Economics of Information Networks

Stephen Turnbull

Division of Policy and Planning Sciences

Lecture 3: November 30, 2021

Abstract

We look at some basic structures of graphs, and applications to social networks, including Granovetter's famous argument about *The Strength of Weak Ties* and approximate network partitioning.

In Part 1, we look at Granovetter's arguments about "the strength of weak ties," and at some quantitative analysis of social network data.

The strength of weak ties

- Granovetter [1973] observed that people frequently learned of opportunities such as job opportunities through personal contacts who were *not* considered friends. That is, the "tie" to these acquaintances was "weak."
- Based on the process of triadic closure, we would expect that
  1. links to "acquaintances" are bridges
  2. friends share much information in common
  3. however, friendly competition is a thing

  concluding that information about such opportunities often would flow from acquaintances naturally.
- Why doesn't triadic closure operate with acquaintances? We propose a *weighted graph* model with strong and weak links. *Strong links* correspond to friends, and *weak links* to acquaintances.

Strong triadic closure

- "Important opportunities" like mid-career job changes are infrequent, so triadic closure converts (local) bridges to non-bridges relatively quickly. We expect information flow via bridges is uncommon.
- The words "friend" and "acquaintance" suggest a solution: a weighted graph, where some links are *strong* and others *weak*.
- *Strong triadic closure* if whenever strong links exist between A and B and between A and C, there is a link (strong *or* weak) between B and C.
- **Theorem:** In a strong triadically closed graph, if A has any other strong links, and is the endpoint of a (local) bridge, the bridge is a *weak* link.
  - Note: the interpretation of a local bridge as an "acquaintance" makes a lot of sense: they are both not part of the (not necessarily strong) triadic closed group of friends and local bridge is a weak link.

Strong triadic closure

- Important opportunities arise infrequently
- Local bridges should convert quickly
- Information shouldn't flow by bridges
- Friends, not acquaintances
- Need to justify acquaintances (weak ties)
- This is why we generalize to a weak-strong graph, weak = acquaintance, strong = friend

A note on "strong"

- "Strong version of a condition" *vs.* usage of "strong" in "strong triadic closure."
  - A *strong condition* usually is more restrictive, and applies to fewer cases.
  - *Strong triadic closure* is the opposite; the condition is less restrictive than in triadic closure (it only applies when the links are both strong, so more graphs are strong triadic closed than are triadic closed).

Graph structure corresponds to real phenomena

- Theorem: *if*
  1. graph is strong triadic closed
  2. node A has a bridge
  3. node A has another strong link

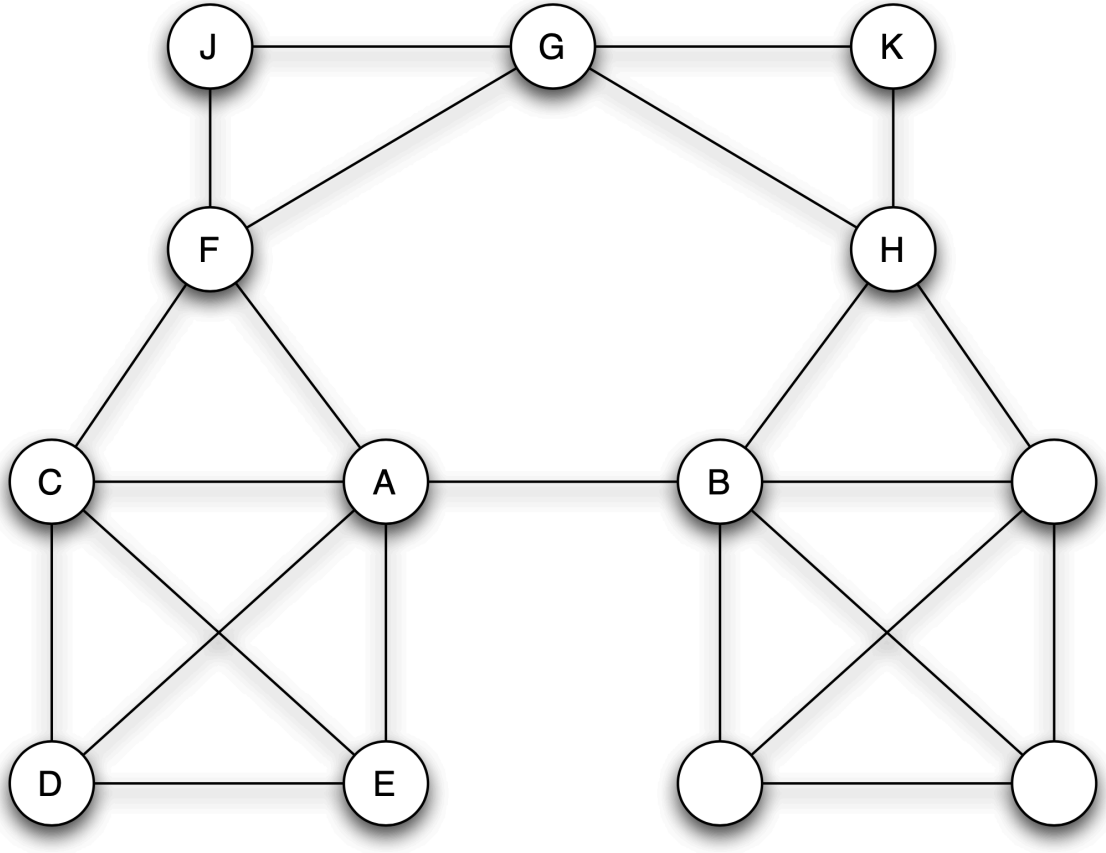
  then the bridge is a weak link.

Interpretation

- Strong triadic closure implies group of friends is closed
- So bridges allow weak links to non-friends (acquaintences)
- The theorem is purely mathematical, but it implies something that we now know about graph structure from Granovetter, namely that there is reason to expect bridges to weak links.
- This is what mathematics is for: helping us understand *why* certain properties of behavior tend to "cluster."

Discussion of graph with local bridge

- Show that graph is strongly triadically closed, though not triadically closed.
- Point out gatekeepers (several).
- Interpretation as members of a college class, with connections across companies.



"Smoothing" strength

- We needed to distinguish strong and weak links to analyze Granovetter's observation. An obvious generalization that is sometimes useful is to make strength a numerical quantity.
  - Practical measures of link strength in communication networks include number of calls, and cumulative length of calls, in a fixed period.
  - In surveys, we might ask whether a person is *unrelated*, an *acquaintance*, or a *friend*.

"Smoothing" bridges

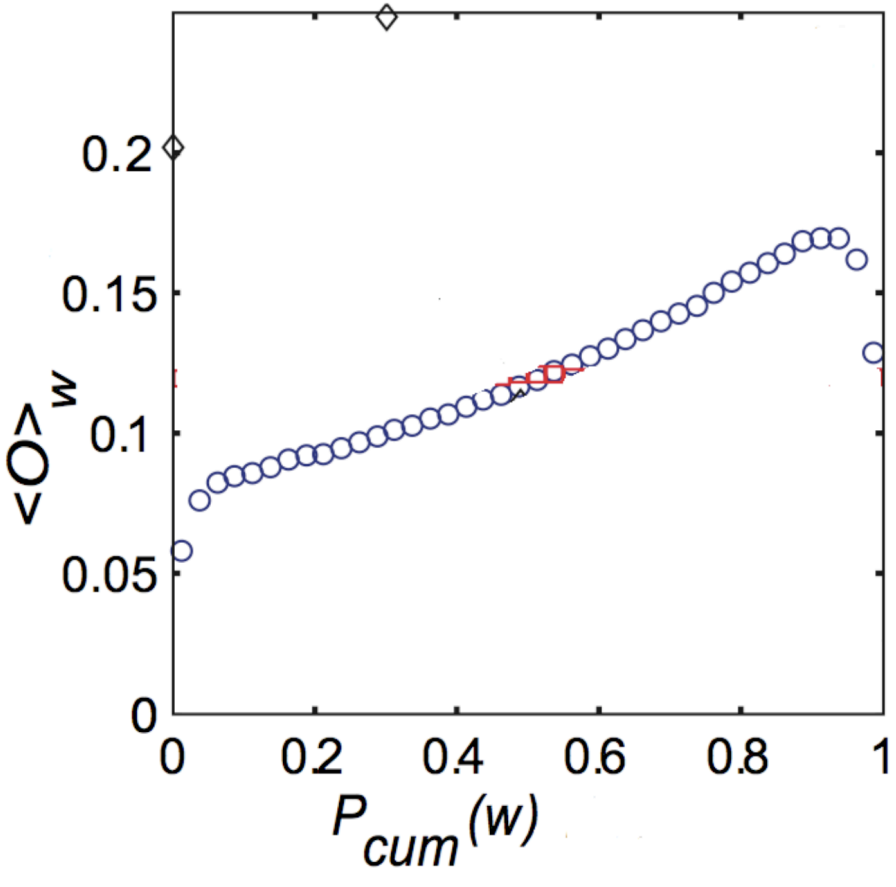
- In many practical applications, only a fraction of edges are local bridges.
- We can make this smoother by defining *neighborhood overlap* of an edge AB as

$$\frac{|N(A) \cap N(B)|}{|N(A) \cup N(B)|}$$

  where $N(X)$ is the set of nodes that are neighbors of X, not including A and B themselves. Note that with this definition, a link has neighborhood overlap of zero exactly when it is a local bridge.
- In data sets where we have a quantitive measure of strength, it often shows a correlation with neighborhood overlap.

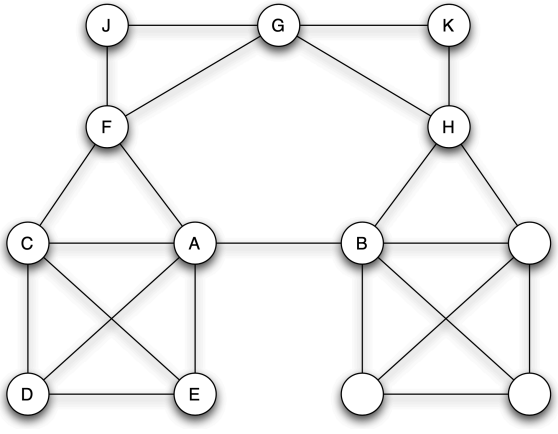
Overlap *vs.* strength in cellphone calls

Onnela *et al.* [2007] worked with cellphone data. An edge exists if two cellphones made calls in both directions in a month. The strength of an edge is measured by its percentile in the distribution of number of minutes used.



Giant component in cellphone data

- Giant components usually don't have many local bridges.
- People usually have many correspondents, and they will typically converse with several. Seems very likely that will be substantial overlap among neighbors (triadic closure).
- However, we expect to see weak ties that link "communities."
  - In the bridge figure we can see three or four such triadically closed "communities."
  - *E.g.* a company with several departments.



Decomposing the giant component

- Onnela *et al.* deleted links one at a time, starting with the strongest.
  - The giant component gradually shrinks as individuals lose all their ties.
- Now try the same process, starting with the weakest.
  - The giant component shrinks more quickly as individuals lose all their ties, but
  - also because it fragments into smaller components when bridges are deleted, and some of these bridges link components of similar size.
  - The fragmentation doesn't happen until late in the strongest first process because these are much less likely to be bridges (or in general have low neighborhood overlap).
- This strongly suggests that weak links join "communities."

Research opportunities

- The theoretical work is not fully worked out.
  - Some advanced mathematics is used, but there remain important concepts (such as "community") that remain undefined.
  - Probably there are elementary theorems (like the one about strong triadic closure and the weakness of bridges) to be proved.
- Experiments like changing the order of link deletion according to link strength often clarify the nature of concepts. They are valuable contributions that non-mathematicians can make.
- Many opportunities to study real networks, as well as to develop theorems that describe why they behave (show relationships) as they do.

Applications of tie strength

- Besides strong/weak, what can we say?
- Try different ways of measuring strength, and how that affects results.
- Social networks allow examination of relationships in ways that subjects define, so are not artifacts of the research design.
  - Consider office colleagues one of whom is a salesperson, they may call each other a lot. But each may rarely call their wives, because they can't do so at work.
- Relationships like Facebook friend or Twitter follower are declared by users themselves, and have well-defined semantics according to the rules of the social network.
  - Be careful: Social network usage of words like "friend" corresponds imperfectly to the usual usage in everyday language.

Friends on Facebook

- Facebook friends aren't necessarily real-life friends, but this is easier to deal with than the reification of "friend" as measured by number of phone calls.
- On the other hand, Facebook friendship is necessarily symmetric, which is not always true in real life.
- With Facebook, you can get an exact list of friends from the Facebook API. Not necessarily true of people trying to list real-life friends.
  - However, in the case of the karate club, there's a small universe known to the research so can be more accurate (*e.g.*, give a list).

Behavior on social networks

- Existence of flamewars on email lists and newsgroups starting from the 1980s.
  - Even virtual violence (deleting files, for example).
- Similar existence of flaming, trolling, doxxing on modern social networks.
- Social networks perhaps even more flammable than mailing lists, since it is (at least, can be) more public.
- Are they really different?
- Are behavioral differences due to changes in individuals' thinking and behavior, or is it emerging from network structure?
  - *E.g.*, text-based communication does not express subtleties of emotion (even with smilies and emoji).

A remark on Twitter feeds

- Compare personal experience. The "algorithm" is part of the network structure.
  - Some people I follow don't seem to post much. Is that because they aren't posting much or because Twitter isn't putting them in my timeline because they don't seem to post things I'm interested in?
  - I don't like Trump, and I don't like my Republican Senator. But I follow my Senators and Representative to be well-informed in future elections. However, I see only nasty opposition in the replies to the Republican Senator. Is that all there is, or has Twitter figured out I don't like him and so only shows the nasty replies?

Homework #9

Due: December 7, 2023 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #9 0AL0200`.

Consider a complex nextwork such as the World Wide Web of documents and links between them. The lecture describes the research of Onnela *et al.*, ordering deletion of links by *strength*.

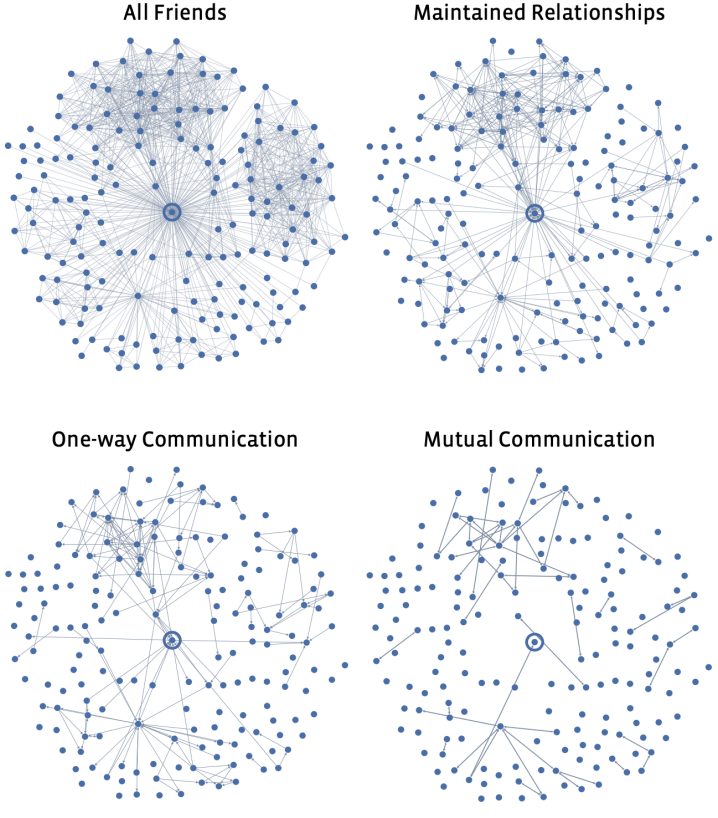
1. What do you think would happen in case of deletion in order of decreasing *neighborhood overlap*? Increasing order?

Interaction on Facebook

- Some researchers looked at intensity of interaction among Facebook users.
- When people have Facebook friends, do they actually interact with them?
- Some people use Facebook to coordinate a group. Interaction looks small, although the group make interact offline frequently and intensely.
- Others post comments each way (*e.g.*, a family spread out geographically.
- Researchers defined three degrees of communication between friends:
  - reciprocal communication (each member sent messages to the other),
  - one-way communication (only one member has sent a message to the other), and
  - maintained relationship (one way), clicking on content or multiple profile visits.

Apparent nesting of relationships by strength

- We look at the *network of neighbors of one member*. The graph shows the various graphs associated with a particular member.
- According to the definitions in this research, *every* reciprocal communication link is also a one-way communication link.

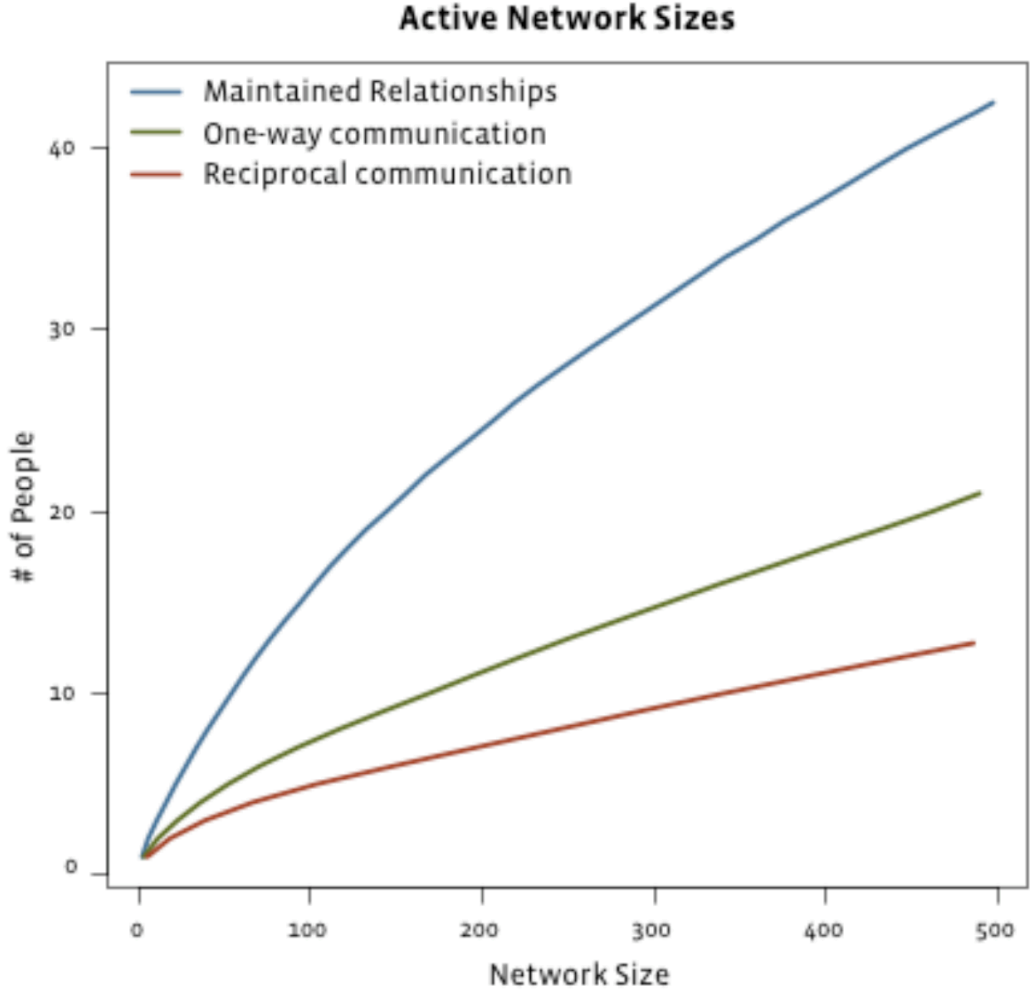


Analysis of nesting of relationships

- For this member, it's easy to visually confirm that the one-way communication links are pretty much all maintained relationship links. I did not expect this. In using Twitter, there are people whose profiles I check *because* I don't "at" them, get "at"-ed by them, *etc.*
- Thinning takes place by "area," not uniformly in the graph. Suggests multiple mechanisms for finding new relationships.
  - Can this be measured statistically? If so, look at distribution of number of mechanisms.

Friendship strength

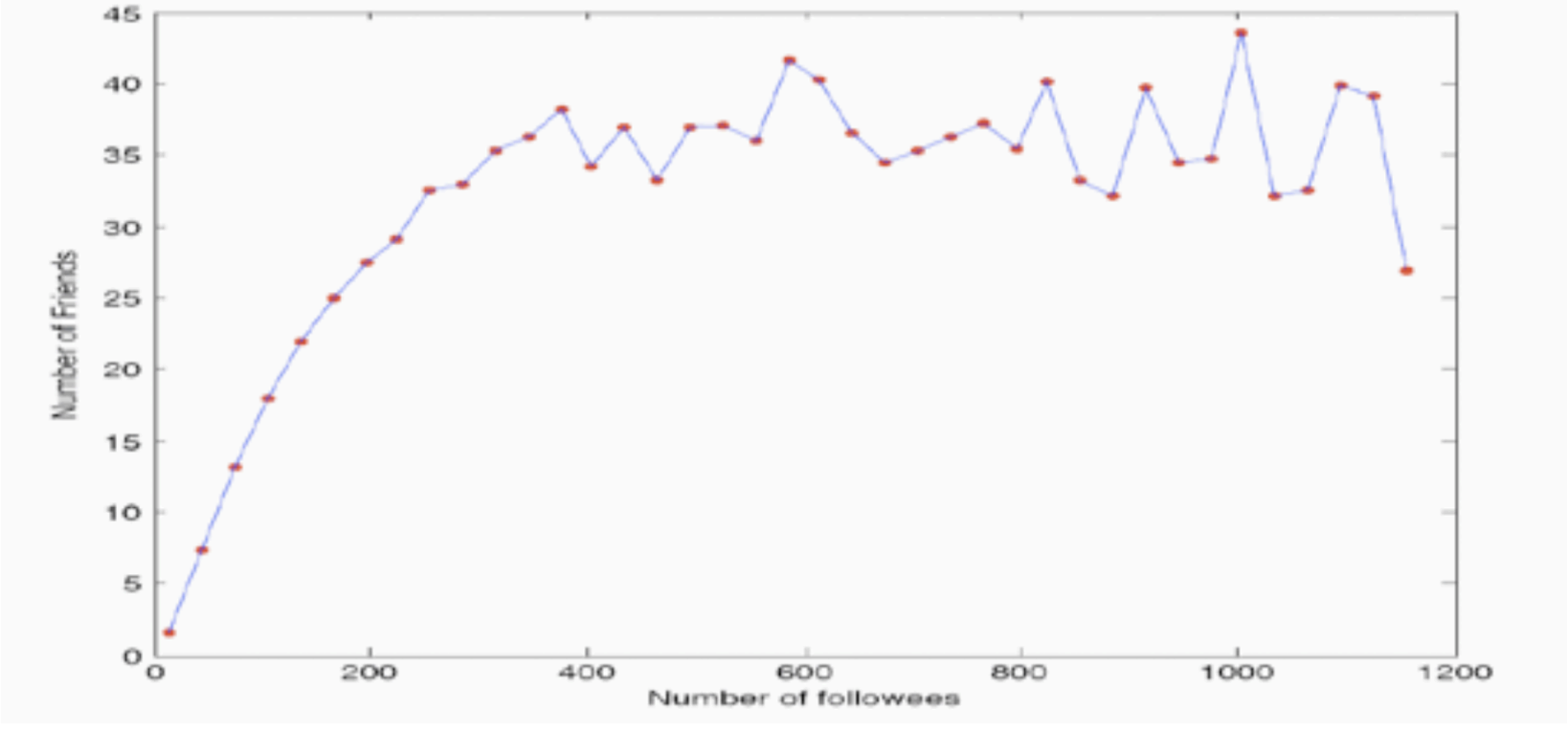
- The graph shows the average count of each kind of link over the members with a particular number of friends.



Analysis of friendship strength

- We see that on average the quantitative relationship is verified.
  - This does *not* mean that the nesting relationship is verified.
  - Reciprocal links are logically a subset of the one-way links, but
  - I would like to see some statistics on the percentage of maintained links that are also one-way and reciprocal links.
- The book points out that even people with 500 friends on Facebook on average limit themselves to about 40 maintained links, 20 one-way links, and 10 mutual links.

Increasing friendship strength on Twitter



Analysis of Twitter friendship strength

- Even people with 500 friends on Facebook on average limit themselves to about 40 maintained links, 20 one-way links, and 10 mutual links.
- I'm more interested that the asymptotes do not seem to be horizontal: people with more friends have more strong relationships of each type.
- This is different for Twitter: the asymptote is horizontal, there seems to be a cap on strong ties.
  - Not clear what this says: could be a limitation of Twitter clients.
  - Could be a limitation of the methodology (no measure of attention paid to tweets, which corresponds to maintained relationships).
  - It could be a different kind of social network, different style of communication: *needs more research*.

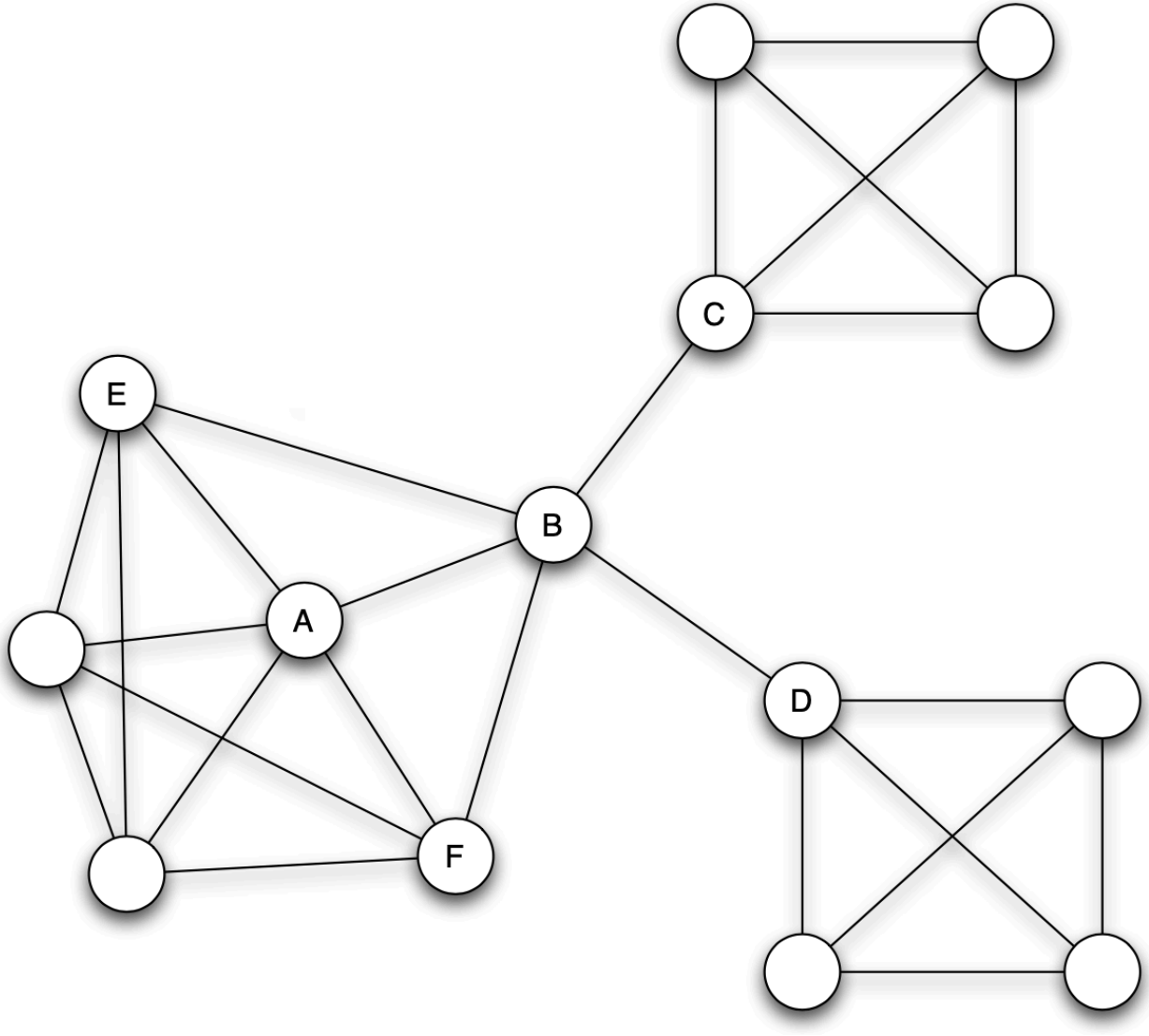
Lecture 3, Part 2

Lecture 3 Part 2 discusses more quantitative measures in networks, such as betweenness and embeddedness, and relates them to ideas in social behavior such as power and trust. Finally we describe the Girvan-Newman graph partitioning algorithm which accurately predicts the split of a club into two.

Structure and behavior

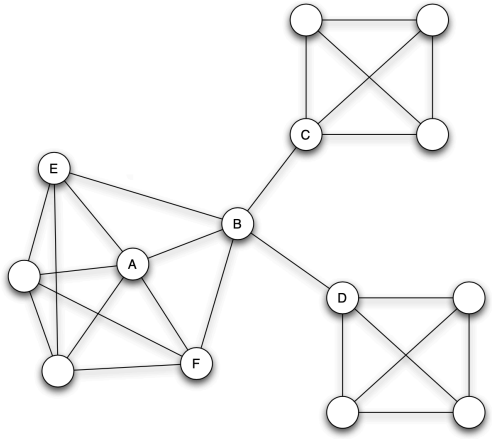
- Network structure induces constraints on behavior.
  - At a granular scale, in a city you can go from anywhere to anywhere by automobile.
  - By contrast you can only go from station to station on a train; you can't even get off between stations as you pass by.
- In this sense, economics (which studies how people deal with constraints) can be considered to "include" social behavior.

Symmetry and asymmetry of nodes



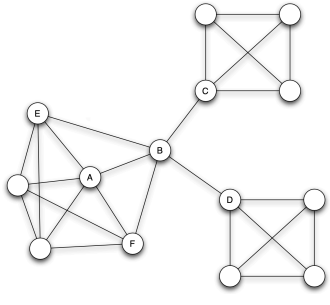
Symmetry and asymmetry of nodes

- Note the symmetry between E and F, and between the two unnamed nodes in that group.
  - E and F are connected to all other group nodes except each other.
  - The unnamed nodes are connected to all other group nodes except B.
- E and an unnamed node are different in a minor way (E is not connected to its twin F, while the unnamed nodes have a common unconnected group member), and in a more important way: E-C is *length 2*, while from an unnamed node to C is *length 3*.



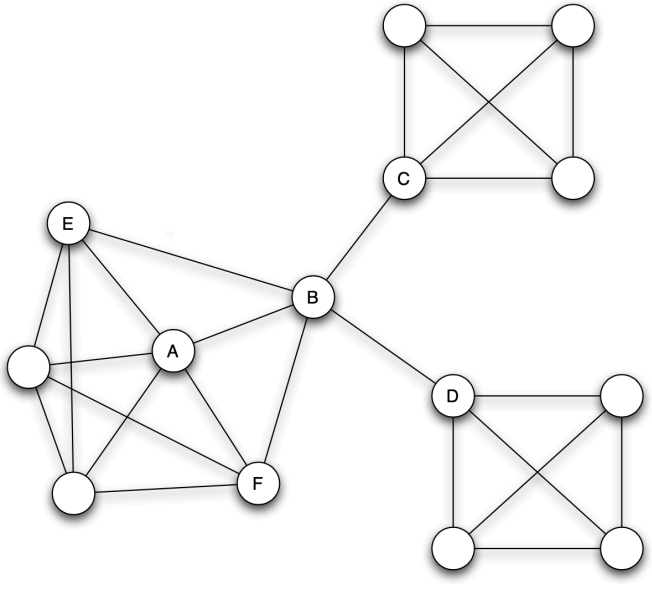
Symmetry and strength of gatekeepers

- One obvious feature of the network is the gatekeeper position of B is stronger than C and D, though they are also gatekeepers.
  - You might think that it matters whether there are *more* nodes in one group than the other, or more *valuable* nodes in one group than the other.
  - But as usual in networks, we focus on the value of the *link* between B and C. Since they are equally able to *deny access* to the link, they are equally strong as gatekeepers for that link.
- B's greater strength is due to its gatekeeping for two subnetworks, while C and D control only one each.
- In fact, B is also a gatekeeper for C's access to D.



Centrality and neighborhood overlap

- Another is the centrality of A *in* its group.
- An important aspect of social (including economic) networks is *centrality*. This applies to both nodes and edges. Simple binary notions are those of nodes that are *pivots* and *gatekeepers*, and edges that are (local) *bridges*.
- Earlier we discussed *neighborhood overlap*, which is a quantitative generalization of bridge. Connected nodes with low neighborhood overlap may function like gatekeepers for each other.



Centrality and embeddedness

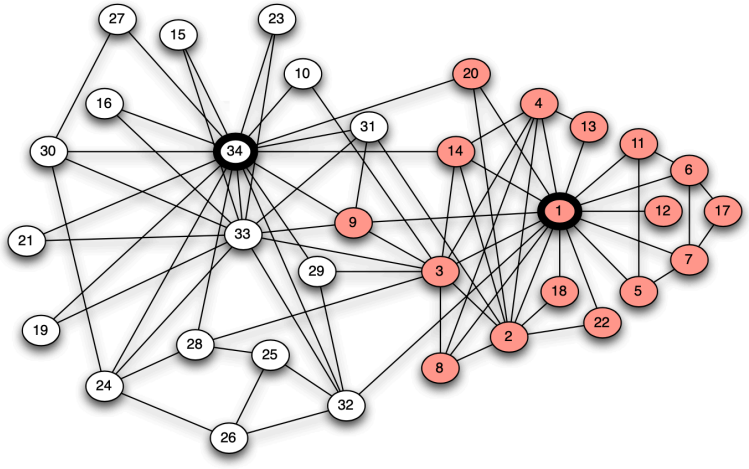
- We define an absolute version of *neighborhood overlap*, called *embeddedness*. The *embeddedness of an edge* is the number of common neighbors of the edge's endpoints, $|N(A) \cap N(B)|$.
  - An edge with embeddedness zero is a local bridge.
- Kleinberg and Easley write:

  > [W]hat stands out about A is the way in which all of his edges have significant embeddedness. A long line of research in sociology has argued that if two individuals are connected by an embedded edge, then this makes it easier for them to trust one another, and to have confidence in the integrity of the transactions (social, economic, or otherwise) that take place between them [117, 118, 193, 194, 395].

- Note that neighborhood overlap is interpreted as *power* while *embeddedness* is interpreted as *trust*.

Graph partitioning

- Links are defined by friendship (each end declares the other to be a friend).
- The two nodes with heavy borders have special roles. Node 1 is the instructor (who is authorized to test candidates and bestow rank), and Node 34 is the club president.
- The club split into two clubs while under study
  The node colors indicate who joined which club.
- Could this split have been predicted from the friendship structure?



Betweenness, flow, and partitioning

- Betweenness can be defined in terms of an abstract flow. Imagine that one unit of fluid flows between each pair of nodes, say A and B. This flow is equally divided among all shortest paths between A and B.
- The *betweenness* of an edge E is computed as follows:
  1. Find all shortest paths between all pairs of nodes.
  2. For each pair of nodes, determine the amount of flow through E. (A pair with no path through E will contribute 0.)
  3. Sum flows through E over all node pairs to get *betweenness of* E.

Calculating betweenness

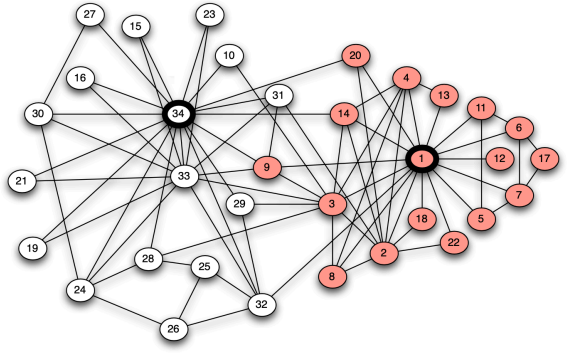
- This is a typical example of how algorithms often build on other algorithms.
- To find *all* the shortest paths from A to *all* the other nodes, perform the breadth-first search algorithm, labeling each edge with its distance from A when it is traversed.
- When each node is reached, check if this is a shortest path. If not, ignore that edge. If so, make a copy of the node reached, and attach the edge to it.
- This new graph is a *tree*, so it is easy to list all the paths. Each node in this tree corresponds to a path, and the path connects A to the node. Group by node labels.

Girvan-Newman partitioning

1. Calculate betweenness for all edges.
2. Remove all edges with highest betweenness.
3. If the graph becomes disconnected, this is the n-th step partition.
4. If there are any edges left, go to 1.
5. Stop.

Test of Girvan-Newman algorithm

- The actual division of the clubs members is given in the image.
- Girvan-Newman predicts almost the same division, except that as suggested by its position in the image, Node #9 is predicted to join the component led by #34, but instead joined the component of #1.
- With other information, we can guess why the algorithm failed. #1 is teacher qualified to evaluate a member in a test for *shodan*, *i.e.*, the "first-degree black belt." At the time of the split, #9 was preparing for the test in a few weeks. #34's group had no qualified teacher.



Algorithms *vs.* intuition

- Don't be mistaken: the "right way" is algorithms *and* intuition.
- A skilled consultant could have *predicted* the outcome. This is the bread-and-butter of the Shako-trained consultant: use a scientific method (computer, data, and algorithm in this case) to get a "good" answer, then apply deep knowledge and human intuition to get a better one.
- It's not obvious, but besides checking whether an *incorrect* prediction is "explained" by other factors, we should be careful to see if *correct* predictions are "explained" by other factors (often called "missing variables").
- For example, suppose we discovered that all of the students in the #34 faction live in one dorm, and all of the students in the #1 faction live in another that is far away. Then it might be that the location of the dorms determines *both* the friendship patterns and the clubs' memberships.

Lecture 3, Part 3

Lecture 3 Part 3 describes the structure of the World Wide Web at several levels of detail, and its relation to the Internet. Then it turns to the problem of automatically determining "importance" of web resources.

The structure of the Internet

- Why is it the *Inter*net? Why is it *the* Internet?

  These are the big questions about the structure of the Internet, which is no longer just a computer-to-computer network. The Internet, the telephone system, and cable television networks have merged (and created new infrastructure such as "the cloud" as well).

- The second we have discussed already. Networks create network externalities. Even if not as dramatic as Metcalfe's Law, each node in the network gets greater benefits as the size of the network (number of nodes) increases.

- There are very large economic benefits to network growth, or to merging networks that were originally separate, even if they seem to be of different "kinds" (such a data networks, voice networks and video networks).

The structure of the Internet, *cont.*

- It's the *Inter*net because it is an *internet*: a network composed by connecting networks together.
- These networks
  1. use different *physical* media (wireless, fiberoptic, metal wires)
  2. use different *protocols* (LTE, 5G, Ethernet, token ring, PPP)
  3. are maintained by different *organizations* (telephone companies, railroads, power companies, universities, corporations, even you.
- This is all very complex. Ask me about *email*, for example (I have about 35 separate standards for email bookmarked!) It isn't really possible to discuss the network structure of the Internet without the details. So for detail, we'll consider a much more uniform network which is more purely an *information* network: the World Wide Web.

What is the World Wide Web?

- The World Wide Web (WWW, or just "the web" for short) is an information network supported by the Internet.
- It is enabled by four central *protocols*, which (on the Internet) is a fancy way of saying "formats for information." The Internet handles the details of moving information around, while the WWW focuses on what information we want to convey.
  1. The *Domain Name Service* (DNS) allows us to access other systems on the Internet by name.
  2. The *Universal Resource Locator* (URL) system gives names to resources we want to access.
  3. The *HyperText Transport Protocol* uses the Internet to move information around as packets of bits.
  4. The *HyperText Markup Language* (HTML) provides the ability to link resources together.

The Domain Name Service

- The Internet Protocol (IP) addresses used by Internet hosts (both servers and clients) are either 32 bits (4 bytes, same as most emoji) in version 4, or 128 bits in version 6.
- These addresses have a numerical structure that made it easy to direct information to the right place—easy for a computer, that is. I used to remember several IPv4 addresses because they were local and I had to teach them to my computers, but 128 bit I wouldn't. Even in my generation few people remembered IPv4 addresses.
- The *Domain Name Service* (DNS) allows us to access other systems on the Internet by name. It translates names humans can remember to IP addresses that allow for fast routing of packets.
- It is implemented by a system of *nameservers* maintained by Internet providers.

Universal Resource Indentifiers

- The *Universal Resource Identifier* (URI) system gives a flexible consistent way to name the resource we want.
- One kind of URI is the Universal Resource Name (URN). You are probably familiar with the DOI (*digital object identifier*) scheme used to identify journal papers as one example. The point is that you don't need to know where the paper is. Each URN scheme makes up its own rules about locating resources.
  - URNs are often used for *distributed archives* or even *peer-to-peer* systems such as Usenet (netnews).
- You're probably more familiar with the *Universal Resource Locator* (URL). A URL has a standardized structure, composed of a scheme, an authority, a path, a query, and a fragment.

Universal Resource Locators

- It lookes like this:

  `scheme://authority/path1/path2?key1=value1;key2=value2#fragment`

- The *scheme* says *how* to use the Internet to fetch the resource.
- The *authority* (usually a DNS *domain* naming a server) decides if access is allowed, and if so, interprets the path, query, and fragment.
- The *path* tells the server where to get the resource. The format is modeled on a file system consisting of a hierarchy of folders.
- The *query* ("key1=value1;key2=value2") directs dynamic features such as database lookups.
- The *fragment* identifies a portion of the whole resource. It's typically used to jump to a position on a page (including internally, for end notes and bibliographies).

The HyperText Transport Protocol

- The *HyperText Transport Protocol* (HTTP) is implemented by every webserver and client. It uses the Internet to move information around as packets of bits. It's very flexible.
- It's *bidirectional*: with permission of the authority, able to upload and download resources.
- It can be used to *stream* resources as they become available, and to conduct "conversations."
- There used to be quite a few protocols used on the Internet to upload and download resources, such as the *File Transfer Protocol* (FTP), Gopher, remote copy (rcp), even email, as well as more specialized protocols such as DNS, and the *Network News Transport Protocol* (NNTP). Recently more and more of these specialized systems have been converted to use HTTP to transfer their data.

The HyperText Markup Language

- The *HyperText Markup Language* (HTML) provides the ability to link resources together.
  - It also provides many features for formatting and presentation, especially in combination with *cascading style sheets* (CSS). These features are not relevant to our concern with information networks.
- Linking is in the name *hypertext*, coined by Ted Nelson to mean a linked document system in 1963.
- In HTML, links are asymmetric (arrows).
  - The WWW is a *directed multigraph.*
  - In Nelson's hypothetical hypertext system *Xanadu*, links were automatically symmetric.
  - In HTML symmetric links can be emulated with a link in each direction, and these are frequently used for footnotes and bibliographic citations.

What is "Web 2.0"?

- "Web 2.0" isn't really a change in the web itself. The Internet itself and the WWW protocols have evolved, but to content creators and users alike, it doesn't look much different.
- The original WWW was "static", because both hardware and browsers were slow.
  - Slow hardware meant that content on *servers* was *static*: stored in files.
  - Slow browsers meant that content as *displayed* was *static*: once a page was displayed, to make it change you needed to fetch a new page (or a new version of the page).

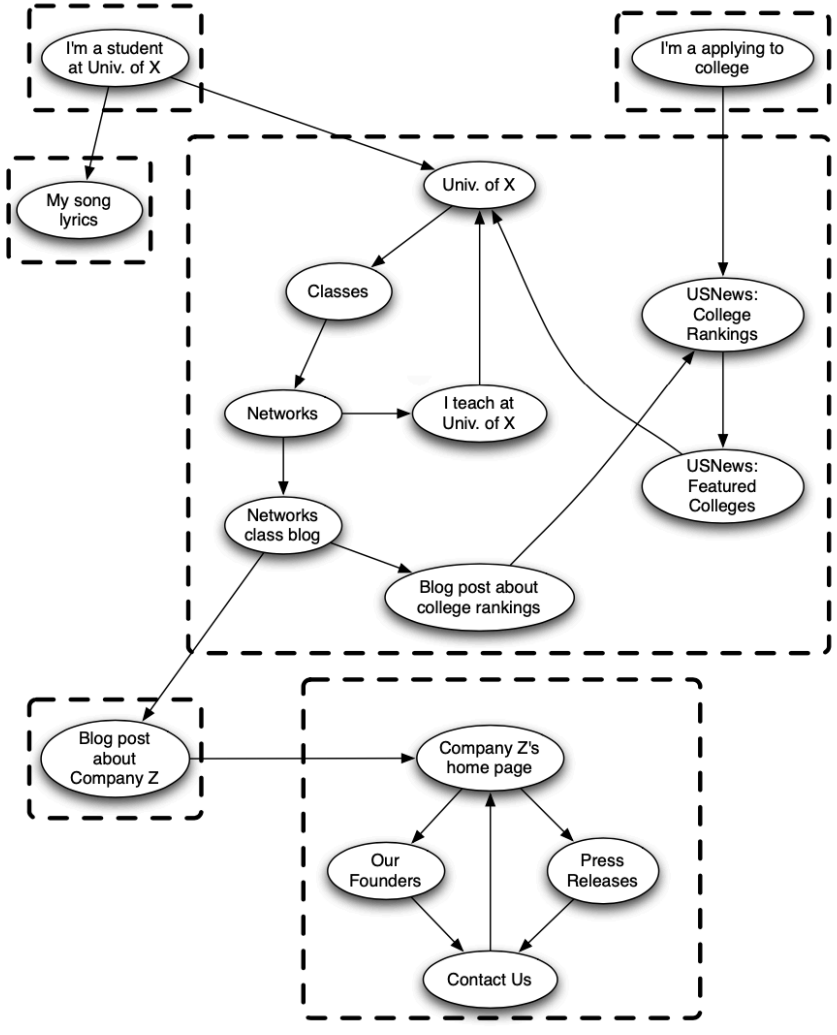
Web 2.0 is dynamic

- The dramatic difference is that "Web 2.0" is *dynamic*.
  - Servers are very likely to generate content by looking it up in a database or computing it by a program on every access, then formatting it for transmission as a web document.
  - These web documents can contain Javascript *code* that is executed by the browser to change the display, rather than fetching new content from the server.
- But the process of fetching documents, and the formatting of the documents, is the same as ever.

Fine structure of the WWW

- The WWW is a *directed multigraph* of *resources.*

There can be multiple links to the same resource from one resource. Having traversed a link, you can't go back without an external memo (such as browser history) or a reverse link (but is there are several, you can't know which is "back."



About resources and links

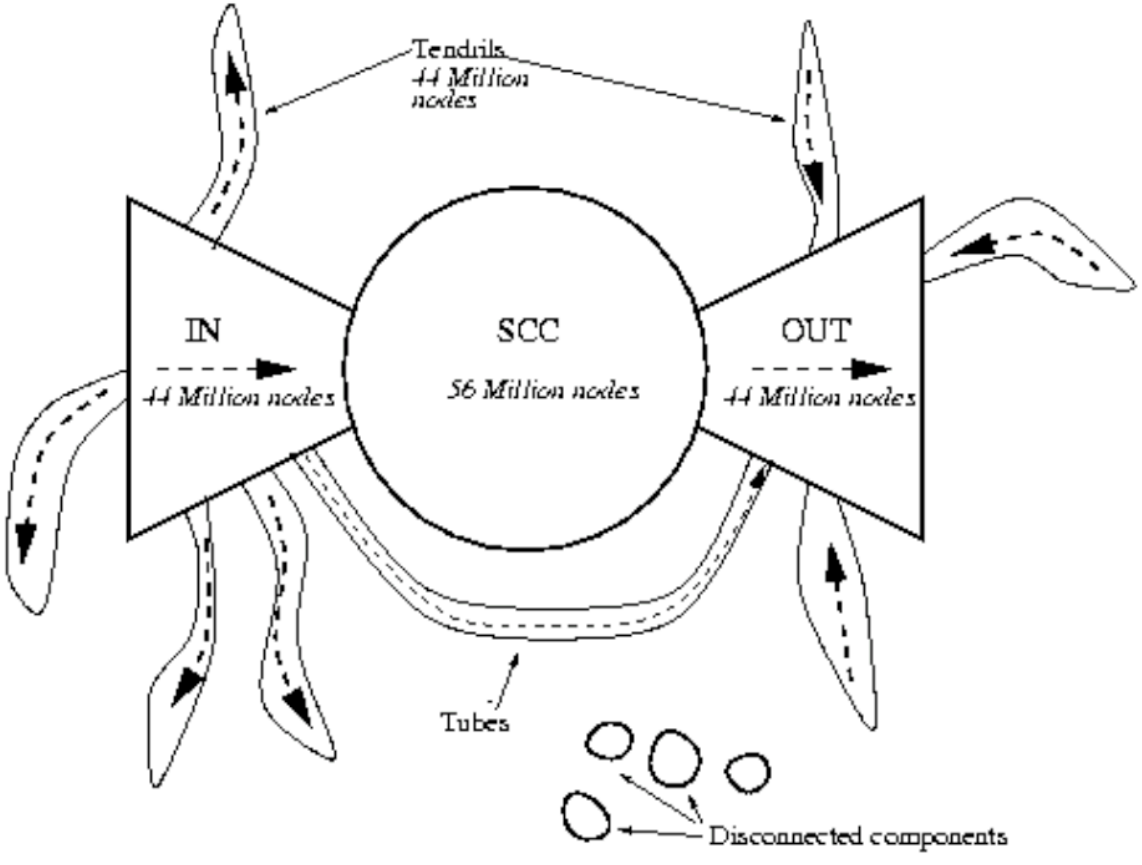
- In the WWW, a *resource* is anything that can be addressed by a URL. It doesn't even need to return anything for the client to display. Examples:
  - Text pages (HTML, plain)
  - Presentation resources (stylesheets, Javascript)
  - Multimedia (images, audio, video)
  - Arbitrary files (OA files, programs)
  - Form submissions (credentials)
  - Database lookups
- *Links* are indicated by URLs. There are two types:
  - User-initiated (`A` or anchor links)
  - Automatic (`IMG` or image links)
  - The distinction is browser-dependent

The global structure of the World Wide Web

- Although it is not a *social networking platform* as we understand that term now (*e.g.*, Facebook, Twitter), it is a platform for creating social networks.
- Linking to another page is a social activity. One can use it as a memo to oneself (that's what your browser's bookmark window is), but generally it's a service for your readers.
  - At present, just *readers*. But we already have voice recognition with Siri, Echo, and so on, and you can imagine virtual reality systems in which you can activate links by some kind of gesture, or even via eye motion.
- The Web truly is global. Some countries (Russia, China, Kazakhstan are well-known) impose substantial restrictions on access to international resources, but even they allow most connections.

Overall structure of the WWW

- The WWW is "almost" connected. It has giant component that contains most of the resources, and all of the links between them.
- The giant component has a sort of "fuzzy bowtie" structure.

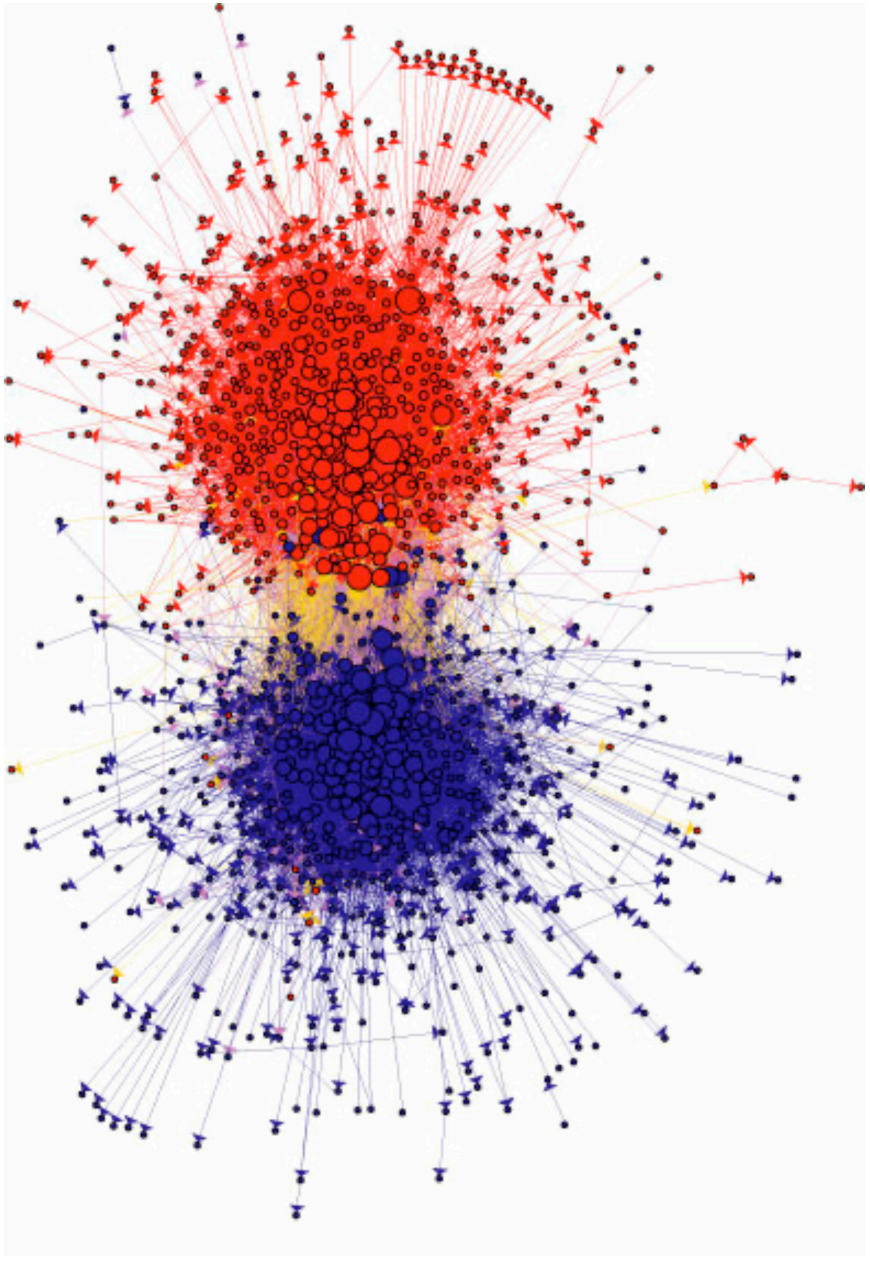


The "bowtie" structure

- This diagram is based on the state of the Web in about 2000, using the index of the AltaVista search engine operated by the Digital Equipment Corporation (DEC).
- The central "knot" of the "bowtie" is a giant *strongly connected* component (SCC, defined as a maximal subset of nodes all of which are reachable by a directed path from any other node in the SCC), containing about 1/4 of the resources (or pages).
- The **In** set of the bowtie is the set of nodes from which the SCC is reachable.
- The definition of the **Out** set is an exercise.
- A *tendril* is a set of related paths not part of the SCC, which either
  1. contain nodes all of which can reach the **Out** set, or
  2. contain nodes all of which are reachable from the **In** set.

Web subsets: Political blogs

Politics is often described as a spectrum from left to right, but the U.S. Constitution leads to a two-party system, and this seems to be reflected in political blogs.



Page Rank and other mysteries

- It's very common for a Google search to claim millions of matches for a search. How do they decide which results to return first?
- It's complex, but one factor is the "importance" of the resource. The measurement of importance is still a hot topic of research, both in academia and as a trade secret of many companies. It is closely related to many other concepts of importance such as the impact of an academic publication.
- The obvious algorithm is described as *voting by in-link* in Kleinberg and Easley. The idea is simple: take all the pages that contain your search key, and rank them according to the number of times they're linked from other pages.
- The obvious algorithm has a major flaw. Some pages get many links, regardless of their content.
- There are two subtly different approaches: *hub and authority* ranking, and *Page Rank*.

Hub and authorities

- This approach divides pages into two kinds.
- *Authorities* are the pages that are likely to give you the information you want.
- *Hubs* are pages that have many links to possible authorities.

Hub and authority algorithm

The algorithm is *iterative*, with two *updating rules* and a *stopping condition*.

1. *Initialization.* Give each page a *hub score* of 1, and an *authority score* of 1.
2. Apply the *authority update rule*: Each page's new *authority* score is the normalized sum of the *hub* scores of pages that link to it.
3. Apply the *hub update rule*: Each page's new *hub* score is the normalized sum of the *authority* scores of pages that it links to.
4. If the stopping condition is satisfied, stop, and rank authorities by their authority score.
5. Otherwise, repeat from Step 2.

Page Rank

- The hub and authority algorithm makes especially good sense when ranking is extremely competitive, for example in *recommender* algorithms for products by competing vendors. The product home pages aren't going to link to each other! So the asymmetry of hubs and authorities in the algorithm matches social behavior.
- But in many other situations, a node will be considered "important" if it's endorsed by other "important" nodes.
- *Page Rank* is an algorithm invented by Google's founders and commonly used as the basis for many commercial and academic endorsement systems, based on this more symmetric idea.
- Page Rank is like a fluid that circulates through the network. It accumulates at the most endorsed pages.

Basic Page Rank algorithm

The algorithm is *iterative*, with one *updating rule* and a *stopping condition*.

1. *Initialization.* Give each of n pages a *Page Rank* of $1/n$.
2. Apply the *basic Page Rank update rule*: Each page divides its Page Rank equally among the pages it points to (keeping it all if it has no out-links), and updates its Page Rank to the sum of Page Rank it receives.
3. If the stopping condition is satisfied, stop, and rank pages by their Page Rank.
4. Otherwise, repeat from Step 2.

This algorithm has a major flaw: if there are strongly connected components with no out-links, they will collect all of the Page Rank. This would be very bad. *E.g.*, for academic papers, this means that student essays with no bibliography would accumulate all the Page Rank!

Revised Page Rank algorithm

The algorithm is *iterative*, with one *updating rule* and a *stopping condition*.

1. *Initialization.* Give each of n pages a *Page Rank* of $1/n$, and choose a number $s \in (0, 1)$.
2. Apply the *revised Page Rank update rule*: Each page takes s times its Page Rank and divides its Page Rank equally among the pages it points to (keeping it all if it has no out-links), takes the remaining $1 - s$ of it Page Rank and divides it equally among *all* pages, and finally updates its Page Rank to the sum of Page Rank it receives.
3. If the stopping condition is satisfied, stop, and rank pages by their Page Rank.
4. Otherwise, repeat from Step 2.

This algorithm gives a different Page Rank vector for each value of s.
For $s = 0$, it's equivalent to in-link voting.
For $s = 1$, it's exactly the basic Page Rank algorithm. In practice, s between 0.8 and 0.9 seems to give good results.

Homework #10

Due: December 7, 2023 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #10 0AL0200`.

Consider the "bowtie" image of the World Wide Web.

1. Why do you think the *disconnected components* at the bottom don't have an estimated number of resources?
2. Give the definition of the "out" side of the bow.
3. How are *tubes* related to *tendrils*?
4. Why can't a tube run from the **Out** set to the **In** set?
5. (Requires knowledge of Internet servers, if you don't know, just say so) How do you think they got information about the **In** set?

Homework #11

Due: December 7, 2023 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #11 0AL0200`.

The graph of the **political blog space** shows **Out** fringes for both red (progressive) and blue (conservative) "components," but no **In** fringe.

1. Why do you think that is? (Hint: It's probably related to data collection method.)
2. Are there any tendrils in this graph? If "yes," describe where they are. If "no," explain why not?
3. What can you say about the depth of the **Out** fringes? Can you explain why this might be?

Homework #12

Due: December 14, 2023 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #12 0AL0200`.

Consider the **hub and authority** algorithm for evaluating page "importance."

1. Why do the update rules have "normalized" in them?
2. Give a formula for normalizing scores. It must preserve ranking according to score.
3. Give as many reasonable stopping rules as you can think of. (I can immediately think of one exact stopping rule and two kinds of approximate rule.)

Homework #13

Due: December 14, 2023 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #13 0AL0200`.

Consider the **Page Rank** algorithm for evaluating page "importance."

1. Why doesn't either Page Rank update rule need normalization?
2. Are the stopping conditions you designed for the hub and authority algorithm suitable for the Page Rank algorithms? Explain.
3. Do you think the "circulating fluid" analogy to Page Rank is a good one? Explain.

Lecture 3, Part 4

In Lecture 3, Part 4 we look at information cascades: dynamic processes where information spreads through a network. These processes are related to "flame wars" and "disinformation campaigns" on social networks, as well as to some kinds of behavior in markets.

Behavior and information

- Networks connect people, and allow them to respond to others' activity.
- This is fundamental to *economics*, to "social engineering," and in fact to any activity that takes place in groups.
- In *social* behavior (including economics), the effect of another's behavior is not physical (except in war and competitive sports); it's informational. We respond to the *information* about others' behavior.
- Sometime the behavior itself is pure information: inspiring speeches, online bullying, or disgusting and dangerous disinformation about COVID vaccines.

Information cascades

- An example most of us have experienced is walking down a "food alley" in a shopping center, mall, arcade, or downtown, and noticing restaurants that have lines, sometimes implying a wait of hours.
- Of course, one response is "the wait is too long, let's go somewhere else." That's based on a physical (or regulatory) fact: the restaurant will allow only so many people in.
- Another, however, is "wow, the food must be delicious." Even if you pass it by today, maybe you put it on your "list" for when you have more time.
- But consider: you're not the only one who thinks that way. Probably, some of the people in line are there for that same reason! So you (who have not eaten there) think that it's delicious because other people (who have not eaten there!) think it's delicious because they joined a long line.
- *That* is an *information cascade.*

A simple numerical example

- Suppose all diners choose at random from the restaurants. Then no information cascade can occur. The distribution of line lengths will be approximately normal, according to the Central Limit Theorem.
- Suppose all the restaurants are different styles, and everyone chooses the food they want. No cascade is possible.
- Suppose all the restaurants are the same style, and everyone chooses the restaurant they think is most delicious. Then you can safely choose the restaurant with the longest line.
- Suppose half the diners know which they think is delicious, and half choose randomly among the restaurants with the longest line. This supports a *pure* cascade.
- Information cascades may be based on *statistical inference*.

"Herding" and information cascades

- *Herding* is a more general term used in psychology and sociology to denote imitative behavior.
- In the case of animals like sheep, herding is instinctive. These fields tend to interpret human herding as either instinctive or socially reinforced conformism.
- There may be some cases of instinctive herding in humans, even in economics, but as we see with the restaurant example an information cascade can induce *rational herding*.
- Kleinberg and Easley give many examples:

  Fashions and fads, voting for popular candidates, the self- reinforcing success of books placed highly on best-seller lists, the spread of a technological choice by consumers and by firms, and the localized nature of crime and political movements can all be seen as examples of herding[.] (p. 484)

Network externalities *vs.* information cascade

- Note that network growth also looks like herding behavior.
- Unlike conformism, it is based on rational choice.
- However, unlike information cascades, it is based on *direct benefits*. You can't go wrong by joining the bigger network under Metcalfe's assumptions about the benefits from the network. You *can* make a mistake in the information cascade case, and under some circumstances you will.

An experiment

Two urns are filled with red and blue colored balls. One is 2/3 red, the other 2/3 blue. We pick one with equal probability, and put the other away. We have different people independently draw a ball, look at it without showing anyone, guess the color of the urn so that all can hear, and replace the ball. Then go on to the next person. Each person who guessed correctly gets ¥1000, the others nothing.

- The first person draws a ball. Suppose they see blue. (The same argument follows with colors exchanged if they saw red.)
- To maximize expected profit, they should guess blue.
- Suppose the second person also sees blue. Obviously they also want to guess blue.
- If they saw red, then they have a split sample, and could guess either. Let's suppose they break the tie by guessing the color they saw.
- Now suppose the first two both saw blue. We know that profit maximizing subjects will both report blue. If the third person *also* sees blue, obviously they should guess blue.

- But suppose the third person saw red. In that case, two people say they saw blue, and they have a strong incentive to guess the color they saw. So it's as if the third person drew three balls, two of which are blue and one red.
- In this case, the evidence is that the urn is majority blue (two blue balls to one red ball), so the third person should guess blue—no matter which color they saw.
- How should the fourth person think? The first two guessed blue, and we know they're trustworthy. But the third always wants to guess blue, so the fourth wants to ignore their guess. But no matter what she sees, a majority of the balls she has seen or can deduce the color are blue, so she wants to guess blue regardless of the color she saw.

Cascades and math: Bayes' Law

- We know that in such situations the quantitive effect of receiving information (denoted **message**) on our belief about the unknown **state** is given for each possible state by *Bayes' Law*:

$$\mathcal{P}[\textbf{state}|\textbf{message}] = \frac{\mathcal{P}[\textbf{message}|\textbf{state}]}{\mathcal{P}[\textbf{message}]}\mathcal{P}[\textbf{state}]$$

- The logic is that if the probability of a **message** in a given situation **state** is higher than our current assessment of the overall probability that we receive **message**, then the fraction is bigger than 1, and tells us how much we should increase the probability we assess for that **state**.
- In this equation, $\mathcal{P}[\textbf{state}]$ and $\mathcal{P}[\textbf{message}]$ are not necessarily *base rates*. We can also think of them as *prior probabilities* based on repeated updating as we receive message after message, up to just before we receive the current **message**.

Cascades and math: Bayesian analysis

- Where do the probabilities come from?
  - *Base rates* are generally measurable by counting.
  - The conditional probability of message given state comes from some scientific model, and some measurement.
  - The conditional probability of state given message is computed from Bayes' Law.

  Kleinberg and Easley, Ch. 16, gives an extended discussion of a realistic example of a witness with imperfect visibility testifying about an accident involving a taxicab.

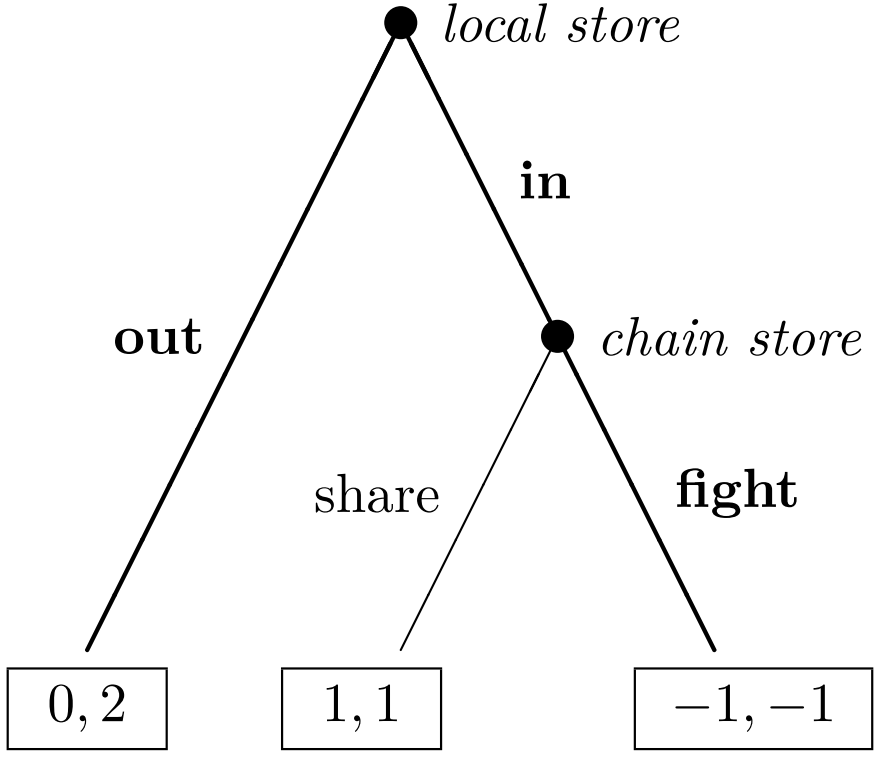
Cascades and math: Bayesian games

- In a situation where people make *interacting decisions* (a *game*), we use optimization theory to characterize their *optimal decisions*, then place probabilities on their choice among optimal decisions.
- An *equilibrium* occurs when the probabilities we place on all optimal choices give all players no reason to change their behavior.
- In *Bayesian game theory*, we suppose people have different types (*e.g.* **informed** *vs.* **uninformed**), and we estimate probabilities of their types and informational state based on what we observe of their behavior and what we know of their incentives for different behavior in each situation.
- A very famous example is the so-called *Chain Store Paradox*, which analyzes an example in which a player strategically induces an information cascade in order to justify behavior that is irrational but profitable.

The chain store paradox stage game

	fight	share
in	$-1, -1$	$1, 1$
out	$0, 2$	$0, 2$

local store payoff listed first.



Summary of information cascades

This list includes that of Kleinberg and Easley section 16.7.

- Some situations make cascades very likely.
- Cascades can be wrong.
- Cascades can be based on very little information.
- Cascades can be fragile.

Homework #14

Due: December 14, 2021 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #14 0AL0200`.

In the restaurant example of an information cascade, consider a situation where there are *two* restaurants, A and B. You think that 50% of all people are **informed**: they know which restaurant is better. 50% are **uninformed**: they guess that the restaurant with the longer line is better. (If the lines are the same length, they choose with uniform probability.) The type of each person who arrives is independent of all the others.

How should you assess the probability that the long line waits at the good restaurant?

1. What are all the possible orders of types **informed** and **uninformed** arriving at the restaurants? (There are 8.)
2. What is the probability of each order of arrival?
3. Considering that the knowers choose A and the guessers may choose randomly, for each order, find the distribution of allocations of diners to

restaurants. "Allocation" means which line each person joins. "Distribution" means that for each allocation you must compute the probability of that allocation.

4. Reduce the distribution of allocations (specific people) to the distribution of line lengths at Restaurant A (of course, Restaurant B has 3 minus that length waiting).

5. For each line length, what is the probability you choose the right restaurant?

6. What is the overall probability you choose the right restaurant? Is it better than 1/2?

7. Is choosing the long line a good strategy if you want to eat good food?

8. Are there *pure* cascades in this example?

Homework #15

Due: December 14, 2021 at 11:00. Submit to `netecon-hw@turnbull.sk.tsukuba.ac.jp` with `Subject: Homework #15 0AL0200`.

Consider the experiment with an urn containing red and blue balls, as presented in the lecture.

1. What does the fifth person think if they see blue? What if red?
2. What happens if the first two split, one red, one blue, and the third sees blue?
3. What can you say about the sequence of guesses in general?