Great Ideas in Computing

University of Toronto CSC196
Fall 2022

Week 11: November 28-December 2

Announcements

Announcements

- I hope everyone did well on the second quiz.
- You might find the following November 2020) article interesting about super infection spreaders and how it relates to the Barabasi and Albert model we will be discussing in regard to social networks. https://www.wired.com/story/covid-19-vaccine-super-spreaders/
- The final assignment is due December 5 at 8AM.

This weeks agenda

- We will first finish up complexity based cryptograph and then try to get to social networks quickly.
- We begin a discussion of graphs/networks in general and social networks in particular.
- The social network question on Assignment A4 is mainly a thought question so you should be able to provide reasonable answers based on your own experience and the W11 slides.
  If you have any question about graph concepts and social networks raise them in class, or on piazza.
- The undergraduate Easley and Kleinberg textbook "Networks, Crowds, and Markets: Reasoning about a Highly Connected World" is an excellent text for understanding the importance of network concepts and applications. We teach an undergraduate course CSC303 devoted to social networks.

WARNING: Real world cryptography is sophisticated

Complexity based cryptography requires careful consideration of the definitions and what precise assumptions are being made.

Complexity based cryptography has led to many important practical protocols and there are a number of theorems. Fortunatley, many complexity assumptions turn out to be equivalent.
In the Rackoff notes, the following theorem is stated as the fundamental theorem of cryptography. (To make this result precise, one needs precise definitions which we are omitting.)
Theorem: The following are equivalent:

- It is possible to do "computationally secure sessions"
- There exists pseudo-random generators; that is, create strings that computationally look random)
- There exist one way functions f; that is functions such that $f(x)$ is easy to compute but given $f(x)$ it is hard to find a z such that $f(z) = f(x)$. Here "hard to find" means not computable in polynomial time.
- There exist computationally secure digital signature schemes.

The discrete log function

RSA is based on the assumed difficulty of factoring. Another assumption that is widely used in cryptography is the discrete log function. Again, we need some facts from number theory.

Let p be a large prime.

- $\mathbb{Z}_p^*$ denotes the set of integers $\{1, 2, \ldots, p-1\}$ under the operations of $+, -, \cdot \mod p$ is a *field*. In particular, for every $a \in \mathbb{Z}_p^*$, there exists a $b \in \mathbb{Z}_p^*$ such that $a \cdot b = 1$; i.e., $b = a^{-1} \mod p$.
- Moroever, $\mathbb{Z}_p^*$ is *cyclic*. That is, there exists a $g \in \mathbb{Z}_p^*$ such that $\{1, g, g^2, g^3, \ldots g^{p-2}\} \mod p = \mathbb{Z}_p^*$. Recall, as a special case of the Euler totient function, $a^{p-1} = 1 \mod p$.

The assumption is that given $(g, p, g^x \mod p)$, it is computationally difficult to find x. This is another example (factoring can also be an example) of a *one-way function*. In fact the discrete log function is a *one-way permutation*.

A pseudo random generator

We started off our discussion of complexity based cryptpgraphy by noting that randomness is essential. We have also noted that it is not clear (or at what cost) one can obtain strings that "look like" truly random strings.

A pseudo random generator G is a *deterministic* function $G : \{0,1\}^k \to \{0,1\}^\ell$ for $\ell > k$. When ℓ is exponential in k, G is called a pseudo random function generator. For now, lets even see how to be able to have $\ell = k + 1$.

The random input string $s \in \{0,1\}^k$ is called the seed and the goal is that $r = G(s)$ should be "computationally indistinguishable" from a truly random string in $t = \{0,1\}^\ell$. This means that no polynomial time algorithm can distinguish between r and t with probability better than $\frac{1}{2} + \epsilon$ for any $\epsilon > 0$. (Here I am being sloppy about the quantification but hopefully the idea is clear.)

A pseudo random generator continued

On the previous slide there was a claim that having a pseudo random generator is equivalent to having a one-way function.

How can we use (for example, the assumption that the discrete log function is a one-way function) to construct a pseudo random generator with $\ell = k + 1$.

The Blum-Micali generator. Assumming the discrete log function is a one-way function then the following is a pseudo random generator:

Let x_0 be a random seed in $\mathbb{Z}_p^*$ by interpeting $(s_1, \ldots, s_k)_2$ as a binary number mod p. Let $x_{k+1} = g^{x_k} \bmod p$. Define $s_{k+1} = 1$ if $x_k \leq \frac{p-1}{2}$.

Manual Blum won the Turing award for his contributions to cryptography and Silvio Micali (along with Shafira Goldwasser) won the Turing award for *interactive zero knowledge proofs*. (Note: The authors on the seminal zero knowledge paper are Goldwasser, Micali, and Rackoff where I am noting that Charlie Rackoff is a UT DCS Professor Emeritus.)

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, . . .

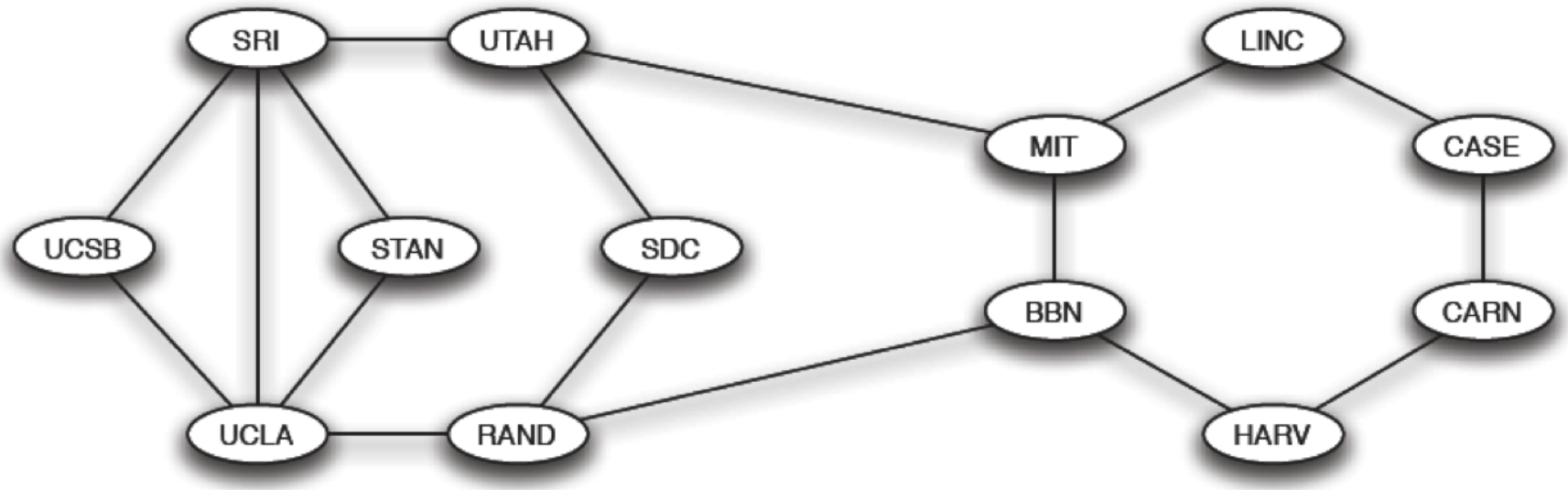
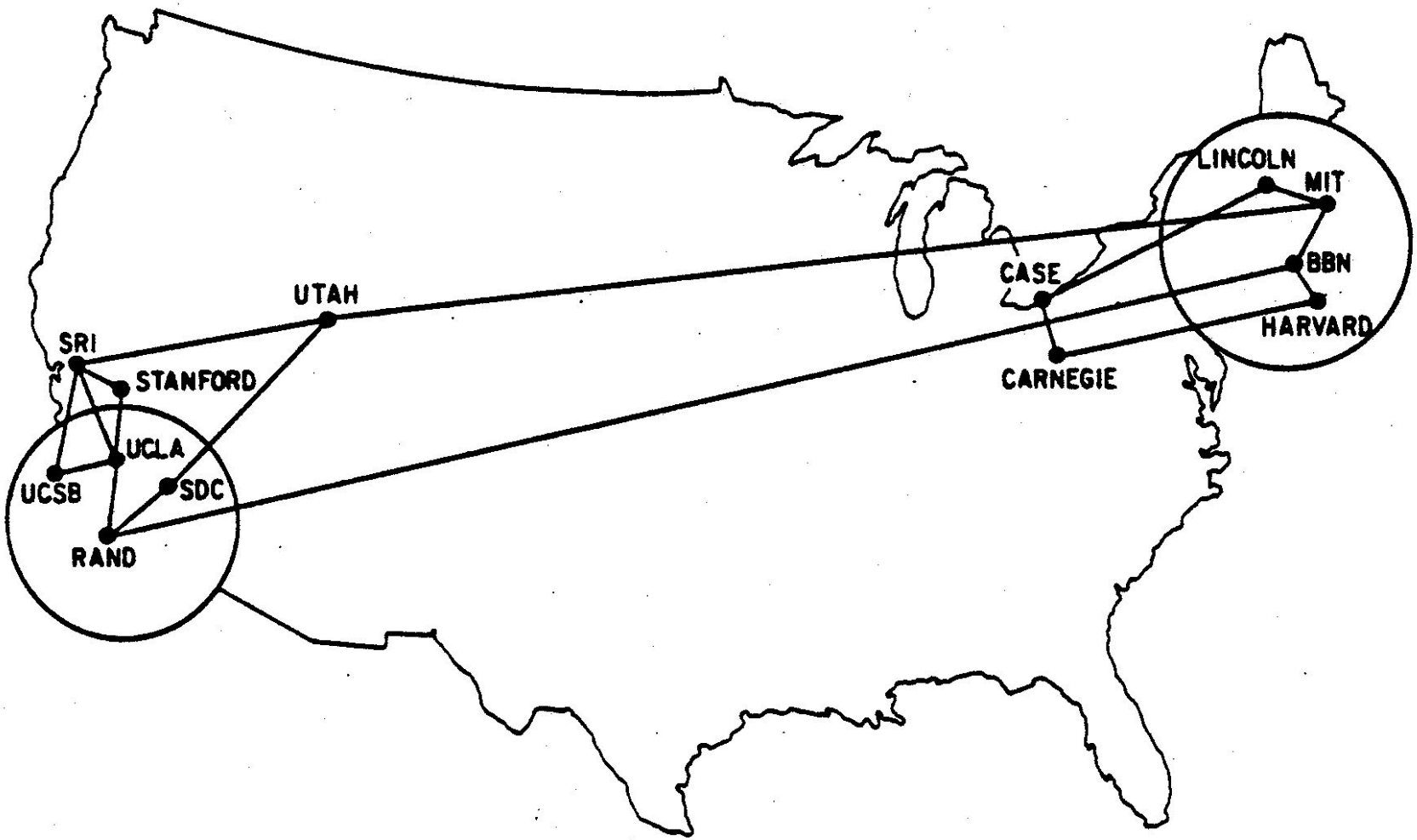


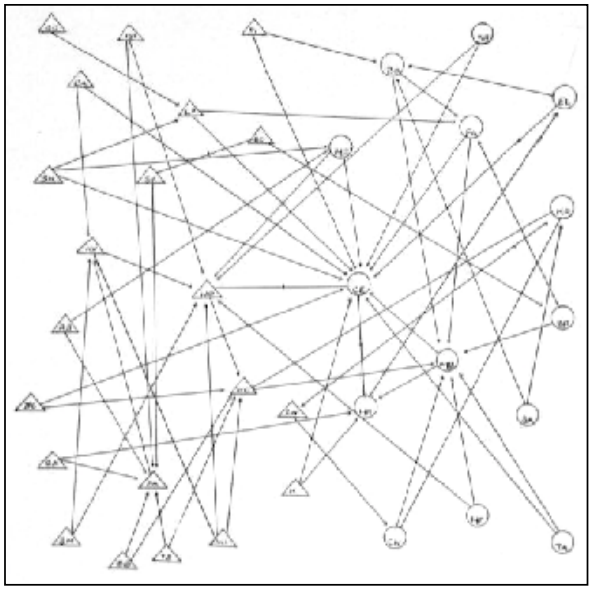
Figure: Initial internet: Dec. 1970 [E&K, Ch.2]

December 1970 internet visualized geographically [Heart et al 1978]

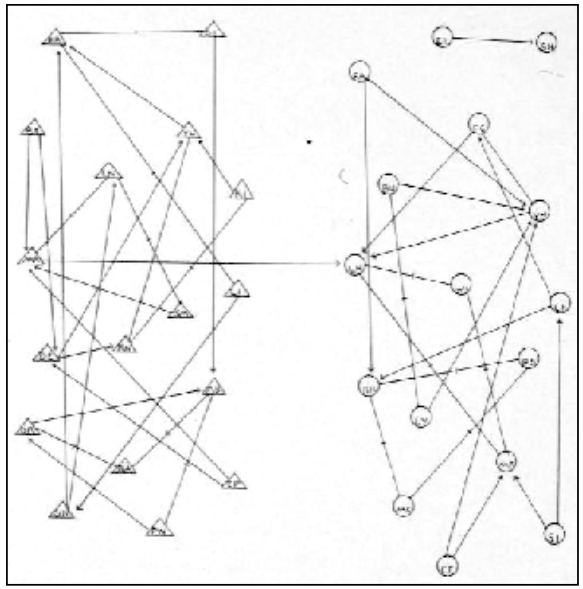


The first social network analysis

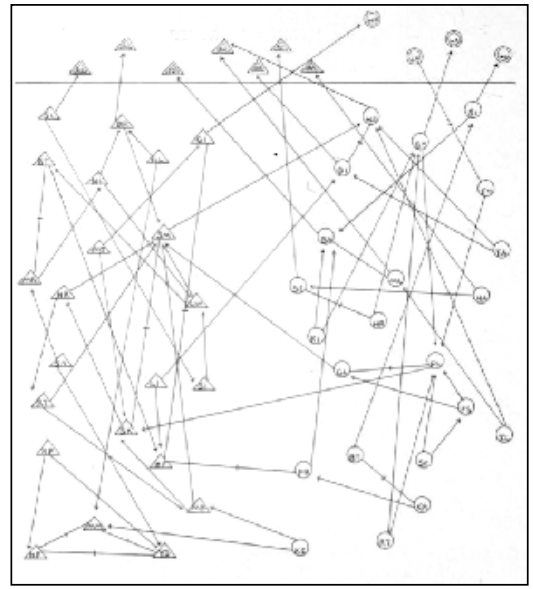
In his **1934** book *Who Shall Survive: A New Approach to the Problem of Human Interrelations*, Jacob Moreno (Romanian-US psychiatrist) introduced *sociograms* and used these graphs/networks to understand relationships. In one study (that was repeated to test changes) he asked each child in various 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

Moreno's sociogram 1934

A closer look at grade 1 in Moreno sociogram

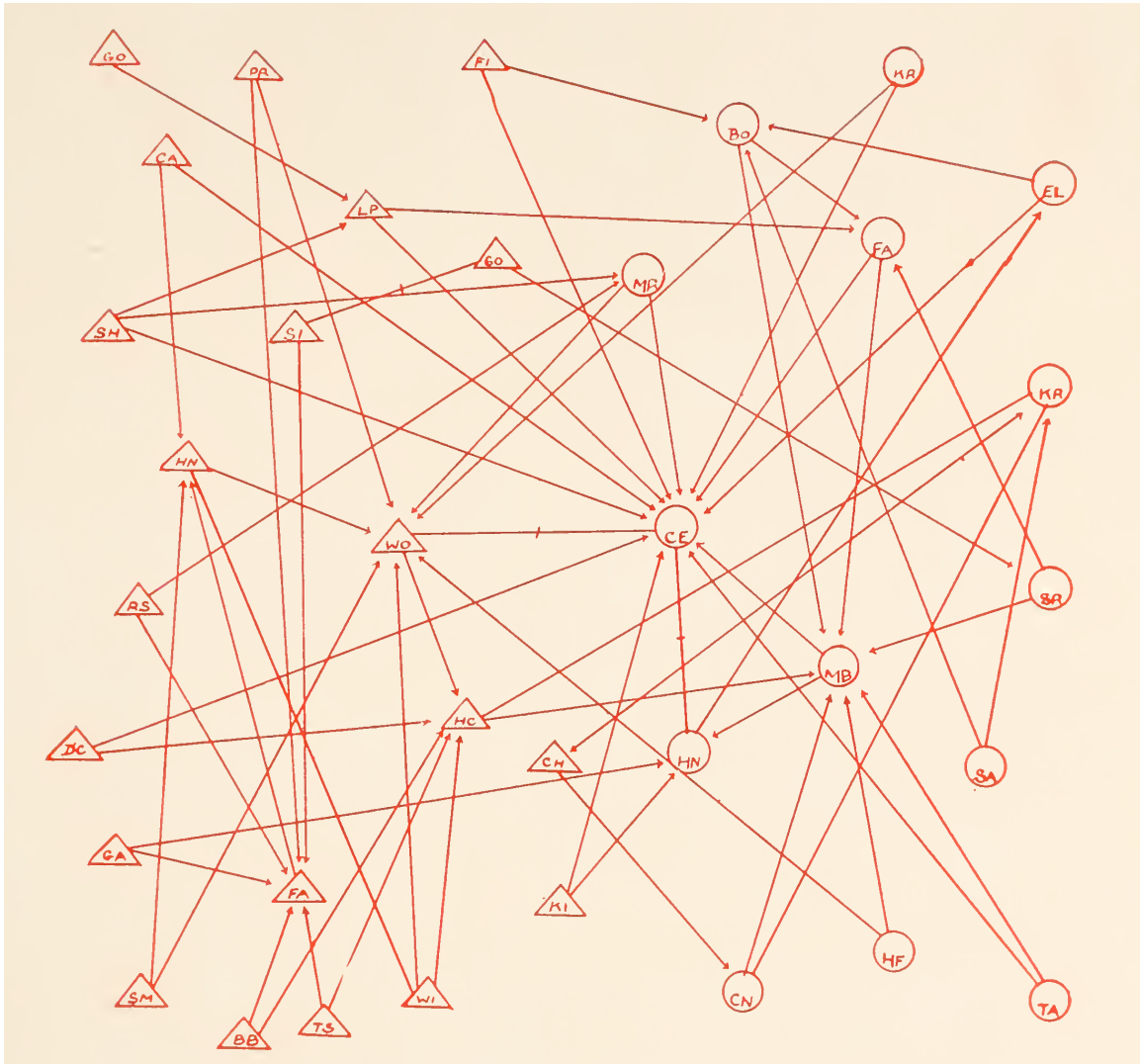


Figure: 21 boys, 14 girls. Directed graph. Every node has out-degree 2. 18 unchosen having in-degree 0. Note also that there are some "stars" with high in-degree.

A closer look at grade 4 in Moreno sociogram

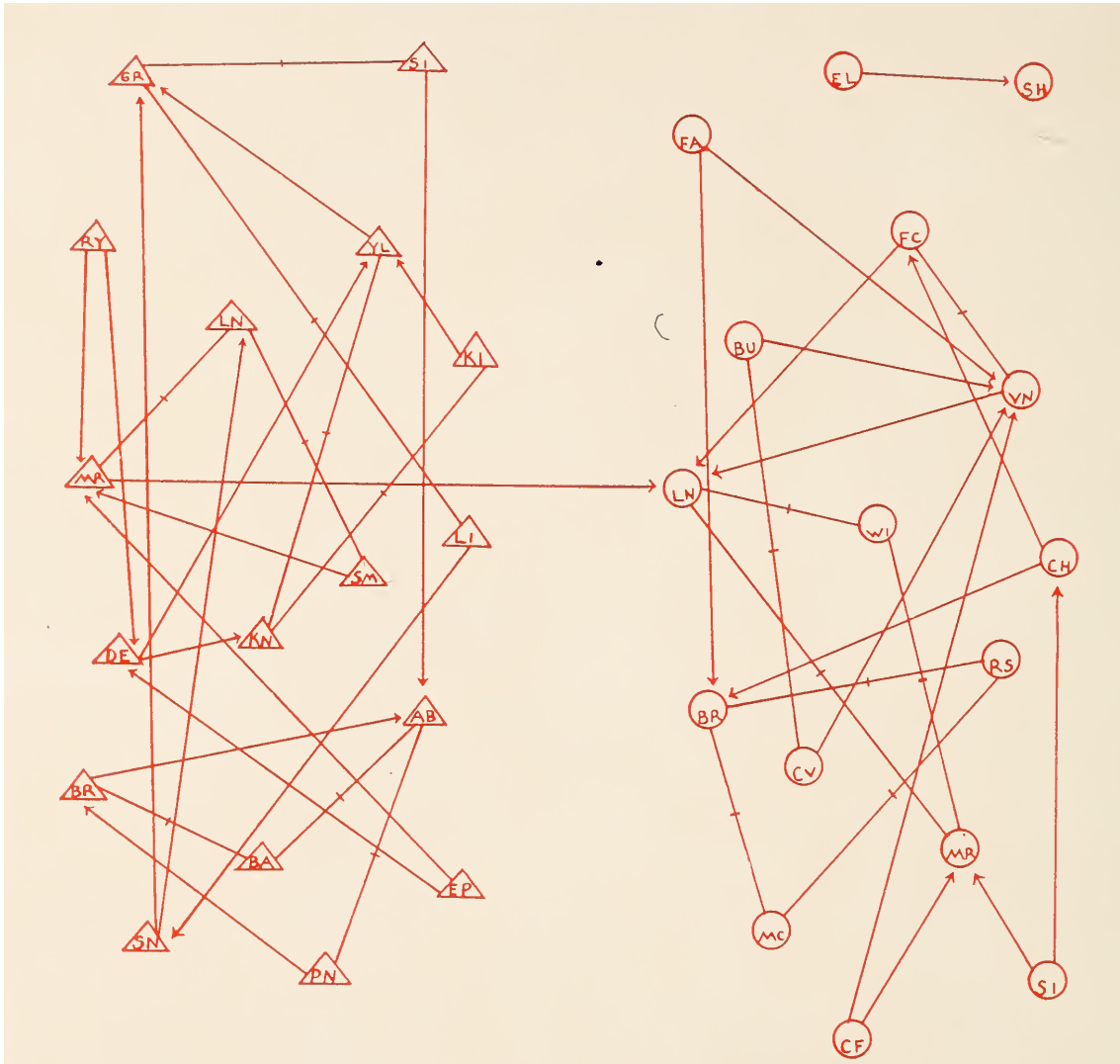


Figure: 17 boys, 16 girls. Directed graph with 6 unchosen having in-degree 0. Moreno depicted his graphs to emphasize inter-gender relations. Note only one edge from a boy to a girl.

A closer look at grade 8 in Moreno sociogram

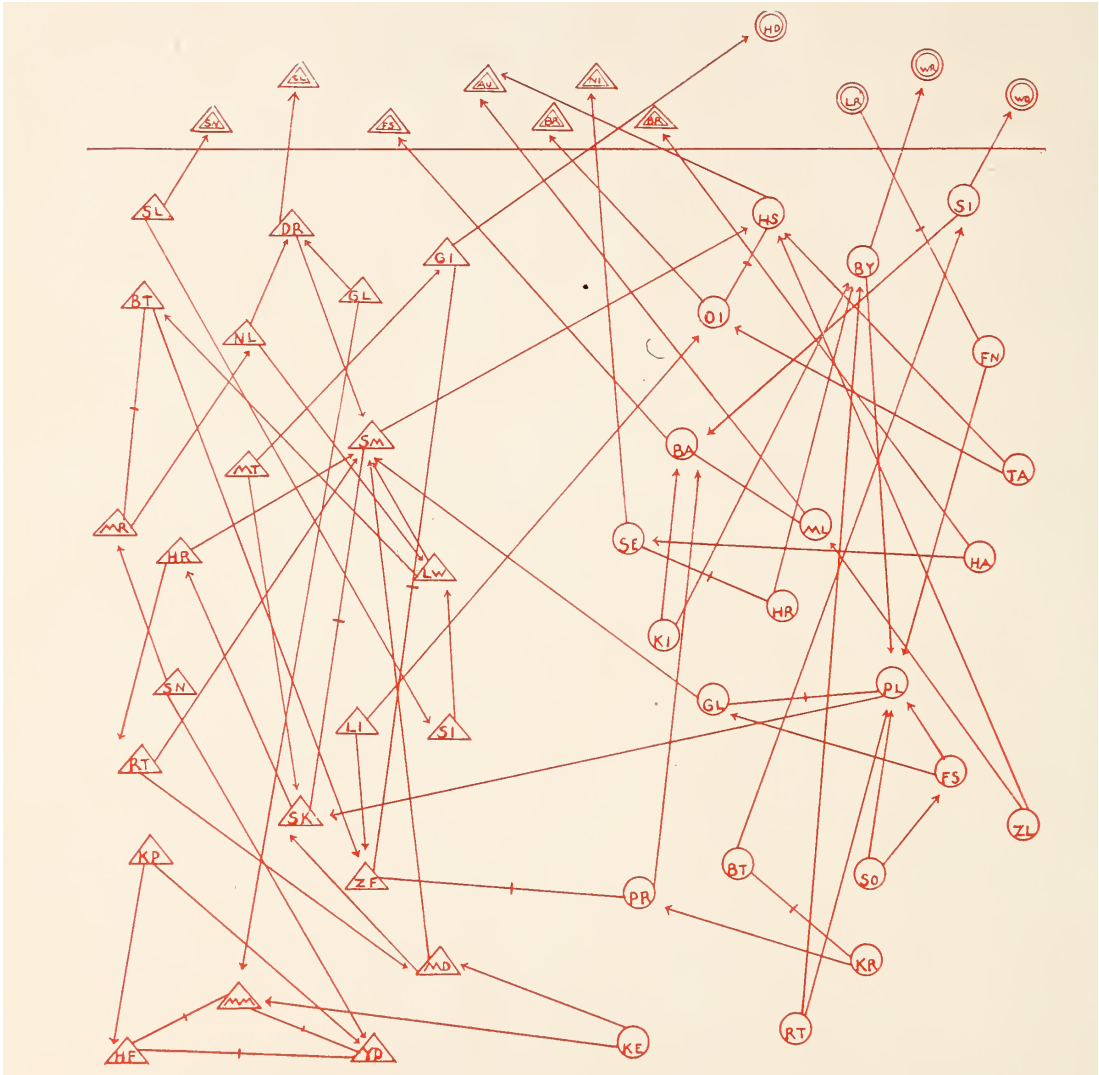


Figure: 22 boys, 22 girls. Directed graph with 12 unchosen having in-degree 0. Some increase in inter-gender relations. Double stars and circles above line indicte different "groups".

Romantic Relationships [Bearman et al, 2004]

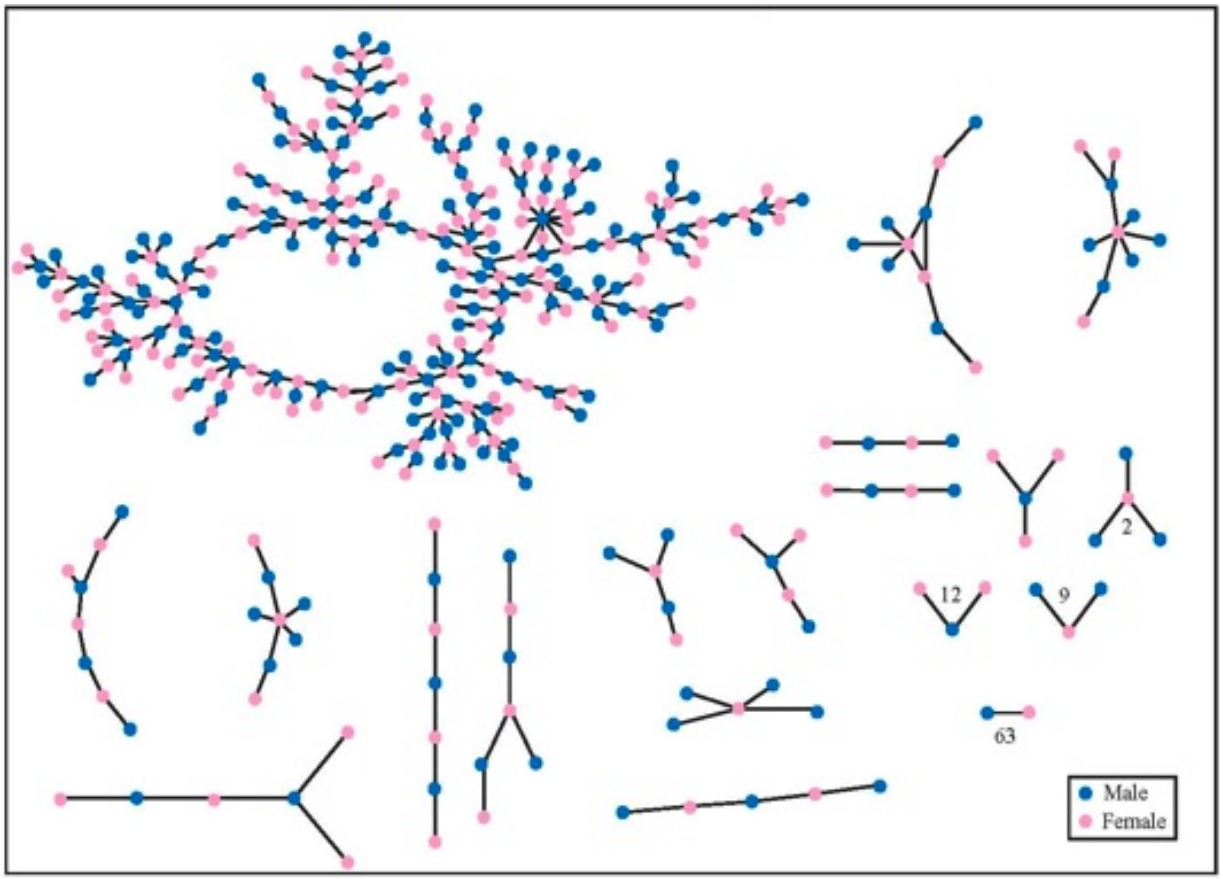


Figure: Dating network in US high school over 18 months.

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

Kidney Exchange: Swap Chains

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

Kidney Exchange: Swap Chains

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

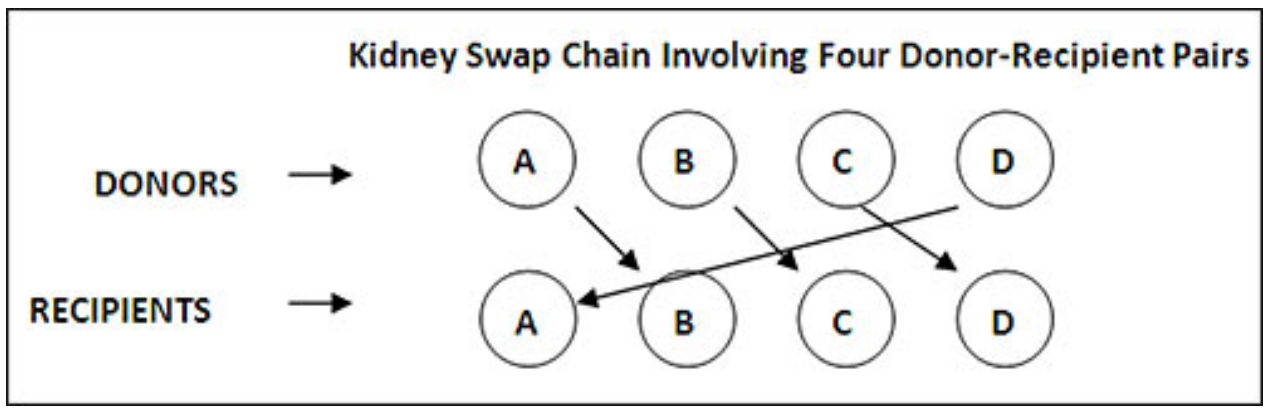
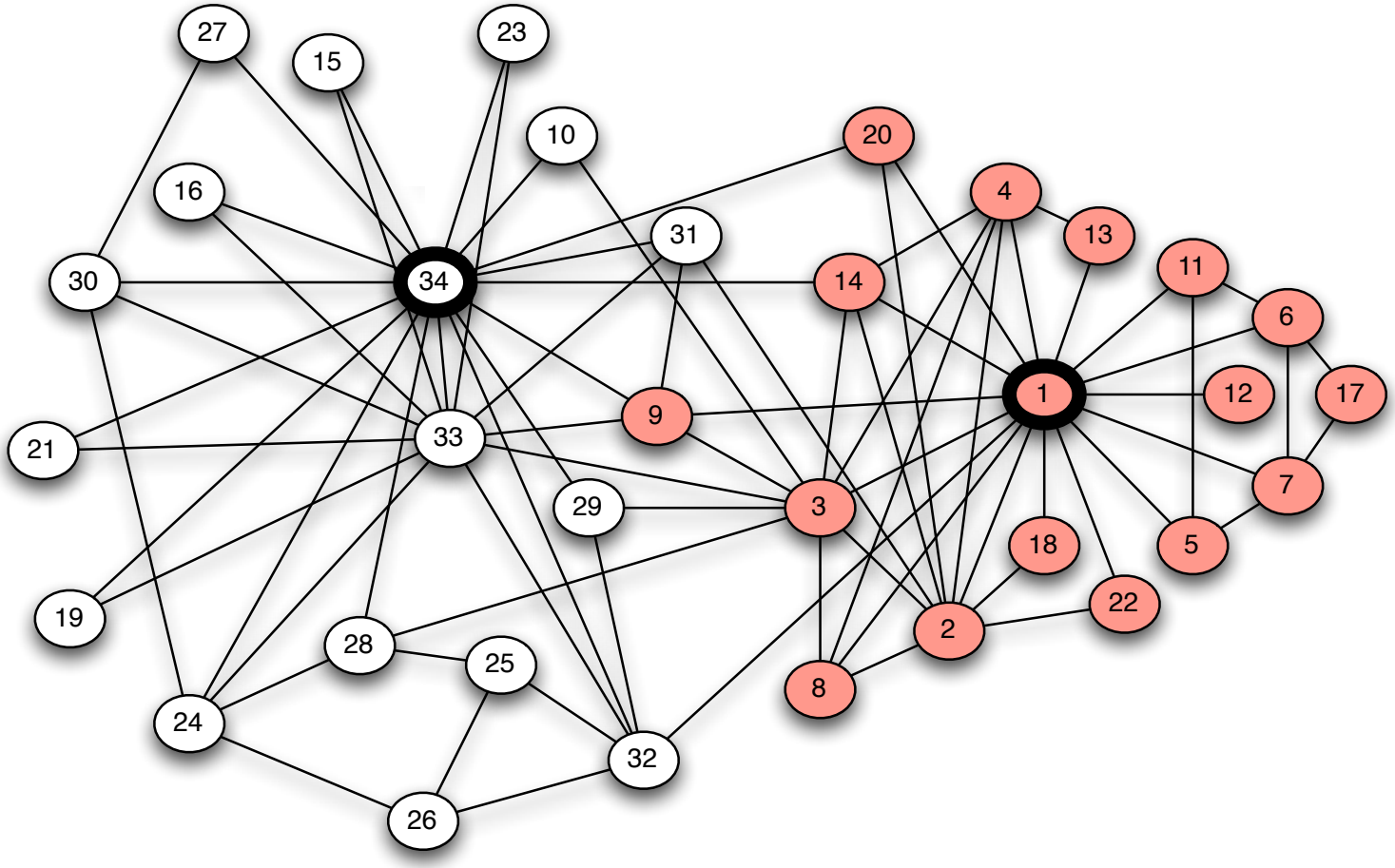


Figure: Dartmouth-Hitchcock Medical Center, NH, 2010

Communities: Karate club division



Karate Club social network, Zachary 1977

Figure: Karate club splis into two clubs (or *communities*)

Communities: 2004 Political blogsphere

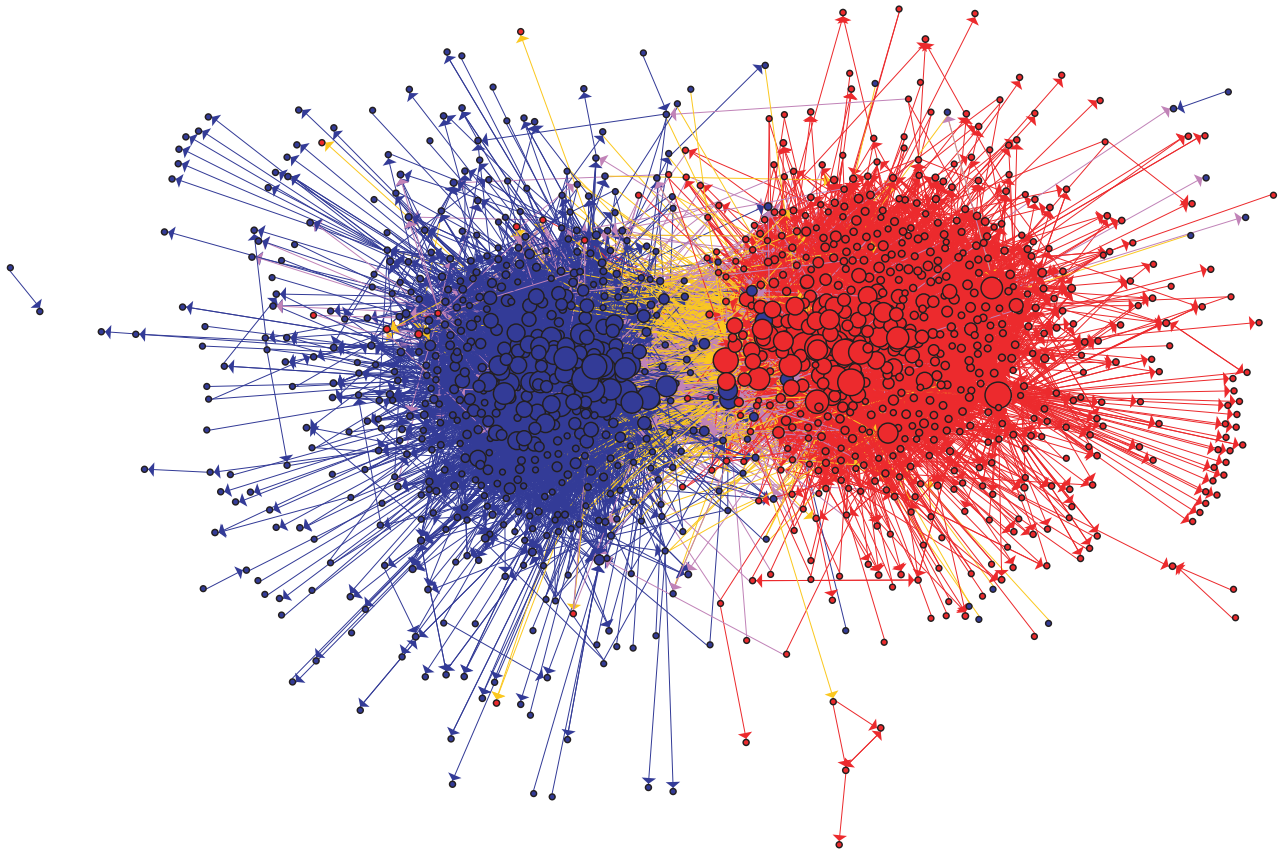


Figure 1: Community structure of political blogs (expanded set), shown using utilizing a GEM layout [11] in the GUESS[3] visualization and analysis tool. The colors reflect political orientation, red for conservative, and blue for liberal. Orange links go from liberal to conservative, and purple ones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

Communities: 2017 Twitter online discourse regarding Black Lives Matter

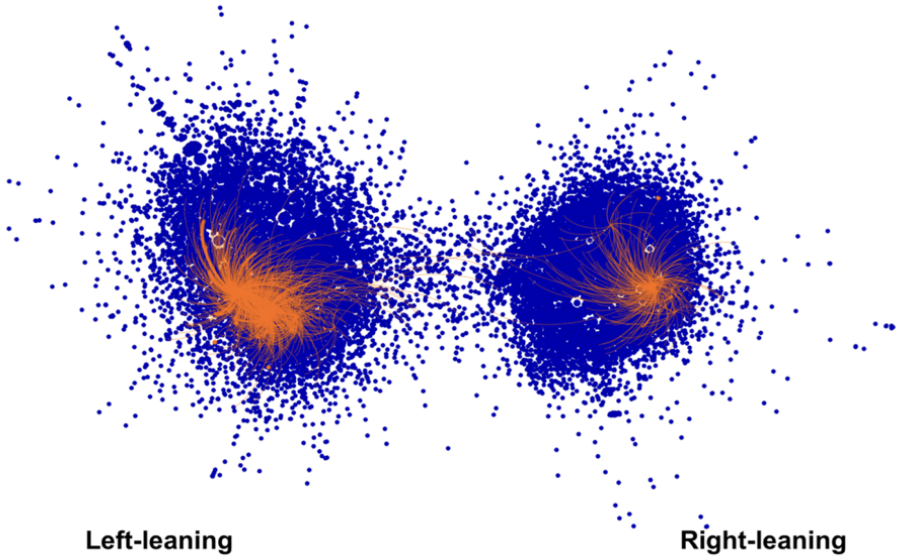


Fig. 1. Retweet Network Graph: RU-IRA Agents in #BlackLivesMatter Discourse. The graph (originally published [3]) shows accounts active in Twitter conversations about #BlackLivesMatter and shooting events in 2016. Each node is an account. Accounts are closer together when one account retweeted another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

The current interest in networks

- Clearly there are complex systems and networks that we are in contact with daily.
- The population of the world can be thought of as social network of approximately 7.8 billion people. AS of January 2020, The people on Facebook are a *subnetwork* of approximatley 2.9 billion active monthly users of which 1.6 billion are daily users. (Different numbers are reported in different sites.)
- 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 ability to store and 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 presure to protect privacy, eliminate "bad actors", and eliminate "divisive policies").

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 neceessarily 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.
- Understand how special interests can target information and misinformation to selected "communities"

And how do we accomplish stated goals

Much of what people do in this research field is empirical analysis. Researchers formulate network models, hypotheses and predictions and then compare against the 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. Like search enginess, social networks 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.

Social networks

A social network is a network $G = (V, E)$ where the nodes in V are people or organizations. Social networks can be undirected or directed networks.

The edges can be relations between people (e.g. friendship) or membership of an individual in an organization.

Social networks can be of any size (e.g., a small network like the Karate Club on slide 18) or enormous networks like Facebook and Twitter. We usually think of Facebook as an undirected graph (where *friendship* is an undirected edge) and Twitter as a directed graph (i.e., where *follows* is a directed edge).

Understanding how networks evolve, the resulting structure of social networks, and computational aspects for dealing with large networks is an active field of study in CS as well as in sociology, political science, economics, epidemiology, and any field that studies human behaviour. J. Kleinberg's 2000 analysis with regard to the six degrees of separation phenomena is an early result that sparked interest in algorithmic aspects of social networks.

The computational challenge presented by super large networks

The size of some modern networks such as the web and social networks such as Facebook are at an unprecedented scale.

As of Februay, 2022, xThe average facebook user has about 155 friends which then implies about $2.9 \cdot \frac{155}{2} \approx 200$ billion edges. It is interesting to note that 90% of daily active users are outside USA and Canada. See https://www.omnicoreagency.com/facebook-statistics/ // for lots of interesting demographic and other facts about Facebook.

What does this imply for the complexity of algorithms involving such super large networks?

Linear is the new exponential

In complexity theory (e.g. in the P vs NP issue that we will be discussing) we say (as an abstraction) that polynomial time algorithms are "efficient" and "exponential time" is infeasible. There are, of course, exceptions but as an abstraction this has led to invaluable fundamental insights.

As problem instances have grown, there was a common saying that "quadratic (time) is the new exponential".

But with the emergence of networks such as the web graph and the Facebook network, we might now say that "linear is the new exponential" when it comes to extracting even the most basic facts about these networks. For example, how do we even estimate the average node degree in a giant network?

There are many facts about large networks that we would like to extract from the network. For example, how do we find "influential" or "interesting nodes" in a social network?

Sublinear time algorithms

What is sublinear time?

In general when we measure complexity, we do so as a funtion of the input/output size. For graphs $G = (V, E)$, the size of the input is usually the number of edges E. (An exception is that when the graph is presented say as an adjacency matrix, the size is n^2 where $n = |V|$.)

Since our interest is in massive information and social networks, we consider sparse graphs (e.g. average constant degree) so that $|E| = O(|V|)$ and hence we will mean sublinear time as a function of n. The desired goal will be time bounds of the form $O(n^\alpha)$ with $\alpha < 1$ and in some cases maybe even $O(\log n)$ or $polylog(n)$.

Given that optimal algorithms for almost any graph property will depend on the entire graph, we will have to settle for approximations to an optimum solution. Furthermore, we will need to sample the graph so as to avoid having to consider all nodes and edges. And we will need a way to efficiently access these massive graphs,

Coping with massive social graphs continued

One way to help coping with massive networks is to hope to utilize some substantial amount of parallelism. There is an area of current research concerning massive parallel computation (MPC) models where (in principle) we can achieve sublinear time by distributing computation amongst a large (i.e., non constant) number of processors.

But even if we could muster and organize thousands of machines, we will still need random samplng, approximation, and have highly efficient "local information algorithms".

Finally, in addition to random sampling and parallelism, we will have to hope that social networks have some nice structural properties that can be exploited to as to avoid complexity barriers that exist for arbitrary (sparse) graphs. These complexity barriers are hopefully clear from our discussion of complexity theory, *NP completeness* and *NP hardness*.

Preferential attachment models

Preferential attachnment models (also called "rich get richer" models) are probabilistic generative models explaining how various networks can be generated. Namely, after starting with some small graph, when we add a new node u, we create a number of links between u to some number m of randomly chosen nodes $v_1, v_2, \ldots, v_m$. The probability of choosing a v_i is proportional to the current degree of v_i. More generally, the probability of choosing a node v_i can be an increasing function of the degree,

These models have been used to help explain the structure of the web as well as social networks. Furthermore, networks generated by such a process have some nice structural properties allowing for substantially more efficient algorithms than one can obtain for arbitrary graphs.

For such models, there are both provable analytic results as well as experimental evidence on synthetic and real networks that support provable results that follow from the model. (Remember, a model is just a model and is not "reality"; as models are implifications of real networks, they may not account for many aspects in a real network. For example, in this basic model, all the edges for a new node are set upon arrival.

Consequences for networks generated by a preferential attachment process

There are many properties, believed and sometimees proven. about preferential attachment network models that do not hold for uniformly generated random graphs (e.g., create sparse graphs with constant average degree by choosing each possible edge with say probability proportional to $\frac{1}{n}$).

One of the most interesting and consequential proerties is that vertex degrees satisfy a *power law distribution* in expectation. Specifically, the expectation fraction $P(d)$ of nodes whose degree is d is proportional to $d^{-\gamma}$ for some $\gamma \geq 1$. Such a distribution is said to have a *fat tail*.

In a uniformaly random sparse graph (with average degree d_{avg}), with high probability , the fraction of nodes having a large degree $d > d_{avg}$ is proportional to c^{-d} for some $c > 1$.

The Barabasi and Albert preferential model

Barabasi and Albert [1999] specified a particular preferential attachment model and conjectured that the vertex degrees satisfy a power law in which the fraction of nodes having degree d is proportional to d^{-3}.

They obtained $\gamma \approx 2.9$ by experiments and gave a simple heuristic argument suggesting that $\gamma = 3$. That is, $P(d)$ is proportional to d^{-3}

Bollobas et al [2001] prove a result corresponding to this conjectured power law. Namely, they show for all $d \leq n^{1/15}$ that the *expected* degree distribution is a power law distribution with $\gamma = 3$ asymptotically (with n) where n is the number of vertices.

Note: It is known that an actual realized distribution may be far from its expectation, However, for small degree values, the degree distribution is close to expectation.

When we say that a distribution $P(d)$ is a power law distribtion this is often meant to be a "with high probaility" whereas results for networks generated by a preferential attachment process the power law is usually only in expectation.

Proven or observed properties of nodes in a social network generated by preferential attachment models

In addition to the power law phenomena suggesting many nodes with high degree, other properies of social networks have been obseerved such as a relatively large number of nodes u having some or all of properties such as the following: .

- high clustering coefficient defined as : $\frac{(u,v),(u,w),(v,w)\in E}{(u,v),(u,w)\in E}$. That is, mutual friend of u are likely to be friends.
- high centrality ; e,g, nodes on many pairs of shortest paths.

Brautbar and Kearns refer to such nodes (as above) as "interesting indiviudals" and these individuals might be candidates for being "highly influential individuals". Bonato et al [2015] refers to such nodes as the *elites* of a social network.

Other proven or observed properties of networks generated by preferentical attachment models

- correlation between the degree of a node u and the degrees of the neighboring nodes.
- graph has small diameter; suggesting "6 degrees of separation phenomena"
- relatively large dense subgraph communities.
- rapid mixing (for random walks to approach stationary distribution)
- relatively small (almost) *dominating sets* .

On my spring 2020 CSC303 web page, I posted a paper by Avin et al (2018) that shows that preferential attachment is the *only* "rational choice" for players (people) playing a simple natural network formation game. It is the rational choice in the sense that the strategy of the players will lead to a unique equilibrium (i.e. no player will want to deviate assumming other players do not deviate). For those intersted, I have posted (in my CSC303 webpage) a number of other papers on elites in a social network and preferential attachment.

The Small World Phenomena

I already mentioned the small worlds phenomena. A mathematical explanation of this phenomiena (expecially how one hones in on a target recipient) was given by J. Kleinberg in a network formation model that explicitly forces a power law property.

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. Moreover, such a path can be found in a decentralized manner

In Milgram's 1967 small world experiment, 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.

Network concepts will be mainly introduced in context

But at the of the slides I will provide an appendix of basic graph definitions.

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

Romantic Relationships [Bearman et al, 2004]

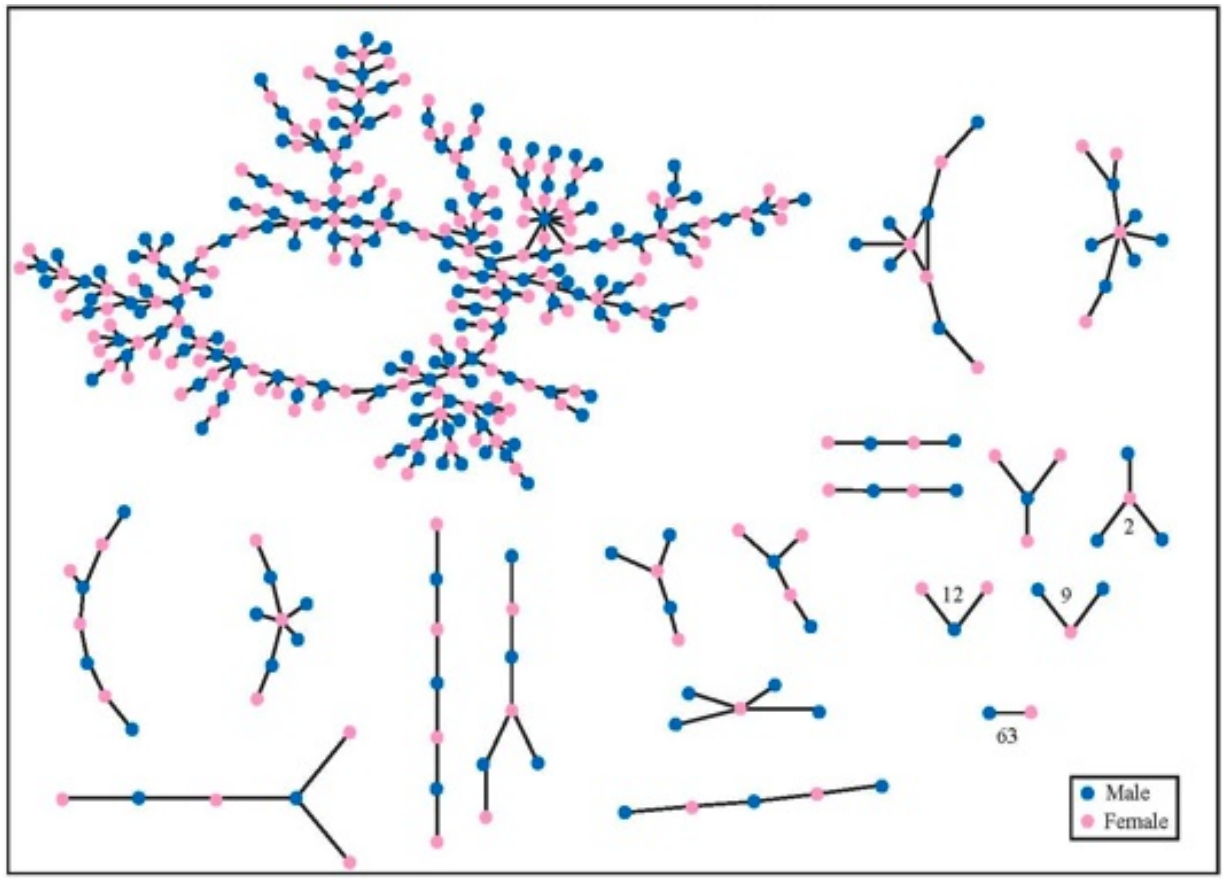
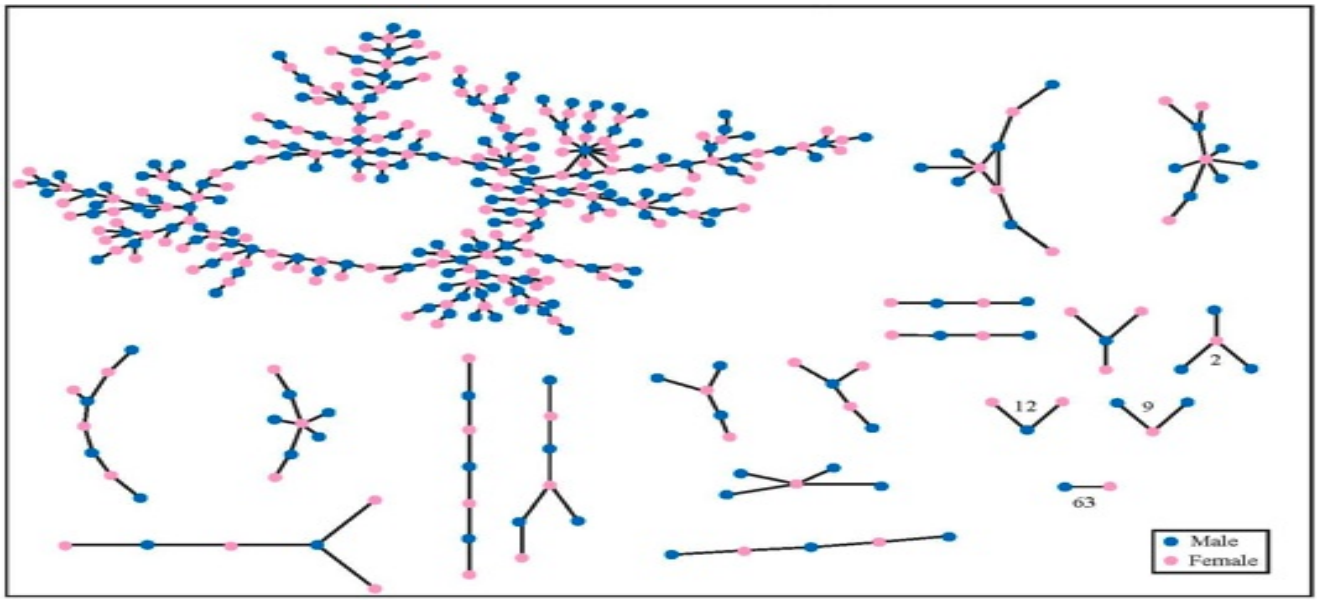


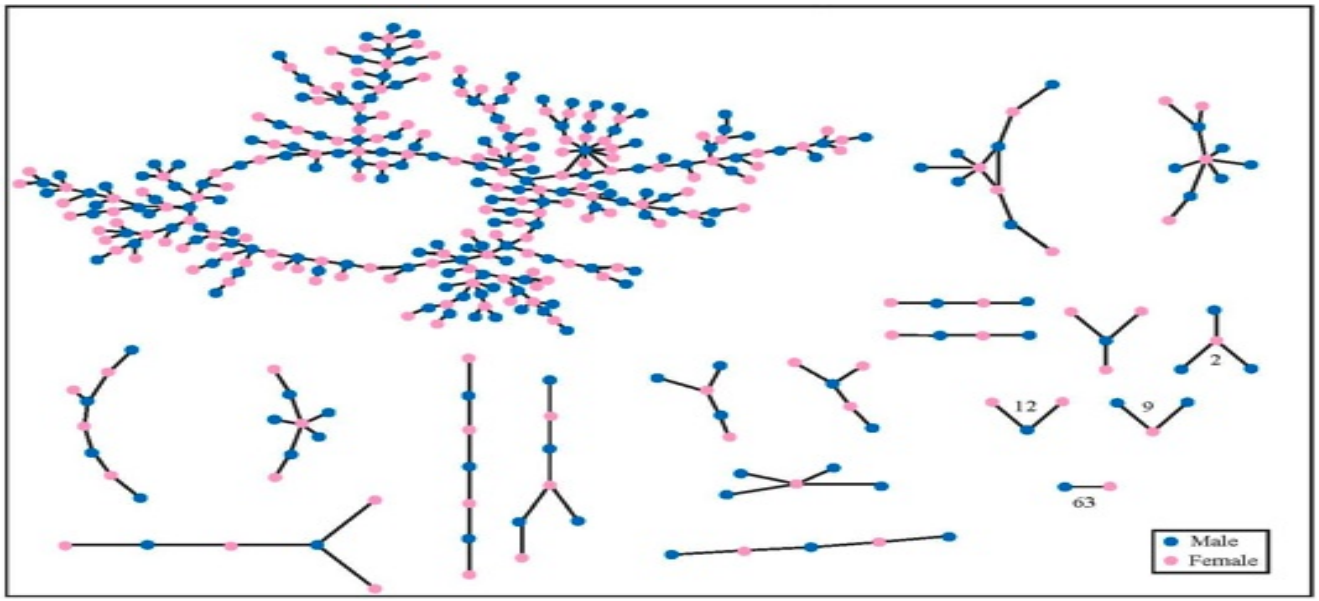
Figure: Dating network in US high school over 18 months.

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

More basic definitions



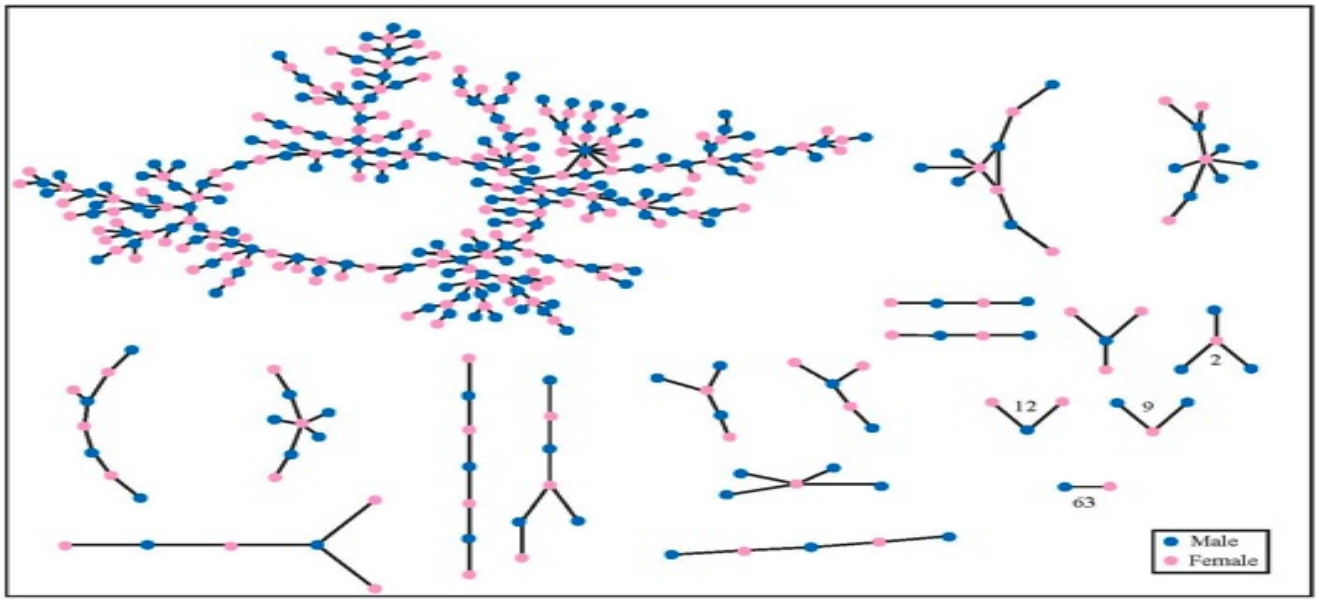
More basic definitions



Observation

Many connected components including one "giant component"

More basic definitions



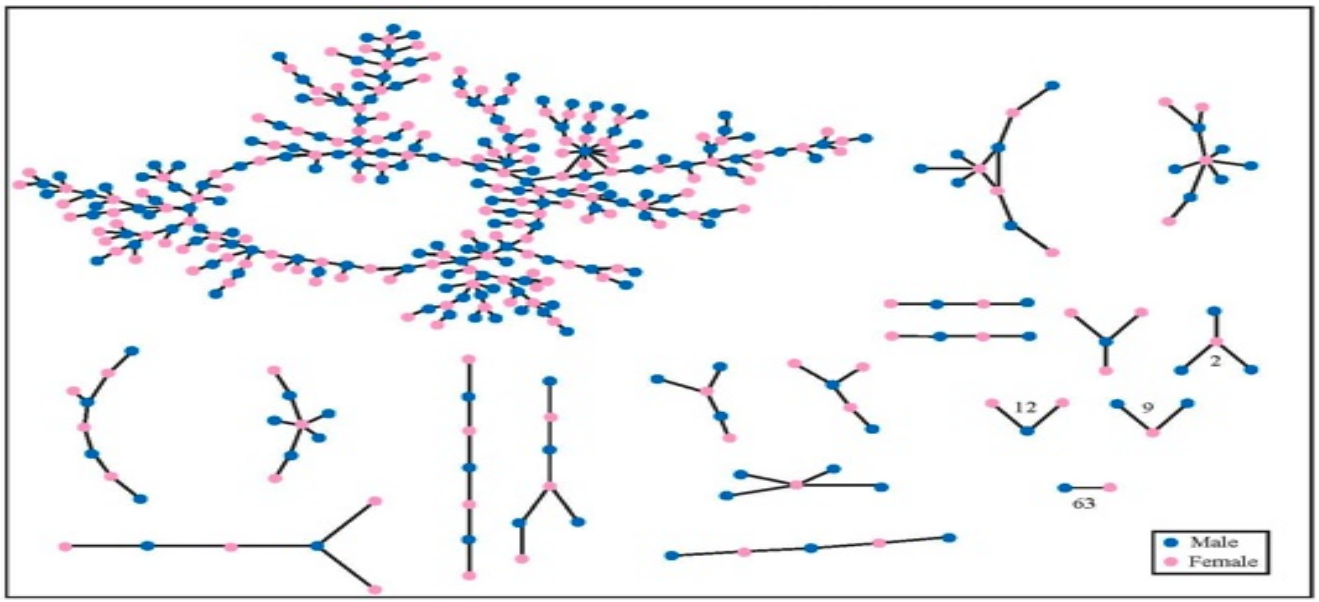
Observation

Many connected components including one "giant component"

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

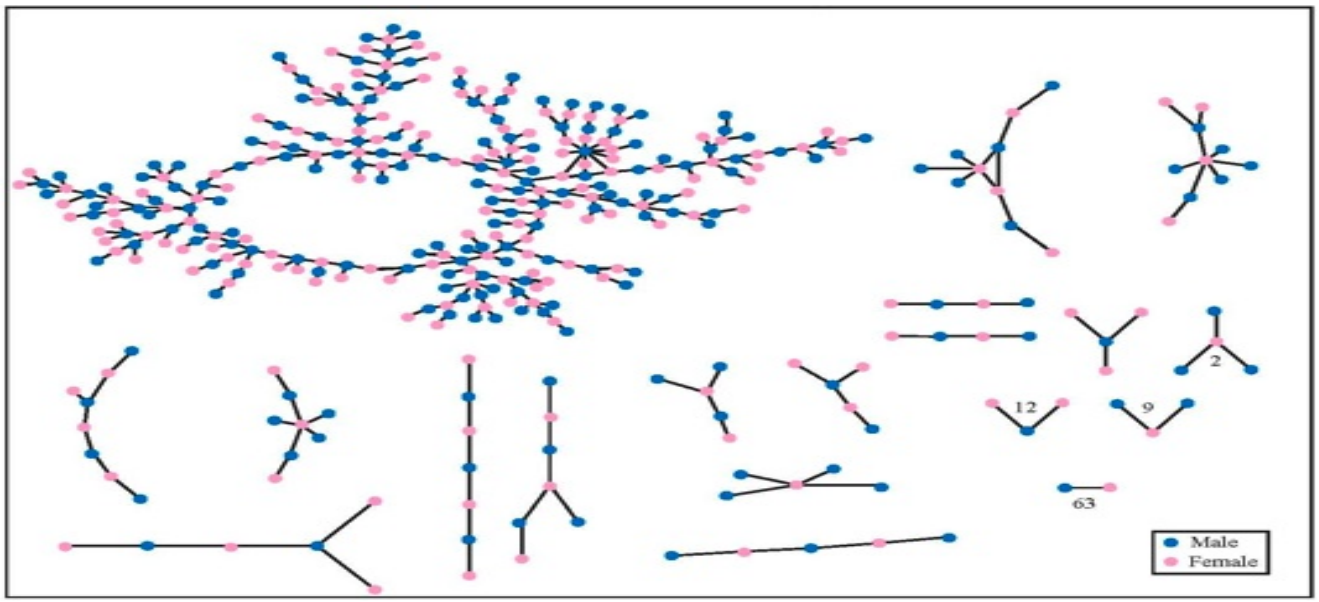
Simple paths and simple cycles

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



Simple paths and simple cycles

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



Observation

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

Example of an acyclic bipartite graph

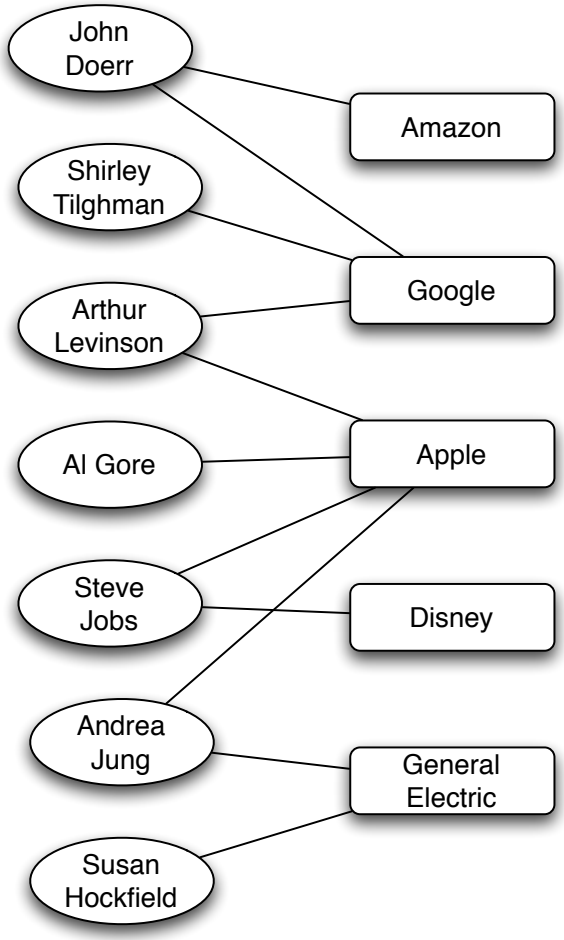


Figure: [E&K, Fig 4.4] One type of affiliation network that has been widely studied is the memberships of people on corporate boards of directors. A very small portion of this network (as of mid-2009) is shown here.

Florentine marriages and "centrality"

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

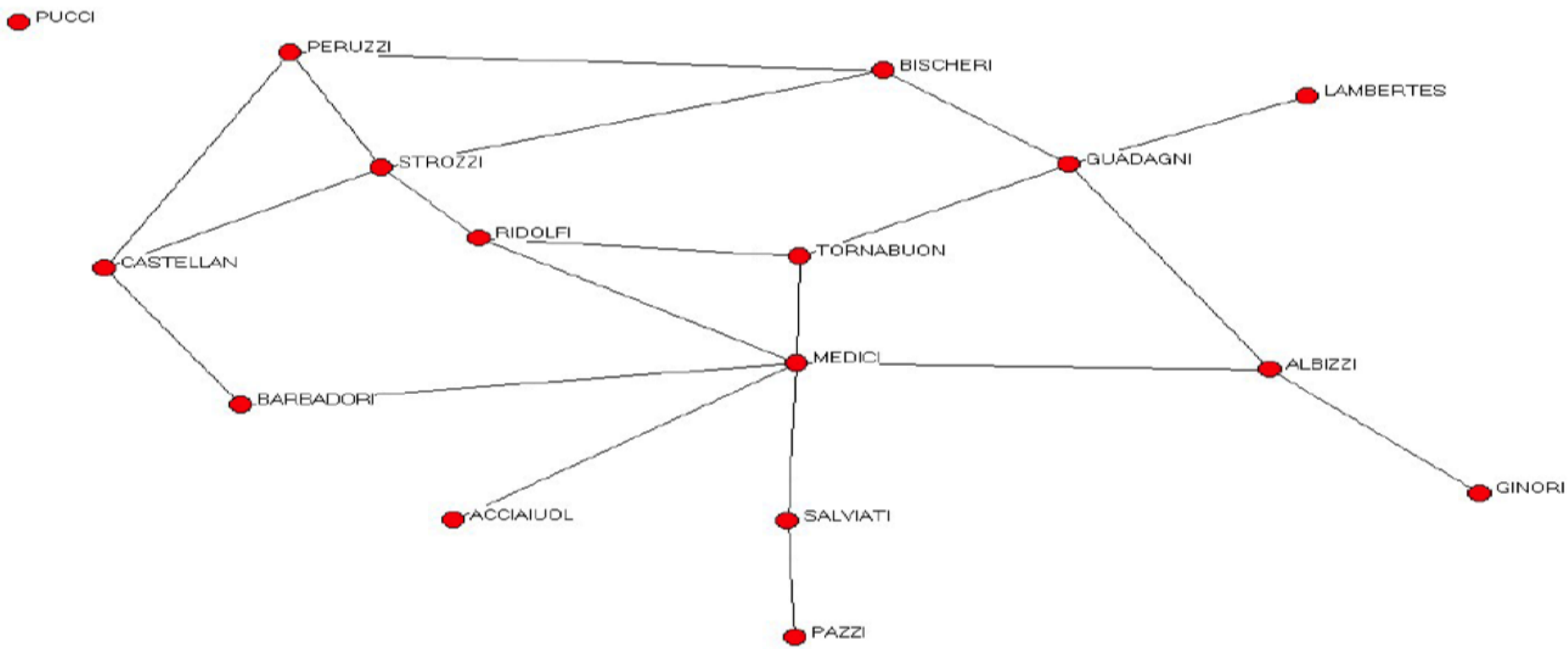


Figure: see [Jackson, Ch 1]

Some additional comments on how graph structure can reveal personal and individual information: Detecting the romantic relation in Facebook

- 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 anbd 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 strucutral 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 disperison example from paper

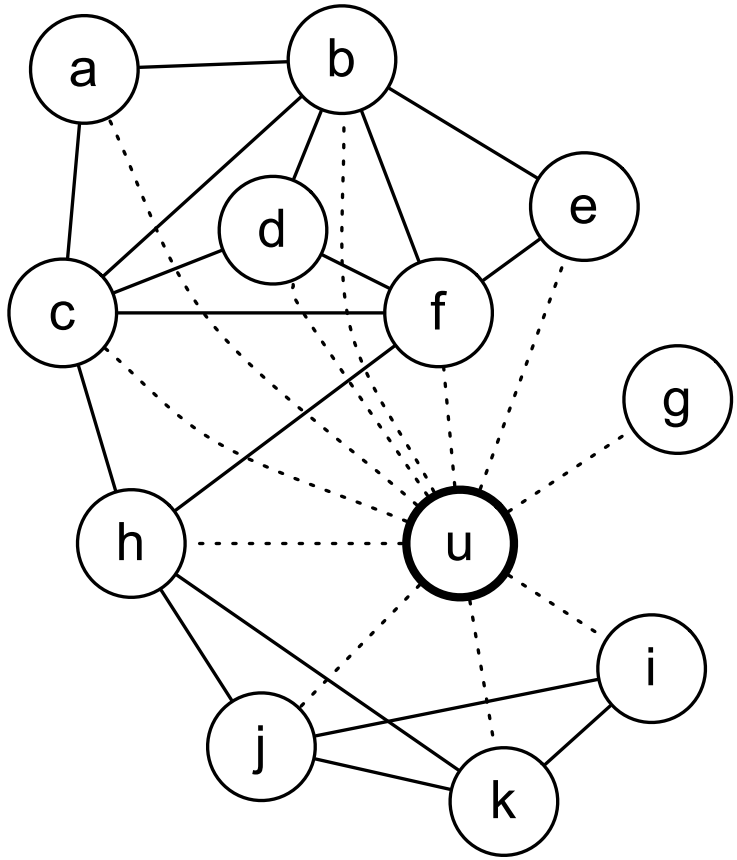


Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes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

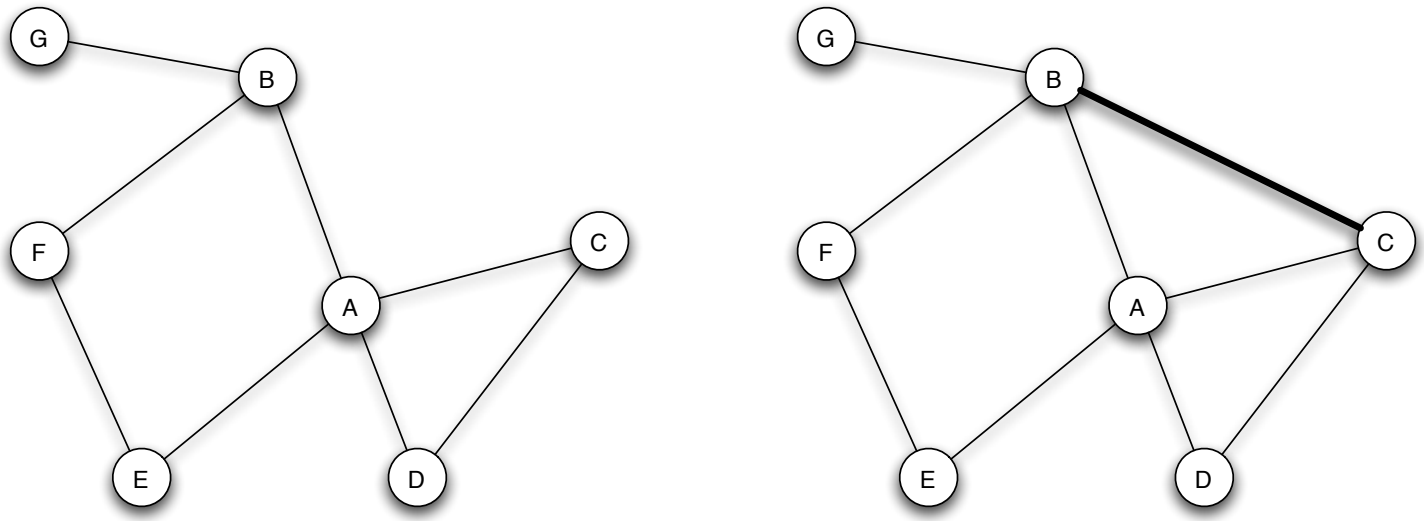
Some experimental results for the fraction of correct predictions

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

Triadic closure (undirected graphs)



(a) *Before B-C edge forms.* (b) *After B-C edge forms.*

Figure: The formation of the edge between B and C illustrates the effects of triadic closure, since they have a common neighbor A. [E&K Figure 3.1]

- Triadic closure: mutual "friends" of say A are more likely (than "normally") to become friends over time.
- How do we measure the extent to which triadic closure is occurring?
- How can we know why a new friendship tie is formed? (Friendship ties can range from "just knowing someone" to "a true friendship" .)

Measuring the extent of triadic closure

- The clustering coefficient of a node A is a way to measure (over time) the extent of triadic closure (perhaps without understanding why it is occurring).

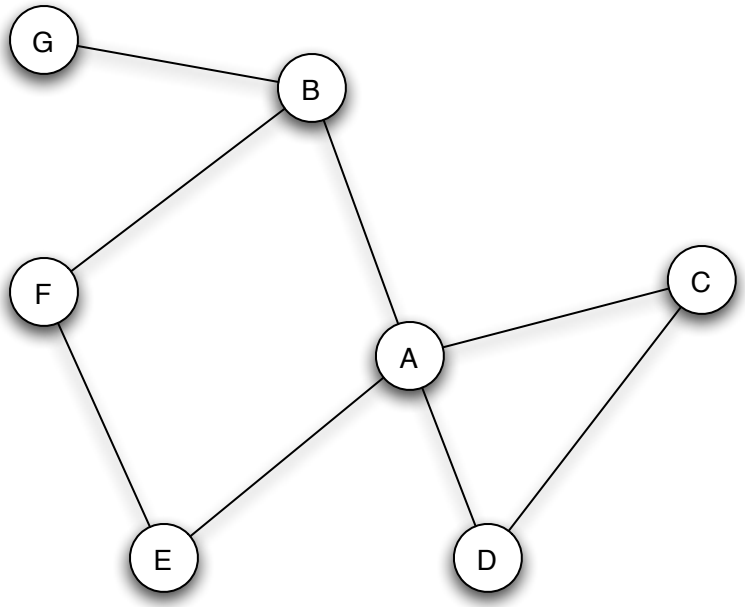
- Let E be the set of an undirected edges of a network graph. (Forgive the abuse of notation where in the previous and next slide E is a node name.) For a node A, the clustering coefficient is the following ratio:

$$\frac{\left|\left\{(B,C) \in E : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}{\left|\left\{\{B,C\} : (B,A) \in E \text{ and } (C,A) \in E\right\}\right|}$$

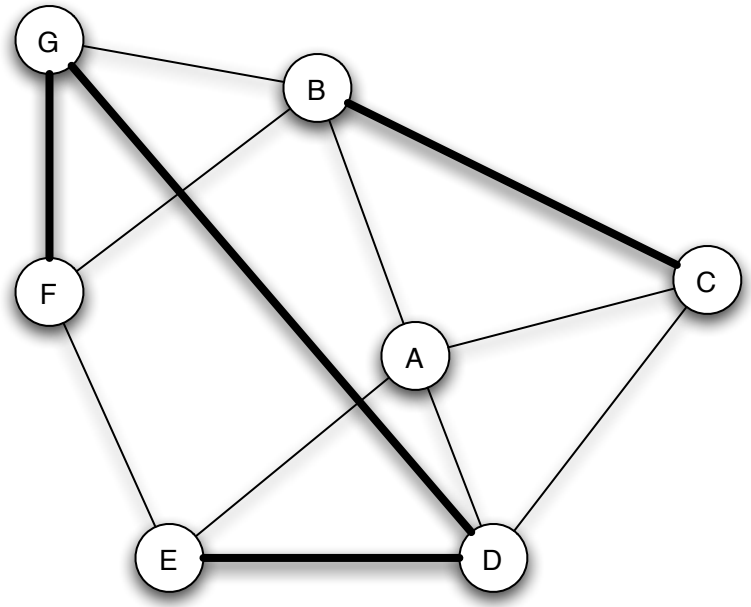
- The numerator is the number of all edges (B, C) in the network such that B and C are adjacent to (i.e. mutual friends of) A.

- The denominator is the total number of all unordered pairs $\{B, C\}$ such that B and C are adjacent to A.

Example of clustering coefficient



(a) *Before new edges form.*



(b) *After new edges form.*

- The clustering coefficient of node A in Fig. (a) is $1/6$ (since there is only the single edge (C, D) among the six pairs of friends: $\{B, C\}$, $\{B, D\}$, $\{B, E\}$, $\{C, D\}$, $\{C, E\}$, and $\{D, E\}$). We sometimes refer to a pair of adjacent edges like $(A, B), (A, C)$ as an "open triangle" if (B, C) does not exist.
- The clustering coefficient of node A in Fig. (b) increased to $1/2$ (because there are three edges (B, C), (C, D), and (D, E)).

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

Interpreting triadic closure

- Does a low clustering coefficient suggest anything?

- Bearman and Moody [2004] reported finding that a low clustering coefficient amongst teenage girls implies a higher probability of contemplating suicide (compared to those with high clustering coeficient). Note:The value of the clustering coefficient is also referred to as the *intransitivity coefficient*.

- They report that " Social network effects for girls overwhelmed other variables in the model and appeared to play an unusually significant role in adolescent female suicidality. These variables did not have a significant impact on the odds of suicidal ideation among boys. "

How can we understand these findings?

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.
  Increased opportunity, trust, incentive ; it can be awkward to have friends (especially good friends with strong ties) who are not themselves friends.

Bearman and Moody study continued

- Triadic closure (or lack thereof) can provide some plausible explanation.
  Increased opportunity, trust, incentive ; it can be awkward to have friends (especially good friends with strong ties) who are not themselves friends.
  As far as I can tell, no conclusions are being made about why there is such a difference in gender results.

The study by Bearman and Moody is quite careful in terms of identifying many possible factors relating to suicidal thoughts. Clearly there are many factors involved but the fact that network structure is identified as such an important factor is striking.

Bearman and Moody factors relating to suicidal thoughts

TABLE 3—Logistic Regression of Suicide Attempts, Among Adolescents With Suicidal Ideation, on Individual, School, Family and Network Characteristics

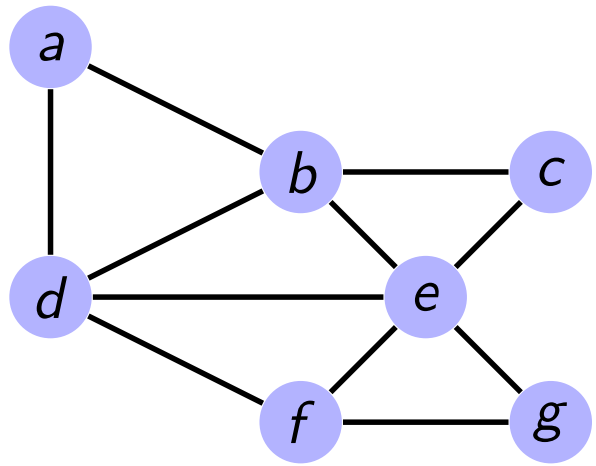
	Suicide Attempts, OR (95% CI)	
	Males	Females
Demographic		
Age	0.956 (0.808, 1.131)	0.920 (0.810, 1.046)
Race/ethnicity		
Black	0.872 (0.414, 1.839)	1.086 (0.680, 1.736)
Other	1.069 (0.662, 1.728)	1.134 (0.810, 1.586)
Socioeconomic status	0.948 (0.872, 1.031)	1.008 (0.951, 1.069)
School and community		
Junior high school	1.588 (0.793, 3.180)	1.271 (0.811, 1.993)
Relative density	0.049 (0.005, 0.521)	0.415 (0.086, 1.996)
Plays team sport	0.985 (0.633, 1.532)	1.020 (0.763, 1.364)
Attachment to school	1.079 (0.823, 1.414)	1.066 (0.920, 1.235)
Religion		
Church attendance	0.975 (0.635, 1.496)	0.818 (0.618, 1.082)
Family and household		
Parental distance	0.925 (0.681, 1.256)	0.955 (0.801, 1.139)
Social closure	1.004 (0.775, 1.299)	0.933 (0.781, 1.115)
Stepfamily	1.058 (0.617, 1.814)	1.368 (0.967, 1.935)
Single-parent household	1.142 (0.698, 1.866)	1.117 (0.800, 1.560)
Gun in household	1.599 (1.042, 2.455)	1.094 (0.800, 1.494)
Family member attempted suicide	1.712 (0.930, 3.150)	1.067 (0.689, 1.651)
Network		
Isolation	0.767 (0.159, 3.707)	1.187 (0.380, 3.708)
Intransitivity index	0.444 (0.095, 2.085)	1.076 (0.373, 3.103)
Friend attempted suicide	1.710 (1.095, 2.671)	1.663 (1.253, 2.207)
Trouble with people	1.107 (0.902, 1.357)	1.119 (0.976, 1.284)
Personal characteristics		
Depression	1.160 (0.960, 1.402)	1.130 (0.997, 1.281)
Self-esteem	1.056 (0.777, 1.434)	0.798 (0.677, 0.942)
Drunkenness frequency	1.124 (0.962, 1.312)	1.235 (1.115, 1.368)
Grade point average	0.913 (0.715, 1.166)	0.926 (0.781, 1.097)
Sexually experienced	1.323 (0.796, 2.198)	1.393 (0.990, 1.961)
Homosexual attraction	1.709 (0.921, 3.169)	1.248 (0.796, 1.956)
Forced sexual relations		1.081 (0.725, 1.613)
No. of fights	0.966 (0.770, 1.213)	1.135 (0.983, 1.310)
Body mass index	0.981 (0.933, 1.032)	1.014 (0.982, 1.047)
Response profile (n = 1/n = 0)	139/493	353/761
F statistic	1.84 ($P = .0170$)	2.88 ($P < .0001$)

Note. OR = odds ratio; CI = confidence interval. Logistic regressions; standard errors corrected for sample clustering and stratification on the basis of region, ethnic mix, and school type and size.

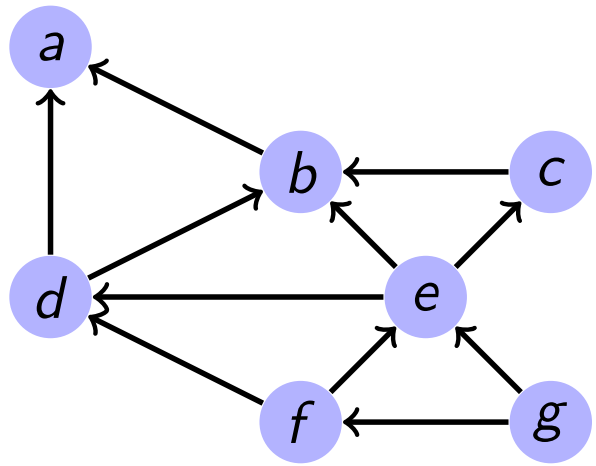
Appendix: Network (graph) definitions and examples

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, . . .

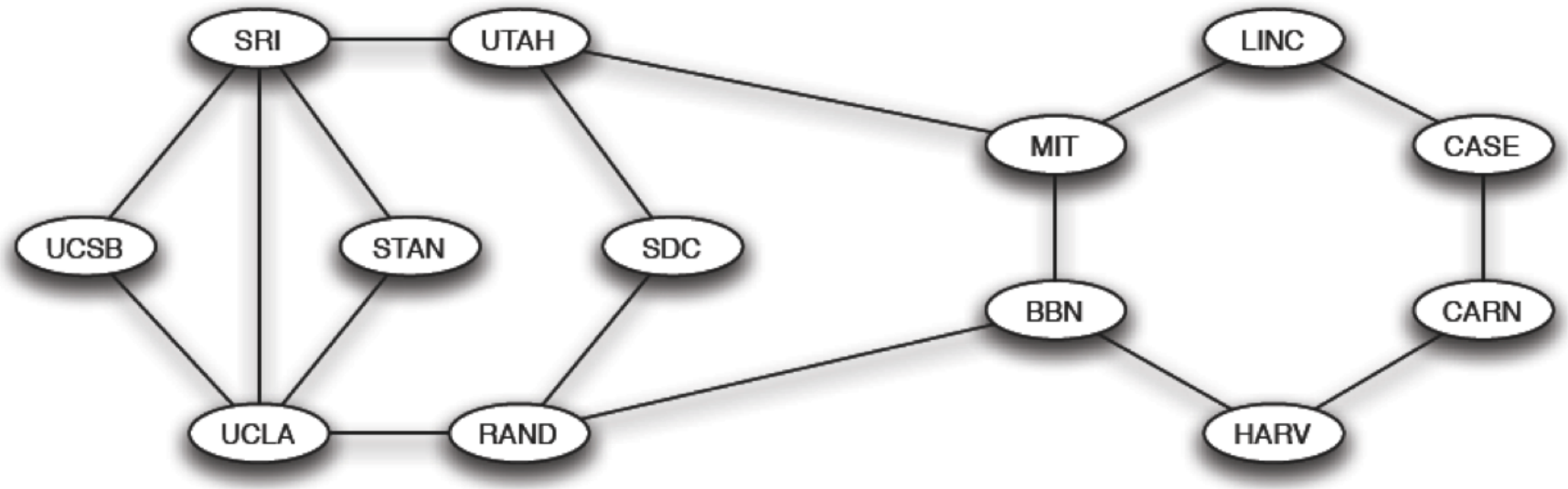


Figure: Internet: Dec. 1970 [E&K, Ch.2]

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

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

- This node induced subgraph (for the sites MIT = 1, LINC = 2, CASE = 3, CARN = 4, HARV = 5, BBN = 6) 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.

The matrix A^2 where $A = A(G)$

Consider squaring the previous matrix $A = A(G)$. That is, $A^2 = A * A$.

$$A^2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Draw a visualization of the graph represented by A^2. If we let $c_{i,j}$ be the i, j entry in A^2, can you desribe the meaning of $c_{i,j}$?

The matrix $B = A + I$

Consider the 6×6 identity matrix $I = (\iota_{i,j})$. That is, $\iota_{i,i} = 1$ for $1 \leq i \leq 6$ and $\iota_{i,j} = 0$ for $1 \leq i, j \leq 6$ and $i \neq j$.

Let $B = A + I$ (as above). That is, $b_{i,j} = a_{i,j} + \iota_{i,j}$ for all i, j. We have

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

Note that now the matrix B has self loops and hence is not a simple graph.

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

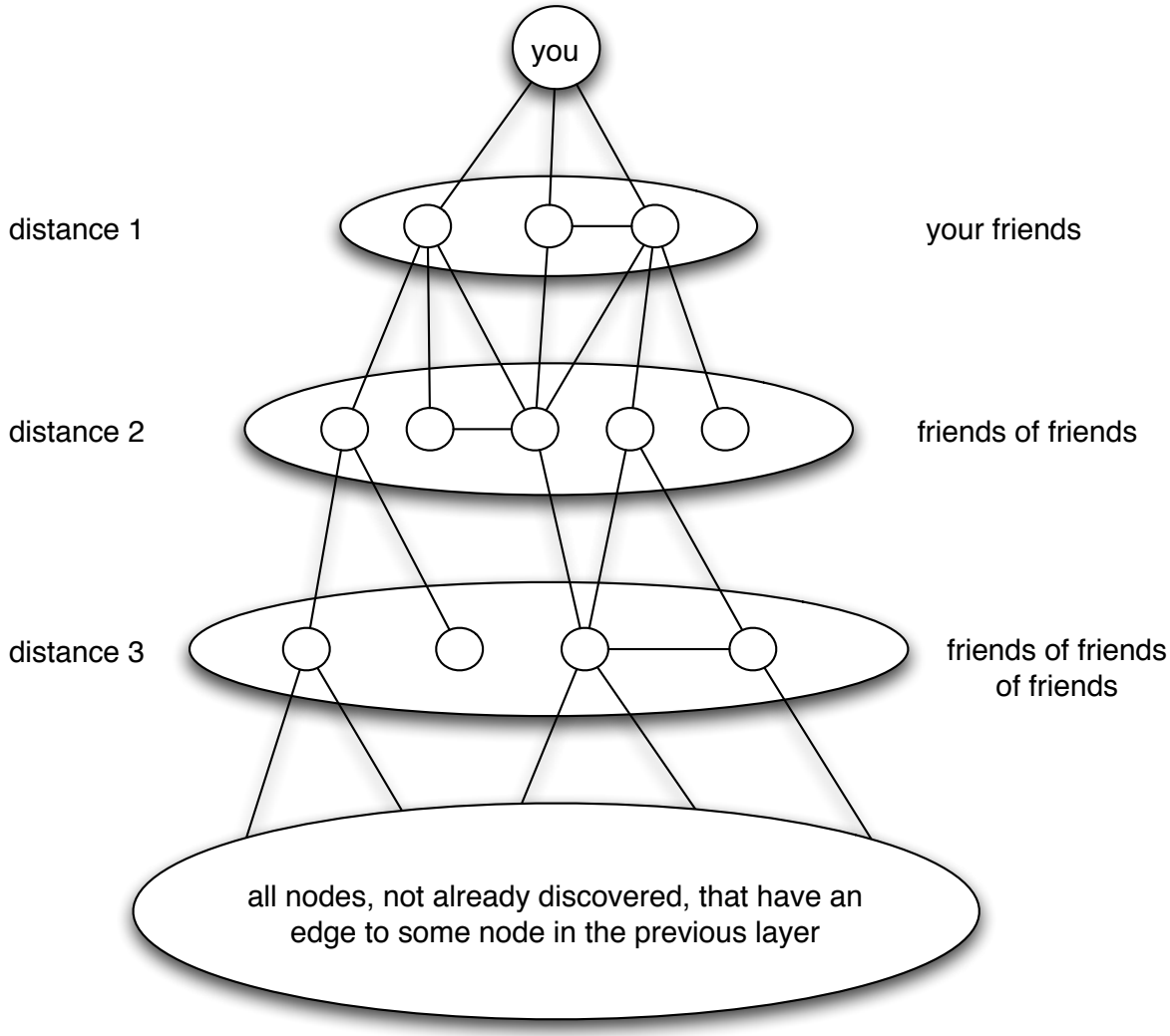


Figure: Breadth-first search discovers distances to nodes one "layer" at a time. Each layer is built of nodes adjacent to at least one node in the previous layer.

Analogous concepts for directed graphs

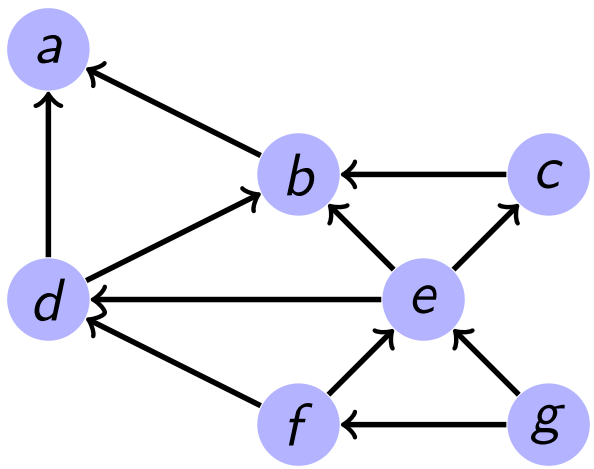
- We use the same notation for directed graphs, i.e. denoting a di-graph as $G = (V, E)$, where now the edges in E are directed.

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).

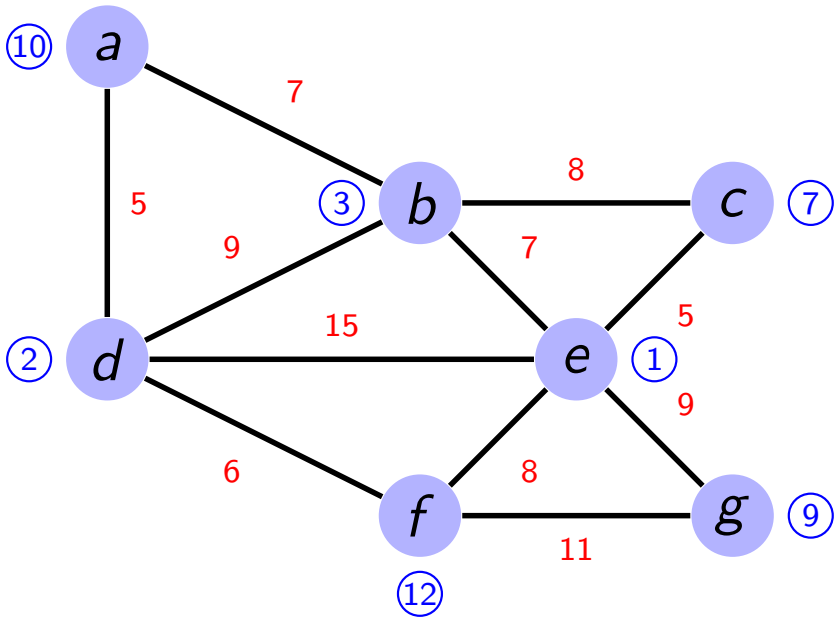
Analogous concepts for directed graphs

- 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

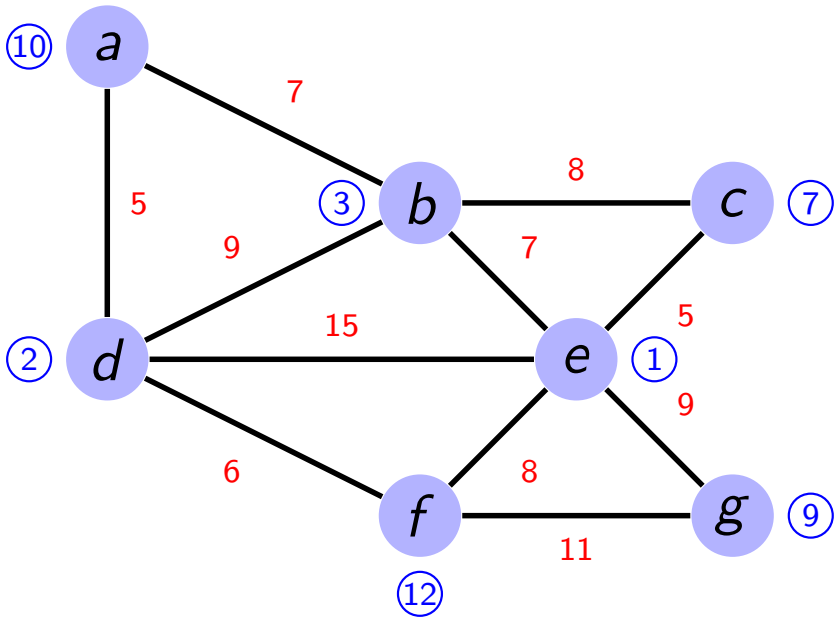
- We will often consider weighted graphs. Lets consider a (directed or undirected) graph $G = (V, E)$. Example:



- 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:

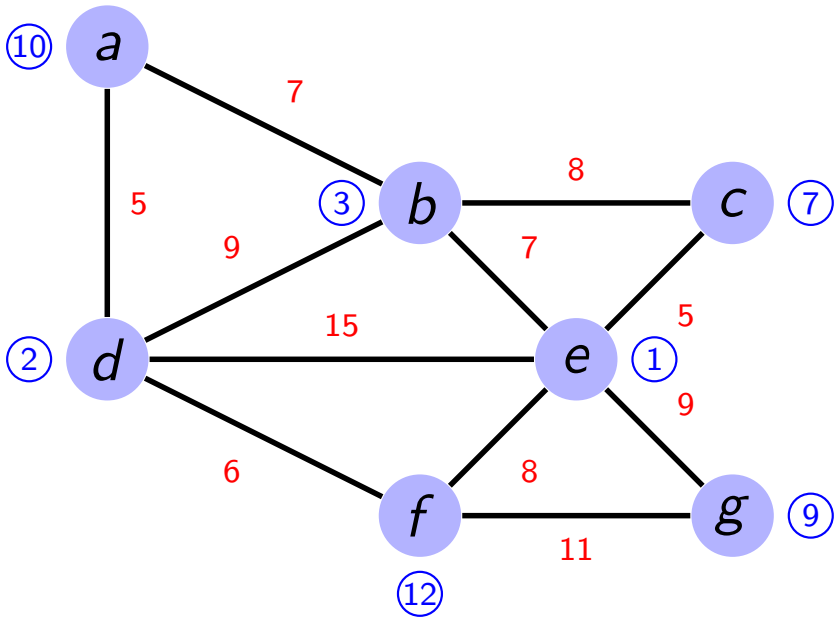


- red numbers: edge weights
- blue numbers: vertex weights

- 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:

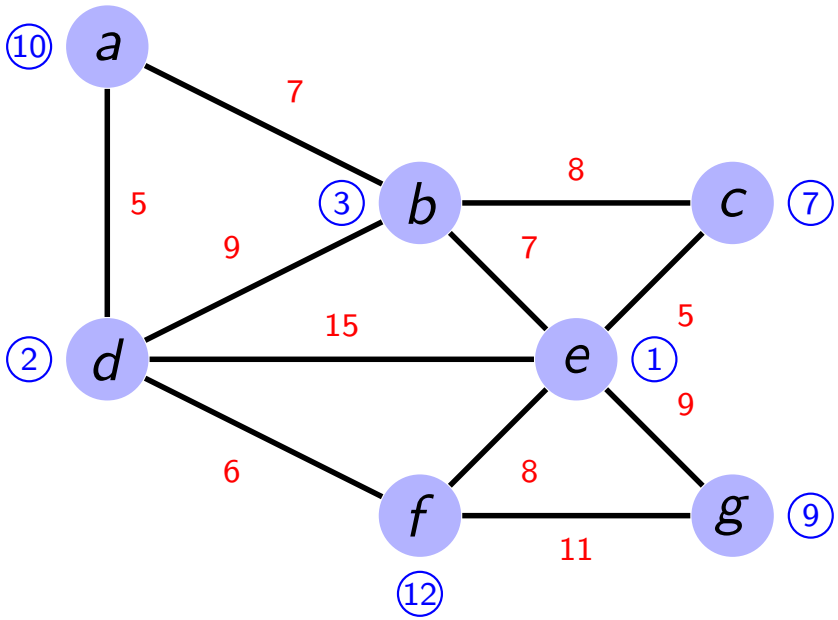


- red numbers: edge weights
- blue numbers: vertex weights

- 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.