CS104 Information and Information Systems

Social Networks and Graph Theory

Morgan Harvey
morgan@cis.strath.ac.uk

University of Strathclyde
Science

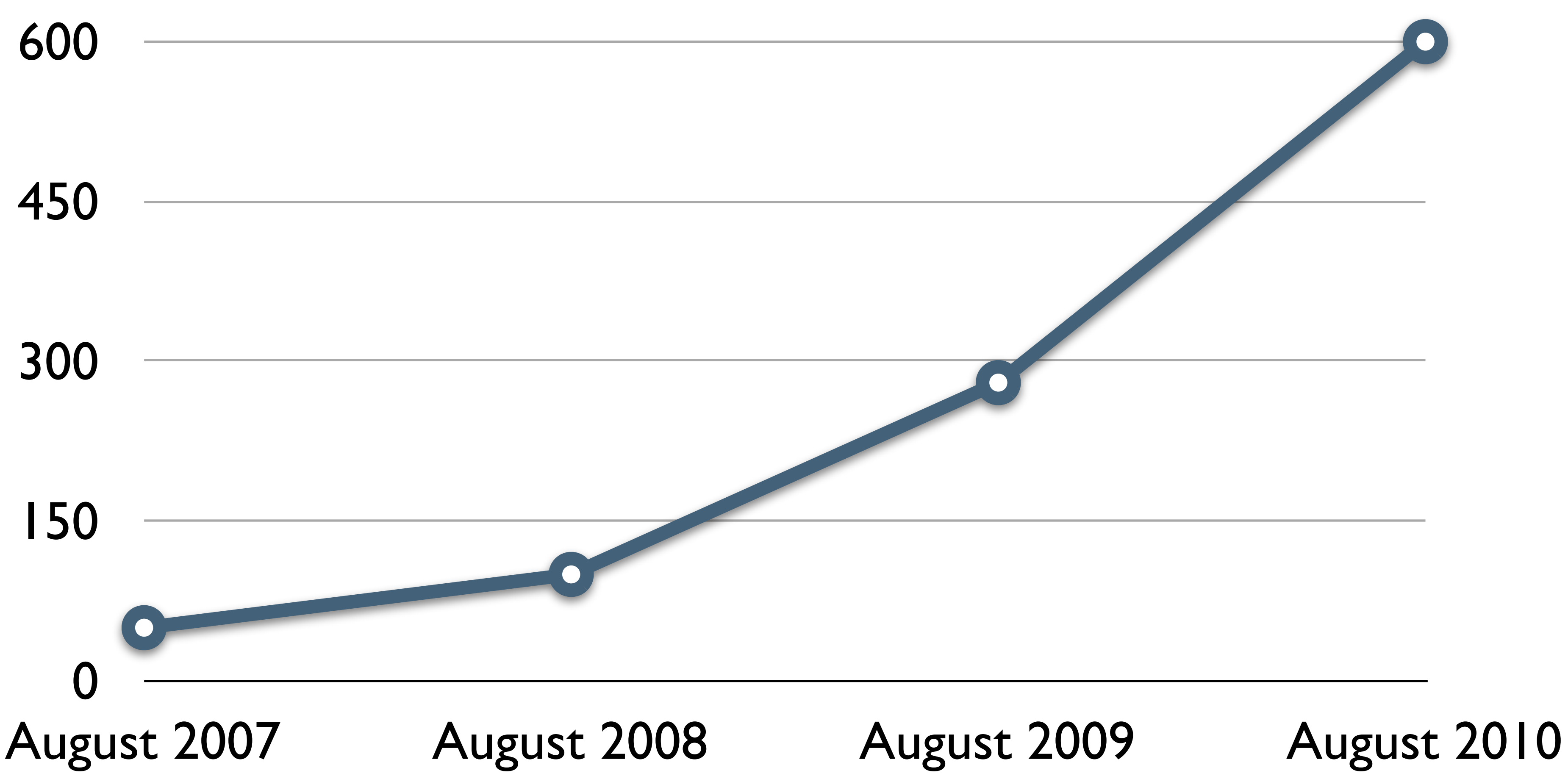
Today’s lecture

- What are social network, why are they important to study?
- How we model them - graphs
- Web as a graph, Google PageRank
- Graph properties and metrics
- Small world phenomena
- Kevin Bacon numbers


Facebook (2) | Barack Obama
http://www.facebook.com/barackobama?ref=ts
Google
LSA Stuff eBay Flickr News (3878) add to salut Research papers Recent Online Lectures current saved Flying
facebook
Search
Home Profile Account
Barack Obama
Like
Wall Info OFA Store Photos Join OFA Video
Barack Obama + others Barack Obama Just others
Suggest to friends
This page is run by Organizing for America, the grassroots organization for President Obama's agenda for change. To visit the White House Facebook page, go to http://bit.ly/2bVCm. OFA is a special project of the Democratic National Committee.
Information
Current office
Office:
President of the United States
6 friends like this.
Barack Obama This Labor Day, we recommit ourselves to this fundamental truth: to heal our economy, we need more than a healthy stock market; we need bustling main streets and a thriving middle class. I will keep working day by day to restore opportunity, economic security, and that basic American Dream.
Weekly Address: Honoring the American Worker
www.youtube.com
Labor Day is a day to honor the American worker—to reaffirm our commitment to the great American middle class that has, for generations, made our economy the envy of the world.
20 hours ago · View feedback (30,143) · Share · Report
Barack Obama We can't afford to go backward to the failed policies of the past. That's what November's elections are about -- moving forward. Commit to cast your vote.
COMMIT TO VOTE
Commit to Vote in the November Elections
my.barackobama.com
This year's election offers a stark choice. Democrats are hard at work trying to move America forward, while Republicans want to take us back to the same policies that led our country into recession.
Friday at 21:37 · View feedback (25,787) · Share · Report
Barack Obama My administration is doing everything in our power to keep people safe in advance of Hurricane Earl along the Eastern
Create an advert
Facebook Pages
Facebook Pages help you discover new artists, businesses and brands, as well as connect with those you already love.
More adverts
Chat (2)
Loading "http://www.facebook.com/barackobama?ref=ts", completed 44 of 45 items

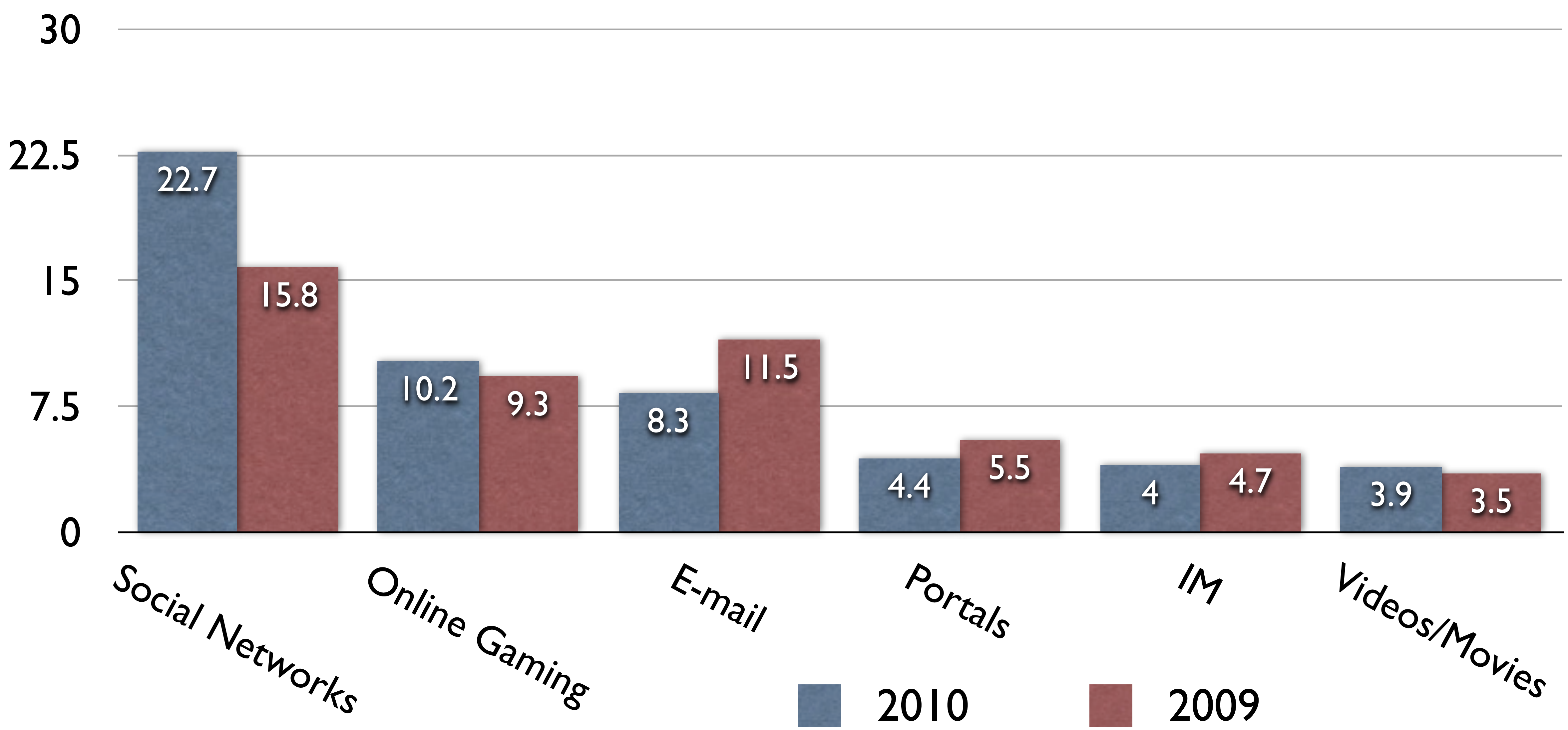
Online Social Networks

- Have recently exploded in popularity, for example FaceBook



Online Social Networks

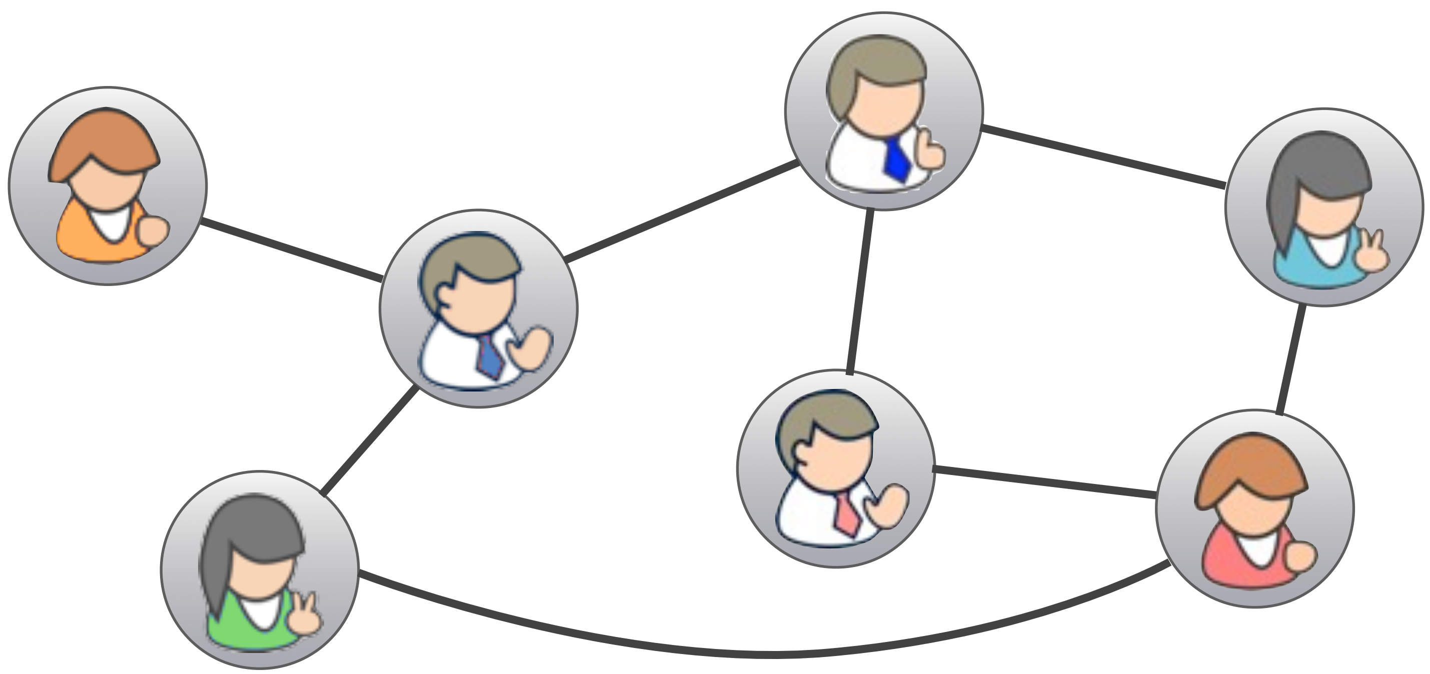
- We now spend more time online using social networks than any other activity!



Source: Neilsen

Social Networks

- Is a social structure, normally represented as a graph with:
    - Individuals (or organisations) as nodes
    - Relationships as edges
- We can usually learn a lot about people from studying their social network - *social network analysis*



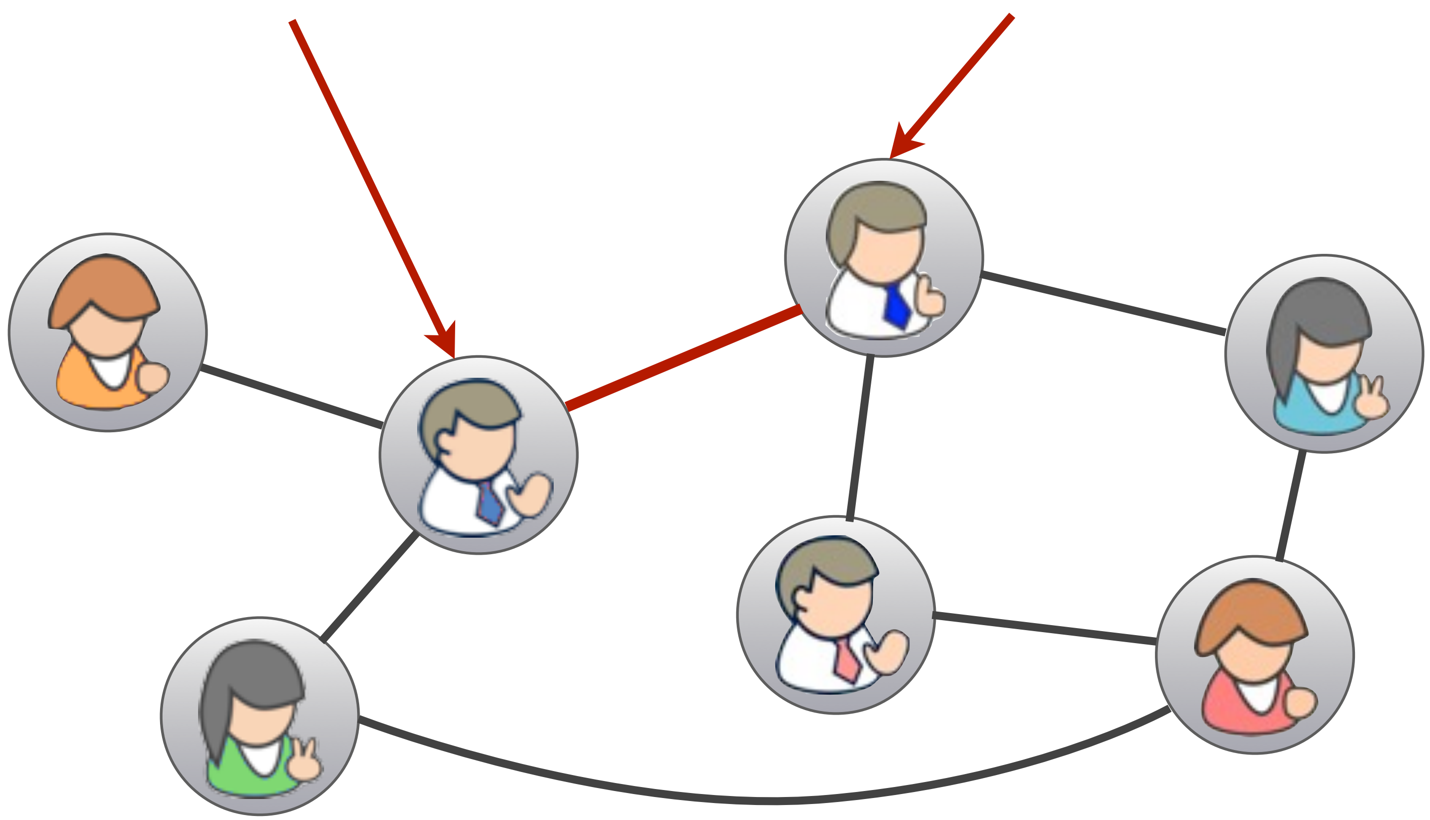
Graphs

- A graph is a set of nodes or vertices *V* and a set of edges or lines
- If an edge exists {*a*,*b*} then we can say that nodes *a* and *b* are related to each other
- The edges themselves can be
  - unordered pairs of nodes
  - or in a directed graph (*digraph*), ordered pairs of nodes where each edge has a direction, sometimes called an *arc*
  - In this case {*a*,*b*} is an arc *from* a *to* b
- Graphs are (generally) non-reflexive; nodes are not related to themselves
- Order is # of nodes, size is # of edges

Social Networks

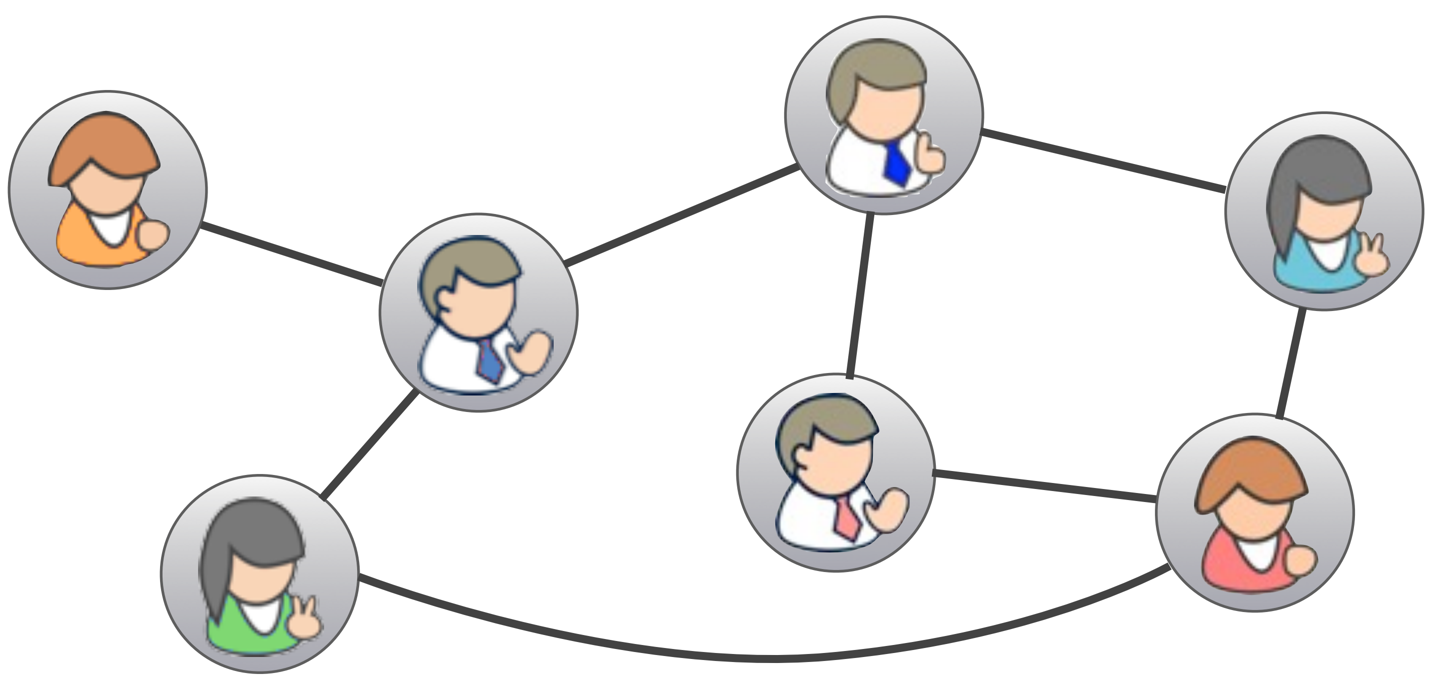
- The graph below shows working relationships between people in an office

Steve works with Dave



Social Networks

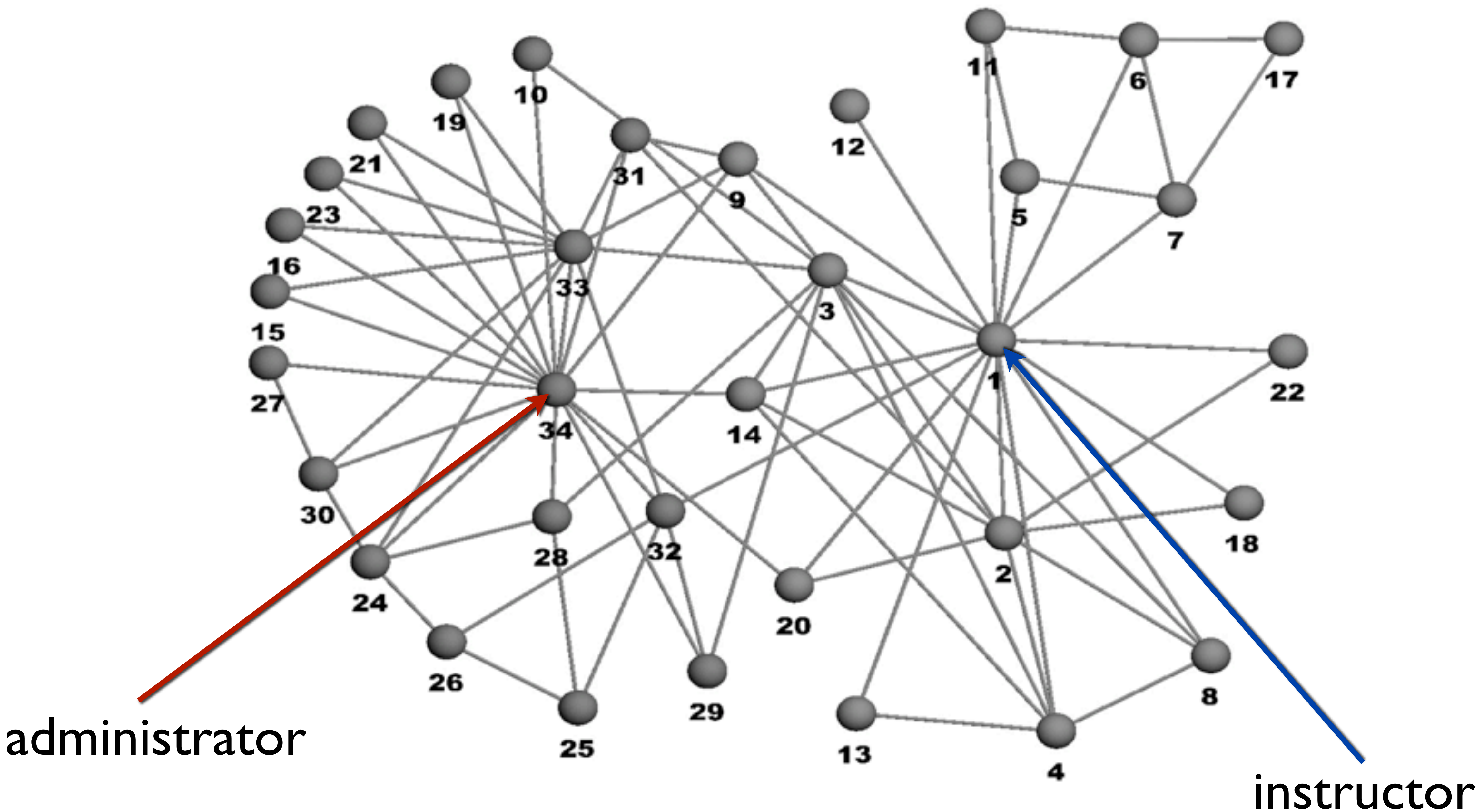
- Existed long before the likes of MySpace, Facebook and Bebo
- Has been used to describe relationships between entities for over a century
- Early research by social scientists and psychologists, now an important field in computer science
- Early online systems included Theglobe.com (1994), Geocities (1994) and Tripod.com (1995)



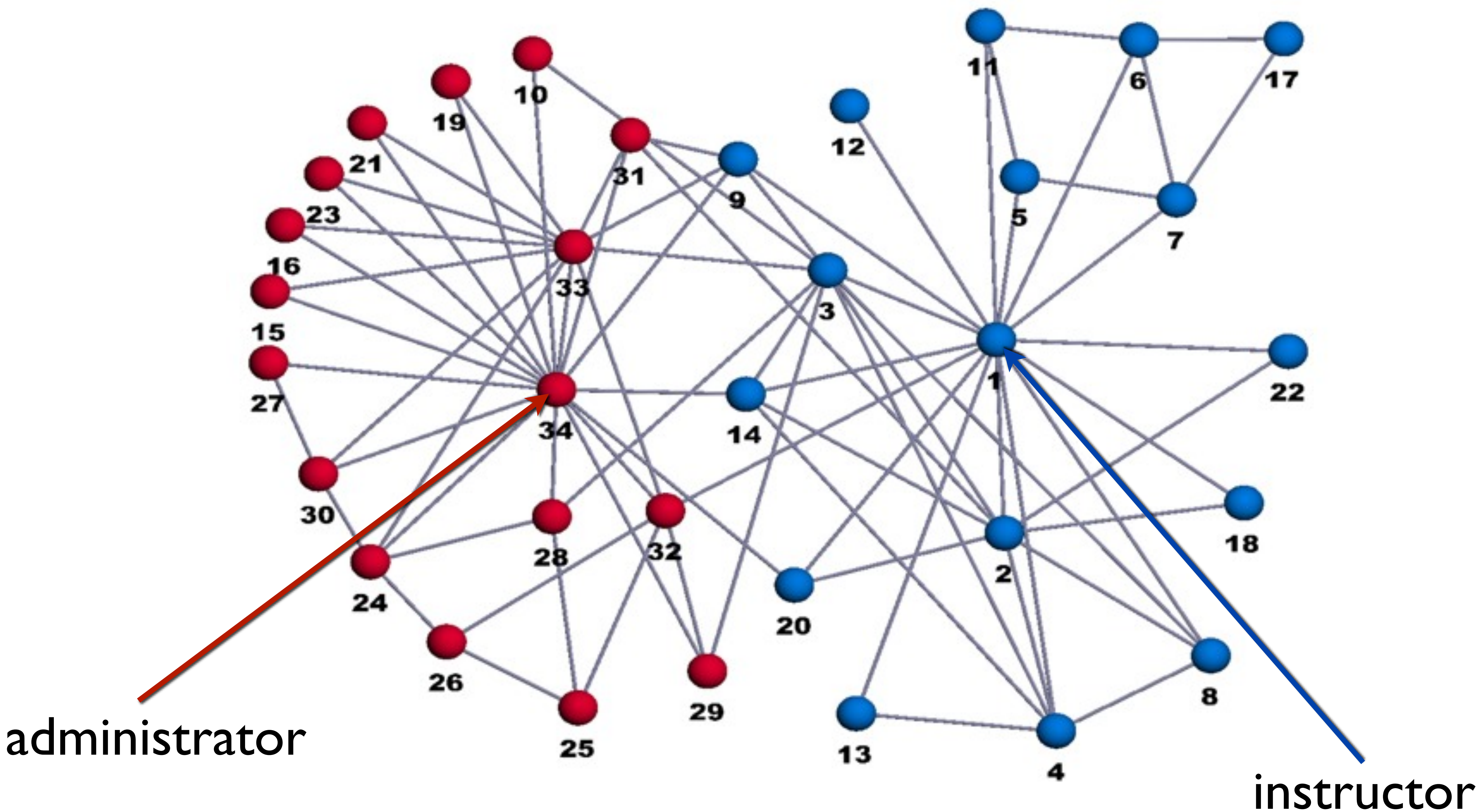
Uses of Social Network Analysis

- **Epidemiology -** to understand how patterns of human contact affect the spread of diseases, such as HIV
- **Marketing and fashion -** to uncover new trends and major influencers
- **Networking -** finding an optimal way of constructing a computer network, locate points of failure and bottlenecks
- **Intelligence -** identifying insurgent networks and determining leaders and active cells
- **Collaborative Filtering -** if your friends like something then there's a good chance you will too
- ... and loads more

Zachary's Karate Club (1977)

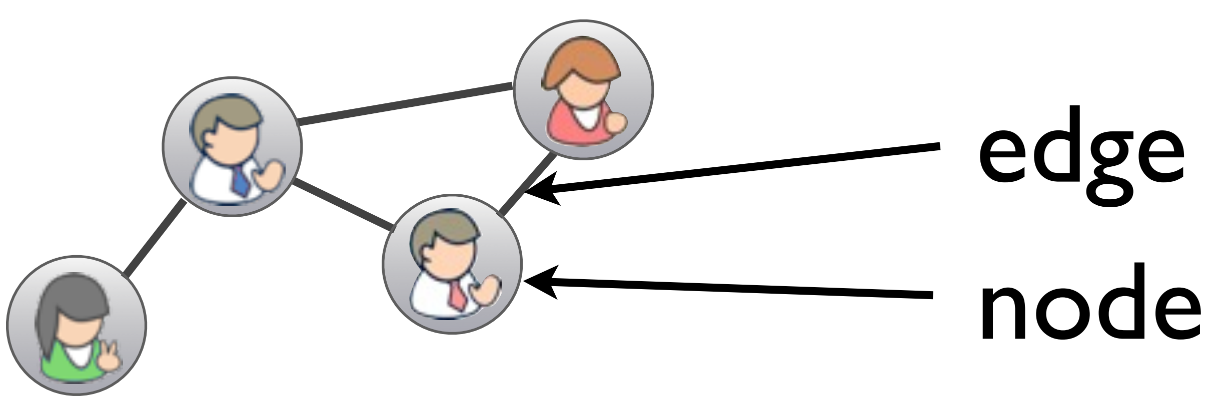


Zachary's Karate Club (1977)



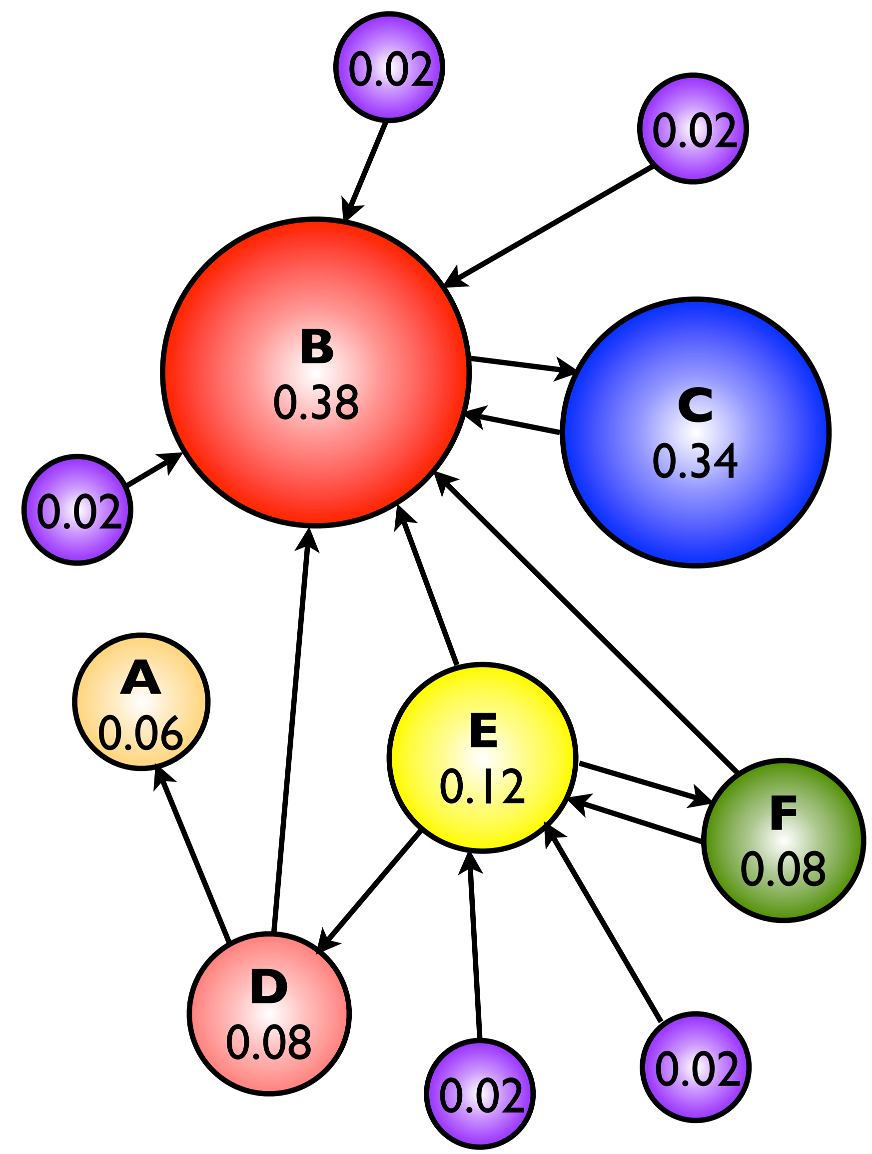
Social Networks

- A graph is a set of nodes or vertices *V* and a set of edges or lines *E*
- The edges themselves can be
  - unordered pairs of nodes
  - or in a directed graph (*digraph*), ordered pairs of nodes where each edge has a direction, sometimes called an *arc*
- If an edge exists {*a*,*b*} then we can say that nodes *a* and *b* are related to each other
- Graphs are (generally) non-reflexive; nodes are not related to themselves



The web is also a graph

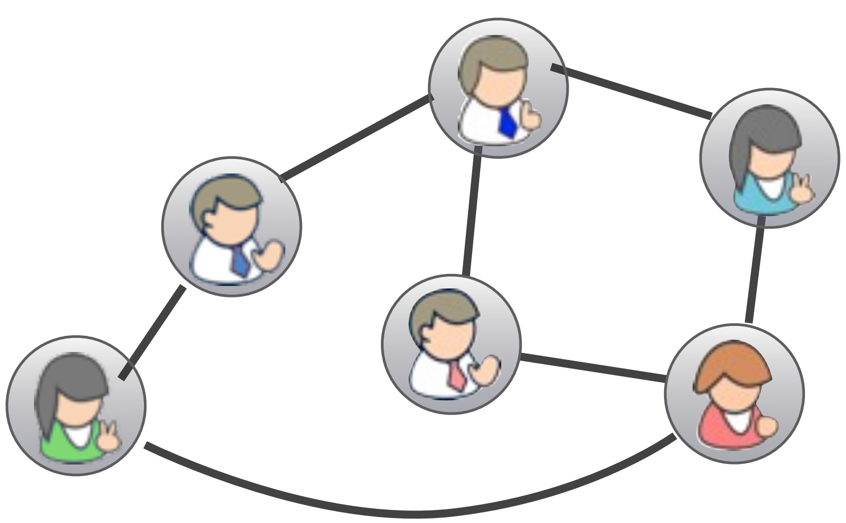
- The web itself can be viewed as a very large graph
- Nodes are individual sites or pages and edges are the links between pages
- This is the basis of Google's page rank algorithm
- The "importance" of a site is determined by the number of sites that link to it weighted by the importance of those sites
- Importance "propagates" around the graph until it stabilises, eventually we end up with probability that a random web surfer will be at a given page
- We can also view other types information as a graph, for example citations of papers



Adjacency Matrix

- We use matrices to represent the relationships within the graph, this is an *adjacency matrix*, denoted g. g_{ij} indicates relationship status for nodes i and j
- 1 indicates a relationship, 0 indicates no relationship

- Each node has a *degree*, the number of other nodes it shares an edge with
-  is degree 3

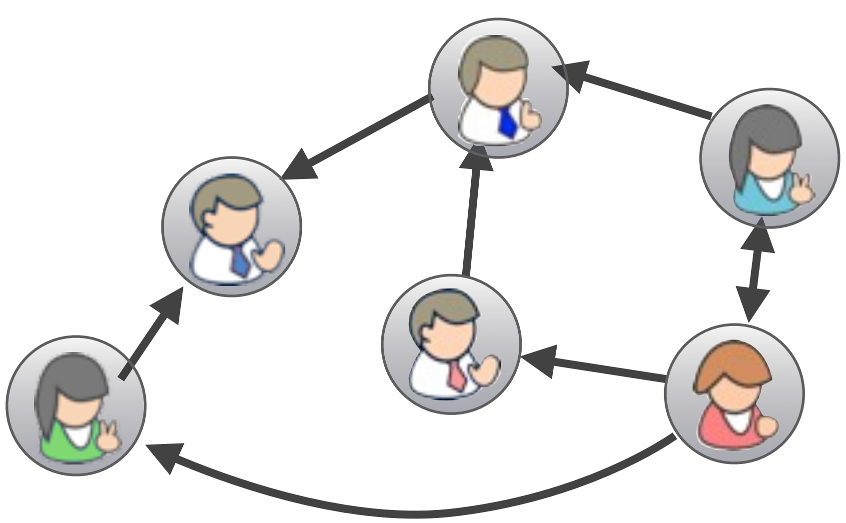


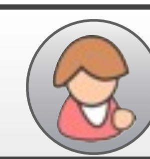
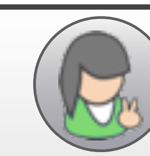
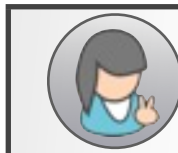
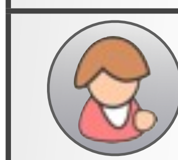
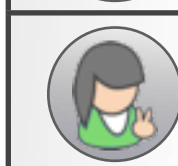
	-	1	1	0	0	0
	1	-	0	1	1	0
	1	0	-	1	0	1
	0	1	1	-	0	0
	0	1	0	0	-	1
	0	0	1	0	1	-

Adjacency Matrix

- Note that for the previous graph the adjacency matrix is symmetrical about the diagonal as it is an undirected graph
- Notice: in the directed graph below the adj matrix is *not* symmetric

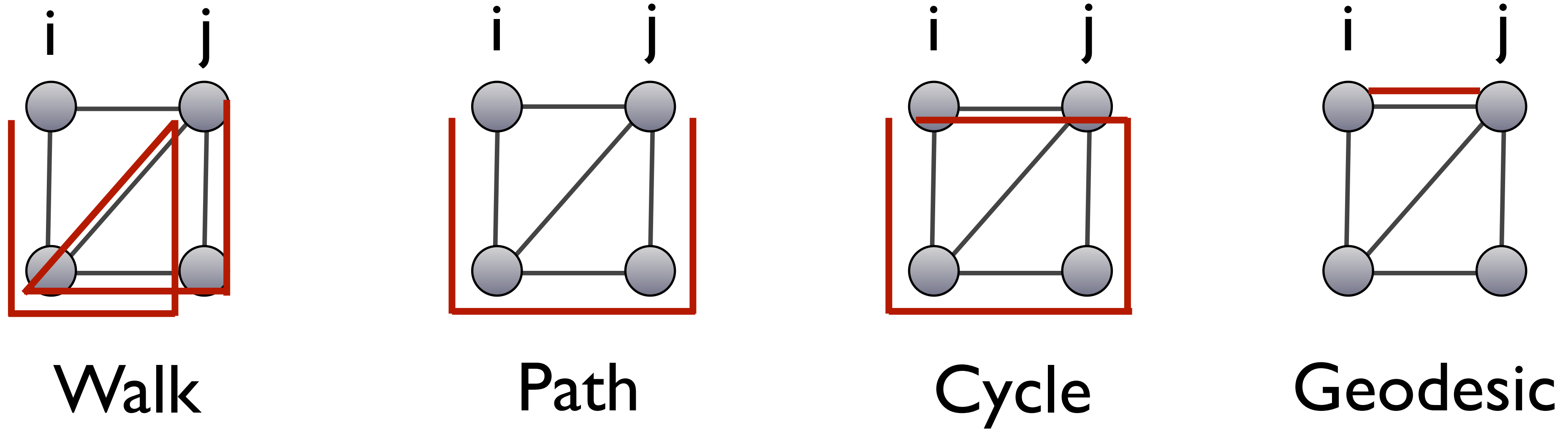
- Nodes in directed graph have outdegree and indegree
-  has outdegree of 3 and indegree of 1



						
	-	1	1	0	0	0
	0	-	0	0	1	0
	1	0	-	1	0	1
	0	1	0	-	0	0
	0	0	0	0	-	0
	0	0	0	0	1	-

Paths traversing the network

- In an un-directed graph:
  - **Walk** - a (connected) sequence of edges
  - **Path** - a connected sequence of edges between 2 nodes, a walk with no repeated edges
  - **Cycle** - a path where the final edge connects to the initial node
  - **Shortest path** - the path with the minimum number of edges connecting 2 nodes (also known as a *geodesic*)

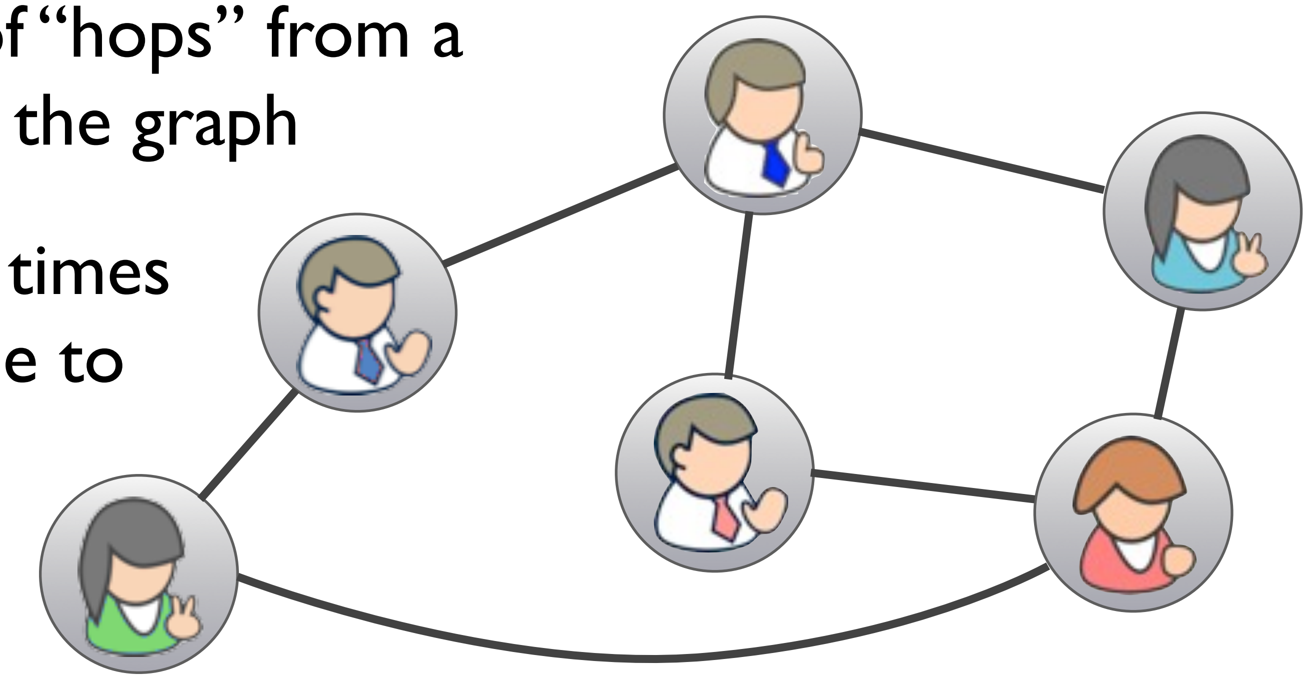


Graph Properties

- If we have a small network (graph) then we can analyse it visually by constructing its graph, however this is impractical for large networks
- We must therefore use *summary statistics* and *performance metrics* in order to describe and compare networks and their graphs such as:
    - Diameter and mean path length
    - Centrality and nodal power
    - Degree distributions
- The *diameter* of the graph is the largest distance between any 2 nodes
- If we let *l(i,j)* be the the length of the geodesic between nodes i and j then the diameter is the maximum *l(i,j)* over all possible node pairs
- The mean path length is the mean distance between all nodes in the graph

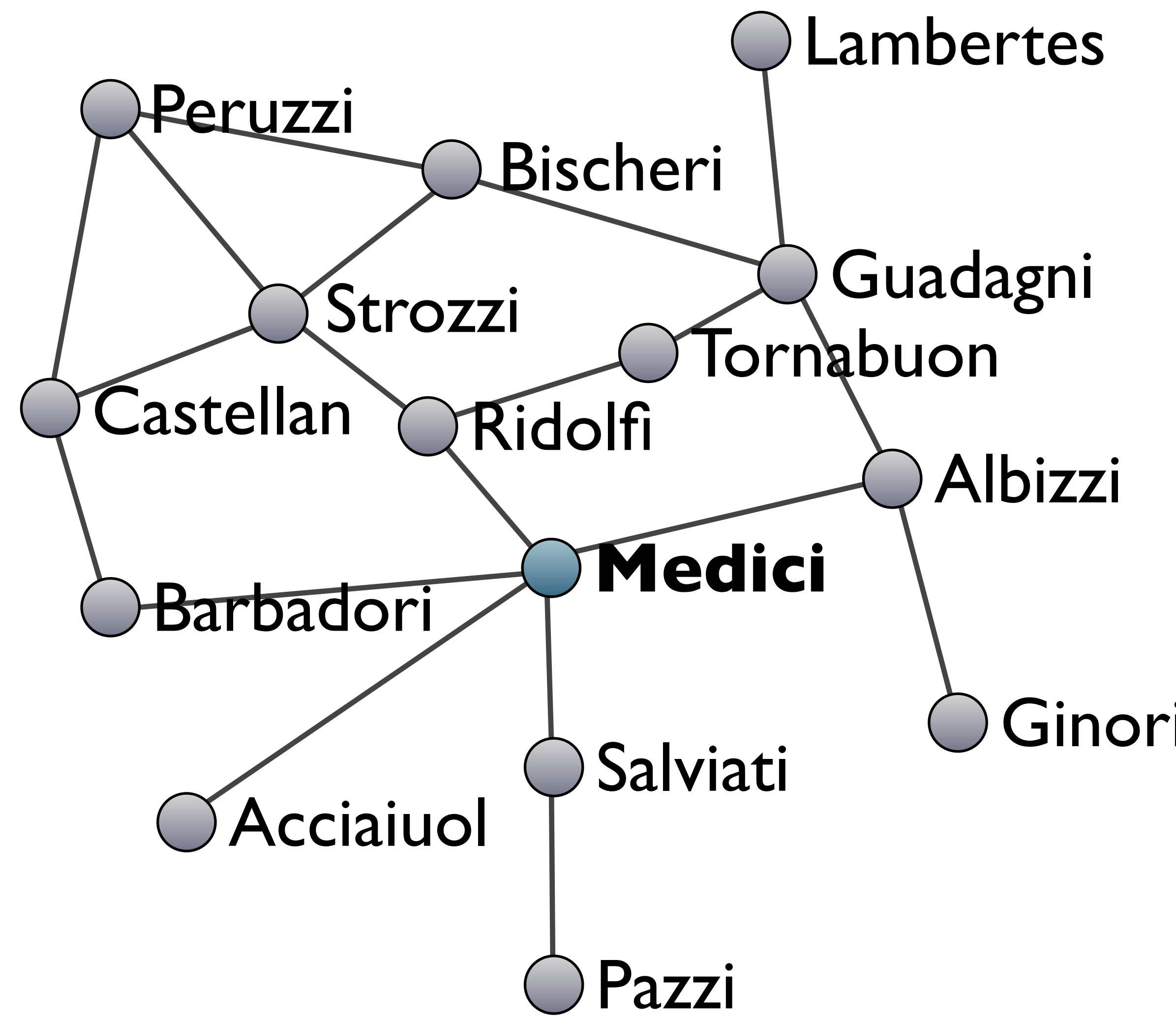
Power and Centrality

- Power is a fundamental property of social structures, related to centrality
- Several techniqes have been developed to study social power
- and we have 3 main measures of power or centrality:
  - **degree -** number of edges a given node has, it's degree, normalised by total number of edges in graph
  - **closeness -** average number of "hops" from a given node to all other nodes in the graph
  - **betweeness -** the number of times that any node needs a given node to reach any other node by the shortest path



Power in a network

- In 15th century Florence the Medici family emerged as the most powerful and ended up dominating trade in the area
- However to start with the family was less powerful than many of the other important families, so how did they achieve so much?
- It has been proposed* that it was their position in the Florentine social network that propelled their success

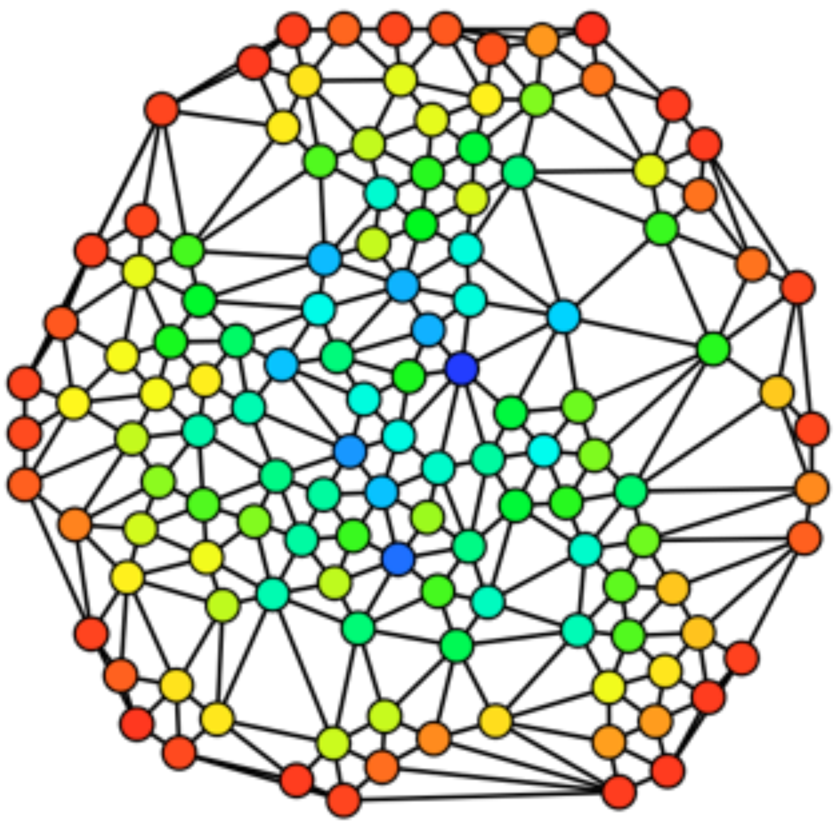


* "Robust Action and the Rise of the Medici" Padgett and Ansell (1993)

Power in a network

- The betweeness measure takes into account the location of a node on a network and how well it acts as a "hub"
- Let *P(i,j)* be the number of shortest paths between nodes *i* and *j*
- Let $P_k(i,j)$ be the number of shortest paths between *i* and *j* that includes nodes *k*

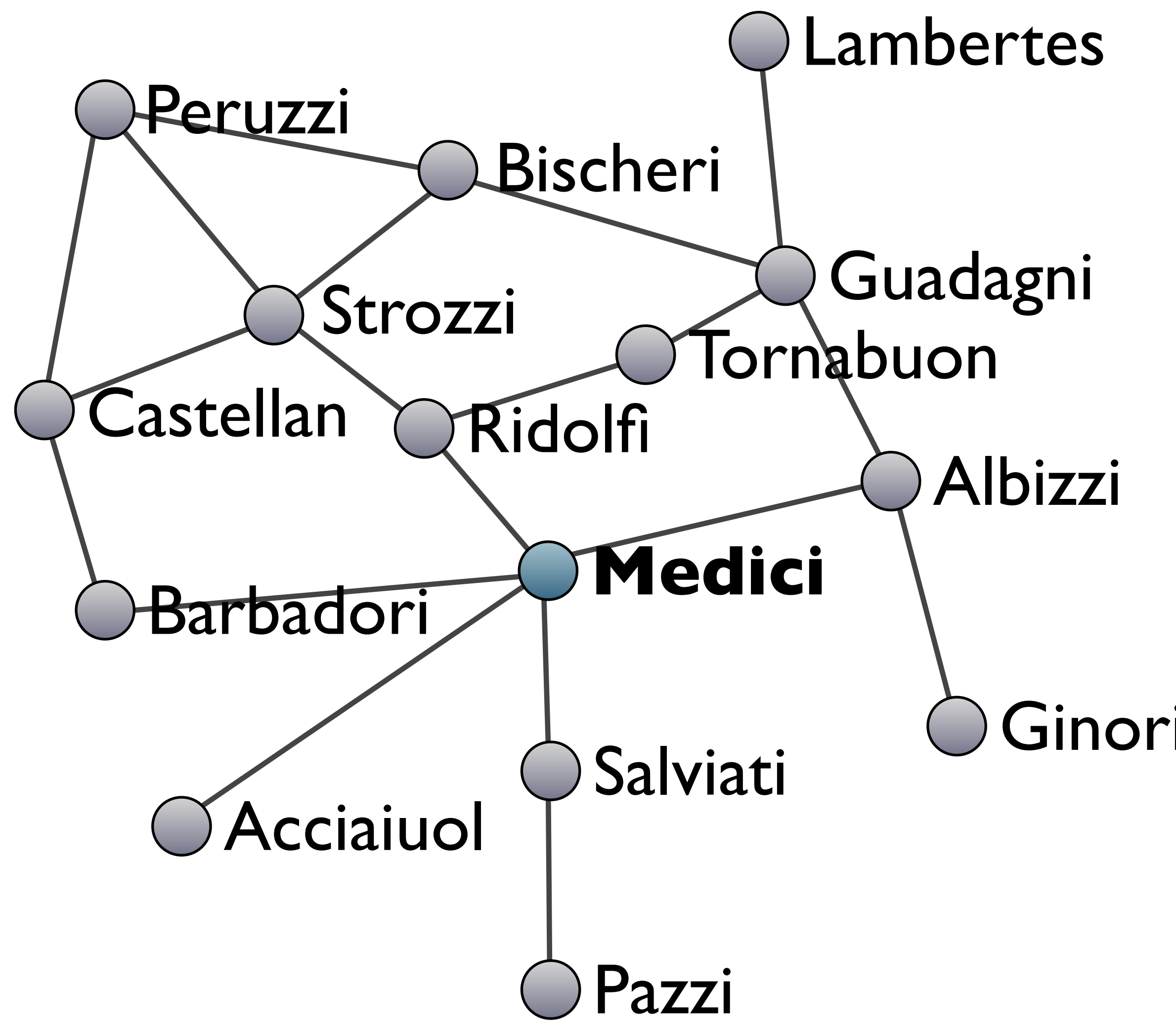
$$B_k = \sum_{(i,j)\in E} \frac{P_k(i,j)/P(i,j)}{(n-1)(n-2)/2}$$



- This gives the fraction of shortest paths (over all possible pairs of nodes) that go through node *k*

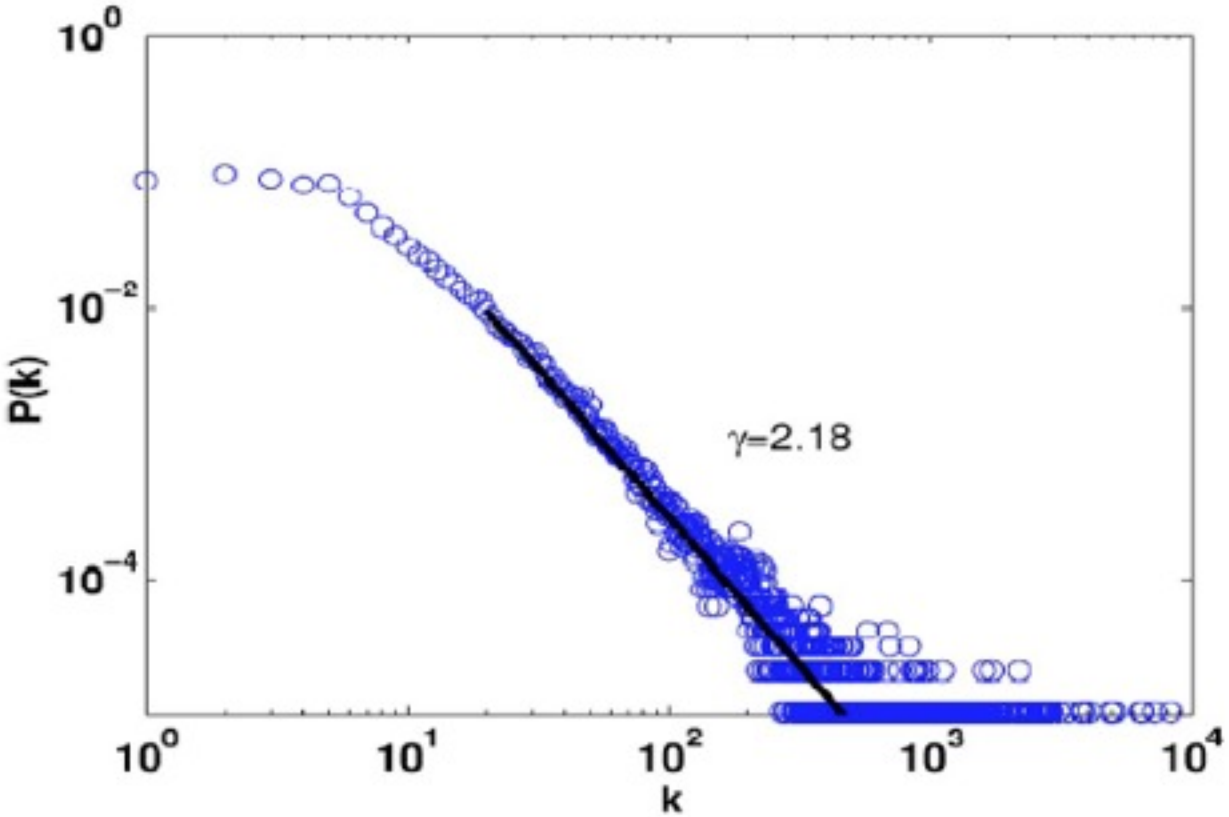
Power in a network

- In the Florientine family network the betweeness for the Medici family is 0.522
- The family with the largest value after the Medicis have a betweeness of only 0.255
- This shows that the Medici family played a very central role in holding this network together and may have gained their power from this



Degree distribution

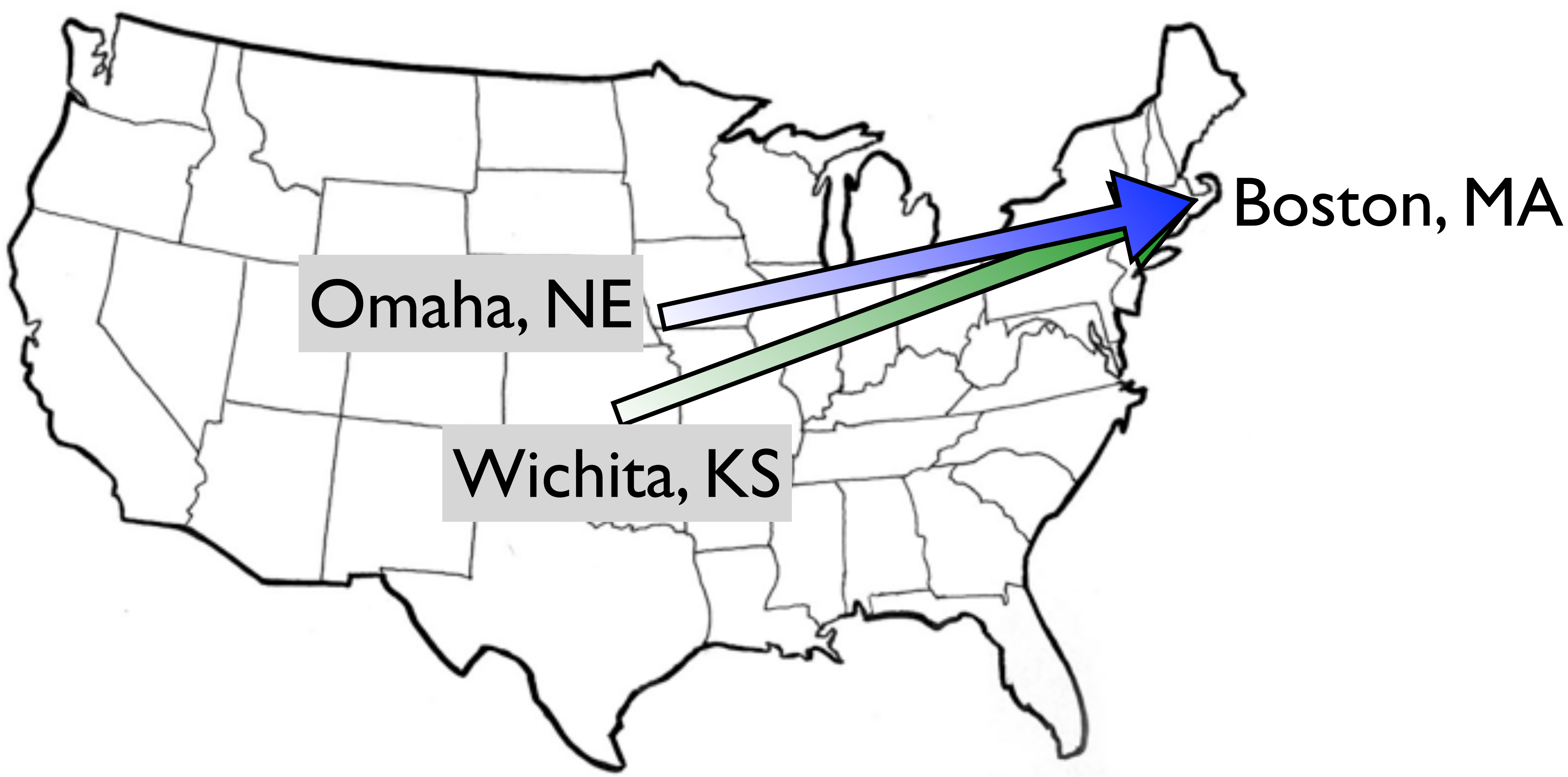
- The distribution of degrees for all nodes in the graph
- For almost all real-world networks this follows a power law pretty closely
- Most nodes have a very small degree
- A small number of nodes have a massive degree
- Examples: wikipedia articles, facebook users, amazon purchases, the web itself



Degree	Probability
2	0.22
3	0.09
5	0.03
10	0.006
100	0.00004

Small World Phenomenon

- This "small world" phenomenon appears in almost all large-scale networks
- Stanley Milgram's 1967 study "The Small World Problem"
- 42 of the 160 letters made it to their target, average number of intermediates was 5.5



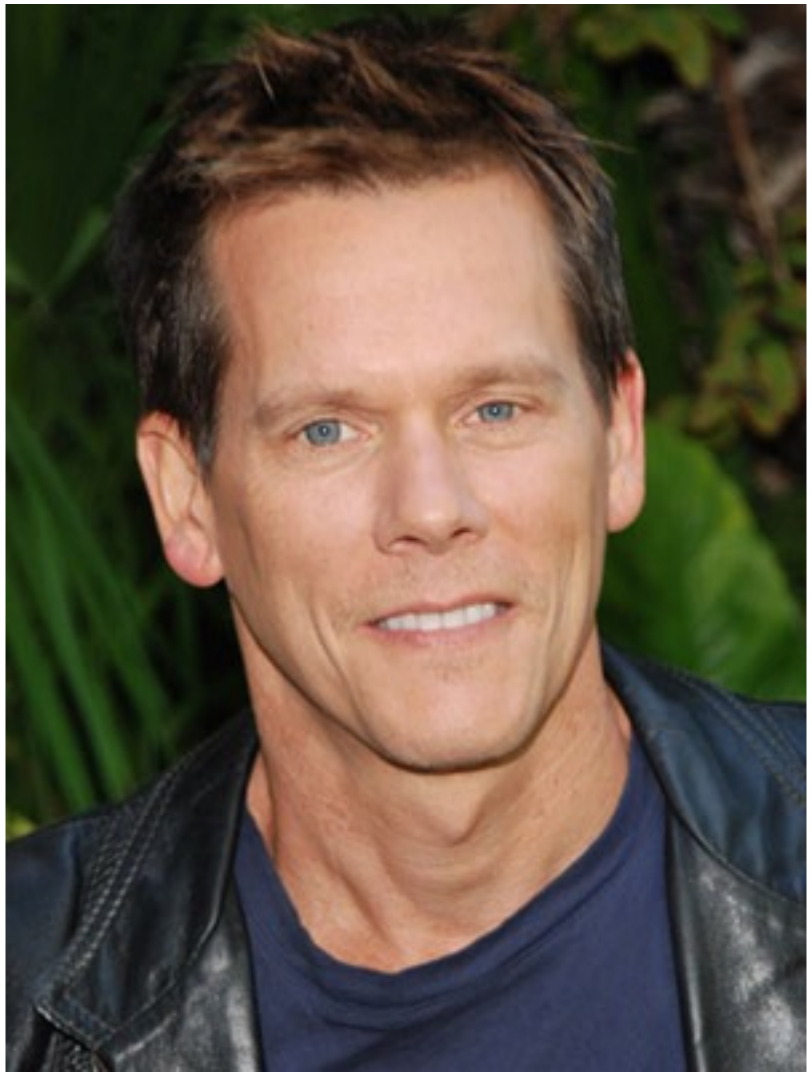
Small World Phenomenon

- Milgram's study and method suffer from a number of key drawbacks which mean we should question his result
  - People will not always know the best person to pass the message on to next
  - Particpants were obtained from advert looking for "well-connected people" so not example of normal case
  - High numbers of non-completion causing iekly under-estimate of mean path length
- Albert, Jeong, and Barabasi (1999) estimated the average path length of the web to be 11 clicks - a lot more than 6!
- but still a surprisingly small number
- What do these "small world" results imply, can we generalise from them?

Small World Phenomenon

- Suppose each node has *k* neighbours
- Then each of those neighbours will also have *k* neighbours and so on
- If we suppose (unrealistically) that neighbours don't share neighbours in common then in just 2 steps we can reach k^2 nodes
- Therefore after s steps we can reach k^S nodes
- If the network has $n = k^S$ nodes then $\mathbb{E}[s] = \dfrac{\ln n}{\ln k}$
- Even though this network is idealistic and unlikely to exist in real life, the average number of steps (s) can still be approximated using this formula
- This would require people in 1967 to have an average of 41 friends
- In reality most nodes are connected by a small number of *key nodes*

Kevin Bacon Number



- Example of an interesting use of graph theory!
- If we have a graph of actors
- Links indicate when 2 actors have worked on the same film
- The number of links between any actor and Kevin Bacon is that actor's Kevin Bacon number
- http://oracleofbacon.org/
- Use imdb for reference
- Let's try a couple....

Scarlett's Bacon Number

Scarlett Johansson

The Black Dahlia (2006)

Steve Eastin

Rails & Ties (2007)

Kevin Bacon

So Scarlett's Bacon number is **2**

Robert's Bacon Number

Robert Webb

Magicians (2007)

David Mitchell

I Could Never Be Your Woman (2007)

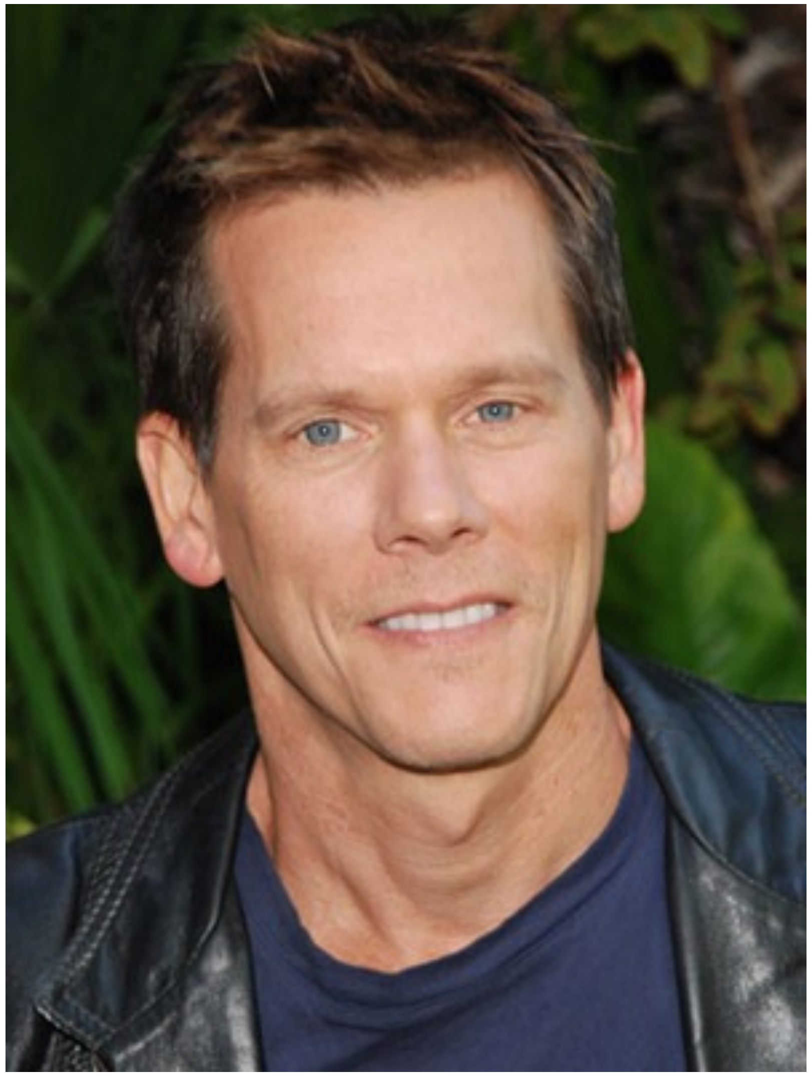
Wallace Shawn

Starting Over (1979)

Kevin Bacon

So Robert's Bacon number is **3**

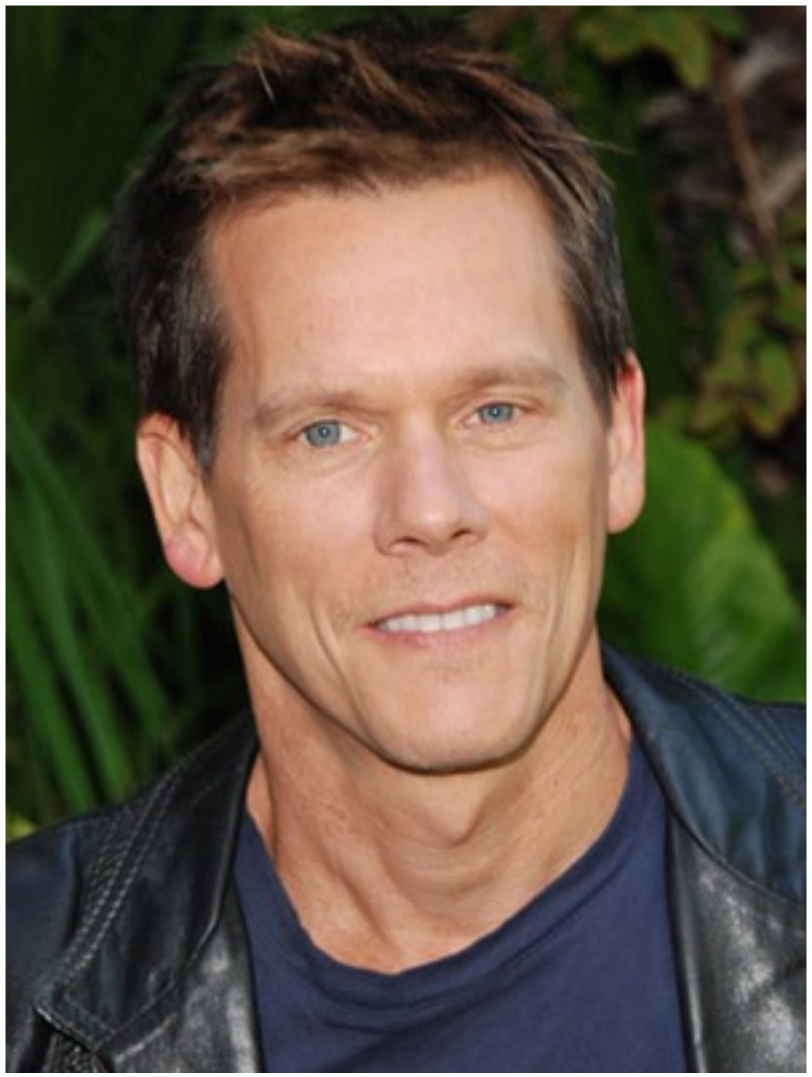
Kevin Bacon Number

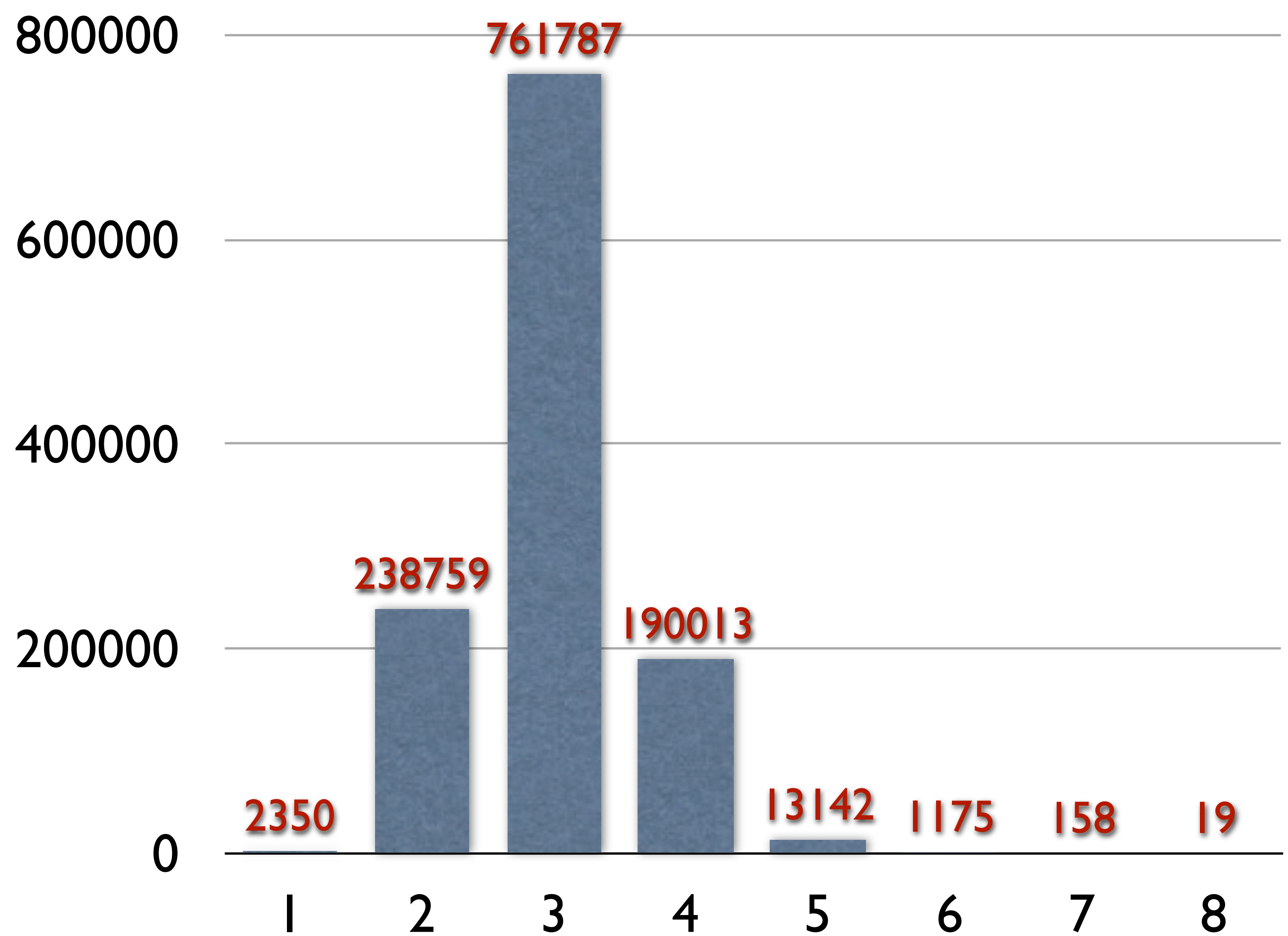


- Total actors: around 1.2 million
- Total films: many millions!
- Average path length to the Bacon: 2.981
- Actor with greatest centrality: Rod Steiger (2.814)
- Kevin's centrality rank: 876
- We could play this game with 1000s of actors and it'd still work!
- Notice that a Bacon number is simply the length of the geodesic between actors

The most important actor in the world?

How Far to the Bacon?





Summary

- Social networks (and relationship networks in general) are abundant and useful sources of information
- We can use graph theory to model them
- However they can be difficult to analyse
- We can learn more about them by calculating metrics and analysing their statistics
- Real graph frequently display power law degree distributions and small-world phenomena
- Kevin Bacon (and of course Rod Steiger) are immensely important!