Social and Information Networks

University of Toronto CSC303
Winter/Spring 2023

Week 1: January 9-13

Welcome! Behind the mask, I am smiling ;)

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

This week's agenda

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

This week's agenda

- Course organization
  - ▶ People
  - ▶ Resources & communication
  - ▶ Course structure
  - ▶ Pre-Term Survey
  - ▶ Key policies
- Basic graph terminology
- What can graphs represent?
- Why study networks?
  - ▶ What are the challenges?

This week's agenda

- Course organization
  - ▶ People
  - ▶ Resources & communication
  - ▶ Course structure
  - ▶ Pre-Term Survey
  - ▶ Key policies
- Basic graph terminology
- What can graphs represent?
- Why study networks?
  - ▶ What are the challenges?
- Example application: Detecting romantic relationships
  - ▶ Embeddedness of an edge
  - ▶ Dispersion of an edge

This week's agenda

- Course organization
  - ▶ People
  - ▶ Resources & communication
  - ▶ Course structure
  - ▶ Pre-Term Survey
  - ▶ Key policies
- Basic graph terminology
- What can graphs represent?
- Why study networks?
  - ▶ What are the challenges?
- Example application: Detecting romantic relationships
  - ▶ Embeddedness of an edge
  - ▶ Dispersion of an edge
- Most likely, we will start next week's material on "the strength of weak ties"

Course Organization

Course Instructor: Ian Berlot-Attwell

- Email: $303s23$ HYPHEN instr AT cs DOT toronto DOT edu

Teaching Assistants: Soroush Ebadian, Parsa Farinneya, Aniket Kali, Fengwei Sun

Communications

1. Course Web page: source of first resort
   https://www.cs.toronto.edu/~ianberlot/303s23/

Communications

1. Course Web page: source of first resort
   https://www.cs.toronto.edu/~ianberlot/303s23/
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

Communications

1. Course Web page: source of first resort
   https://www.cs.toronto.edu/~ianberlot/303s23/
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://piazza.com/utoronto.ca/winter2023/csc303/home
   - 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!

Communications

1. Course Web page: source of first resort
   https://www.cs.toronto.edu/~ianberlot/303s23/

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://piazza.com/utoronto.ca/winter2023/csc303/home
   - 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.

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

- The course (both lectures & tutorials) are in-person
- The only in-person assessment is the midterm

Lecture/Tutorial/Course Structure

- The course (both lectures & tutorials) are in-person
- The only in-person assessment is the midterm

- 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

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - ▶ Lectures Mondays & Wednesdays
  - ▶ Tutorials on Fridays; starting Jan 20th in GB248, GB244, and GB120
    - ★ You should have been automatically enrolled in a tutorial section on ACORN
    - ★ In general, you are free to attend whichever room you prefer
    - ★ GB120 will not be recorded

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - Lectures Mondays & Wednesdays
  - Tutorials on Fridays; starting Jan 20th in GB248, GB244, and GB120
    - ★ You should have been automatically enrolled in a tutorial section on ACORN
    - ★ In general, you are free to attend whichever room you prefer
    - ★ GB120 will not be recorded
  - **This Friday, Jan 13th, is a lecture in MP202**

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - ▶ Lectures Mondays & Wednesdays
  - ▶ Tutorials on Fridays; starting Jan 20th in GB248, GB244, and GB120
    - ★ You should have been automatically enrolled in a tutorial section on ACORN
    - ★ In general, you are free to attend whichever room you prefer
    - ★ GB120 will not be recorded
  - ▶ **This Friday, Jan 13th, is a lecture in MP202**
  - ▶ Zoom Links for lecture are Quercus
  - ▶ When necessary, we will rearrange the schedule – therefore you should be available M,W,F 13:00-14:00 each week whether it is a lecture or a tutorial.

Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - ▶ Lectures Mondays & Wednesdays
  - ▶ Tutorials on Fridays; starting Jan 20th in GB248, GB244, and GB120
    - ★ You should have been automatically enrolled in a tutorial section on ACORN
    - ★ In general, you are free to attend whichever room you prefer
    - ★ GB120 will not be recorded
  - ▶ **This Friday, Jan 13th, is a lecture in MP202**
  - ▶ Zoom Links for lecture are Quercus
  - ▶ When necessary, we will rearrange the schedule – therefore you should be available M,W,F 13:00-14:00 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 (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

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

- Feedback, suggestions, & ideas for improving the course are welcome via
  - Email
  - Anonymously via https://forms.gle/fN3dYF2PrWY8WH8H8

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

- Found under the Quercus Quizzes tab
- https: //q.utoronto. ca/courses/ 293521/ quizzes/300331
- 5 minutes
- Closes Saturday Jan 14th
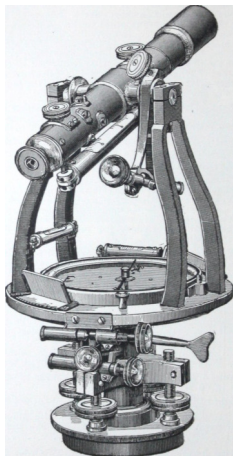
Preparation

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

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

`https://www.cs.toronto.edu/~ianberlot/303s23/material.html`

Grading scheme and schedule

Grading Scheme

1. Assignments: Two, each worth 15% = 30%
   Due dates: Thur. Feb. 16th and Thur. Mar 23rd
2. Critical review of a current article (groups 3-4): Worth 15%
   - Draft is due Fr. Mar. 24th, final submission due Thur. April 6th
3. Term Test (closed book, in tutorial rooms): Worth 20%
   Fri. Mar. 10th
4. Final Exam (open book, take-home): Worth 35%
   Date: TBD, 48 hour window, should take you 3-5 hours

Be careful! Feb 16 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!

1. For 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...

Submitting work late – this slide is important!

1. For 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

Submitting work late – this slide is important!

1. For 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
   - Late work is unlikely to be graded in time for the solution take-up in tutorials

Submitting work late – this slide is important!

1. For 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
   - Late work is unlikely to be graded in time for the solution take-up in tutorials
   - 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

Policies

1. All requests for remarking must be submitted on Markus within one week of work being graded.

Policies

1. All requests for remarking must be submitted on Markus within one week of work being graded.

2. Collaboration and Plagiarism:
   - We encourage discussion of course materials, but any work submitted must be your own!
   - Any material from a published source 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.

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/303s23/CSC303s23.pdf
   ```

Recap

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

Announcements

- We're looking for volunteer notetakers! See Quercus announcement for details

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

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"

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

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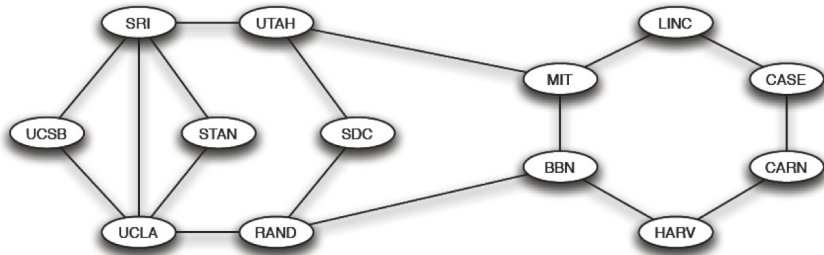
Figure: Initial internet: Dec. 1970 [E&K, Ch.2]

December 1970 internet visualized geographically [Heart et al 1978]
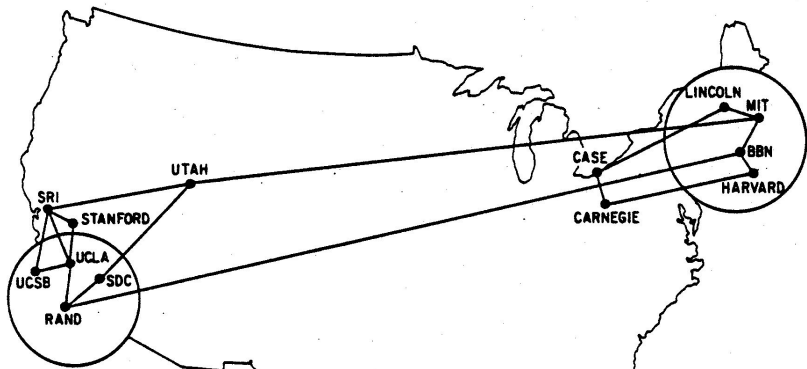
The first social network analysis

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.

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).

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 classrooms at a public school and he asked each child to choose two children to sit next to in class. He used this study to understand relationships (and other relationships).

Here boys are depicted by triangles and girls by circles.
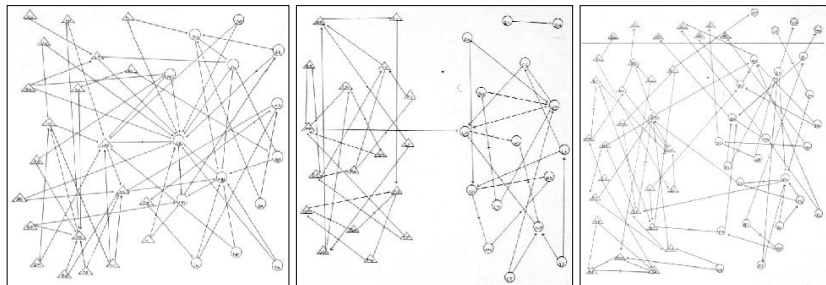
1st grade

8th grade

A closer look at grade 1 in Moreno sociogram
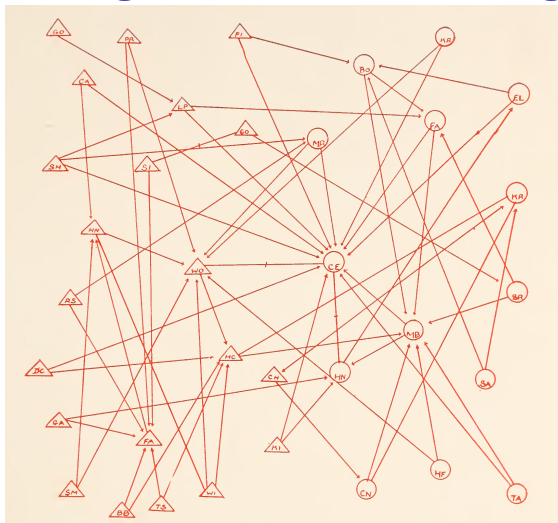
Figure: 21 boys, 14 girls. Directed graph. Most nodes have out-degree 2. 18 are not chosen, thus having in-degree 0. Note also that there are some "stars" with high in-degree.

A closer look at grade 4 in Moreno sociogram
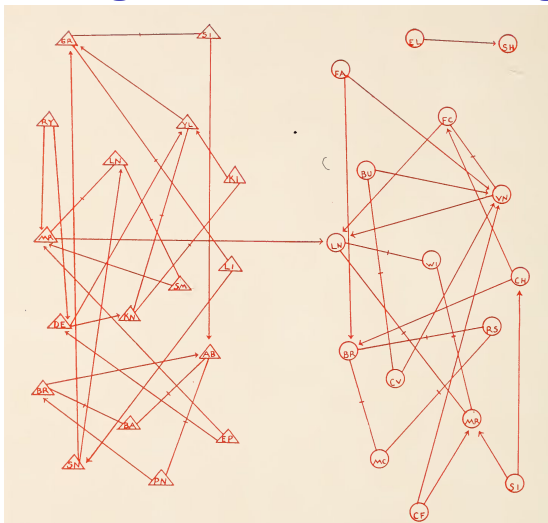
Figure: 17 boys, 16 girls. Directed graph with 6 unchosen having in-degree 0. Moreno depicted his graphs to emphasize inter-gender relations. Note only one edge from a boy to a girl.

A closer look at grade 8 in Moreno sociogram
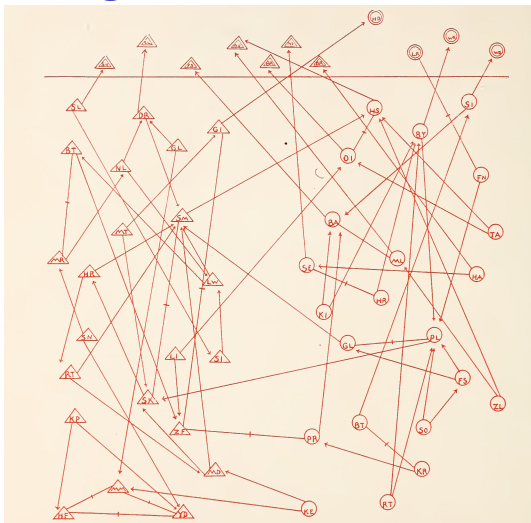
Figure: 22 boys, 22 girls. Directed graph with 12 unchosen having in-degree 0. Some increase in inter-gender relations. Double triangles and circles above line indicate individuals outside of the class.

Romantic Relationships [Bearman et al, 2004]
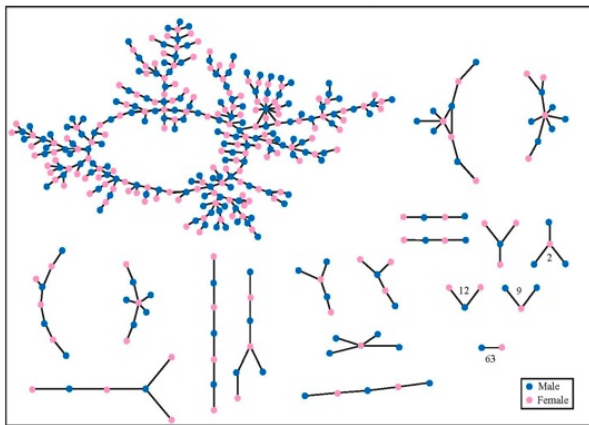
Figure: Dating network in US high school over 18 months.

- 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. What if they are incompatible?

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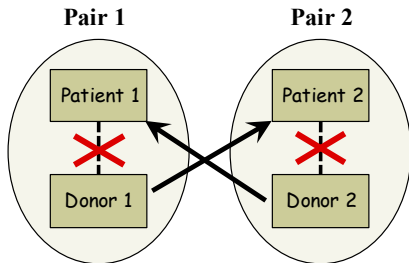
- Exchange: supports willing pairs who are incompatible
  1. allows multiway-exchange
  2. supported by sophisticated algorithms to find matches

Kidney Exchange: Swap Chains

- But what if someone renegs? $\Rightarrow$ Cycles require simultaneous transplantation; Paths require an altruistic donor!
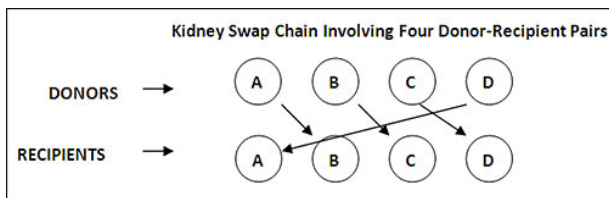
Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

Communities: Karate club division
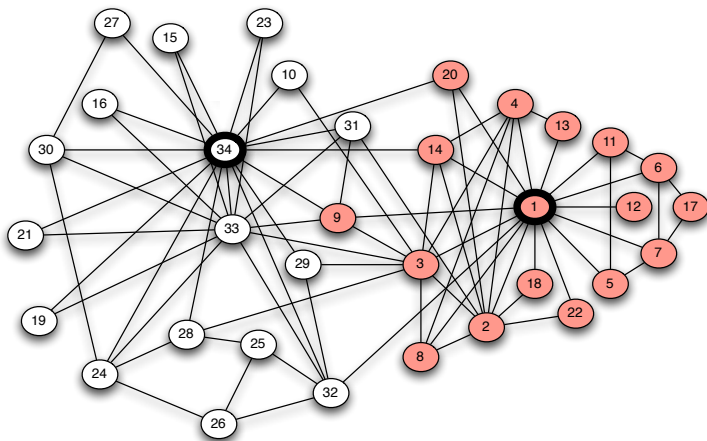
Karate Club social network, Zachary 1977

Figure: Karate club splis into two clubs

Communities: 2004 Political blogsphere
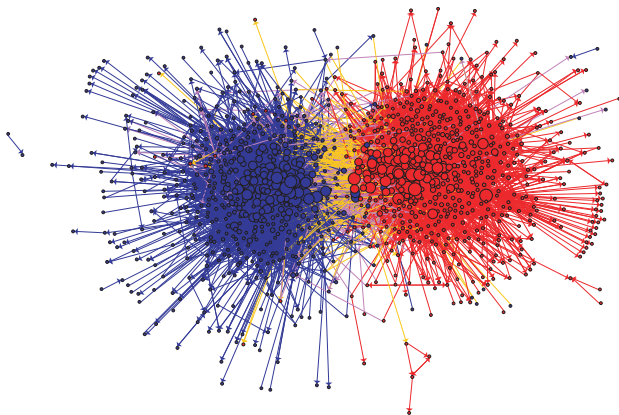
Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones 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
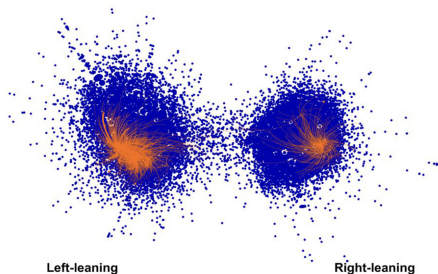
Fig. 1. Retweet Network Graph: RU-IRA Agents in #BlackLivesMatter Discourse. The graph (originally published [3]) shows accounts active in Twitter conversations about #BlackLivesMatter and shooting events in 2016. Each node is an account. Accounts are closer together when one account retweeted another 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
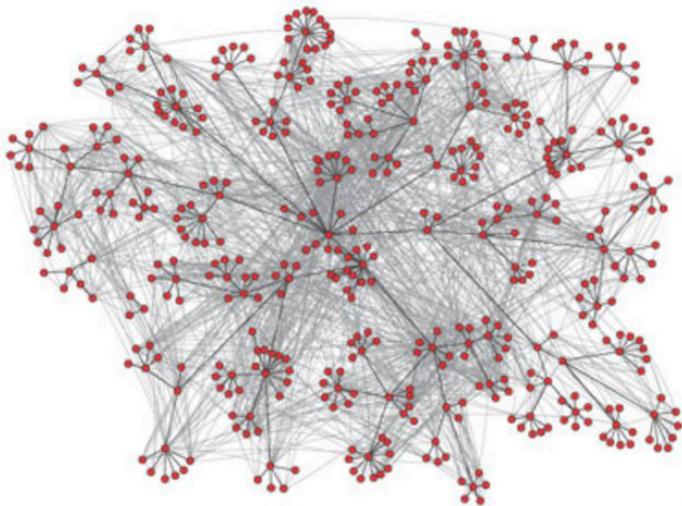
Figure: Email communication among 436 employees of Hewlett Packard Research Lab, superimposed on the organizational hierarchy [Fig 1.2, EK textbook]
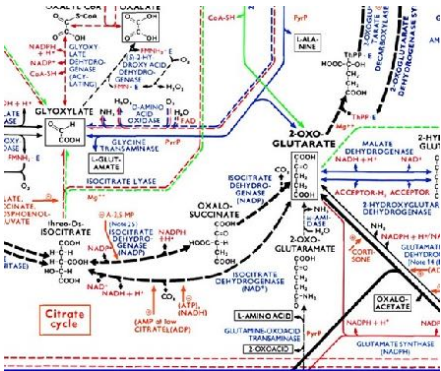
Protein-protein interaction network
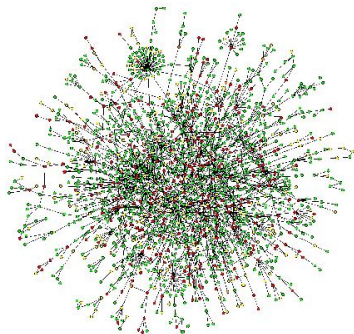
Protein-Protein Interaction Networks
Nodes: Proteins
Edges: 'physical' interactions

Metabolic network

Metabolic networks
Nodes: Metabolites and enzymes
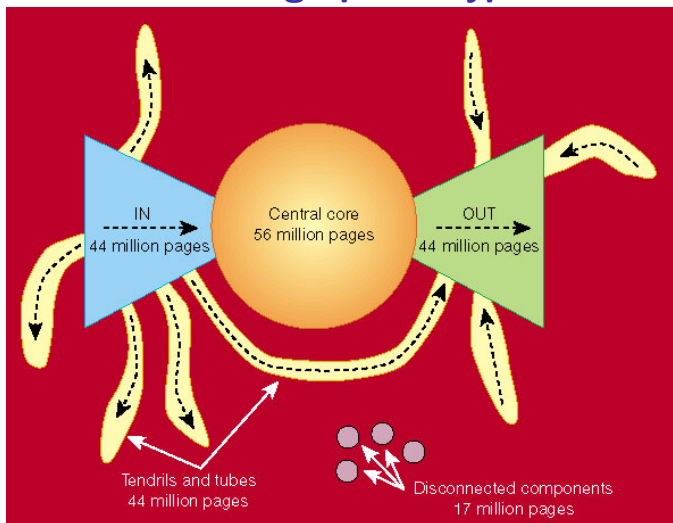Edges: Chemical reactions

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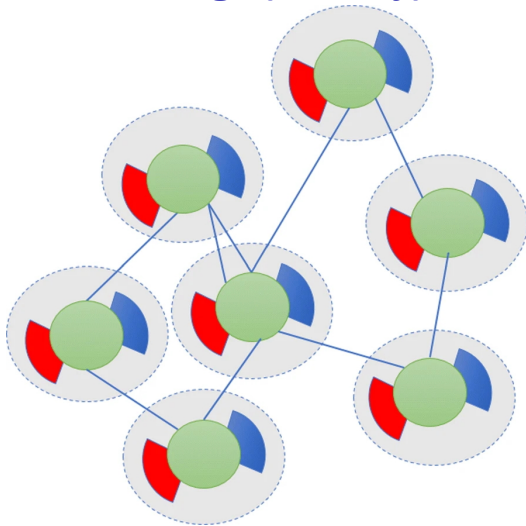
Figure: A schematic picture of the local bow tie structure of the modern web
Fujita et al. 2019]

The current interest in networks

- Clearly there are complex systems and networks that we are in contact with daily.

- The population of the world can be thought of as social network of approximately 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* analyse properties of a network, and then compare to real or synthetic data.

What are the challenges?

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.

Wed. Jan 11th: Announcements & Corrections

- A big thank you to those on Zoom this Monday
  - ▸ Hopefully the mic and chat issues are now resolved

Wed. Jan 11th: Announcements & Corrections

- A big thank you to those on Zoom this Monday
  - Hopefully the mic and chat issues are now resolved

- Monday's lecture got both positive and negative feedback
  - One piece of feedback was to "please stop being so excited [..]"
    - ★ First and foremost, thank you for letting me know!
    - ★ I can't improve if I don't know what doesn't work, and I know that giving suggestions for improvements can be very hard
    - ★ Having said that, I don't know how actionable this feedback is; some more specificity (e.g., gestures, pace, volume, etc...) would be very helpful for me

Wed. Jan 11th: Announcements & Corrections

- A big thank you to those on Zoom this Monday
  - ▸ Hopefully the mic and chat issues are now resolved

- Monday's lecture got both positive and negative feedback
  - ▸ One piece of feedback was to "please stop being so excited [..]"
    - ★ First and foremost, thank you for letting me know!
    - ★ I can't improve if I don't know what doesn't work, and I know that giving suggestions for improvements can be very hard
    - ★ Having said that, I don't know how actionable this feedback is; some more specificity (e.g., gestures, pace, volume, etc...) would be very helpful for me

  Monday's recording is available on Quercus
  - ▸ See the OCCS Student App for streaming
  - ▸ See the Files tab for a textfile with unedited download links
  - ▸ I'm working on a better download option

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.)
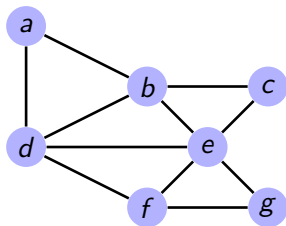
2. directed graphs (often called di-graphs).
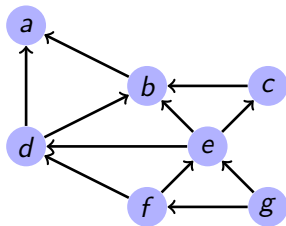
Visualizing Networks as Graphs

- 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, . . .
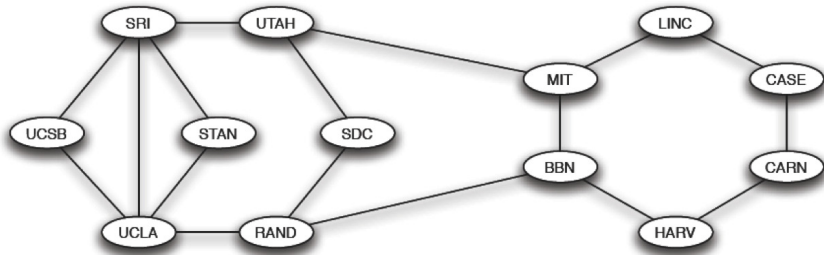 - membership in a club, class, political party, . . .

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.

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 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!
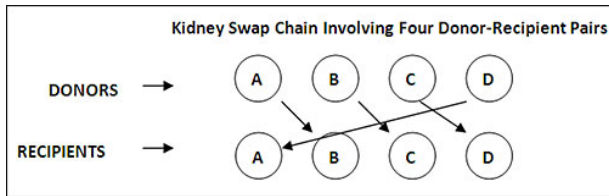

Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

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 $1 \le i \le k - 1$,
  - 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 $1 \le i \le k-1$,
  - 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
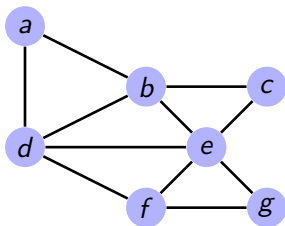
Romantic Relationships [Bearman et al, 2004]
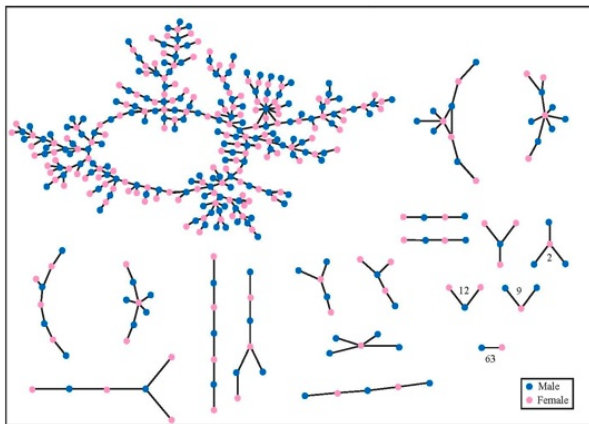
Figure: Dating network in US high school over 18 months.

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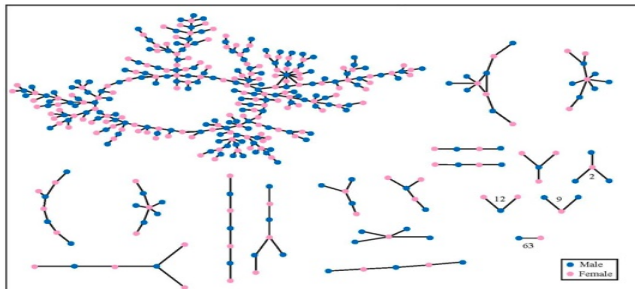- What predictions could you use this for?

More basic definitions
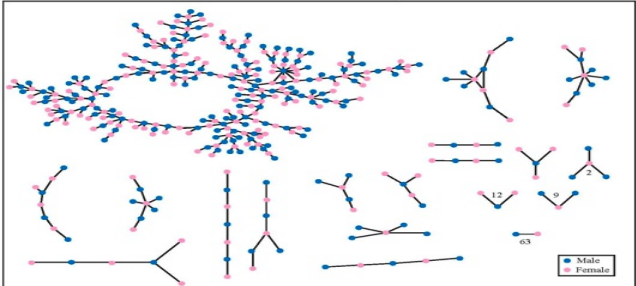
More basic definitions
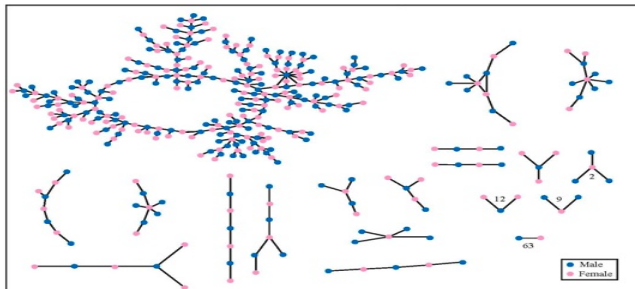
Observation

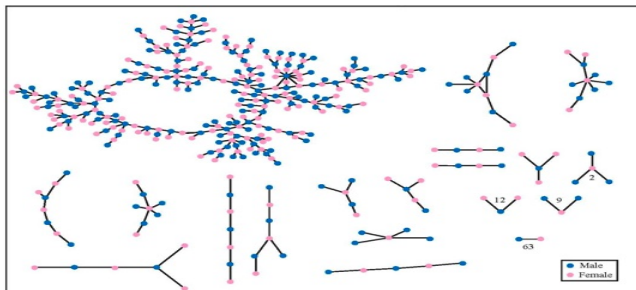Many connected components including one "giant component"

More basic definitions

Observation

Many connected components including one "giant component"

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.)
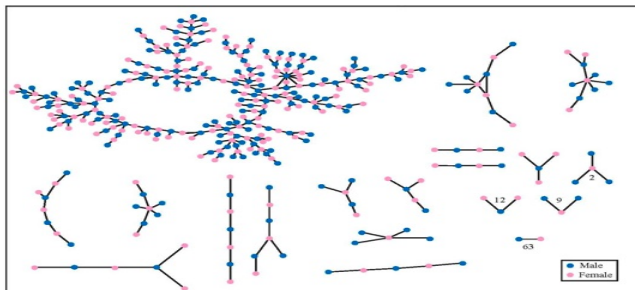
Simple paths and simple cycles

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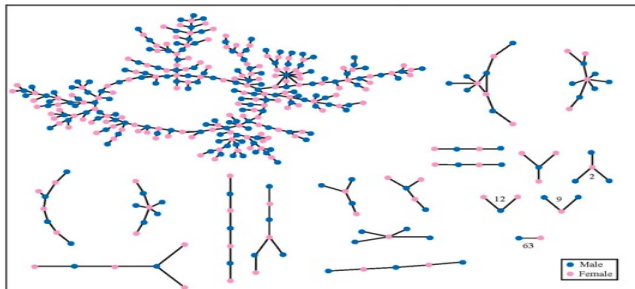
Observation

- 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.

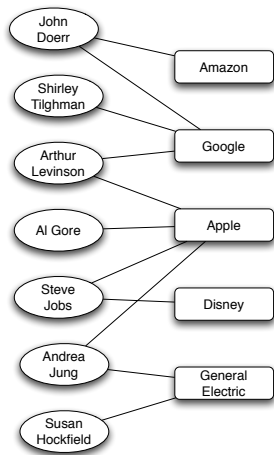# Example of an acyclic bipartite graph

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.
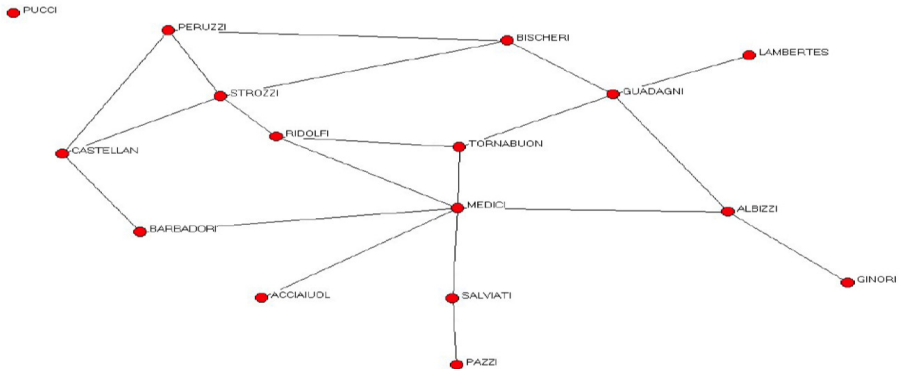
Florentine marriages and "centrality"

Figure: see [Jackson, Ch 1]
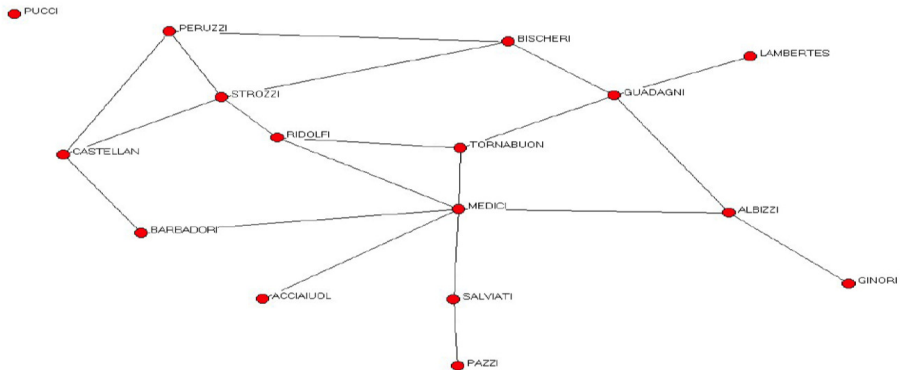
Florentine marriages and "centrality"

Figure: see [Jackson, Ch 1]

- Medici connected to more families, but not by much
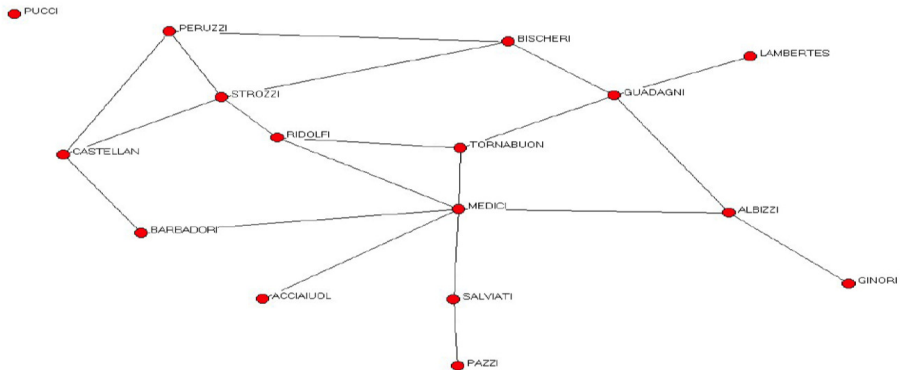
Florentine marriages and "centrality"

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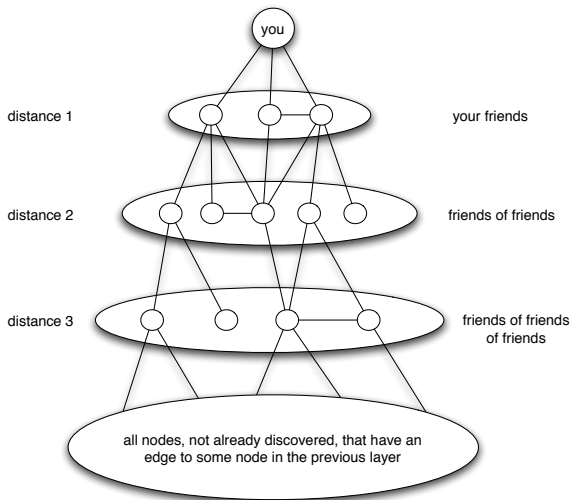
Breadth first search and path lengths [E&K, Fig 2.8]

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

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

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.

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]
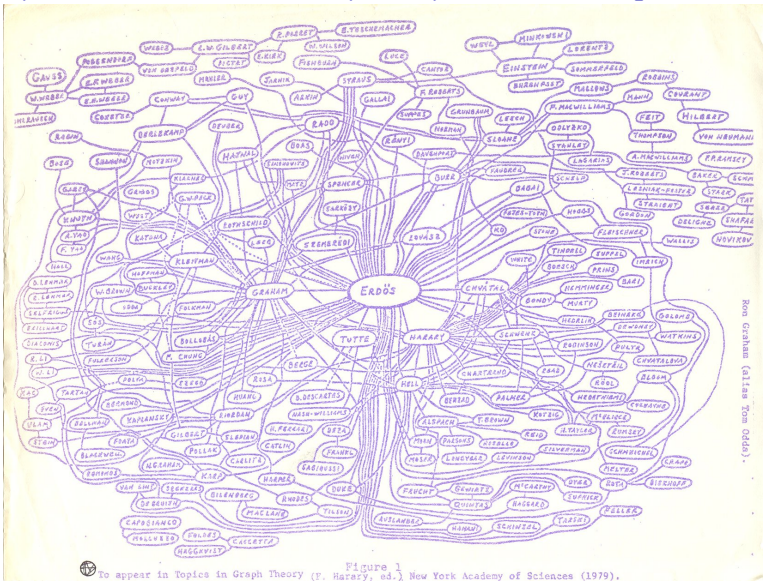
Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

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.

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

- 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).
- We now have directed paths and 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
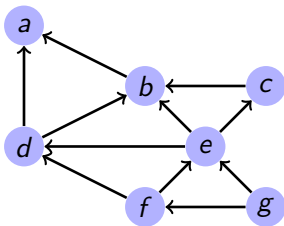
Weighted graphs

- We will often consider weighted graphs. Lets consider a (directed or undirected) graph $G = (V, E)$. Example:
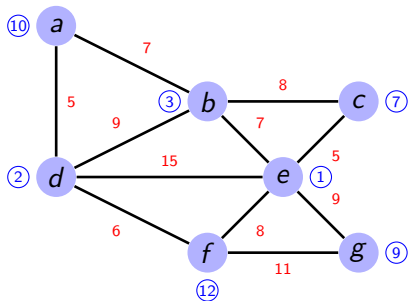
- ▶ red numbers: edge weights
- ▶ blue numbers: vertex weights

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 $e \in E$.

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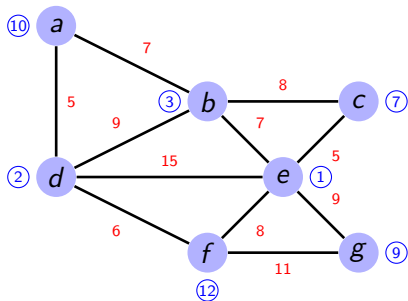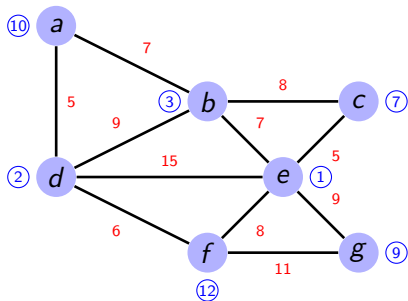
- We will often consider weighted graphs. Lets consider a (directed or undirected) graph $G = (V, E)$. Example:

- ▶ 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.

Weighted graphs

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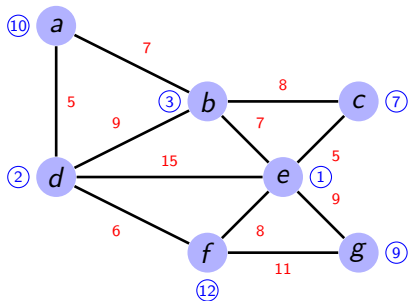

- red numbers: edge weights
- blue numbers: vertex weights

- 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).

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).

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.

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).

Graph anatomy: summary thus far
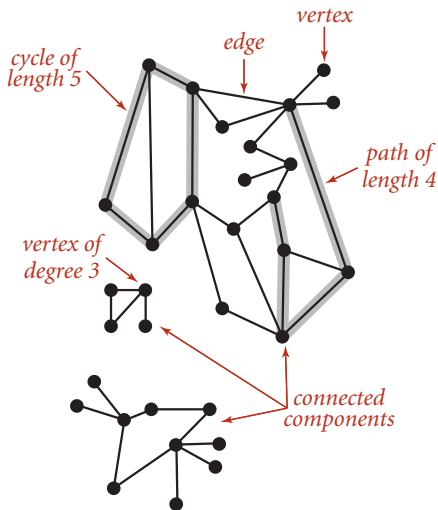
[from Algorithms, 4th Edition by Sedgewick and Wayne]

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.

acyclic

18 edges

19 vertices

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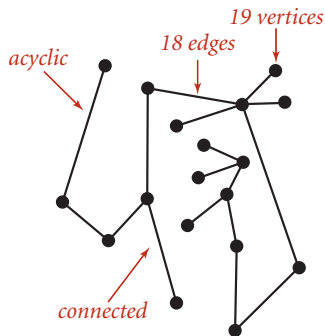
- Thus, a forest is simply a collection of trees.

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?
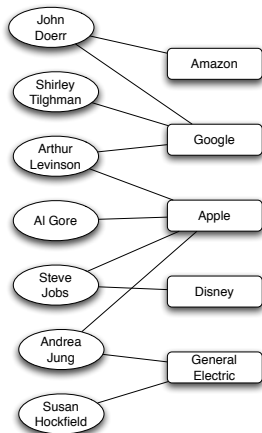
Facts

- 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.

directed edge

vertex

directed cycle of length 3

directed path of length 4

vertex of indegree 3 and outdegree 2
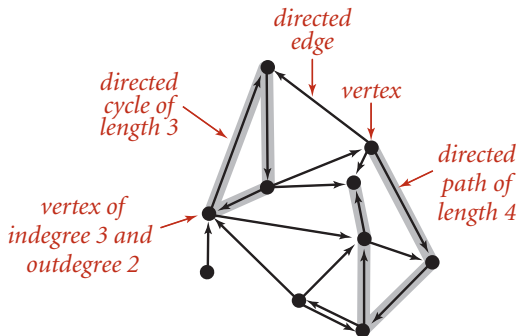
Figure: Directed graph antonomy [from Sedgewick and Wayne]

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/]

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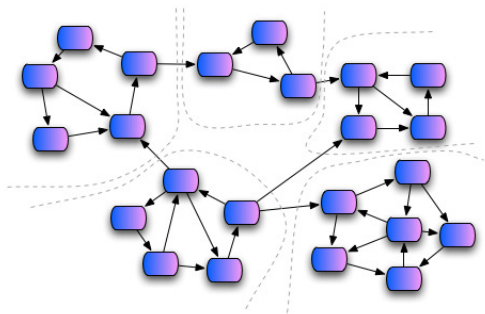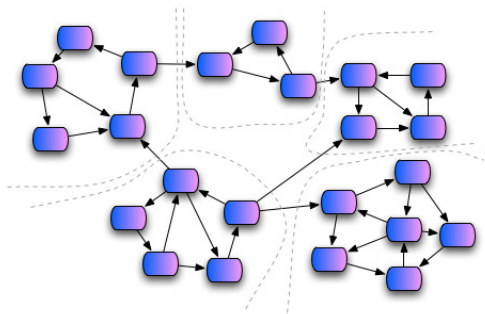
- Instead of trees, we have directed (i.e. rooted) trees which have a unique root node with in-degree 0 and having a unique path from the root to every other node.
- Question: What is a natural example of a rooted tree?

Recap

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

Fri. Jan 13th: Announcements & Corrections

- More feedback! :)
    - Again, a big thank you to everyone who gave suggestions for improvement, or let me know what they enjoyed :)
    - I got some really solid suggestions for improvements
        - Adjusting the mic to try and avoid volume spikes
        - Trying to repeat questions from the audience, for those in the back of the room, on Zoom, or watching recordings

Fri. Jan 13th: Announcements & Corrections

- More feedback! :)
  - ▶ Again, a big thank you to everyone who gave suggestions for improvement, or let me know what they enjoyed :)
  - ▶ I got some really solid suggestions for improvements
    - ★ Adjusting the mic to try and avoid volume spikes
    - ★ Trying to repeat questions from the audience, for those in the back of the room, on Zoom, or watching recordings
- The Pre-Term survey closes tomorrow; thank you to those who've finished it :)

Fri. Jan 13th: Announcements & Corrections

- More feedback! :)
  - ▸ Again, a big thank you to everyone who gave suggestions for improvement, or let me know what they enjoyed :)
  - ▸ I got some really solid suggestions for improvements
    - ⋆ Adjusting the mic to try and avoid volume spikes
    - ⋆ Trying to repeat questions from the audience, for those in the back of the room, on Zoom, or watching recordings
- The Pre-Term survey closes tomorrow; thank you to those who've finished it :)

- A belated happy 26th (31rst?) birthday Hal!

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.

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.

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.

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 (A, B) which is the number of mutual friends of A and B.
    2. various forms of a new *dispersion* measure of an edge (A, B) where high dispersion intuitively means that the mutual neighbours of A and B are not "well-connected" to each other (in the graph without A and B).
    3. One definition of dispersion given in the paper 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.

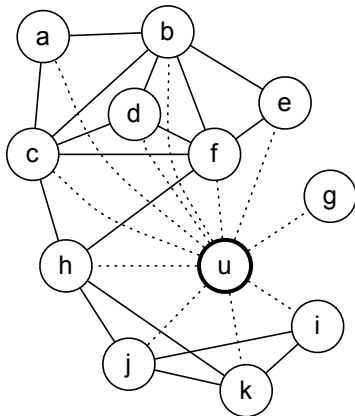# Embeddedness and dispersion example from paper

Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes c and f to the nodes j and k; the u-h link has greater dispersion than the links from u to b, c, and f.

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 $1/200 = .5\%$

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?
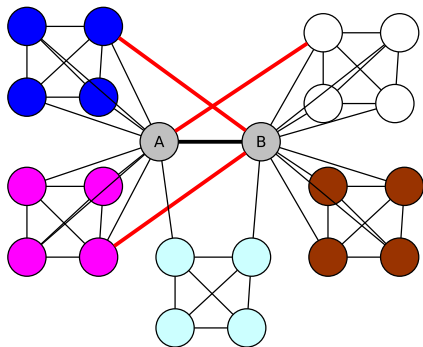
Dispersion Example
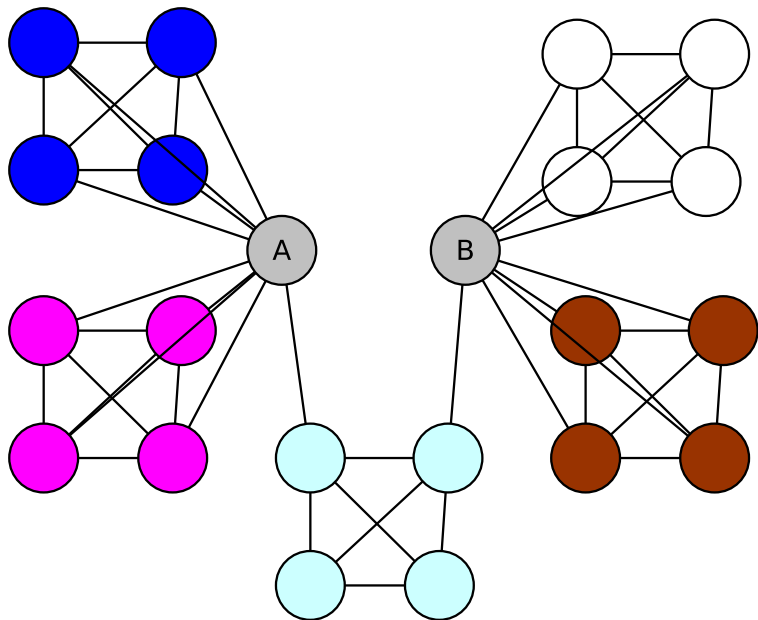
Dispersion Example
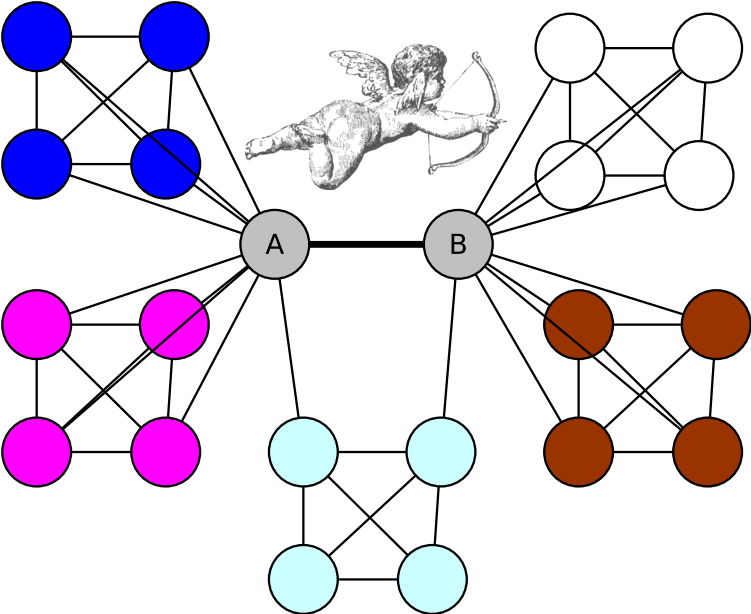
Dispersion Example

- 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 Example
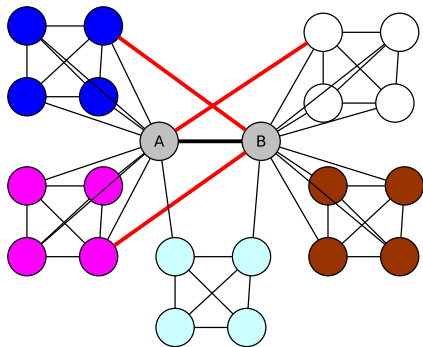
- Dispersion is how poorly connected A and B's mutual friends are if A and B are removed from the graph
- Romantic relationship can create pairs of mutual friends in different "social focii" (e.g. A's friend in chess club & B's dark blue friend)

Dispersion Example
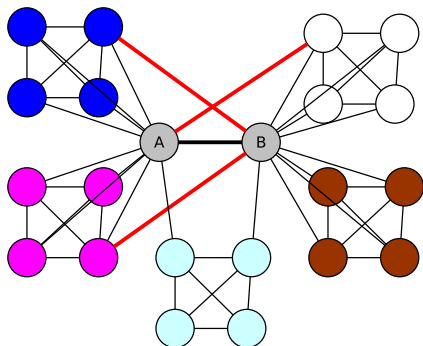
- Dispersion is how poorly connected A and B's mutual friends are if A and B are removed from the graph
- Normal friendships are less likely to form these connections (e.g. unlikely to be connections from white nodes to brown nodes if brown nodes are extended family)

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?

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.

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

- Course organization
  - ▶ People
  - ▶ Resources & communication
  - ▶ Course structure
  - ▶ Survey on office Hours & course organization
- Basic graph terminology
- What can graphs represent?
- Why study networks?
  - ▶ What are the challenges?
- Example application: Detecting romantic relationships
  - ▶ Embeddedness of an edge
  - ▶ Dispersion of an edge