

# Social and Information Networks

University of Toronto CSC303  
Winter/Spring 2019

Week 1: January 6-10 (2020)

# Course Organization

Course Instructor: Allan Borodin

- Office phone: 416 978 6416
- Office: (Sandford Fleming) SF 2303B
- Email: bor@cs.toronto.edu

Teaching Assistants : Ruijian An, Ian Berlot-Attweil, Reid McIlroy  
Email : 303s20tas@cs.toronto.edu

**Note: These course emails may take a couple of days to set up**

# Communications and Course Materials

- Communication:

- 1 Course Web page: source of first resort

<http://www.cs.toronto.edu/~csc303s20/>

- 2 Discussion board: **piazza** for questions of general interest

<https://piazza.com/utoronto.ca/winter2020/csc303s20>

<https://piazza.com/utoronto.ca/winter202/csc303>

Instructor and TA will monitor and respond as appropriate. I welcome (encourage) questions and responses to questions in class which leads to less confusion especially with regard to technical questions.

- 3 Office hours: TBA

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

- 1 Text: D. Easley, J. Kleinberg. Networks, Crowds, and Markets:

Reasoning About a Highly Connected World. Cambridge University Press, 2010. Online version available at

<http://www.cs.cornell.edu/home/kleinber/networks-book/>

- 2 We will supplement with some topics and material not in the text.

- 3 Additional materials will be linked on course web page.

# Lecture/Tutorial/Course Structure

- Times for lectures and tutorials
  - ▶ Lectures or Tutorials Monday, Wednesday and Friday.  
Although the Faculty calendar indicates tutorials on Fridays, we will usually have the tutorials on Wednesdays, and lectures Mondays and Fridays. Rooms will be announced. This week we will meet in GB248 for all classes.
  - ▶ However, if necessary, we will sometimes rearrange the schedule between tutorial time and lecture times. You should be available M,W,F 15:00-16:00 each week whether it is a lecture or a tutorial.
- More generally
  - ▶ Readings posted on web site usually (but not always) posted in advance. Initially I am posting slides from last year.
  - ▶ When available, read assigned sections of text prior to class, come prepared to discuss!
  - ▶ Lecture slides (some detailed, some less so) will usually be posted one or two days *after* the class. **The slides are not a reason to miss lectures or tutorials; the class discussions are part of the course and you are responsible (i.e., can be tested) for information that occurs in lectures and tutorials.**

# Preparation, reading scheme and schedule

You should be comfortable with very basic probability and discrete math concepts (in particular, some basic graph theory) as would be covered in the prerequisites. I have posted a probability primer on the course web page.

## Grading Scheme

- ① Assignments: Two, each worth  $15\% = 30\%$   
Tentative due dates: February 14 and March 15
- ② One critical review of a current article: Worth  $10\%$   
Tentative due date: March 29
- ③ Term Test: Worth  $20\%$   
Tentative date: March 4
- ④ Final Exam: Worth  $40\%$

# Policies

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

# What's in a name? Graphs or Networks?

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

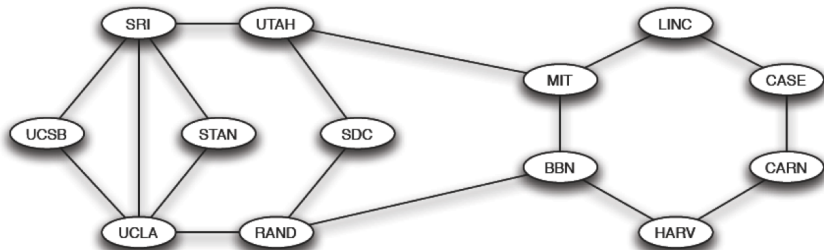
Here is one explanation for the different terminology: We use networks for settings where we think of links transmitting or transporting “things” (e.g. information, physical objects, friendship).

## Many different types of networks

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

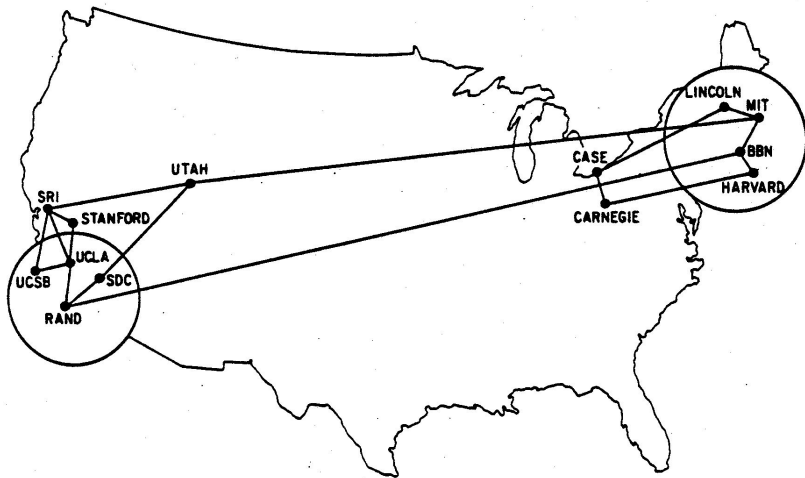
# Visualizing Networks

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



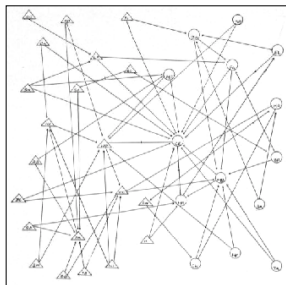
**Figure:** Initial internet: Dec. 1970 [E&K, Ch.2]

## December 1970 internet visualized geographically [Heart et al 1978]

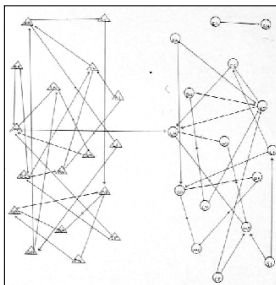


# The first social network analysis

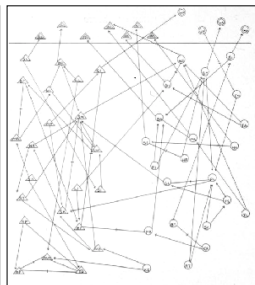
In his **1934** book *Who Shall Survive: A New Approach to the Problem of Human Interrelations*, Jacob Moreno (Romanian-US psychiatrist) introduced *sociograms* and used these graphs/networks to understand relationships. In one study (that was repeated to test changes) he asked each child in various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



1st grade



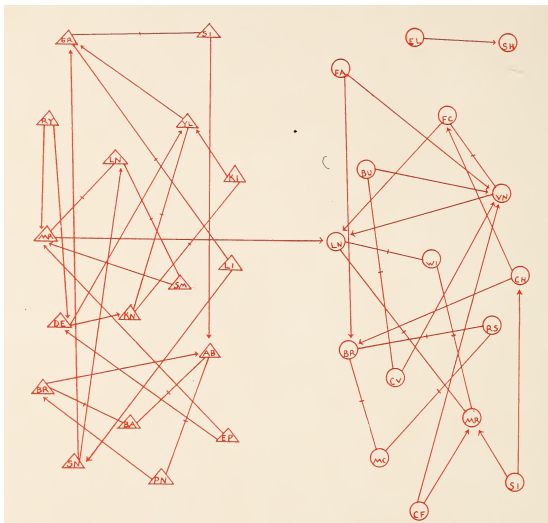
4th grade



8th grade



## A closer look at grade 4 in Moreno sociogram



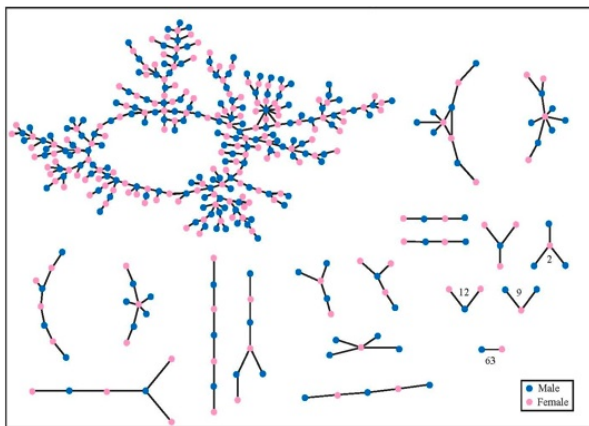
**Figure:** 17 boys, 16 girls. Directed graph with 6 unchosen having in-degree 0. Moreno depicted his graphs to emphasize inter-gender relations. Note only one edge from a boy to a girl.



# Today's agenda

- More network examples. Note: Any numbers being stated may not be very current. Social and information networks are usually very dynamic and the numbers change rapidly.
- What is this course about?
- Basic concepts in graph theory

# Romantic Relationships [Bearman et al, 2004]



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

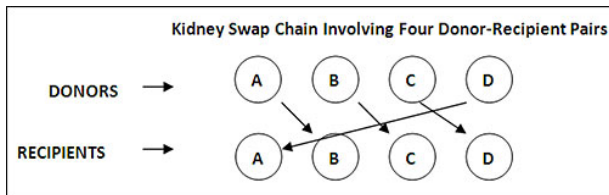
- Illustrates common “structural” properties of many networks
- What is the benefit of understanding this network structure?

## Kidney Exchange: Swap Chains

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

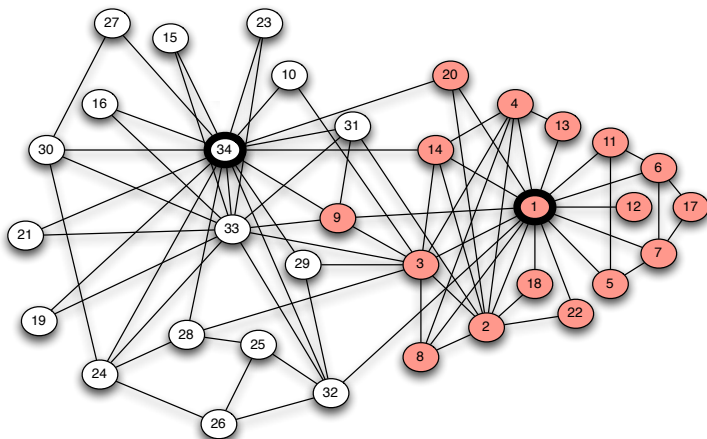
# Kidney Exchange: Swap Chains

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



**Figure:** Dartmouth-Hitchcock Medical Center, NH, 2010

## Communities: Karate club division



Karate Club social network, Zachary 1977

**Figure:** Karate club splits into two clubs

# Communities: 2004 Political blogosphere

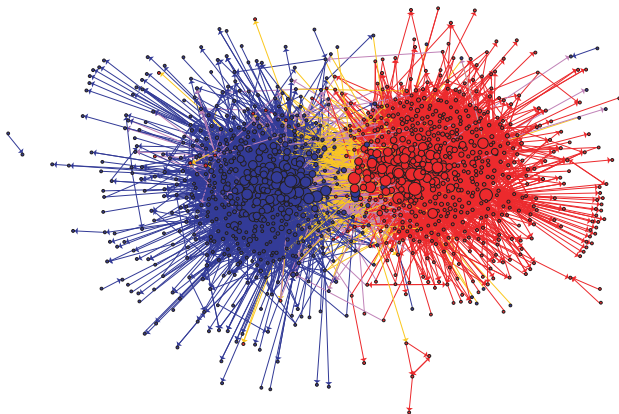
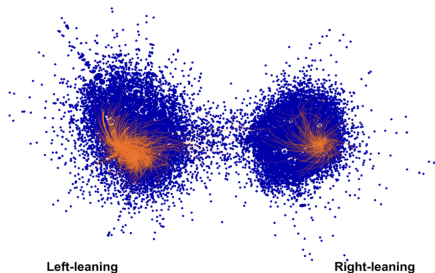


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones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

# 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.

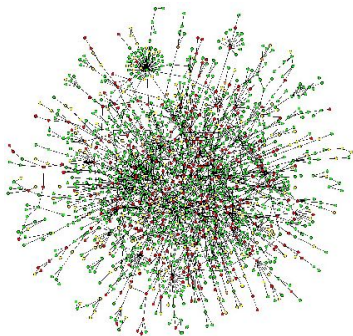
**Figure:** From Starbird et al [2017, 2019]

## Communities and hierarchical structure: Email communication



**Figure:** Email communication amongst 436 employees of Hewlett Packard Research Lab, superimposed on the Lab organizational hierarchy

# Protein-protein interaction network

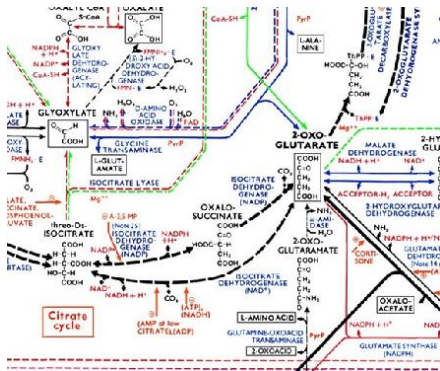


## Protein-Protein Interaction Networks

Nodes: Proteins

Edges: 'physical' interactions

# Metabolic network

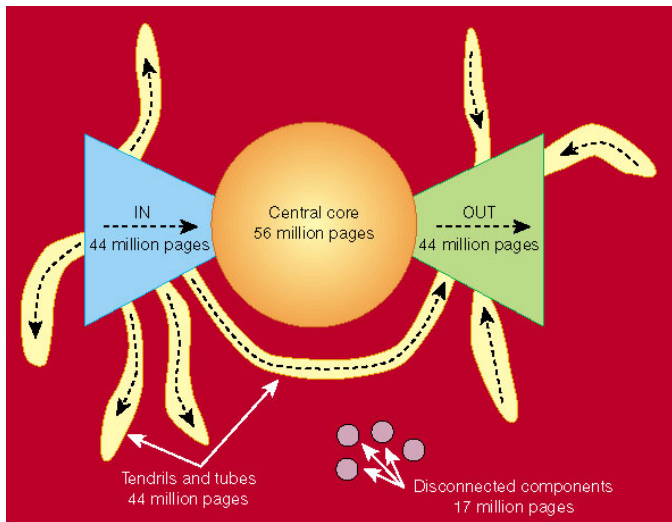


## Metabolic networks

Nodes: Metabolites and enzymes

Edges: Chemical reactions

# The web as a directed graph of hyperlinks



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

# Why study networks

- Clearly there are complex systems and networks that we are in contact with daily.
- The population of the world can be thought of as social network of approximately 7.75 billion people. AS OF January 2020, The people on Facebook are a *subnetwork* of approximately 2.45 billion active monthly users of which 1.62 billion are daily users. AS OF THIRD QUARTER 2019.
- 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 availability and ability to process massive amounts of data, makes computational aspects of networks essential.
- The current impact of social and information networks will almost surely continue to escalate (even if Facebook and other social networks are under increasing pressure to protect privacy and eliminate “bad actors”).

# What can one accomplish by studying networks

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

In studying social and information networks we can hopefully

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

## And how do we accomplish stated goals

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

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

What are the challenges?

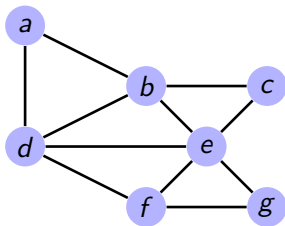
- Real world data is sometimes hard to obtain. For example, search engine companies treat much of what they do as proprietary.
- Many graph theory problems are known to be computationally difficult (i.e., *NP* hard) and given the size of many networks, results can often only be approximated and even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

# Network concepts used in this course

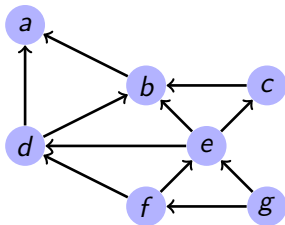
- Two main mathematical subjects of primary relevance to this course:
  - ① graph theoretic concepts
  - ② probability
- In motivating the course, we have already seen a number of examples of networks and hinted at some **basic graph-theoretic concepts**. We will now continue that discussion (i.e. material from Chapter 2 of the text) and for part of the next lecture before moving on to Chapter 3.
- We use the previous examples and some new ones to illustrate the basic graph concepts and terminology we will be using.

# Graphs: come in two varieties

- 1 undirected graphs (“graph” usually means an undirected graph.)

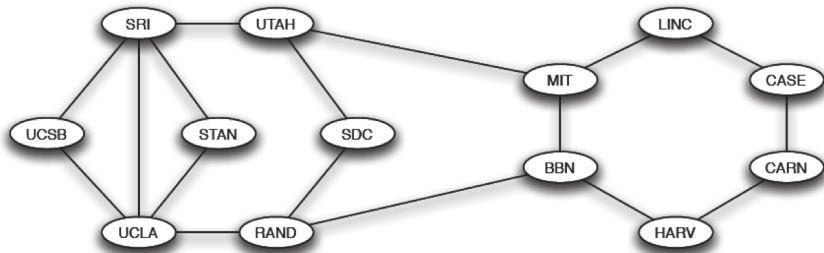


- 2 directed graphs (often called di-graphs).



# Visualizing Networks as Graphs

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



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

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

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

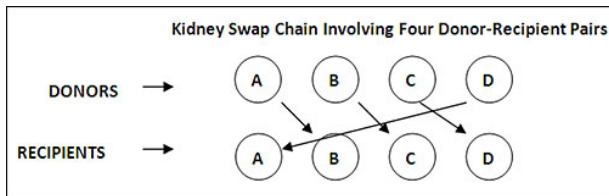
- This **node induced subgraph** is a 6 node **regular graph** of **degree 2**. It is a **simple graph** in that there are no self-loops or multiple edges.
- Note that the adjacency matrix of an (undirected) simple graph is a symmetric matrix (i.e.  $A_{i,j} = A_{j,i}$ ) with  $\{0,1\}$  entries.
- To specify distances, we would need to give weights to the edges to represent the distances. (As you will see, I will use both edges and links as terminology.)

## Kidney Exchange: Swap Cycles

- Live kidney donation common in N.A. 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

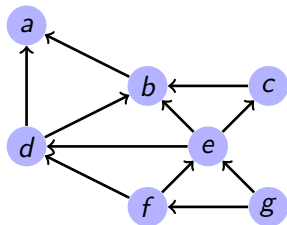
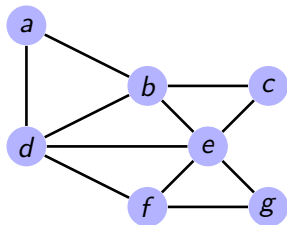
# Kidney Exchange: Swap Cycles

- Live kidney donation common in N.A. 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

## Recall: undirected graphs vs. directed graphs



# 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

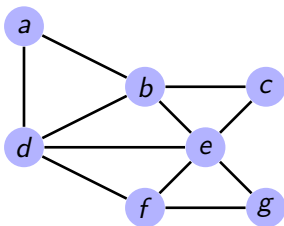
- First start with **undirected** graphs  $G = (V, E)$ .
- A **path** between two nodes, say  $u$  and  $v$  is a sequence of nodes, say  $u_1, u_2, \dots, u_k$ , where for every  $1 \leq i \leq k - 1$ ,
  - ▶ the pair  $(u_i, u_{i+1})$  is an edge in  $E$ ,
  - ▶  $u = u_1$  and  $v = u_k$

## Basic definitions continued

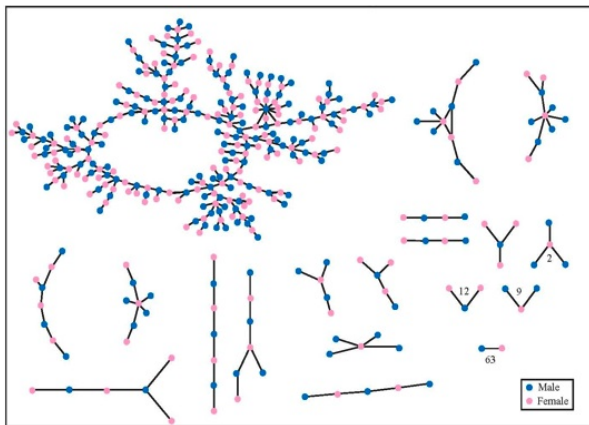
- First start with **undirected** graphs  $G = (V, E)$ .
- A **path** between two nodes, say  $u$  and  $v$  is a sequence of nodes, say  $u_1, u_2, \dots, u_k$ , where for every  $1 \leq i \leq k - 1$ ,
  - ▶ the pair  $(u_i, u_{i+1})$  is an edge in  $E$ ,
  - ▶  $u = u_1$  and  $v = u_k$
- The **length**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

## Basic definitions continued

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



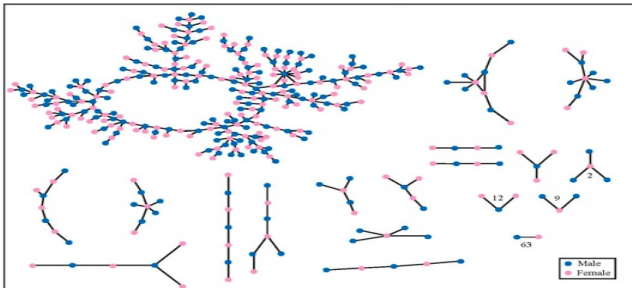
# Romantic Relationships [Bearman et al, 2004]



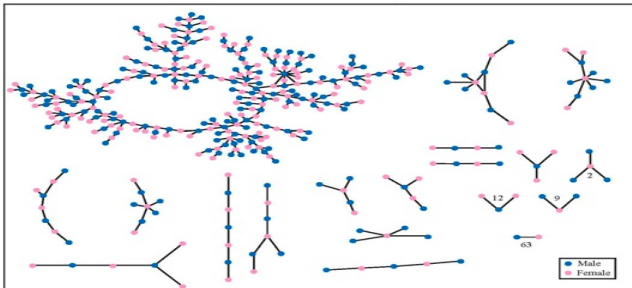
**Figure:** Dating network in US high school over 18 months.

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

# More basic definitions



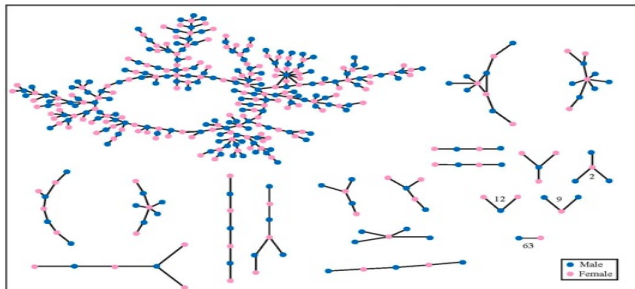
# More basic definitions



## Observation

Many **connected components** including one “**giant component**”

# More basic definitions



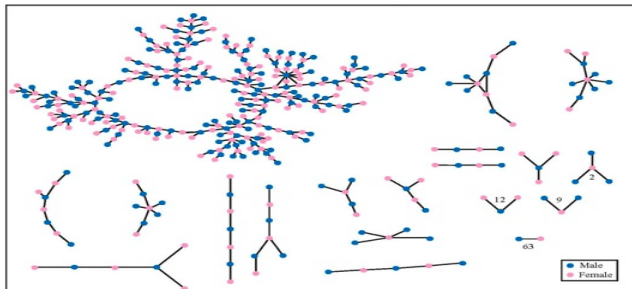
## Observation

Many **connected components** including one “**giant component**”

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

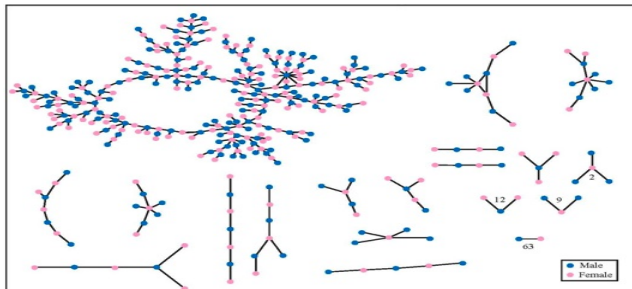
# Simple paths and simple cycles

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



# Simple paths and simple cycles

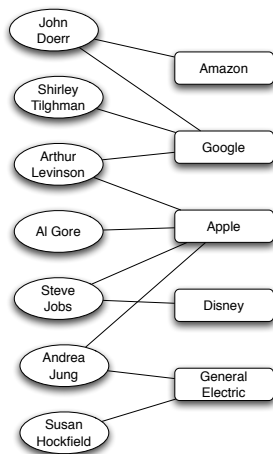
- Usually only consider **simple paths** and **simple cycles**: **no repeated nodes** (other than the start and end nodes in a simple cycle.)



## Observation

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

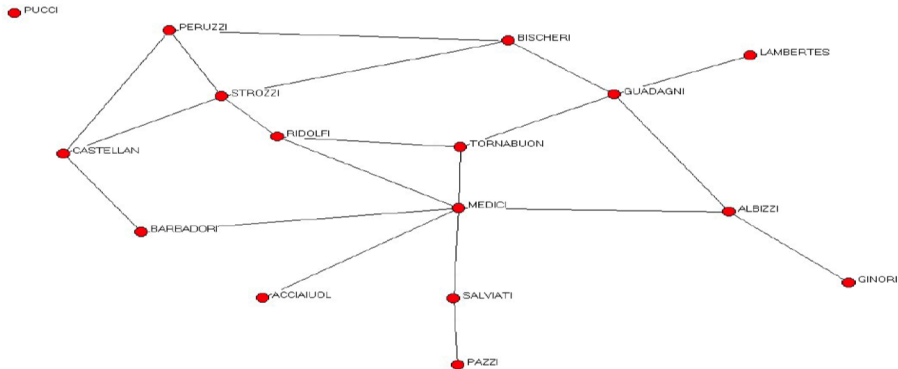
## Example of an acyclic bipartite graph



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

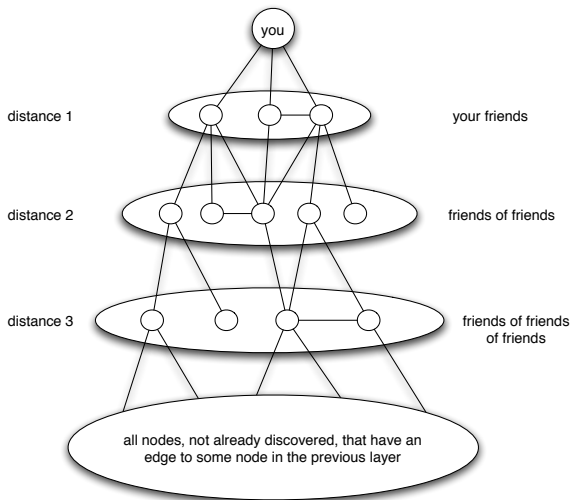
# Florentine marriages and “centrality”

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



**Figure:** see [Jackson, Ch 1]

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



**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 like that of mathematicians or actors. For mathematicians (or more generally say scientists) we co-authorship on a published paper. For actors, we can form a collaboration network where an edge represents actors performing in the same movie. For mathematicians one considers their Erdos number which is the length of the shortest path to Paul Erdos. 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>]

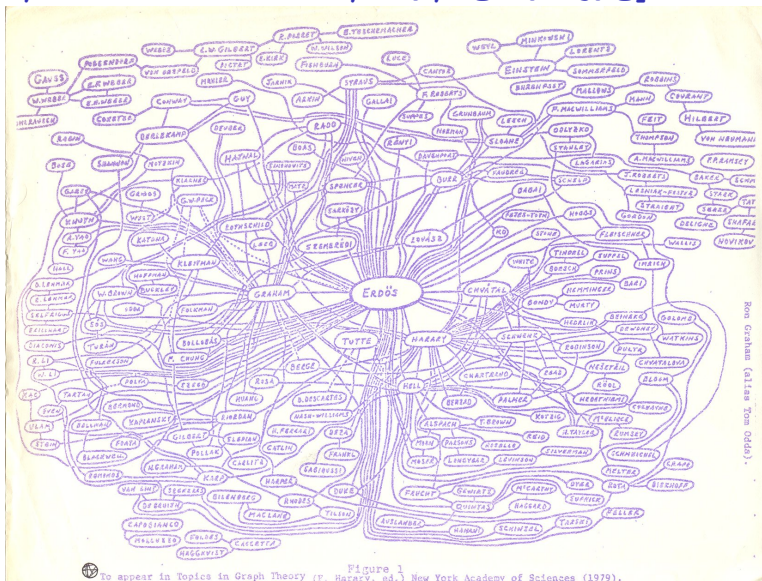


Figure 1  
To appear in Topics in Graph Theory (F. Harary, ed.), New York Academy of Sciences (1979).

## Analogous concepts for directed graphs

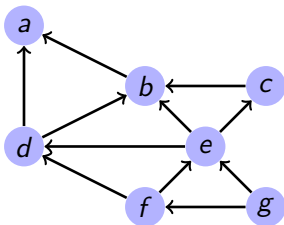
- We use the same notation for directed graphs, i.e. denoting a di-graph as  $G = (V, E)$ , where now the edges in  $E$  are **directed**.

# 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  $(u, v)$ .

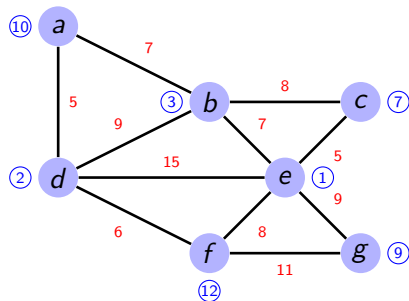
# Analogous concepts for directed graphs

- 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**.



# Weighted graphs

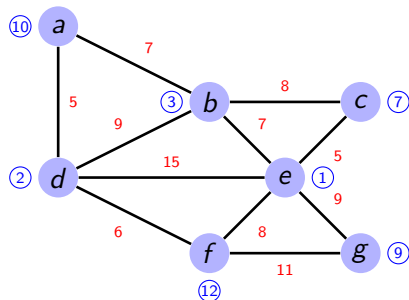
- We will often consider **weighted graphs**. Lets consider a (directed or undirected) graph  $G = (V, E)$ . Example:



- ▶ **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:

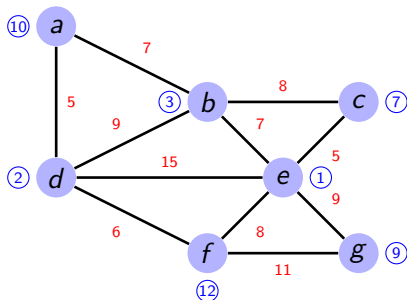


- ▶ **red numbers**: edge weights
- ▶ **blue numbers**: vertex weights

- 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:

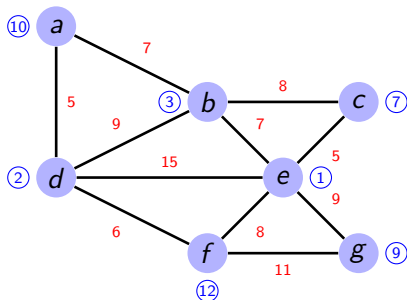


- ▶ **red numbers**: edge weights
- ▶ **blue numbers**: vertex weights

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

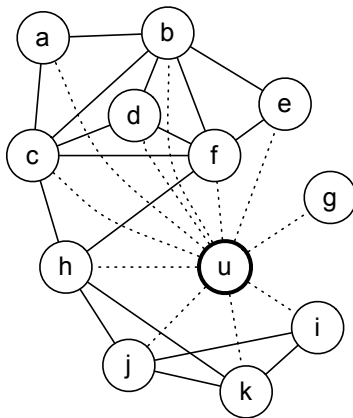
# Detecting the romantic relation in Facebook

- As previously mentioned, 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 and (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 (continued)

- They consider 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  $u$  and  $v$  such that  $(s, t) \notin E$  and  $s, t$  have no common neighbours except for  $u$  and  $v$ .
- They also 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.

## 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 users, 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 users, a random guess would have prediction accuracy of  $1/200 = .5\%$
- Various dispersion measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. 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 users, a random guess would have prediction accuracy of  $1/200 = .5\%$
- Various dispersion measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. **Why would high dispersion be a better measure than high embeddedness?**
- By itself, dispersion outperforms various interaction 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 users, a random guess would have prediction accuracy of  $1/200 = .5\%$
- Various dispersion measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. **Why would high dispersion be a better measure than high embeddedness?**
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.

## 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 users, a random guess would have prediction accuracy of  $1/200 = .5\%$
- Various dispersion measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. **Why would high dispersion be a better measure than high embeddedness?**
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- 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 users, a random guess would have prediction accuracy of  $1/200 = .5\%$
- Various dispersion measures do better than the embeddedness measure in its ability to predict the correct romantic relationship. **Why would high dispersion be a better measure than high embeddedness?**
- By itself, dispersion outperforms various interaction features.
- For most measures, performance is better for male users and also better for data when restricted to marriage as the relationship.
- 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.**

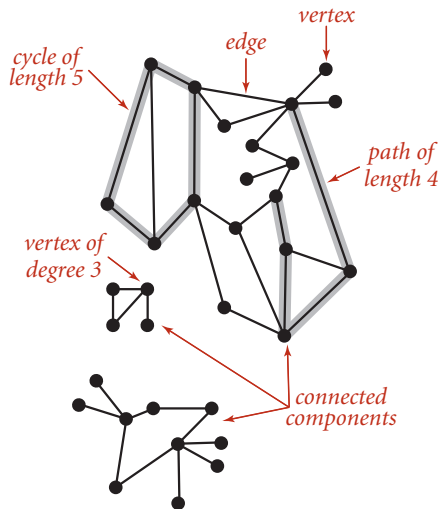
# Some experimental results for the fraction of correct predictions

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

# Graph Anatomy: summary thus far

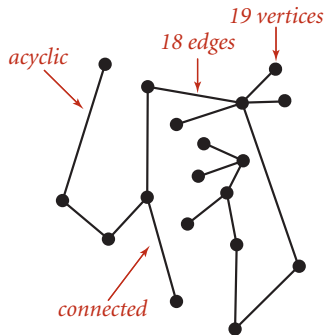


[from Algorithms, 4th Edition by Sedgewick and Wayne]

# Acyclic graphs (forests)

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

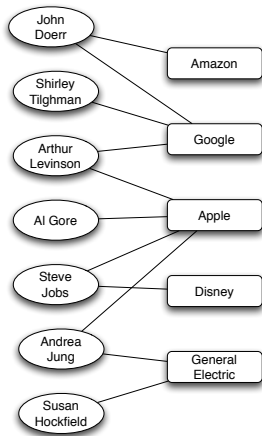
- ▶ A tree is a **connected acyclic** graph.
- ▶ **Question:** Why are such graphs called trees?
- ▶ **Fact:** There are always  $n - 1$  edges in an  $n$  node tree.



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

## Another tree [E&K Figure 4.4]

- The bipartite graph from last class (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?

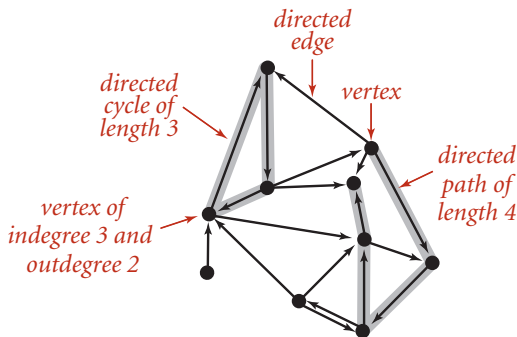


### Facts

- 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.

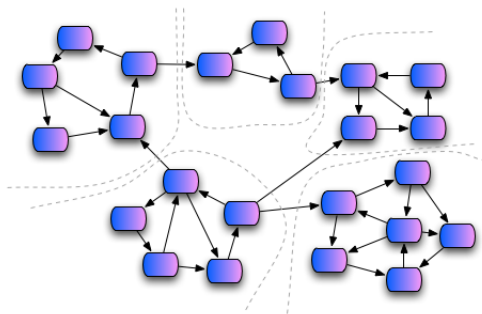


**Figure:** Directed graph anatomy [from Sedgewick and Wayne]

# 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?