Data and Networks
Part II

Dr. Franck Kalala Mutombo

African Institute for Mathematical Sciences (AIMS) South Africa

franckm@aims.ac.za

May 19, 2017

Outline

Centrality measures in complex networks

Complex networks: motivation and background

- The XXI century is becoming the century of networks. The concept of network is very intuitive to everyone in our modern society.
- The word network, meaning a net-like arrangement of threads, wires, etc., appeared by the first time in the English language in 1560 in the Geneva bible, Exodus xxvii 4: And thou shalt make unto it a grate like networke of grass.
- A search for the word 'network' through Google (on 10th May 2016) gives rise to 222×10^7 items. Just for comparison a similar search for the words 'coffee', 'football' and 'sex' gives rise to 113×10^7, 116×10^7, 175×10^7 of items, respectively.

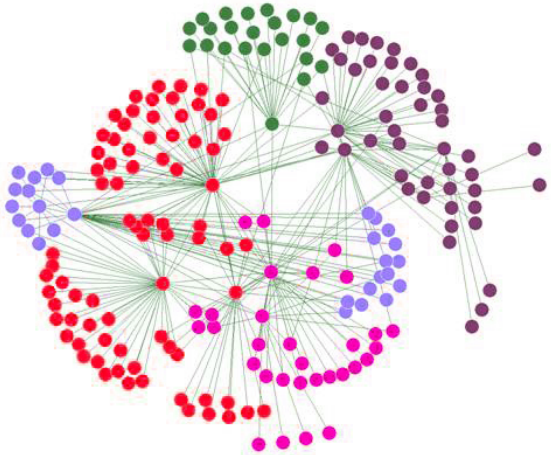
Complex networks: motivation and background

1. Networks, in particular complex networks, provide for a wide variety of physical, biological, engineered or social systems.
2. For example: molecular structure, gene and protein interaction, anatomical and metabolic networks, food webs, transportation networks, power grids, financial and trade networks, social networks, the internet, the WWW, Facebook, Twitter,..
3. Network Science is the study of networks, both as mathematical structures ans as concrete, real world objects. It is a growing multidisciplinary field, with important contributions not just from mathematicians, computer scientists and physicists but also from social scientists, biologists, public health researchers and even from scholars in the humanities.

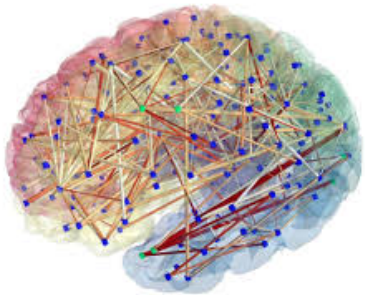
Network today

(a) social

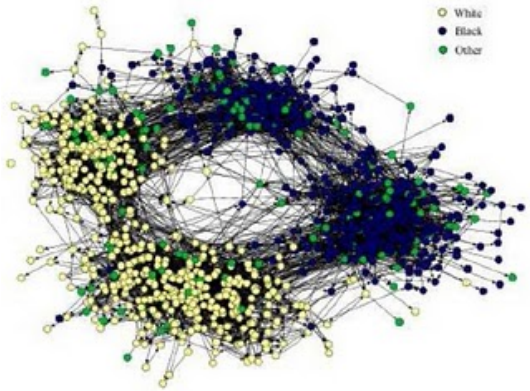
(b) biological



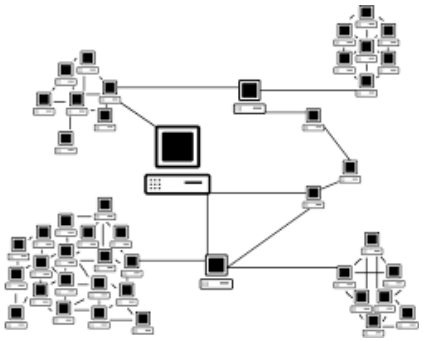
(c) social with communities



(d) brain



(e) friendship



(f) internet

Complex networks: motivation and background (cont.)

1. The field has its origins in the work of psychologists, sociologists, economists, anthropologists and statisticians dating back to the late 1940s and early 1950s. In the last 15 years or so, physicists, computer scientists and mathematicians have entered the scene and made the field more mathematically sophisticated.
2. Basic tools for the analysis of networks include graph theory, linear algebra, probability, numerical analysis, and of course algorithms and data structures from discrete mathematics. More advanced techniques include statistical mechanics and multilinear algebra.

Basic references

Some classic early references:

1. J. R. Seely, *The net of reciprocal influence: A problem in treating sociometric data*, Canadian J. Psychology, 3 (1949), pp. 234-240.
2. L. Katz, *A new status index derives from sociometric data analysis*, Psychometrika, 18 (1953), pp. 39-43.
3. A. Rapoport, *Mathematical models of social interaction*, in Handbook of Mathematical Psychology, vol. 2, pp. 493-579. Wiley, New York, 1963.
4. D. j. de Solla Price, *Networks of scientific papers, Science*, 149(1965), pp. 510-515.
5. S. Milgram, *The small world problem*, Psychology Today, 2 (1967), pp. 60-67.
6. J. Travers and S. Milgram, *An experimental study of the small world problem, Sociometry,* 32 (1969), pp. 425-443.

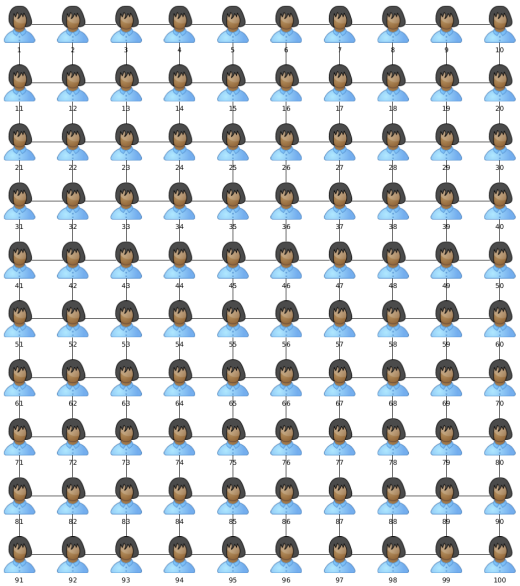
Basic references (cont.)

The field exploded in the late 1990s due to several breakthroughs by physicists, applied mathematicians and computer scientists.
Landmark pappers include

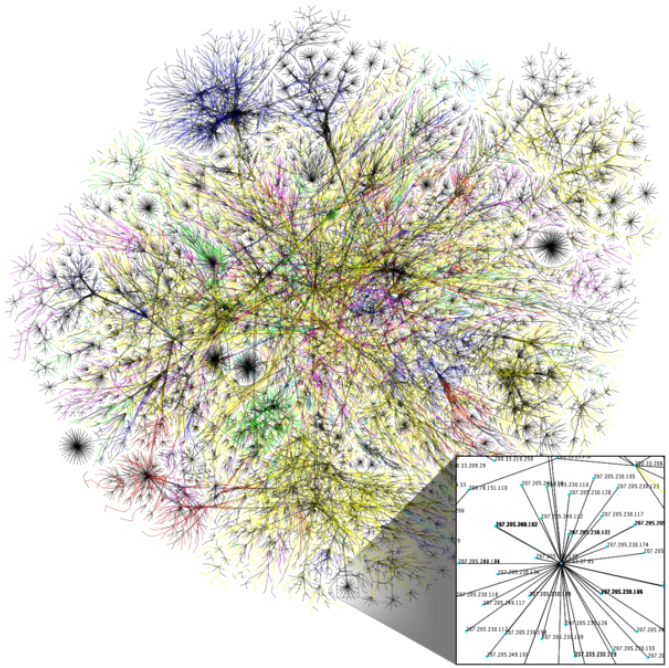
1. D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks*, Nature, 393 (1998), pp. 440-442.
2. L. Katz, *A new status index derives from sociometric data analysis*, Psychometrika, 18 (1953), pp. 39-43.
3. L.-A. Barabaási and R. Albert,*Emergence of scaling in random netwoks*, Science, 386 (1999), pp. 509-512.
4. M. E. J. Newman,*Models of the small world*, J. Stat. Phys., 101 (2000), pp. 819-841.
5. J. Kleinberg, *Navigation in small world problem*, Nature, 406 (2000), p. 845.
6. R. Albert and L.-A. Barabási, *Statistical mechanics of complex networks*, Rev. Modern Phys. 74 (2002), pp. 47-97.

Complex networks: motivation and background

what exactly is complex network?



(g) grid lattice



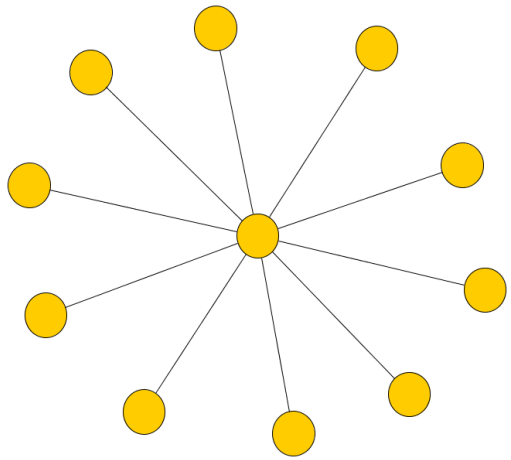
(h) internet

It is easy to tell which graphs are not complex networks.

Complex networks: motivation and background

what exactly is complex network?

(i) Transportation network

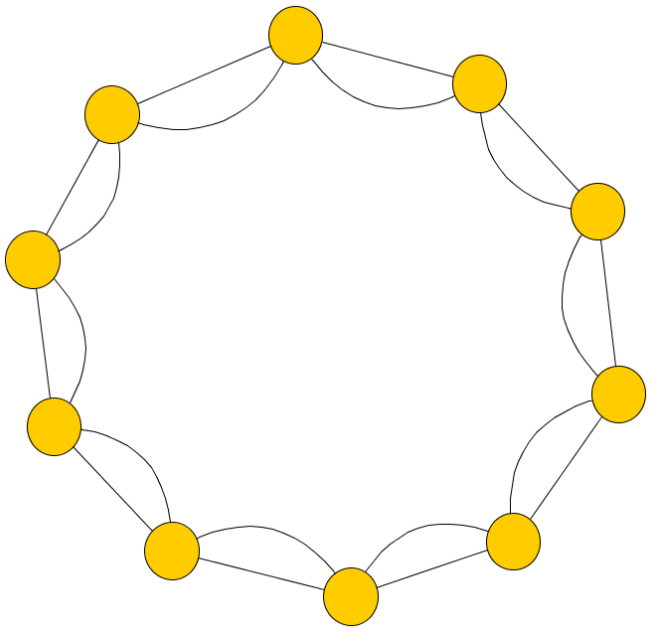


(j) Star graph

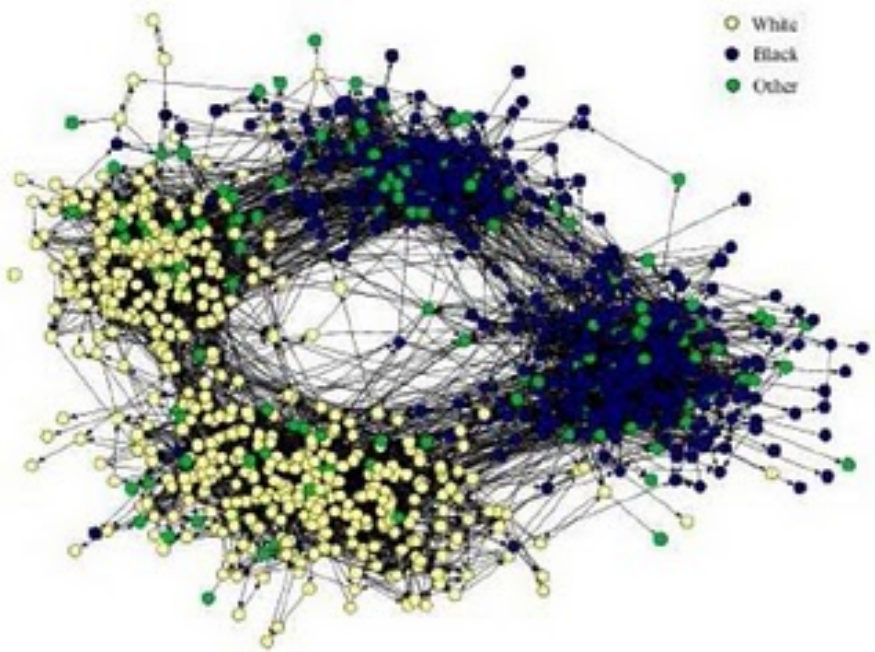
It is easy to tell which graphs are not complex networks.

Complex networks: motivation and background

what exactly is complex network?



(k) Ring lattice



(l) Star

It is easy to tell which graphs are not complex networks.

Complex networks: motivation and background

※

Unfortunately,no precise definition exists, although there is some ongoing work on characterizing (and quantifying) the degree of complexity in a network.

☒

Regular lattices, ring lattice, star graph are not considered complex networks, and neither are completely random graphs such as the Erdos-Rényi model.

⊎

Random graphs, however, are useful as null models against which compare (possible examples of) complex networks.

Some common features of complex networks

Some of the attributes typical of many real-world complex networks are:

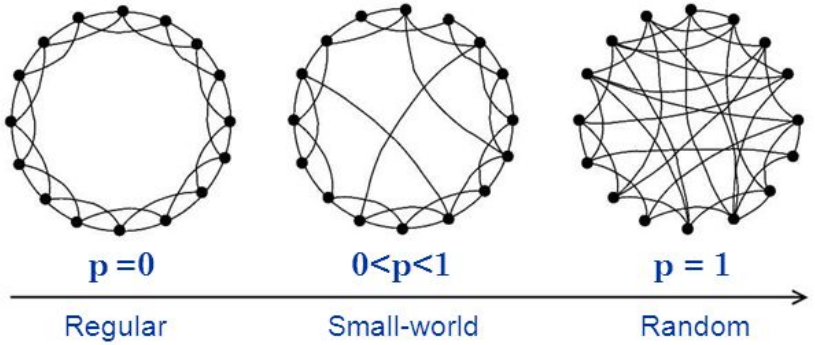
- "Scale-free": the degree distribution follows a power law (Paretos curve)
- "Small-world":
  - Small graph diameter, short average distance between nodes
  - High clustering coefficient: many triangles, hubs, ...
- Hierarchical structure
- Rich in "motifs"
- Self-similar (as in fractals)

Briefly stated: complex networks exhibit a non-trivial topology.

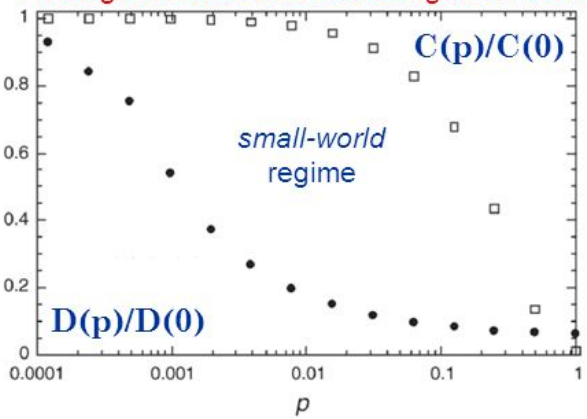
The Watts-Strogatz random rewire model

SMALL-WORLD model (Watts, Strogatz *Nature* 1998)

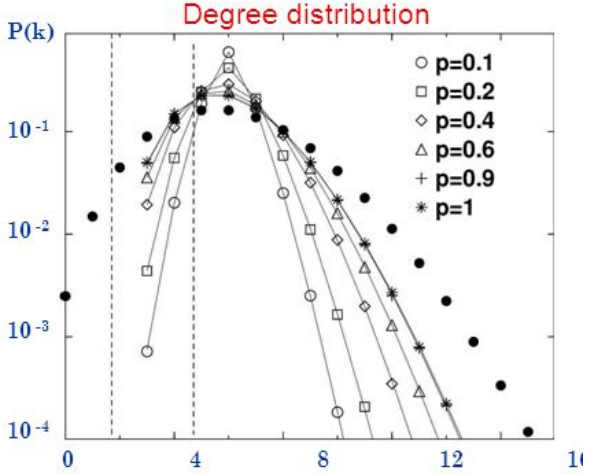
- Start with a regular ***d***-dimensional lattice, connected up to ***q*** nearest neighbours;
- With probability ***p***, an end of each link is *rewired* to a new randomly chosen vertex.



Average distance and clustering coefficient



Degree distribution



The Barabási Albert model: The rich always get richer!

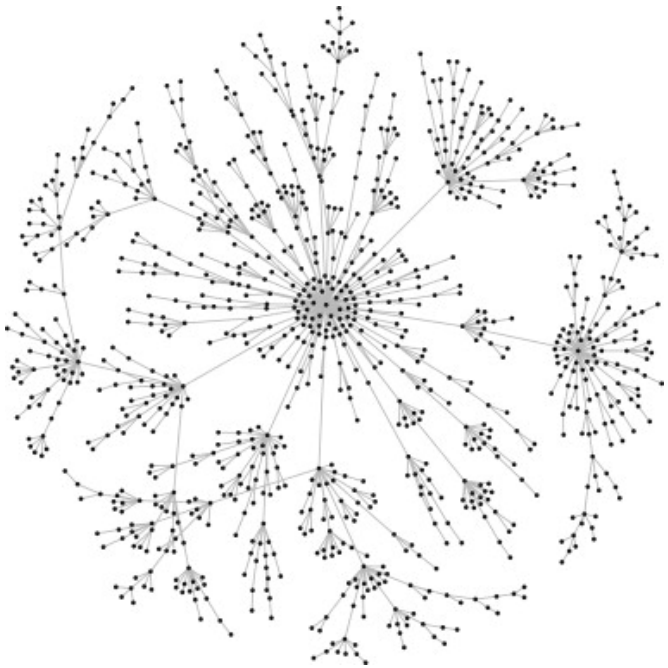
The Barabási-Albert model is based on the notion of preferential attachment:

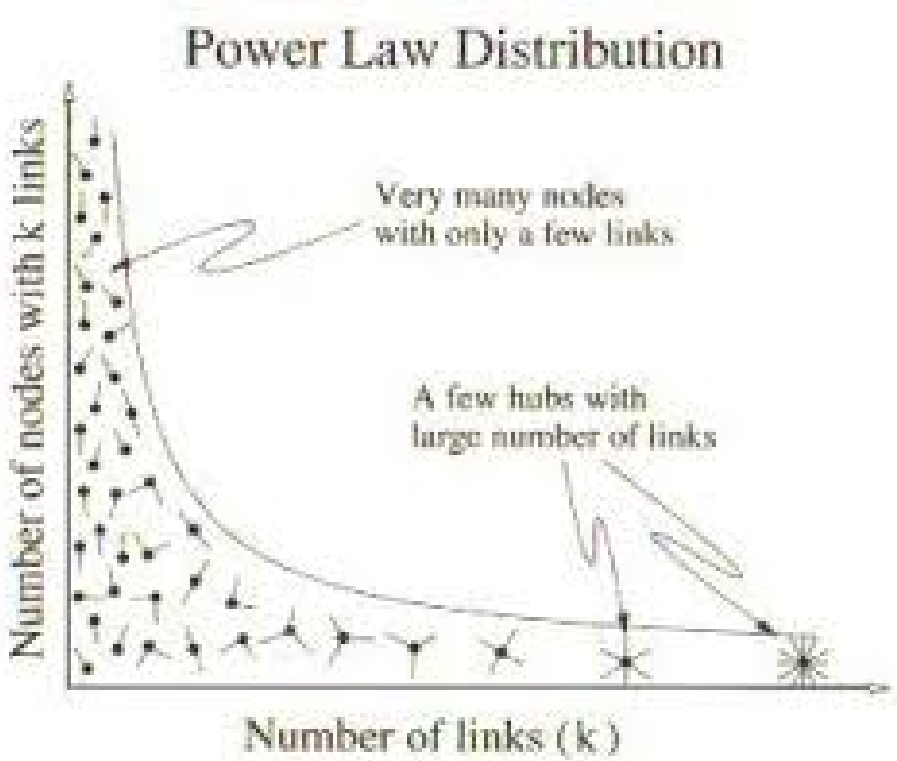
- starting from a given sparse graph, new nodes are added (one at a time) to the network, joining them to existing nodes with a probability proportional to the degree (the number of adjacent nodes) of such nodes.
- Hence, nodes that are rich of connections to other nodes have a high probability of attracting new "neighbours", while "poor" nodes tend to remain such.
- One can show that the Barabási-Albert model leads to a highly skewed degree distribution given by a power law of the form

$$p(k) \sim k^{-\gamma}$$

where $p(k)$ denotes the fraction of nodes of degree k and γ is a constant, typically with $2 \leq \gamma \leq 3$.

The Barabási Albert model: The rich always get richer!





what is network structure or topology

- Complex networks are the skeletons of complex systems
- Complex systems are composed of interconnected parts which display some properties that are not obvious from the properties on individual parts.

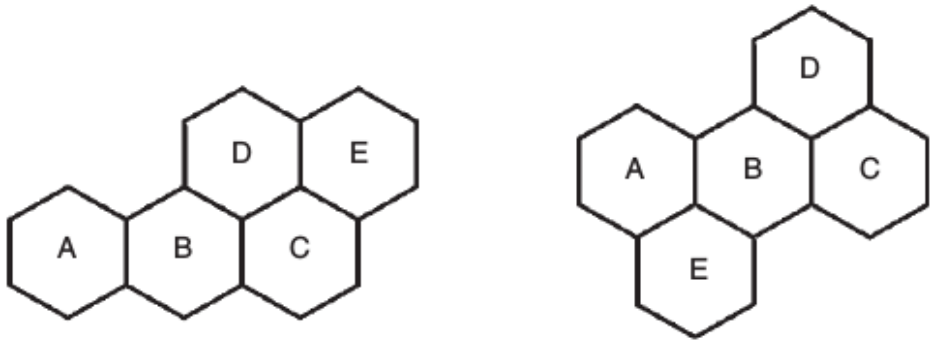
In their essay "Models of structure", Cottrell and Pettifor (2000) have written

> There are three main frontiers of science today. First, the science of the very large, i.e. cosmology, the study of the universe. Second, the science of the very small, the elementary particles of matter. Third, and by far the largest, is the science of the very complex, which includes chemistry, condensed- matter physics, materials science, and principles of engineering through geology, biology, and perhaps even psychology and the social and economic sciences.

Network structure

÷

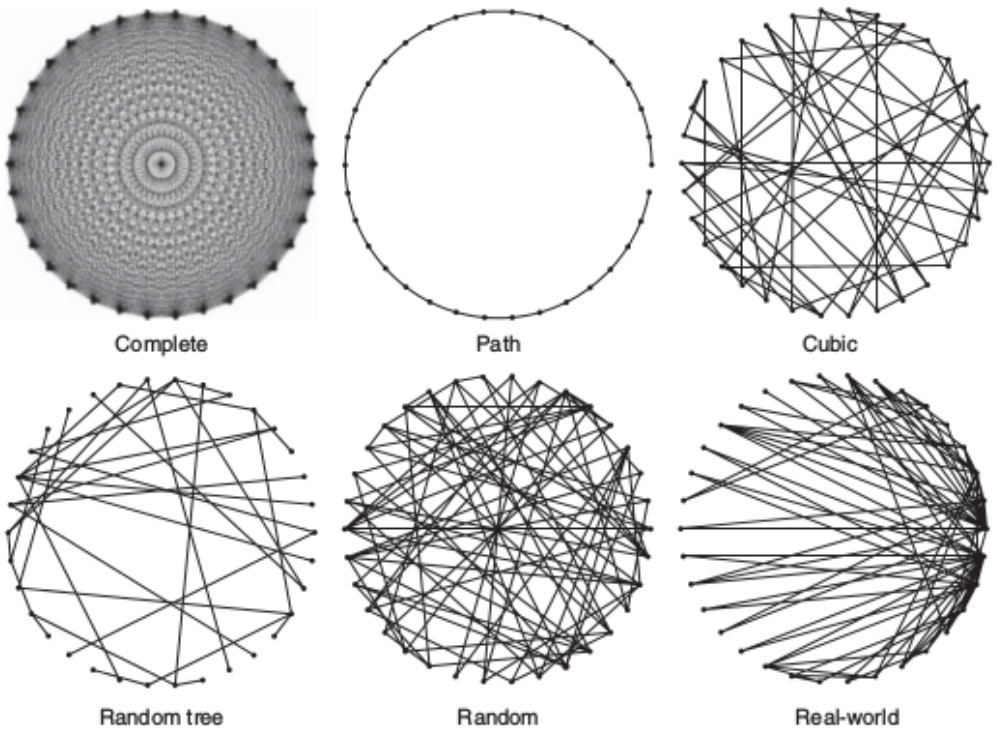
Structure of a network can determine many, if not all, of the properties of the complex system represented by it. It is believed that network theory can help in many important areas of molecular sciences, including the rational design of new therapeutic drugs (Csermely et al., 2005).



(m) benzene-a-pyrene (n) pyrene

Figure : Molecular isomeric network. Network representation of two polycyclic aromatic compound having each 20 nodes. Left, beneze[a]-pyrene highly carcinogenic found in tobacco smoke and pyrene left non carcinogenic compound.

Network structure



Network structure

Network analysis

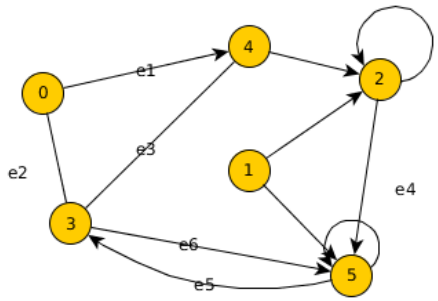
Basic questions about network structure include centrality, robustness, communicability and community detection issues:

1. Which are the most "important" nodes?
   - Network connectivity and robustness/vulnerability
   - Identification of influential individuals in social networks
   - Essential proteins in PPI networks (lethality)
   - Identification of keystone species in ecosystems
   - Author centrality in collaboration networks
   - Ranking of documents/web pages on a given topic
2. How do "disturbances" spread in a network?
   - Spreading of epidemics, beliefs, rumors, fads,...
   - Routing of messages; bottlenecks, returnability
3. How to identify "community structures" in a network?
   - Clustering, triadic closure (transitivity)
   - Partitioning

Formal definitions

Real-world networks are usually modelled by means of graphs.

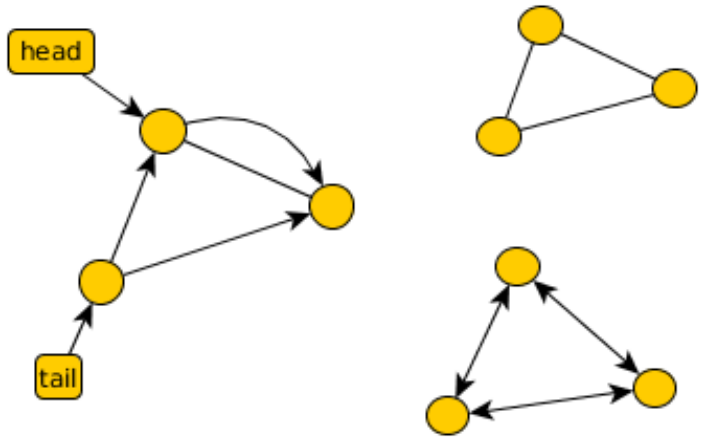
- A graph is a set of vertices and a set of lines between pairs of vertices.
- A graph represents the structure of network. It need a set of vertices and a set of lines.
- A vertex is the smallest unit of a network and a line is tie between two vertices in the network.
- A loop is a special kind of line that connect a vertex to itself.



- A line can be directed (edge) or undirected (arc).

Formal definitions (cont.)

- A directed graph or a digraph contains one or more arcs.
- An undirected graph contains no arcs: all its lines are edges.
- In a graph multiple lines are allowed.
- A graph is simple if it has no multiple lines.
- A simple undirected graph contains no loops.
- A simple directed graph can contains loops.



- A simple undirected graph contains neither multiple edges nor loops.
- A simple directed graph contains no multiple arcs.

Formal definitions (cont.)

Definition

A graph is a pair $G = (V, E)$, where V is a set of vertices or nodes, and E is a set of edges between the vertices $E \subseteq \{(u, v)|u, v \in V\}$. The number of node is $n = |V|$ and the number of edges $m = |E|$.

- A graph may be undirected, that is edges have no orientation or directed i.e. edges have direction and are called arcs.
- A graph is simple if it has no loops and no more than one edge between any two different vertices.
- The degree of a vertex is the number of edges that connect to it. We shall consider here simple and undirected graphs.

Adjacency matrix

1. To every unweighted graph $G(V, E)$ we associated is adjacency matrix A defined as follow

$$A_{ij} = \begin{cases} 1 & \textit{if } i \textit{ and } j \textit{ are connected} \\ 0 & \textit{otherwise} \end{cases}$$

2. If G is an undirected graph, A is symmetric with zeros along the main diagonal. In this case, the eigenvalues of A are all real.
3. We label the eigenvalues of A in non-increasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$.
4. If G is connected, then λ_1 is simple and satisifies $\lambda_1 > |\lambda_i|$ for $2 \leq i \leq N$ (Perron-Frobenius Theorem).

Formal definitions (cont.)

- A walk of length k in G is a series of edges $(u_1, v_1), (u_2, v_2), \cdots, (u_p, v_p)$ for which $v_i = v_{i+1}$.
- A closed walk is a walk where $v_p = v_1$.
- A path is a walk with no repeated nodes.
- A cycle is a path with and edge between the first and the last node. In other words, a cycle is closed path.
- A triangle in G is a cycle of length 3.
- The Shortest path distance is the number of links/Edges in the shortest path connecting two nodes. This is also known as the geodesic distance.
- The diameter of a graph $G(V, E)$ is defined as

$$diam(G) = \max_{v_i, v_j \in V} d(v_i, v_j) \tag{1}$$

Degree, simple graph

Definition

Let $v \in G$ be a vertex of a graph G. The *neighbourhood* of v is the set

$$N_G(v) = \{u \in G | vu \in E(G)\}.$$

The degree of the node v is defined to be

$$k_v = |N_G(v)|.$$

$$k_{\min}(G) = \min\{k_v | v \in G\} \quad \text{and} \quad k_{\max}(G) = \max\{k_v | v \in G\}.$$

The column vector of node degrees for a graph G is given by

$$\mathbf{k} = (\mathbf{1}^T \mathbf{A})^T = \mathbf{A1}, \text{ where}$$

$\mathbf{1}$ is $| V | \times 1$ all-one vector.

Degree, directed graph

For an directed graph we define two types of degree; the in-degree which is the number of links pointing towards a given vertex defined by

$$\mathbf{k}^{in} = (\mathbf{1}^T \mathbf{A})^T,$$

or, for each component,

$$k_i^{in} = \sum_j a_{ji},$$

and the out-degree which is the number of links departing from the corresponding node and defined by

$$\mathbf{k}^{out} = \mathbf{A1}$$

or, for each component,

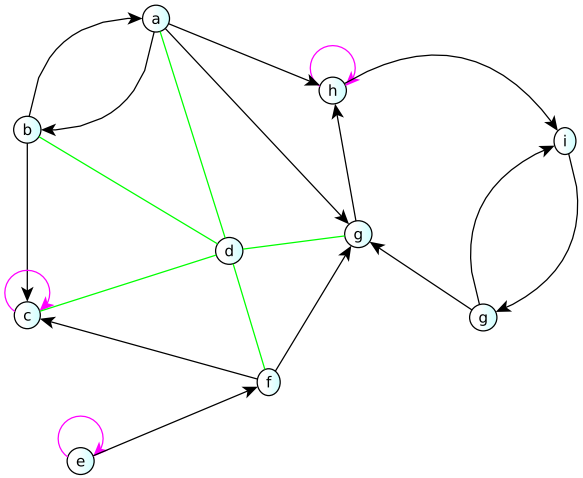
$$k_i^{out} = \sum_j a_{ij}.$$

The total degree of a node in this case is then given by

$$\mathbf{k} = \mathbf{k}^{in} + \mathbf{k}^{out}$$

Degree, in-degree, out-degree, total degree

1. $k_g^{in} = 4$, the in-degree g
2. $k_g^{out} = 1$, the out-degree of g
3. $k_g = k_g^{in} + k_g^{out} = 5$, total degree



○

The average node degree in a graph is defined by

$$\overline{k} = \frac{1}{n}\mathbf{1}^T\mathbf{k} = \frac{1}{n}\sum_{i=1}^{n} k_i.$$

clustering

- A clustering coefficient measures the degree to which the nodes in a network tend to cluster together. For a node v i with degree d i , it is defined as

$$CC(i) = \frac{2\Delta_i}{d_i(d_i - 1)}$$

  where Δ_i is the number of triangles in G having node v_i as one of its vertices.
- The clustering coefficient of a graph G is defined as the average of the clustering coefficients over all the nodes of degree ≥ 2.
- Many real world small-world networks, and particularly social networks, tend to have high clustering coefficient.
- This is not the case for random networks. For example, Erdös-Rényi (ER) graphs. are small-world graphs but have very low clustering coefficients. Also, the degree distribution in ER graphs falls off exponentially (does not follow a power law).

- The number of triangles in G that a node participates in is given by

$$\Delta_i = \frac{1}{2}\left(A^3\right)_{ii},$$

while the total number of triangles in G is given by

$$\Delta(G) = \frac{1}{6} Tr(A^3).$$

- Hence, computing clustering coefficients for a graph G requires estimating $Tr(A^3)$, which for very large networks can be a challenging task.
- We note for many networks, A^3 is a rather dense matrix.
- For example, for the PPI network of beer yeast the percentage of non-zero entries in A^3 is about 19%, compared to 0.27% for A. This fact is related to the small-world property.

Average path length

Definition

Let $G = (V, E)$ be a graph the average path length l_G is defined by

$$l_G = \frac{1}{n(n-1)} \sum_{i,j} d(v_i, v_j) \tag{2}$$

where $d(v_i, v_j)$ is the shortest path between vertex v_i and v_j and n is the total number of nodes in G.

- In words is defined as the average number of steps along the shortest paths for all possible pairs of network nodes

Network properties

Complex graphs arising in real-world applications tend to be highly irregular and exhibit a non-trivial topologyin particular, they are far from being either highly regular, or completely "random". Complex networks are very often

- Scale-free, meaning that their degree distribution tends to follow a power law: $p(k) =$ number of nodes of degree $k \approx ck^{-\gamma}$, $\gamma > 0$. Frequently, $2 \leq \gamma \leq 3$. This implies sparsity but also the existence of several highly connected nodes (hubs).
- Small-world, meaning that the diameter grows very slowly with the number N of nodes; e.g.,

$$diam(G) = \mathcal{O}(logN), \quad N \to \infty. \tag{3}$$

- Highly clustered, i.e., they contain a very large proportion of triangles (unlike random graphs).

Summary of complex networks characteristics

Random networks

completely random graphs (like Erdös-Rényi graphs) are not scale-free and have low clustering coefficients. This makes them ill-suited as models of real-world complex networks.

Small-World networks

- The Watts-Strogatz (WS) model starts with a regular graph (say, a ring), which is then "perturbed" by rewiring some of the links between nodes in a randomized fashion.
- The WS model interpolates between a regular and a random graph model.
- With this technique, one can obtain small-world graphs with high clustering coefficients; the degree distribution, however, is rather homogeneous (i.e., WS graphs are not scale-free).

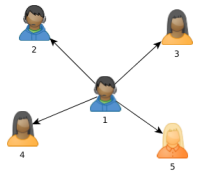
Summary of complex networks characteristics (cont.)

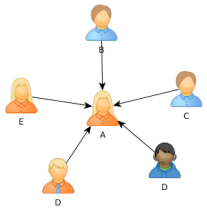
Scale-free networks

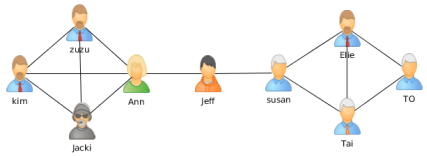
- The Barabási-Albert (BA) model uses a preferential attachment, or rich get richer, mechanism to evolve a given initial graph.
- The resulting networks are small-world, scale-free, and have high clustering coefficients.
- The study of generative models for constructing complex graphs with prescribed properties is still undergoing intensive development.

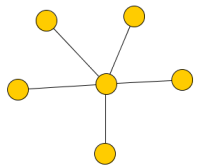
Centrality measures

- A central node is import/or powerful
- A central node has an influential position in the network
- A central node has an advantageous position in the network







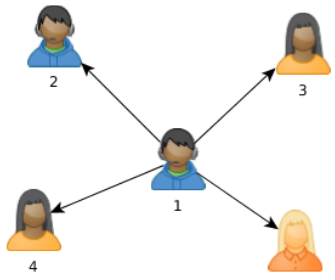


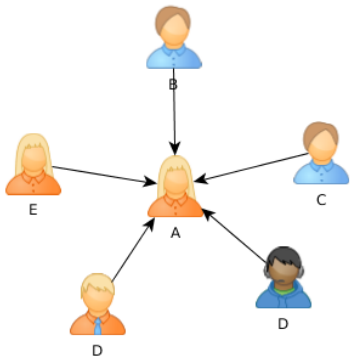
Degree Centrality (force/power through links)

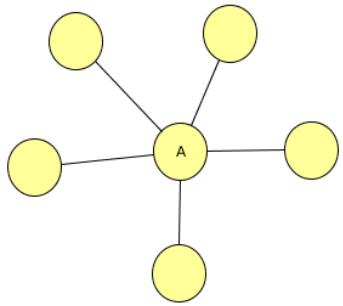
$$C_D(i) = \sum_j A_{i,j} = k(i)$$

$$C_D^{in}(i) = \sum_j A_{i,j} = k^{in}(i)$$

$$C_D^{out}(i) = \sum_j A_{i,j} = k^{out}(i)$$

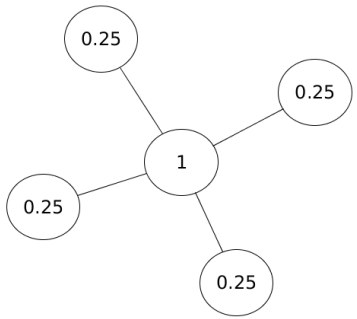


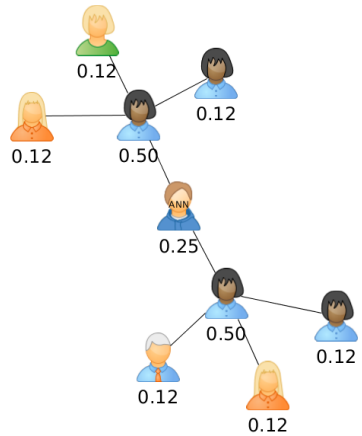




Degree Centrality (force/power through links)

We can normalise the degree centrality by dividing it by the maximum centrality value possible; $n - 1$ (so values are in between 0 and 1)

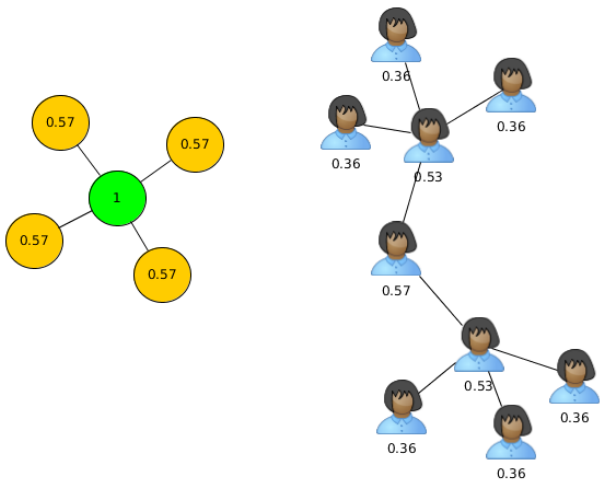




The degree is very cheap to compute but is unable to recognize the centrality of certain nodes: its a purely local notion.

Closeness Centrality
power through proximity to others

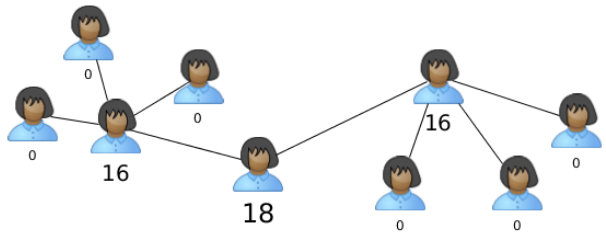
$$C_c(i) = \left(\frac{\sum_{i \neq i} d(i,j)}{n-1} \right)^{-1} = \frac{n-1}{\sum_{i \neq i} d(i,j)}$$



The most important node is the one which close to everybody else, i.e., to one which is easily reachable or have the power to quickly reach other.

Betweenness Centrality

A node is important if it lies on many short path. It is playing an important role on passing/spreading information through the network.



$$B_c(i) = \sum_{j<k} \frac{L_{jk}(i)}{L_{jk}}$$

1. L_{jk} is the number of shortest-paths between j and k, and
2. $L_{jk}(i)$ is the number of shortest-paths through i.

Often normalised as

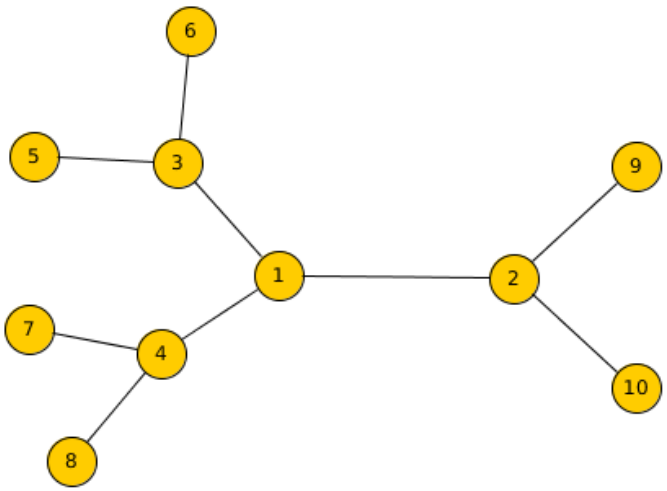
$$NB_C(i) = \frac{2B_c(i)}{(n-1)(n-2)}$$

Betweenness Centrality

- Betweenness and closeness centrality assume that all communication in the network takes place via shortest paths, but this is often not the case.
- This observation has motivated a number of alternative definitions of centrality, which aim at taking into account the global structure of the network and the fact that all walks between pairs of nodes should be considered, not just shortest paths.

Eigenvector centrality (improvement of degree centrality)

The most important node is the one which is connected to the most important. Important node contribute more to centrality. A central node is the one that is connected to other central nodes.



$$Ev_c(i) \propto \sum_{j \neq i} A_{ij} Ec_v(j)$$

Eigenvector centrality (improvement of degree centrality)

Suppose that we have an initial value for all $\mathbf{x}_i(0)$. Then, we compute next iteration of values using the formula

$$\mathbf{x}_i(t+1) = \sum_{j \neq i} A_{ij}\mathbf{x}_i(t) \quad \text{or} \quad \mathbf{x}(t+1) = A\mathbf{x}(t),$$

$$\mathbf{x}(t) = A^t\mathbf{x}(0)$$

Let express $\mathbf{x}(0)$ in terms of the eigenvectors $\mathbf{v}_i$ of A,

$$\mathbf{x}(0) = \sum_i c_i\mathbf{v}_i$$

Let λ_i be the eigenvalues of A and λ_1 be the spectral radius of A,

$$\mathbf{x}(0) = A^t\mathbf{x}(0) = \sum_i c_i\lambda_i^t\mathbf{v}_i = \lambda_1^t \sum_i c_i \left(\frac{\lambda_i}{\lambda_1}\right)^t \mathbf{v}_i$$

Eigenvector centrality (improvement of degree centrality)

Since $\frac{\lambda_i}{\lambda_1} < 1$ for all $i > 1$, all terms (other than the first) decay exponential as t grows. Therefore,

$$\mathbf{x}(t) \rightarrow c_1 \lambda_1 \mathbf{v}_1 \quad \text{as } t \rightarrow \infty$$

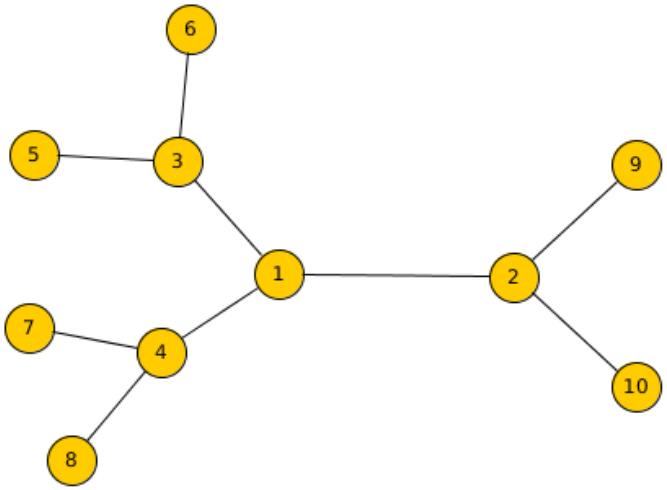
Eigenvector centrality is proportional to the leading eigenvector of A (and hence, the name!).

Equivalently, define centrality vector $\mathbf{x}$ satisfying:

$$A\mathbf{x} = \lambda_1 \mathbf{x}$$

$$Ev_c(i) = \mathbf{x}(i)$$

Eigenvector centrality (improvement of degree centrality)



i	1	2	3	4	5	6	7	8	9	10
Ev(i)	0.55	0.41	0.41	0.41	0.18	0.18	0.18	0.18	0.18	0.18

Centrality: Subgraph Estrada-Rodrìguez-Velsquez, Phys. Rev. E, 2005

- Subgraph centrality measures the centrality of a node by taking into account the number of subgraphs the node "participates" in.
- This is done by counting, for all $k = 1, 2, \cdots$ the number of closed walks in G starting and ending at node i, with longer walks being penalized (given a smaller weight).
- $(A^k)_{ii}$ = number of closed walks of length k based at node i.
- $(A^k)_{ij}$ = number of walks of length k that connect nodes i and j.

Using $^k/k!$ as weights leads to the notion of subgraph centrality:

$$SC(i) = \left(\sum_{k=0}^{\infty} \frac{A^k}{k!}\right)_{ii} = \left(I + \frac{A}{1!} + \frac{A^2}{2!} + \frac{A^3}{3!} + \cdots\right)_{ii} = (e^A)_{ii}.$$

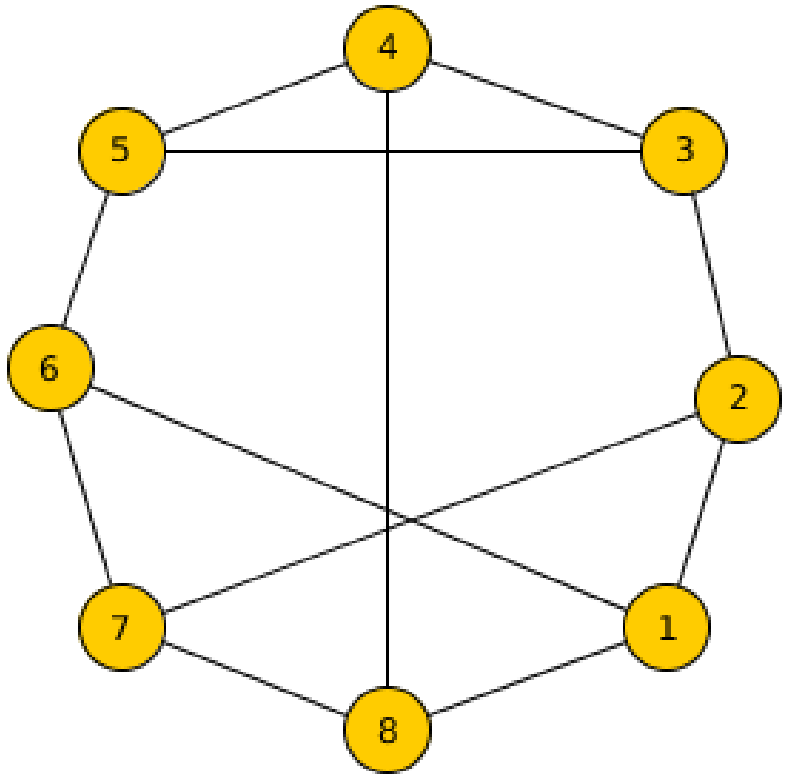
Centrality: Subgraph Estrada-Rodrìguez-Velsquez, Phys. Rev. E, 2005

It is sometimes desirable to normalize the subgraph centrality of a node by the sum

$$EE(G) = \sum_{i=1}^{N} SC(i) = \sum_{i=1}^{N} \left(e^{A}\right)_{ii} = Tr(e^{A}) = \sum_{i=1}^{N} e^{\lambda_i}$$

of all the subgraph centralities. The quantity $EE(G)$ is known as the Estrada index of the graph G.

i	DC	CC	BC	EVC	SC
1	0.43	0.63	0.07	0.35	3.71
2	0.43	0.63	0.1	0.35	3.64
3	0.43	0.63	0.1	0.35	3.90
4	0.43	0.63	0.1	0.35	3.90
5	0.43	0.63	0.1	0.35	3.90
6	0.43	0.63	0.1	0.35	3.64
7	0.43	0.63	0.07	0.35	3.71
8	0.43	0.63	0.1	0.35	3.64



THANK YOU!

AIMS | African Institute for Mathematical Sciences SOUTH AFRICA