Network Science
Introduction to Complex Networks

Franck Kalala M

AIMS Sénégal
University of Lubumbashi

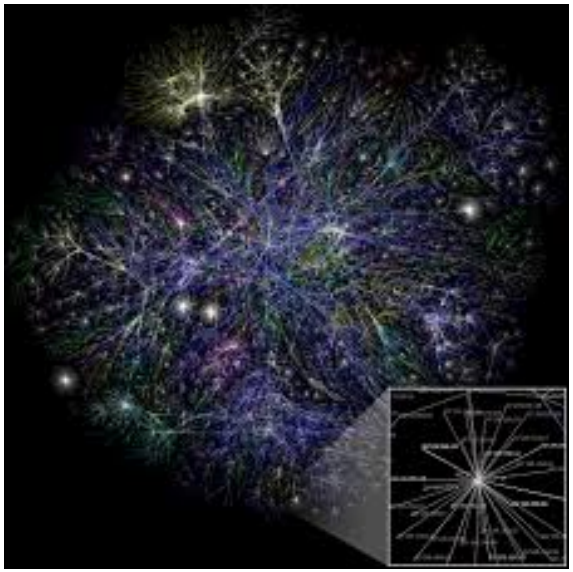
franckm@aims.ac.za | *franckm@aims-senegal.org*

April 10, 2018

Table of contents

network:

all the time, everywhere, with everybody



Complex Networks

Stephen Hawking (8 January 1942 - 14 March 2018)

I think the next century will be the century of complexity

Complexity

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
–adjective

1.
composed of many interconnected parts; compound; composite: a complex highway system.

2.
characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3.
so complicated or intricate as to be hard to understand or deal with: a complex problem.

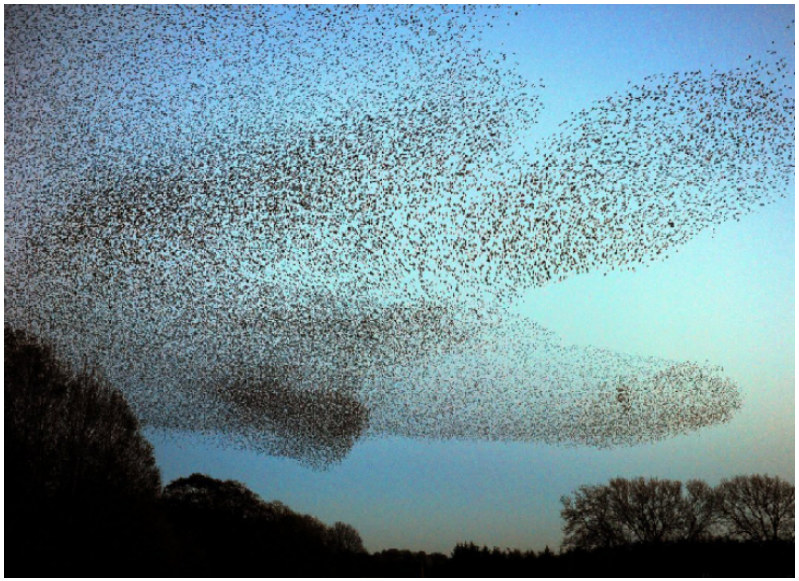
Source: Dictionary.com

Complexity, a **scientific theory** which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

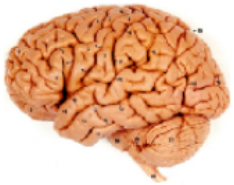
Complexity

Complex systems



1. Self-organised
2. Evolving
3. Adaptive
4. No central organising mind
5. No conventional way of description

Complex Systems: How to approach?


$\neq$



$\neq$



$\neq$


1. Interactions between elements give rise to emergent behaviour
2. This behaviour is apparent at the system level
3. Studying isolated elements is not enough
4. Variations in behaviour of elements often average out at the system level
5. A "holistic", system-level viewpoint is needed!

Complex systems: how to approach

Statistical description

- Systems with random features
- One sample does not characterise the typical behaviour
- Statistical averages of quantities

Empirical data analysis

- How to detect patterns and structure in information?
- How to characterize the system instead of its building blocks?
- Multivariate methods etc

Analytical approach

- Write down (coupled) differential equations for interactions
- Attempt to solve
- Usually no closed-form solutions; numerical solutions, phase space analysis, etc

Simulations

- Postulate rules (e.g. the ant raids)
- Simulate and observe system behaviour
- Try to match empirical observations

Complex systems: network approach

...a way of mapping complexity

Each complex system can be interpreted as a complex network, which identifies the interactions between the interconnected components

The network approach

- Combines the elements of all the other approach
- Disregards (unnecessary) details of the system
- Focuses on the structure of interactions
- Statistical characterisation of system

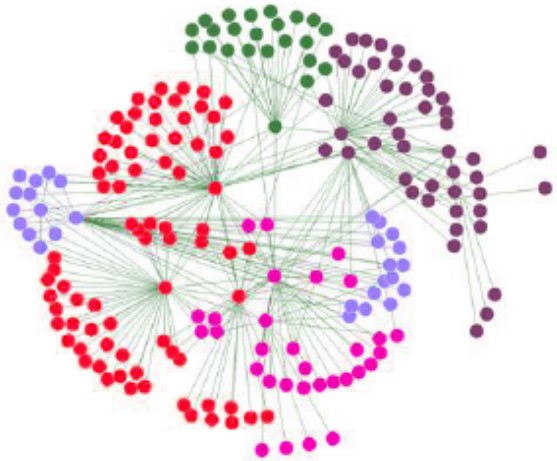
Complex systems: network approach

1. **Measuring** - make observations on Nature
2. **Modelling** - attempt to explain observations:
   - 2.1. Choose the right level of coarse-graining
     - Units: Vertices or nodes ⇔ interacting elements
     - Edges or links ⇔ interactions
   - 2.2. Strip the problem to its simplest form
     - Interaction structure ⇔ evolution and behaviour of system
   - 2.3. Formulate the problem in mathematical terms
     - Statistical analysis of network structure
     - Dynamics of processes taking place on networks
3. **Validating** - check if calculations or simulations can
   - reproduce the observations
   - explain the observations
4. Go back to 1. & 2. and rethink

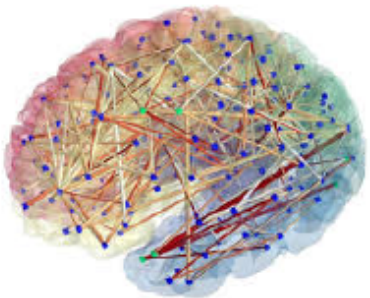
Networks are everywhere

(a) social

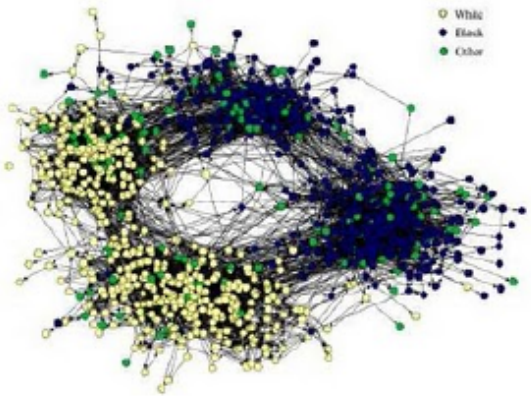
(b) biological



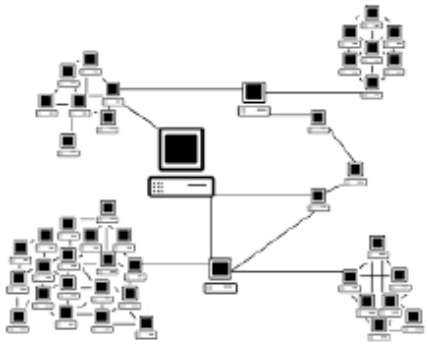
(c) social with communities



(d) brain



(e) friendship



(f) internet

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 and 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 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

Some common features of complex networks

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

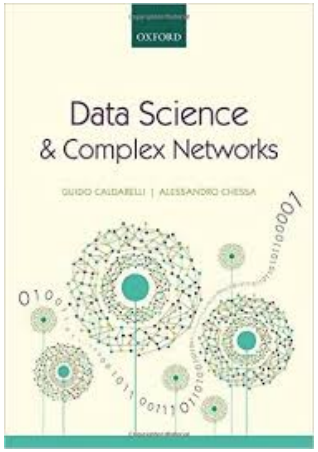
- "Scale-free": the degree distribution follows a power law (Paretos curve)
- "Small-world":
  - Small graph diameter, short average distance between nodes
  - High clustering coefficient: many triangles, hubs, ...
- Hierarchical structure
- Rich in "motifs"
- Self-similar (as in fractals)

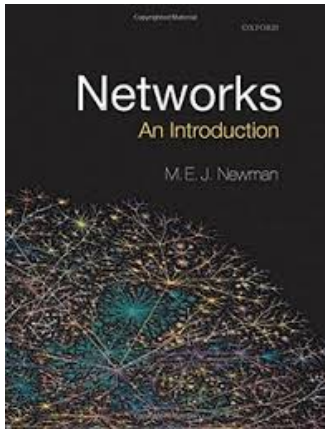
Briefly stated: complex networks exhibit a non-trivial topology.

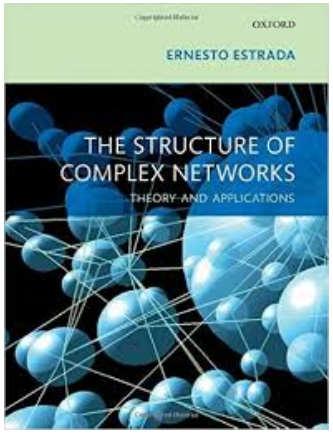
Caveat

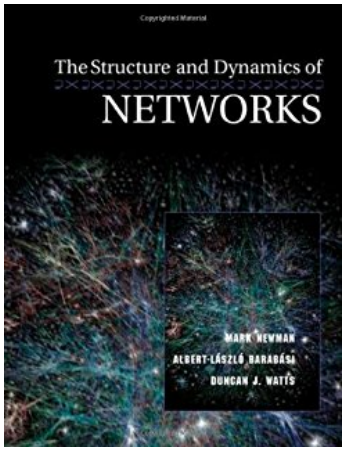
there are important examples of real-world complex networks lacking one or more of these attributes.

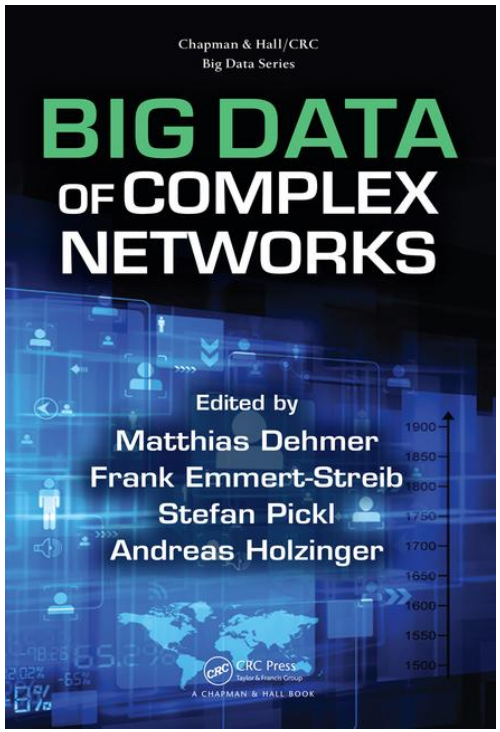
Materials


OXFORD
Data Science
& Complex Networks

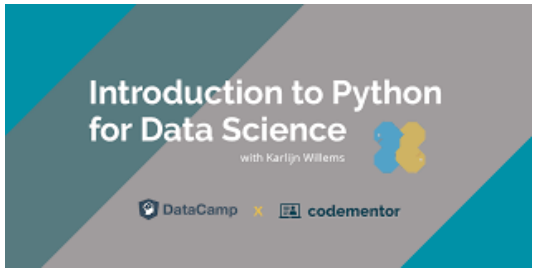

Networks
An Introduction
M. E. J. Newman

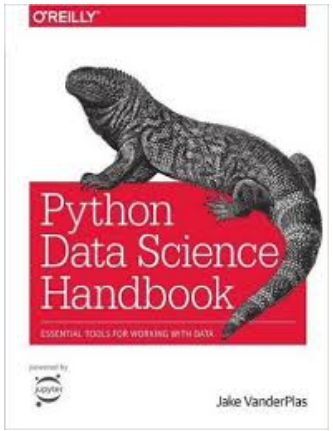

OXFORD
ERNESTO ESTRADA
THE STRUCTURE OF
COMPLEX NETWORKS
THEORY AND APPLICATIONS

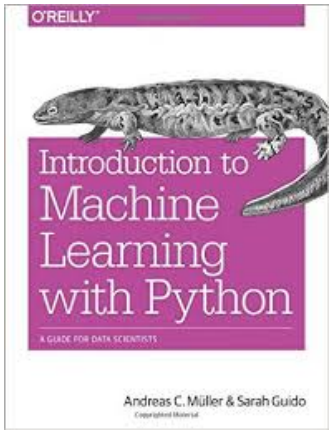

The Structure and Dynamics of
NETWORKS
MARK NEWMAN
ALBERT-LÁSZLÓ BARABÁSI
DUNCAN J. WATTS

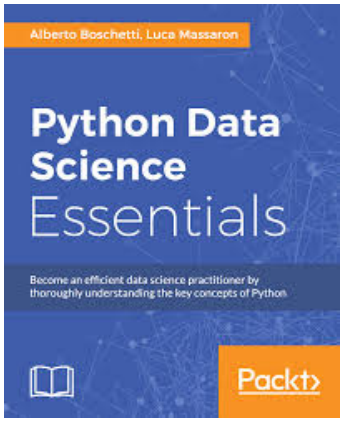

Chapman & Hall/CRC
Big Data Series
BIG DATA
OF COMPLEX
NETWORKS
Edited by
Matthias Dehmer
Frank Emmert-Streib
Stefan Pickl
Andreas Holzinger
CRC Press
A CHAPMAN & HALL BOOK

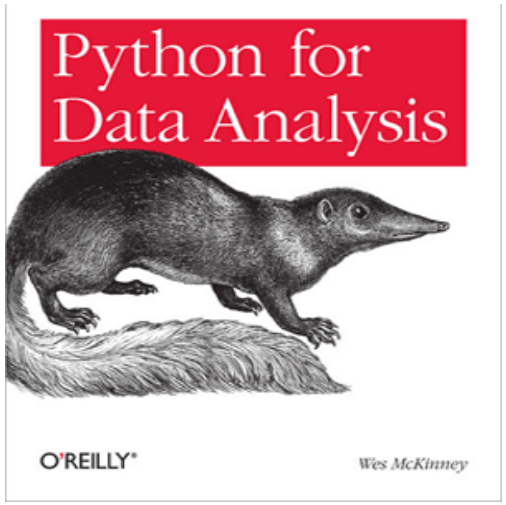
Materials


Introduction to Python for Data Science
with Karlijn Willems
DataCamp x codementor


O'REILLY
Python Data Science Handbook
ESSENTIAL TOOLS FOR WORKING WITH DATA
Jake VanderPlas

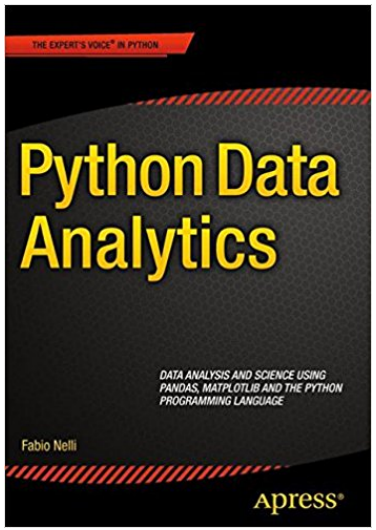

O'REILLY
Introduction to Machine Learning with Python
A GUIDE FOR DATA SCIENTISTS
Andreas C. Müller & Sarah Guido

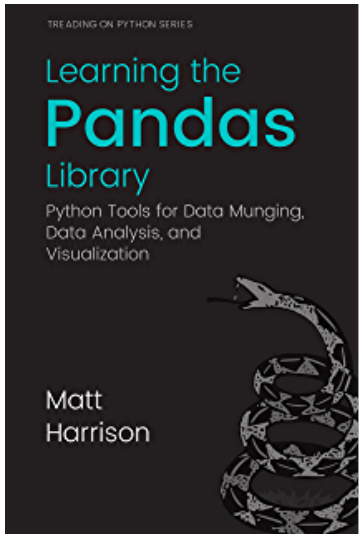

Alberto Boschetti, Luca Massaron
Python Data Science Essentials
Become an efficient data science practitioner by thoroughly understanding the key concepts of Python
Packt

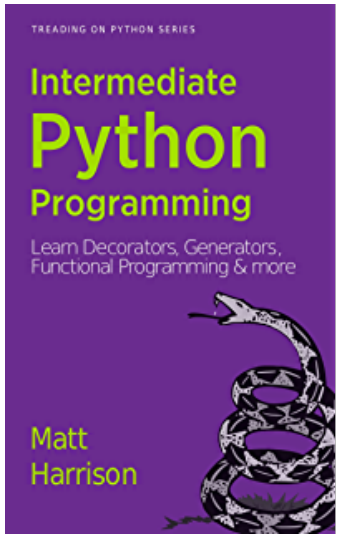

Python for Data Analysis
O'REILLY
Wes McKinney


pandas:
POWERFUL
PYTHON
DATA
ANALYSIS
TOOLKIT

Materials


THE EXPERT'S VOICE® IN PYTHON
Python Data Analytics
DATA ANALYSIS AND SCIENCE USING PANDAS, MATPLOTLIB AND THE PYTHON PROGRAMMING LANGUAGE
Fabio Nelli
Apress®


TREADING ON PYTHON SERIES
Learning the
Pandas
Library
Python Tools for Data Munging, Data Analysis, and Visualization
Matt
Harrison


TREADING ON PYTHON SERIES
Intermediate
Python
Programming
Learn Decorators, Generators, Functional Programming & more
Matt
Harrison

Supervised essays

http://archive.aims.ac.za/structured-masters-research-projects

1. Fidele Tubananza (from Rwanda), AIMS-Senegal, 2014
   Network approach of epidemic spreading (stochastic model)
2. Reham Bashir (from Soudan), AIMS-Tanzania,2015
   Centrality in complex networks and application
3. Alice Nayanzi (from Ouganda), AIMS-South Africa, 2016
   Laplacian matrix of a networks and applications, (Research Center)
4. Laeticia Shoma (from DRC), AIMS-South Africa, 2016
   Network spectra and applications.
5. Emily S. Muller (from South Africa), AIMS-South Africa, 2017
   Dynamic Model for social network: The stochastic actor-oriented approach.
6. Boipelo Sihle Ncube (from South Africa), AIMS-South Africa, 2017
   Communicability in Complex networks and applications
7. Mulalo Netshithuthuni (from South Africa), joint supervision with Dr. Bah AIMS-South Africa, 2017
   Prediction of credit card default using Machine learning

Supervised Research Master jointly with Dr. Simukai Utete

- Alice Nanyanzi, AIMS South Africa Research Center
  Extended diffusion process in complex networks

Softwares

1. Networkx (Pyhton) (https://networkx.github.io/)
   1. Cross-platform python library
   2. Examples/introduction in this presentation
2. NodeXL (http://nodexl.codeplex.com/):
   1. Add-in for Microsoft Excel (Windows only)
   2. Good data collection options (Twitter, Facebook, YouTube, . . .)
   3. Basic visualization
3. igraph (Python, C, R) (http://igraph.org/redirect.html)
   1. install.packages(igraph) in R (r-project.org) (or python)
   2. Cross-platform
   3. Text driven: powerful for analysis
4. pajet Program for Large Network Analysis http://vlado.fmf.uni-lj.si/pub/networks/pajek/
   1. Standalone, Windows (or Linux with wine)
   2. Good interactive environment for metrics and basic visualization
5. Gephi https://gephi.org/

Source of network data

Precollated data

1. Stanford Large Network Dataset Collection, https://snap.stanford.edu/data/.
2. Network Repository, http://networkrepository.com/.
3. Pajek datasets, http://vlado.fmf.uni-lj.si/pub/networks/data/
4. http://www.diggingintodata.org/Repositories/tabid/167/
5. http://www-personal.umich.edu/~mejn/netdata
6. http://networkdata.ics.uci.edu/index.php

Collecting your own

1. Basic point-and-click options in NodeXL (more) and Gephi
2. Twitter data with a small amount of programming https://github.com/computermacgyver/twitter-python/
3. A good general resource: Russell (2011), Mining the Social Web, OReilly Media.

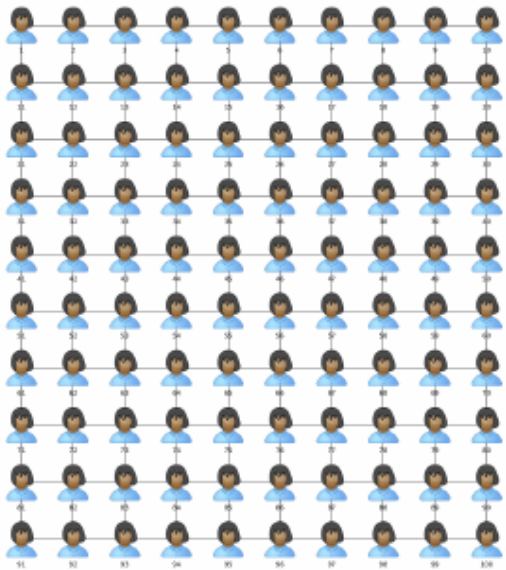
Complex networks: Definition

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

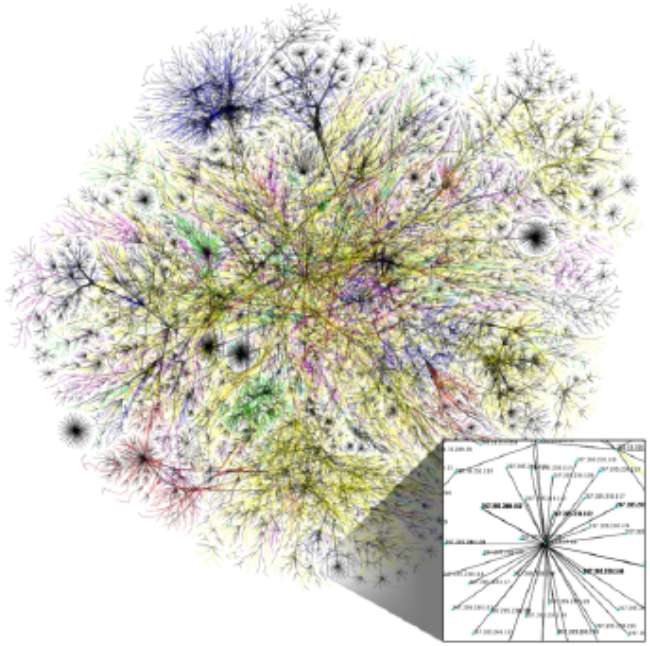
A network consists of a graph and additional information on the vertices or the lines of the graph.

Complex Networks

what exactly is complex network?



(g) grid lattice



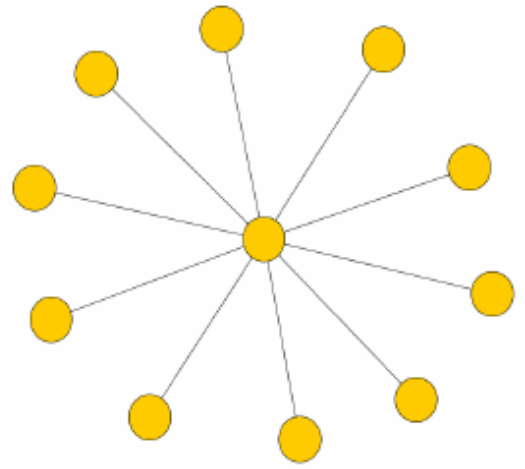
(h) internet

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

Complex Networks

what exactly is complex network?

(i) Transportation network

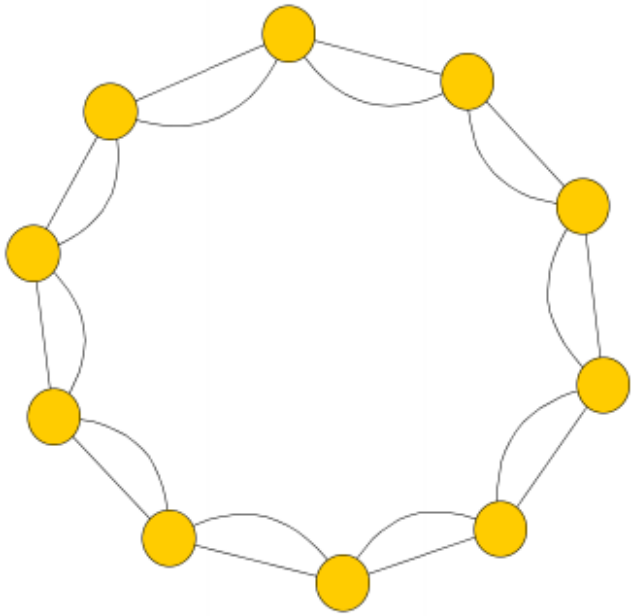


(j) Star graph

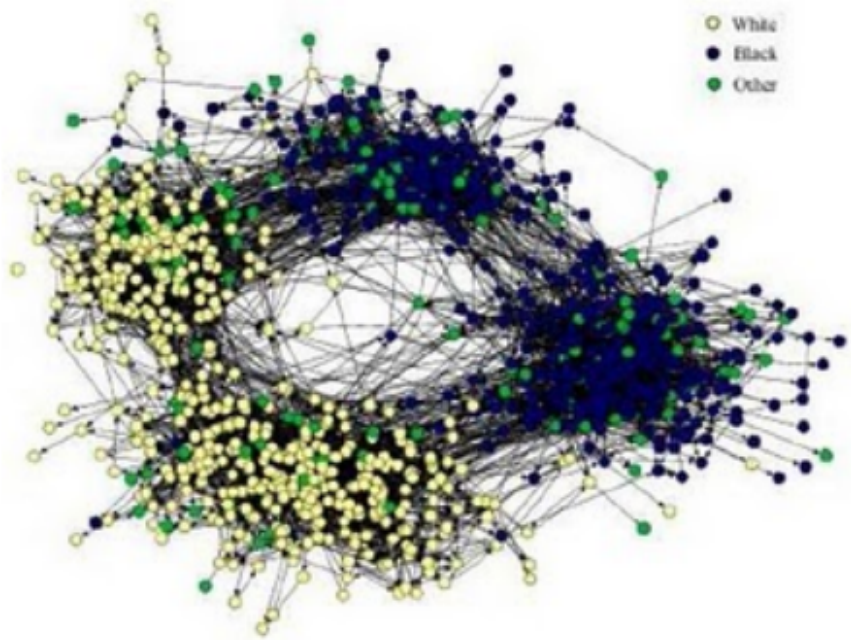
It is easy to tell which graphs are not complex networks.

Complex Networks

what exactly is complex network?



(k) Ring lattice



(l) Social network in school

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

Basic references

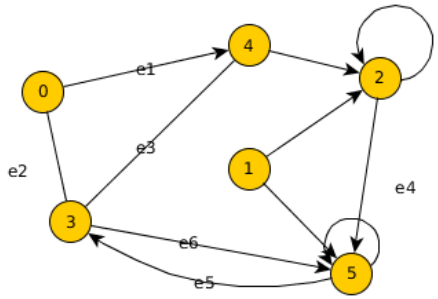
Some classic early references:

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.

Formal definitions

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

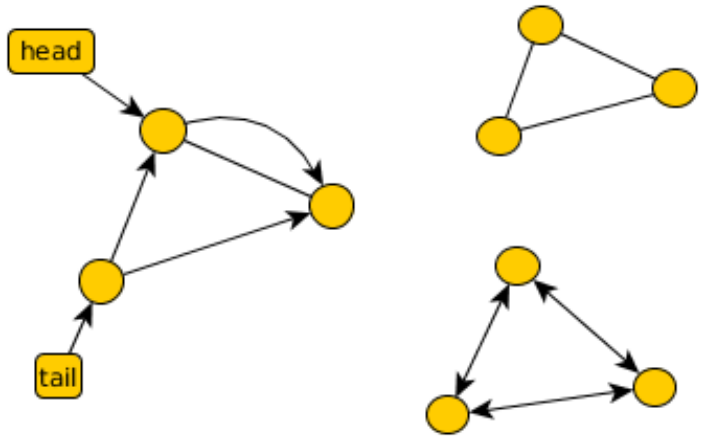
- A graph is a set of vertices and a set of lines between pairs of vertices.
- A graph represents the structure of network. It need a set of vertices and a set of lines.
- A vertex is the smallest unit of a network and a line is tie between two vertices in the network.
- A loop is a special kind of line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

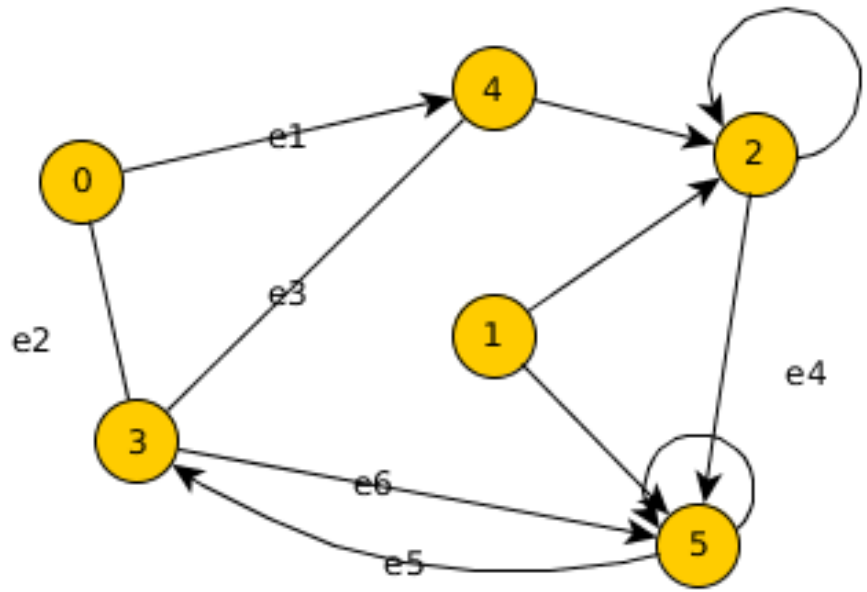
1. To every unweighted graph $G(V, E)$ we associated is adjacency matrix A defined as follow

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

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).
5. The diagonal entries of A are zeroes and if the network is undirected, A is a symmetric matrix.

Adjacency matrix

$$A = \begin{pmatrix} & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \\ & & & & & \end{pmatrix},$$



For multigraphs and graphs with loops, the entries are the number of edges between each pair of vertices and the diagonal entries are non-zero due to self-loops which may be counted once or twice based on whether the network is directed or undirected.

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 $\mid V \mid \times 1$ 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

$$\mathbf{k}^{out} = \mathbf{A}\mathbf{1}$$

or, for each component,

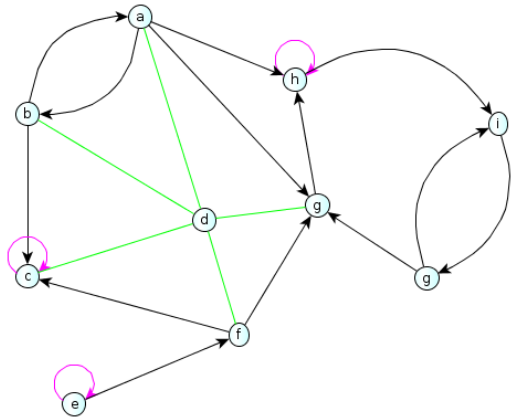
$$k_i^{out} = \sum_j a_{ij}.$$

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

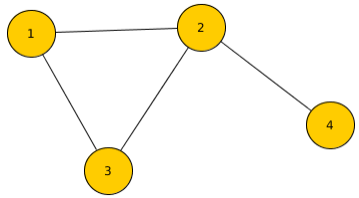


○

The average node degree in a graph is defined by

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

Degree



Nodes degrees

$$k_1 = 2\ , k_2 = 2\ , k_3 = 2\ , k_4 = 1.$$

Handshaking lemma

For any given undirected network $G = (V, E)$, where V is the set of nodes and E the set of edges, the sum of all vertex degrees is equal to twice the number of edges.

$$\sum_{v \in V} k_v = 2\ |E|. \tag{2}$$

Therefore, the total number of links, L, in an undirected network can be expressed in term of the sum of the node degrees:

$$L = |E| = \frac{1}{2} \sum_{v=1}^{N} k_v \tag{3}$$

Descriptive statistic

For a sample having n units, $x_1, \cdots, x_N$ we have these four parameter that characterise it:

$$\text{Average (mean)} \quad \langle k \rangle = \frac{x_1 + \cdots + x_n}{N} = \frac{1}{N} \sum_i^N x_i \tag{4}$$

$$n^{th} \text{ moment} \quad \langle k^n \rangle = \frac{x_1^n + \cdots + x_n^n}{N} = \frac{1}{N} \sum_i^N x_i^n \tag{5}$$

$$\text{standard deviation} \quad \sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)^2} \tag{6}$$

$$\text{distribution de } x \quad p_x = \frac{1}{N} \sum_i \delta_{x,x_i} \tag{7}$$

$$\sum p_x = 1 \quad \left(\int p_x dx = 1 \right) \tag{8}$$

Degree distribution

.

The degree distribution of a network is obtained in terms of the probability p_k and is defines as the probability that a node chosen uniformly at random has degree k or equivalently as the fraction of nodes in the graph having degree k.

$$\sum_{k=1}^{\infty} p_k = 1 \tag{9}$$

For a network with N nodes the degree distribution is the normalized histogram is given by

$$p_k = \frac{N_k}{N} \tag{10}$$

where N_k is the number of nodes having degree. Hence the number of nodes having degree k can be obtained from the degree distribution as

$$N_k = Np_k. \tag{11}$$

Degree distribution

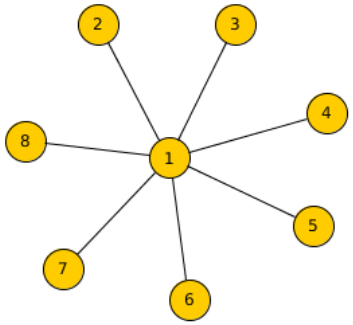
The degree distribution is the most fundamental topological characterisation of a network. It assumed a central role in network theory following the discovery of scale-free networks.

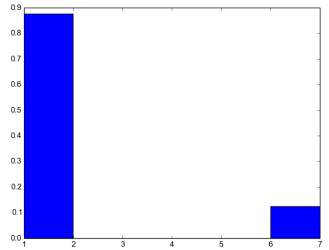
One reason is that the calculation of most network properties requires us to know pk. For example, the average degree of a network can be written as

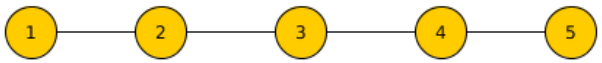
$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k \tag{12}$$

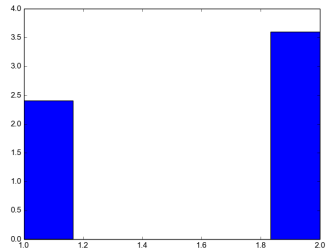
The other reason is that the precise functional form of pk determines many network phenomena, from network robustness to the spread of viruses.

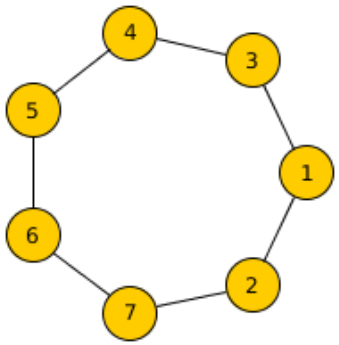
Degree distribution

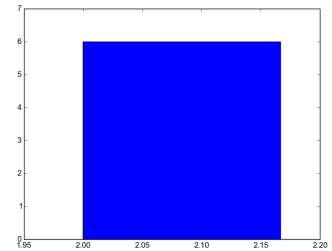












Sparsity of real networks

In real networks the number of nodes (N) and links (L) can vary widely. For example:

1. The neural network of the worm C. elegans, has $N = 302$ neurons (nodes). (this the only fully mapped nervous system of a living organism).
2. The human brain is estimated to have about a hundred billion ($N \sim 10^{11}$) neurons.
3. The genetic network of a human cell has about $20,000$ genes as nodes;
4. WWW is estimated to have over a trillion web documents ($N > 10^{12}$).

The number of links in a network can also varies widely, between $L = 0$ (null graph, without edges/links) and

$$L_{max} = \frac{N(N-1)}{2} \tag{13}$$

in a complete graph having N nodes.

Sparsity of real networks

- The number of link L in real networks is much smaller than L_{max}, reflecting the fact that most real networks are sparse.
- A network is sparse if $L \ll L_{max}$. For example, This is true for all of the networks in described earlier. One can check that their number of links is only a tiny fraction of the expected number of links for a complete graph of the same number of nodes.

WWW graph

The WWW graph has about 1.5 million links. Yet, if the WWW were to be a complete graph, it should have $L_{max} \approx 5 \times 10^{10}$ links according. Consequently the web graph has only a 3×10^{-5} fraction of the links it could have.

Sparsity of real networks

1. The sparsity of real networks implies that the adjacency matrices are also sparse.
2. Indeed, a complete network has $A_{ij} = 1$, for all (i, j), i.e. each of its matrix elements are equal to one.
3. In contrast in real networks only a tiny fraction of the matrix elements are nonzero.

show the adjacency matrix of C. elangs network and for a complete network having the same number of nodes.

Sparsity of real networks

How we store very large network

Sparseness has important consequences on the way we explore and store real networks. For example, when we store a large network in our computer, it is better to store only the list of links (i.e. elements for which $A_{ij} \neq 0$), rather than the full adjacency matrix, as an overwhelming fraction of the A_{ij} elements are zero. Hence the matrix representation will block a huge chunk of memory, filled mainly with zero.

This can be applied to facebook network for example.

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).

clustering

- 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.

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

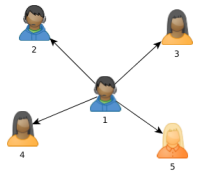
- Scale-free, meaning that their degree distribution tends to follow a power law: $p(k) =$ number of nodes of degree $k \approx ck^{-\gamma}$, $\gamma > 0$. Frequently, $2 \leq \gamma \leq 3$. This implies sparsity but also the existence of several highly connected nodes (hubs).
- Small-world, meaning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

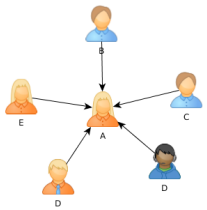
$$diam(G) = \mathcal{O}(logN), \quad N \to \infty. \tag{14}$$

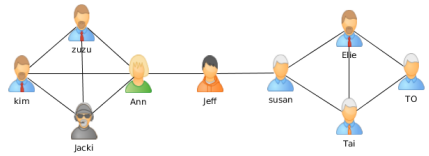
- Highly clustered, i.e., they contain a very large proportion of triangles (unlike random graphs).

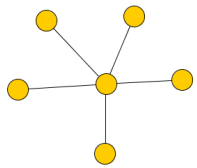
Centrality measures

- 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







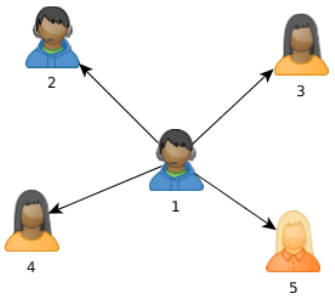


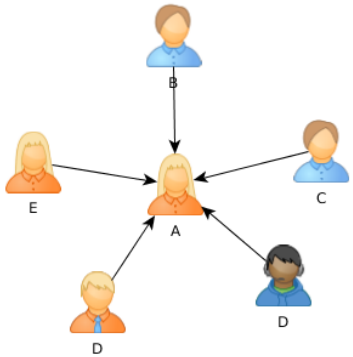
Degree Centrality (force/power through links)

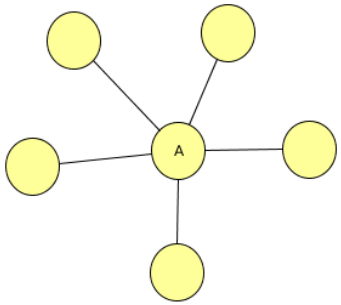
$$C_D(i) = \sum_j A_{i,j} = k(i)$$

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

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

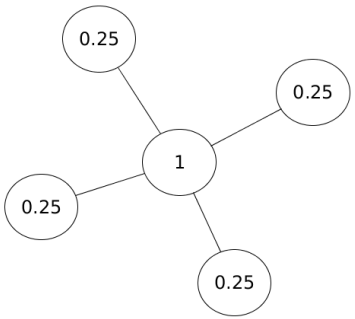


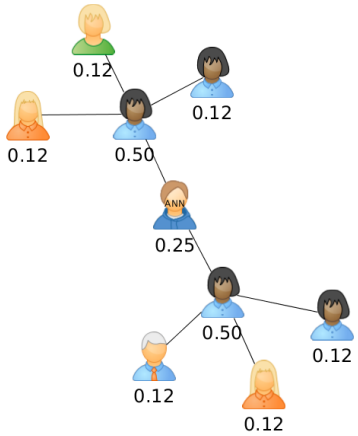




Degree Centrality (force/power through links)

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)

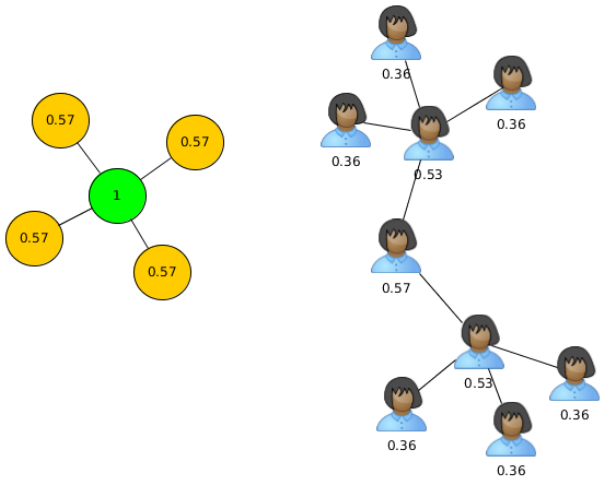




The degree is very cheap to compute but is unable to recognize the centrality of certain nodes: its a purely local notion.

Closeness Centrality
power through proximity to others

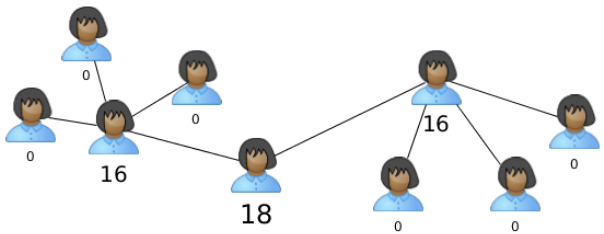
$$C_c(i) = \left(\frac{\sum_{i \neq i} d(i,j)}{n-1} \right)^{-1} = \frac{n-1}{\sum_{i \neq i} d(i,j)}$$



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

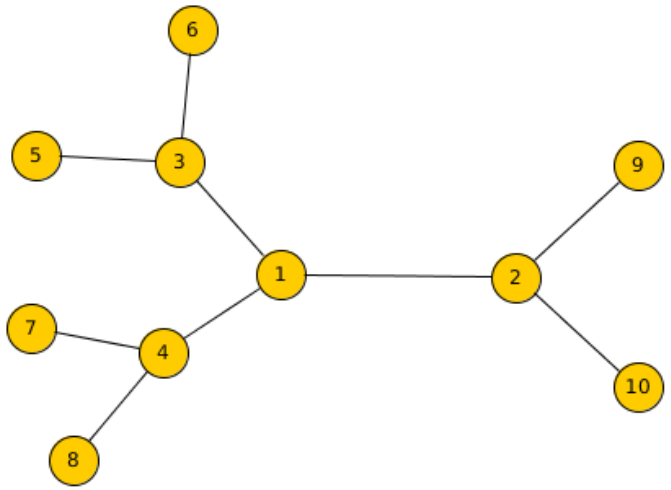
$$NB_C(i) = \frac{2B_c(i)}{(n-1)(n-2)}$$

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,

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

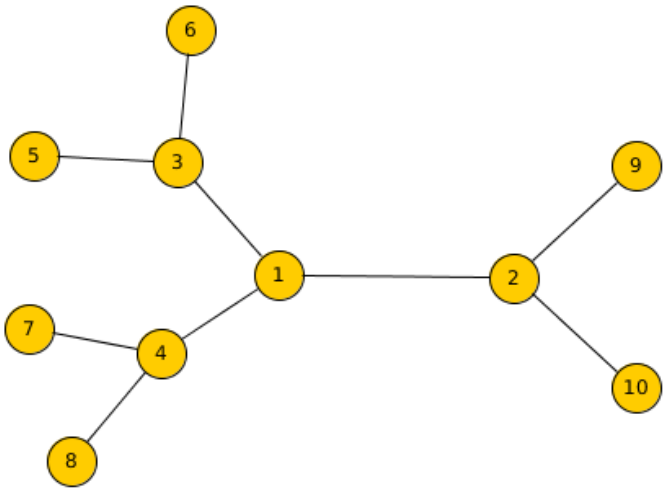
Eigenvector centrality is proportional to the leading eigenvector of A (and hence, the name!).

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:

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

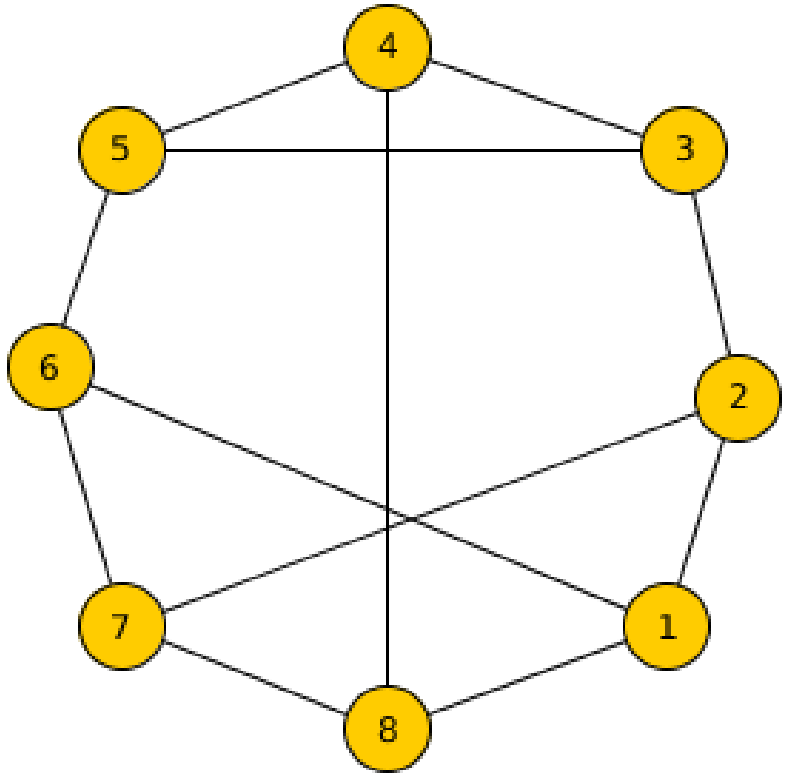
Centrality: Subgraph
Estrada-Rodrìguez-Velsquez, Phys. Rev. E, 2005

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.

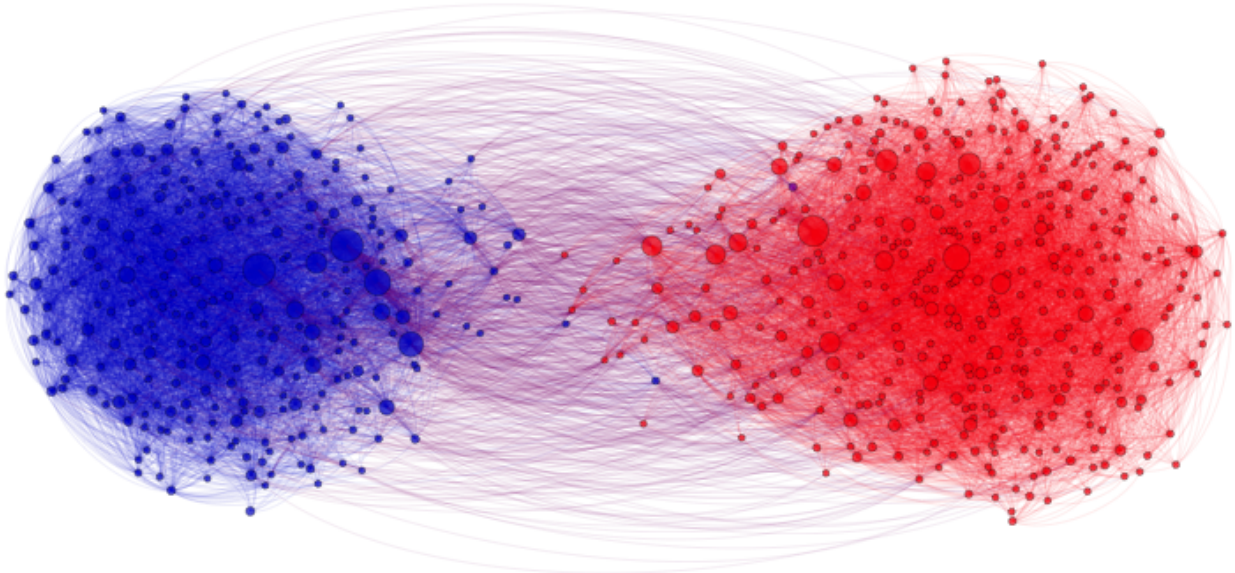
i	DC	CC	BC	EVC	SC
1	0.43	0.63	0.07	0.35	3.71
2	0.43	0.63	0.1	0.35	3.64
3	0.43	0.63	0.1	0.35	3.90
4	0.43	0.63	0.1	0.35	3.90
5	0.43	0.63	0.1	0.35	3.90
6	0.43	0.63	0.1	0.35	3.64
7	0.43	0.63	0.07	0.35	3.71
8	0.43	0.63	0.1	0.35	3.64



Community Finding

Community Dectection

1. We will often be interested in identifying communities of nodes in a complex networks..



2. Example: Two distinct communities based on political blogs (Democratic vs. Republican) in US.

Community

In network science we call a community a group of nodes that have a higher likelihood of connecting to each other than to nodes from other communities.

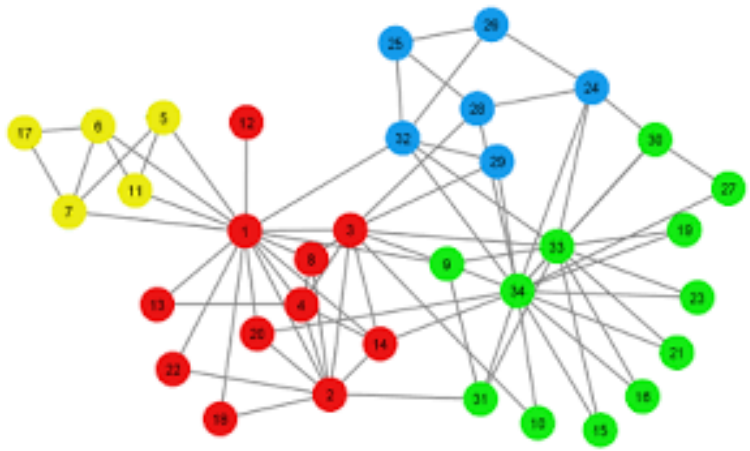
Community in social networks

1. Employees of a company are more likely to interact with their coworkers than with employees of other companies.
2. Work places appear as **densely** interconnected communities within the social network.

3. Communities could also represent circles of friends, or a group of individuals who pursue the same hobby together, or individuals living in the same neighborhood.

Community in social networks: Zachary Karate Club

1. Zacharys Karate Club has received particular attention in the context of community detection.
2. Capturing the links between 34 members of a karate club. Given the club's small size, each club member knew everyone else.
3. To uncover the true relationships between club members, sociologist Wayne Zachary documented 78 pairwise links between members who regularly interacted outside the club

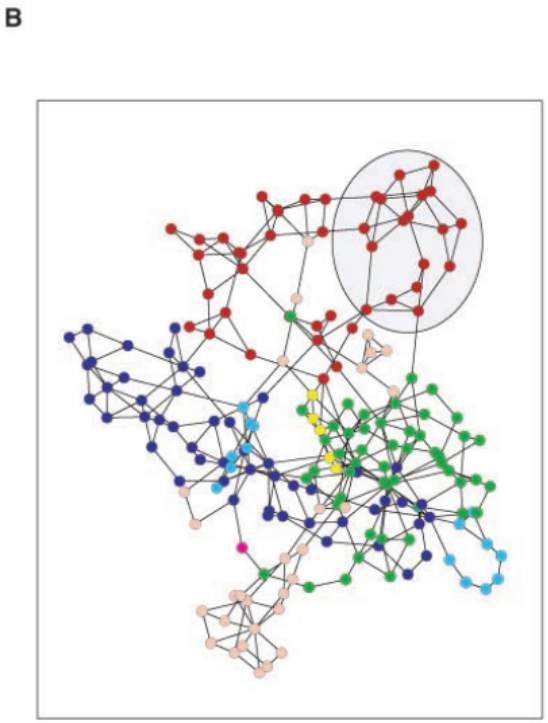


Community in biological networks

From molecular to modular cell biology. [L.H. Hartwell et al. Nature 1999]

1. Communities play a particularly important role in our understanding of how specific biological functions are encoded in cellular networks.
2. Lee Hartwell argued that biology must move beyond its focus on single genes. [The Nobel Prize in Medicine.]
3. Biology must explore instead how groups of molecules form functional modules to carry out a specific cellular functions

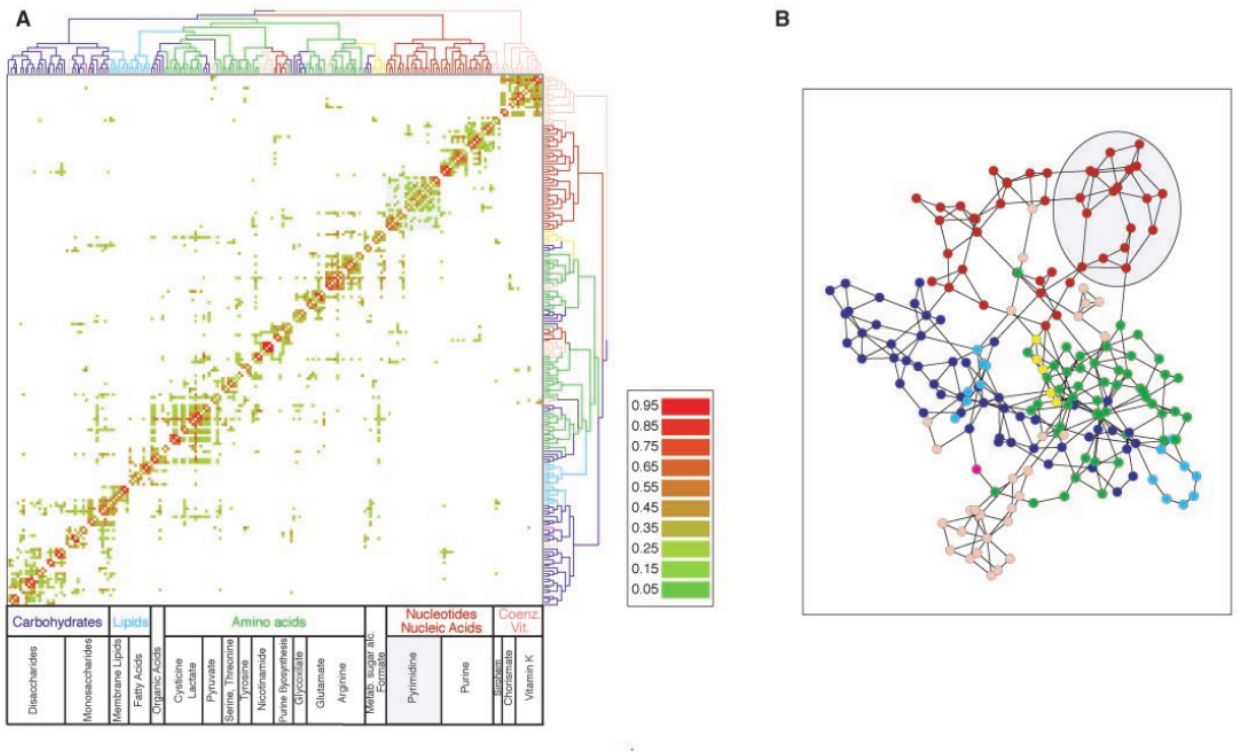
The biological modules (communities) identified by the Ravasz algorithm [E. Ravasz et al. 2002].



- The biological modules (communities) identified by the Ravasz algorithm.
- The color of each node, capturing the predominant biochemical class to which it belongs, indicates that different functional classes are segregated in distinct network neighborhoods.
- The highlighted region selects the nodes that belong to the pyrimidine metabolism, one of the predicted communities.

Community : metabolic networks

Ravasz and al. made the first identification of such modules in metabolic networks (for E. coli metabolism) by building an algorithm to identify groups of molecules that form locally dense communities.



[E. Ravasz et al. 2002].

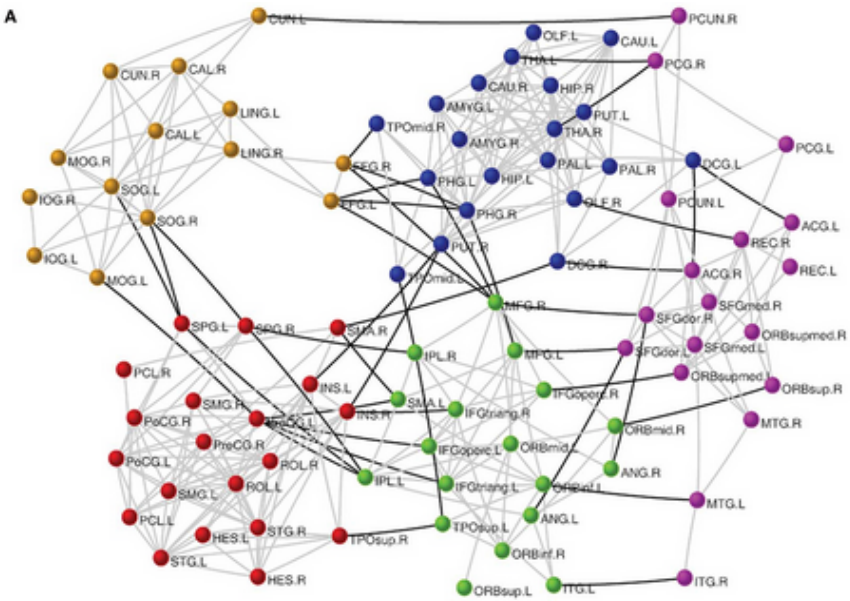
Community : metabolic networks

1. Communities play a particularly important role in understanding human diseases.
2. Indeed, proteins that are involved in the same disease tend to interact with each other.
3. This finding inspired the disease module hypothesis, stating that each disease can be linked to a well-defined neighborhood of the cellular network.

ref

- K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabsi. The human disease network. PNAS, 104:8685-8690, 2007.
- Menche, A.Sharma, M. Kitsak, S. Ghiassian, M. Vidal, J. Loscalzo, A.-L. Barabsi. Oncovering disease-disease relationships through the human interactome. 2014.
- A.-L. Barabsi, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. Nature Review Genetics, 12:56-68, 2011.

Community: brain network, The modular architecture of resting-state functional brain network [He et al., 2009b]



Five modules in a functional network of the human brain, represented by five different colors. intra-module and inter-module connections are shown in gray and dark lines.

Surface representation of modular architecture of a functional brain network. All 90 brain regions are marked by using different colored spheres (different colors represent distinct network modules).

Community detection

1. Discussed above examples illustrate diverse motivations that drive community identification.
2. Existence of communities is rooted in who connects to whom, hence they cannot be explained based on the degree distribution alone.
3. To extract communities we must therefore inspect a networks detailed wiring diagram.

H1: Fundamental Hypothesis

A networks community structure is uniquely encoded in its wiring diagram.

There is a ground truth about a networks community organization, that can be uncovered by inspecting A_{ij}.

H2: Connectedness and Density Hypothesis

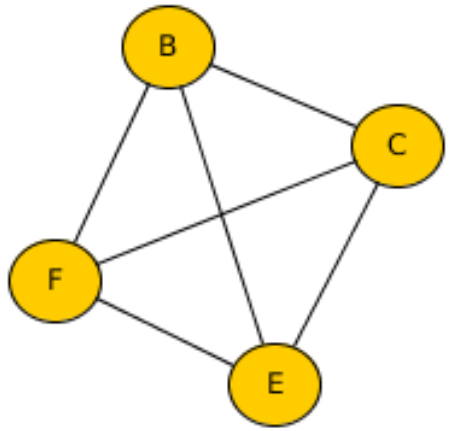
A community is a locally dense connected subgraph in a network.

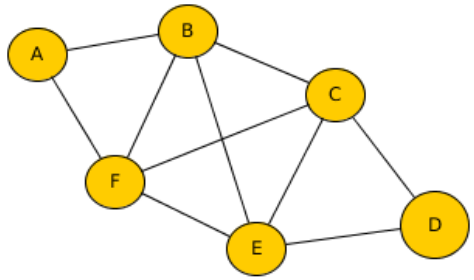
1. All members of a community must be reached through other members of the same community (connectedness).
2. At the same time we expect that nodes that belong to a community have a higher probability to link to the other members of that community than to nodes that do not belong to the same community (density).
3. This hypothesis narrows what would be considered a community, it does not uniquely define it.
4. Several community definitions are consistent with H2.

Community detection: Cliques

Following H2, defined a community as group of individuals whose members all know each other [R.D. Luce and A.D. Perry]. In graph theoretic terms this means that a community is a complete subgraph, or a clique.

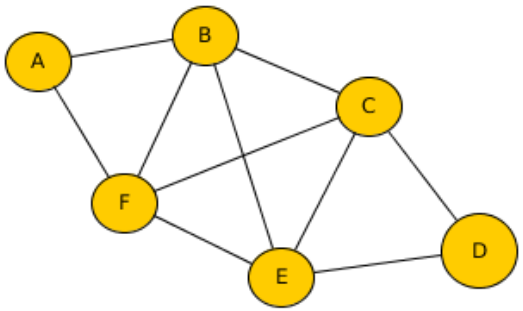
1. A clique is a social grouping where everyone knows everyone else (i.e. there is an edge between each pair of nodes).
2. A maximal clique is a clique that is not subset of any other clique in the graph.
3. A clique with size greater than or equal to that of every other clique in the graph is called a maximum clique.

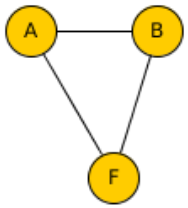


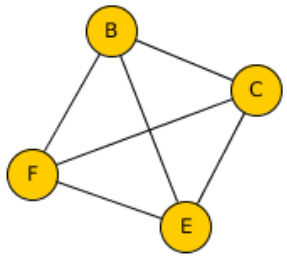


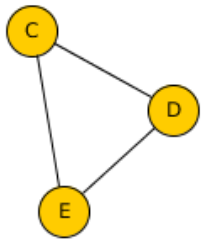
Cliques: Find all maximal cliques in the second graph

```
>>> import networkx as nx
>>> L = [b', 'b', 'e', 'f']
>>> g = nx.complete_graph(len(L))
>>> nx.relabel_nodes(G,dict(enumerate(L)), copy = False)
>>> g.add_edges_from([('a','b'),('a','f'),('c','d'),('c','f')]
>>> cl = list(nx.networkx.find_cliques(g))
>>> print cl
[['a', 'b', 'f'], ['b', 'b', 'e', 'f'], ['c', 'd', 'e']]
```



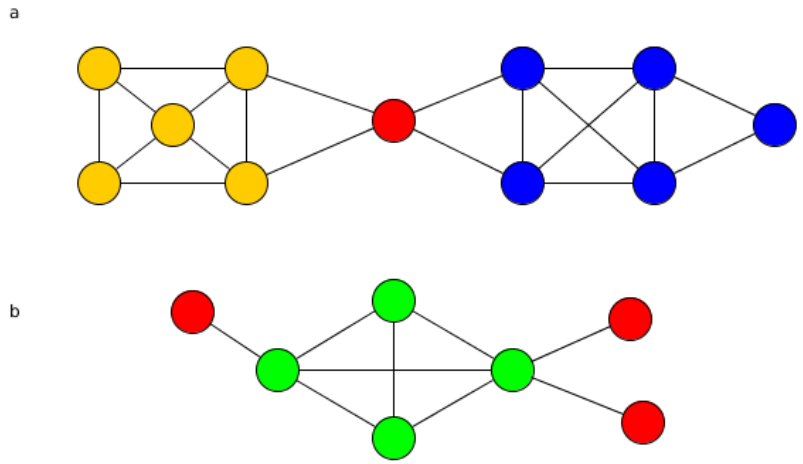






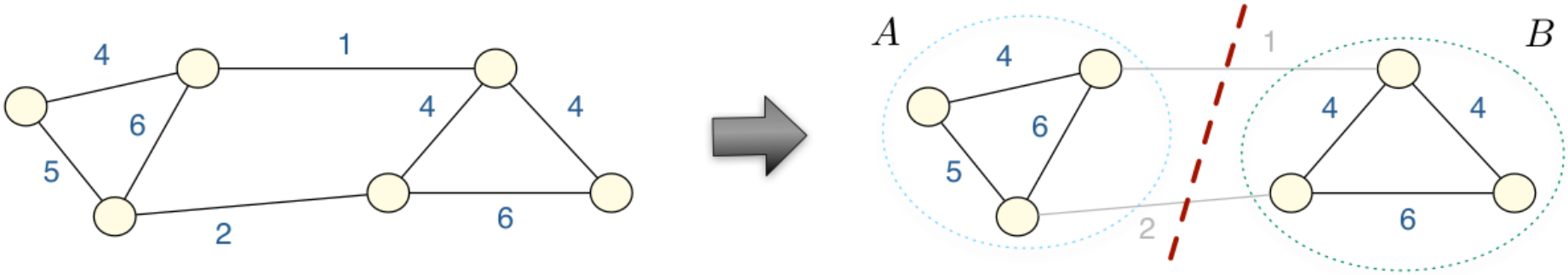
Cliques: Find all maximal cliques in the second graph

Try it yourself...



Graph Partionning

- **Goal:** Divide the nodes in a graph into a user-specified number of disjoint groups to optimise a criterion related to number of edges cut.



- Min-cut simply involves minimising number (or weight) of edges cut by the partition.

$$cut(A, B) = 3$$

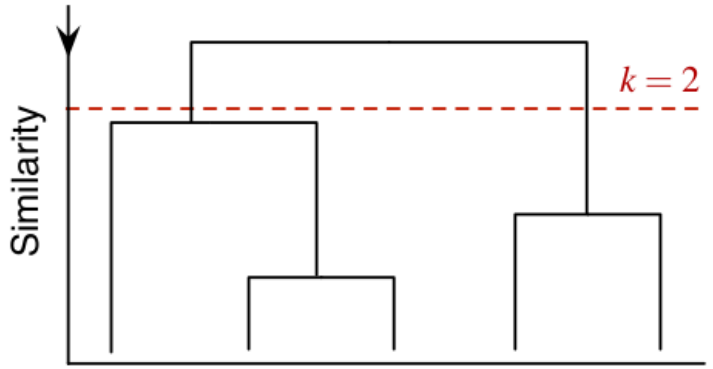
- Recent approaches use more sophisticated criteria (e.g. normalised cuts) and apply multi-level strategies to scale to large graphs.

Hierarchical Clustering

- Construct a tree of clusters to identify groups of nodes with high similarity according to some similarity measure.
- Two basic families of algorithm...
  1. Agglomerative: Begin with each node assigned to a singleton cluster. Apply a bottom-up strategy, merging the most similar pair of clusters at each level.
  2. Divisive: Begin with single cluster containing all nodes. Apply a top-down strategy, splitting a chosen cluster into two sub-clusters at each level.

Issues for Community Detection:

- How do we choose among many different possible clusterings?
- Is there really a hierarchical structure in the graph?
- Often scales poorly to large graphs.



Hierarchical Clustering

- We can apply agglomerative clustering to a NetworkX graph by calling functions from the NumPy and SciPy numerical computing packages.

```
import networkx
import numpy, matplotlib
from scipy.cluster import hierarchy
from scipy.spatial import distance
```


```
g = networkx.read_edgelist("karate.edgelist")
```

```
path_length=networkx.all_pairs_shortest_path_length(g)
n = len(g.nodes())
distances=numpy.zeros((n,n))
for u,p in path_length.iteritems():
   for v,d in p.iteritems():
      distances[int(u)-1][int(v)-1] = d
sd = distance.squareform(distances)
```

Build pairwise distance matrix based on shortest paths between nodes.

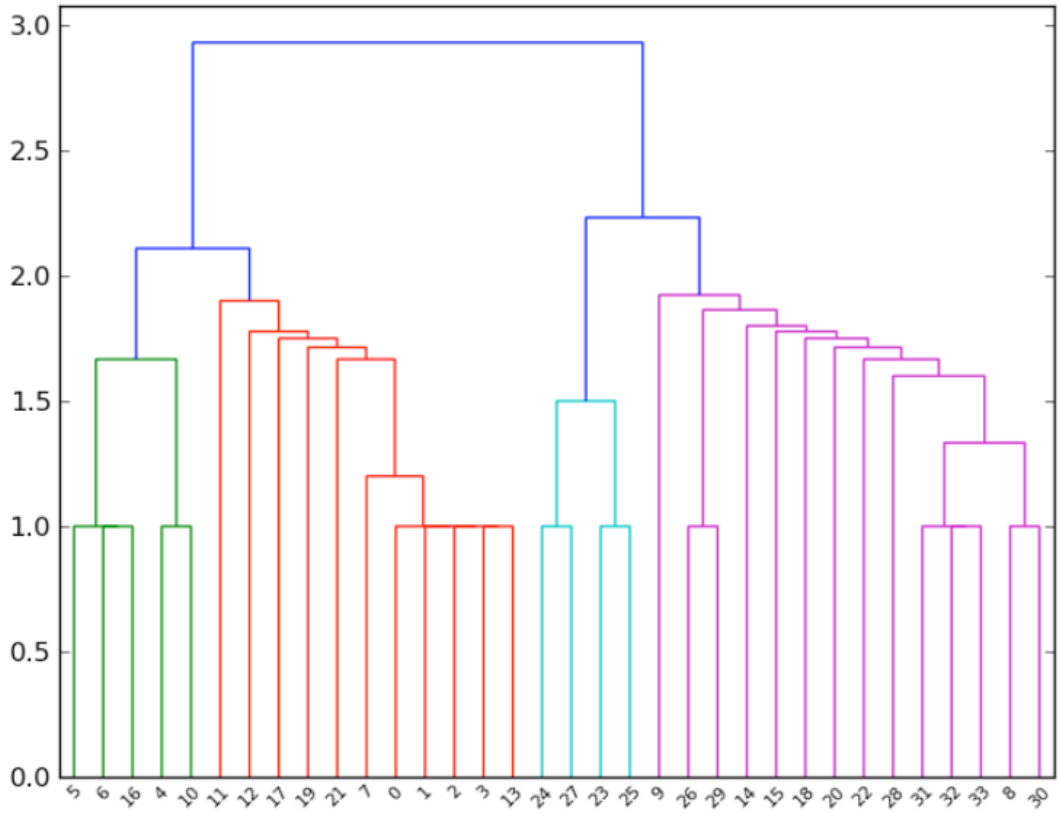
```
hier = hierarchy.average(sd)
```

Apply average-linkage agglomerative clustering.

Hierarchical Clustering

```
hierarchy.dendrogram(hier)
matplotlib.pylab.savefig("tree.png",format="png")
```

Build the dendrogram, then write image to disk.



Modularity Optimization

- Newman & Girvan [2004] proposed measure of partition quality....
  - ➡ Random graph shouldn't have community structure.
  - ➡ Validate existence of communities by comparing actual edge density with expected edge density in random graph.

$$Q = (\text{number of edges within communities}) - (\text{expected number within communities})$$

- Apply agglomerative technique to iteratively merge groups of nodes to form larger communities such that modularity increases after merging.
- Recently efficient greedy approaches to modularity maximisation have been developed that scale to graphs with up to 10^9 edges.

Issues for Community Detection:

- Total number of edges in graph controls the resolution at which communities are identified [Fortunato, 2010].
- Is it realistic/useful to assign nodes to only a single community?

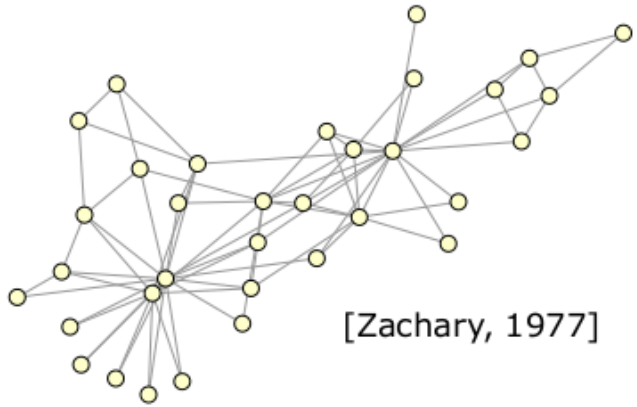
Networkx: Modularity - Optimization

Python Implementation for the Louvain algorithm are available
http://perso.crans.org/aynaud/communities/community.py

```
g = networkx.read_edgelist("karate.edges")
```

Apply Louvain algorithm to the graph

```
import community
partition = community.best_partition( g )
```



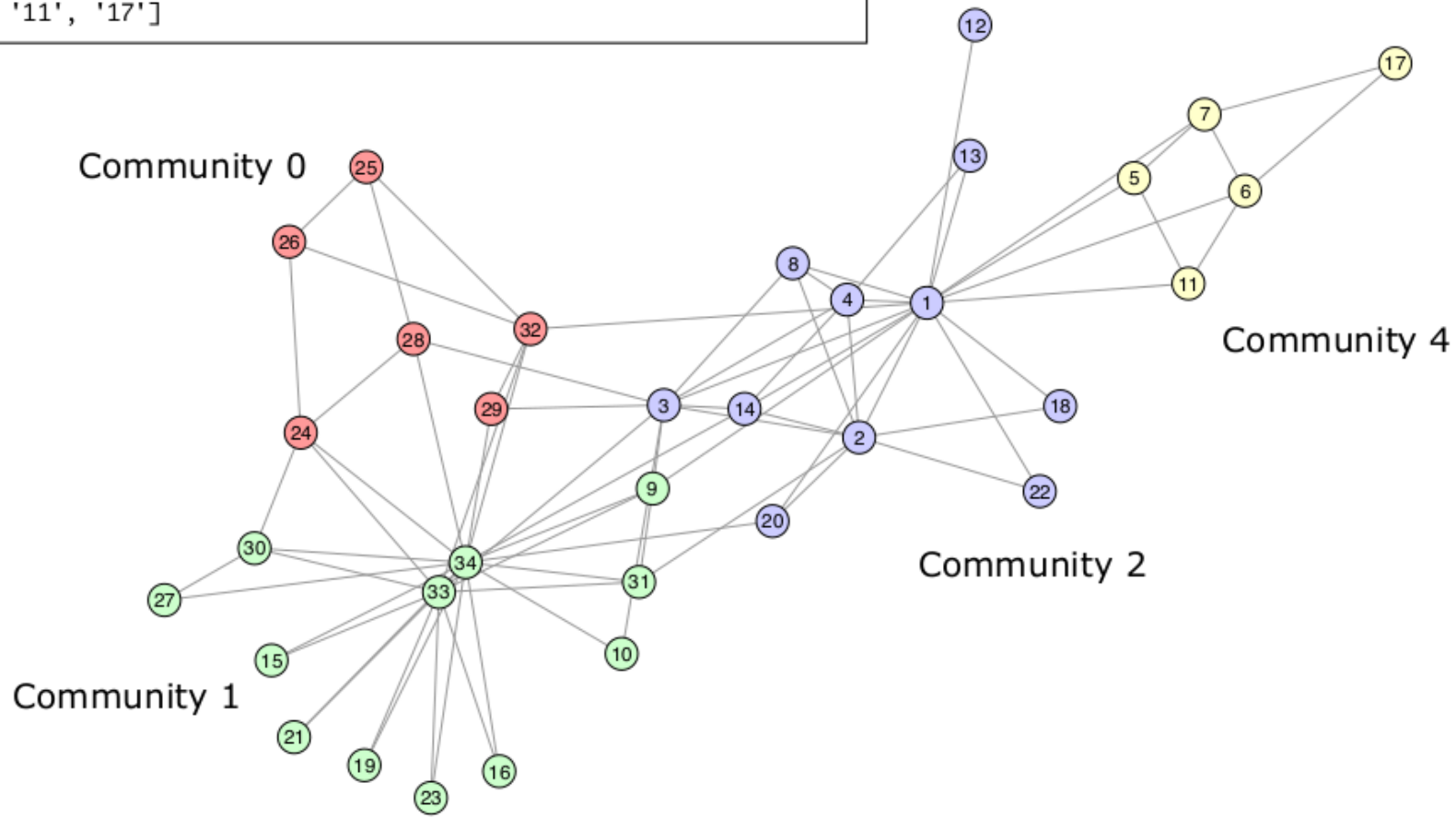
Print nodes assigned to each community in the partition

```
for i in set(partition.values()):
   print "Community", i
   members = list_nodes = [nodes for nodes in partition.keys() if partition[nodes] == i]
   print members
```

```
Community 0
['24', '25', '26', '28', '29', '32']
Community 1
['27', '21', '23', '9', '10', '15', '16', '33', '31', '30', '34', '19']
Community 2
['20', '22', '1', '3', '2', '4', '8', '13', '12', '14', '18']
Community 3
['5', '7', '6', '11', '17']
```

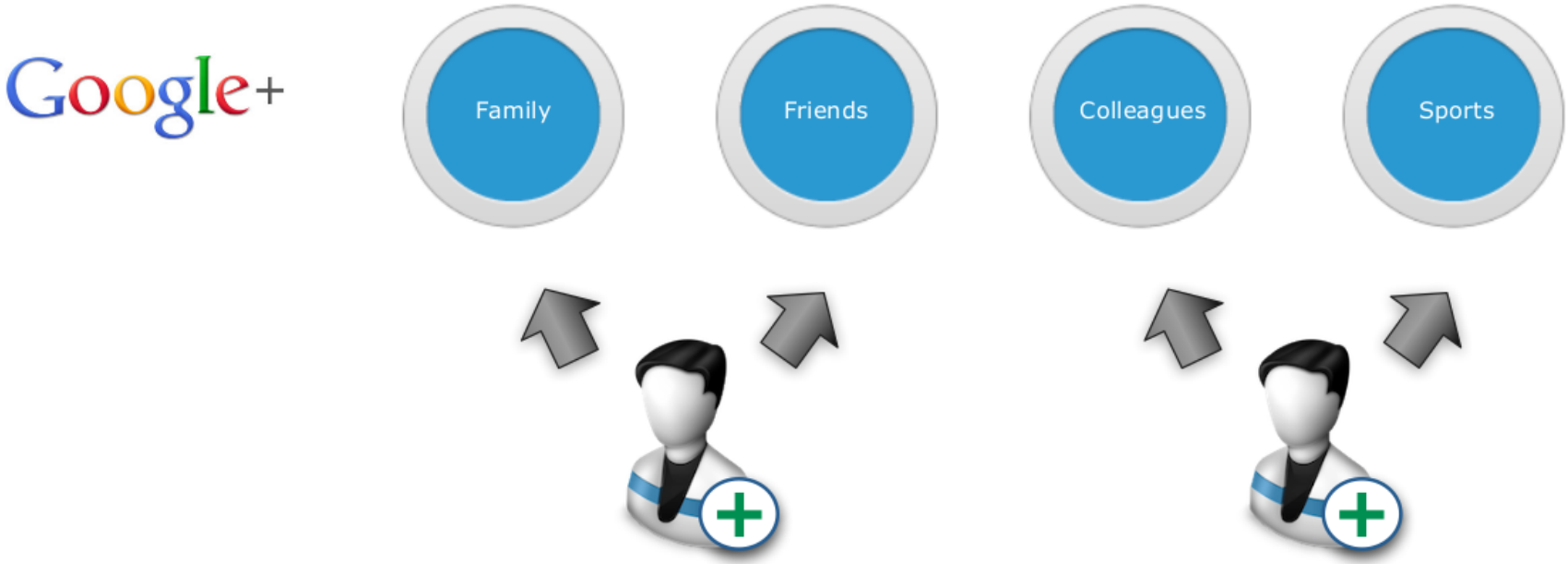
Networkx: Modularity - Optimization

```
Community 0
['24', '25', '26', '28', '29', '32']
Community 1
['27', '21', '23', '9', '10', '15', '16', '33', '31', '30', '34', '19']
Community 2
['20', '22', '1', '3', '2', '4', '8', '13', '12', '14', '18']
Community 3
['5', '7', '6', '11', '17']
```



Overlapping vs non overlapping

- Do disjoint non-overlapping communities make sense in empirical social networks?

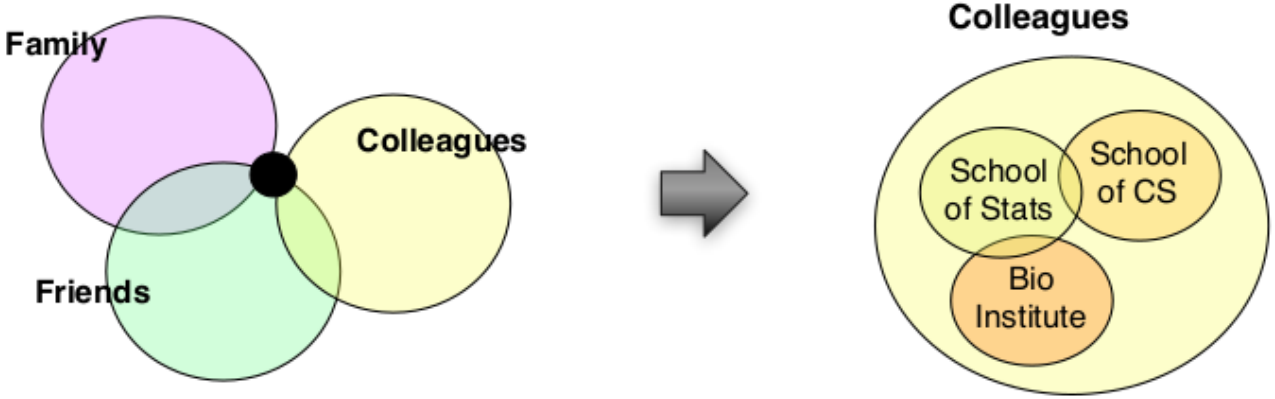


Overlapping vs non overlapping

- Do disjoint non-overlapping communities make sense in empirical social networks?

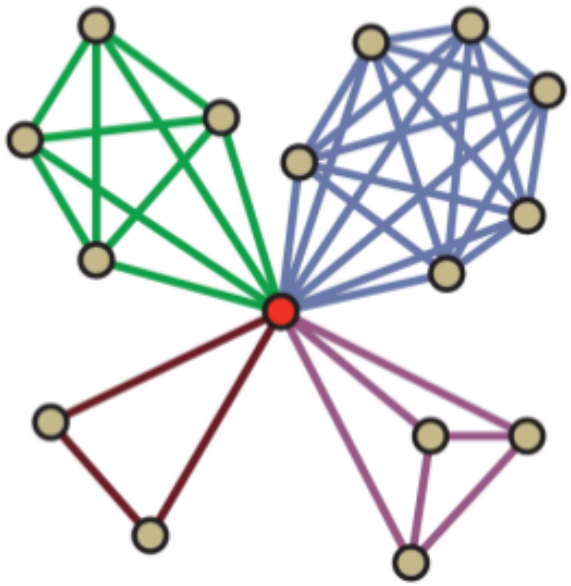
Google+

Overlapping communities may exist at different resolutions.

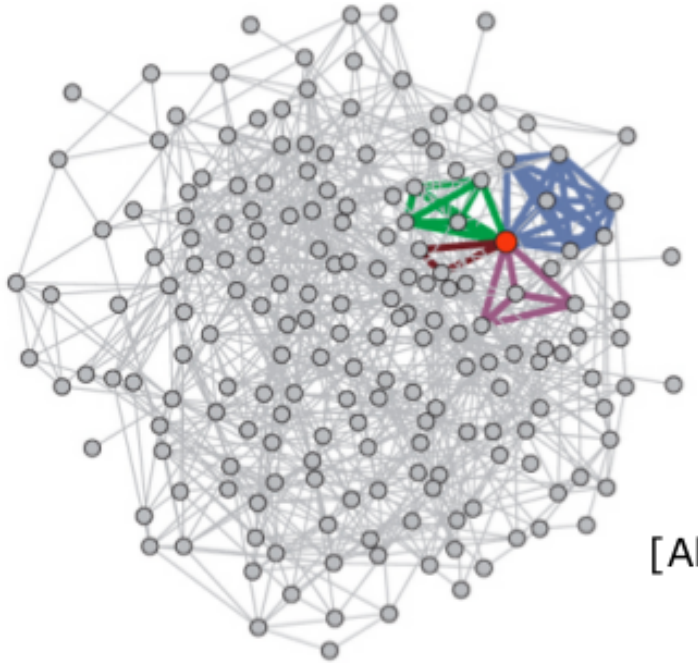


Overlapping vs non overlapping

- Distinct "non-overlapping" communities rarely exist at large scales in many empirical networks [Leskovec et al, 2008].
- ➡ Communities overlap pervasively, making it impossible to partition the networks without splitting communities [Reid et al, 2011].



Community overlap at an ego level



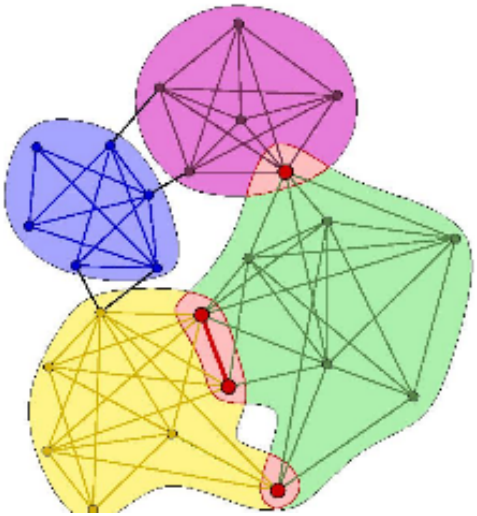
[Ahn et al, 2010]

Community overlap at a global level

Overlapping vs non overlapping

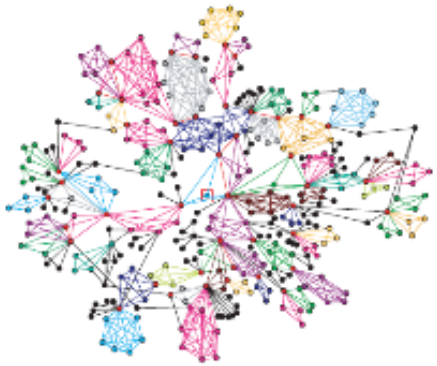
- **CFinder**: algorithm based on the clique percolation method [Palla et al, 2005].
- Identify *k*-cliques: a fully connected subgraph *k* nodes.
- Pair of *k*-cliques are "adjacent" if they share *k−1* nodes.
- Form overlapping communities from maximal union of *k*-cliques that can be reached from each other through adjacent *k*-cliques.

http://cfinder.org



Set of overlapping communities built from 4-cliques.

Co-authorship Network



[Palla et al, 2005]

Overlapping vs non overlapping

- **Greedy Clique Expansion (GCE)**: identify distinct cliques as seeds, expands the seeds by greedily optimising a local fitness function [Lee et al, 2010].

https://sites.google.com/site/greedycliqueexpansion

- **MOSES**: scalable approach for identifying highly-overlapping communities [McDaid et al, 2010].
  - Randomly select an edge, greedily expand a community around the edge to optimise an objective function.
  - Delete "poor quality" communities.
  - Fine-tune communities by re-assigning individual nodes.

https://sites.google.com/site/aaronmcdaid/moses

Overlapping community detection in Networkx

Algorithmes for detecting communites as well overlapping communities using NetworkX can be found here:

Community algorithms in NetworkX

https://networkx.github.io/documentation/stable/reference/algorithms/community.html

The End