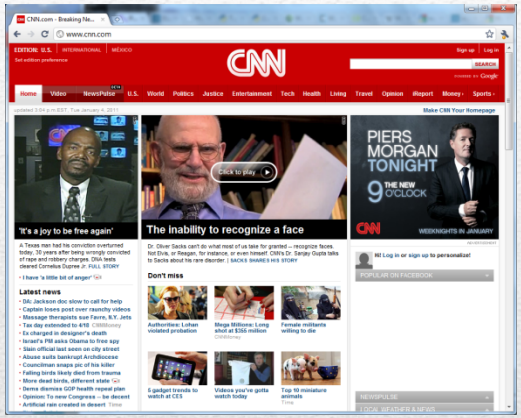
Web Graph

COMP6218

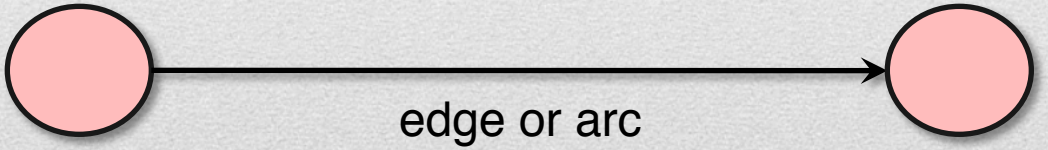
How is the Web structured?

page



link

page



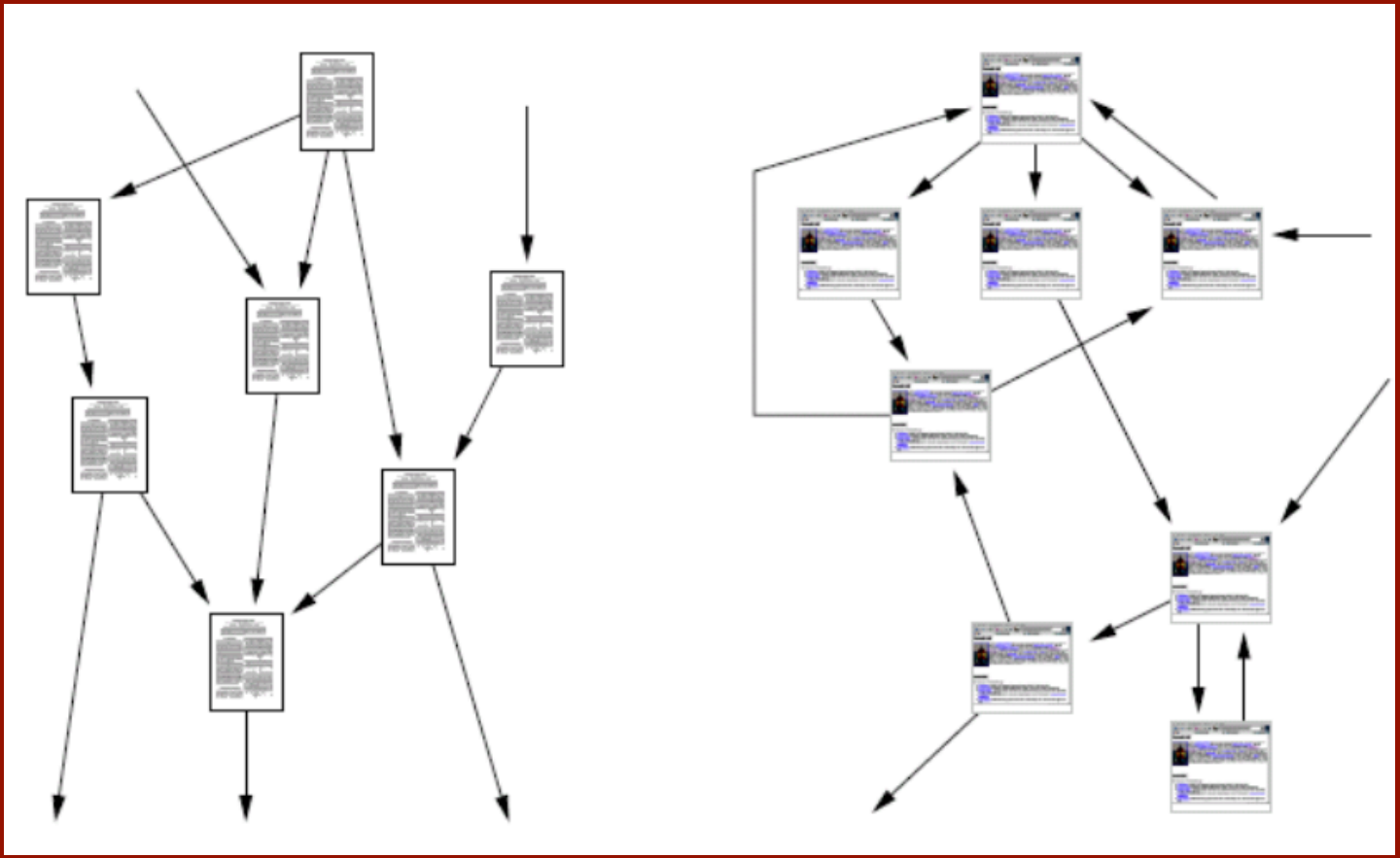
node or vertex

edge or arc

node or vertex

Graph Theory: Pages are nodes & links are directed edges

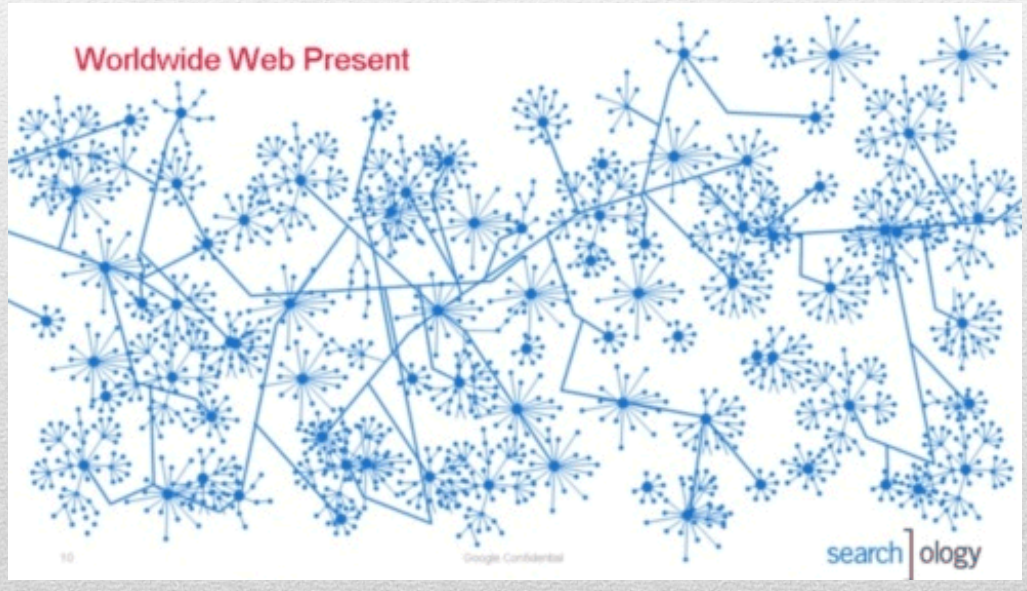
The Web Graph



- Both citations and links form graph structures.

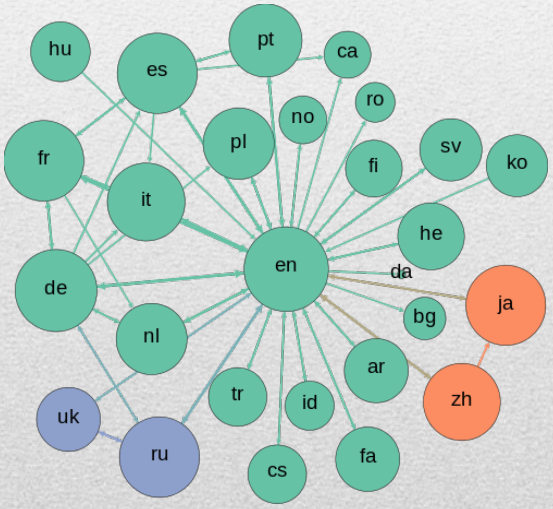
At scale

- Web graph ~ directed graph that is formed by webpages and their hyperlinks



http://googlesystem.blogspot.com/2007/05/world-wide-web-as-seen-by-google.html

http://en.wikipedia.org/wiki/Graph_(data_structure)

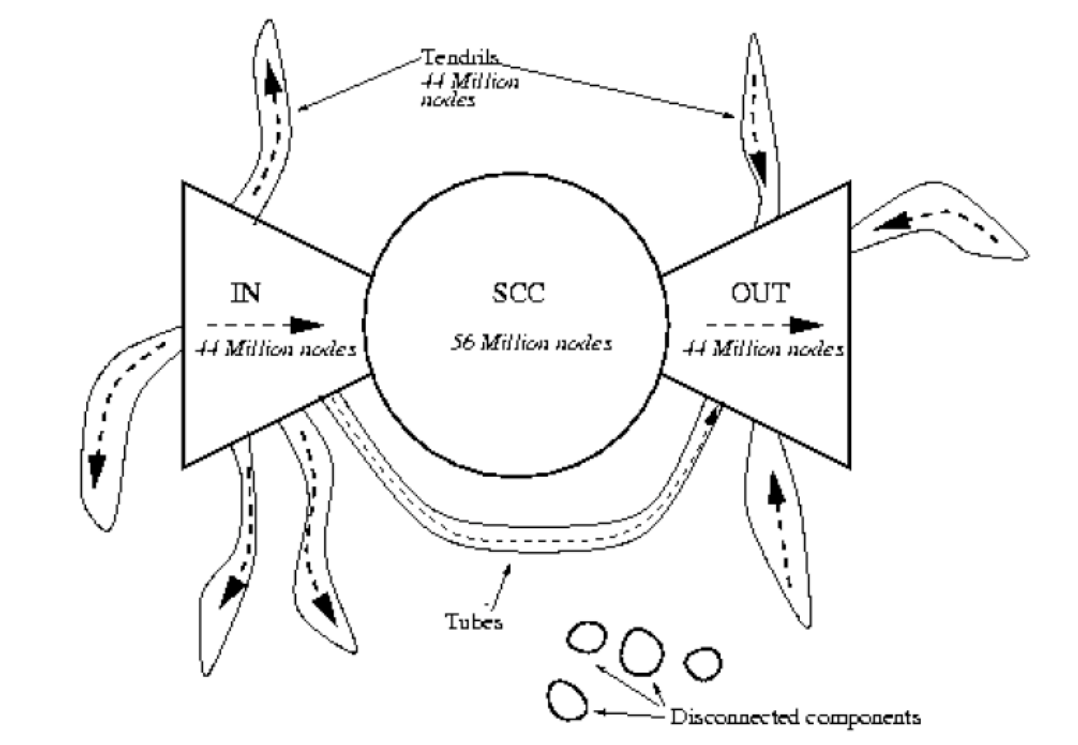


Small World Network

- Six degrees of separation
- Most pages are not neighbors but most pages can be reached from others by a small number of hops
- Many hubs- pages with many inlinks
- Robust for random node deletions
- Other examples: road maps, networks of brain neurons, voter networks, and social networks

Global Web Structure

- What does the Web look like to a search engine?



SCC = Strongly Connected Core

Graph Terminology

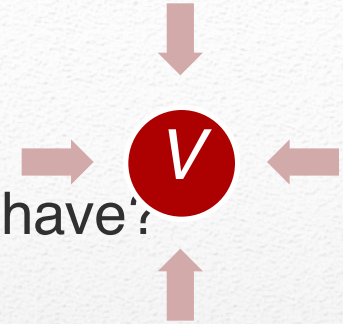
- A graph is made from
  - nodes or vertices (pages, documents) connected by
  - edges (links, arcs)
- Network size
  - total number of vertices (N)
- Distance between two vertices (*v* and *v'*)
  - shortest path between the pair
- Diameter of a network
  - longest distance between any two vertices
  - *ie* the distance between the pair of nodes who are furthest apart in the network
- Average-case diameter
  - average distance between any pair of nodes *v* and *v'*
- Degree of vertex v
  - number of edges connected to v
- Density of network
  - ratio of edges to vertices

Graph Importance Measures

- Four types of centrality*:
  - Degree centrality
  - Betweenness centrality
  - Nearness centrality
  - Eigenvector centrality

- **Central as in important or vital or 'at the heart of'. Not 'near the centre'*

Degree Centrality



- How many edges does a vertex *v* have?
  - How many friends does a Facebook user have?
- In a directed graph:
  - *in degree* – how many links to this vertex?
    - A web page is important if lots of things link to it.
    - A twitter user is important if lots of other users follow it.
    - A journal article is important if lots of other articles cite it.
  - *out degree* – how many links from this vertex?
    - A review paper is useful if it links to lots of other papers.
    - A search engine is useful if it links to other sites.

Betweeness Centrality

- How many pairs of nodes is *v* between?
  - A vertex is between two others if it is on the shortest path from one to another.
- A node that has high 'betweenness' is vital to the communication of the network
  - Many / most messages between two nodes must travel through it.

Closeness Centrality

- The *farness* of a vertex with respect to a graph is the sum of its distances to every other vertex in the graph.
- The *closeness* of a vertex is the inverse of its *farness*.

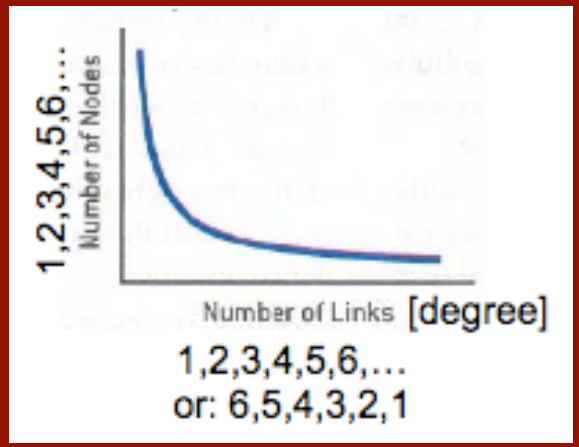
- Closeness determines how quickly a message can travel from one node to the whole graph.

Eigenvector centrality

- Instead of considering all vertexes equivalent to each other, assume that some are more important.
  - There is some kind of *ranking* function
- Eigenvector centrality determines the connection of a vertex v to the *important* nodes in the network.
  - PageRank is an example

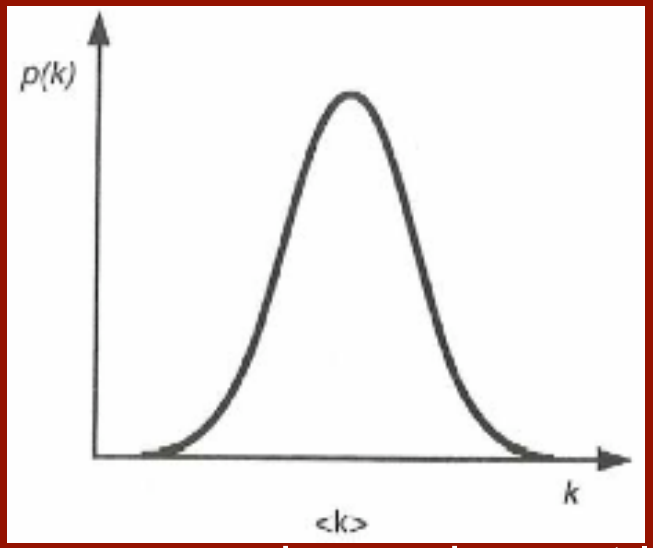
Degree Distributions

- Degree distribution p(k)
  - A plot showing the fraction of nodes in the graph of degree k, for each value of k



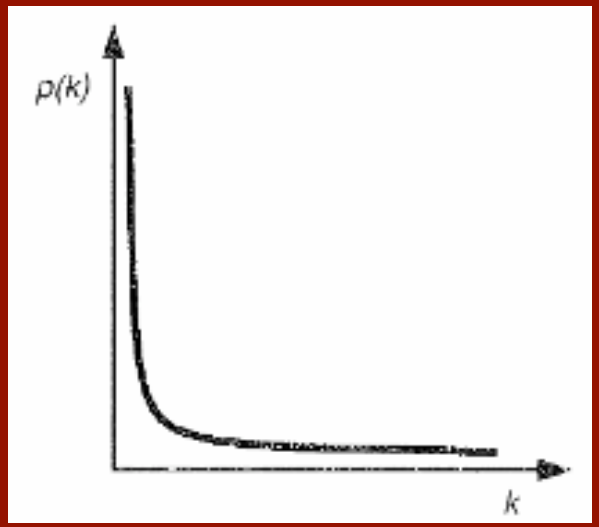
Degree Distribution Examples

- Normal distribution:



- The average degree is most likely. Very high and very low degrees are highly unlikely.

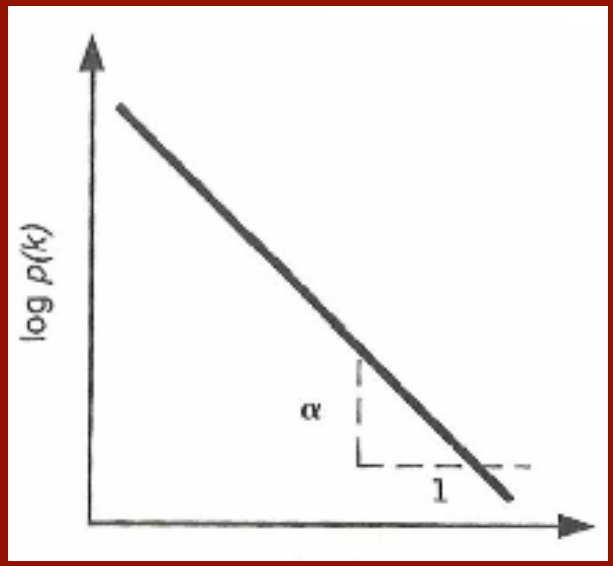
- Power law distribution



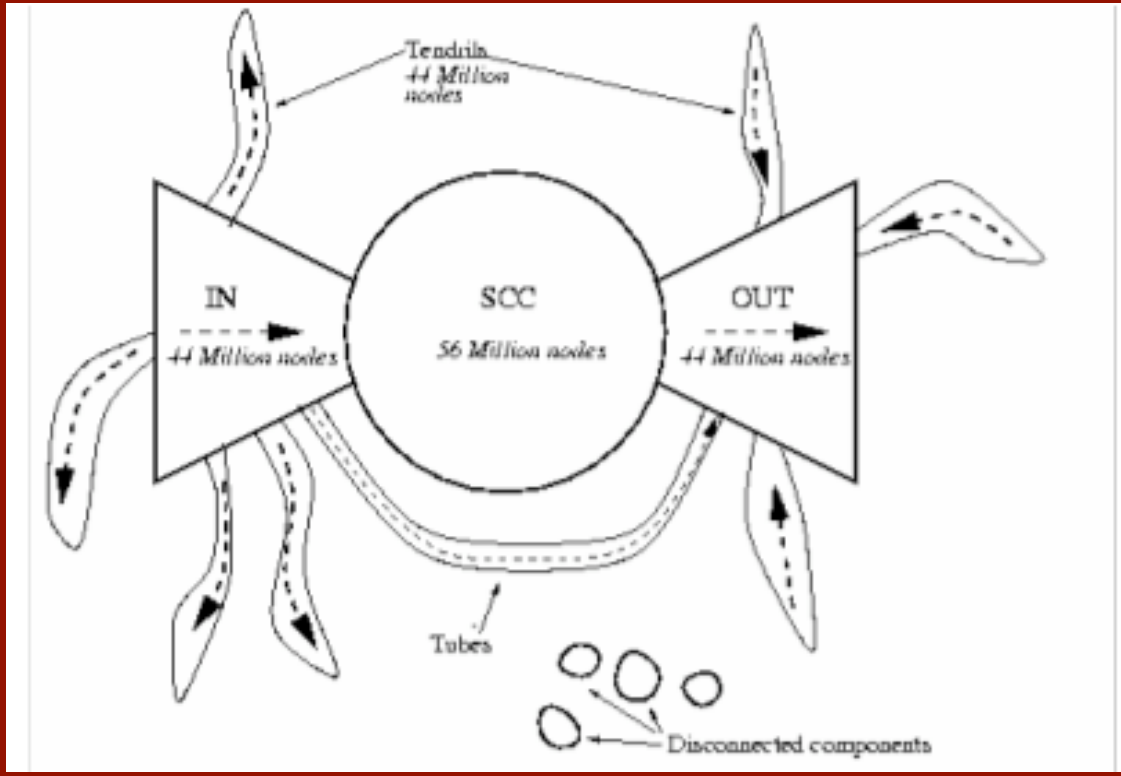
- No meaningful average degree. Very low degrees are likely, but very high degrees are likely in aggregate.

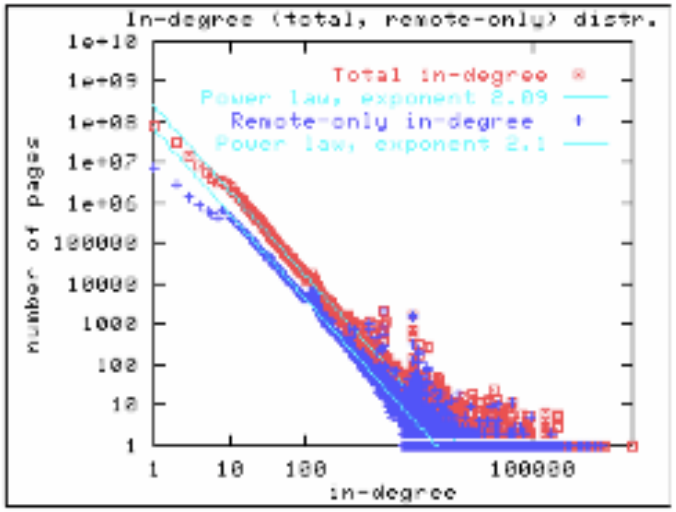
Power Law Networks

- Some nodes have a tremendous number of connections to other nodes (hubs)
- most nodes have just a handful
- Robust against accidental failures, but vulnerable to coordinated attacks
- Popular nodes can have millions of links: The network appears to have no scale (no limit)



Web Graph Revisited





- Power law describes graph

Preferential Attachment

- Preferential attachment is process by which items are distributed to objects according to how many items they already have
- Preferential attachment can lead to power law distribution
- E.g.
  - Links on the web
  - Citations to scientific papers

Web Graph Results

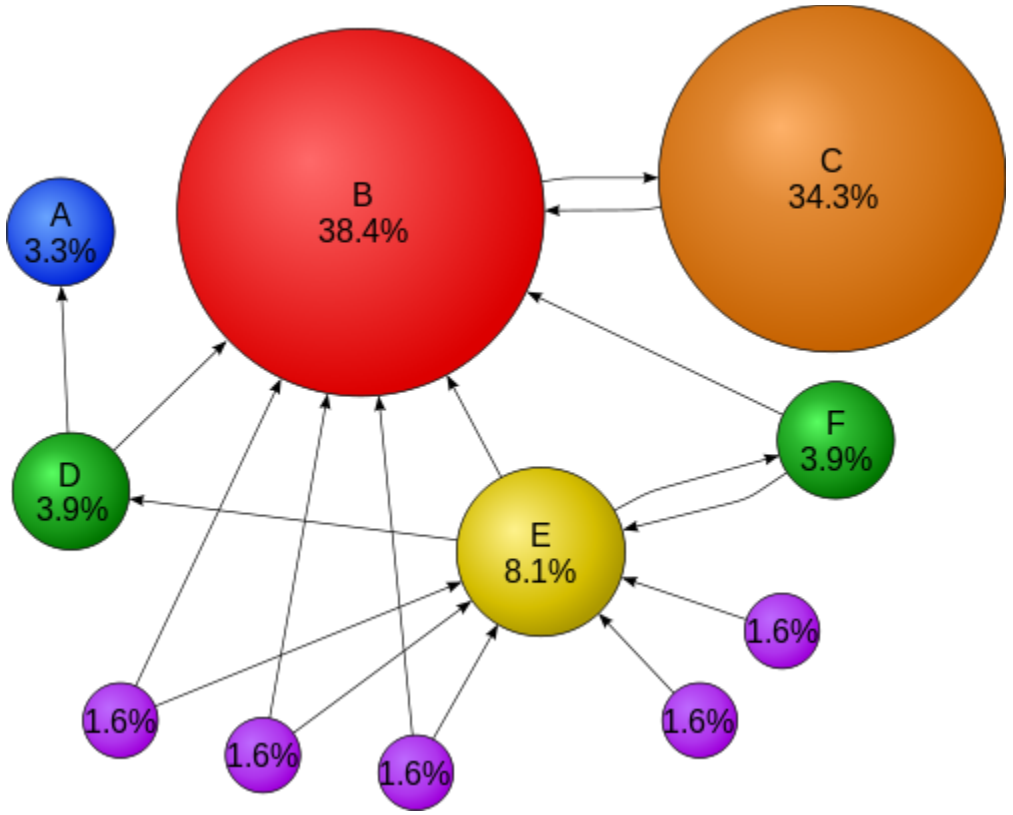
- the diameter of the central core (SCC) is at least 28, and the diameter of the graph as a whole is over 500
- the probability of a path between randomly chosen pairs is only 24%
  - average directed path length is about 16
  - average undirected path length is about 6

Google Relevancy (1)

- Google ranks web pages based on the number, quality and content of *links pointing at them (citations)*.
- Number of Links
  - All things being equal, a page with more links pointing at it will do better than a page with few or no links to it.
- Link Quality
  - Numbers aren't everything. A single link from an important site might be worth more than many links from relatively unknown sites.

Web Graph Application to PageRank

- Google PageRank algorithm used web graphs



http://en.wikipedia.org/wiki/PageRank

Google Relevancy (2)

- Link Content
  - The text in and around links relates to the page they point at. For a page to rank well for "travel," it would need to have many links that use the word travel in them or near them on the page. It also helps if the page itself is textually relevant for travel
- Ranking boosts on text styles
  - The appearance of terms in bold text, or in header text, or in a large font size is all taken into account. None of these are dominant factors, but they do figure into the overall equation.

PageRank: Original Formula

- PageRank of page P_i is given by the summation of the PageRank of all pages P_j that link to P_i divided by the set of outbound links of P_j

$$r(P_i) = \sum_{P_j \in B_{P_i}} \frac{r(P_j)}{|P_j|}$$

- Iterative formula, starting with rank 1/n for all n pages

PageRank: Surfer Model

- Usage simulation:
  - Based on a model of a Web surfer who follows links and makes occasional haphazard jumps, arriving at certain places more frequently than others.
- User randomly navigates
  - Jumps to random page with probability p
  - Follows a random hyperlink with probability 1-p
  - Never goes back to a previously visited page by following a previously traversed link backwards

PageRank

- Google finds a single type of universally important page--intuitively, locations that are heavily visited in a random traversal of the Web's link structure.