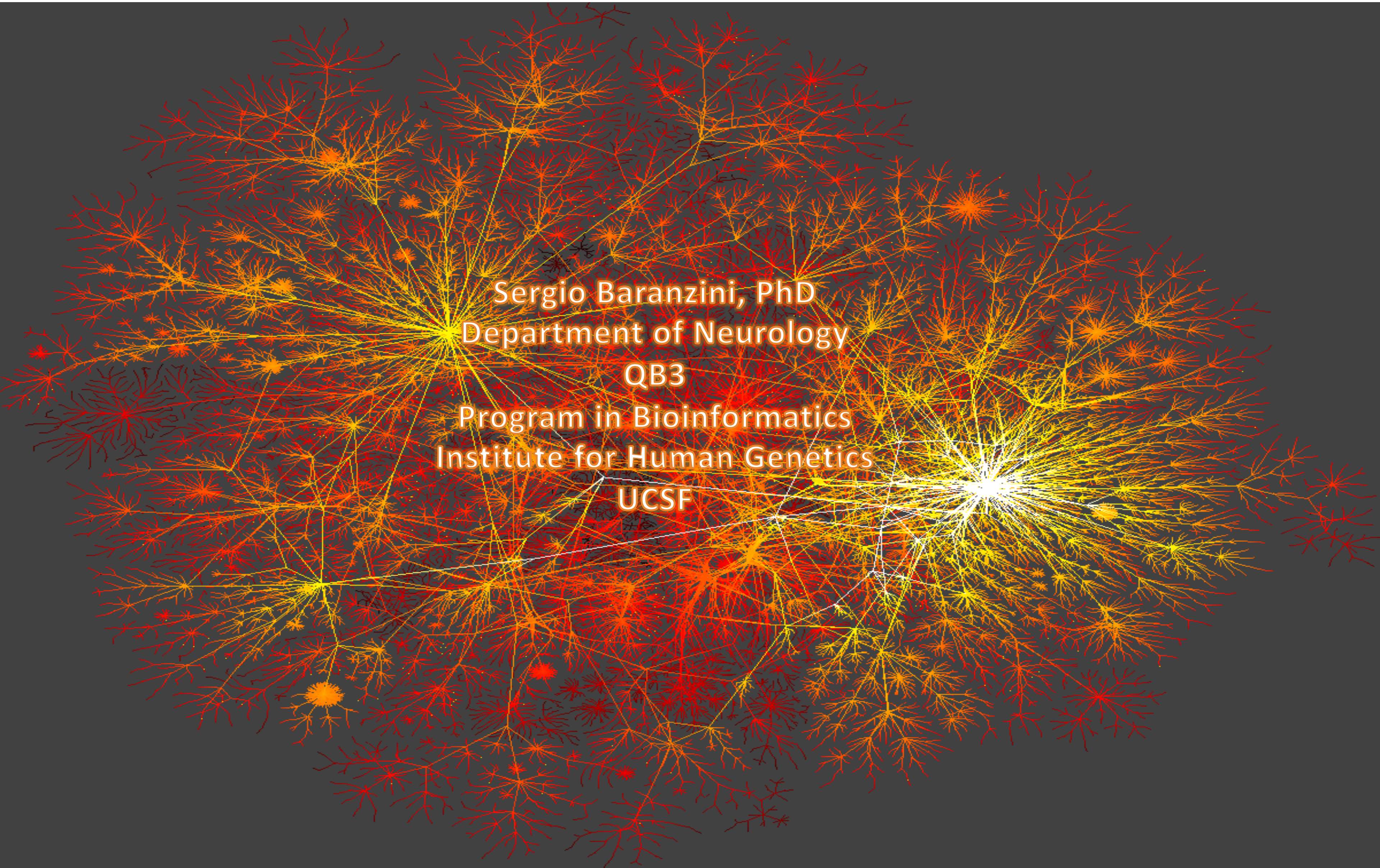
Introduction to network science

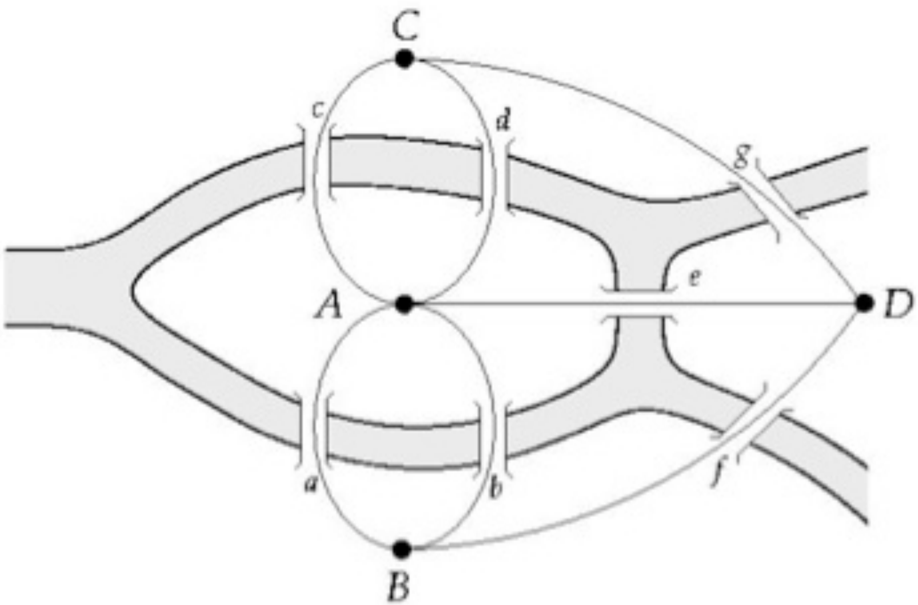


Sergio Baranzini, PhD
Department of Neurology
QB3
Program in Bioinformatics
Institute for Human Genetics
UCSF

The Bridges of Konigsberg

Can one walk across the seven bridges and never cross the same bridge twice?

The problem as a graph



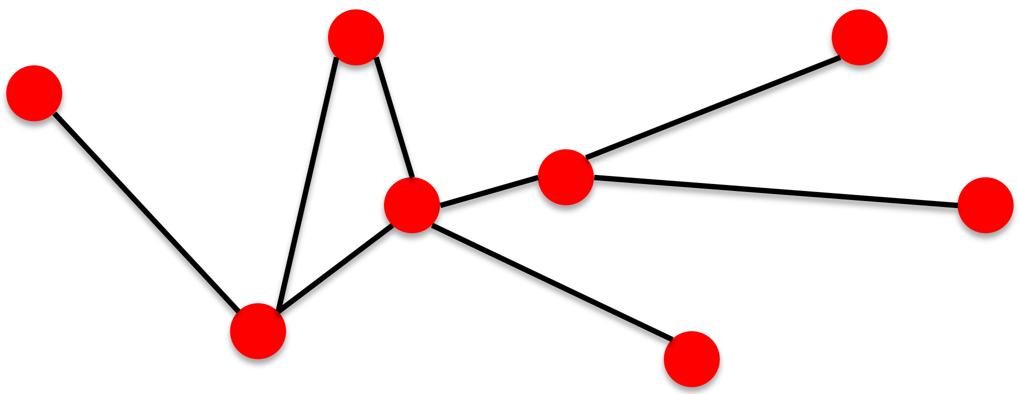
Can one walk across the seven bridges and never cross the same bridge twice?

1735: **Euler's theorem:**

(a) If a graph has more than two nodes of odd degree, there is no path.

(b) If a graph is connected and has no odd degree nodes, it has at least one path.

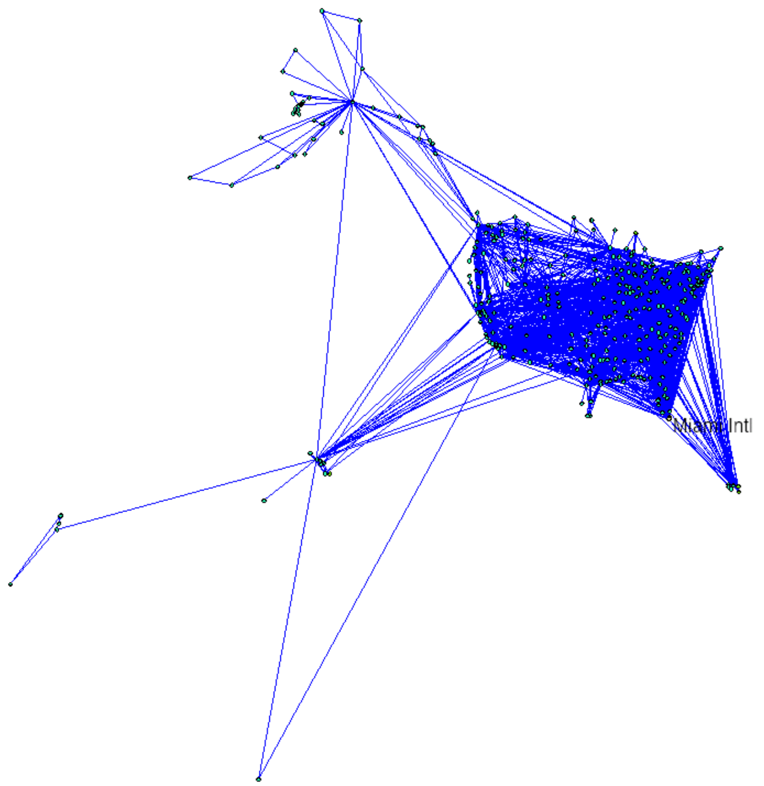
Networks as complex systems



- **components**: nodes, vertices N
- **interactions**: links, edges L
- **system**: network, graph (N,L)

Examples of real-life networks

Social networks
- -connections among people
- -trade among organizations, countries
- -citation networks
- -computer networks
- -telephone calls



- Organic molecules in chemistry
- Genes and proteins in biology
- Connections among words in text
- Transportation (airlines, streets, electric networks, etc)

Types of networks

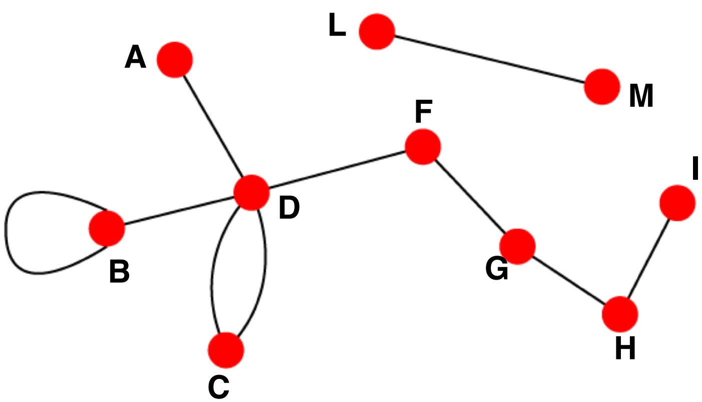
- Directed vs undirected
- Random vs scale-free
- Homogeneous vs bi-partite vs heterogeneous

Undirected vs directed networks

Undirected

Links: undirected (*symmetrical*)

Graph:

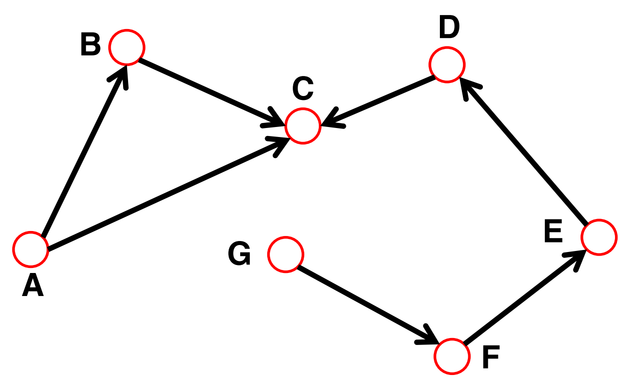


Undirected links :
coauthorship links
Actor network
protein interactions

Directed

Links: directed (*arcs*).

Digraph = directed graph:



An undirected link is the superposition of two opposite directed links.

Directed links :
URLs on the www
phone calls
metabolic reactions

Network topology metrics

- Degree (k) and distribution
- Path length
- Clustering Coefficient
- Eccentricity
- Radius
- Diameter
- Centrality
  - Closeness
  - betweenness

- **Setup in R**
  - Install and load SNA package in R
  - Create a test graph (10 nodes, edges generated randomly)

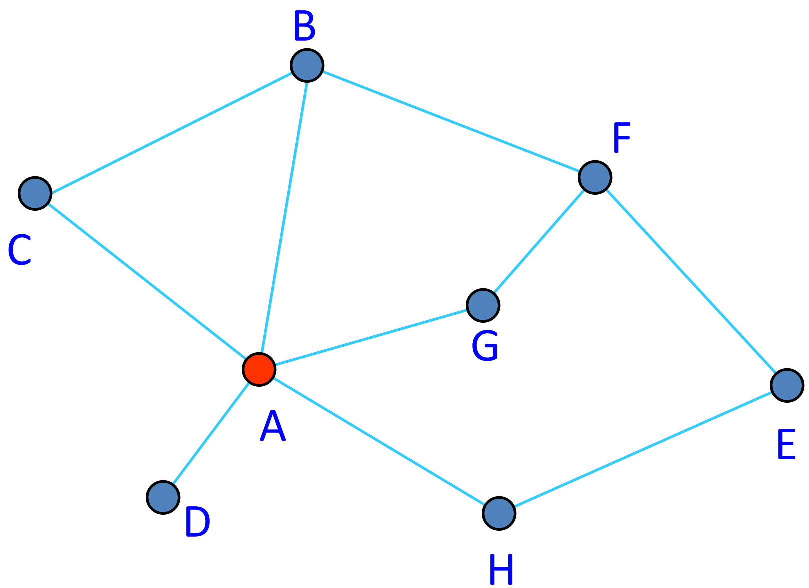
```
> #Load the sna(social network analysis) library
> library(sna)
> #Parameters required for the graph
> #N(number of vertices in the graph)
> #plink(probability of a link between any 2 vertices)
> N=10
> plink=0.2
> #sna::rgraph() -- Generate Bernoulli Random Graphs
> #2nd argument(1) for one graph is to generated
> #4th argument("graph") for the graph to be undirected
> #5th argument(FALSE) for the possibility of loops
> graph=rgraph(N,1,plink,"graph",FALSE)
> #generated graph in a matrix format
> graph
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,] 0    1    1    1    0    0    0    0    0    0
[2,] 1    0    0    0    0    0    0    0    0    0
[3,] 1    0    0    0    0    0    0    0    1    0
[4,] 1    0    0    0    1    1    0    0    0    0
[5,] 0    0    0    1    0    0    1    0    0    0
[6,] 0    0    0    1    0    0    0    0    1    0
[7,] 0    0    0    0    1    0    0    0    0    0
[8,] 0    0    0    0    0    0    0    0    1    0
[9,] 0    0    1    0    0    1    0    1    0    0
[10,]0    0    0    0    0    0    0    0    0    0
```

gplot (graph) for visualization

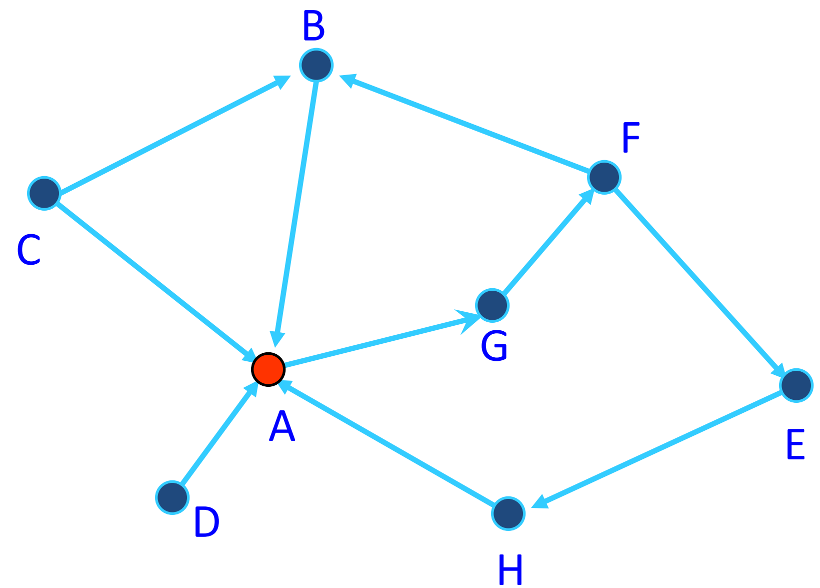
Degree

```
> degree(graph)
[1] 6 2 4 6 4 4 2 2 6 0
```

Undirected



Directed

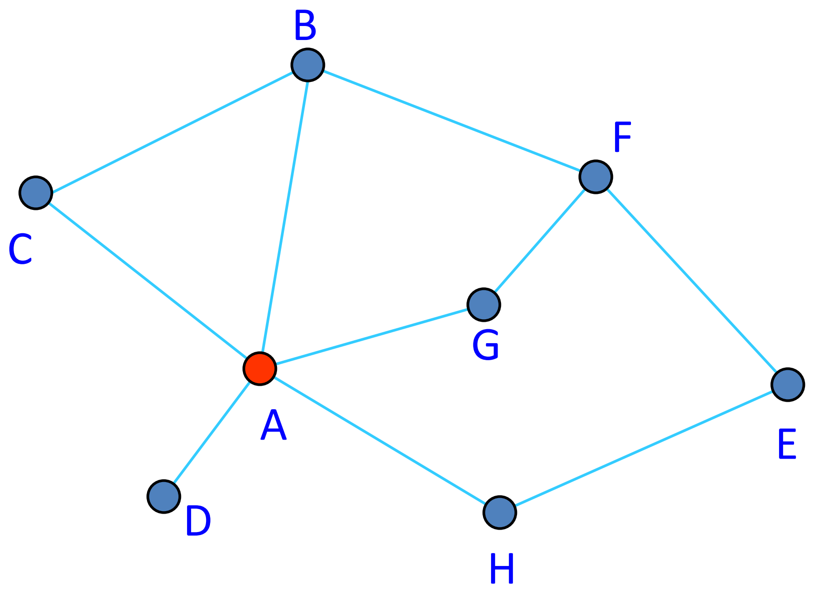


$k_A=5$

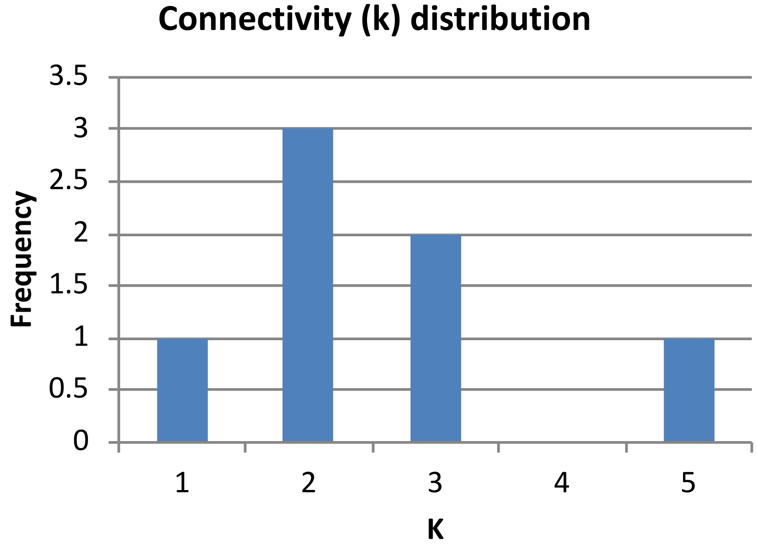
$k_{Ain}=5$

$k_{Aout}=1$

Degree Distribution

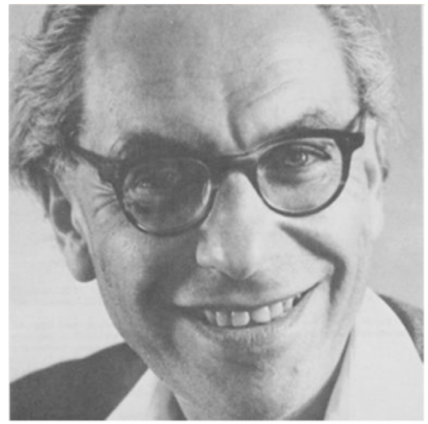


Node	k
A	5
B	3
F	3
C	2
E	2
G	2
D	1

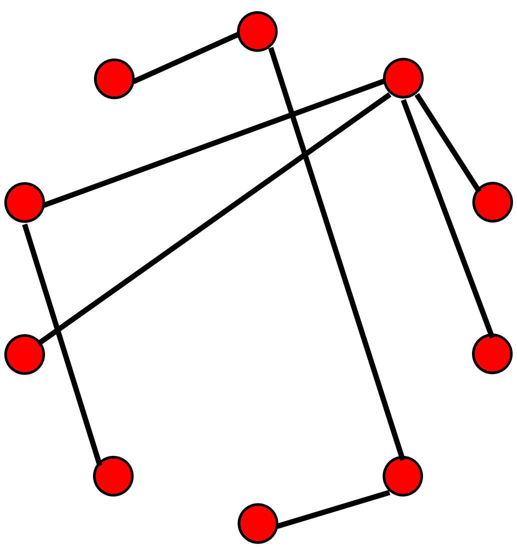


Random network model

Pál Erdös
(1913-1996)



Alfréd Rényi
(1921-1970)



Erdös-Rényi model (1960)

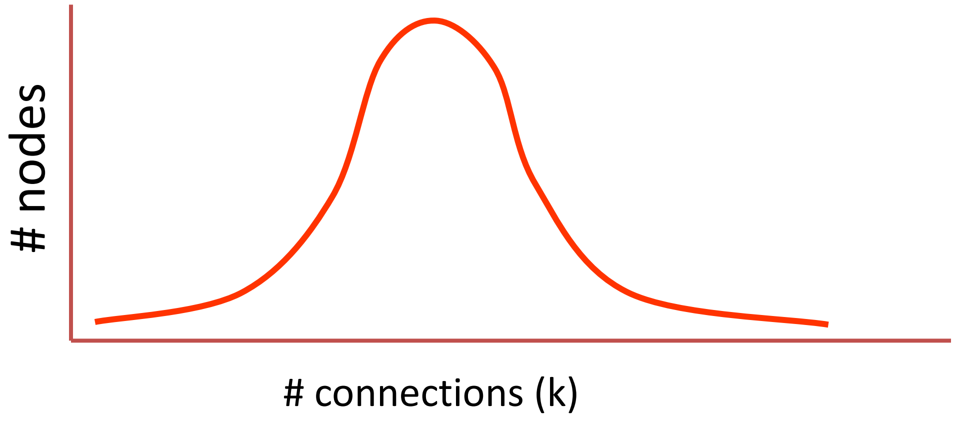
Connect with probability p

$p = 1/6 \quad N = 10$

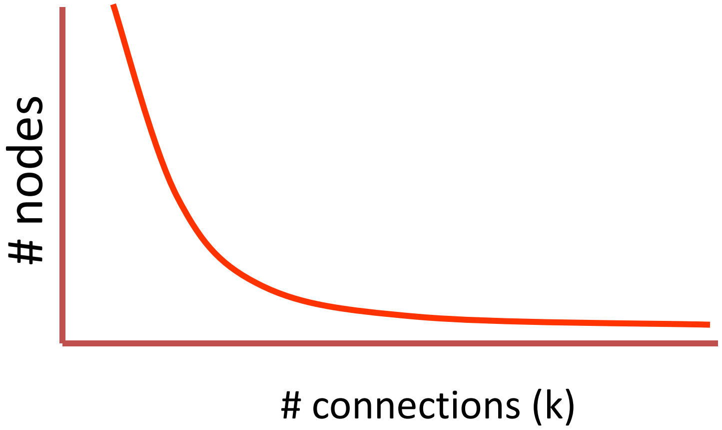
$\langle k \rangle \sim 1.5$

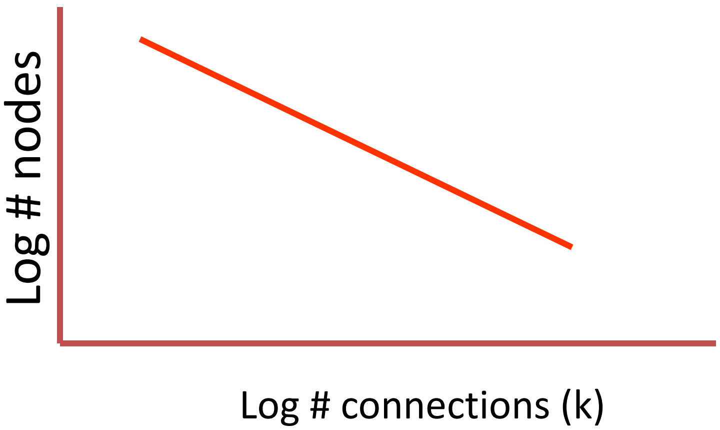
Random vs scale-free

- E-R: connectivity per node follows normal distribution

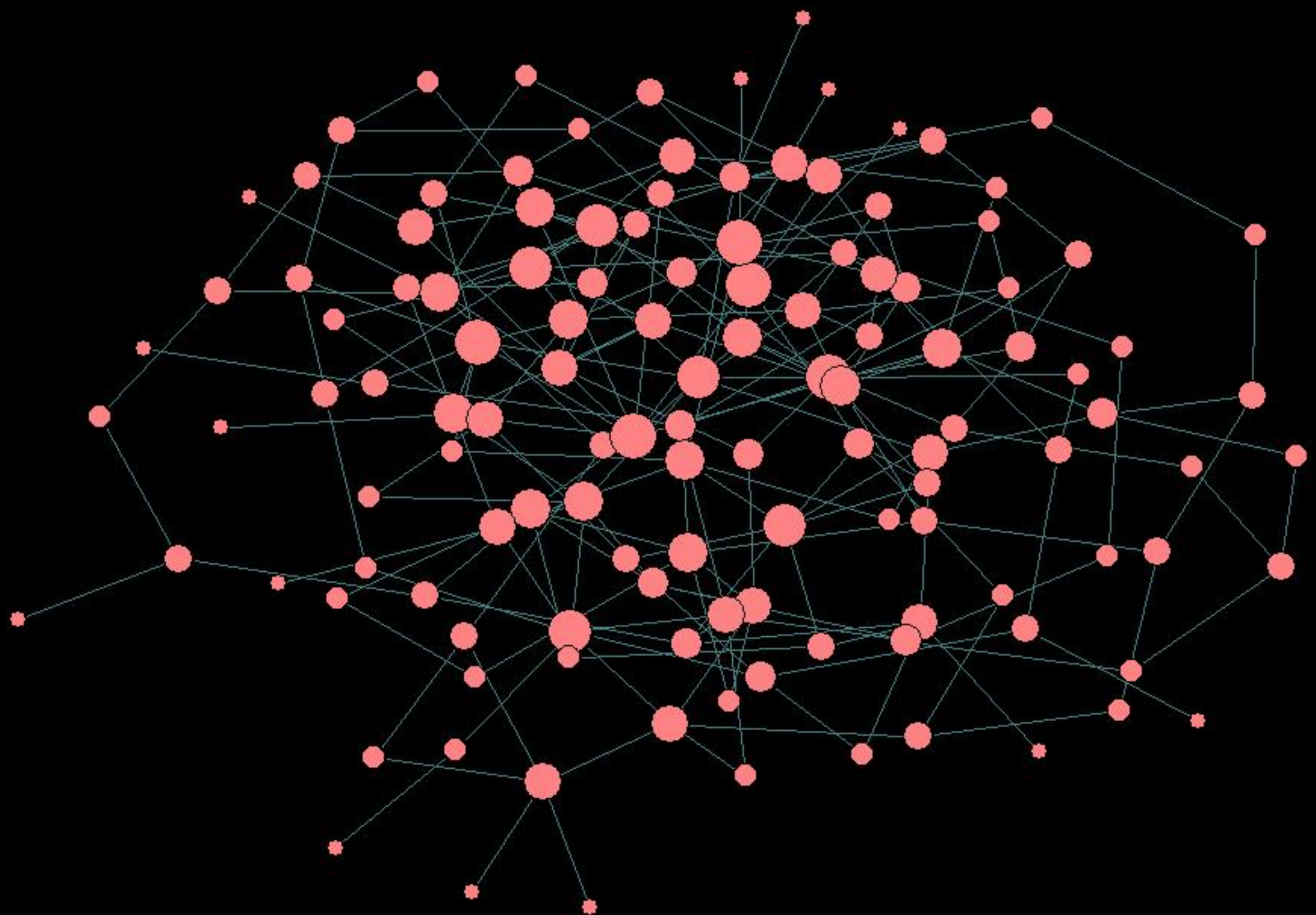


- Scale-free: Connectivity per node follows power law distribution

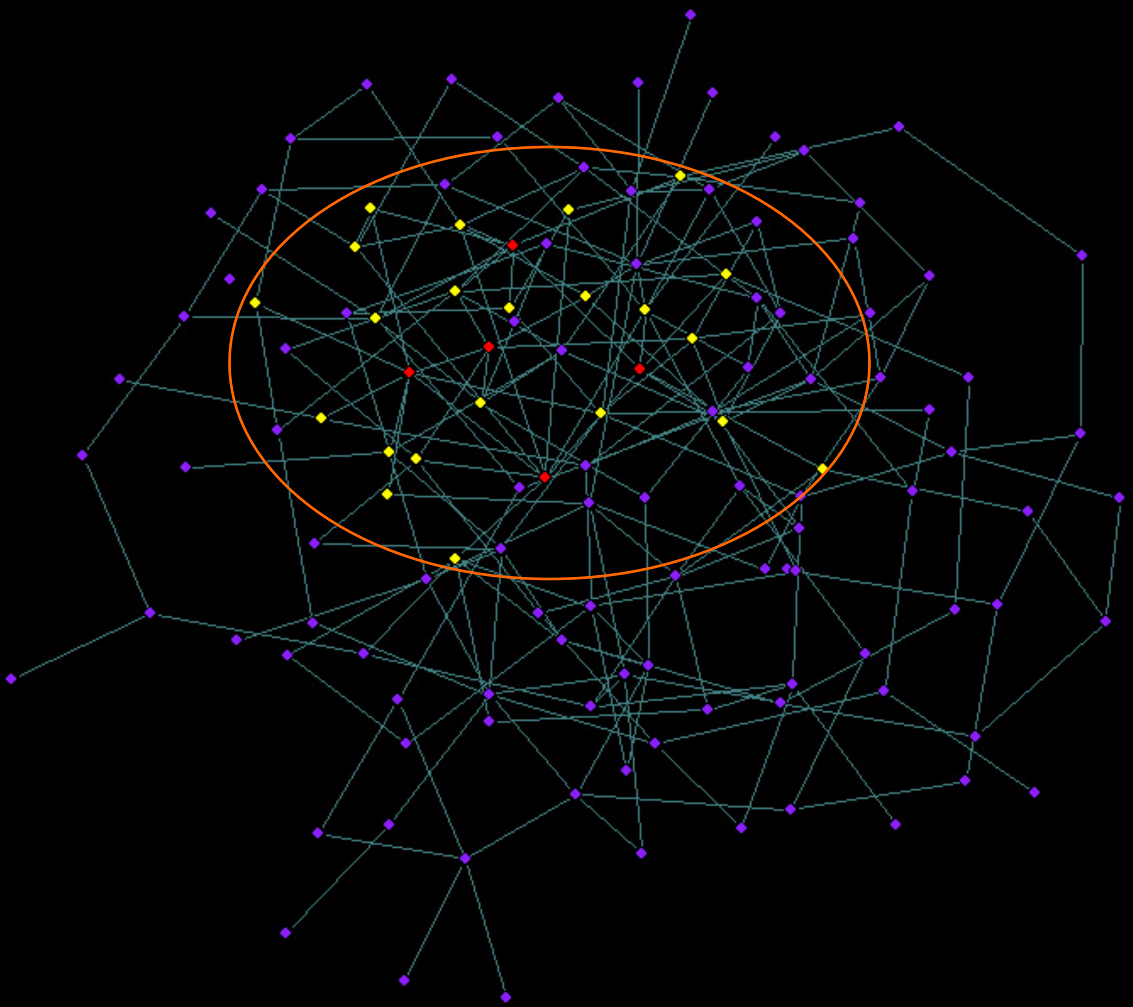




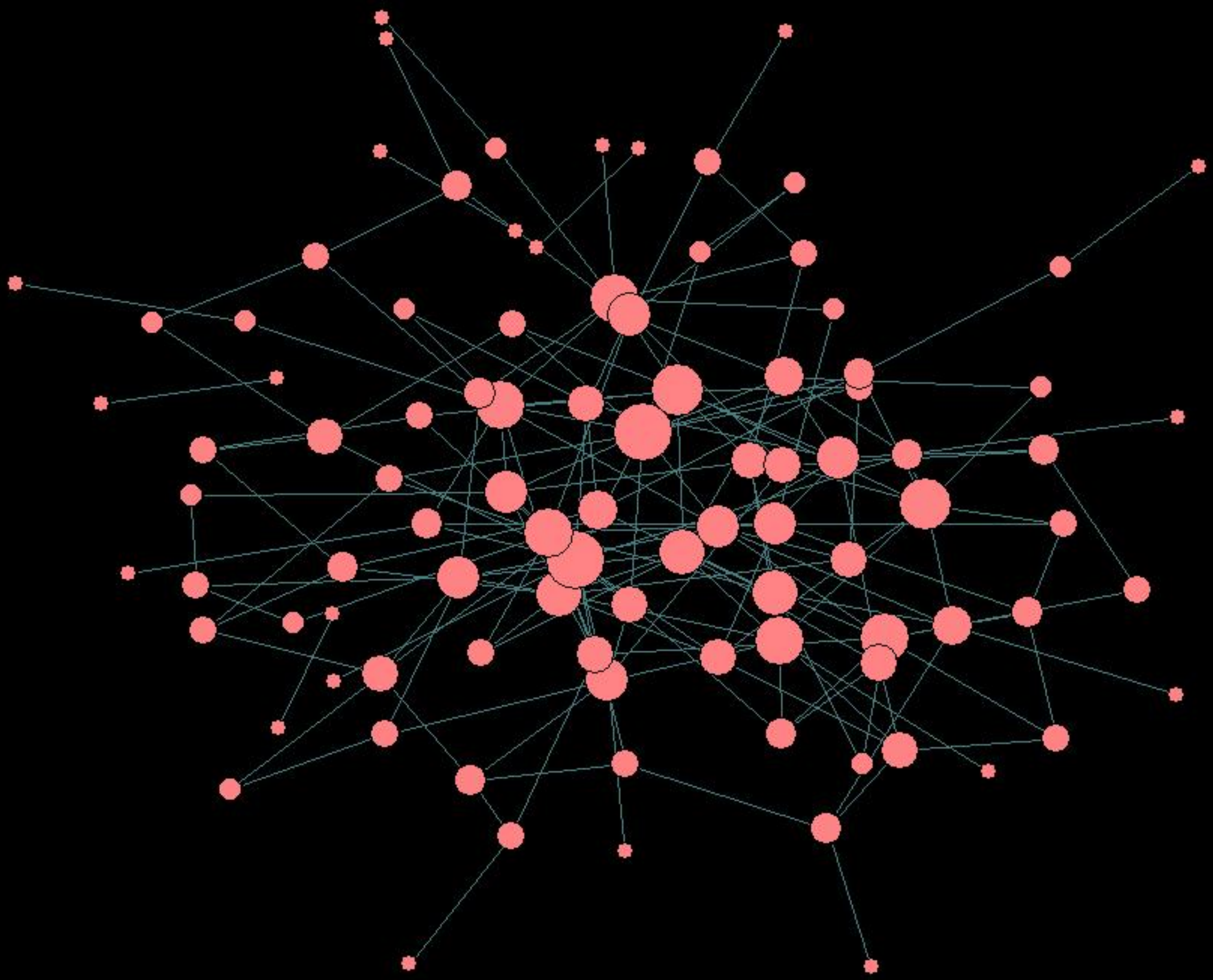
Random (E&R) network: An example



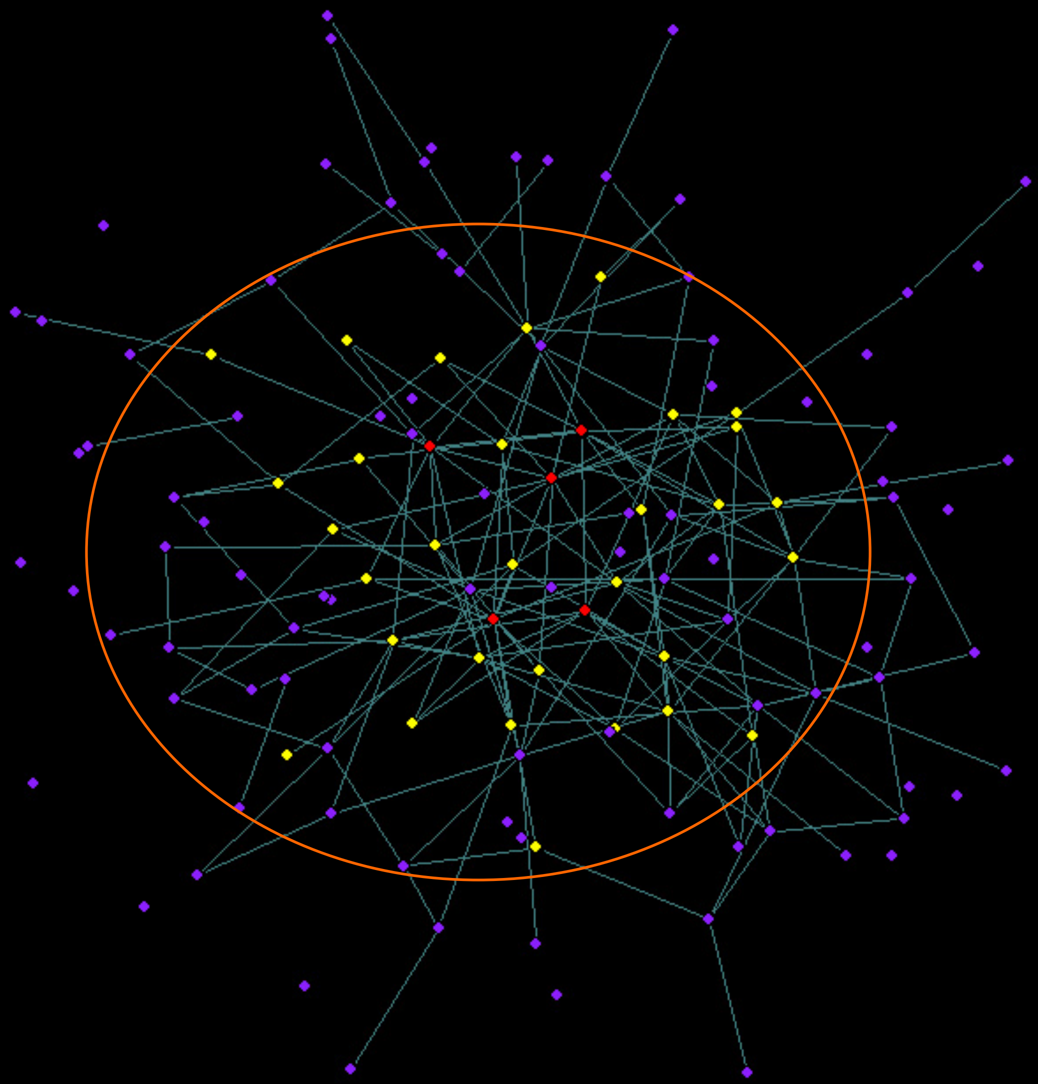
Random (E&R) network: limited reach



scale-free network: An example

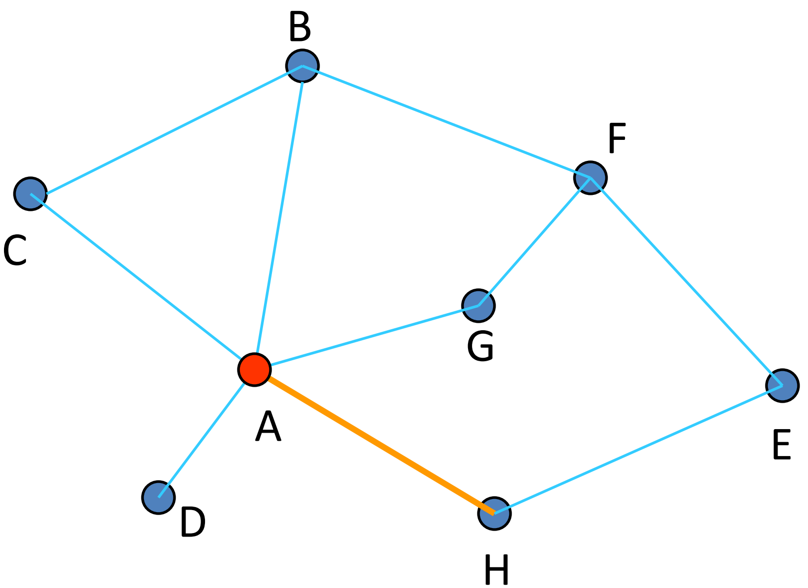


scale-free network: wider reach



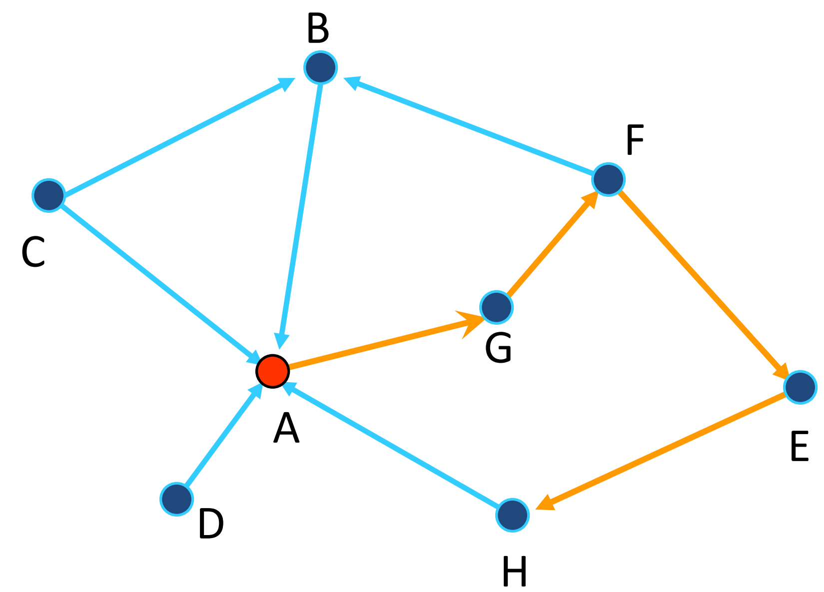
Shortest path

Undirected



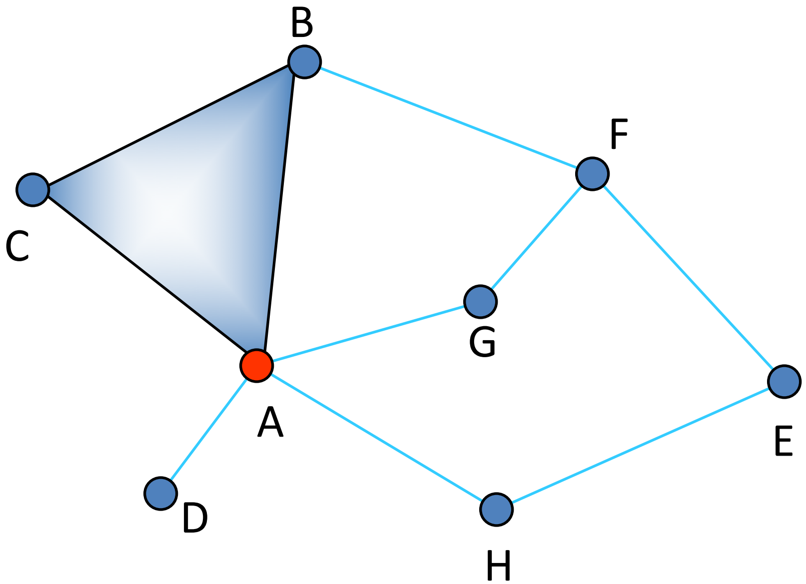
$l_{AH}=1$

Directed



$l_{AH}=4$

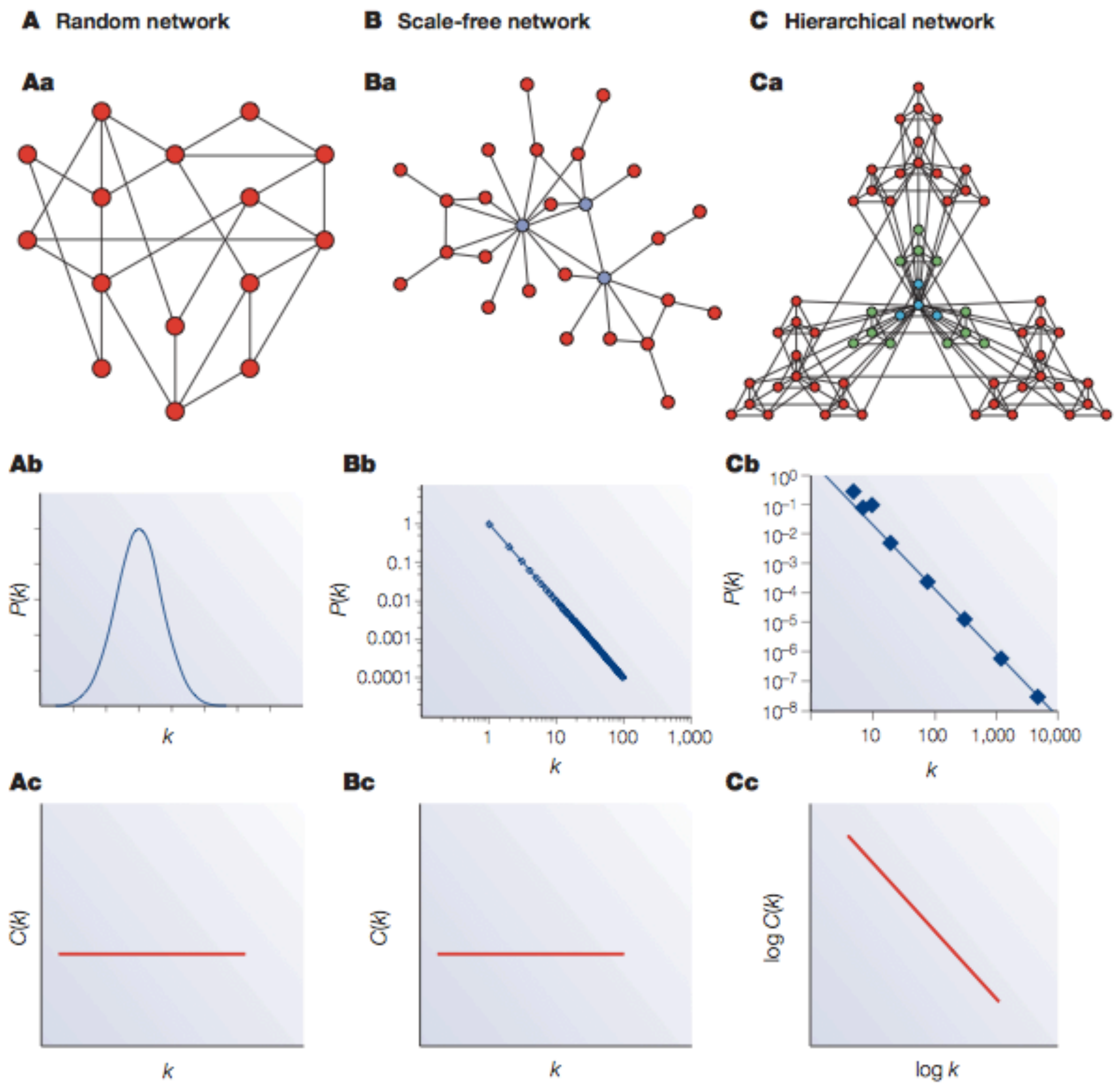
Clustering coefficient



$C_I=2n_I/k(k-1)$

$C_A=2*1/5(5-1)= 0.1$

Network characterization by degree and clustering coefficient



Eccentricity

- The **eccentricity** of a vertex is the greatest geodesic distance between a given node and any other node. It can be thought of as how far a node is from the node most distant from it in the graph.

Diameter

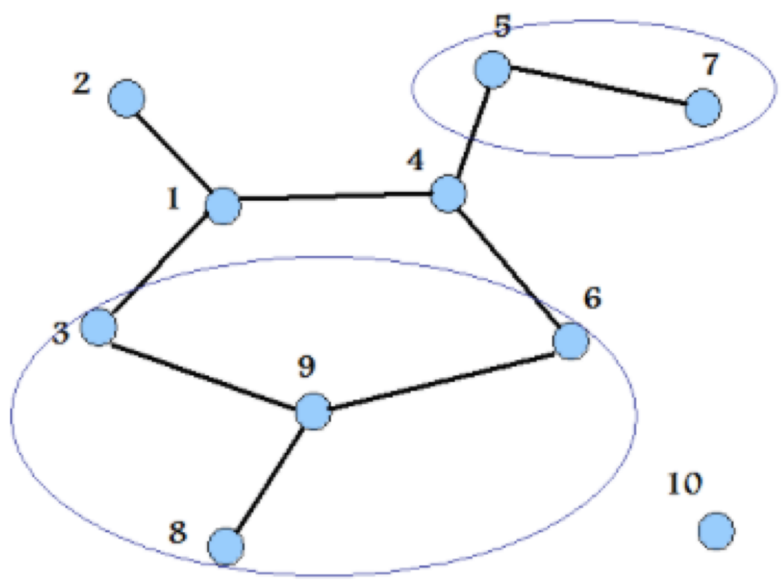
- The diameter of a graph is the **maximum eccentricity** of any vertex in the graph. That is, it is the greatest distance between any pair of vertices.
- To find the diameter of a graph, first find the shortest path between each pair of vertices. The greatest length of any of these paths is the diameter of the graph.

Radius

The radius of a graph is the **minimum eccentricity** of any vertex

Network Metrics in R: Egocentricity

- **Egocentric Network**
  - The egocentric network (or ego net) of vertex v in graph G is defined as the subgraph of G induced by v and its neighbors
  - It can be used to compute metrics over a local neighborhood, especially useful when dealing with large networks



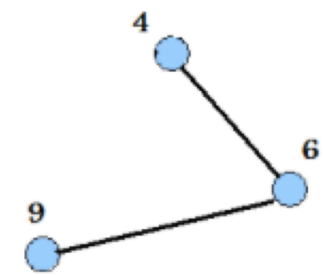
Egocentric networks for nodes 9 and 7

As depicted in this figure, the egocentric network of 9 has nodes 3, 6 and 8 (in addition to 9). Similarly, the ego net of 7 includes node 5.

Network Metrics in R: Egocentricity

- **Example: ego.extract()**

```
> #ego.extract takes one or more input graphs and
  returns a list containing the egocentric networks
  centered on vertices named in ego, using adjacency
  rule neighborhood to define inclusion.
> ego.extract(graph,6)
$`6`
     [,1] [,2] [,3]
[1,] 0    1     1
[2,] 1    0     0
[3,] 1    0     0
```



  - The ego-centric network of node 6 has nodes 6, 4 and 9
  - Note that the sub-graph extracted in this example has the original nodes 6, 4, 9 renamed to 1, 2, 3, respectively
  - Looking at the adjacency matrix, it can be inferred that node 6 is connected to both nodes 4 and 9, whereas nodes 4 and 9 are not directly connected to each other

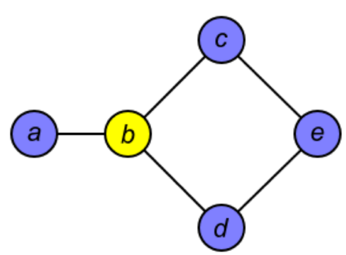
Network Metrics in R: Betweenness

- **Betweenness Centrality**
  - A measure of the degree to which a given node lies on the shortest paths (geodesics) between other nodes in the graph
  - For node v in graph G, betweenness centrality (C_b) is defined as:

$$C_b(v) = \sum_{s,t \neq v} \frac{\Omega_v(s,t)}{\Omega(s,t)}$$

where $\Omega(s,t)$ is the number of distinct geodesics from s to t and $\Omega_v(s,t)$ is the number of geodesics from s to t that pass through v.

  - A node has high betweenness if the shortest paths (geodesics) between many pairs of other nodes in the graph pass through it
  - Thus, when a node with high betweenness fails, it has a greater influence on the information flow in the network



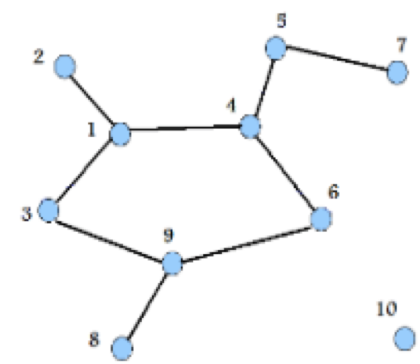
Network Metrics in R: Betweenness

- **Example: betweenness()**

```
> #Here node 4 has the highest betweenness
> betweenness(graph)
 [1] 20 0 8 28 14 12 0 0 16 0
```

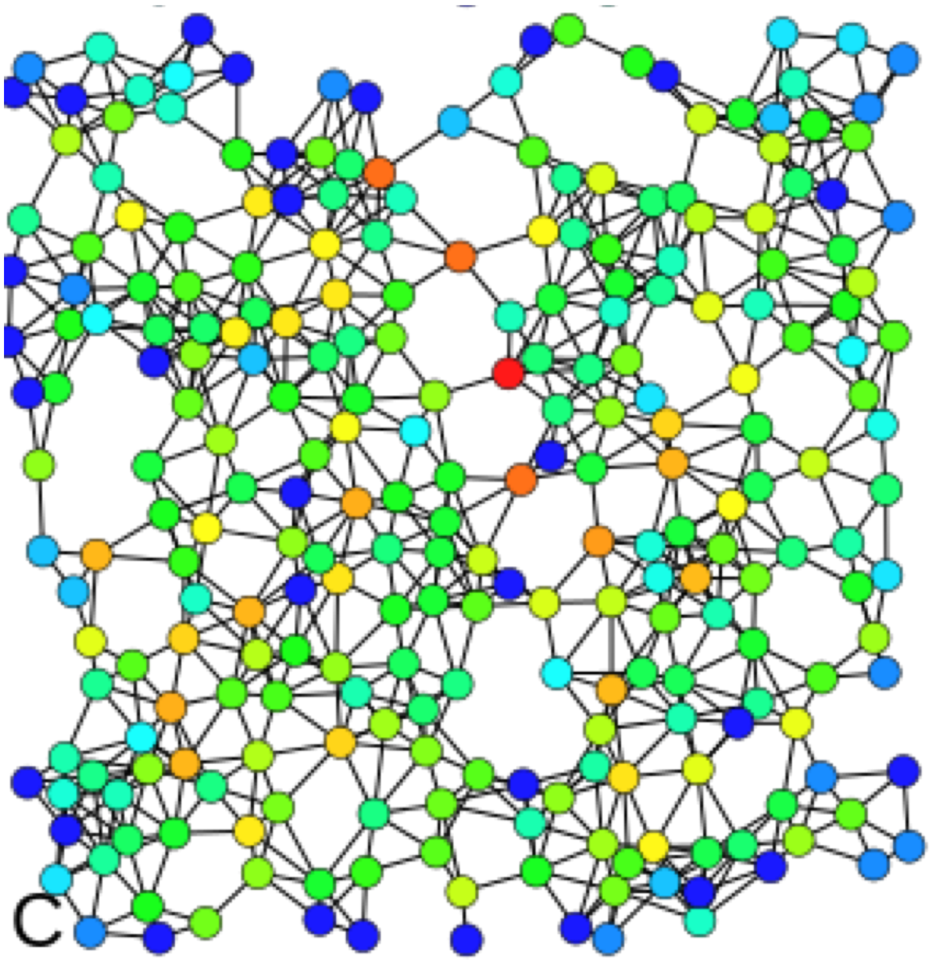
  – Path lengths/geodesic distances can be calculated using geodist()

```
> geo=geodist(graph)
> geo$gdist
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
 [1,]   0    1    1    1    2    2    3    3    2   Inf
 [2,]   1    0    2    2    3    3    4    4    3   Inf
 [3,]   1    2    0    2    3    2    4    2    1   Inf
 [4,]   1    2    2    0    1    1    2    3    2   Inf
 [5,]   2    3    3    1    0    2    1    4    3   Inf
 [6,]   2    3    2    1    2    0    3    2    1   Inf
 [7,]   3    4    4    2    1    3    0    5    4   Inf
 [8,]   3    4    2    3    4    2    5    0    1   Inf
 [9,]   2    3    1    2    3    1    4    1    0   Inf
[10,]Inf  Inf  Inf  Inf  Inf  Inf  Inf  Inf  Inf     0
```



  – It could be inferred that node 5 requires two hops to reach node 1 and node 10 is not reachable by any other node

Betweenness centrality



Network Metrics in R: Closeness

- **Closeness Centrality**
  - Closeness Centrality (CLC) is a category of measures that rate the centrality of a node by its closeness (distance) to other nodes
  - CLC of a node v is defined as:

$$CLC(v) = \frac{|V| - 1}{\sum_{i, v \neq v_i} distance(v, v_i)}$$

where $|V|$ is the number of nodes in the given graph and v_i is the node i of the given graph.

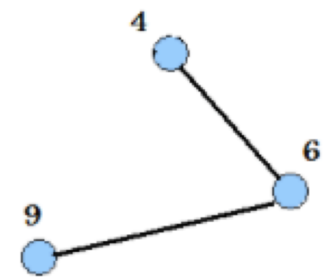
  - Closeness Centrality decreases if either the number of nodes reachable from the node in question decreases, or the distances between the nodes increases

Network Metrics in R: Closeness

- **Example: closeness()**
  - The 10-node graph we have been using has one disconnected node; the resulting infinite distances thus created invalidate any aggregate measure over all nodes such as Closeness Centrality
  - So, we choose a sub-graph – the egocentric network of node 6

```
> #closeness centrality measures how many steps are
  required to access every other vertex from a given
  vertex
> closeness(graph)
 [1] 0 0 0 0 0 0 0 0 0 0
> #We now consider a sub-graph of the graph
 generated for easy understanding of closeness
> graph1=ego.extract(graph,6)
> graph1
$`6`
     [,1] [,2] [,3]
[1,]    0    1    1
[2,]    1    0    0
[3,]    1    0    0

> closeness(graph1)
             6
[1,] 1.0000000
[2,] 0.6666667
[3,] 0.6666667
```



The closeness centrality of node 6 is:

CLC(6) = (3-1) / (1+1) = 1

Incidentally, this means node 6 can reach all other nodes in one hop.

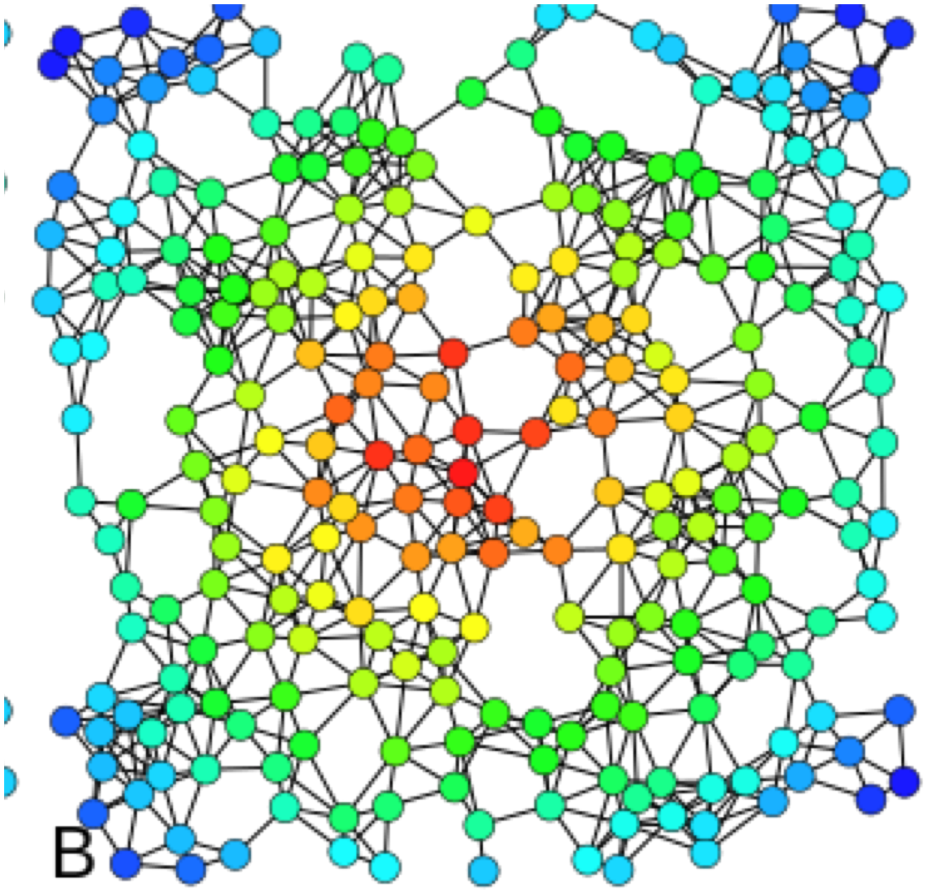
Now, considering node 4:

CLC(4) = (3-1) / (1+2) = 2 / 3
= 0.667

Similarly for node 9:

CLC(9) = 0.667

Closeness Centrality



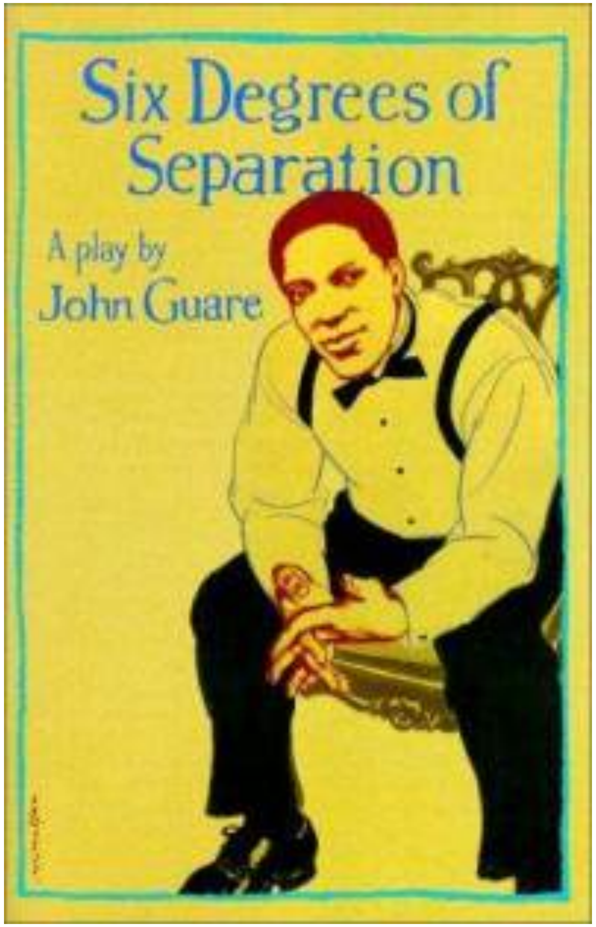
Six degrees of separation

Milgram's experiment

1. ADD YOUR NAME TO THE ROSTER AT THE BOTTOM OF THE SHEET. So that the next person who receives the letter will know where it came from
2. DETACH ONE POSTCARD. FILL IT OUT AND RETURN IT TO HARVARD UNIVERSITY. To allow us to keep track of the folder as it moves toward the target person
3. IF YOU KNOW THE TARGET PERSON ON PERSONAL BASIS, MAIL THIS FOLDER DIRECTLY TO HIS/HER.
4. IF YOU DO NOT KNOW THE TARGET PERSON, MAIL THIS FOLDER TO A PERSONAL ACQUAINTANCE WHO IS MORE LIKELY THAN YOU TO KNOW THE TARGET PERSON

Milgram, S (1967). Psychol. Today, 2, 60-67)

1991: John Guare



"Everybody on this planet is separated by only six other people. Six degrees of separation. Between us and everybody else on this planet. The president of the United States. A gondolier in Venice…. It's not just the big names. It's anyone. A native in a rain forest. A Tierra del Fuegan. An Eskimo. I am bound to everyone on this planet by a trail of six people. It's a profound thought. How every person is a new door, opening up into other worlds."

WWW: 19 DEGREES OF SEPARATION

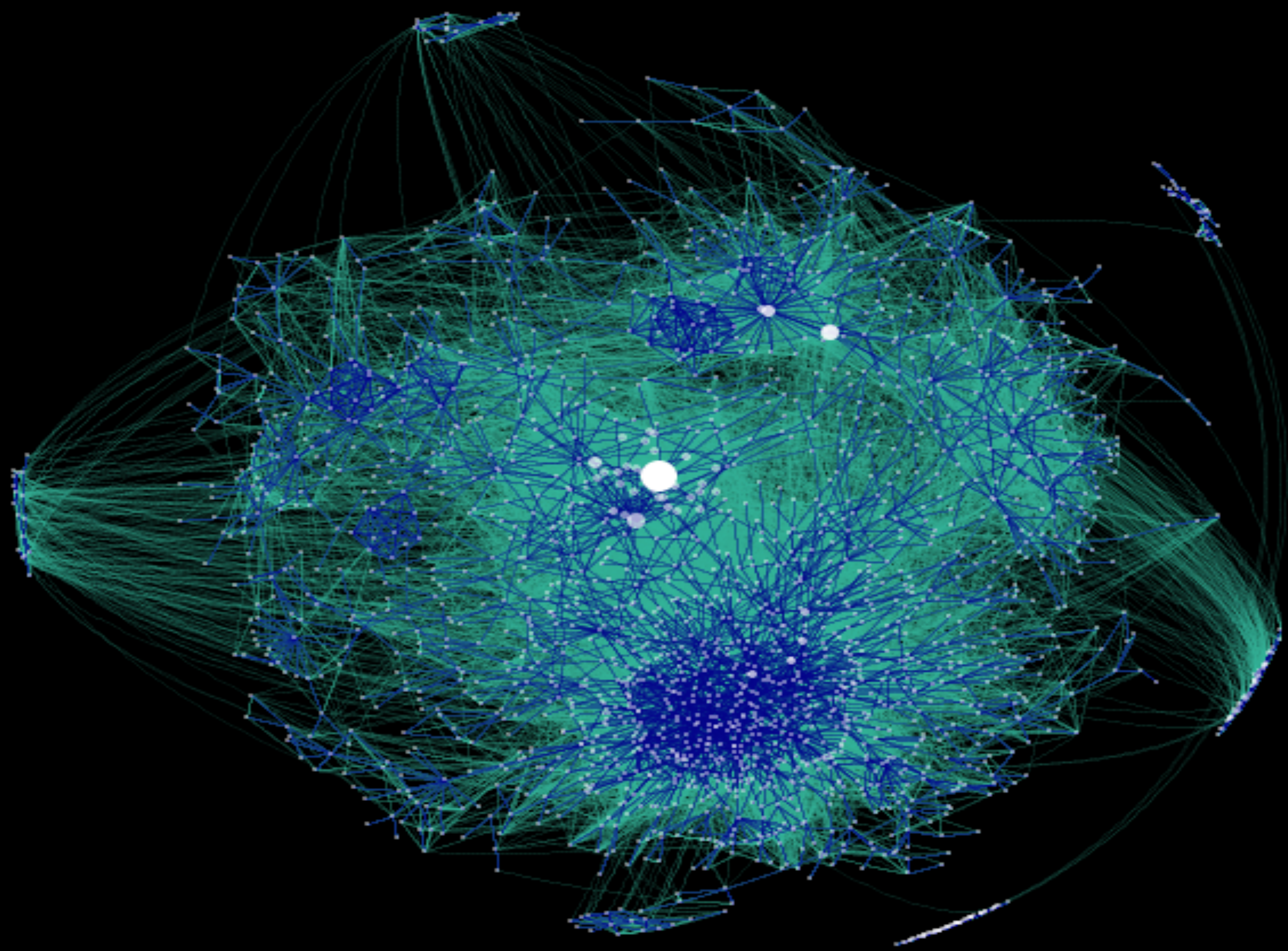
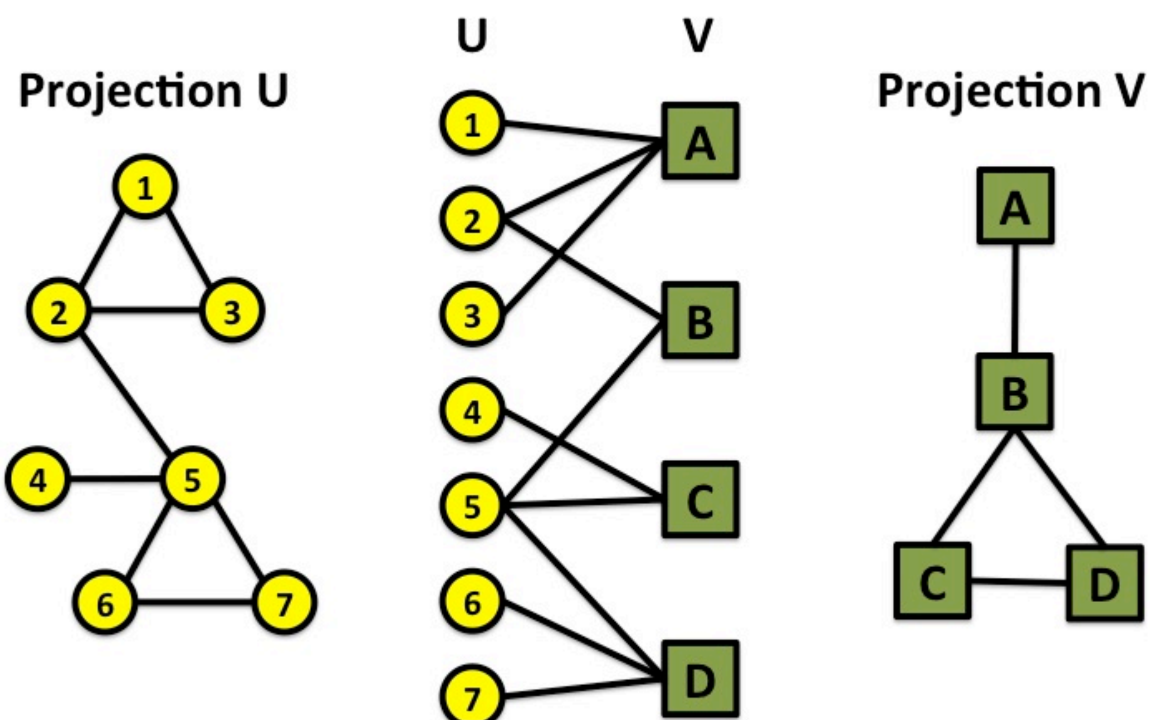


Image by **Matthew Hurst**
Blogosphere

Bi-partite networks

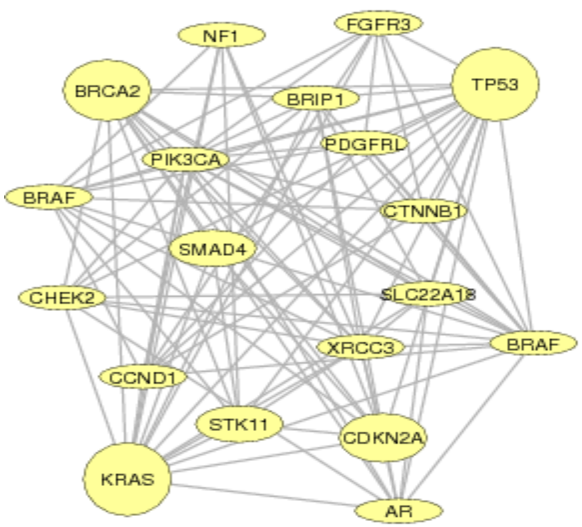
bipartite graph (or bigraph) is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V; that is, U and V are independent sets.



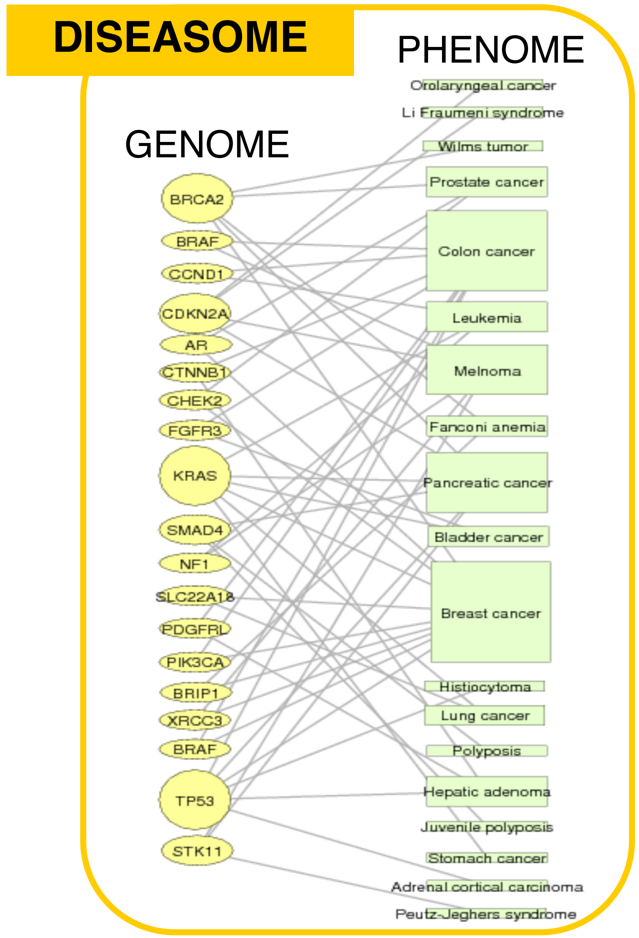
Examples:

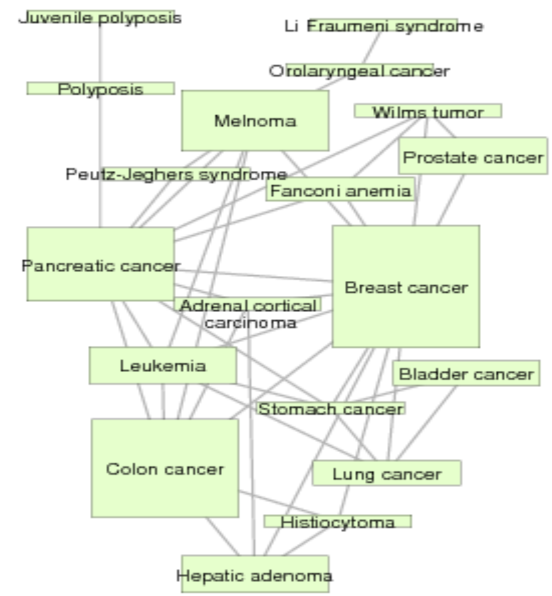
Hollywood actor network
Collaboration networks
Disease network (diseasome)

GENE NETWORK – DISEASE NETWORK



Gene network

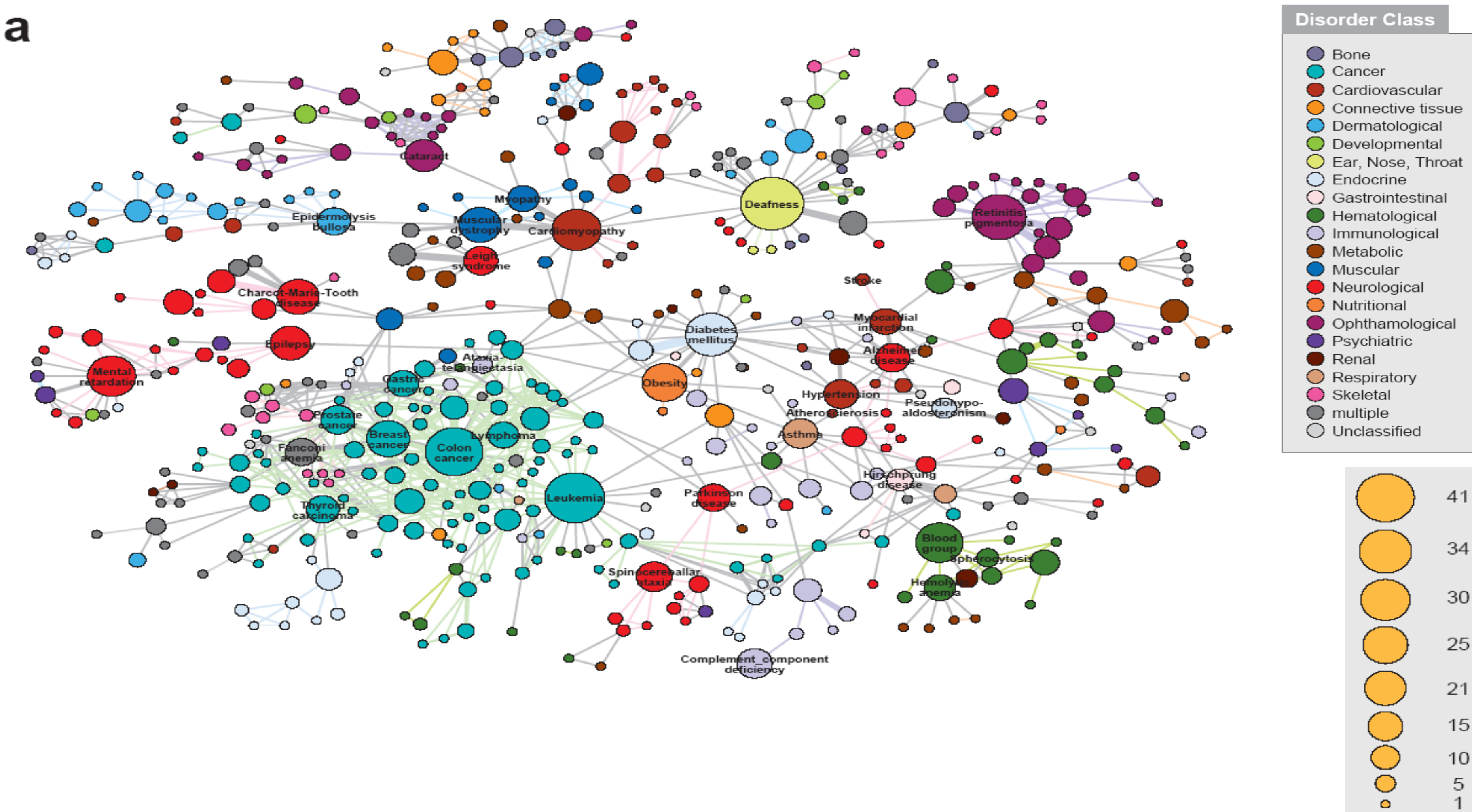




Disease network

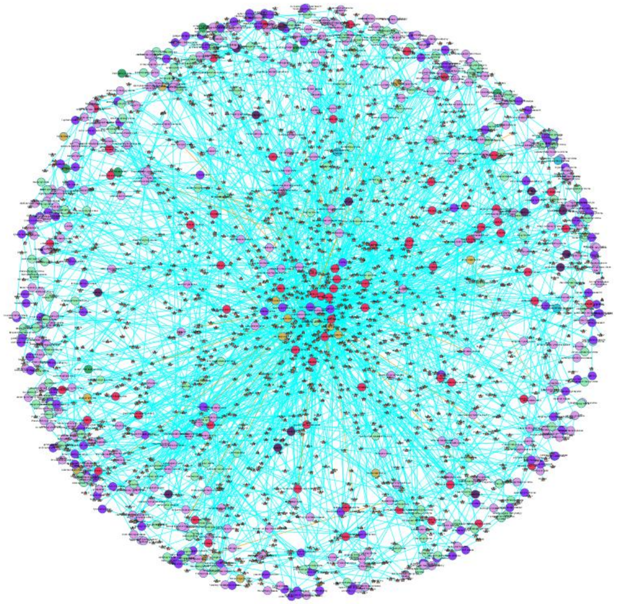
Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)

The diseasome



GWAS

OMIM

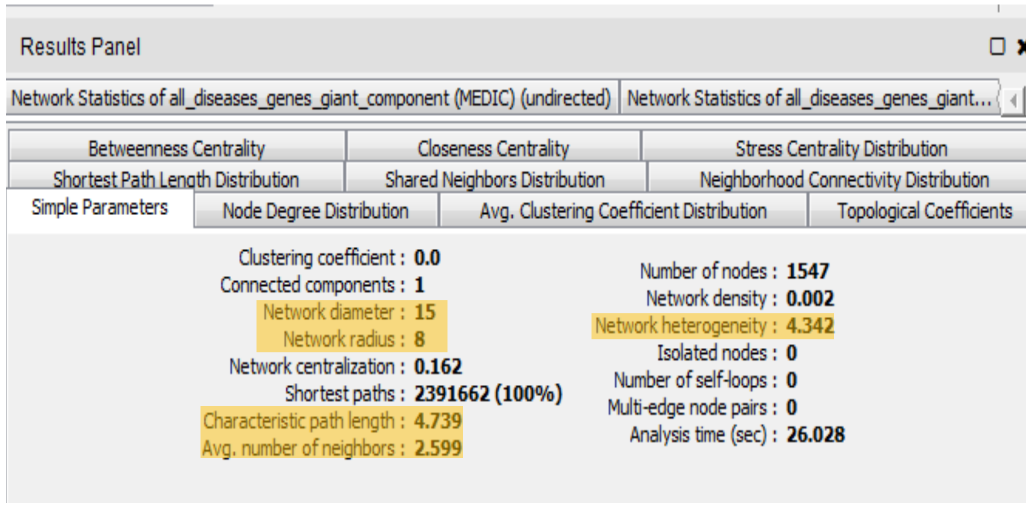


1547 nodes
2010 edges
Ratio N/E= 0.77

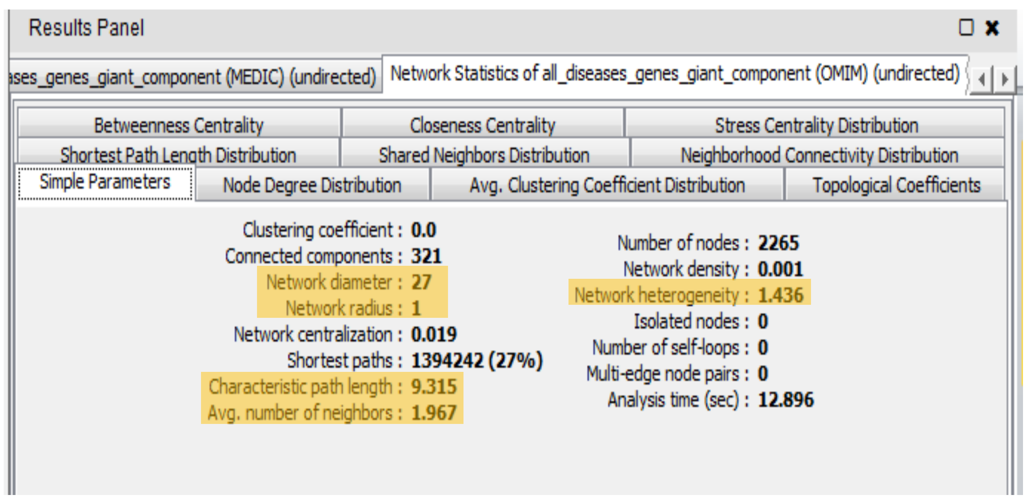
2265 nodes
2228 edges
Ratio N/E= 1.01

Summary network statistics

GWAS

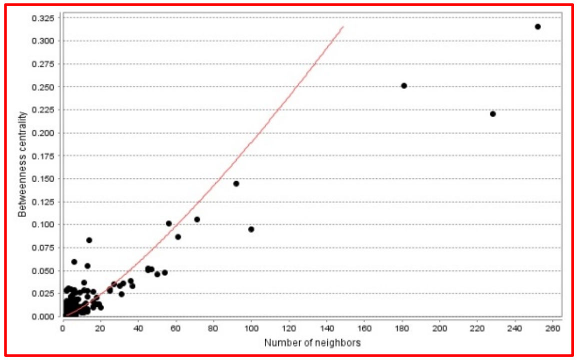


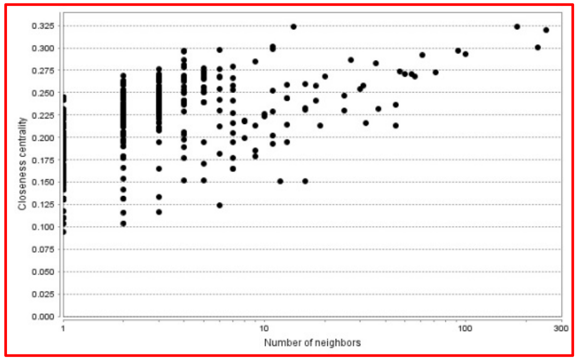
OMIM

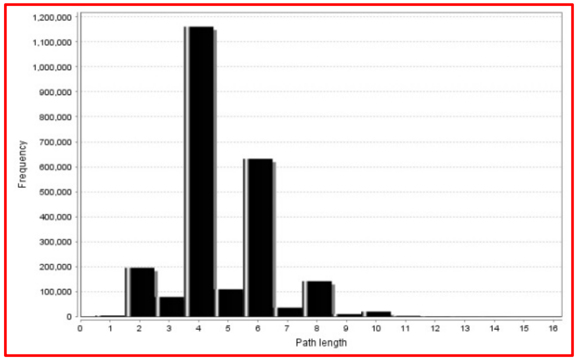


Betweeness centrality | Closeness centrality | Shortest path length distribution

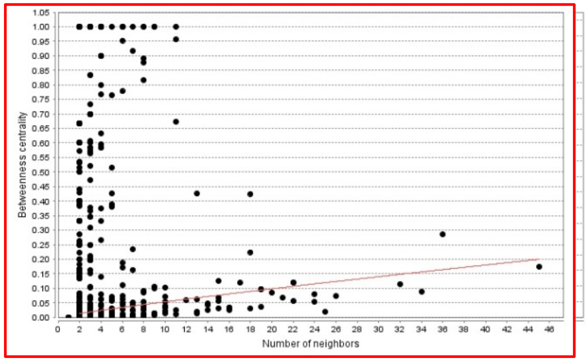
GWAS

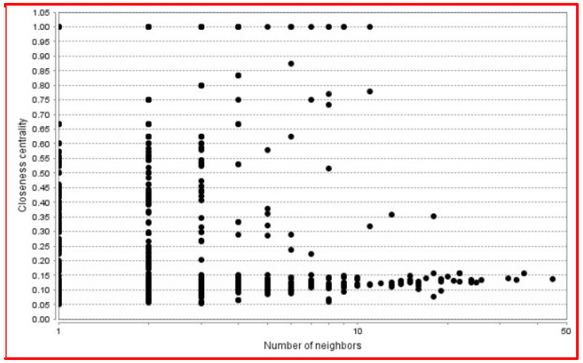


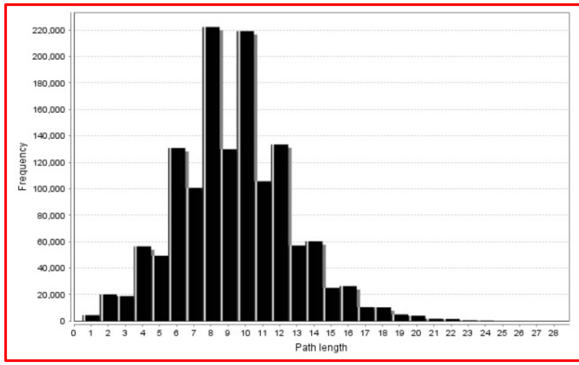




OMIM







ROBUSTNESS IN COMPLEX SYSTEMS

Complex systems maintain their basic functions even under errors and failures

Cell → mutations

There are uncountable number of mutations and other errors in our cells, yet, we do not notice their consequences.

Internet → router breakdowns

At any moment hundreds of routers on the internet are broken, yet, the internet as a whole does not loose its functionality.

Where does robustness come from?

There are feedback loops in most complex systems that keep tab on the component's and the system's 'health'.

Could the network structure affect a system's robustness?

Attack threshold for arbitrary P(k)

Attack problem: we remove a fraction *f* of the hubs.

At what threshold f_c will the network fall apart (no giant component)?

Hub removal changes

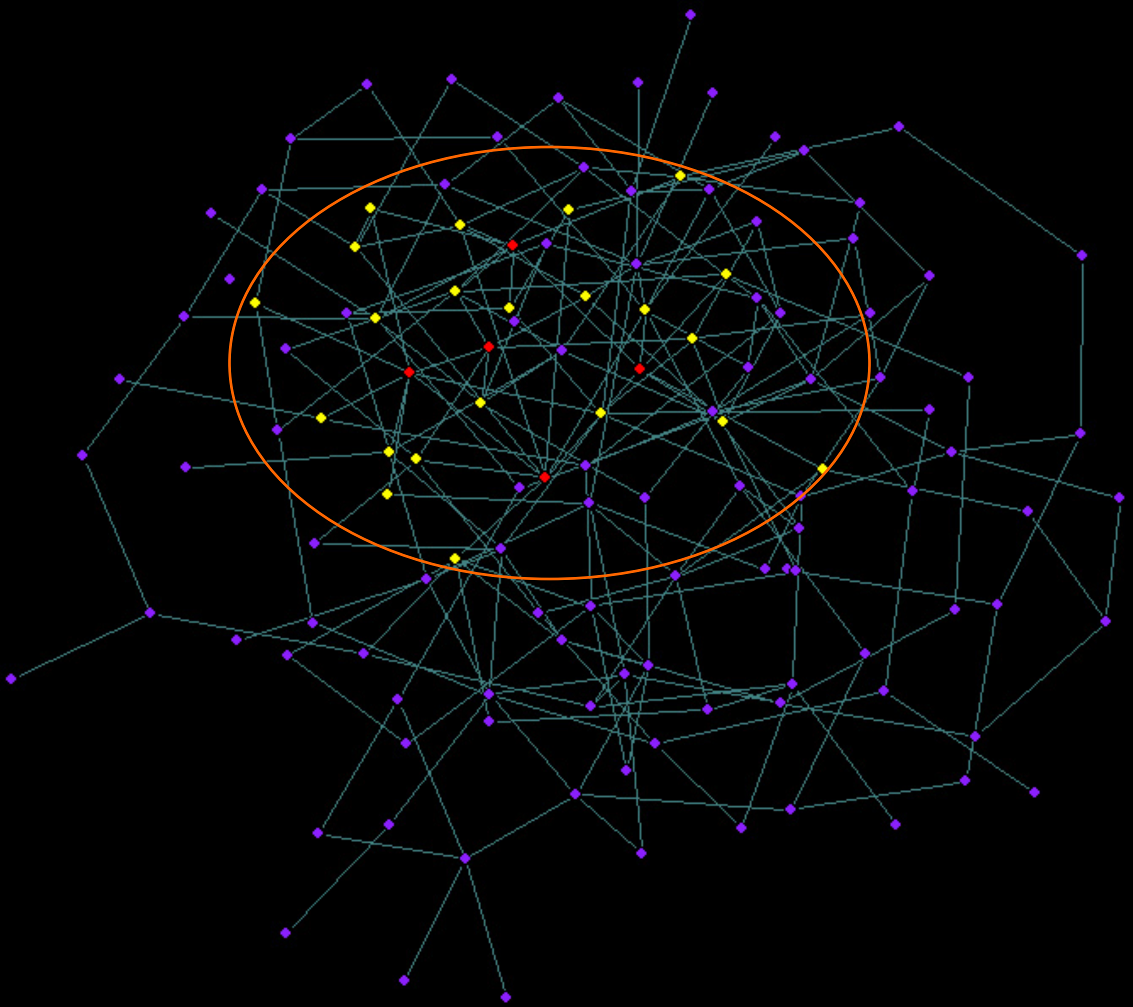
the maximum degree of the network [$K_{max} \rightarrow K'_{max} \leq K_{max}$)

the degree distribution [$P(k) \rightarrow P'(k')$]

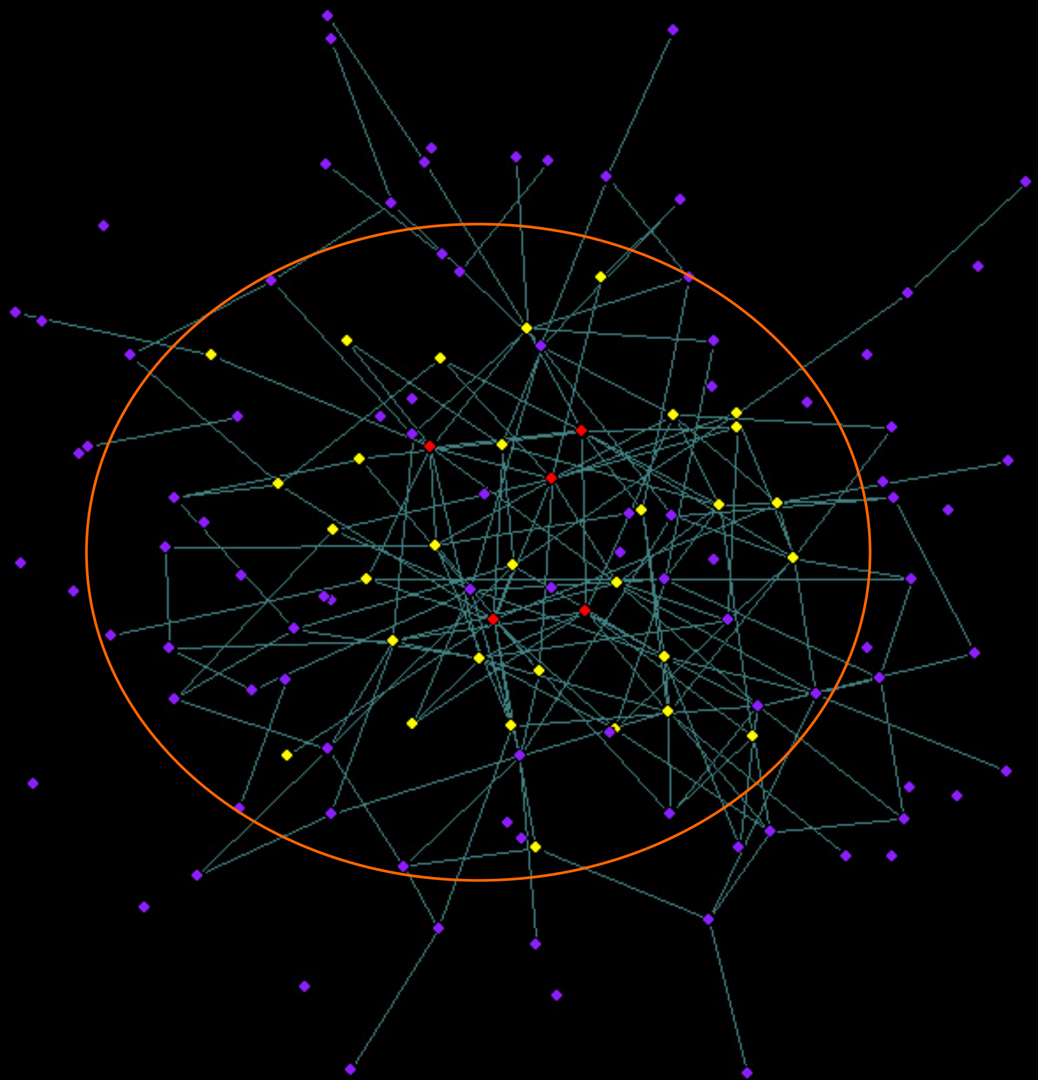
A node with degree k will loose some links because some of its neighbors will vanish.

Cohen et al., Phys. Rev. Lett. 85, 4626 (2000).

Random (E&R) network: limited reach



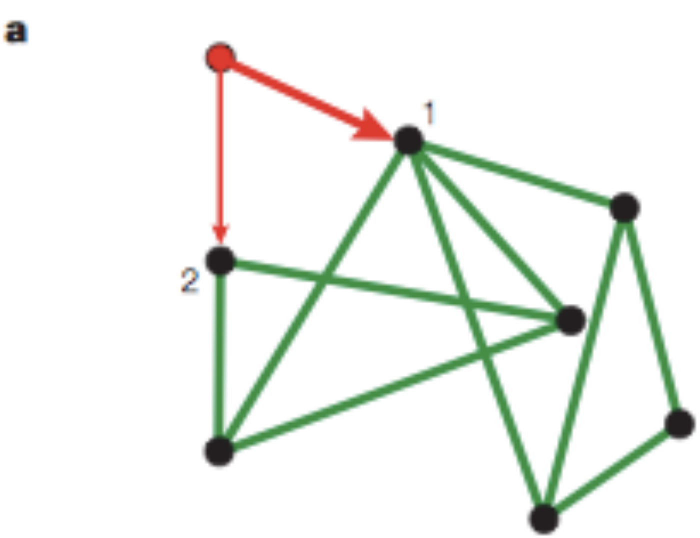
scale-free network: wider reach

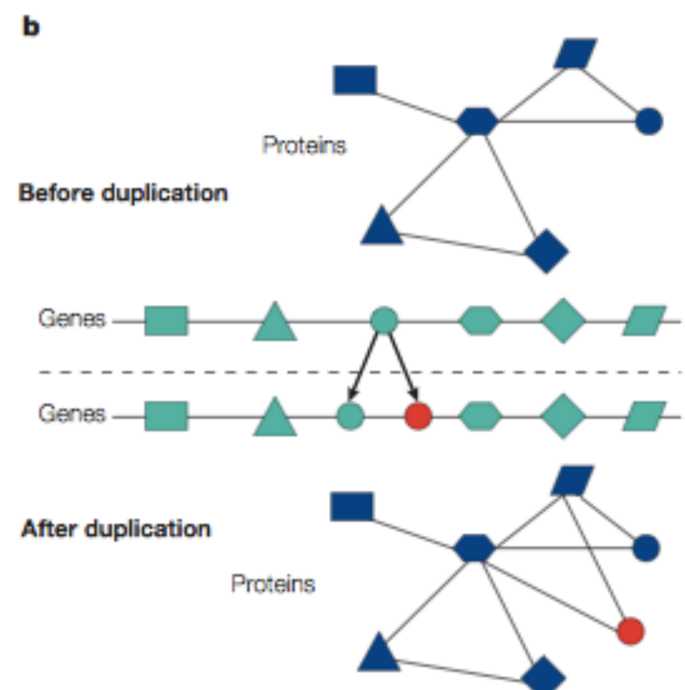


Evolution of scale-free networks

1. duplication

2. Preferential attachment





Google page rank: an example of preferential attachment

- Preferential attachment will favor older nodes (e.g. journal article citations). Early journal articles on a given topic more likely to be cited. Once cited, this material is more likely to be cited again in new articles, so original articles in a field have a higher likelihood of becoming hubs in a network of references.

- The Google search engine (PageRank) interprets a link from page A to page B as a vote, by page A, for page B. It also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important”.

Useful links on networks

- http://barabasilab.neu.edu/courses/phys5116/
- http://math.nist.gov/~RPozo/complex_datasets.html
- http://www2.econ.iastate.edu/tesfatsi/netgroup.htm
- http://www.visualcomplexity.com/vc/about.cfm
- http://necsi.edu/publications/dcs/
- http://cnets.indiana.edu