Social network analysis

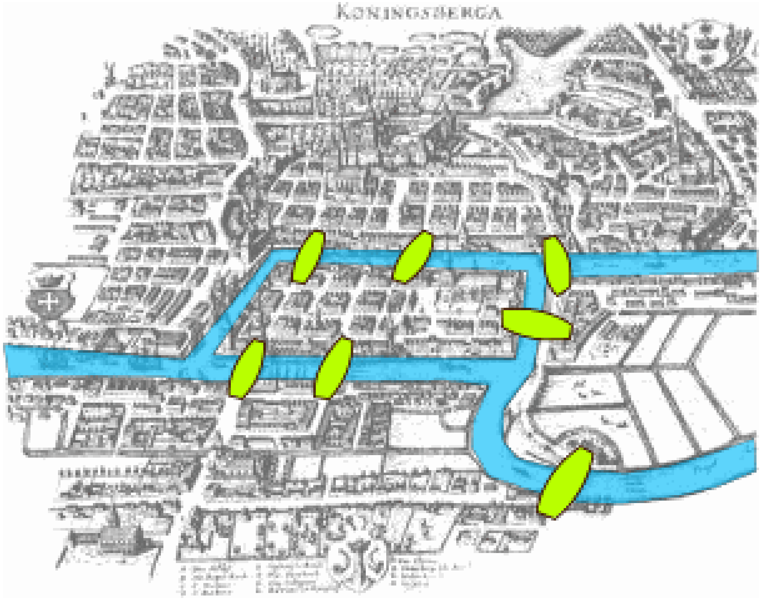
Miguel Rebollo (@mrebollo)
Univ. Politècnica de València

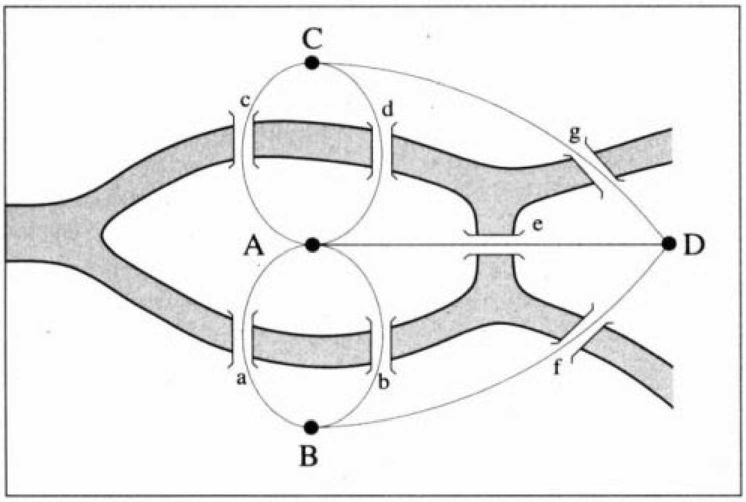
UPV. ETSINF. SWB 2021

Table of contents

1. Introduction to Complex Networks

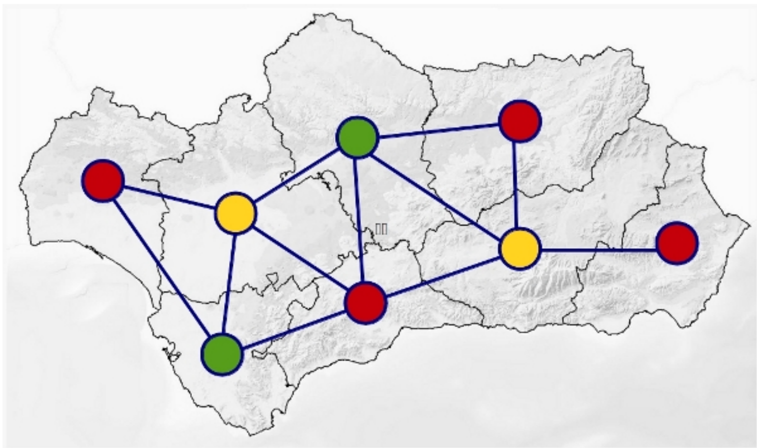
Origin





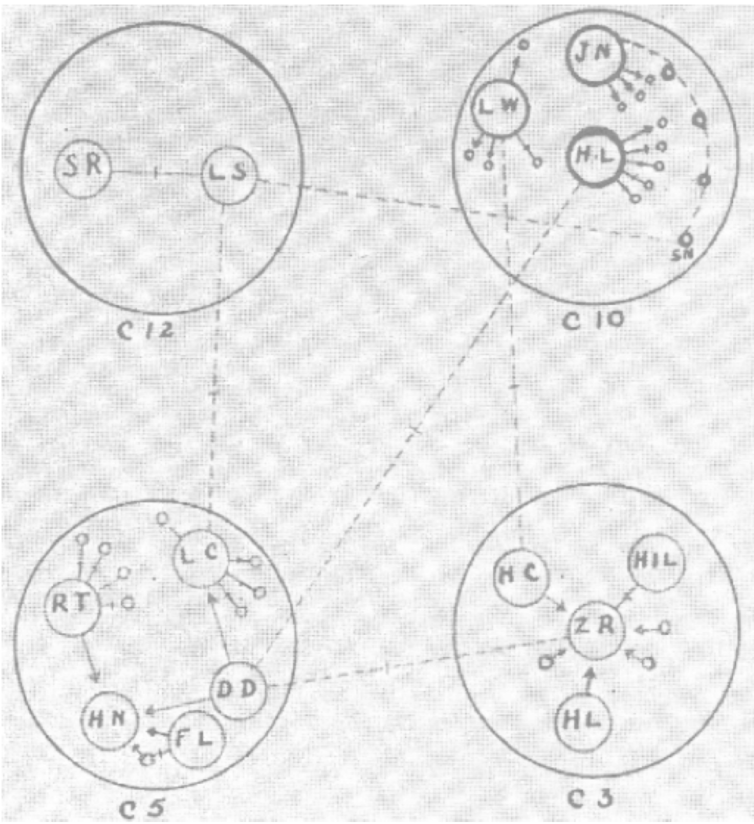
The Königsberg bridge problem (Euler, 1736). The problem was to find a walk through the city that would cross each bridge once and only once.

Classic graph theory



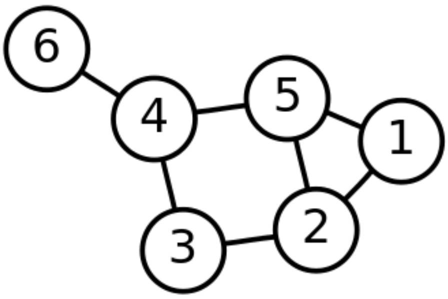
The map coloring problem was proposed in 1834, but the solution couldn't be proved until 1976. Any map can be colored with 4 colors.

Fisrt application to human beings



Used to study why a group of girls run away from a college at Hudson (NY St) in 1934. The relations among them and the cottages they live in where represented

Graph representation

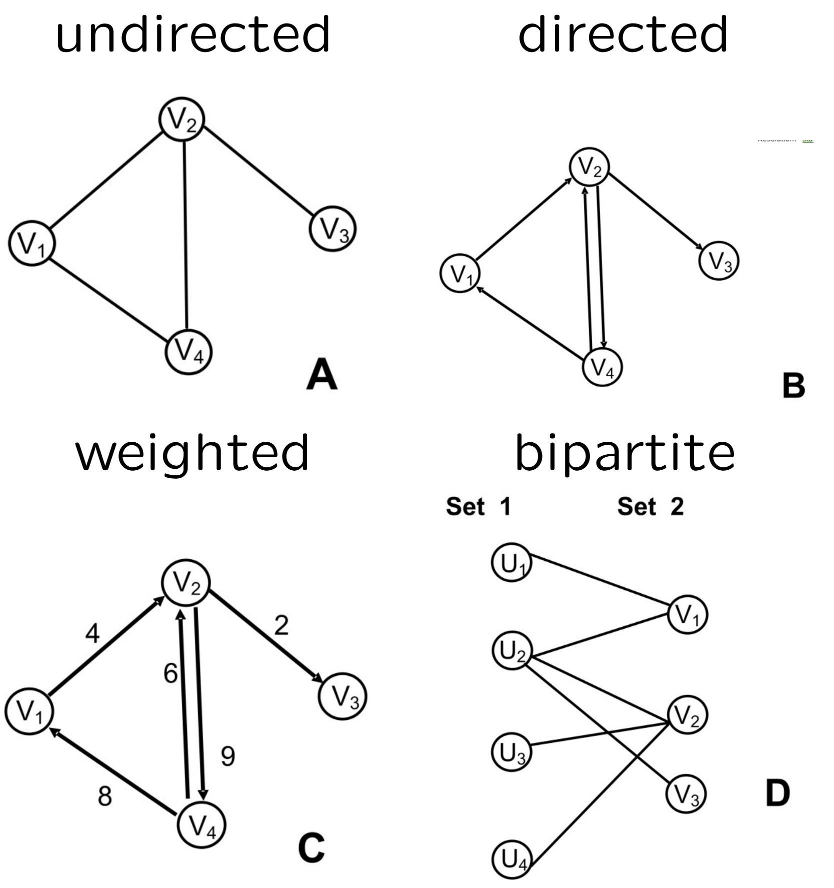


$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

The *adjacency matrix* $A = (a_{ij})$ where

- $a_{ij} = 1$ if there is an edge between nodes i and j
- $a_{ij} = 0$ otherwise

Types of graphs



Basic concepts

Shortst path d_{ij} minimal distance (weight) between nodes i and j

Average path length l average of d_{ij} between all the nodes

Diameter D the longest (maximum) of the shortest paths $\max d_{ij}$

Degree d_i number of neighbors of the node i

Clustring coeficient C number of triangles of all the possible ones

Algorithms

Bellman-Ford & Dijkstra computes shortest paths between two nodes

Floyd-Warshall computes all shortest paths

Primm finds the minimum cost spanning tree

Kruskal finds all minimum cost spanning trees (connected components)

Nearest neighbors solves the traveling salesman problem

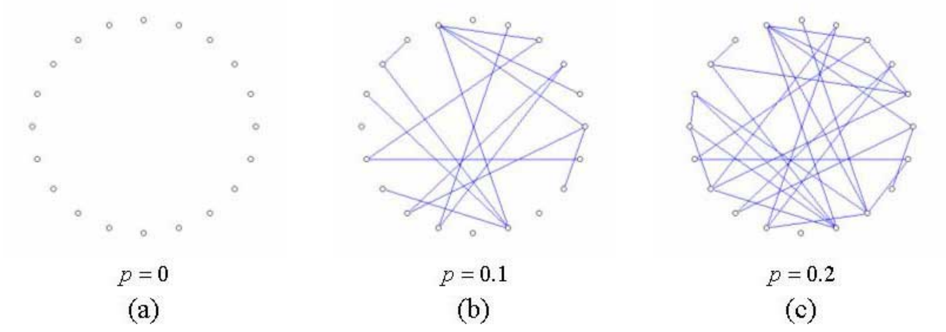
Depth-first search explores complete branches

Breadth-first search explores complete neighborhood

Hungarian perfect matching in bipartite graphs

2. Complex Networks Models

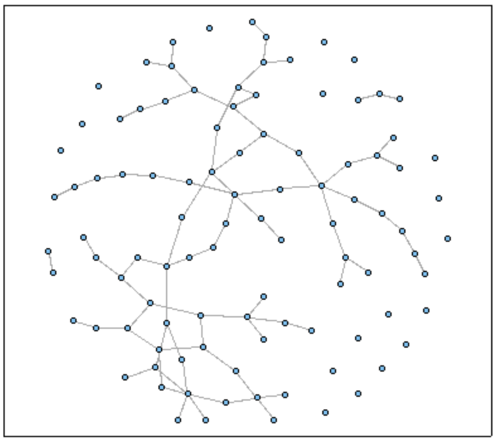
Random graphs

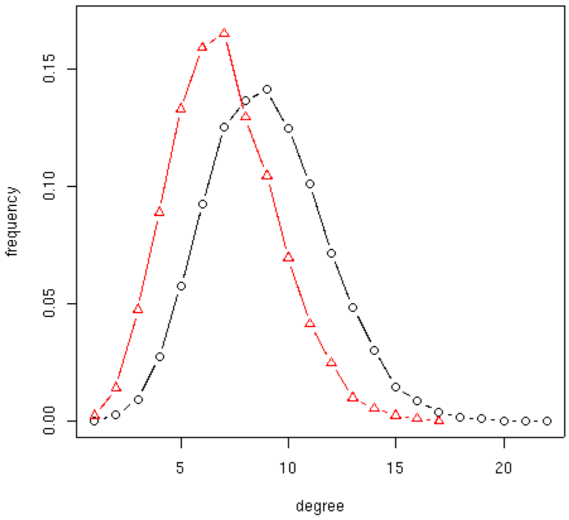


Erdös-Renyi model (1959). A new edge is added between two nodes with probability p.

Emergence of the *giant component*: when $p > \frac{1}{n}$ (or $\frac{n}{2}$ edges have been added). The size of the component is $n^{\frac{2}{3}}$. The complete network is connected after $n \log n$ edges

Random graphs. Degree distribution





The degree distribution of the Erdös-Renyi model is a Poisson one

6 degrees of separation

Short Story 'Chains' (F. Karinthy, 1929)

> A fascinating game grew out of this discussion. One of us suggested performing the following experiment to prove that the population of the Earth is closer together now than they have ever been before. We should select any person from the 1.5 billion inhabitants of the Earth – anyone, anywhere at all. He bet us that, using no more than five individuals, one of whom is a personal acquaintance, he could contact the selected individual using nothing except the network of personal acquaintances

6 degrees of separation

Milgram's experiment (1967).

- people at Omaha, Nebraska and Wichita was chosen

- they have to send a letter to one person in Boston or Massachusetts

- if they know the target, they send them the letter

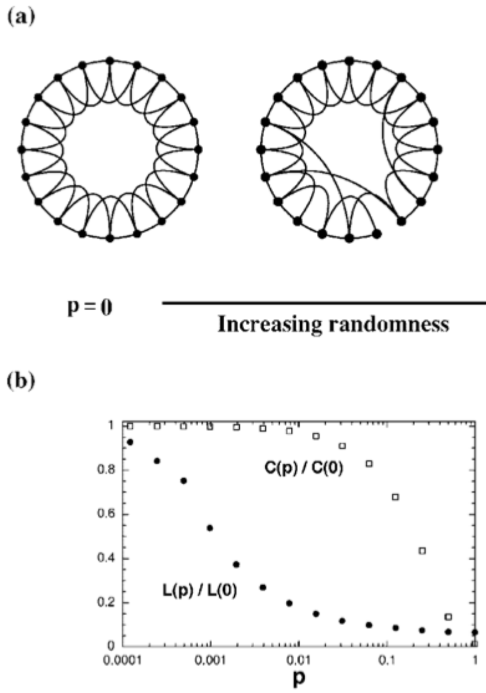
- if not, they send it to an friend who is more likely to know the target

- 64 letters reached the target using between 2 an 10 steps $\rightarrow$ average path length in $[5.5, 6]$

Samples: Kevin Bacon or Erdös numbers

Small World network

The Watts-Strogatz model. It begins with a regular lattice and rewires edges at random.

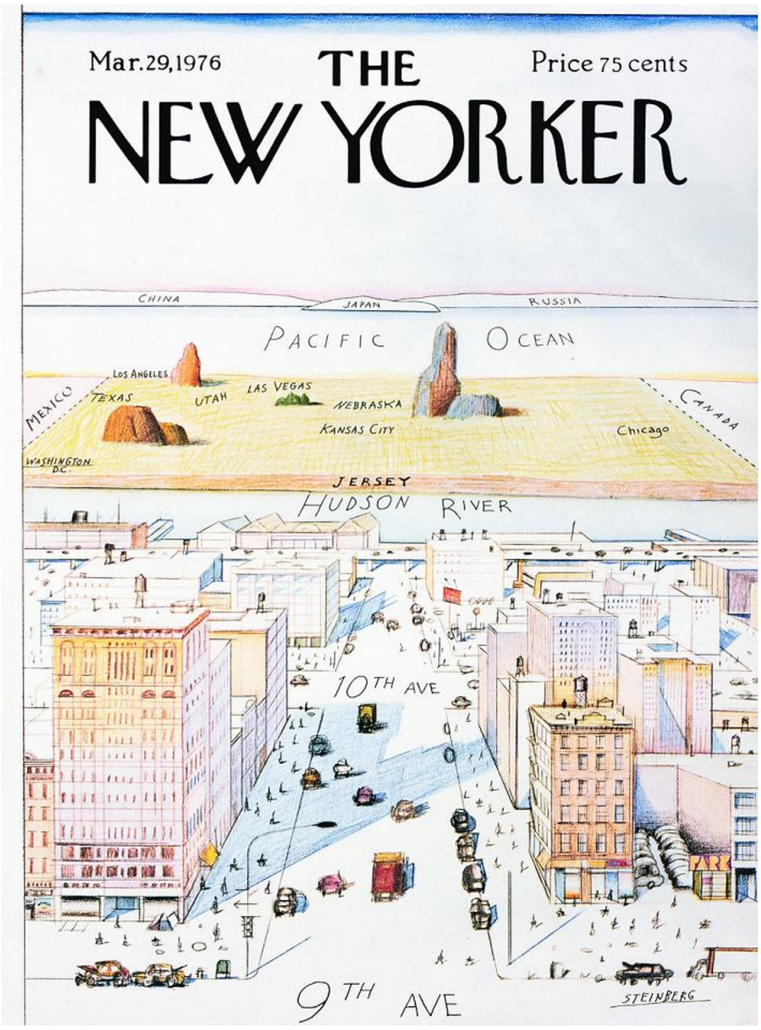


Small-world effect appears because of

- high clustering
- short path lengths

. . . but it still has a Poisson degree distribution

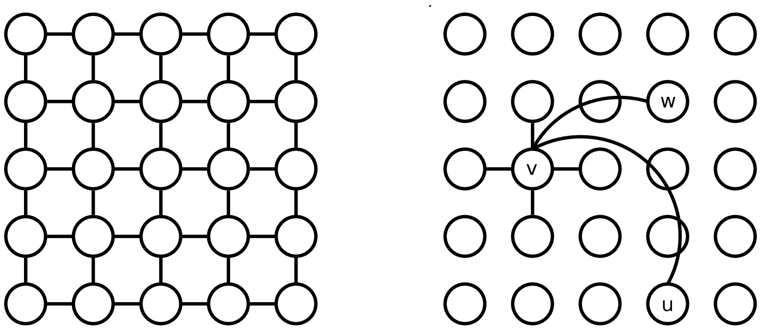
Searchability



Kleinberg proved that, besides these short paths exists, furthermore, people can locate this paths using only their local information.

Searchability

It begins with a regular, 2-d lattice and add q long-range edge at random with a probability proportional to the Manhattan distance



The network shows

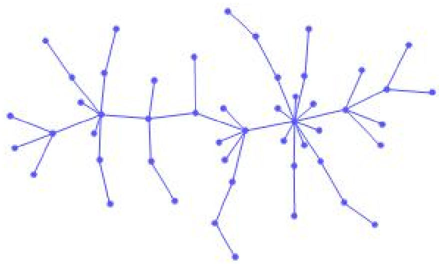
- short path lengths

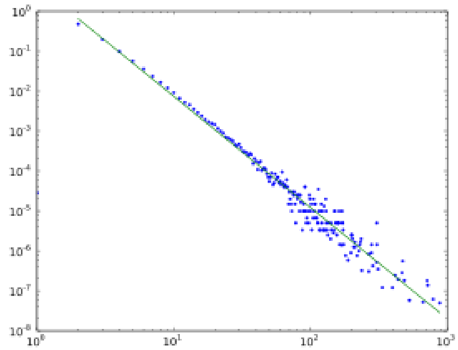
- searchability (distance)

. . . but a low clustering

Preferential attachment

The Barabasi-Albert model. One node added at a time (population growth), tied to a random node with a probability that depends on its degree (preferential attachment).





- high clustering, but dependent on the network size

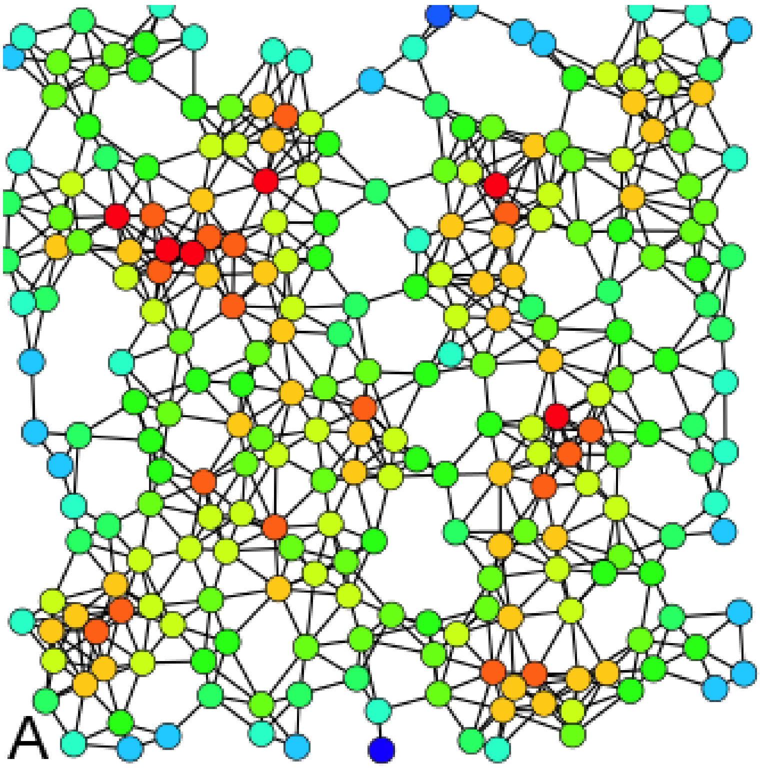
- short path lengths

- power law degree distribution $p(k) \sim k^{-\alpha}$, $2 < \alpha < 3$

- Matthew effect (the richest get richer)

3. Network Characterization

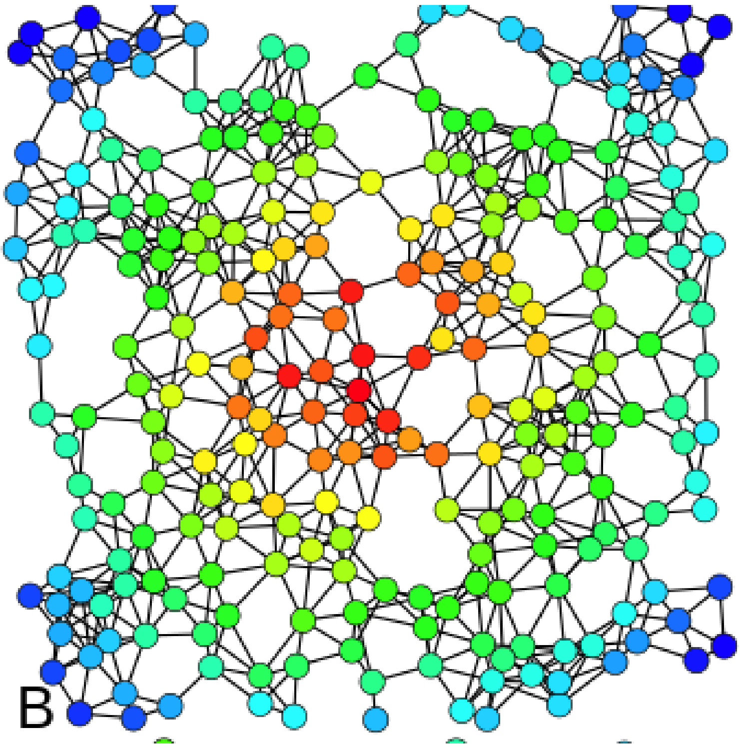
Centrality. Degree



The importance of a node in the network depends on its degree

$$C_D(i) = d_i$$

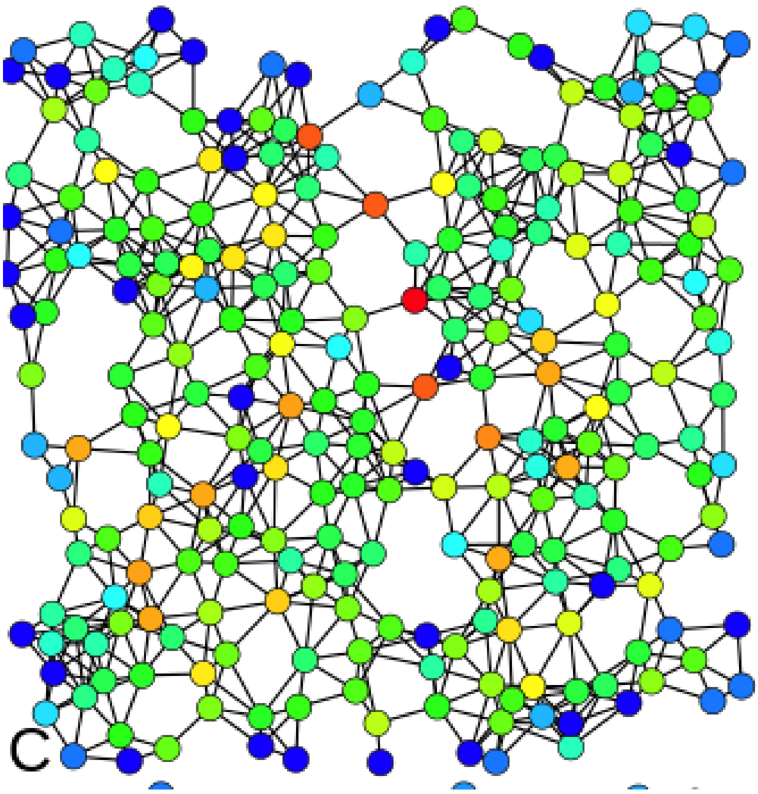
Centrality. Closeness



The most central node is the nearest one to any other node

$$C_C(i) = \frac{1}{\sum_{j \neq i} d(i,j)}$$

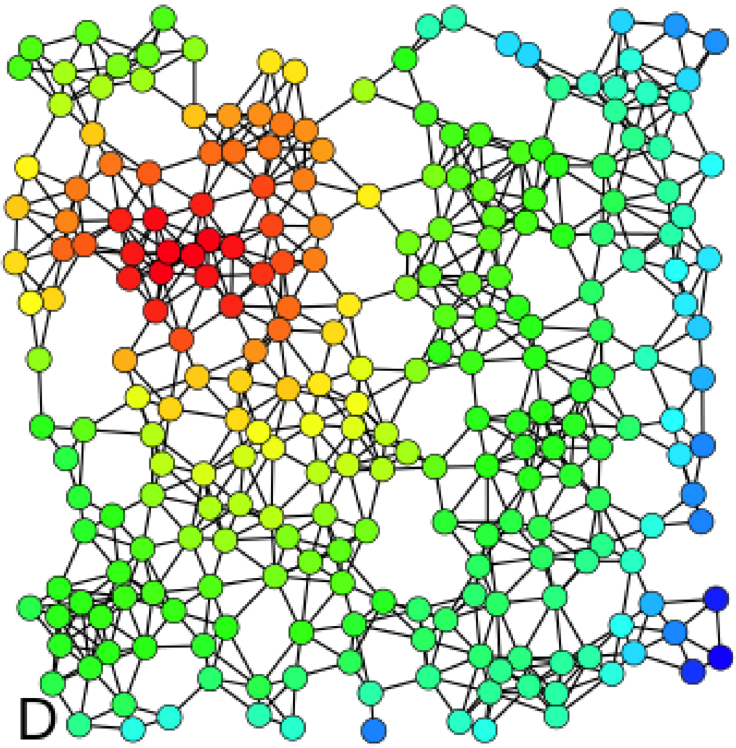
Centrality. Betweenness



The importance of a node in the network depends on how many shortest paths pass through it (bridge)

$$C_B(i) = \sum \frac{\#shortestpaths_{st}(i)}{\#shortestpaths_{st}}$$

Centrality. Eigenvalue

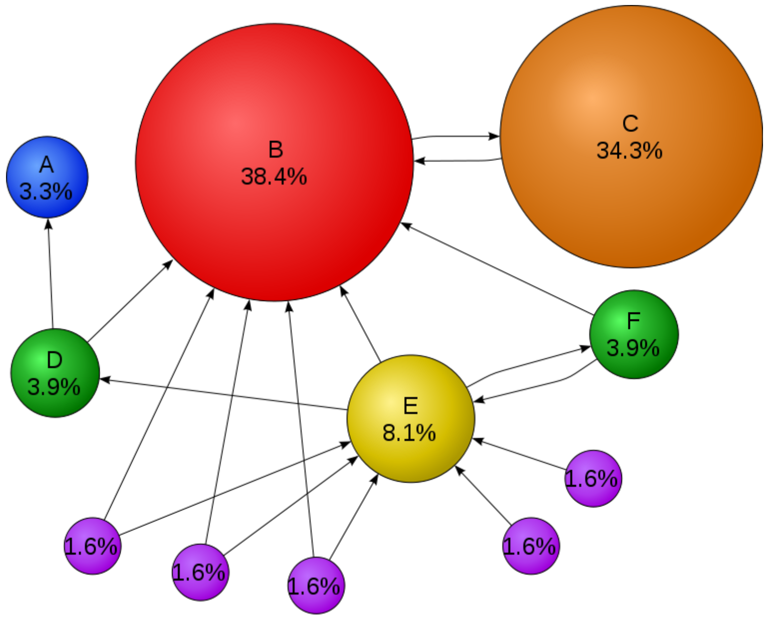


The importance of a node depends on the importance of its neighbors

$$C_E(i) = \frac{1}{\lambda} \sum_{j \in N(i)} C_E(j)$$

and $Av = \lambda v$, with λ the highest eigenvalue (dominant) of matrix A

Eigenvalue Centrality. Pagerank

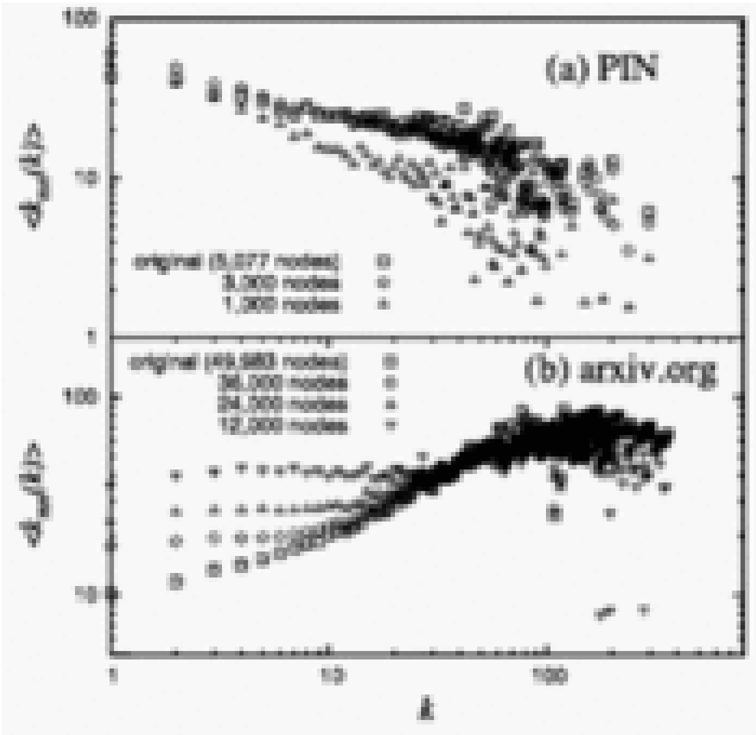


Used by Google to rank the importance of web pages.

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in N(i)} \frac{PR(j)}{d_j)}$$

being $d \simeq 0.85$ a dumping factor

Assortativity (degree correlation)

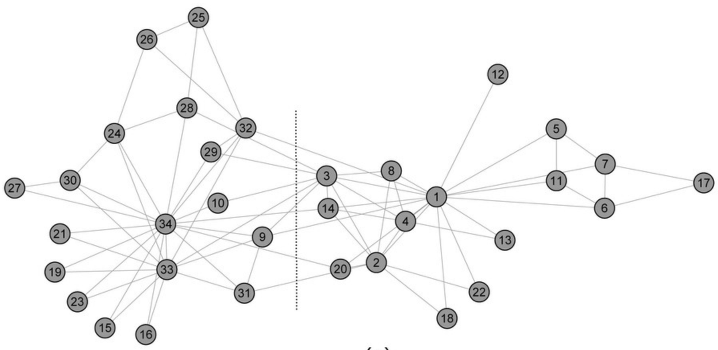


- Determines if the nodes with highest degree are connected with nodes with high degree (positive) or with nodes with low degree (negative).

- It is a correlation measure.

- Homophily: related with other characteristic different from degree.

Community detection



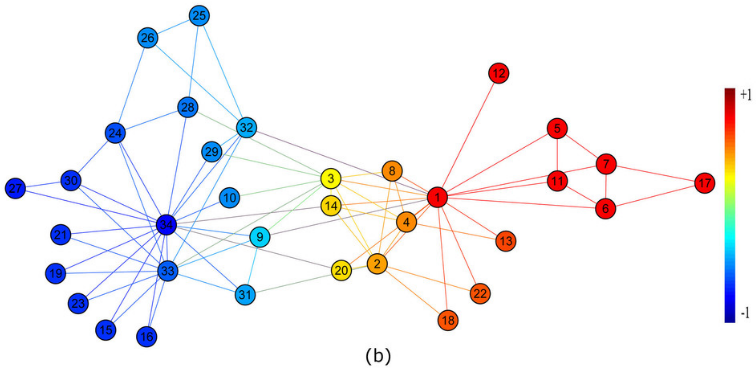
(a)

A common problem is try to find communities with strong ties inside a bigger network

Hierarchical clustering based on a similarity measure: cosine distance, Jaccard similarity or Hamming distance usually

Girvan–Newman algorithm remove edges by its betweenness (break ties among communities)

Community detection



(b)

A common problem is try to find communities with strong ties inside a bigger network

Hierarchical clustering based on a similarity measure: cosine distance, Jaccard similarity or Hamming distance usually

Girvan–Newman algorithm remove edges by its betweenness (break ties among communities)

Network efficience

$$E = \frac{1}{N(N-1)} \sum_{i,j \neq i} \frac{1}{l_{ij}}$$

The bigger the distance d_{ij} is, the less efficient the transmission of information is and, therefore, the network is less efficient. Small-world networks are efficient because they have a small diameter.

Network resilience

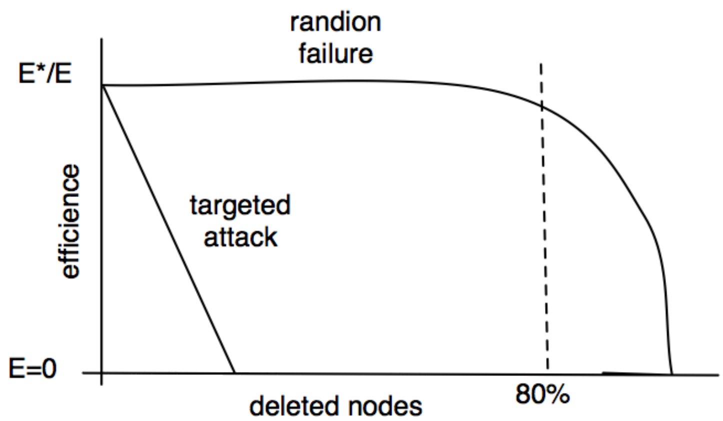
Vulnerability measures haw affects to the efficiency the failures in the network. If $\Delta E(j)$ is the variation of the efficiency when link j disappears,

$$\Delta E(j) = E - E^*_{(j)}$$

the vulnerability of the network is defined as

$$V_{(d)} = \max_j \ \Delta E(j)$$

Dependence on the topology: random or preferential attachment networks behaves differently under random or deliberated (targeted) attacks.



4. References

Lectures. Books

Duncan J. Watts. *Six Degrees: The Science of a Connected Age*, W. W. Norton and Company. 2003.

A.L. Barabasi. *Linked: The New Science of Networks*, Perseus, Cambridge, MA, 2002.

Mark Buchanan. *Nexus: Small Worlds and the Groundbreaking Theory of Networks*, W. W. Norton and Company. 2003.

Lectures. Review papers

Duncan J. Watts. The 'New' Science of Networks'. *Annu. Rev. Sociol.* 2004, 30:243–70.

M.E.J. Newman. The structure and function of complex networks. *SIAM Review* 2003, 45:167–256.

S. Boccaletti, et al. . Complex networks: Structure and dynamics, *Phys. Rep.* 2006, 424, 175.]

Databases

University of Michigan (Newman)
http://www-personal.umich.edu/~ mejn/netdata/

University of Notre Dame (Barabasi)
http://www.nd.edu/~ networks/resources.htm

Universitat Rovira y Virgili (A.Arenas)
http://deim.urv.cat/~ aarenas/data/welcome.htm

Indiana University
http://iv.slis.indiana.edu/db/index.html