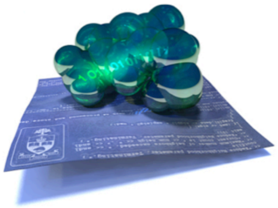
A
Bioinformatics
Course

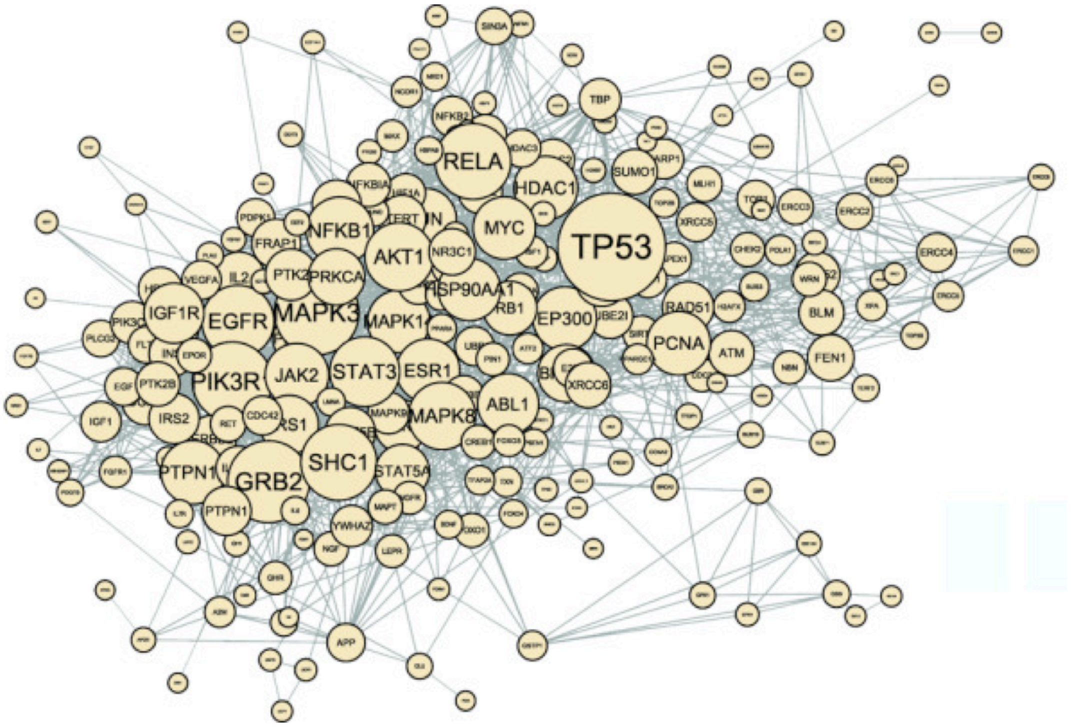
Graphs and Networks



Boris Steipe

department of biochemistry – department of molecular genetics
University of Toronto

Protein-Protein Interactions



The human protein-protein interaction network of aging-associated genes. A total of 261 aging-associated genes were assembled using the GenAge Human Database. Protein-protein interactions of the human interactome were collected from the 8.0 version of the STRING database using physical contacts only. The network was visualized using the Cytoscape program. The degree (number of neighbors) of nodes is represented by the size of the circle and the font. Note the high number of signaling pathway proteins among hubs (nodes with degrees - and therefore size - much greater than average), exemplified by the MAPK/ERK and PI3K/AKT proteins.

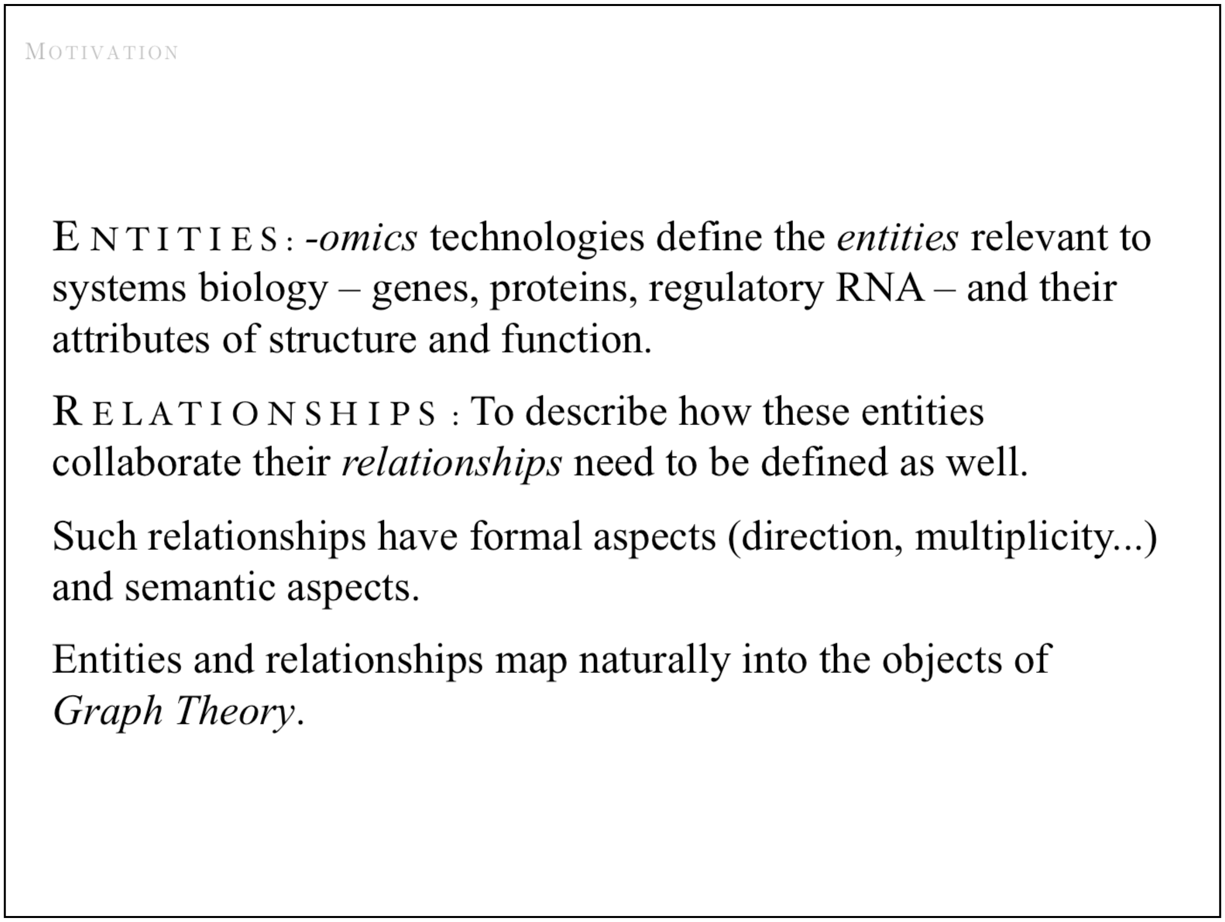
Figure and caption from:

Simkó G, Gyurkó D, Veres DV, Nánási T, and Peter Csermely P (2009) Network strategies to understand the aging process and help age-related drug design. Genome Med. 2009; 1(9): 90

Plots like these have become ubiquitous in the litearture. They have an undeniable aesthetic quality, but the information about biological processes is limited. The authors mention “the high number of signalling pathway proteins” among the nodes of large degree, but they have not (i) mentioned how a “signalling pathway protein” is defined, (ii) described in detail how those “aging-related” proteins were selected in the first place, (iii) compared this fraction against the fraction of signalling pathway proteins among all genes, (iv) quantified their finding, (v) emphasized this class of proteins by colour in the plot.

Plotting networks is in general not a good way to analyze them: the inference process has to go the other way around: analyze the network with computational means, then visualize the results in a plot. This plot may or may not be a graph similar to the one above, in this particular case a simple boxplot of node-degree by GO biological process category would have been more effective. Network plots show interactions, but no hypothesis about those interactions has been suggested that the graph could help us visualize.

Let us thus discuss the backgrounds of graph theory and useful measures that we can apply towards biological inference from **relationship data**.



Motivation

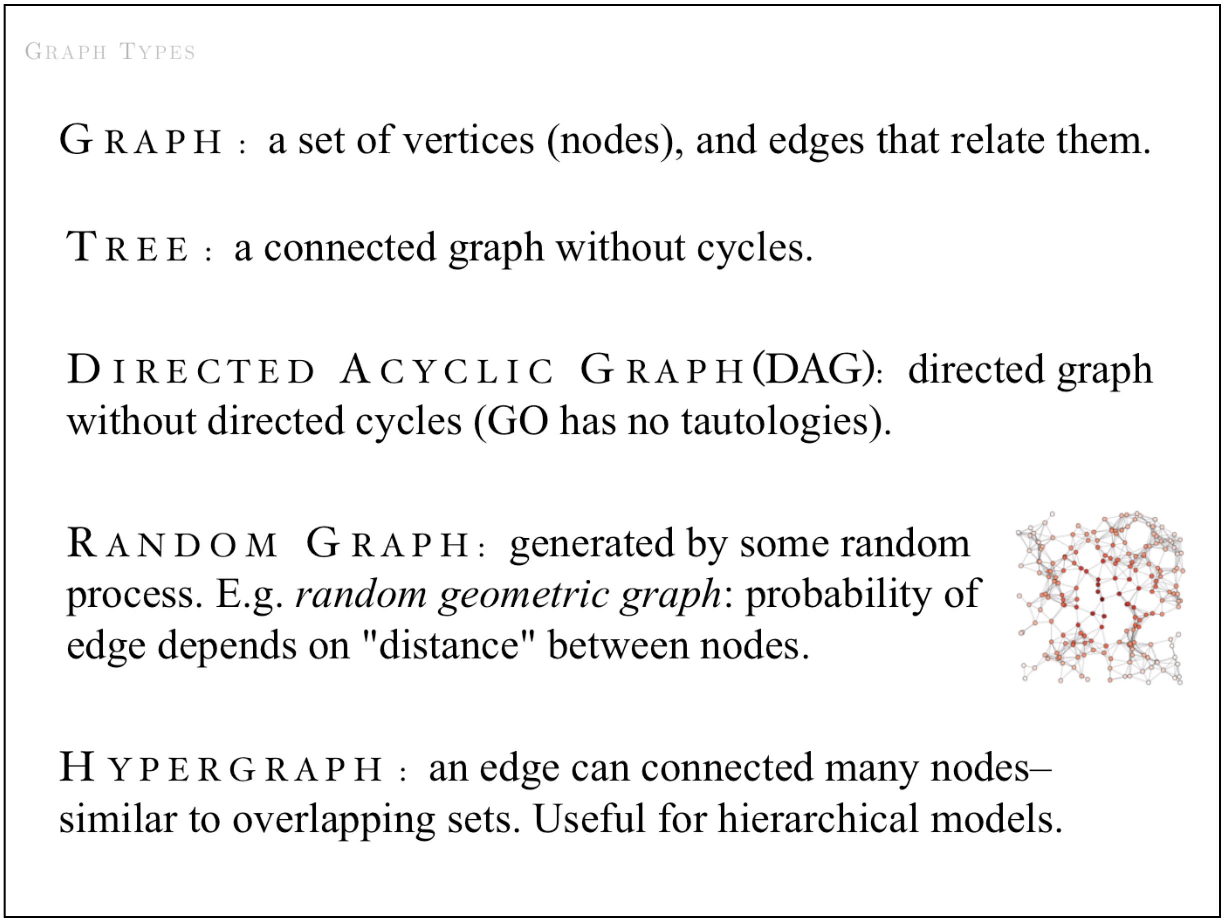
Entities: *-omics* technologies define the *entities* relevant to systems biology – genes, proteins, regulatory RNA – and their attributes of structure and function.

Relationships : To describe how these entities collaborate their *relationships* need to be defined as well.

Such relationships have formal aspects (direction, multiplicity...) and semantic aspects.

Entities and relationships map naturally into the objects of *Graph Theory*.

The quantitative analysis of interactions takes bioinformatics to the next higher dimension: we go from 1D to 2D with graph theory.



Graph Types

Graph : a set of vertices (nodes), and edges that relate them.

Tree : a connected graph without cycles.

Directed Acyclic Graph (DAG): directed graph without directed cycles (GO has no tautologies).

Random Graph: generated by some random process. E.g. *random geometric graph*: probability of edge depends on "distance" between nodes.

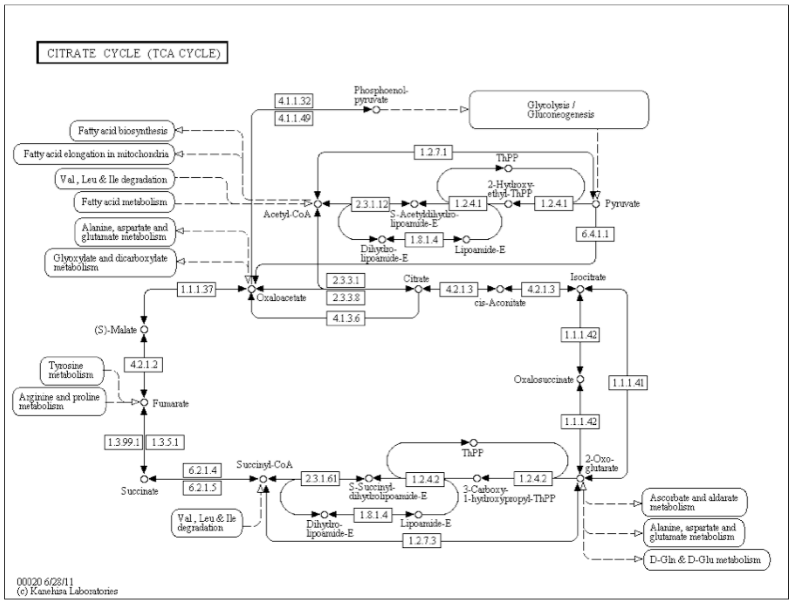
Hypergraph : an edge can connected many nodes– similar to overlapping sets. Useful for hierarchical models.

Definitions.

Biartite Graph (bigraph)

Metabolic Network:

A *bipartite* graph that contains metabolites, enzymes and reactions. Metabolites and enzymes are both nodes, but of a different class. Reactions are edges.

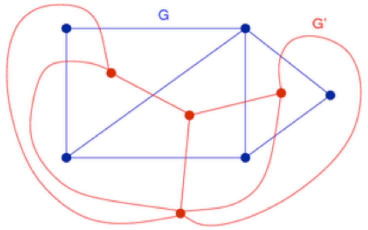


00020 6/28/11
(c) Kanehisa Laboratories

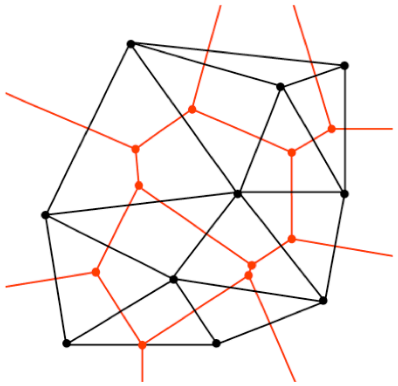
Distance Parity can be used to test for bipartiteness.

Dual Graph

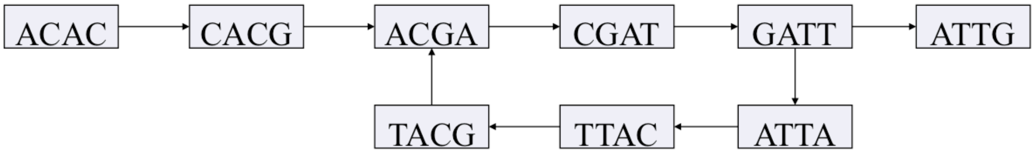
A Dual Graph of a planar graph *G* has a vertex for each plane region of *G*, and an edge for each edge in *G* joining two neighboring regions.



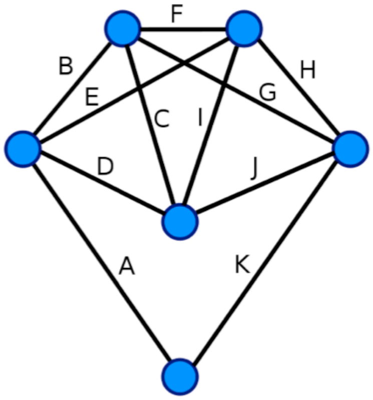
Application: Voronoi Tesselation and Delaunay Triangulation.



De Bruijn Graphs are directed, labeled graphs where each node is a sequence and each edge connects nodes whose sequences overlap with an offset of one character. This is useful for short-read sequence assembly.

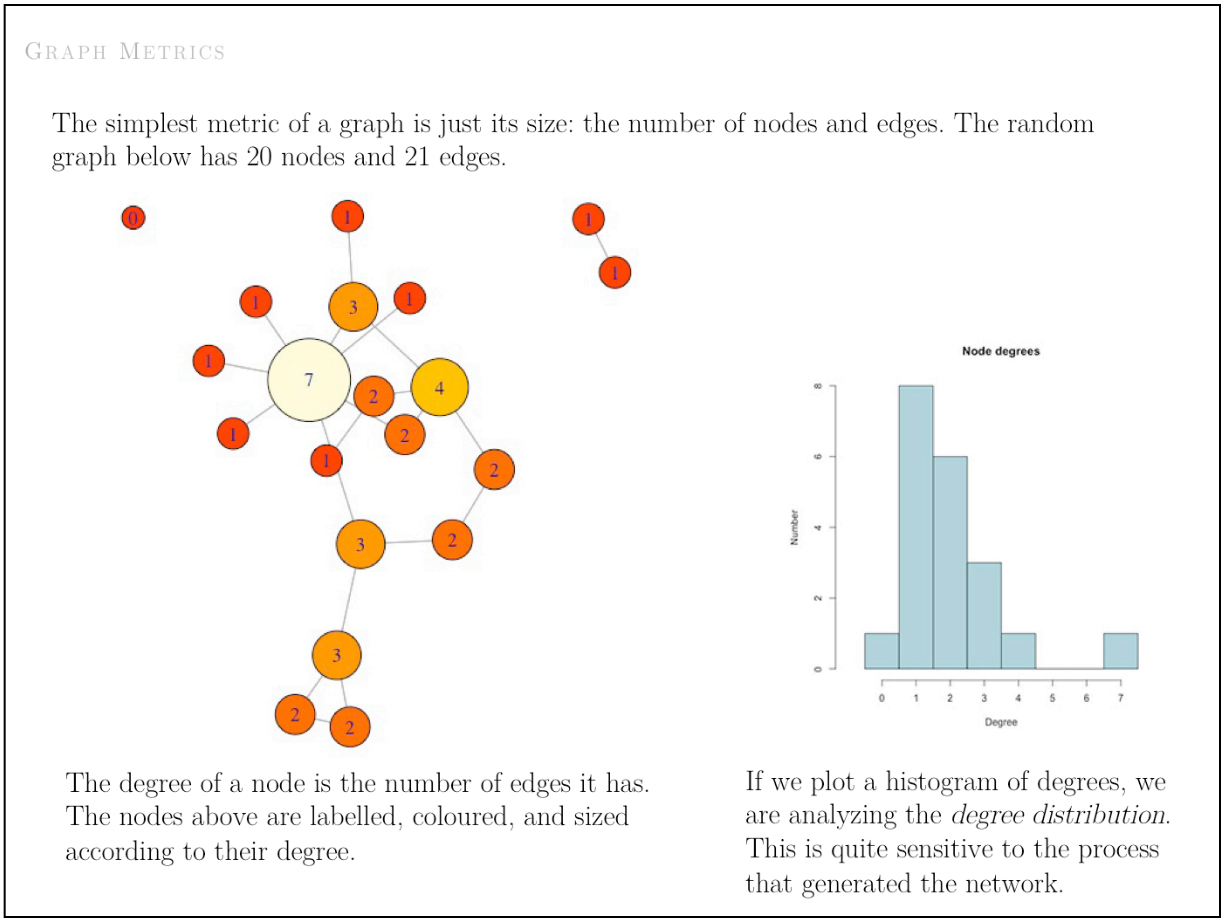


The Velvet Assembler uses De Bruijn graphs. Advantages include: reduced memory requirements, easy access of nodes through hash tables, easier treatment of repetitive sequences.



EULER CYCLE: visits every edge exactly once.

HAMILTON CYCLE: visits every node exactly once. cf. *Travelling Salesman Problem* – finding the shortest *Hamiltonian* Cycle is NP-hard, because all combinatorially many solutions have to be considerd



Graph Metrics

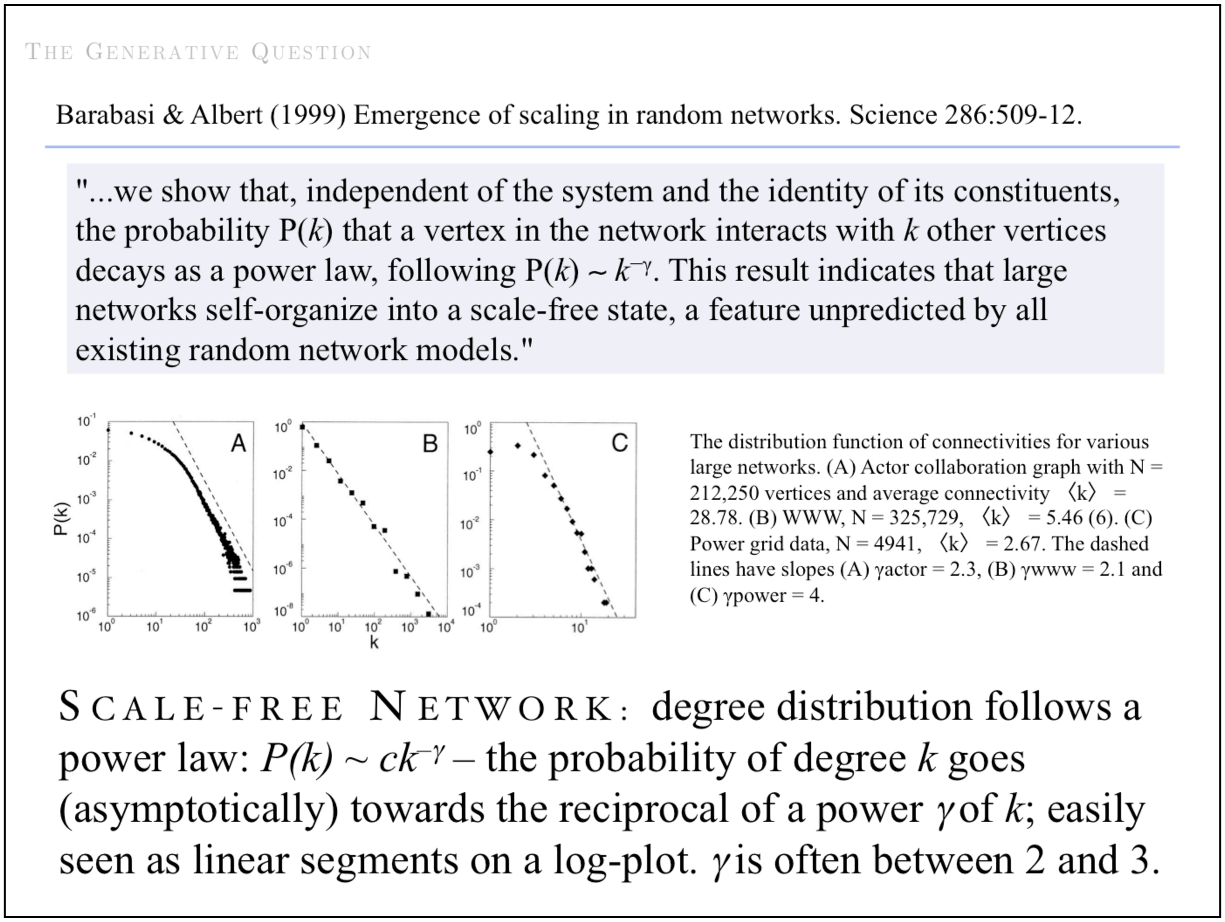
The simplest metric of a graph is just its size: the number of nodes and edges. The random graph below has 20 nodes and 21 edges.

The degree of a node is the number of edges it has. The nodes above are labelled, coloured, and sized according to their degree.

If we plot a histogram of degrees, we are analyzing the *degree distribution*. This is quite sensitive to the process that generated the network.

If we have a *directed graph*, a node may have incoming edges and outgoing edges. In that case, we distinguish between “in-degree” and “out-degree” of the node.

We can conceptualize that nodes with a high degree lie in the centre of a network, and that nodes with degree 1 constitute the boundary of the network. Therefore the degree of a node is also a *topological* measure of the graph, i.e. it describes the contribution of nodes to the overall “shape” of the graph: we call this interpretation *degree centrality*.



THE GENERATIVE QUESTION

Barabasi & Albert (1999) Emergence of scaling in random networks. Science 286:509-12.

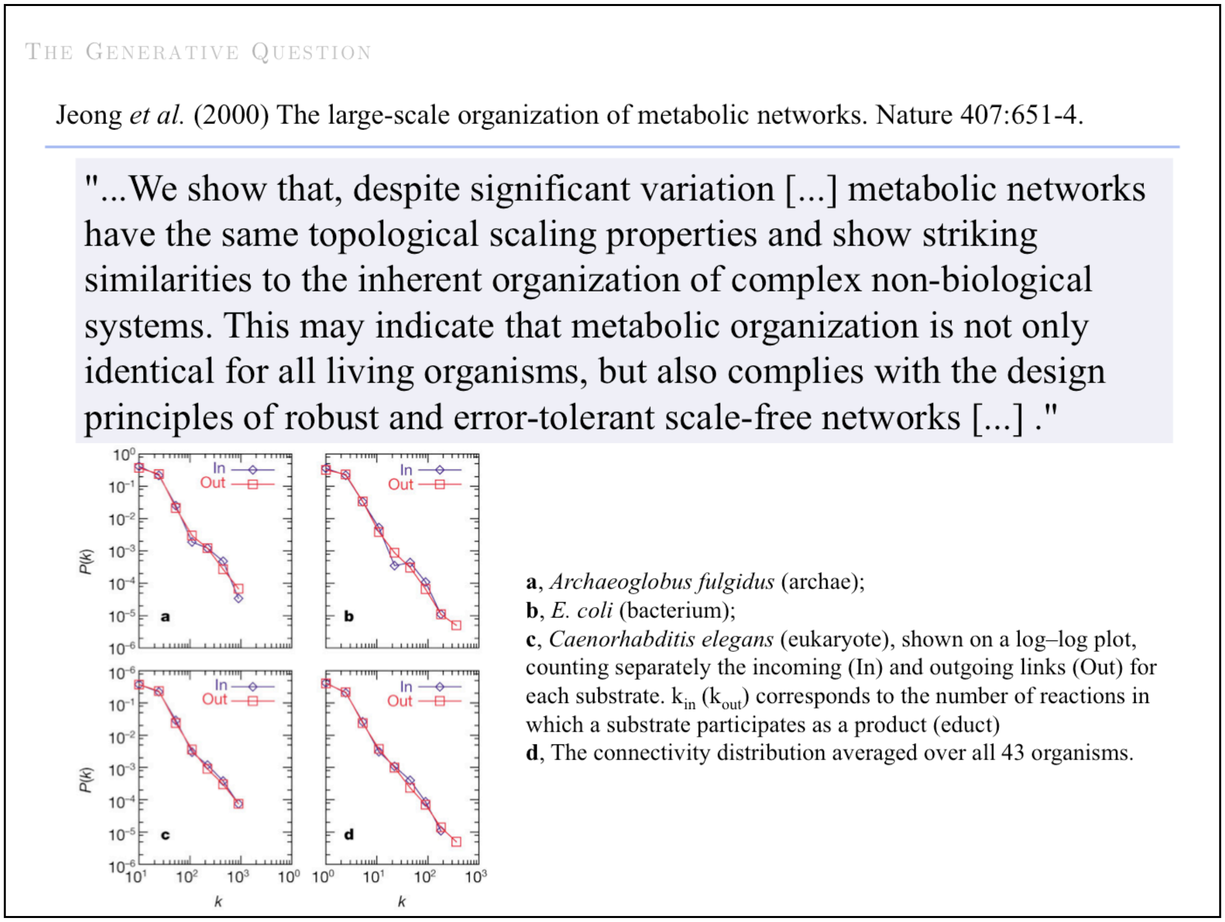
> "...we show that, independent of the system and the identity of its constituents, the probability P(*k*) that a vertex in the network interacts with *k* other vertices decays as a power law, following P(*k*) ~ $k^{-\gamma}$. This result indicates that large networks self-organize into a scale-free state, a feature unpredicted by all existing random network models."

The distribution function of connectivities for various large networks. (A) Actor collaboration graph with N = 212,250 vertices and average connectivity 〈k〉 = 28.78. (B) WWW, N = 325,729, 〈k〉 = 5.46 (6). (C) Power grid data, N = 4941, 〈k〉 = 2.67. The dashed lines have slopes (A) γactor = 2.3, (B) γwww = 2.1 and (C) γpower = 4.

SCALE-FREE NETWORK: degree distribution follows a power law: $P(k) \sim ck^{-\gamma}$ – the probability of degree *k* goes (asymptotically) towards the reciprocal of a power γ of *k*; easily seen as linear segments on a log-plot. γ is often between 2 and 3.

Degree distributions give us a way to reason about the circumstances that could have produced a network of interactions in the real world.

Barabasi and Albert showed scale-free properties for movie-actor networks, pages in the WWW, and the electric power grid.



THE GENERATIVE QUESTION

Jeong *et al.* (2000) The large-scale organization of metabolic networks. Nature 407:651-4.

> "...We show that, despite significant variation [...] metabolic networks have the same topological scaling properties and show striking similarities to the inherent organization of complex non-biological systems. This may indicate that metabolic organization is not only identical for all living organisms, but also complies with the design principles of robust and error-tolerant scale-free networks [...] ."

a, *Archaeoglobus fulgidus* (archae);
b, *E. coli* (bacterium);
c, *Caenorhabditis elegans* (eukaryote), shown on a log–log plot, counting separately the incoming (In) and outgoing links (Out) for each substrate. k_{in} (k_{out}) corresponds to the number of reactions in which a substrate participates as a product (educt)
d, The connectivity distribution averaged over all 43 organisms.

It was soon appreciated, that many biological networks also have scale-free properties.

This is not trivial, and begs the question what the WWW and metabolic networks could have in common with power-grid layouts and developmental signalling pathways.

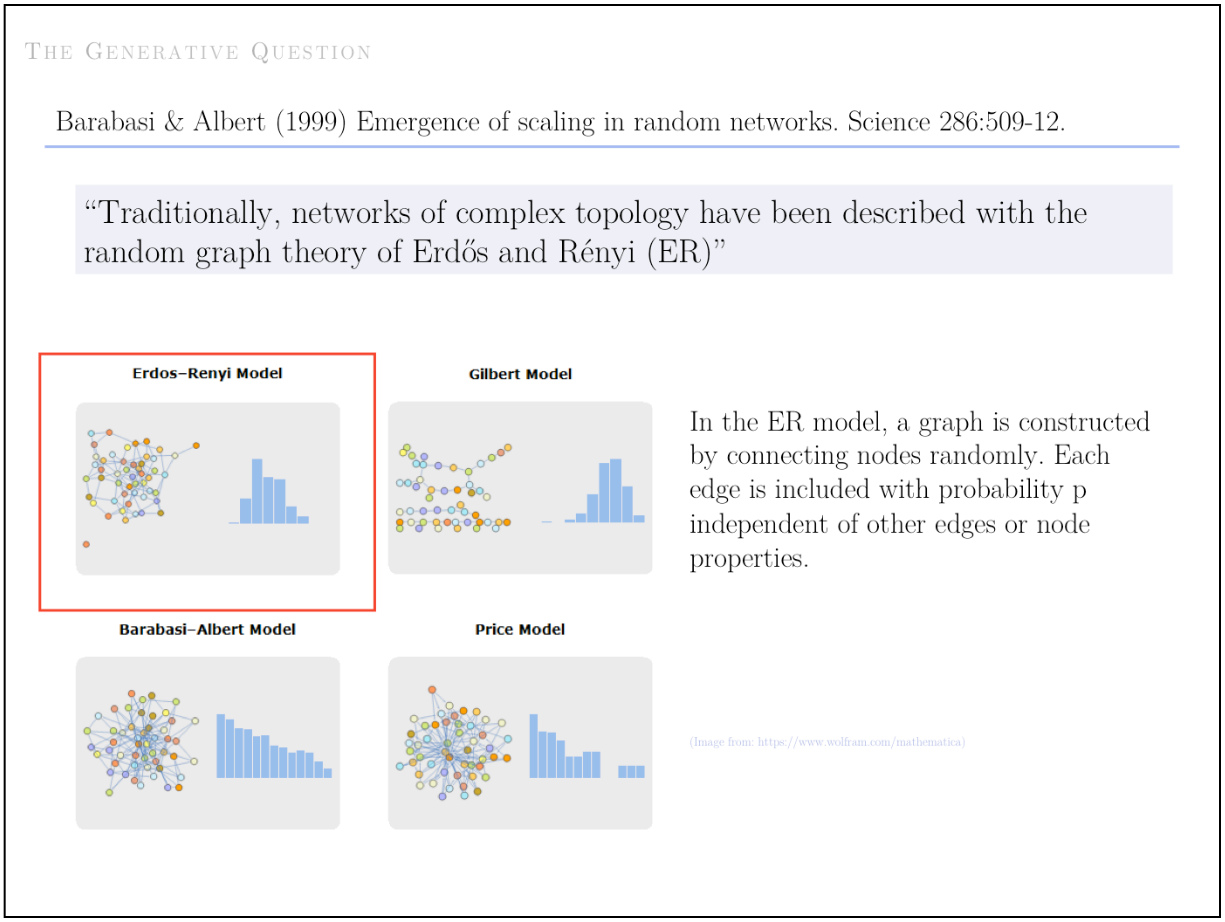
How is a Scale-free Network Generated in Biology?

This is the *crucial* question that connects the mathematics and biology of network analysis. If it has a non-trivial answer, it would shed light on the objective function of biological complexity!

Scale-free Network: degree distribution follows a power law: $P(k) \sim ck^{-\gamma}$ – the probability of degree k goes (asymptotically) towards the reciprocal of a power γ of k; easily seen as linear segments on a log-plot. γ is often between 2 and 3.

Small-world Network: distance between two randomly chosen nodes grows with the log of the network size:

$$D = a + b \log | V |$$



THE GENERATIVE QUESTION

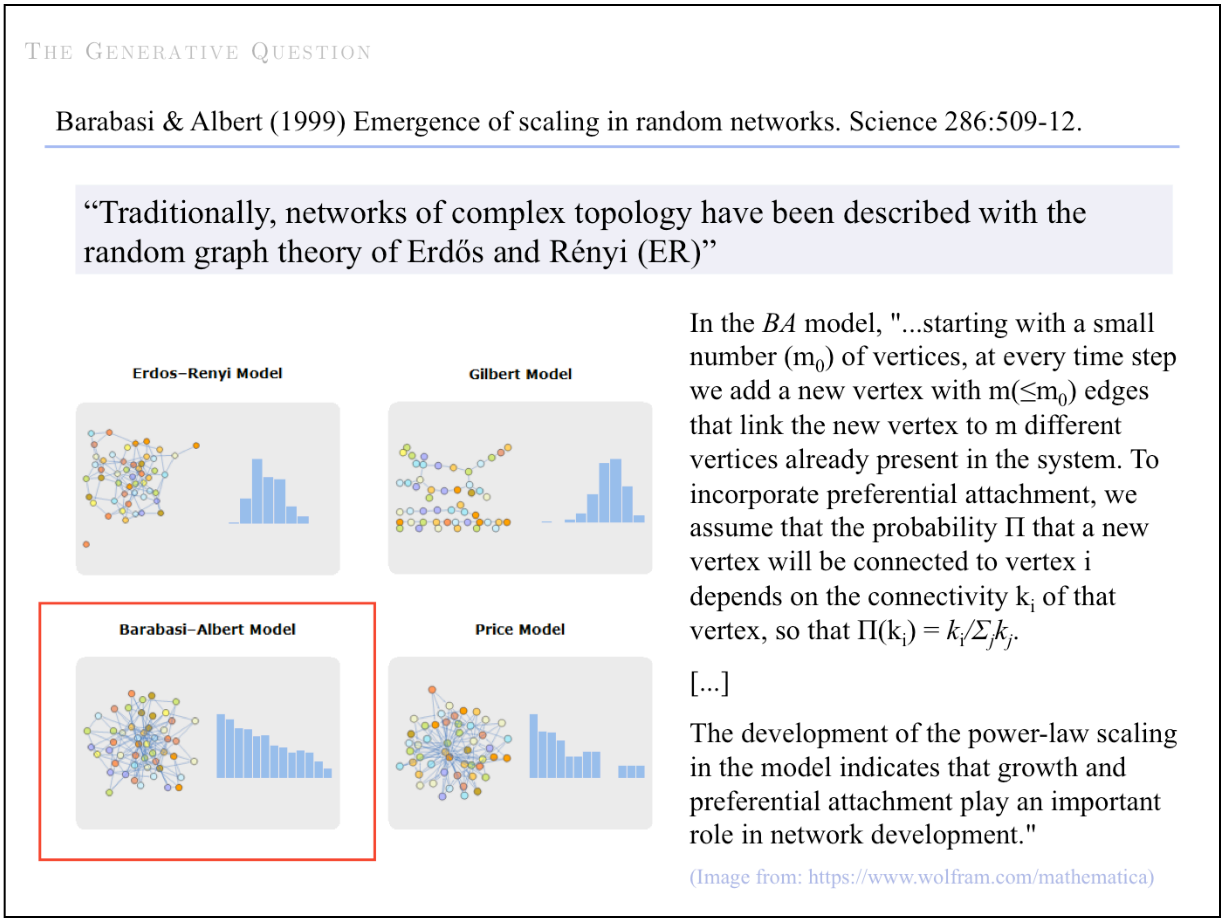
Barabasi & Albert (1999) Emergence of scaling in random networks. Science 286:509-12.

> “Traditionally, networks of complex topology have been described with the random graph theory of Erdős and Rényi (ER)”

In the ER model, a graph is constructed by connecting nodes randomly. Each edge is included with probability p independent of other edges or node properties.

(Image from: https://www.wolfram.com/mathematica)

Note that the degree-distribution histograms (frequency is on a log-scale) are very characteristic of the generative process.



THE GENERATIVE QUESTION

Barabasi & Albert (1999) Emergence of scaling in random networks. Science 286:509-12.

> “Traditionally, networks of complex topology have been described with the random graph theory of Erdős and Rényi (ER)”

Erdos–Renyi Model
Gilbert Model
Barabasi–Albert Model
Price Model

In the *BA* model, "...starting with a small number (m_0) of vertices, at every time step we add a new vertex with $m(\leq m_0)$ edges that link the new vertex to m different vertices already present in the system. To incorporate preferential attachment, we assume that the probability Π that a new vertex will be connected to vertex i depends on the connectivity k_i of that vertex, so that $\Pi(k_i) = k_i / \Sigma_j k_j$.

[...]

The development of the power-law scaling in the model indicates that growth and preferential attachment play an important role in network development."

(Image from: https://www.wolfram.com/mathematica)

The problem here is that it is not obvious why protein-protein interactions should be subject to preferential attachment. Nor is it entirely clear whether actual interaction graphs are indeed scale-free.

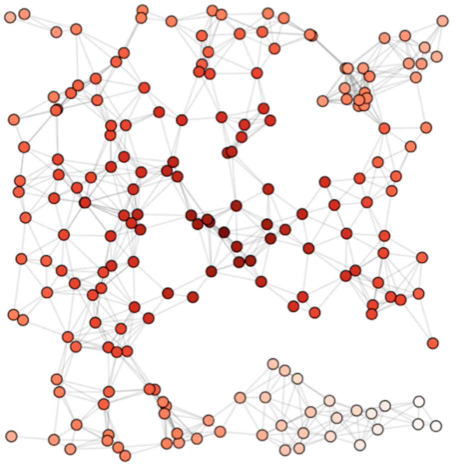
I do not agree with Barabasi and Albert that the model “*indicates that growth and preferential attachment play an important role in network development*”. This would be confusing correlation with causation. They have indeed shown that a model, based on preferential attachment, can reproduce characteristics of many networks. However we must keep an open mind about whether there are other mechanisms that also could give rise to scale-free properties. And indeed there are.

Alternative models include: the *copy model* that creates a scale-free distribution by adding new nodes through copying a fraction of the links of an existing node. Rewiring of random networks towards game-theoretic optimal objective functions also creates scale-free networks. *Hierarchical network models* are scale-free, as are *hyperbolic geometric graphs*. All of these have much more straightforward biological analogies than preferential attachment.

Besides, the actual networks of biology are not necessarily formed by **optimization** with a single mechanism towards a common objective function, but are the result of a messy, stochastic, vaguely conserved process of achieving **sufficiency** of purpose.

See also: https://en.wikipedia.org/wiki/Scale-free_network

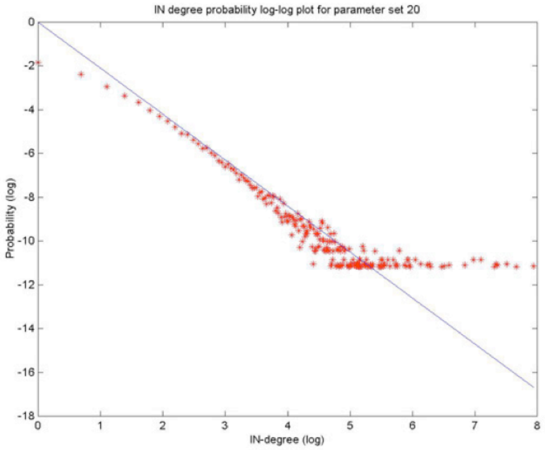
But: there are important alternatives.



In **random geometric graphs**, two nodes are connected if the distance between them is smaller than a threshold.

Distance should be seen as a generalizable property - this could be distance in time post activation signal or cell-cycle, distance in space or compartment, distance in a separate graph of metabolic or signalling nodes *etc*.

But: there are important alternatives.

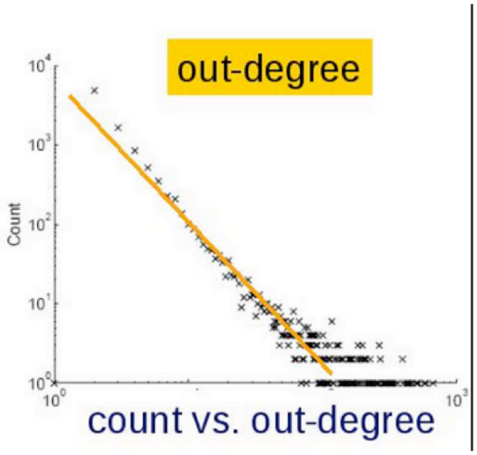


Copying models ...

Add a node and choose the number of edges to add. Choose a random node and copy its edges.

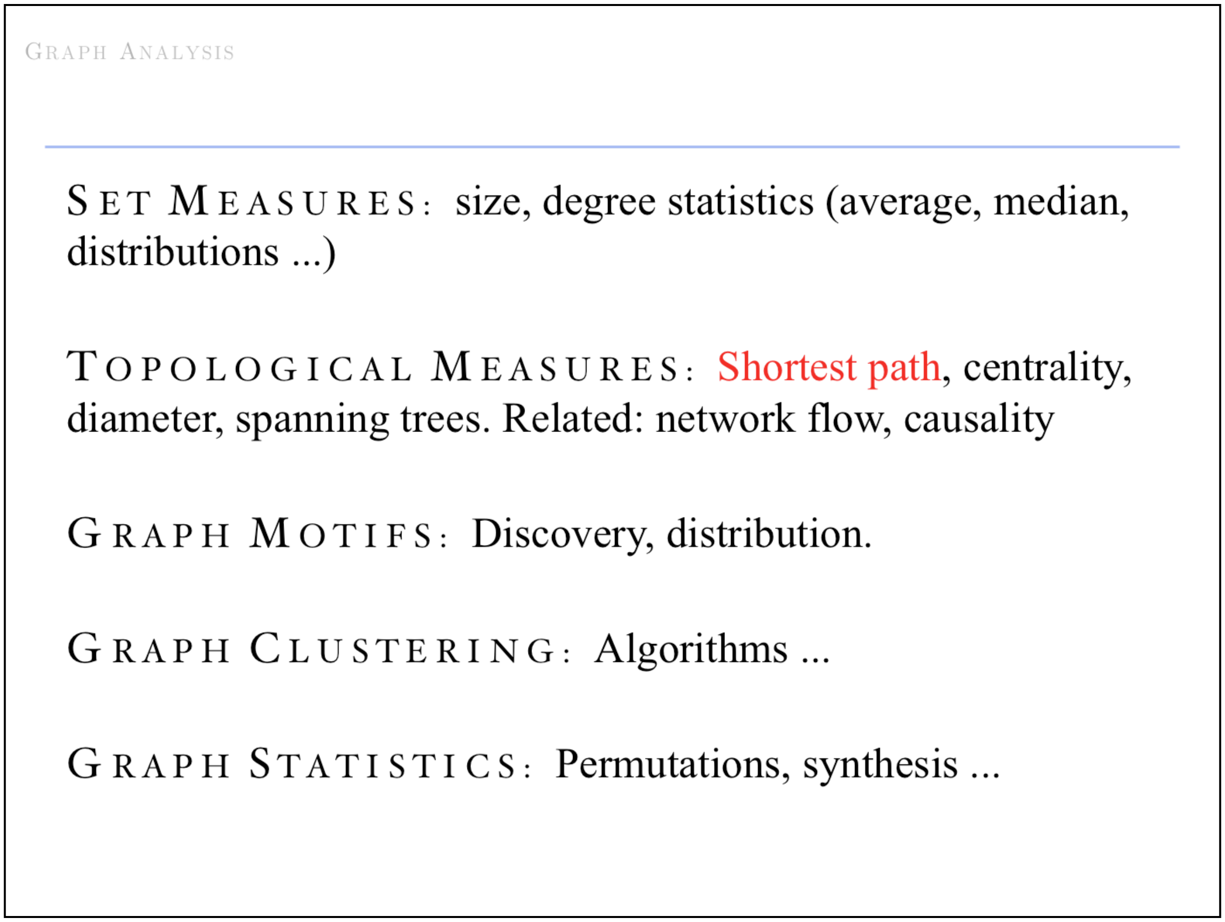
This generates scale free **and** community structures

But: there are important alternatives.



Forest fire model graphs are generated from a cellular automaton:

- A burning cell turns into an empty cell.
- A cell will burn if at least one neighbor is burning.
- A cell ignites with probability f even if no neighbor is burning
- An empty cell fills up with probability p.



Graph Analysis

SET MEASURES: size, degree statistics (average, median, distributions ...)

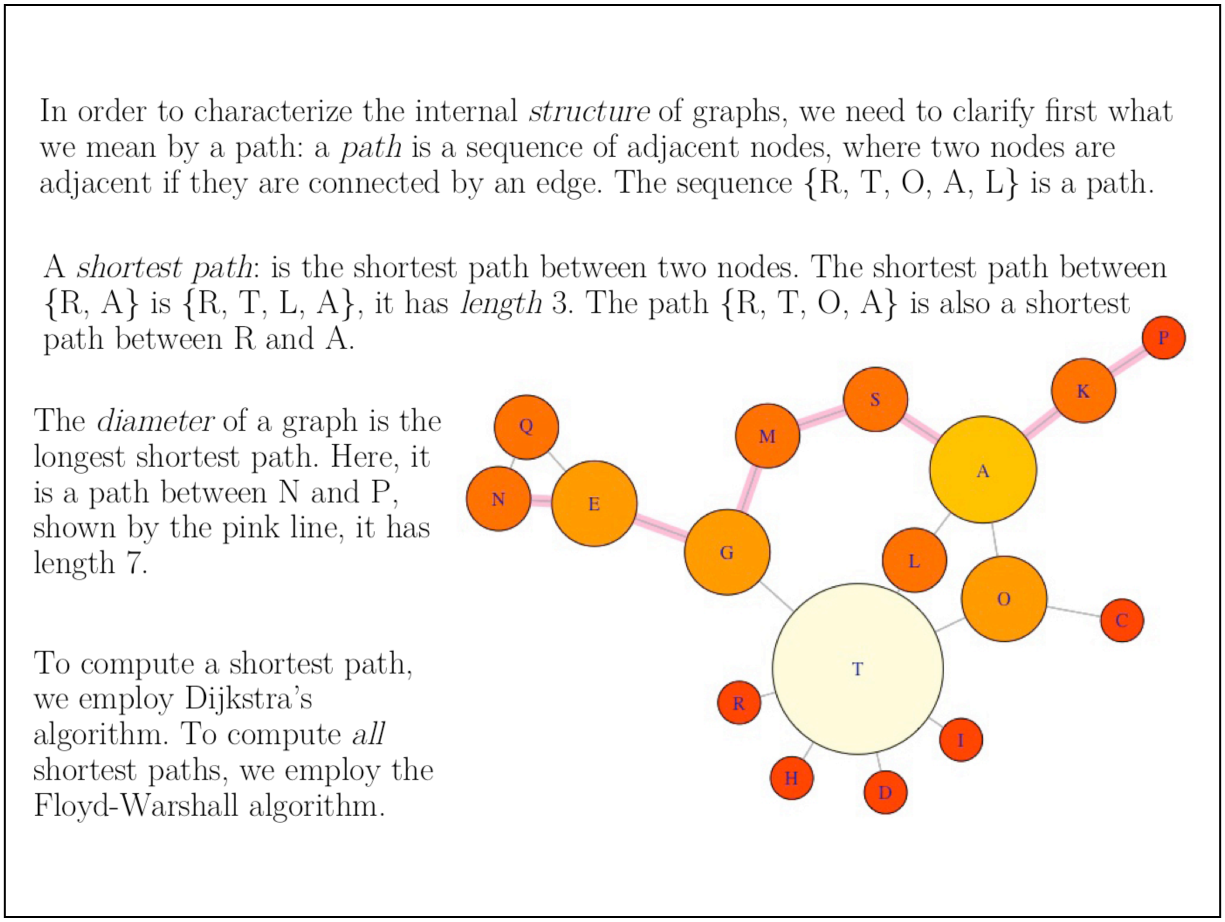
TOPOLOGICAL MEASURES: Shortest path, centrality, diameter, spanning trees. Related: network flow, causality

GRAPH MOTIFS: Discovery, distribution.

GRAPH CLUSTERING: Algorithms ...

GRAPH STATISTICS: Permutations, synthesis ...

Besides degree distributions, there are several other approaches to the quantitative analysis of biological networks.



In order to characterize the internal *structure* of graphs, we need to clarify first what we mean by a path: a *path* is a sequence of adjacent nodes, where two nodes are adjacent if they are connected by an edge. The sequence {R, T, O, A, L} is a path.

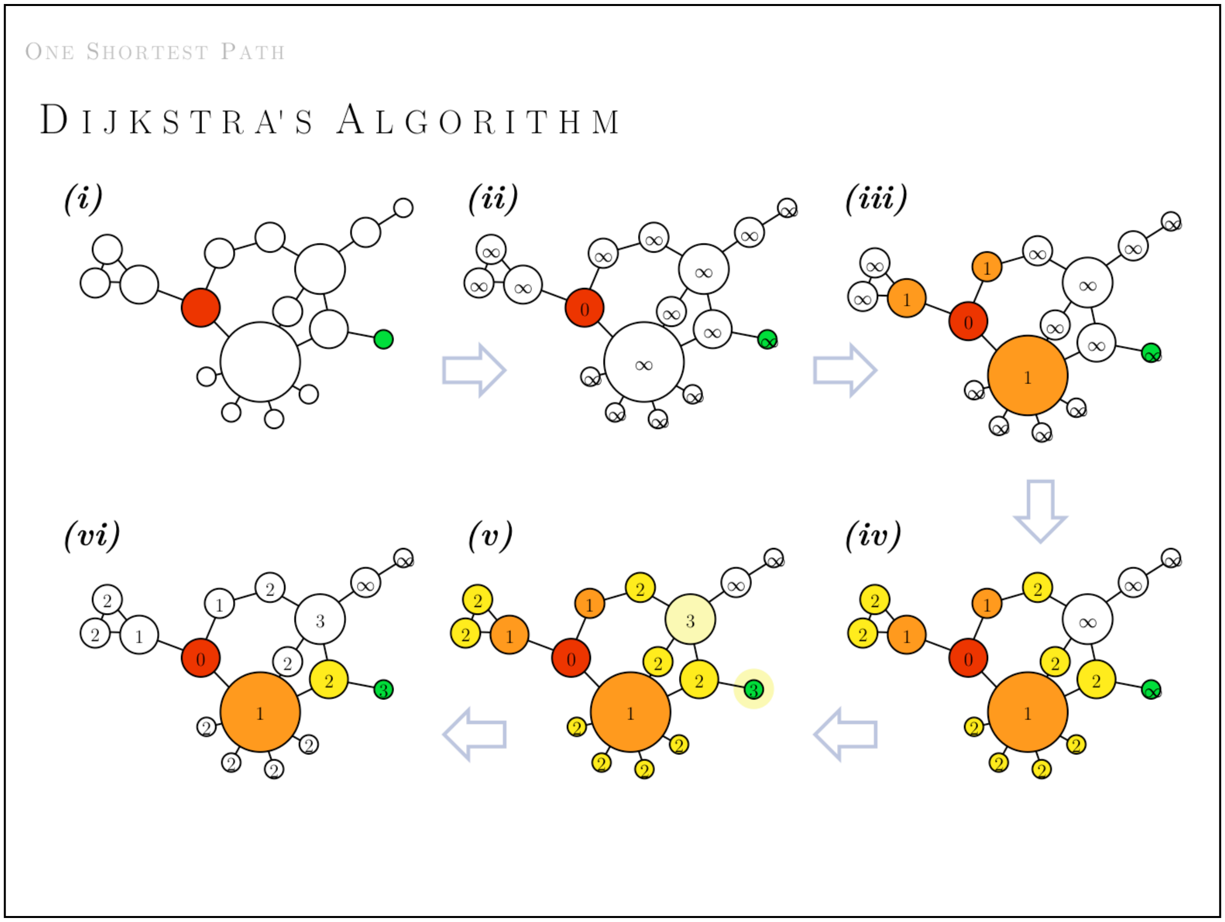
A *shortest path*: is the shortest path between two nodes. The shortest path between {R, A} is {R, T, L, A}, it has *length* 3. The path {R, T, O, A} is also a shortest path between R and A.

The *diameter* of a graph is the longest shortest path. Here, it is a path between N and P, shown by the pink line, it has length 7.

To compute a shortest path, we employ Dijkstra's algorithm. To compute *all* shortest paths, we employ the Floyd-Warshall algorithm.

How many diameters does this graph have?

I say: six.



ONE SHORTEST PATH

DIJKSTRA'S ALGORITHM

Dijkstra's algortihm finds a shortest path between two nodes.

(*i*) Suppose we want to find the shortest path between the red ("origin") and the green ("target") node. (*ii*) We set the distance of the origin to 0, all other distances to ∞. The origin is our "current node".

(*iii*) Next we collect all neighbors of the current node, and set their distance to one-more than the distance of the current node.

(*iv*) We repeat what we did in (*iii*) by considering all neighbors in turn. However, we never consider a node that we visited previously, i.e. we include only nodes whose distance is ∞.

(*v*) We repeat what we did in (*iii*) one more time, This is a loop. Lo-and-behold, this time we encounter the target node.

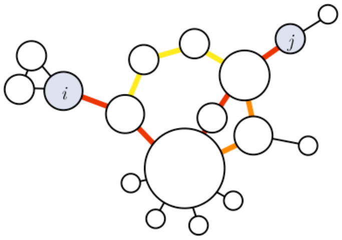
(*vi*) From the target node, we look for a node whose distance is one-less and add it to the shortest path. We repeat this, always going to a node closer to the origin, and adding that to the path. This is called "backtracking". Once we have reached the origin, the shortest path is defined.

This is generally considered an O(|V|^2) algorithm, but the implementation based on a min-priority queue implemented by a Fibonacci heap runs in O(|E|+|V|\log |V|)

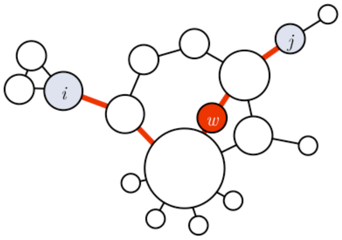
Fredman ML and Tarjan RE. (1984). Fibonacci heaps and their uses in improved network optimization algorithms. 25th Annual Symposium on Foundations of Computer Science. IEEE. pp. 338–346.

Betweenness centrality is a useful *topological* measure on graphs. It is high for nodes that lie between many other nodes, i.e. on the shortest path that connects these. You can think of nodes with high betweenness centrality to constitute bottlenecks in the connections between other nodes.

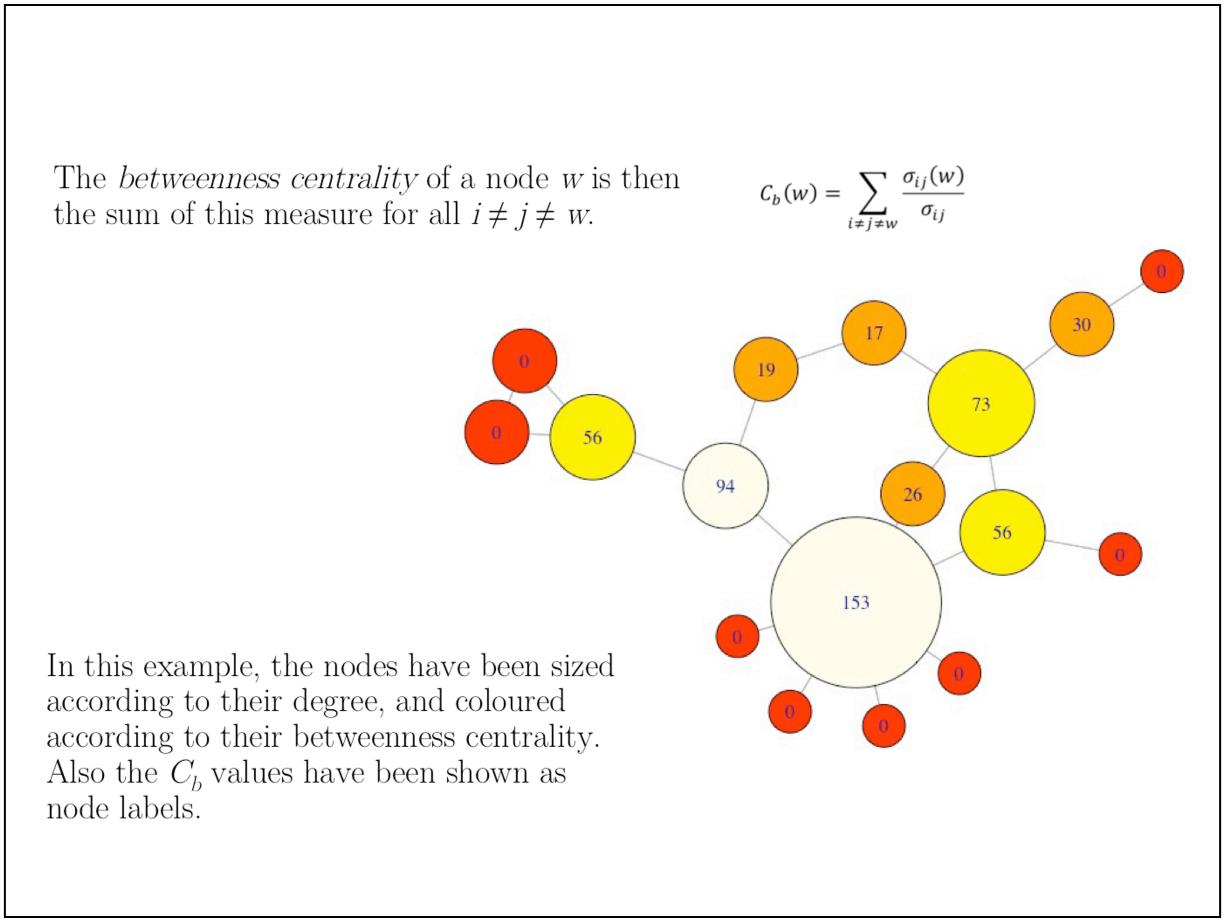
We define $\boldsymbol{\sigma}_{ij}$ to be the set of all shortest paths between nodes i and j. In the example below there are three.



We define $\boldsymbol{\sigma}_{ij}(w)$ to be the set of all shortest paths between nodes i and j that pass through node w. In the example below this is one path.



Then the ratio $\frac{\sigma_{ij}(w)}{\sigma_{ij}}$ measures what fraction of shortest paths between i and j pass through w. In our example this is 1/3.



The *betweenness centrality* of a node w is then the sum of this measure for all $i \neq j \neq w$.

$$C_b(w) = \sum_{i \neq j \neq w} \frac{\sigma_{ij}(w)}{\sigma_{ij}}$$

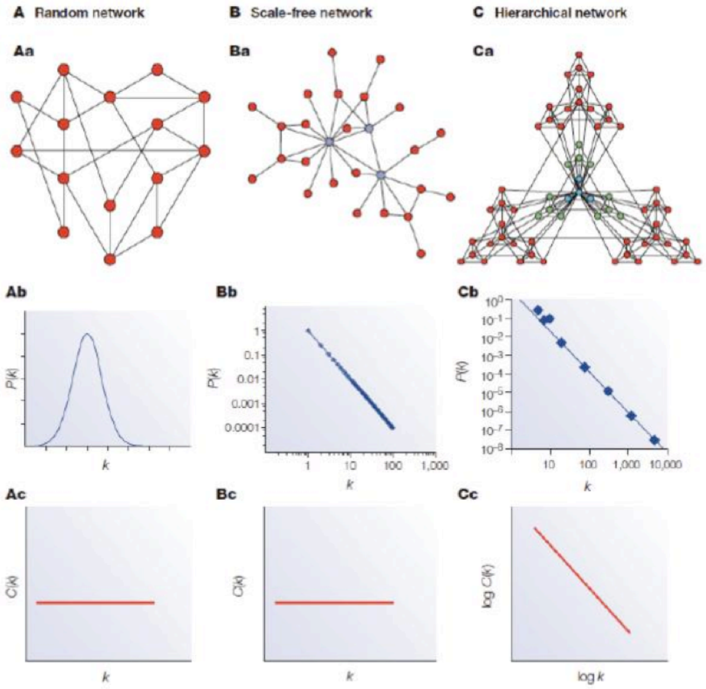
In this example, the nodes have been sized according to their degree, and coloured according to their betweenness centrality. Also the C_b values have been shown as node labels.

Note that in general, betweeness centrality and degree centrality go in the same direction – nodes with high degree tend to have high C_b values. *B*ut the relationship is not absolute – e.g. the second highest C_b value, 94, is in a node with degree 3, like two other nodes who only have a C_b of 56. This node also has a higher C_b than the node with C_b 73, which has a degree of 4.

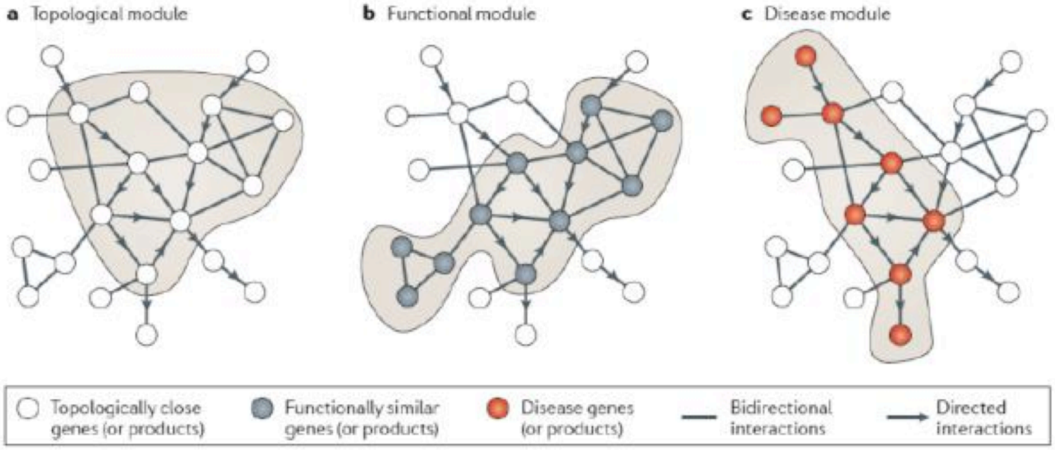
Stress centrality is a related concept: the stress centrality of a node w is the number of shortest paths passing through w.

Properties of protein-protein interaction networks

The properties of random networks (A), scale-free networks (B), and hierarchical networks (C). Protein networks are also called scale free, because it is not possible to define a meaningful average node in these networks. Plotting the degree k of nodes in protein interaction networks against the probability of observing that degree P(k), follows a power law (Bb). In these networks the clustering coefficient C(k) does not change as the function of the nodes degree (Bc), meaning that nodes with few interactions and a lot of interactions alike tend to participate in highly connected topological modules in the network. These properties are different for random networks (Aa, Ab, Ac) where edges are randomly distributed across nodes, and hierarchical networks (Ca, Cb, Cc), where clusters are united in an iterative manner.



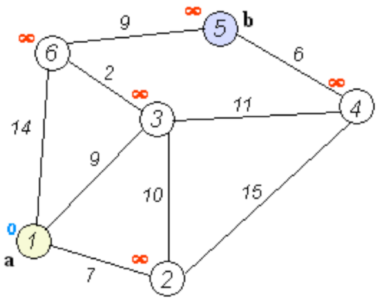
Analysis


a Topological module
b Functional module
c Disease module
Topologically close genes (or products)
Functionally similar genes (or products)
Disease genes (or products)
Bidirectional interactions
Directed interactions

Protein interaction networks have topological modules in which proteins are more connected to each other than to the reset of the network (a). These represent genes in the same pathways, molecular machines, or rigid architectural structures, i.e., functional modules (b). This has implications for human disease biology, as genes involved in the same disease tend to fall into the same clusters or functional modules. Modules enriched for genes from a particular disease are termed disease modules (c).

FLOYD-WARSHALL ALGORITHM: …

```
let dist be a |V| × |V| array of minimum distances initialized to ∞ (infinity)
for each vertex v
    dist[v][v] ← 0
for each edge (u,v)
    dist[u][v] ← w(u,v)  // the weight of the edge (u,v)
for k from 1 to |V|
    for i from 1 to |V|
        for j from 1 to |V|
            if dist[i][j] > dist[i][k] + dist[k][j]
                dist[i][j] ← dist[i][k] + dist[k][j]
            end if
```



$\Theta(|V|^3)$

http://steipe.biochemistry.utoronto.ca/abc

BORIS.STEIPE@UTORONTO.CA

DEPARTMENT OF BIOCHEMISTRY & DEPARTMENT OF MOLECULAR GENETICS
UNIVERSITY OF TORONTO, CANADA