Walk-Based Centrality and Communicability Measures for Network Analysis

Michele Benzi

Department of Mathematics and Computer Science

Emory University

Atlanta, Georgia, USA

Workshop on Innovative Clustering Methods for Large Graphs and Block Methods

Toulouse, France

July 6-8, 2015

Outline

- Complex graphs and networks
- Clustering coefficient
- Graph spectra
- Centrality measures
- Network communicability

Some features of complex networks

Complex networks provide models for a wide variety of physical, biological, engineered or social systems.

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 ...

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.

Some features of complex networks (cont.)

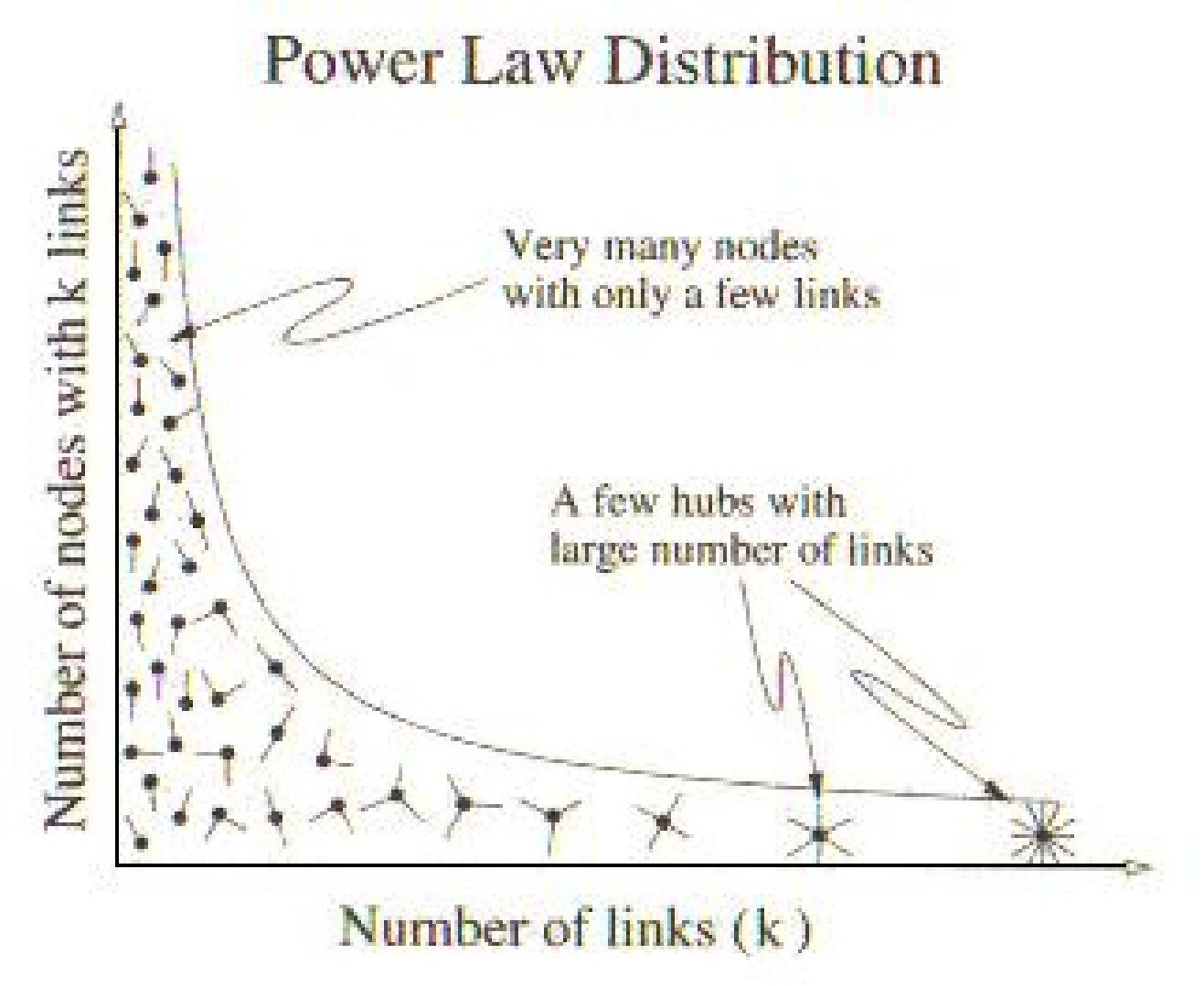
Some of the attributes typical of many real-world complex networks are:

- "Scale-free": the degree distribution follows a power law $p(d) \approx d^{-\gamma}$ (highly skewed, heavy-tailed)
- "Small-world":
  - Small graph diameter, short average distance between nodes
  - High clustering coefficient: many triangles, hubs, ...
- Hierarchical structure
- Rich in "motifs"
- Overlapping communities

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

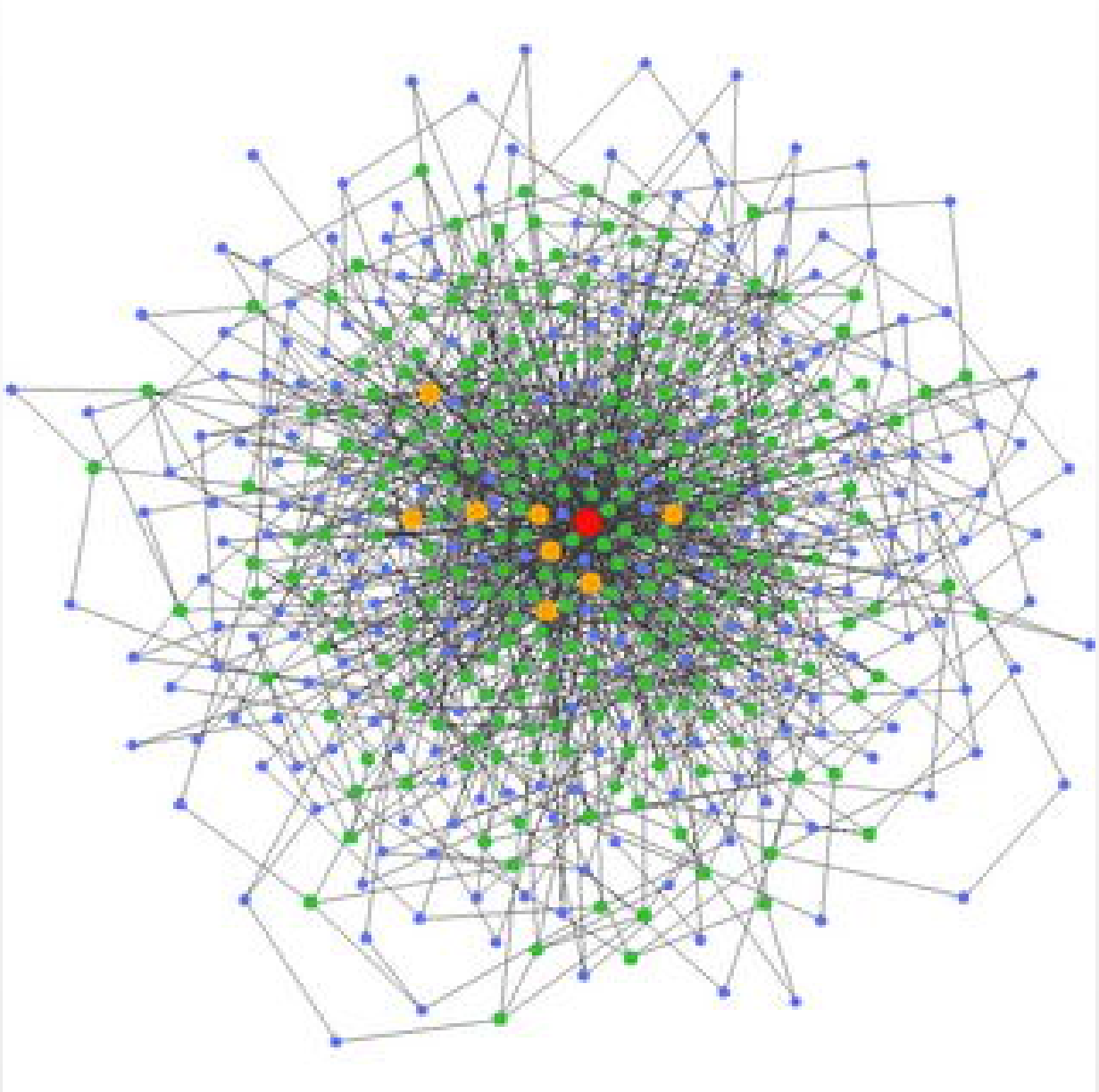
Caveat: there are important examples of real-world complex networks lacking one or more of these attributes.

Power law degree distribution

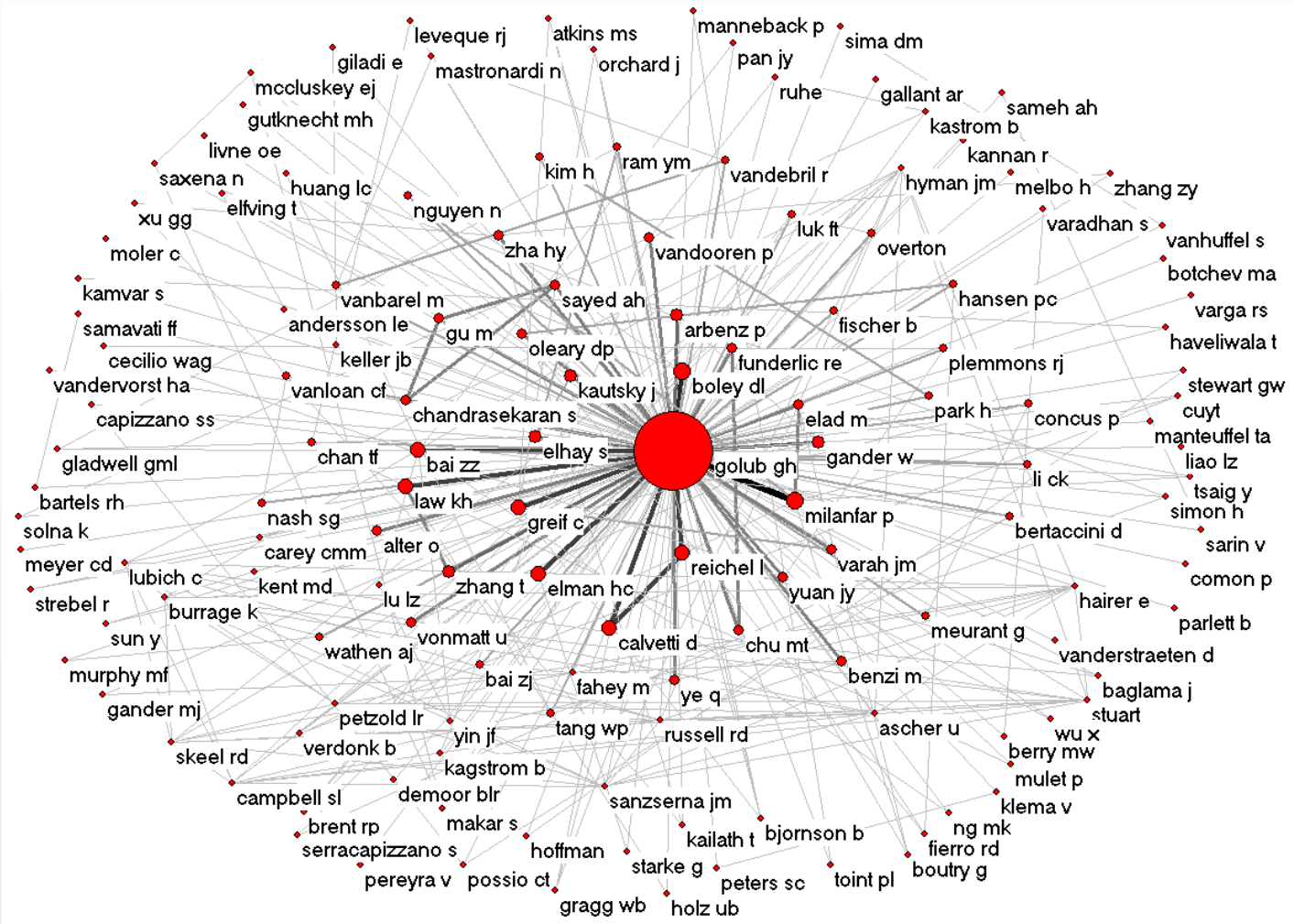


From L.-A. Barabási, *Linked. The New Science of Networks*, Perseus Publishing, Cambridge, MA, 2002.

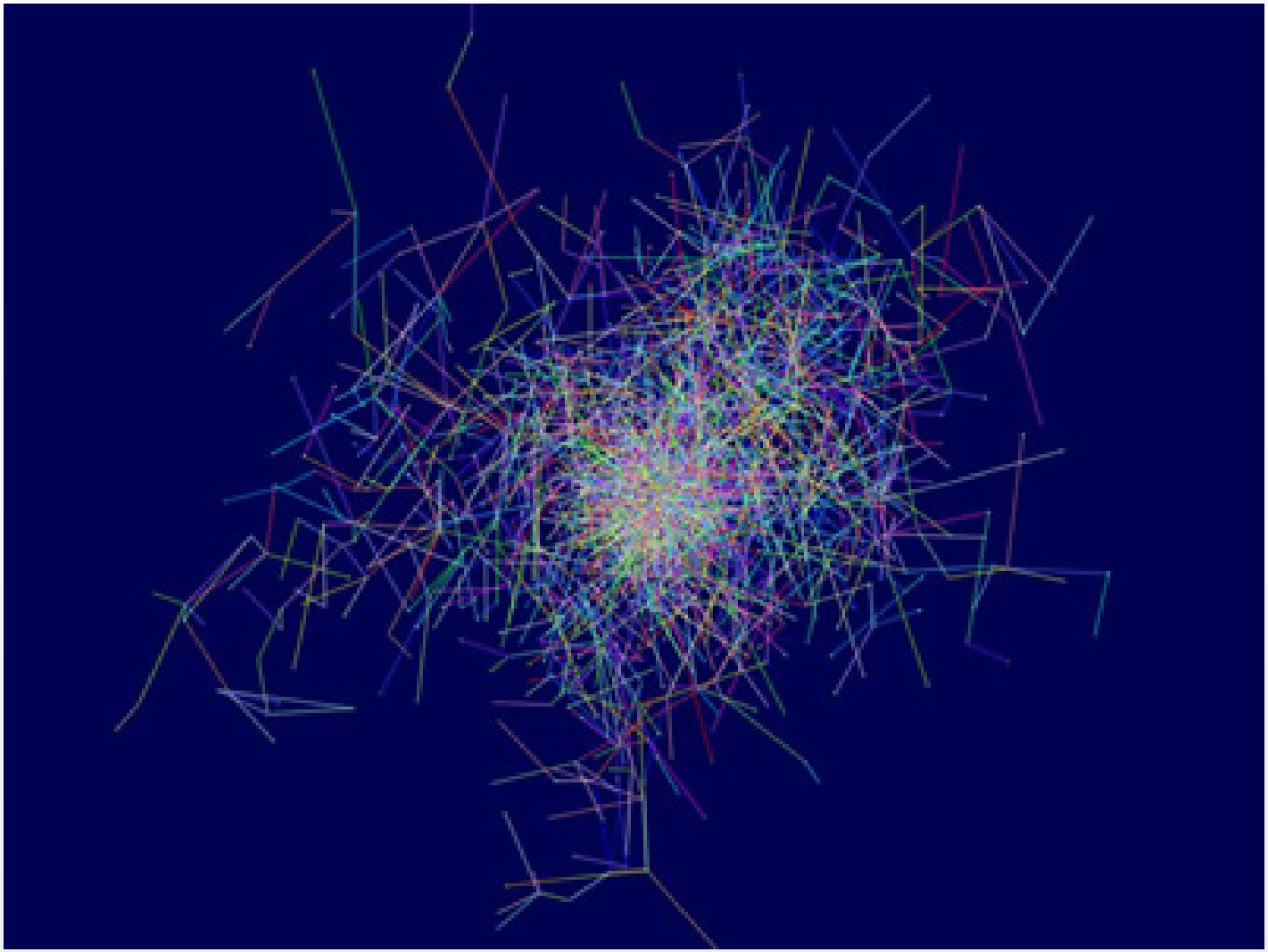
Barabási–Albert model (preferential attachment)



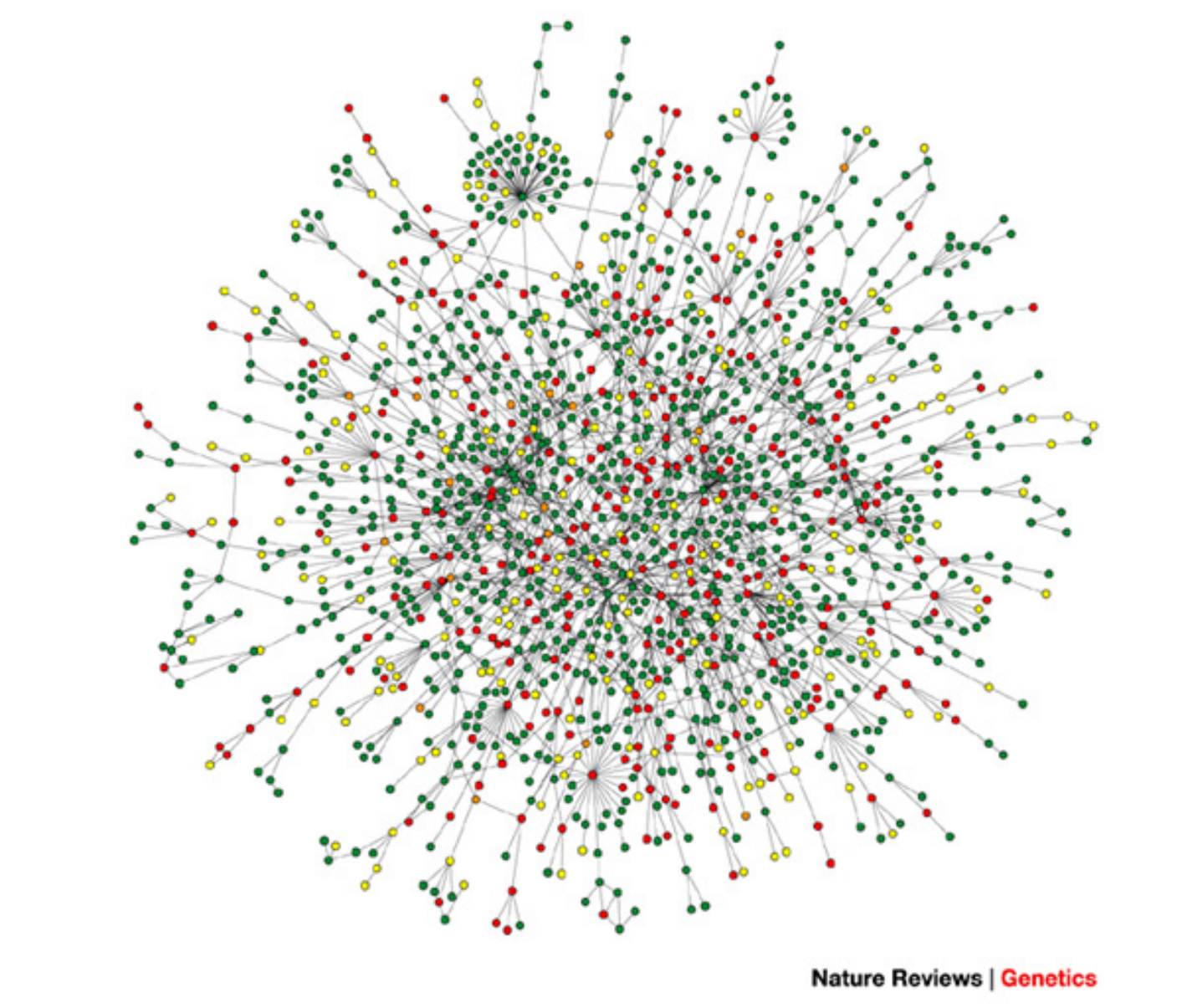
A real-world complex network: Golub collaboration graph



A real-world complex network: Erdös collaboration graph



PPI network of *Saccharomyces cerevisiae* (beer yeast)



Sexual contact network of injecting drug users

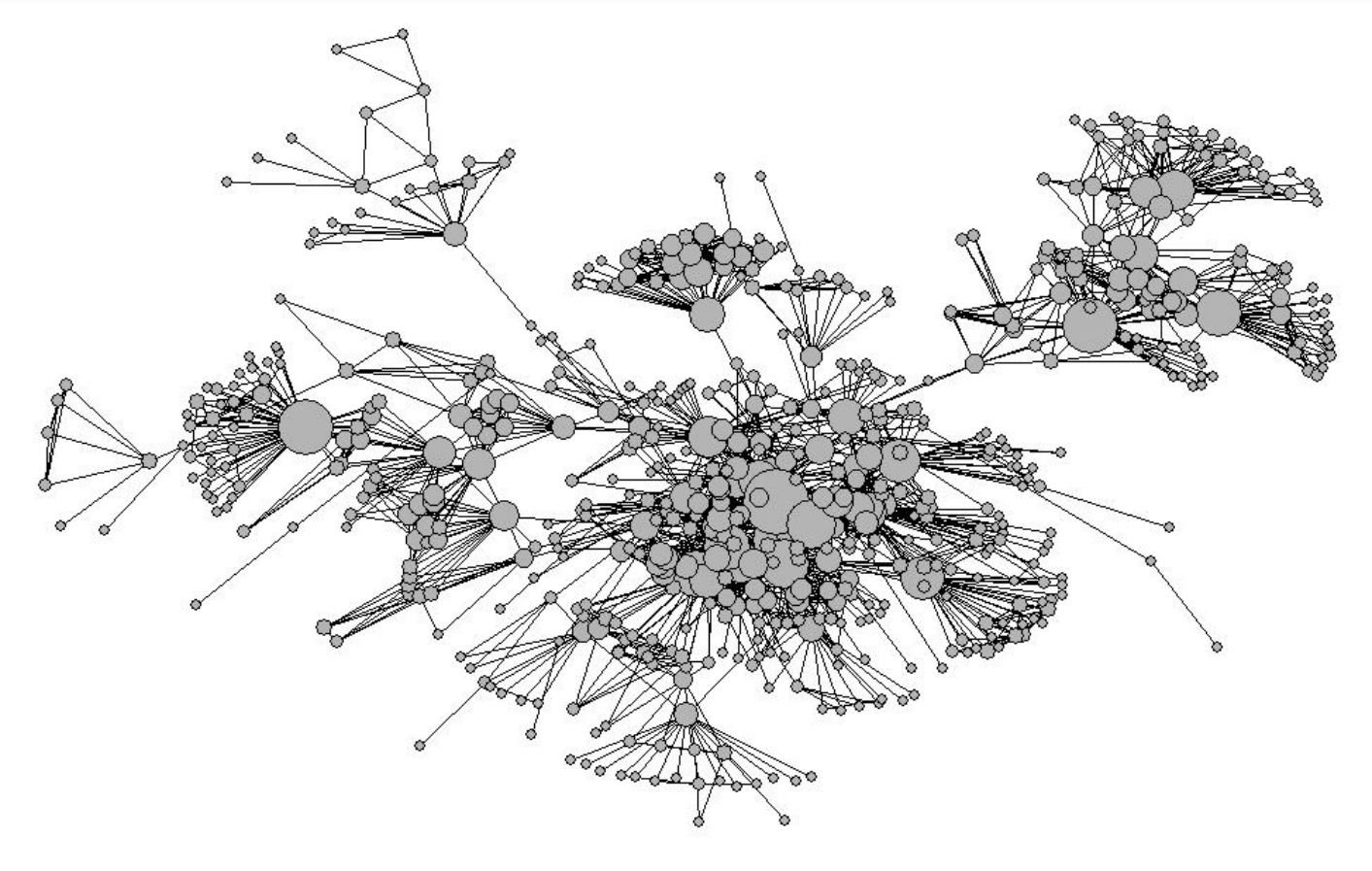
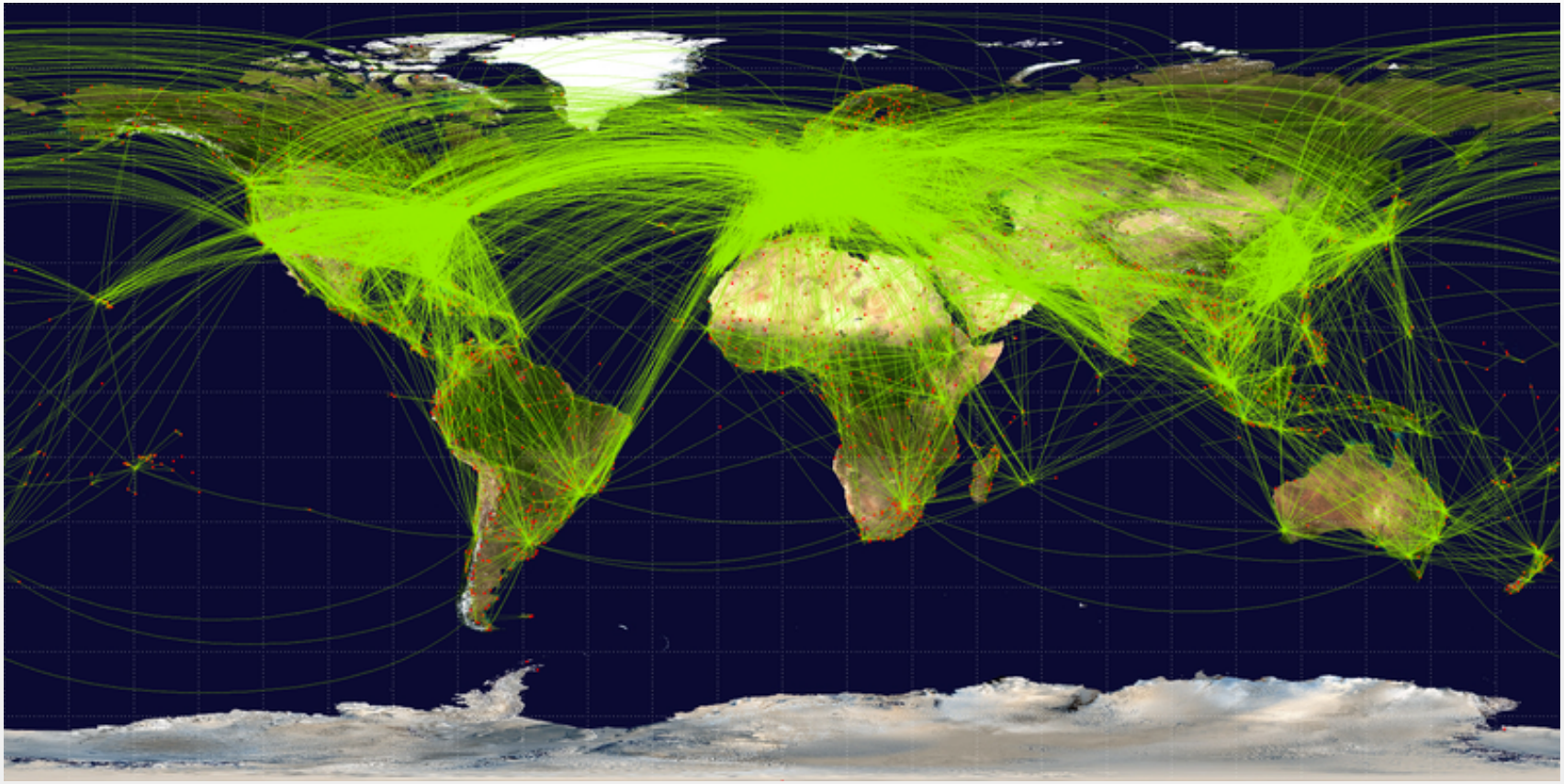
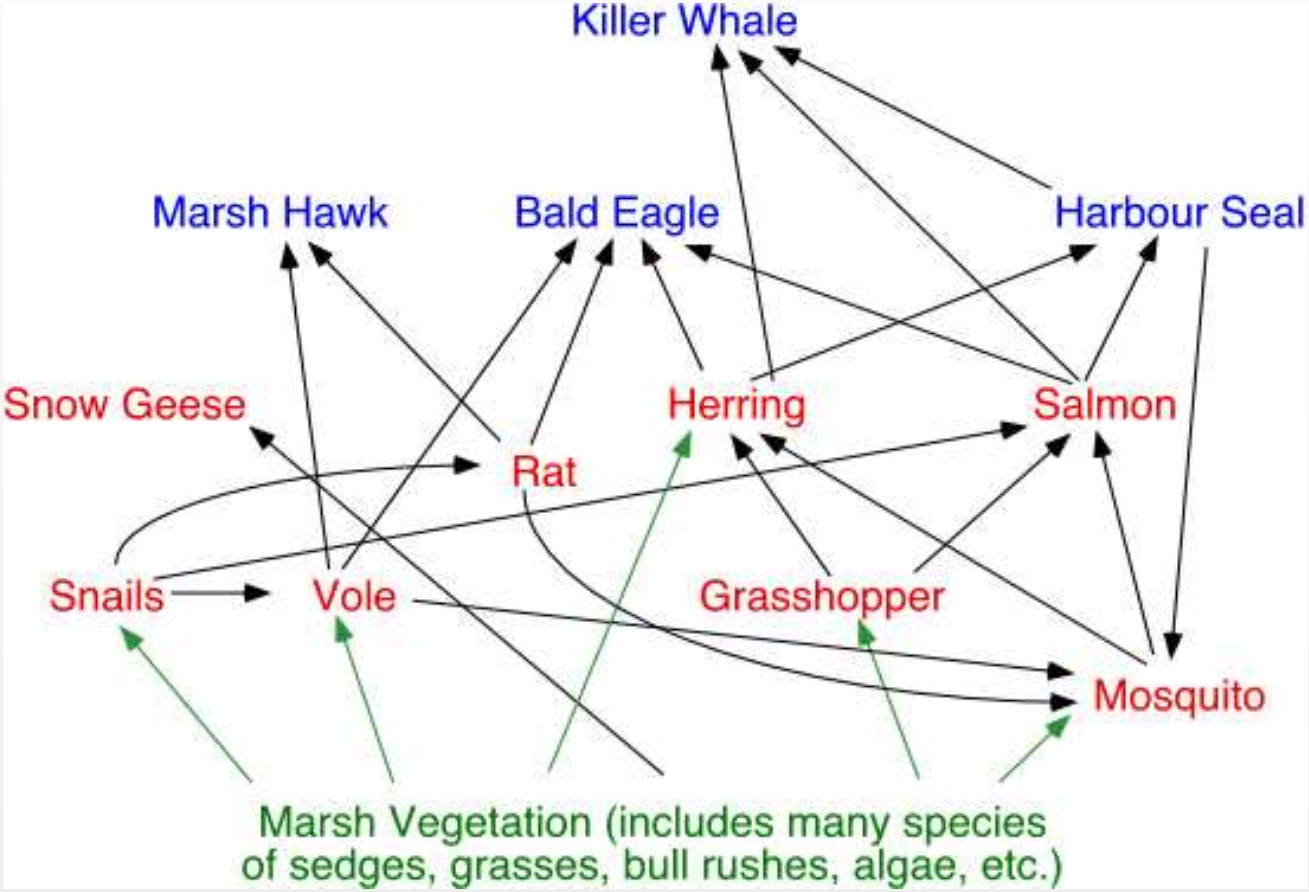


Figure courtesy of Ernesto Estrada.

The world airline routemap



Example of (directed) complex network: a food web



Network analysis

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

- 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
- How do "disturbances" spread in a network?
  - Spreading of epidemics, beliefs, rumors, fads,...
  - Routing of messages; bottlenecks, returnability
- How to detect "community structures" in a network?
  - Clustering, triadic closure (transitivity)
  - Partitioning

Formal definitions

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

A graph $G = (V, E)$ consists of a (finite) set $V = \{v_1, v_2, \ldots, v_N\}$ of nodes (or vertices) and a set E of edges (or links), which are pairs $\{v_i, v_j\}$ with $v_i, v_j \in V$.

The graph G is directed if the edges $\{v_i, v_j\} \in E$ are ordered pairs $= (v_i, v_j) \in V \times V$, otherwise G is undirected. A directed graph is often referred to as a digraph.

A loop in G is an edge from a node to itself. Loops are often ignored or excluded.

A graph G is weighted if numerical values are associated with its edges. If all the edges are given the same value 1, we say that the graph is unweighted.

A simple graph is an unweighted graph without multiple edges or loops.

Formal definitions (cont.)

A walk of length k in G is a set of nodes $v_{i_1}, v_{i_2}, \ldots v_{i_k}, v_{i_{k+1}}$ such that for all $1 \leq j \leq k$, there is an edge between v_{i_j} and $v_{i_{j+1}}$.

A closed walk is a walk where $v_{i_1} = v_{i_{k+1}}$.

A path is a walk with no repeated nodes.

A cycle is a path with an edge between the first and last node. In other words, a cycle is a closed path.

A triangle in G is a cycle of length 3.

Formal definitions (cont.)

The geodesic distance $d(v_i, v_j)$ between two nodes is the length of the shortest path connecting v_i and v_j. We let $d(v_i, v_j) = \infty$ if no such path exists.

The diameter of a graph $G = (V, E)$ is defined as

$$\text{diam}(G) := \max_{v_i, v_j \in V} d(v_i, v_j)\,.$$

A graph G is connected if for every pair of nodes v_i and v_j there is a path in G that starts at v_i and ends at v_j; i.e., $\text{diam}(G) < \infty$.

These definitions apply to both undirected and directed graphs, though in the latter case the orientation of the edges must be taken into account.

Formal definitions (cont.)

To every unweighted graph $G = (V, E)$ we associate its adjacency matrix $A = [a_{ij}] \in \mathbb{R}^{N \times N}$, with

$$a_{ij} = \begin{cases} 1, & \text{if } (v_i, v_j) \in E, \\ 0, & \text{else.} \end{cases}$$

Any renumbering of the graph nodes results in a symmetric permutation $A \longrightarrow PAP^T$ of the adjacency matrix of the graph.

If G is an undirected graph, A is symmetric with zeros along the main diagonal (A is "hollow"). In this case, the eigenvalues of A are all real. We label the eigenvalues of A in non-increasing order: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Note that A is always indefinite if $E \neq \emptyset$.

If G is connected and acyclic, then λ_1 is simple and satisfies $\lambda_1 > |\lambda_i|$ for $2 \leq i \leq N$. Also, the associated eigenvector can be chosen to be strictly positive (Perron–Frobenius Theorem).

Formal definitions (cont.)

If G is undirected, the degree d_i of node v_i is the number of edges incident to v_i in G. In other words, d_i is the number of "immediate neighbors" of v_i in G. A regular graph is a graph where every node has the same degree d.

Note that in terms of the adjacency matrix, $d_i = \sum_{j=1}^{N} a_{ij}$.

For a directed graph, we define the in-degree of node v_i as the number d_i^{in} of edges ending in v_i, and the out-degree of v_i as as the number d_i^{out} of edges originating at v_i.

In terms of the (nonsymmetric) adjacency matrix,

$$d_i^{in} = \sum_{i=1}^{N} a_{ij}, \quad d_i^{out} = \sum_{j=1}^{N} a_{ij}.$$

Hence, the column sums of A give the in-degrees and the row sums give the out-degrees.

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 like the Erdös–Rényi (ER) graphs.

Clustering (cont.)

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

$$\Delta_i = \frac{1}{2}[A^3]_{ii},$$

while the total number of triangles in G is given by

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

Hence, computing clustering coefficients for a graph G requires estimating the diagonal of A^3, which for very large networks can be a challenging task.

Some of the most successful approaches are based on randomized algorithms (can be used to estimate $[f(A)]_{ii}$ for fairly general f).

Graph spectra

The adjacency matrix A of an undirected network is always symmetric, hence its eigenvalues are all real. The eigenvalue distribution reflects global properties of G; as we shall see, the eigenvectors also carry important information about the network structure.

The spectrum of a network is the spectrum of the corresponding adjacency matrix A.

Spectral graph theory is devoted to the study of the eigenvalues and eigenvectors of graphs. Much work has been done in characterizing the spectra (eigenvalue distributions) of random graphs and of certain classes of complex graphs (e.g., scale-free graphs).

The adjacency matrix of a directed network, on the other hand, is typically nonsymmetric and will have complex (non-real) eigenvalues in general. In this case, the singular values of A are often useful.

Graph Laplacian spectra

For an undirected network, let $D = \text{diag}(d_1, \ldots, d_N)$. The spectrum of the graph Laplacian $L = D - A$ also plays an important role. For instance, the connectivity properties of G can be characterized in terms of spectral properties of L; moreover, L plays an important role in the study of diffusion processes on G.

The eigenvalues of L are all real and nonnegative, and since $L\,\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ denotes the vector of all ones, L is singular. Hence, $0 \in \sigma(L)$. The graph Laplacian L is a singular M-matrix.

Theorem. The multiplicity of 0 as an eigenvalue of L coincides with the number of connected components of the network.

Corollary. For a connected network, the null space of the graph Laplacian is 1-dimensional and is spanned by $\mathbf{1}$. Thus, $\text{rank}(L) = N - 1$.

Graph Laplacian spectra (cont.)

Let G be a simple, connected graph.

The smallest nonzero eigenvalue of L is called the algebraic connectivity of the graph, and the associated eigenvector is called the Fiedler vector. Since this eigenvector must be orthogonal to the vector of all ones, it must contain both positive and negative entries.

There exist elegant graph partitioning algorithms that assign nodes to different subgraphs based on the sign of the entries of the Fiedler vector.

These methods, however, tend to work well only in the case of fairly regular graphs, such as those arising from the discretization of PDEs. In general, partitioning complex graphs (scale-free graphs in particular) is very hard.

See V. Kuhlemann & P. Vassilevski, *Improving the communication pattern in matrix–vector operations for large scale-free graphs by disaggregation*, SISC 35 (2013).

Graph Laplacian spectra (cont.)

If the graph G is regular of degree d, then $L = dI_N - A$ and the eigenvalues of the Laplacian are just

$$\lambda_i(L) = d - \lambda_i(A), \quad 1 \leq i \leq N\,.$$

If G is not a regular graph, there is no simple relationship between the eigenvalues of L and those of A.

Also useful is the notion of normalized Laplacian:

$$\widehat{L} := I_N - D^{-1/2} A D^{-1/2}\,.$$

In the case of directed graphs, there are several distinct notions of graph Laplacian in the literature.

Three nice books on graph spectra

F. Chung and L. Lu, *Spectral Graph Theory*, American Mathematical Society, 1997.

F. Chung and L. Lu, *Complex Graphs and Networks*, American Mathematical Society, 2006.

P. Van Mieghem, *Graph Spectra for Complex Networks*, Cambridge University Press, 2011.

Centrality measures

There are dozens of different definitions of centrality for nodes in a graph. The simplest is degree centrality, which is just the degree d_i of node i. This does not take into account the "importance" of the nodes a given nodes is connected to—only their number.

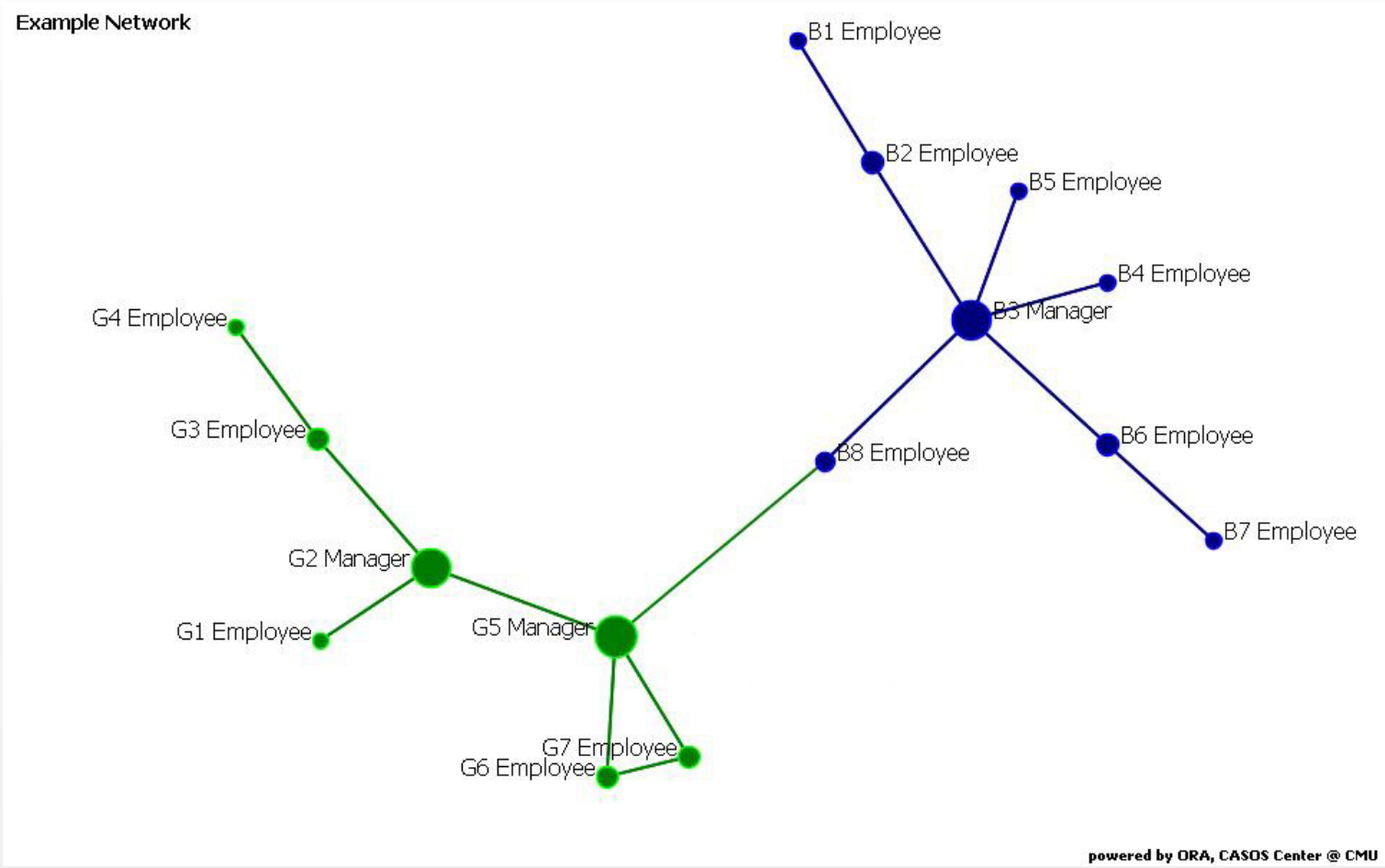
A popular notion of centrality is betweenness centrality (Freeman, 1977), defined for any node $i \in V$ as

$$C_B(i) := \sum_{j \neq i} \sum_{k \neq i} \delta_{jk}(i)\,,$$

where $\delta_{jk}(i)$ is the fraction of all shortest paths in the graph between nodes j and k which contain node i:

$$\delta_{jk}(i) := \frac{\#\,\text{of shortest paths between}\, j,k\,\text{containing}\, i}{\#\,\text{of shortest paths between}\, j,k}\,.$$

Centrality measures (cont.)



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

Centrality measures (cont.)

Another centrality measure popular in social network analysis is closeness centrality (Freeman, 1979), defined as

$$C_C(i) = \frac{1}{\sum_{j \in V} d(i,j)} .$$

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.

Centrality measures (cont.)

We mention here that for directed graphs it is often necessary to distinguish between hubs and authorities. Indeed, in a directed graph a node plays two roles: broadcaster and receiver of information.

Crude broadcast and receive centrality measures are provided by the out-degree d_i^{out} and by the in-degree d_i^{in} of the node, respectively.

Other, more refined receive and broadcast centrality measures have been introduced.

Spectral centrality measures

Bonacich's eigenvector centrality (1987) uses the entries of the dominant eigenvector $\mathbf{x}$ to rank the nodes in the network in order of importance: the larger x_i is, the more important node i is considered to be. By the Perron–Frobenius Theorem, the vector $\mathbf{x}$ is positive and unique provided the network is connected.

The underlying idea is that "a node is important if it is linked to many important nodes." This *recursive definition* corresponds to the fixed-point iteration

$$\mathbf{x}^{(k+1)} = A\mathbf{x}^{(k)}, \quad k = 0, 1, \ldots$$

which converges, upon normalization, to the dominant eigenvector of A as $k \to \infty$. The rate of convergence depends on the spectral gap $\gamma = \lambda_1 - \lambda_2$. The larger γ, the faster the convergence.

In the case of directed networks, the dominant left and right eigenvectors of A provide authority and hub scores, respectively.

Spectral centrality measures

Eigenvector centrality has the following interpretation in terms of walks on G:

The eigenvector centrality x_i of node i is the limit as $k \to \infty$ of the percentage of walks of length k which start at node i among all walks of length k.

See for example D. Cvetković, P. Rowlinson, and S. Simić, *Eigenspaces of Graphs*, Cambridge University Press, 1997.

Thus, the eigenvector centrality measures the global influence of node i.

Spectral centrality measures (cont.)

Google's PageRank algorithm (Brin & Page, 1998) is a variant of eigenvector centrality, applied to the (directed) graph representing web pages (documents), with hyperlinks between pages playing the role of directed edges. Since the WWW graph is *not* connected, some tweaks (in the form of a rank-one modification to the hyperlink matrix) are needed to have a unique PageRank eigenvector.

PageRank has a probabilistic interpretation in terms of random walks on the web graph, a special type of Markov chain. The PageRank eigenvector is the stationary probability distribution of this Markov chain.

An alternative approach (HITS), proposed by J. Kleinberg in 1998, uses the dominant left and right singular vectors of the (nonsymmetric) adjacency matrix of the graph in order to obtain both hub and authority scores. We will return to this topic in Part II.

Subgraph centrality

We now turn to walk-based centrality measures.

Subgraph centrality (Estrada & Rodríguez-Velásquez, *Phys. Rev. E*, 2005) 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 $k = 1, 2, \ldots$ the number of closed walks in G starting and ending at node i, with longer walks being penalized (given a smaller weight).

It is sometimes useful to introduce a tuning parameter $\beta > 0$ ("inverse temperature") to simulate external influences on the network, for example, increased tension in a social network, financial distress in the banking system, etc.

Subgraph centrality (cont.)

Recall that

- $(A^k)_{ii} = \#$ of closed walks of length k based at node i,
- $(A^k)_{ij} = \#$ of walks of length k that connect nodes i and j.

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

$$\begin{aligned} SC(i) &= \left[I + \beta A + \frac{\beta^2}{2!} A^2 + \frac{\beta^3}{3!} A^3 + \cdots \right]_{ii} \\ &= [\mathrm{e}^{\beta A}]_{ii} \, . \end{aligned}$$

Note that $SC(i) \geq 1$. Subgraph centrality has been used successfully in various settings, including proteomics and neuroscience.

Note: the weights are needed to "penalize" longer walks, and to make the power series converge.

Katz centrality

Of course different weights can be used, leading to different matrix functions, such as the resolvent (Katz, 1953):

$$(I - \alpha A)^{-1} = I + \alpha A + \alpha^2 A^2 + \cdots = \sum_{k=0}^{\infty} \alpha^k A^k, \quad 0 < \alpha < 1/\lambda_1.$$

Note that $I - \alpha A$ is a nonsingular M-matrix, in particular $(I - \alpha A)^{-1} \geq 0$.

Originally, Katz proposed to use the row sums of $(I - \alpha A)^{-1}$ as a centrality measure:

$$C_K(i) = \mathbf{e}_i^T (I - \alpha A)^{-1} \mathbf{1}\,.$$

Resolvent-based subgraph centrality uses instead the diagonal entries of $(I - \alpha A)^{-1}$ (E. Estrada & D. Higham, SIAM Rev., 2010).

In the case of a directed network one can use the solution vectors of the linear systems

$$(I - \alpha A)\mathbf{x} = \mathbf{1} \quad \text{and} \quad (I - \alpha A^T)\mathbf{y} = \mathbf{1}$$

to rank hubs and authorities. These are the row and column sums of the matrix resolvent $(I - \alpha A)^{-1}$, respectively.

Comparing centrality measures

Different centrality measures correspond to different notions of what it means for a node to be “influential"; some (like degree) only look at local influence, others (like eigenvector centrality and PageRank) emphasize long distance influences, and others (like subgraph and Katz centrality) try to take into account short, medium and long range influences.

Others yet, like closeness and betweenness centrality, emphasize nodes that are a short distance away from most other nodes, or that are central to the flow of information along shortest paths in G.

And these are just a few of the many centrality measures proposed in the literature!

Comparing centrality measures (cont.)

In the case of very large-scale networks (with millions or even billions of nodes/edges), computational cost becomes a limiting factor.

Of course, degree centrality is very cheap, but as we know it is not very satisfactory.

At the other end of the spectrum, subgraph centrality (which scales roughly like $\mathcal{O}(N^2)$) is very satisfactory but is too expensive for very large networks.

For large networks, methods like eigenvector and Katz centrality, while still potentially expensive, are often used, together with closeness and betweenness centrality.

Comparing centrality measures (cont.)

Theorem: Let A be the adjacency matrix for a simple, connected, undirected graph G. Then

- For $\alpha \to 0+$, Katz centrality reduces to degree centrality;
- For $\alpha \to \frac{1}{\lambda_1}-$, Katz centrality reduces to eigenvector centrality;
- For $\beta \to 0+$, subgraph centrality reduces to degree centrality;
- For $\beta \to \infty$, subgraph centrality reduces to eigenvector centrality.

Note: A similar result holds for directed networks, in which case we need to distinguish between in-degree, out-degree, left/right dominant eigenvectors, and row/column sums of $(I - \alpha A)^{-1}$ and $e^{\beta A}$.

M. Benzi & C. Klymko, *On the limiting behavior of parameter-dependent network centrality measures*, SIMAX, 36 (2015).

Communicability

Communicability measures "how well" two nodes $i \in V$ and $j \in V$ communicate. It is defined as (Estrada & Hatano, 2008):

$$\begin{aligned} C(i,j) &= [\mathrm{e}^{\beta A}]_{ij} \\ &= \left[I + \beta A + \frac{\beta^2}{2!} A^2 + \frac{\beta^3}{3!} A^3 + \cdots \right]_{ij} \\ &\approx \text{weighted sum of walks joining nodes } i \text{ and } j. \end{aligned}$$

Note that as the temperature increases ($\beta \to 0$) the communicability between nodes decreases ($\mathrm{e}^{\beta A} \to I$).

Communicability has been successfully used to identify bottlenecks in networks (e.g., regions of low communicability in brain networks) and for community detection.

E. Estrada & N. Hatano, *Phys. Rev. E*, 77 (2008), 036111;
E. Estrada, N. Hatano, & M. Benzi, *Phys. Rep.*, 514 (2012), 89–119.

Communicability (cont.)

See also

E. Estrada, *Community detection based on network communicability*, Chaos 21, 016103 (2011).

The basic idea is that two nodes (i, j) that are in the same community should display a (much) higher value of $(\mathrm{e}^A)_{ij}$ than two nodes that belong to different communities.

Estrada's algorithm performs very well in practice but is prohibitive for large graphs, since it requires estimating $(\mathrm{e}^A)_{ij}$ for all $i > j$.

Communicability (cont.)

The total communicability of a node is defined as

$$TC(i) := \sum_{j=1}^{N} C(i,j) = \sum_{j=1}^{N} [\mathrm{e}^{\beta A}]_{ij}.$$

This is a node centrality measure that is based on the intuition that a node that communicates well with all other nodes (as a whole) should be an important node in terms of the spreading of information on the network.

It can be computed efficiently, even for graphs with millions of nodes, since it involves computing $\mathrm{e}^{\beta A}\,\mathbf{1}$, and there are very efficient Krylov subspace methods for this task.

Versions of TC for digraphs also exist.

Communicability (cont.)

The normalized total communicability of G:

$$TC(G) := \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} C(i,j) = \frac{1}{N}\mathbf{1}^T \mathrm{e}^{\beta A}\,\mathbf{1}$$

provides a global measure of how "well-connected" a network is, and can be used to compare different network designs.

Thus, highly connected networks, such as small-world networks without bottlenecks, can be expected to have a high total communicability.

Conversely, large-diameter networks with a high degree of locality (such as regular grids, road networks, etc.) or networks containing bottlenecks are likely to display a low value of $TC(G)$.

There are algorithms that can be used to design sparse networks with high total communicability (F. Arrigo & M. Benzi, submitted to SISC).

Global communicability measures (cont.)

Note that $TC(G)$ can be easily bounded from below and from above:

$$\frac{1}{N} EE(G) \leq TC(G) \leq \mathrm{e}^{\beta \lambda_1}$$

where $EE(G) = \mathrm{Tr}(\mathrm{e}^{\beta A})$ is the Estrada index of the graph. These bounds are sharp, as can be seen from trivial examples.

In the next Table we present the results of some calculations (for $\beta = 1$) of the normalized Estrada index, global communicability and e^{λ_1} for various real-world networks.

M. Benzi & C. F. Klymko, *J. Complex Networks*, 1 (2013), 124–149.

Communicability (cont.)

Table: Comparison of the normalized Estrada index $E(G) = \mathrm{Tr}(e^A)/N$, the normalized total network communicability $TC(G)$, and e^{λ_1} for various real-world networks.

Network	$E(G)$	$TC(G)$	e^{λ_1}
Zachary Karate Club	30.62	608.79	833.81
Drug User	1.12e05	1.15e07	6.63e07
Yeast PPI	1.37e05	3.97e07	2.90e08
Pajek/Erdos971	3.84e04	4.20e06	1.81e07
Pajek/Erdos972	408.23	1.53e05	1.88e06
Pajek/Erdos982	538.58	2.07e05	2.73e06
Pajek/Erdos992	678.87	2.50e05	3.73e06
SNAP/ca-GrQc	1.24e16	8.80e17	6.47e19
SNAP/ca-HepTh	3.05e09	1.06e11	3.01e13
SNAP/as-735	3.00e16	3.64e19	2.32e20
Gleich/Minnesota	2.86	14.13	35.34

Total communicability (cont.)

The total network communicability is a good measure of network connectivity and robustness. It is also inexpensive to compute, the cost often being $O(n)$ in practice.

It is also possible to define the notion of communicability “block-wise", for a given partitioning of the vertex set V of G into subsets $V_1, \ldots, V_p$.

It would be interesting to explore possible uses of communicability and related ideas for the purpose of identifying blocks of tighly connecte nodes, or groups of nodes, in a graph.