Do their social circles have a large overlap?
  2. The *dispersion* of an edge
     - ★ Intuitively: Are the shared friends of A and B poorly connected to each other, in the graph *without* A and B?
     - ★ Put differently: Is it often the case that A and B are the only thing that their mutual friends have in common?

Embeddedness and dispersion example from paper

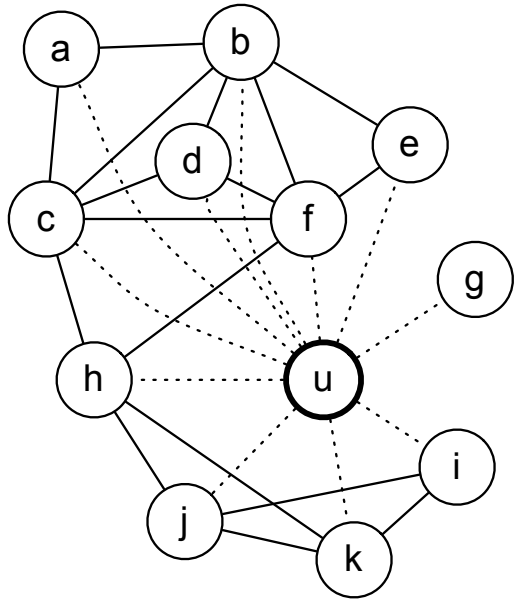


Definition (Embeddedness)

The *embeddedness* of an edge (A, B) is the number of mutual friends of A and B.

Links from u to b,c, and f all have embeddedness 5, the link from u to h has embeddedness of ... 4.

Embeddedness and dispersion example from paper



Definition (Dispersion)

The *dispersion* of an edge (A, B) is the number of pairs $\{s, t\}$ of mutual friends of A and B such that $(s, t) \notin E$ and s, t have no common neighbours except for A and B.

The dispersion of the edge (u, h) is ... 4. We have the unique pairs $\{f, j\}$, $\{f, k\}$, $\{c, j\}$, $\{c, k\}$.
Note that the nodes u and h are the unique pair of intermediaries from the nodes c and f to the nodes j and k. In contrast with embeddedness, the u-h link has greater dispersion than the links from u to b, c, and f.

Aside: Dispersion

Note: In this course, unless otherwise noted we'll be using the definition of dispersion from the paper, given in the slide before. However, other definitions exist – all follow the intuition that high-dispersion means that the mutual neighbours of A and B are not "well-connected" to each other (in the graph without A and B).

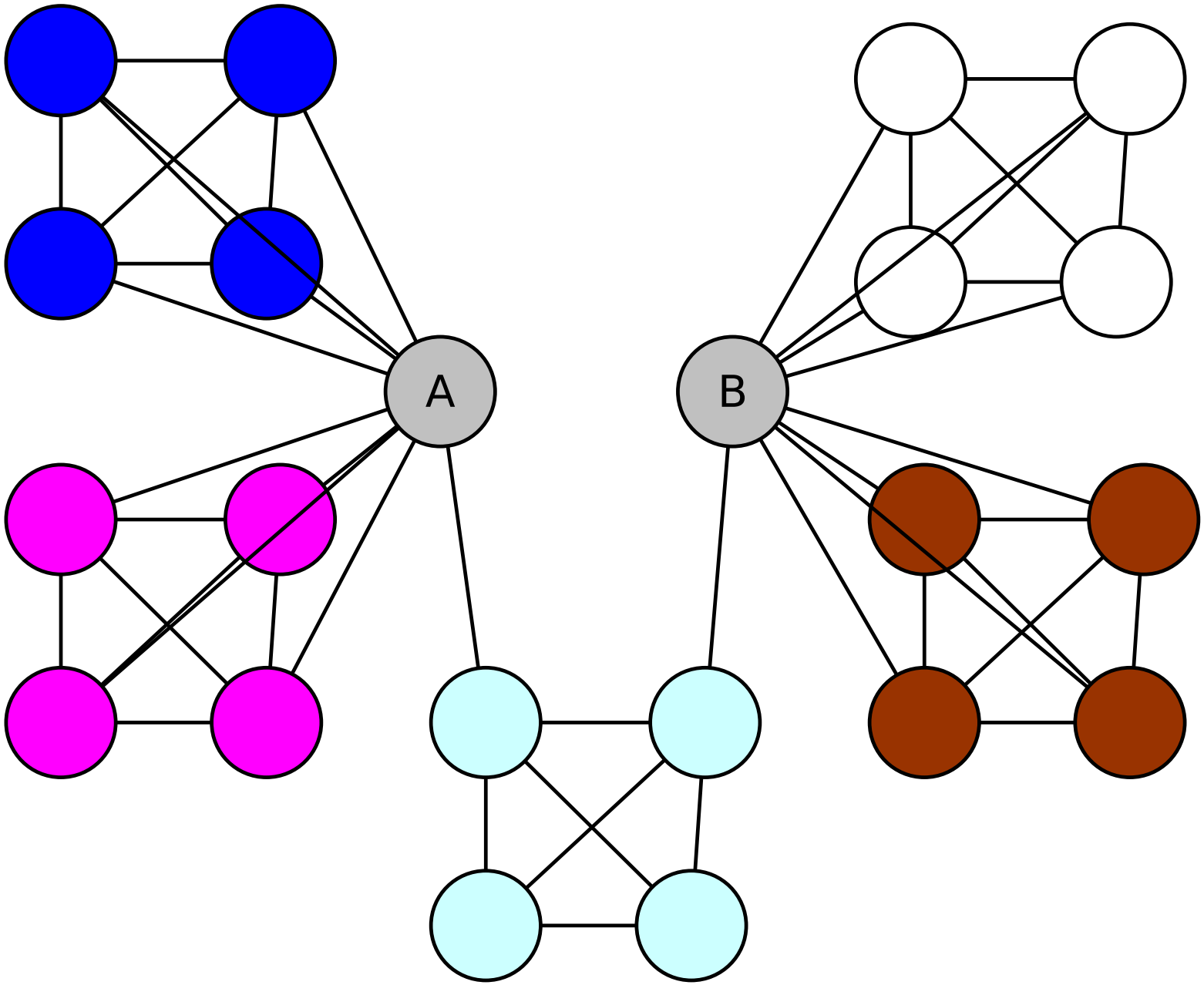
Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 friends, a random guess would have prediction accuracy of $1/200 = .5\%$

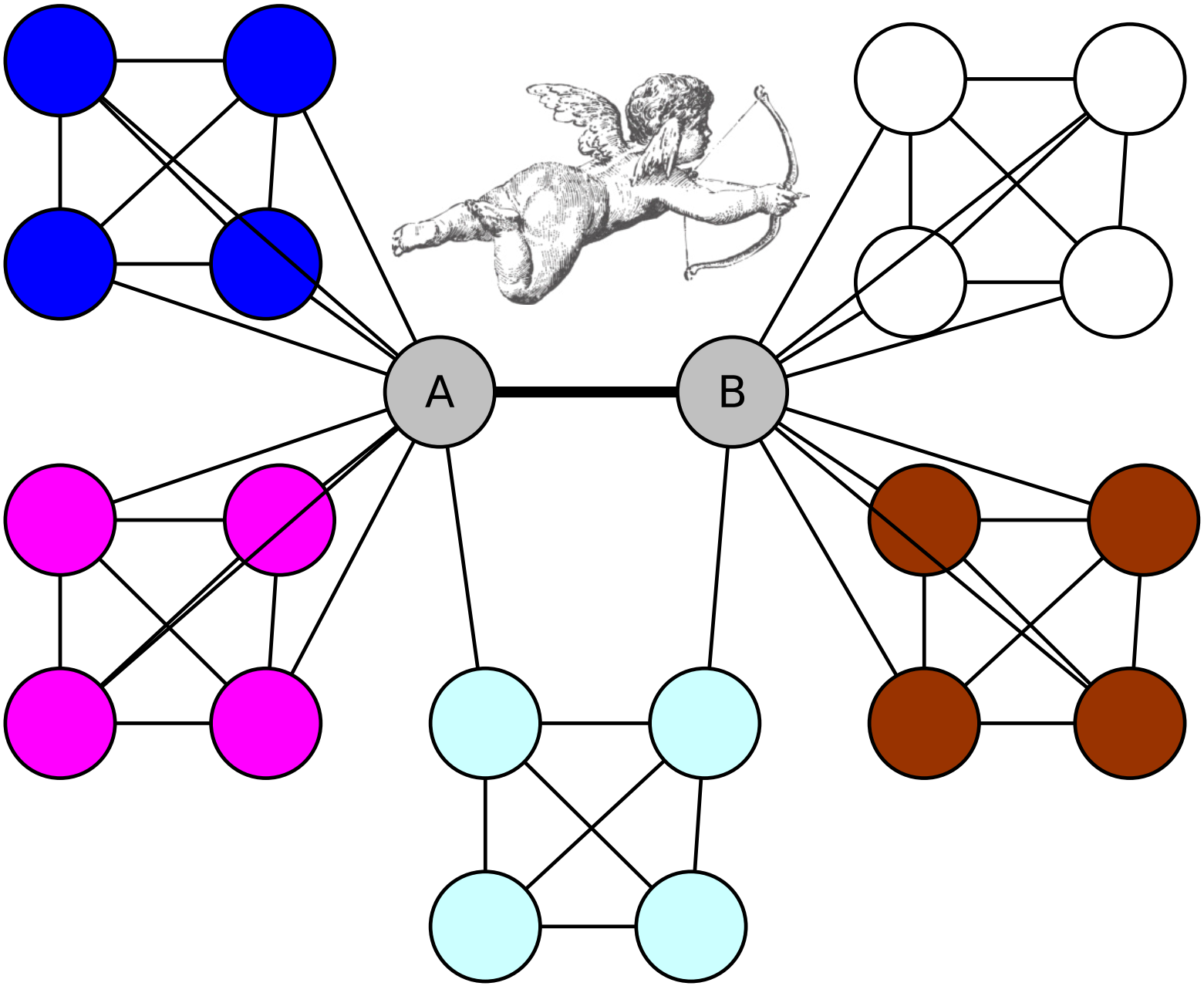
type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship.
- By itself, dispersion outperforms various interaction features.
- Why would high dispersion be a better measure than high embeddedness?

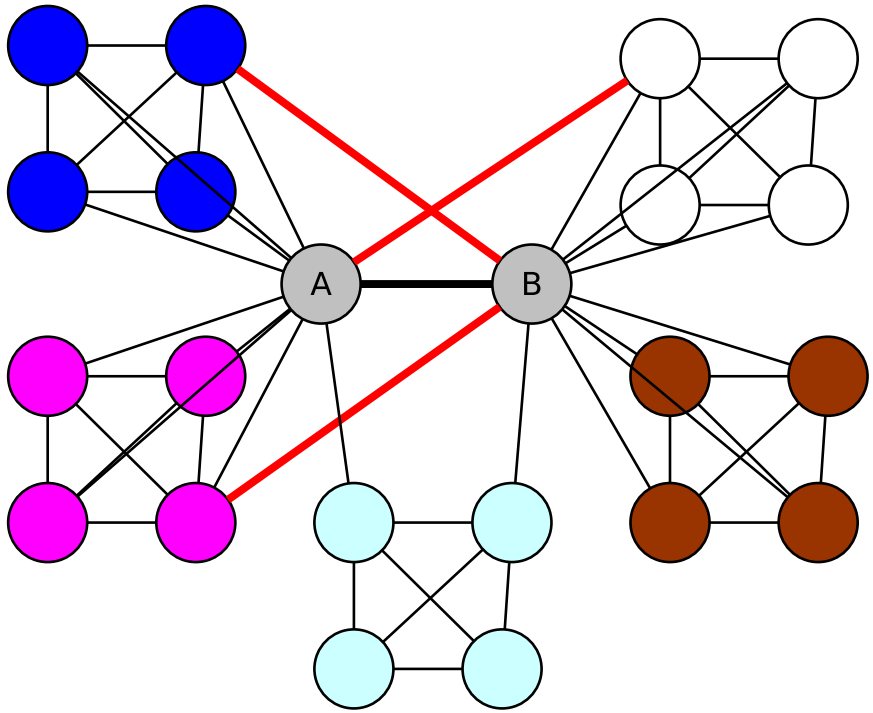
Dispersion Example



Dispersion Example



Dispersion Example



- Embededness of (A, B) is their number of mutual friends
- Easily skewed by dense groups (e.g. white nodes are chess club where everyone is eachothers' friend)
- Dispersion is how poorly connected A and B's mutual friends are if A and B are removed from the graph
- Romantic relationship can create pairs of mutual friends in different "social focii" (e.g. A's friend in chess club & B's dark blue friend)

Qualitative results from Backstrom and Kleinberg

- The goal is to predict (for each user in the data set) which of their friendship edges is the romantic relation. Note that each user has between 50 and 2000 friends and assuming say a median of 200 friends, a random guess would have prediction accuracy of $1/200 = .5\%$
- Various disperson measures do better than the embeddedness measure in its ability to predict the correct romantic relationship.
- By itself, dispersion outperforms various interaction features.
- Why would high dispersion be a better measure than high embeddedness?
- By combining many features, structural and interaction, the best performance is achieved using machine learning classification algorithms based on these many features.
- There are a number of other interesting observations but for me the main result is the predictive power provided by graph structure although there will generally be a limit to what can be learned solely from graph structure.

Some experimental results for the fraction of correct predictions

Recall that we argue that the fraction might be .005 when randomly choosing an edge. Do you find anything surprising?

type	embed	rec.disp.	photo	prof.view.
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (fem)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
engaged	0.179	0.446	0.442	0.391
engaged (fem)	0.171	0.399	0.386	0.401
engaged (male)	0.185	0.490	0.495	0.381
relationship	0.132	0.344	0.347	0.441
relationship (fem)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

type	max. struct.	max. inter.	all. struct.	all. inter.	comb.
all	0.506	0.415	0.531	0.560	0.705
married	0.607	0.449	0.624	0.526	0.716
engaged	0.446	0.442	0.472	0.615	0.708
relationship	0.344	0.441	0.377	0.605	0.682

Recap

With practice & review, you'll be able to:

- Recall or locate the course organization
  - People
  - Resources & communication
  - Course structure
  - Pre-Term Survey
  - Key policies
- Define basic graph terms & concepts, and apply them to example graphs
- Provide examples of what can graphs represent
- Recognize strengths and weaknesses of network modelling
  - Assess how these may apply to a new problem
- Contrast *embeddedness* and *dispersion* in the example application of detecting romantic relationships
  - Define & calculate the *embeddedness* of an edge
  - Define & calculate the *dispersion* of an edge
  - Conceptually explain what they measure