CSC200: Lecture 3

Allan Borodin

Announcements

- This week, tutorial on Wednesday. Next lecture on Friday in this room. Thereafter we will mainly stay with scheduled lectures on Monday and Wednesday and tutorials on Fridays.
- We will start with two tutorial sections and if warranted start a third section.
- North side of room will have tutorial in the lecture room. South side of room will have tutorial in SS 1088.
- My office hours will be Tuesday 2-4 (SF 2303B) or by appointment; note that I may have to move office hours on occasion but will notify class (on web page and/or in lecture).

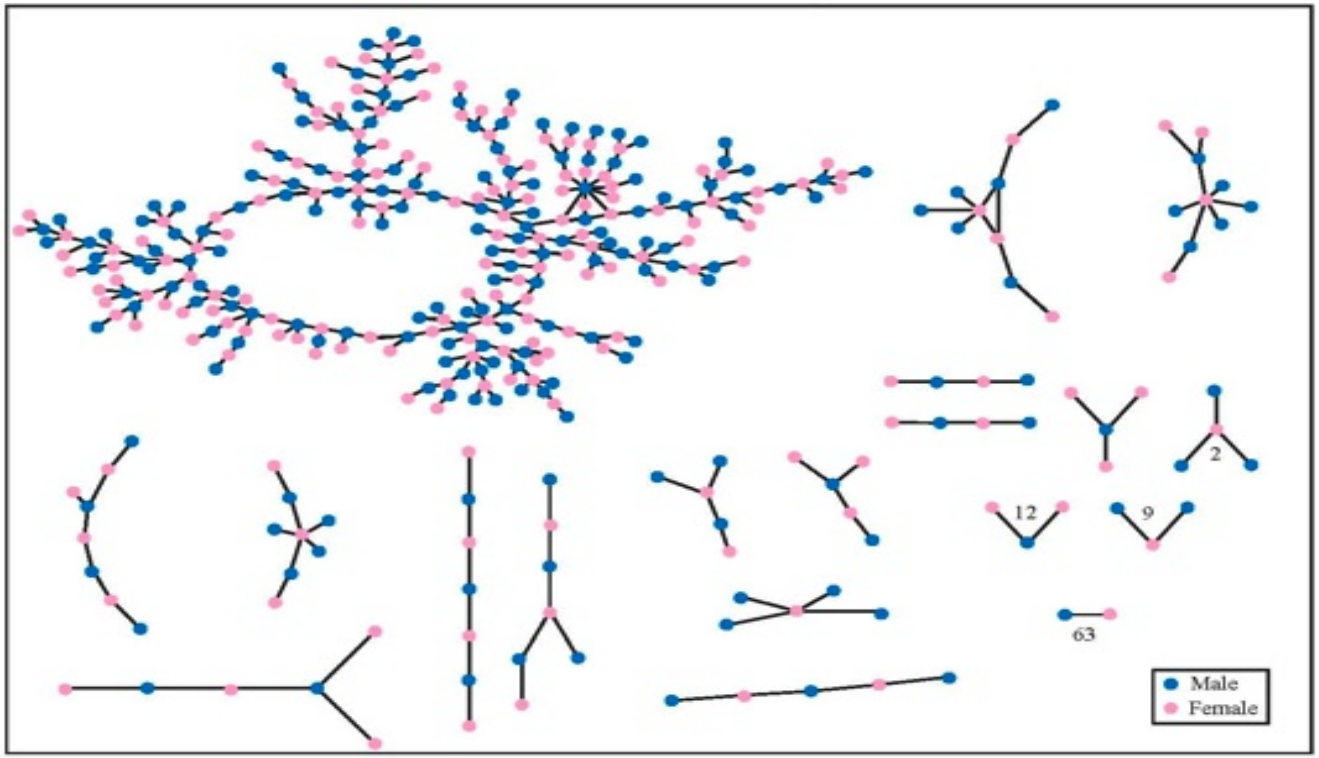
Todays agenda

- Last lecture: a number of basic graph-theoretic concepts:
  - undirected vs directed graphs
  - paths and cycles
  - bipartite graphs

- This lecture: a few more graph-theoretic concepts and then move on to Chapter 3 of the textbook on "Strong and Weak Ties".

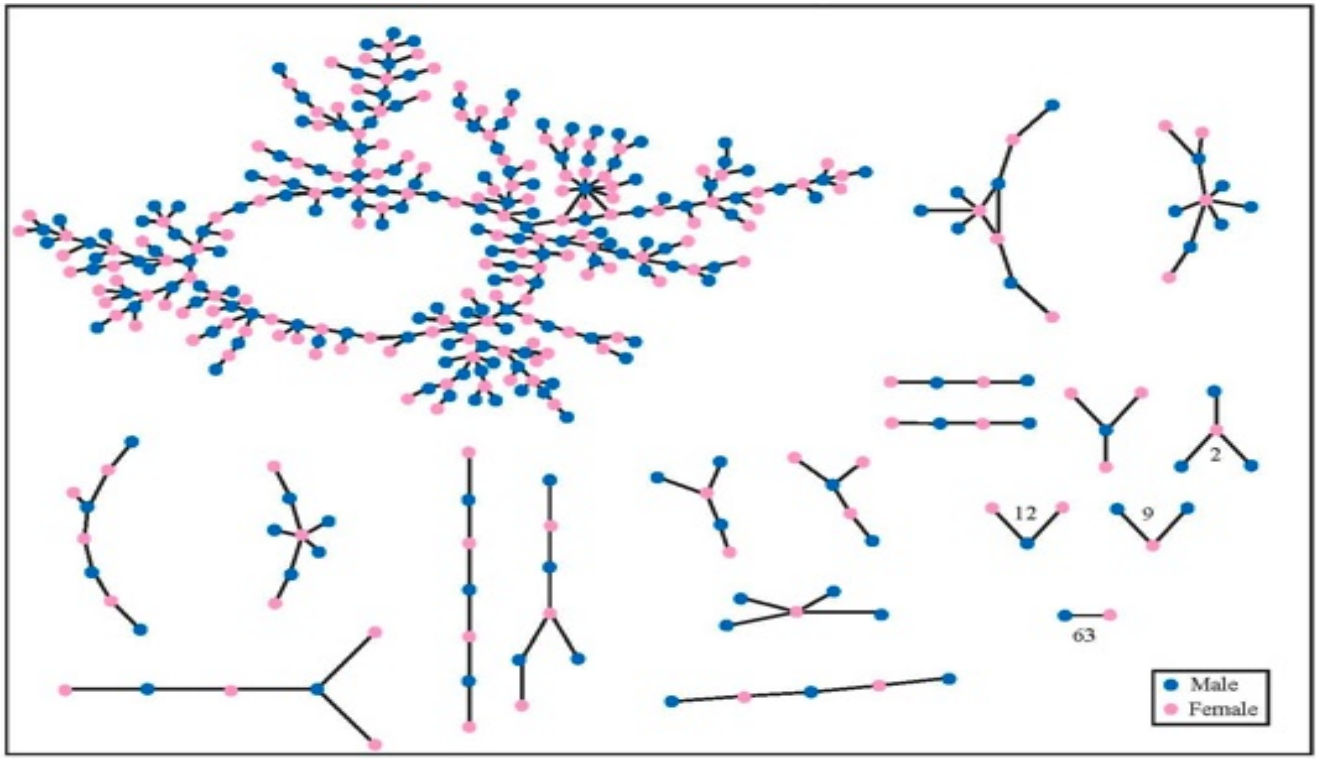
- Let's briefly return to the "romantic relations" graph to see how graph structure may or may not align with our understanding of sociological phenomena.

Observations from the last lecture



- The "giant component" has one big cycle and very few small cycles.
- Another component has a 3-cycle (i.e. triangle).
- The graph was "almost" bipartite and "almost" acyclic; indeed, from each of these cycles, there is an edge we can remove so that the graph becomes acyclic.

More observations



- Most nodes have small degree.

  degree = the number of neighbors of a node

- Observation: there are obvious reasons for not having many small cycles or nodes having large degree?

Detecting the romantic relation in Facebook

- As mentioned last Wednesday, 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

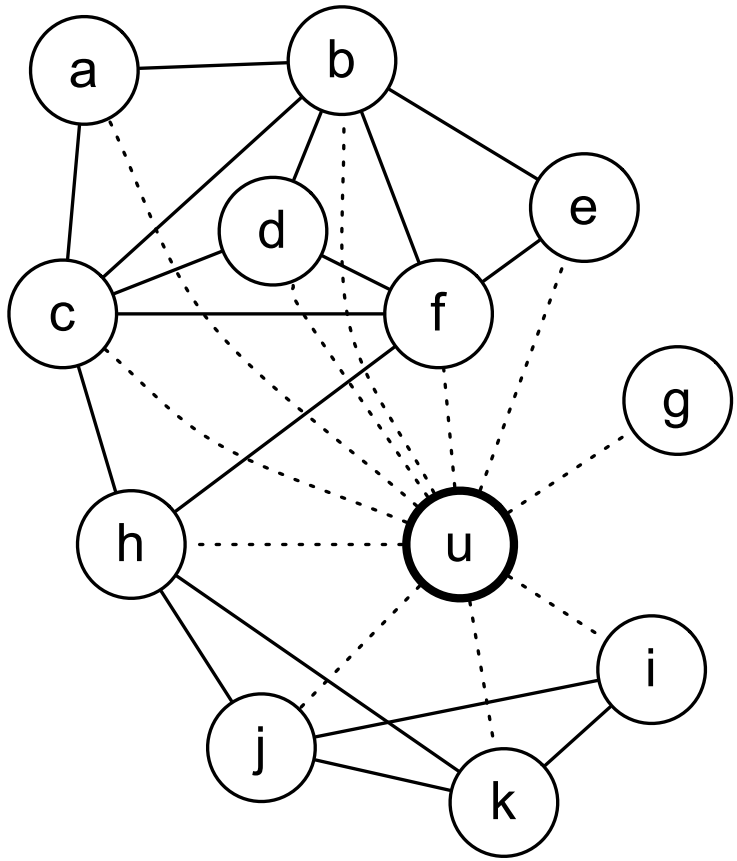


Figure 2. A synthetic example network neighborhood for a user u; the links from u to b, c, and f all have embeddedness 5 (the highest value in this neighborhood), whereas the link from u to h has an embeddedness of 4. On the other hand, nodes u and h are the unique pair of intermediaries from the nodes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

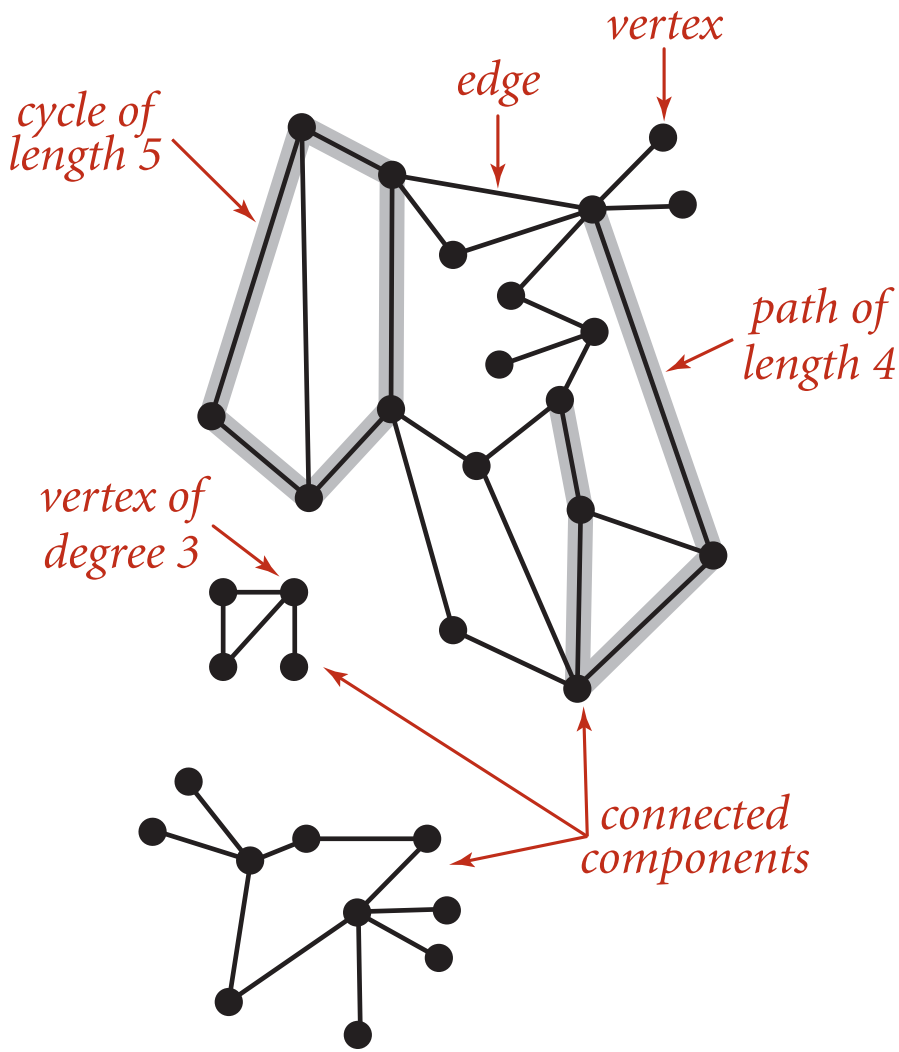
Some experimental results for the fraction of correct predictions

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

Graph Anatomy: summary thus far

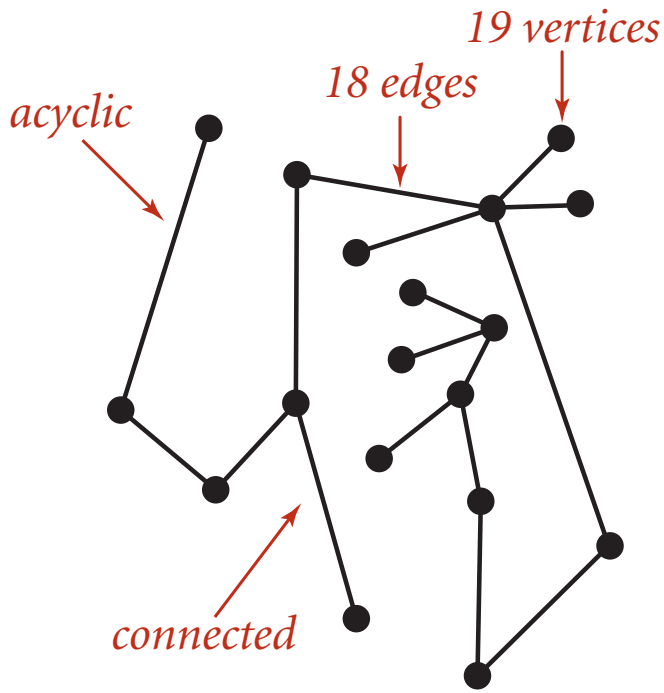


[from Algorithms, 4th Edition by Sedgewick and Wayne]

Acyclic graphs (forests)

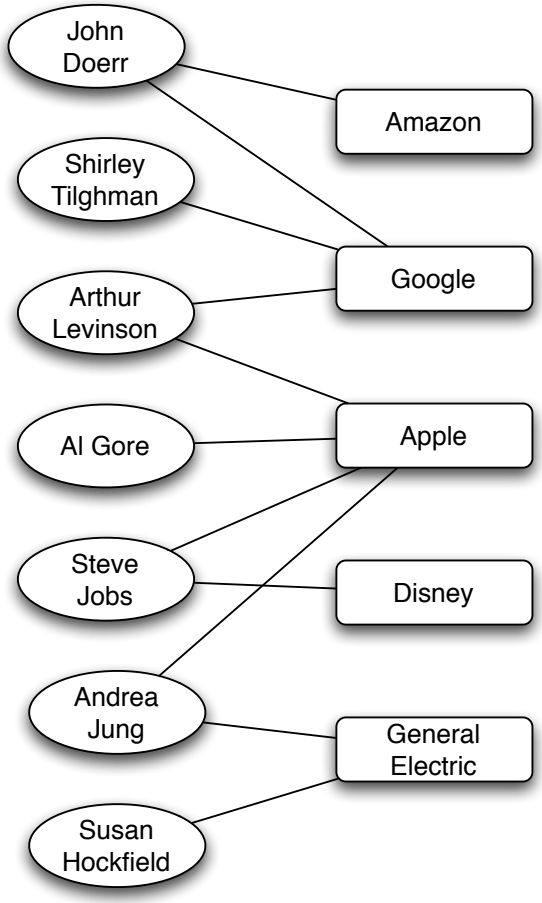
- A graph that has no cycles is called a forest.
- Each connected component of a forest is a tree.

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



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

Another tree [E&K Figure 4.4]



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

Facts

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

Analogous concepts for directed graphs

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

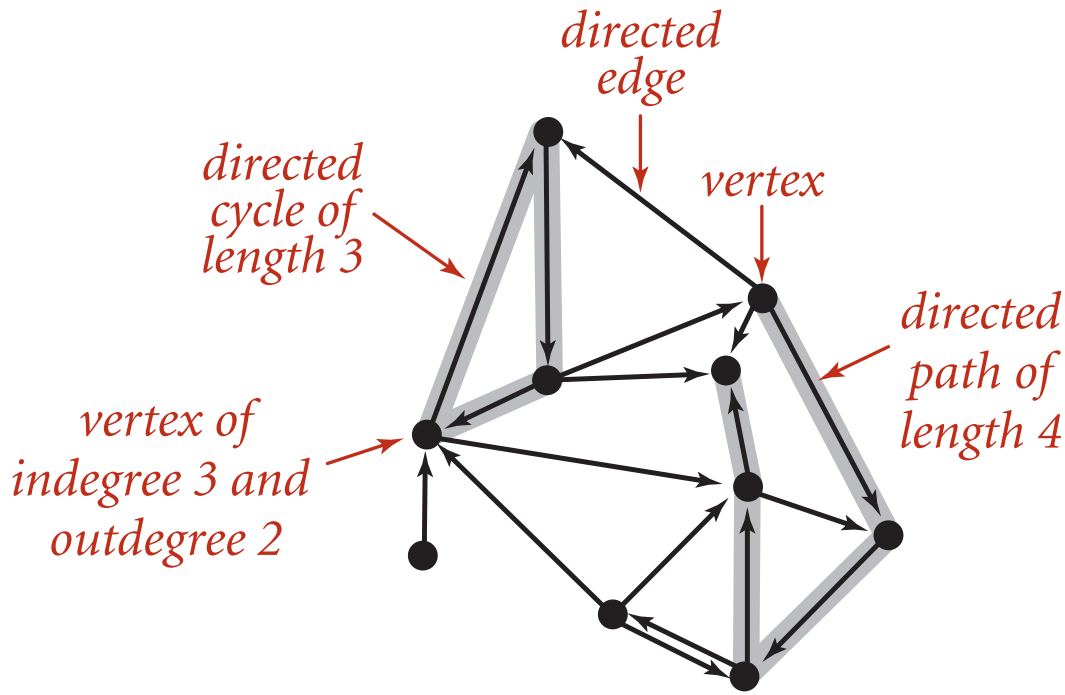
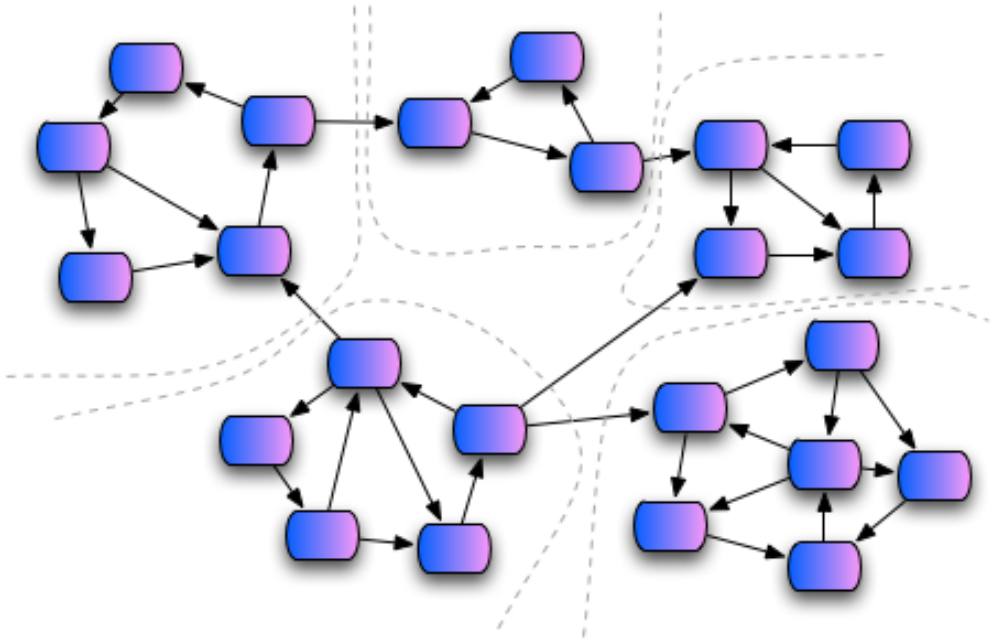


Figure : Directed graph antonomy [from Sedgewick and Wayne]

More analogous concepts for directed graphs

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

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



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

Chapter 3: Strong and Weak Ties

There are two themes that run throughout this chapter.

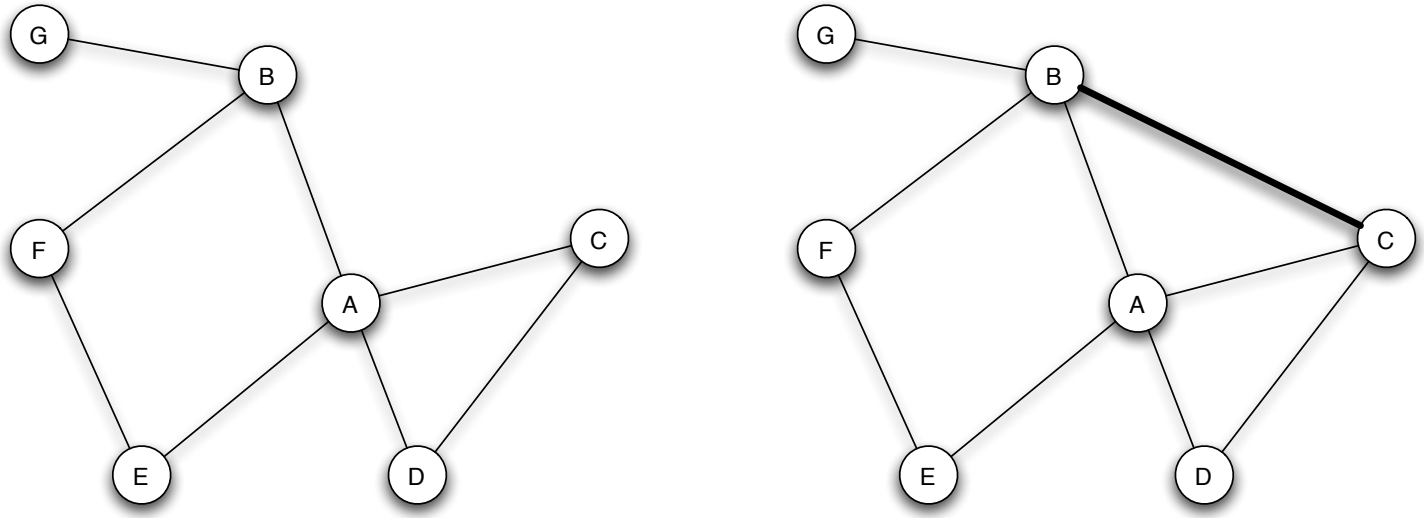
1. Strong vs. weak ties and "the strength of weak ties" is the specific defining theme of the chapter. The chapter also starts a discussion of how networks evolve.

2. The larger theme is in some sense "the scientific method".
   - Formalize concepts, construct models of behaviour and relationships, and test hypotheses.
   - Models are not meant to be the same as reality but to abstract the important aspects of a system so that it can be studied and analyzed.
   - See the discussion of the strong triadic closure property on pages 49-50 of text.

Informally

- strong ties: stronger links, corresponding to friends
- weak ties: weaker links, corresponding to acquaintances

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).

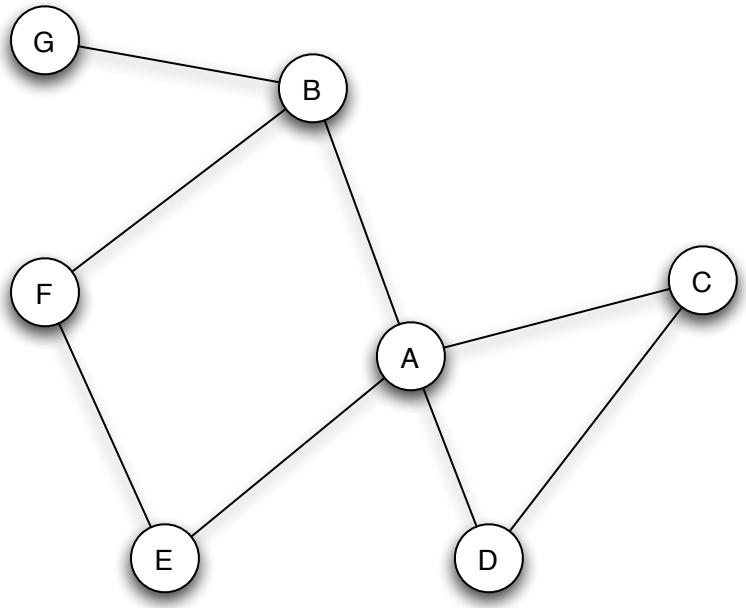
- Let E be the set of an undirected edges of a network graph. For a node A, the clustering coefficient is the following ratio:

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

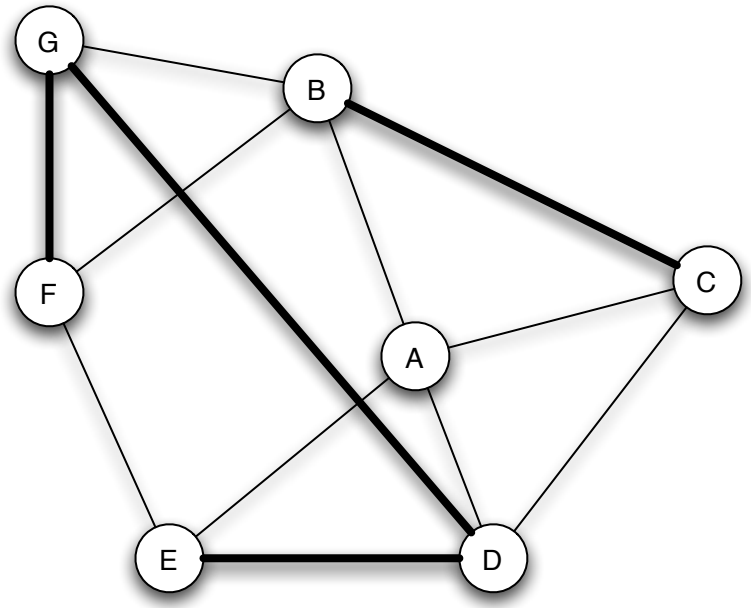
- The numerator is the number of all edges (B, C) in the network such that B and C are adjacent to (i.e. mutual friends of) A.

- 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 $\{D, E\}$)
- 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)).