Web Structure Mining

Graph & Network Mining: Concepts

Li Xiaoli

Information about Instructors

- Senior scientist and Dept Head at Institute for Infocomm Research (I2R), A*Star
- Adj AP Li Xiaoli at NTU
- Email: xlli@i2r.a-star.edu.sg (I2R email)
- Web Webpage: Google Xiaoli Li
  - http://www.ntu.edu.sg/home/xlli/

Plans for Web Structure Mining

- Graph & Network Mining: Concepts
- Rising Stars and Community Detection
- Link Analysis Algorithms

Objective

- At the end of the course, students should be able to
  - Have a good knowledge of graph mining concepts.
  - Be familiar with various graph mining algorithms.
  - Given a graph dataset or task, understand how to use appropriate graph mining techniques to analyze data and address the problem.

Target Audiences

- Students who intend to do research in graph data mining or related fields – Hot topics!
- Students who intend to become a data scientist to apply graph mining techniques to solve real-world applications – Higher Paid Job!
- Take your time to practice to sharpen your skills – you need to do a bit programming!
- Remember you need to integrate with domain knowledge - mind your application domains!

Evaluation

- CA: Assignment Projects (40%)
  - 20% from Barry
  - 20% from Xiaoli

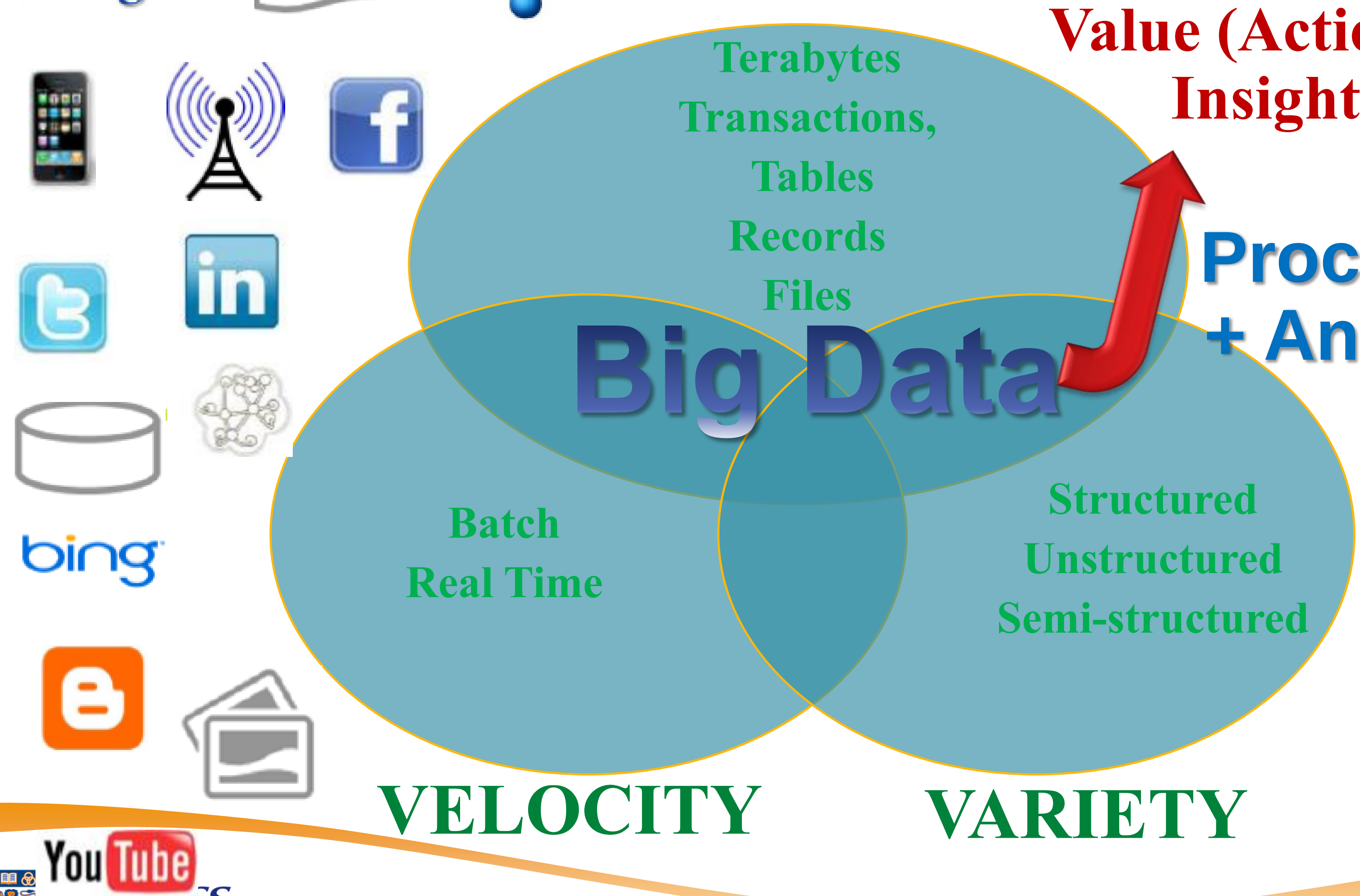
- Final Exam (60%)
  - 30% from Barry
  - 30% from Xiaoli

To Be Confirmed

Big Data Analytics

VOLUME

Value (Actionable Insights)



Terabytes
Transactions,
Tables
Records
Files

Big Data

Processing + Analytics

Batch
Real Time

Structured
Unstructured
Semi-structured

VELOCITY

VARIETY

Motivation: Why Mine Graph Data?

- Example:

Facebook

- 800 million active users
- 60 billion photos in total, 250 million photos uploaded per day
- 80 groups/events per user (till Feb 2011)

Flickr

- 60 million users
- Five billion photos
- 10 million groups (till Feb 2011)

Twitter

- 175 million users (registered)
- 140 million tweets per day

Weibo

- 200 million users (till June 2011)

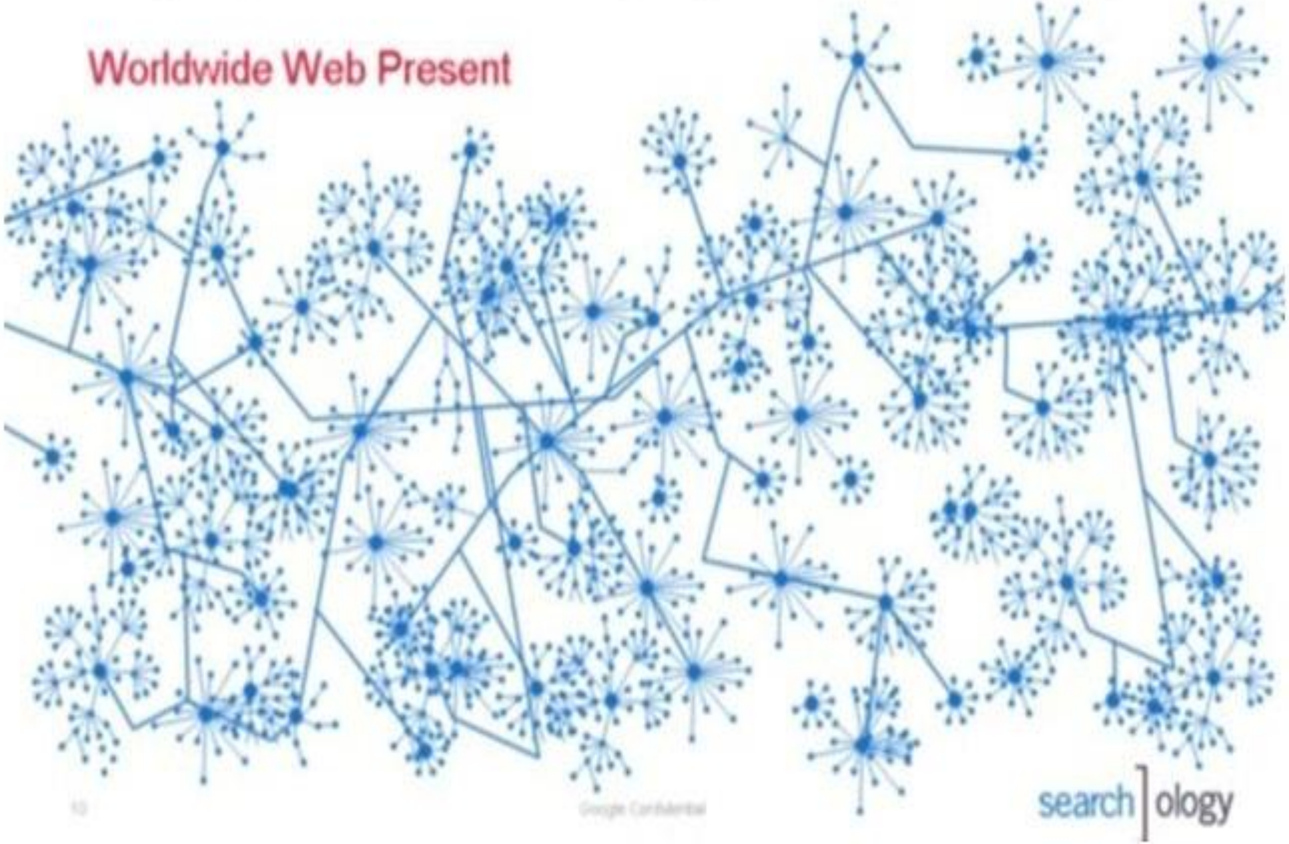
Telecom CDR Data: Huge data sets with location information

“Necessity is the mother of invention”

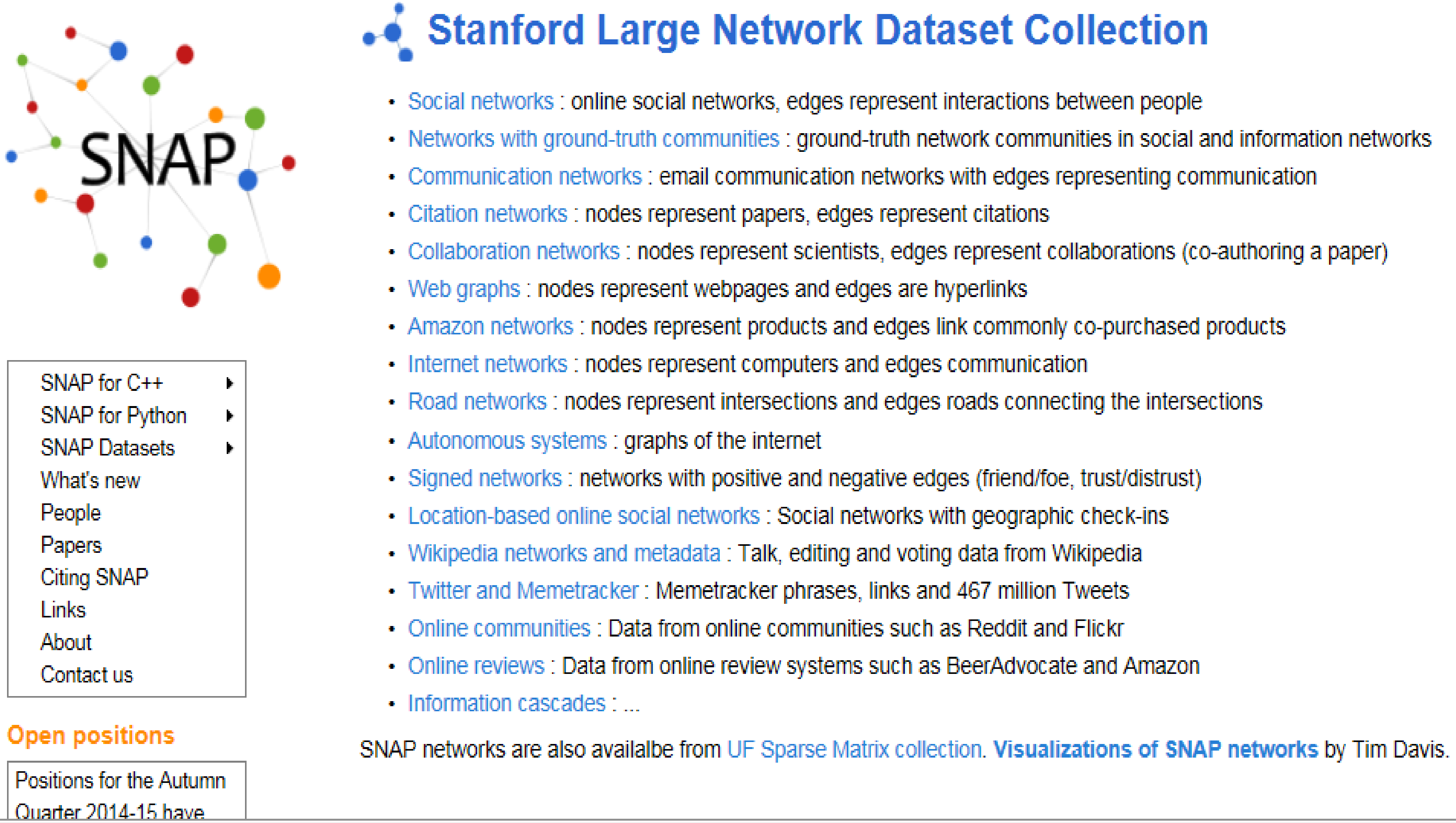
Graph Mining—Automated analysis of massive graph data

The Web is a Graph

- Web graph ~ Directed graph that is formed by webpages and their hyperlinks
- Sub-graph is a set of pages linked to one specific topic



Stanford Large Network Dataset Collection
https://snap.stanford.edu/data/



Stanford Large Network Dataset Collection

- Social networks : online social networks, edges represent interactions between people
- Networks with ground-truth communities : ground-truth network communities in social and information networks
- Communication networks : email communication networks with edges representing communication
- Citation networks : nodes represent papers, edges represent citations
- Collaboration networks : nodes represent scientists, edges represent collaborations (co-authoring a paper)
- Web graphs : nodes represent webpages and edges are hyperlinks
- Amazon networks : nodes represent products and edges link commonly co-purchased products
- Internet networks : nodes represent computers and edges communication
- Road networks : nodes represent intersections and edges roads connecting the intersections
- Autonomous systems : graphs of the internet
- Signed networks : networks with positive and negative edges (friend/foe, trust/distrust)
- Location-based online social networks : Social networks with geographic check-ins
- Wikipedia networks and metadata : Talk, editing and voting data from Wikipedia
- Twitter and Memetracker : Memetracker phrases, links and 467 million Tweets
- Online communities : Data from online communities such as Reddit and Flickr
- Online reviews : Data from online review systems such as BeerAdvocate and Amazon
- Information cascades : ...

SNAP networks are also availalbe from UF Sparse Matrix collection. **Visualizations of SNAP networks** by Tim Davis.

Open positions

Positions for the Autumn Quarter 2014-15 have

Commonly used Network Examples

- **Social networks**. Social networks link people according to various social relationships, like acquaintance, friendship, and collaboration. It is important understand and anticipate the spread of ideas, innovations, fads, as well as biological and computer viruses.
- **The World Wide Web**. This is a directed network in which nodes represent Web pages and edges are the hyperlinks between pages.

Commonly used Network Examples (Cont.)

- **The Internet**. This is a collection of routers linked by various physical lines. The study of Internet topology is crucial to investigate the robustness of the network under failures
- **Citation networks.** An article citation network links scholarly papers through bibliographic references contained in the bibliography of the papers.
- **Language networks**. In these networks the nodes are words and the links represent relationships among words (e.g. "is a" or "part of" relationships, significant co-occurrence in texts)

Commonly used Network Examples (Cont.)

- **Economic networks**. Market can be viewed as a huge directed multi-relational network. Companies, firms, financial institutions, governments play the role of nodes. Links symbolize different interactions between them.
- **Metabolic and protein networks**. The nodes are simple molecules. The links are the biochemical reactions that take place between these molecules
- **Transportation networks**
- **Powerline**
- **Food webs**

Some Applications on Graph Analysis

- These networks contain valuable information for many network applications:
  - Recommendation systems – collaborative filtering
  - Classification – classify the nodes
  - Key players identification – find important nodes
  - Community detection – find interesting subgraphs
  - Web search – enhance web search results
  - Trust and reputation – find experts
  - ……

Graph & Network Mining: Concepts

- 1. General Network Measures
- 2. Typical Architecture of Networks
- 3. Social network analysis

A Small Graph (undirected/directed)

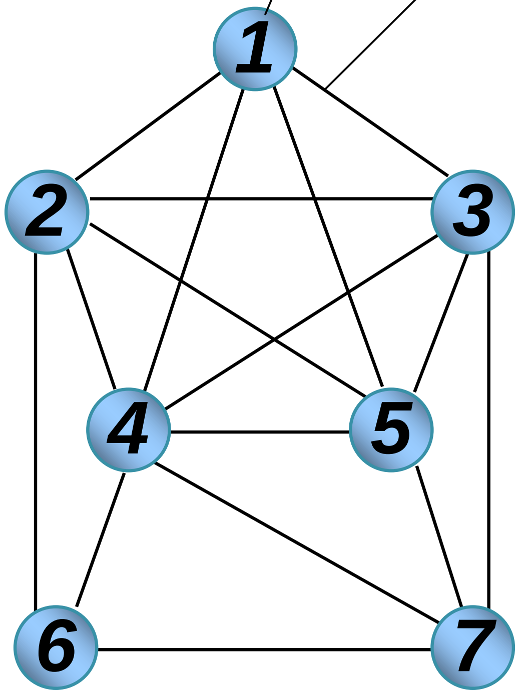
Graph with 7 nodes and 16 edges

$$G = (V, E)$$
$$V = \{v_1, v_2, \ldots, v_n\}$$
$$E = \{e_k = (v_i, v_j) \mid v_i, v_j \in V, k = 1, \ldots, m\}$$

Nodes / Vertices

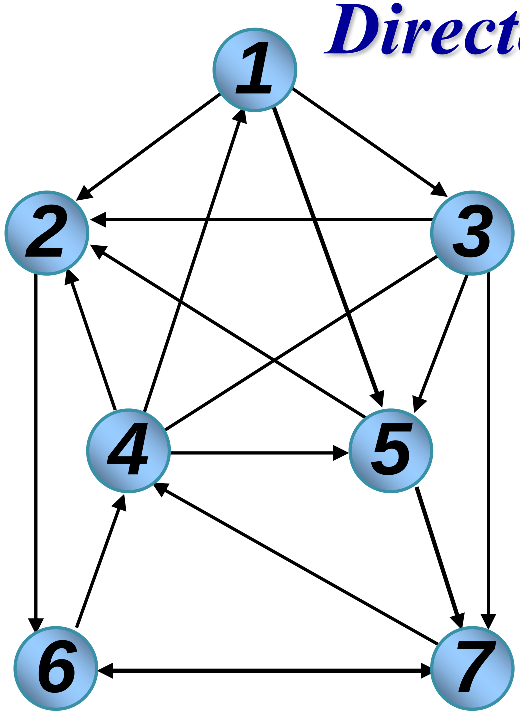
Edges/links

Undirected



$$(v_i, v_j) = (v_j, v_i)$$

Directed



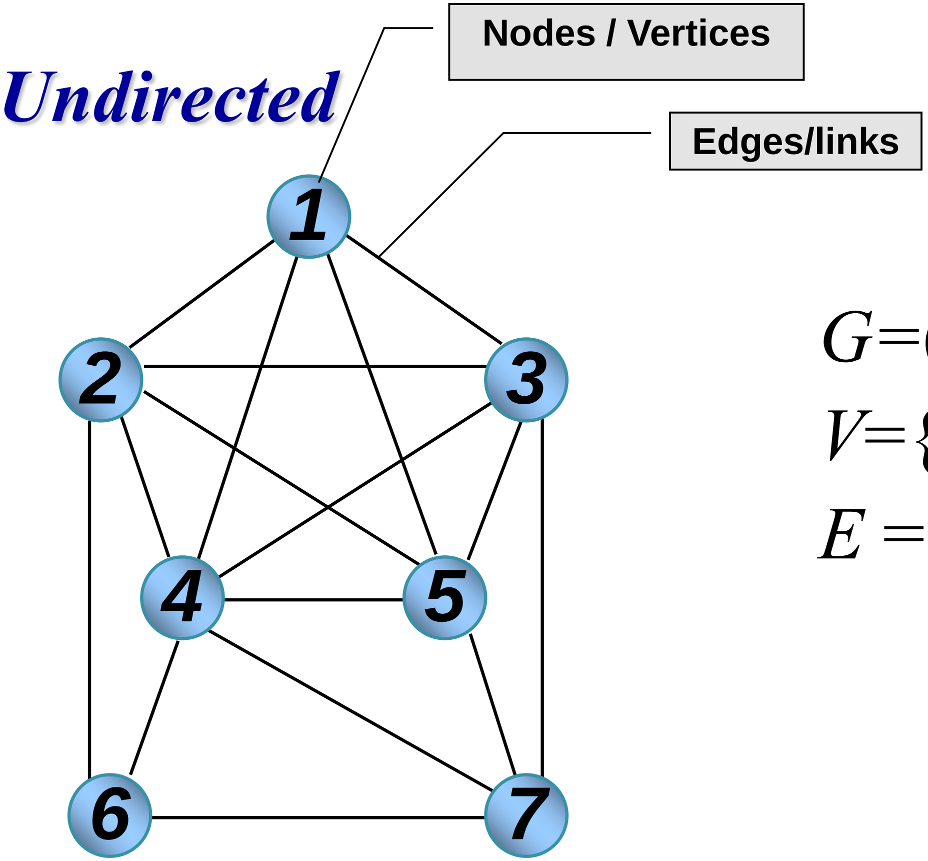
$$(v_i, v_j) \neq (v_j, v_i)$$

A Small Graph:
Example of its mathematical representation

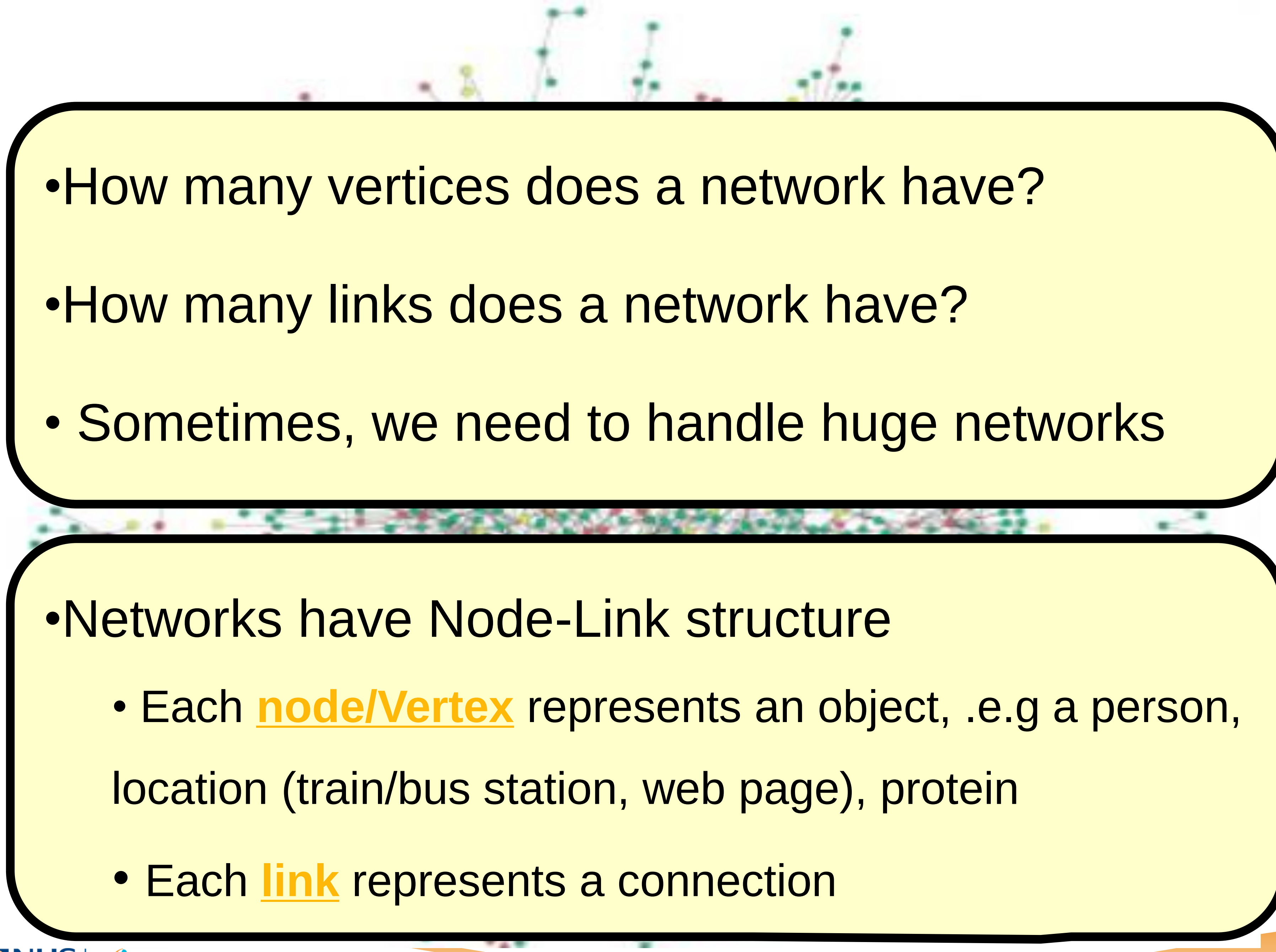
Graph with 7 nodes and 16 edges

$$G = (V, E)$$
$$V = \{v_1, v_2, \ldots, v_n\}$$
$$E = \{e_k = (v_i, v_j) \mid v_i, v_j \in V, k = 1, \ldots, m\}$$

Undirected



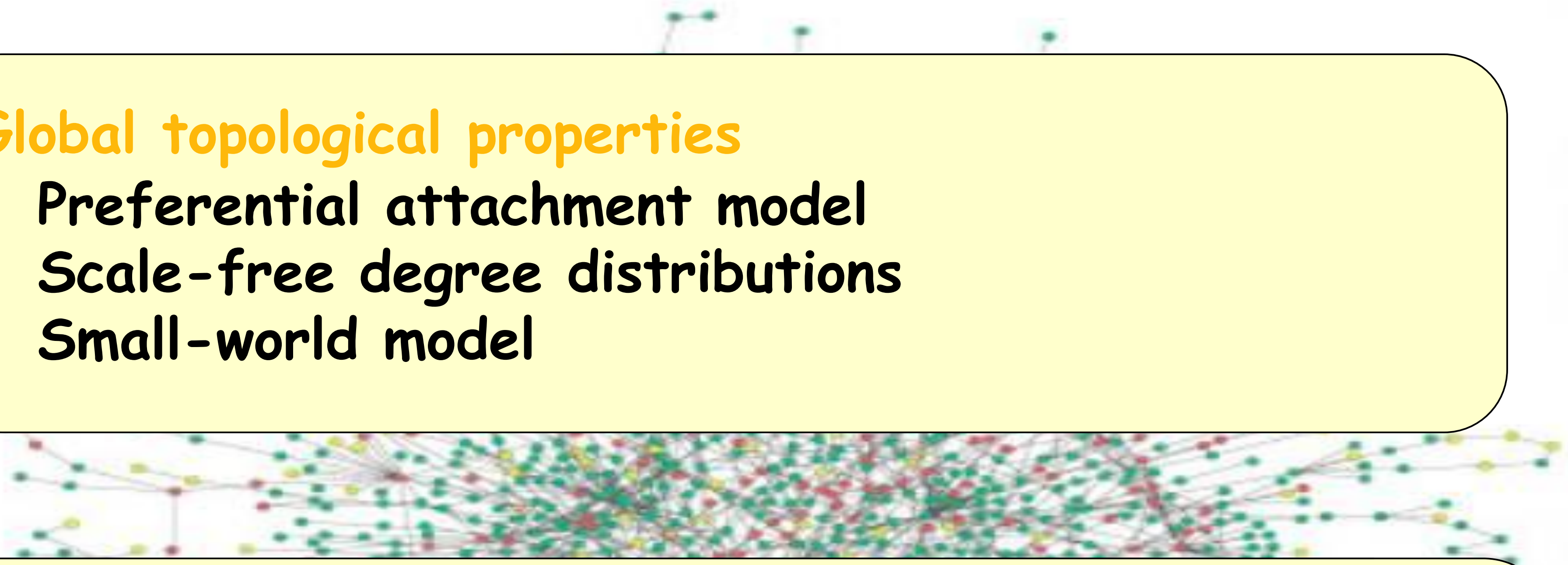
$$G=(V, E)$$
$$V=\{1, 2, 3, 4, 5, 6, 7\}$$
$$E = \{(1, 2), (1, 3) (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (2, 6), (3, 4), (3, 5), (3, 7), (4, 5), (4, 6), (4, 7), (5, 7), (6, 7)\}$$



- How many vertices does a network have?
- How many links does a network have?
- Sometimes, we need to handle huge networks

- Networks have Node-Link structure
  - Each **node/Vertex** represents an object, .e.g a person, location (train/bus station, web page), protein
  - Each **link** represents a connection

Topological Properties of Networks



Global topological properties

- Preferential attachment model
- Scale-free degree distributions
- Small-world model

Local topological properties

•Key nodes in the network (key players)

•Closely connected node groups (communities)

•Recurring patterns of local connections within the network (Network motifs)

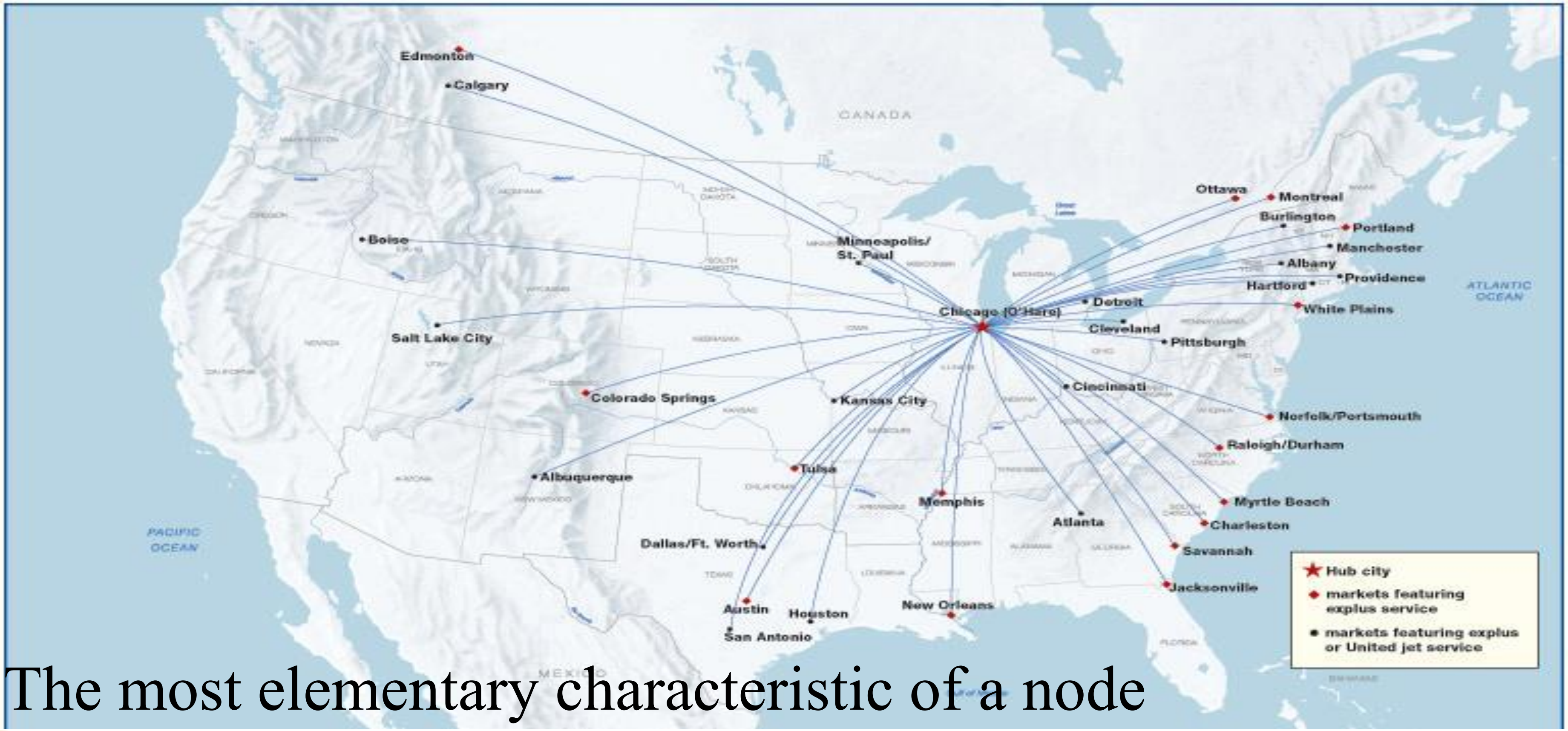
1. General Network Measures

- Important network measures to understand networks:
    1. **degree** k
    2. **degree distribution** $P(k)$
    3. **scale-free networks and degree exponent** $P(k) \sim k^{-r}$
    4. **shortest path length** l **and mean path length** $<l>$
    5. **diameter** (d)
    6. **density** (cc)
    7. **clustering coefficient** $C(k)$
- These measures can be used to compare and characterize different complex networks

•A.-L.Barabási & Z.N. Oltvai, **Nature**, 2004

General Network Measure (1): Degree (*k*) or Connectivity

- How many links a vertex has to other vertices

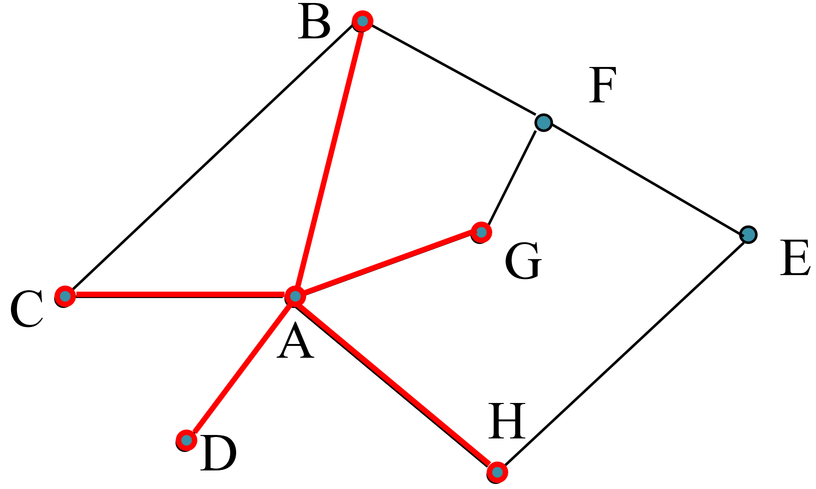


The most elementary characteristic of a node

General Network Measure (1): Example: Degree (*k*) or Connectivity

Undirected network

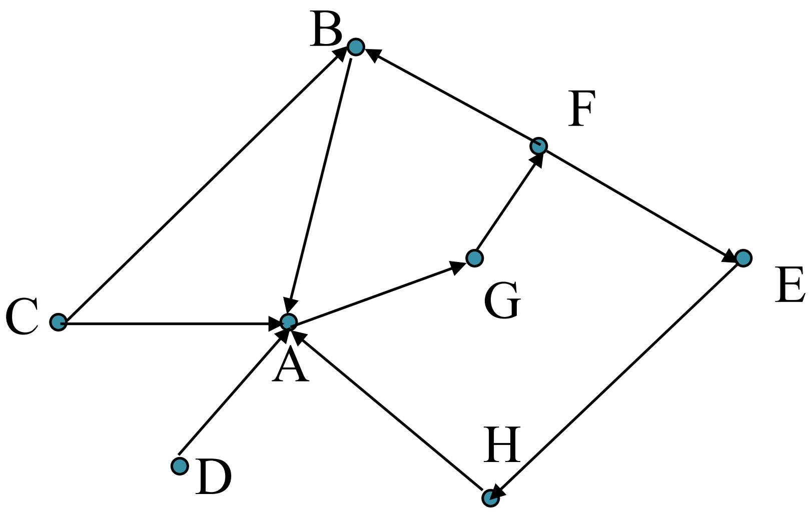
Directed network



- For node A, its **degree *k*** = 5
-

How to compute average degree of all nodes???

One link contributes two degrees (10 links and 8 vertices)



- For node A, $k_{in}=4$, $k_{out}=1$
- Average in-degree and out-degree???

Normalized degree

- Sometimes, node degree is normalized with the maximal degree $N-1$, where N is the total number of vertices
- $k(v) = k(v)$ / $(N-1)$, so normalized degree ranged [0,1].

General Network Measure (2): Degree Distribution *P*(*k*)

- $P(k)$ **gives the probability that a selected node has exactly *k* links**: $\boldsymbol{P(k) = N(k) / N}$,

$k = 1, 2, \ldots, m$ (m is the maximal degree)

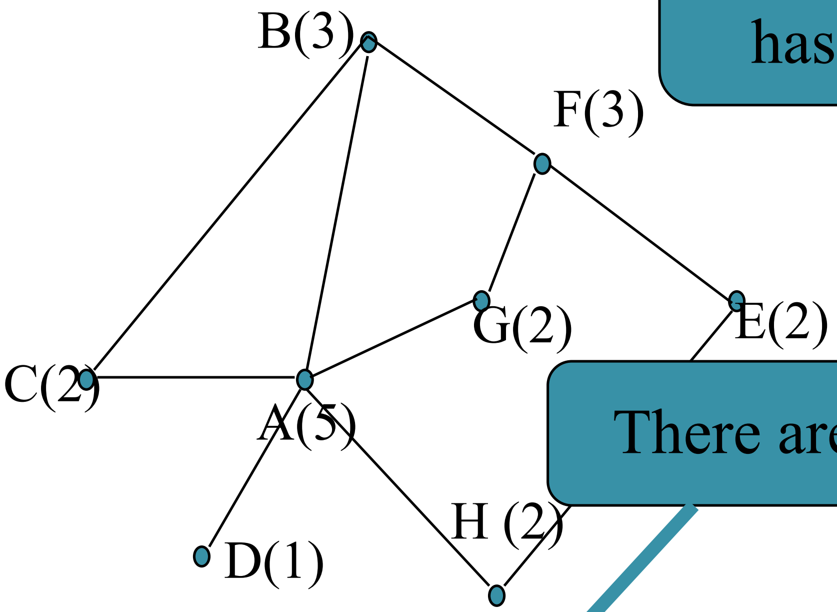
  - **$N(k)$** is *the number of nodes with degree k*
  - **N**: total number of nodes
  - Typically, the sum of degree distribution satisfied

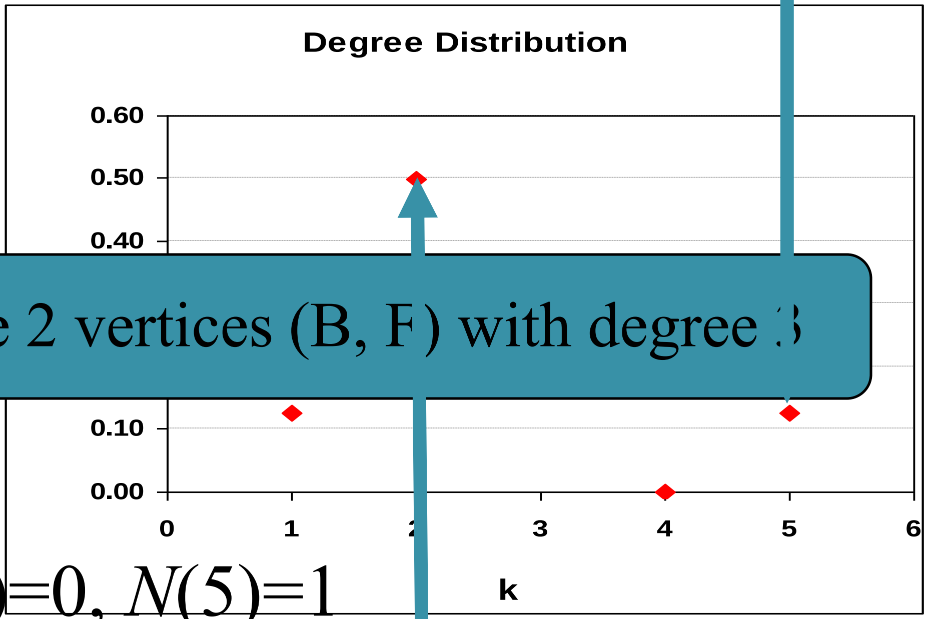
$$P(1) + P(2) + \ldots + P(m) = 1$$

- $P(k)$ can be used to classify various networks

Example of Degree Distribution $P(k)$

$P(k) = N(k) / N$, in this graph, $N=8$





The probability that a selected node has exactly maximal (5) links is small

There are 2 vertices (B, F) with degree 3

$N(1)=1$, $N(2)=4$, $N(3)=2$, $N(4)=0$, $N(5)=1$

$P(1)=1/8=0.13$, $P(2)=4/8=0.5$,

$P(3)=2/8=0.25$, $P(4)=0/8=0$

The probability that a selected node has exactly 2 links is big

NUS National University of Singapore

ISS INSTITUTE OF SYSTEMS SCIENCE

General Network Measure (3): scale-free networks and degree exponent

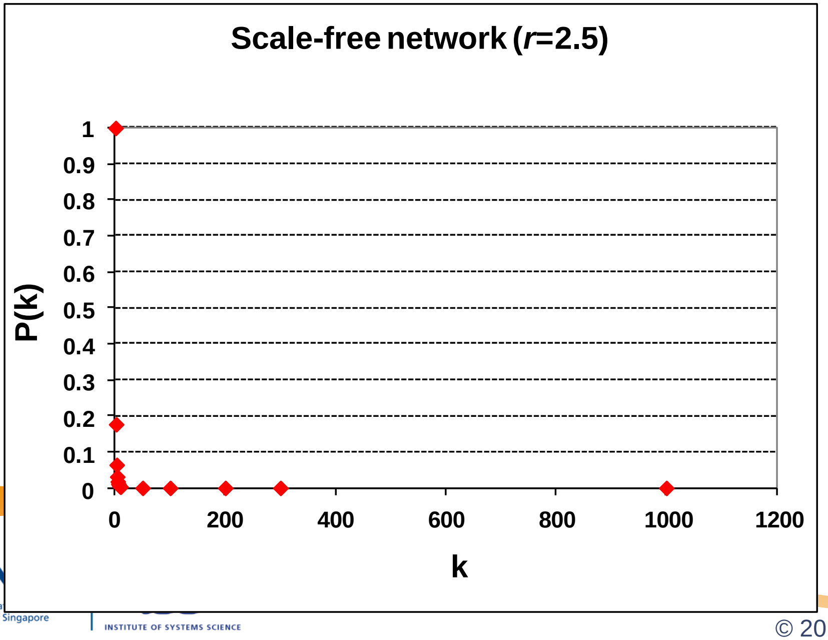
- A scale-free network: network's degree distribution $P(k)$ ***approximates*** a power law, $P(k) \sim k^{-r}$
  - Where *r is the degree exponent*
  - *'~' indicates 'proportional to'*
  - *2<r<3, which is observed in many complex social, biological and other networks*
  - ***Approximate*** means it doesn't strictly follow

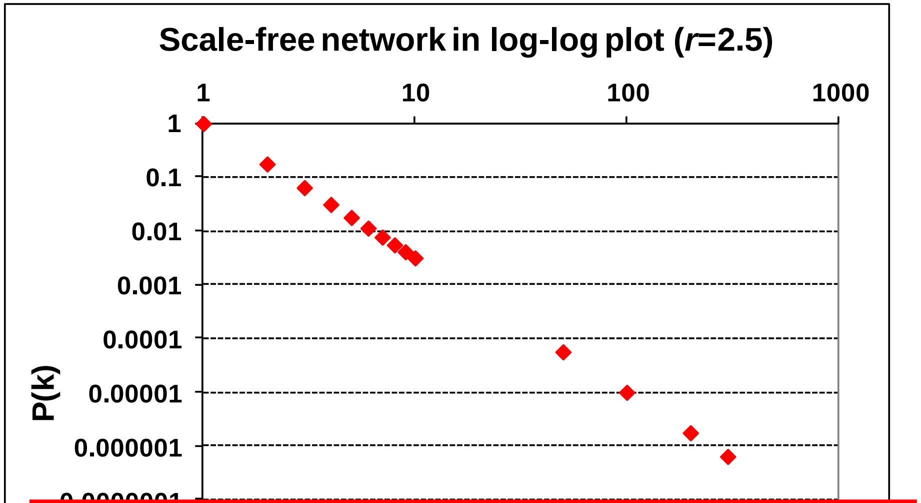
Example: scale-free network $P(k)=k^{-2.5}$

Equation P(1) + P(2) + … + P(m) = 1 does not hold

k	1	2	3	4	5	6	7	8	9	10	50	100	200	300	1000
***P*(*k*)**	1	0.177	0.064	0.031	0.018	0.011	0.008	0.006	0.004	0.003	5.66E-05	1E-05	1.77E-06	6.42E-07	3.16E-08

Most nodes in the network have small degree; for big networks, there are relatively small number of hubs: highly connected nodes (*P*(*k*)*N: estimated number of nodes with degree *k*)





Log-log plot
-2.5 is gradient of the line

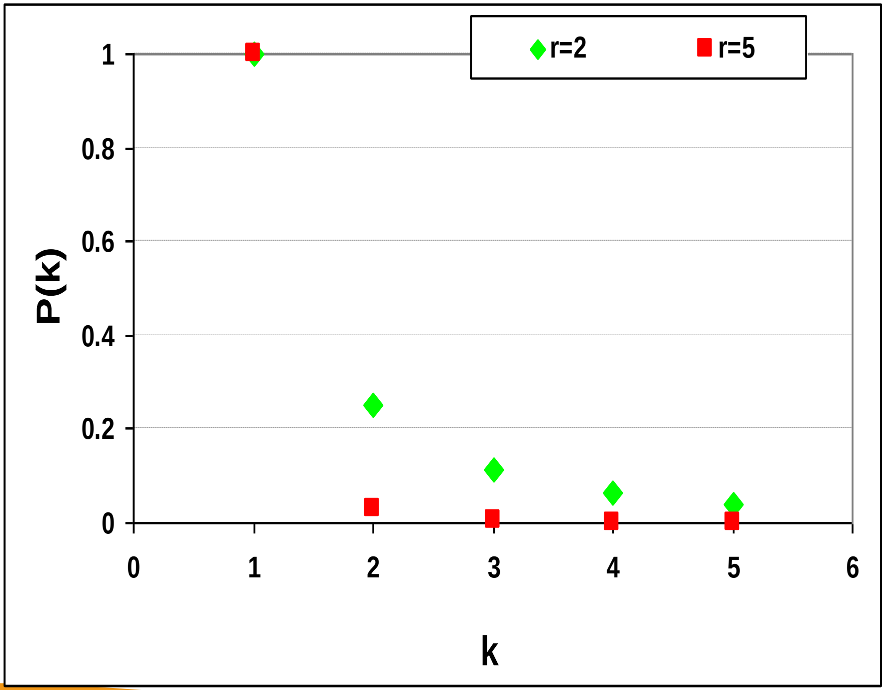
Why Log-log plot of $P(k)=k^{-r}$ is a line

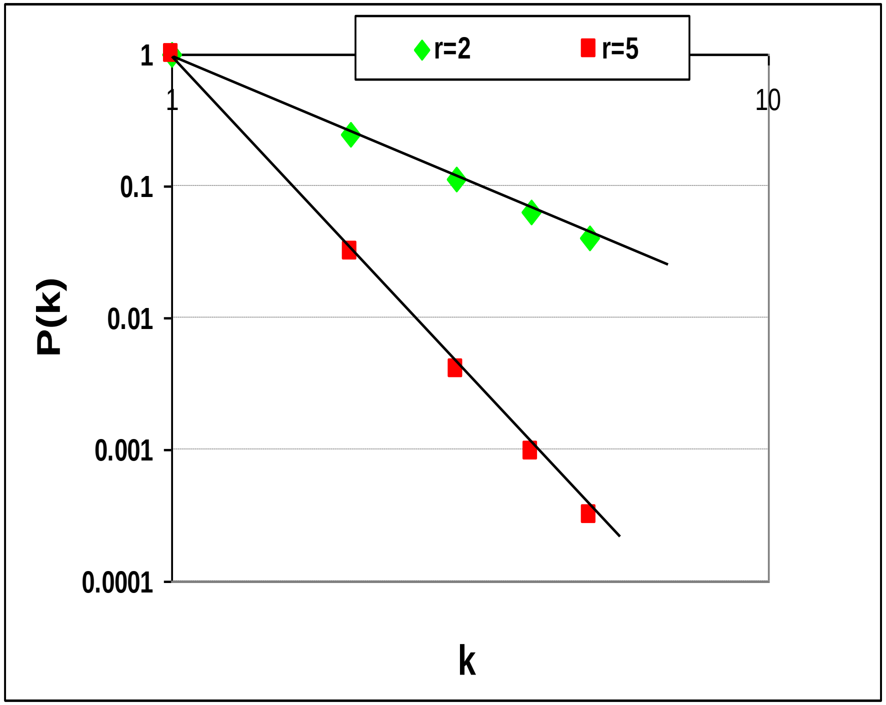
- $P(k) = k^{-2.5}$
- $log(P(k)) = log(k^{-2.5})$
- $log(P(k)) = -2.5 * log(k)$
- $Let\ Y = log(P(k))$

  $X = log(k)$
- $Y = m * X$ -> $m = -2.5$ is gradient

Compare $P(k) \sim k^{-2}$ and $P(k) \sim k^{-5}$

k	1	2	3	4	5
$P(k) \sim k^{-2}$	1	0.25	0.111111	0.0625	0.04
$P(k) \sim k^{-5}$	1	0.03125	0.004115	0.000977	0.00032





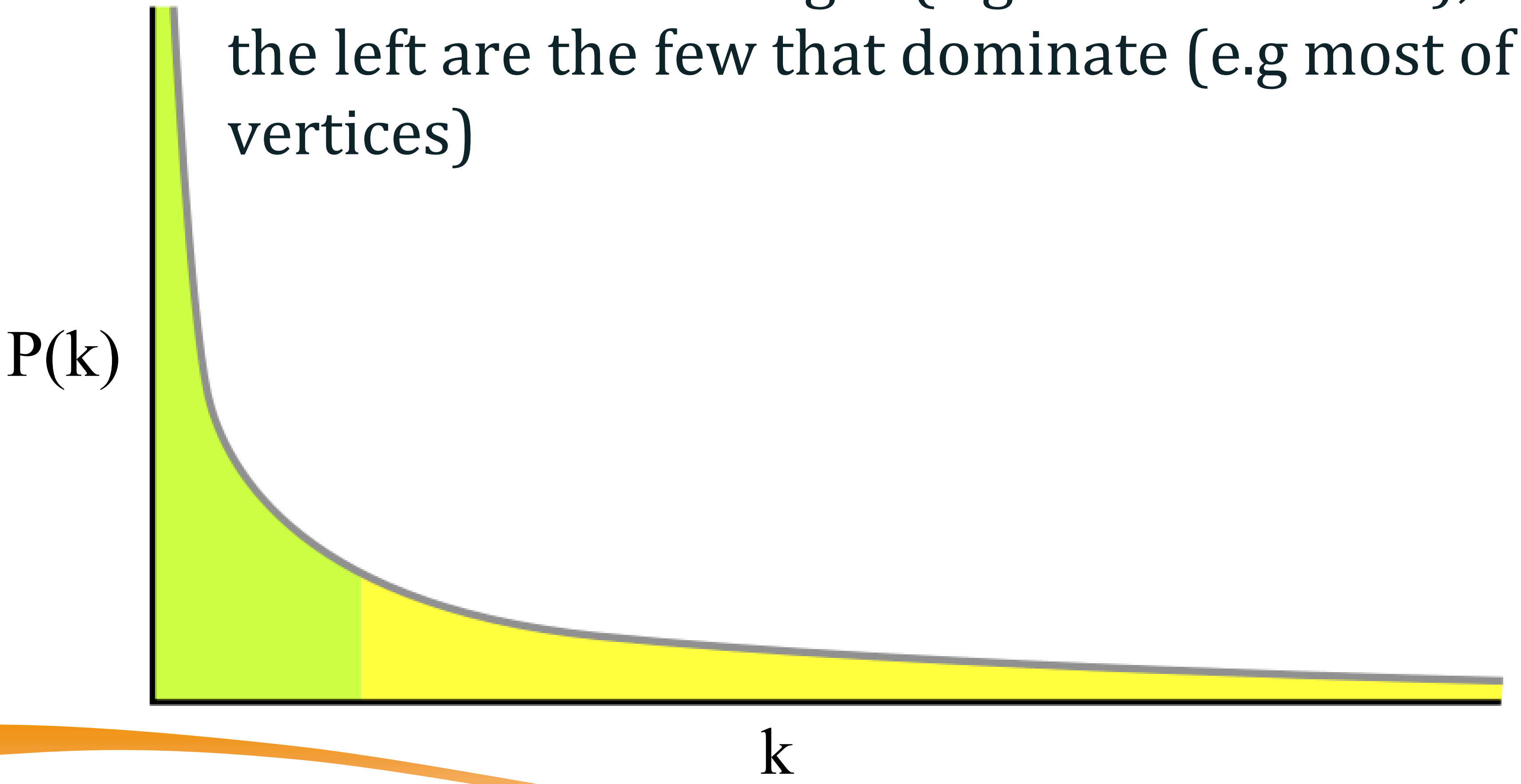
Log-log plot

Degree Exponent *r* for Scale-free Networks *P*(*k*)~*k* $^{-r}$

- *What does* degree exponent *'r' mean? How different if we have* ***r*** **=**5 *and* ***r*** **=** *2?*
  - *Given a bigger* ***r*** **(say *r* =5,** *P*(*k*)~*k*$^{-5}$**)**, *for a larger degree* ***k*** (e.g. ***k*=10**), *it has very very small number of vertices that have degree k* (the probability that a selected node has exactly 10 links is very small, *P*(10)~10 $^{-5}$ =1e-5=0.00001).
  - *Given a smaller* ***r*** **(say *r* =2,** *P*(*k*)~*k* $^{-2}$**)**, *for a larger degree* ***k*** (e.g. ***k*=10**), *it has some vertices that have degree k* (the probability that a selected node has exactly 10 links is 0.01, *P*(10)~10 $^{-2}$=0.01).

The Shape of Power law distribution

- Power law distribution is a long tail distribution to the right (e.g. a few vertices); to the left are the few that dominate (e.g most of vertices)



Why we call them *scale-free* networks?

- In scale-free networks, some nodes act as "highly connected hubs" (high degree), although most nodes are of low degrees.
- It is because there are no **typical** vertices in networks. It can also be called **scale-rich networks**. Nodes have widely different degrees (*scales*), from nodes with one or two links to major hubs joining the vertices together into a network. **Scale-*free*** *means scale-rich*

Scale-free network $P(k)=k^{-r}$

- *Given a network, typically Equation, when we compute the degree distribution,*

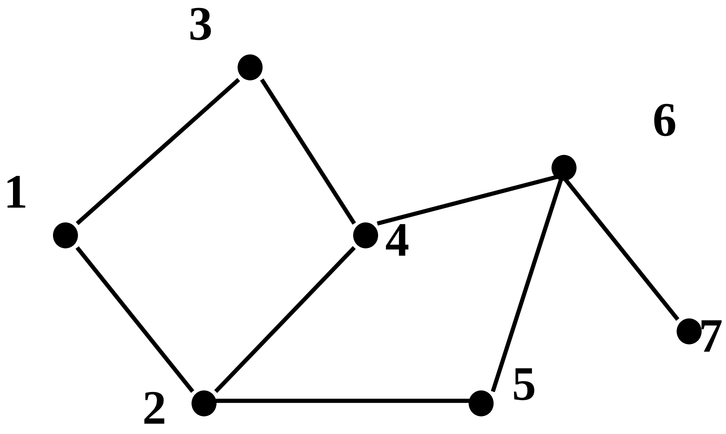
  $P(1) + P(2) + \ldots + P(m) = 1$ will hold

- *If a network* ***approximately follows*** *power low, such as* $P(k)=k^{-2.5}$*, then* $P(1) + P(2) + \ldots + P(m) = 1$ **does not hold**. The reason is $P(1)=1^{-2.5}=1$ already, so $P(1) + P(2) + \ldots + P(m) > 1$.
- A scale-free network just means that its degree distribution curve and power law curve look very similar (long tail curve in original degree distribution curve and like a line in log-log plot)

General Network Measure (4):
shortest path length *l* and mean path length <*l*>

- **Shortest path** between two nodes (*u*,*v*): the path with the smallest number of links between *u* and *v*. *l* is used to represent the **length** of a shortest path
- The **mean path length**, <*l*> represents the *average over the shortest paths between all pairs of nodes* and offers a measure of a network's overall navigability

Example of Path Length



Shortest path

$l_{12}=1$ [1↔2]

$l_{13}=1$ [1↔3]

$l_{14}=2$ [1↔3↔4] or [1↔2↔4]

$l_{15}=2$ [1↔2↔5]

$l_{16}=3$ [1↔3↔4↔ 6] or […]

$l_{17}=4$ [1↔3↔4↔6↔7] or […]

……

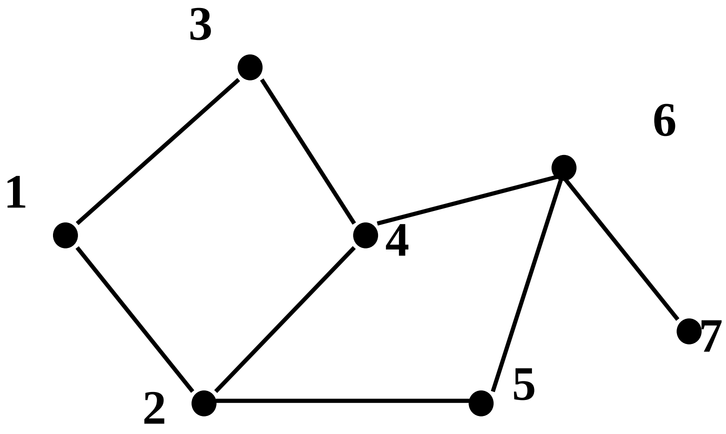
$l_{67}=1$ [6↔7]

Mean Path Length: **$< l> = (l_{12}+l_{13}+\ldots+l_{67})/(7*6/2)$**

General Network Measure (5): Network diameter *D*

- Network diameter *D*
  - Definition 1: the mean path length, <*l*> , of shortest path over all pair-wise vertices *u* and *v.*
  - Definition 2: the maximum path length of shortest path over all pair-wise vertices *u* and *v.*
- If a graph is disconnected, we assume that its diameter is equal to the diameters of *its largest* ***connected components***.

Example of Network diameter *D*



Shortest path

$l_{12}=1$ [1↔2],

$l_{13}=1$ [1↔3]

$l_{14}=2$ [1↔3↔4] or [1↔2↔4]

$l_{15}=2$ [1↔2↔5]

$l_{16}=3$ [1↔3↔4↔6]

$l_{17}=4$ [1↔3↔4↔6 ↔7]

……

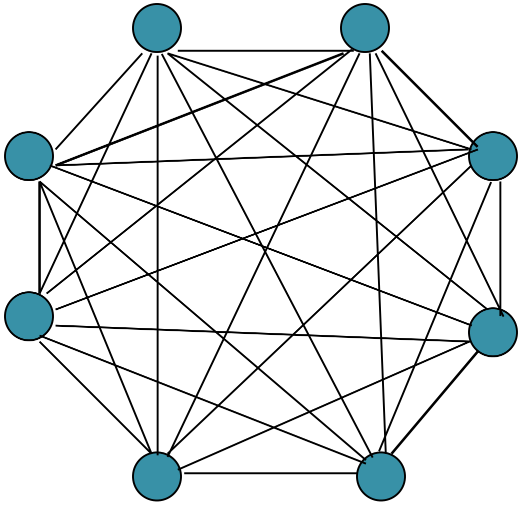
$l_{67}=1$ [6 ↔ 7]

1. Mean Path Length: $\mathbf{D} = (l_{12}{}^{+}l_{13}+\ldots+l_{67})/(7*6/2)$

2. If *maximum Path Length* is used as diameter, ***D***= 4

General network measure (6): density for a given graph

- What is the maximum number of edges for a graph with 8 vertexes?

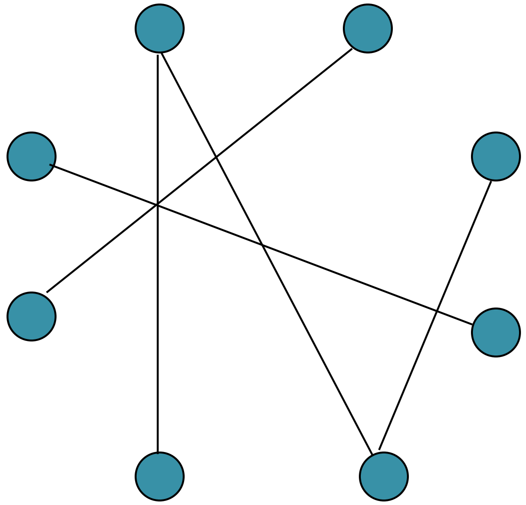


Ans: 7+6+5+…+1 = 8*(8-1)/2=28

More generally = (|V|*(|V|-1)/2)

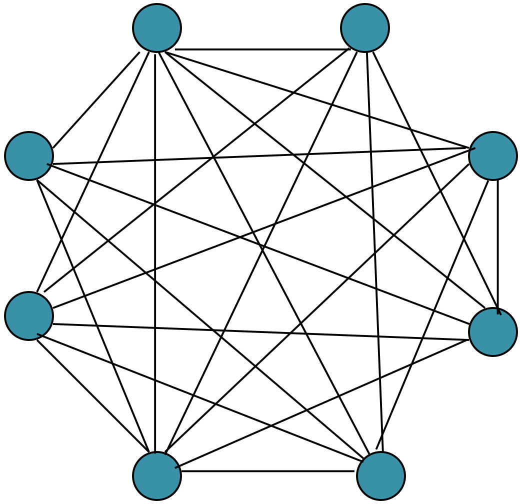
General Network Measure (6)
Example: Density for a Given Graph

- The maximum number of edges for a graph with 8 vertexes is 8*(8-1)/2=28 (|V|*(|V|-1)/2)



G1: 5 edges

Density(G1)=5/28=17.85%



G2: 24 edges

Density(G2)=24/28=85.71%

Computing the Density for a Given Graph (Cont.)

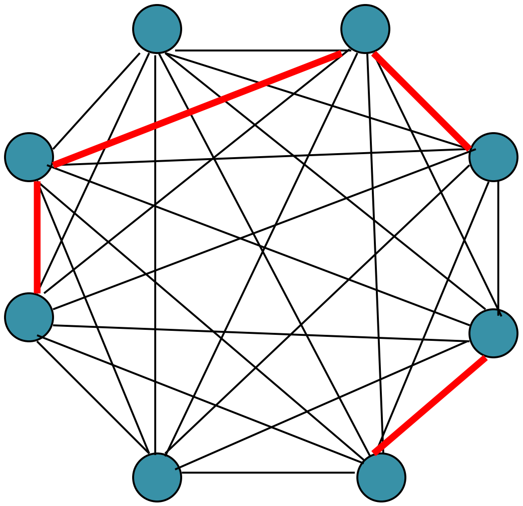
- Density of a graph G=(V,E) is defined as:

$$\text{density(G)}=\frac{|E|}{|V|*(|V|-1)/2}$$

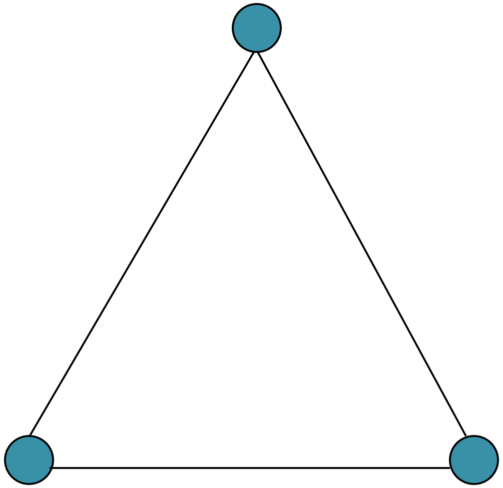
$$=2*|E|/[|V|*(|V|\text{-}1)]$$

- Basically, 0≤density(G) ≤1. If density(G) = 1, then G is fully connected graph or a clique, which has maximal number of edges.

Computing the Density for a Given Graph (Cont.)



G3



G4

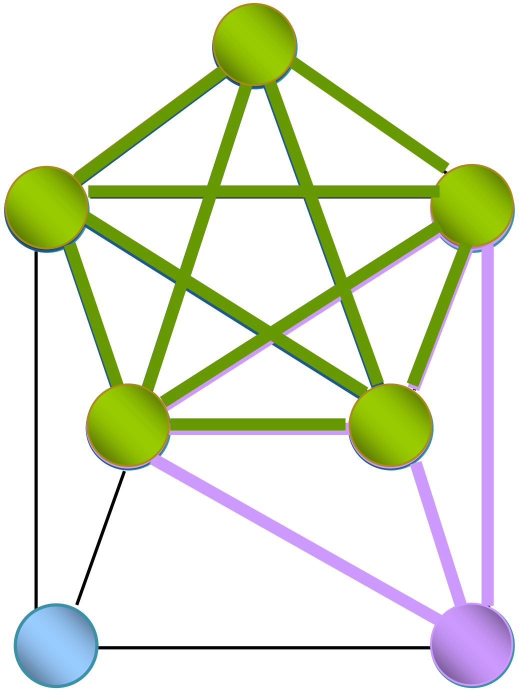
Density(G3)=28/28=1

Density(G4)=3/3=1

Both graphs G3 and G4 are cliques--- fully connected graphs

The CLIQUE Problem

$$\boxed{CLIQUE = \{<G,k>|\, G \text{ has a clique of size } k\}}$$



Maximum Clique of Size 5

Clique: a complete subgraph

Maximal Clique: a *clique* cannot be enlarged by adding any more vertices

Maximum Clique: the largest *maximal clique* in the graph

General Network Measure (7): clustering coefficient for a vertex

- Clustering coefficient (C) of a vertex v characterizes v's tendency to form clusters or groups (cliques):

$$C(v) = 2*n / (k*(k-1))$$

$$=n/(k*(k-1)/2)=2n/(k*(k-1))$$

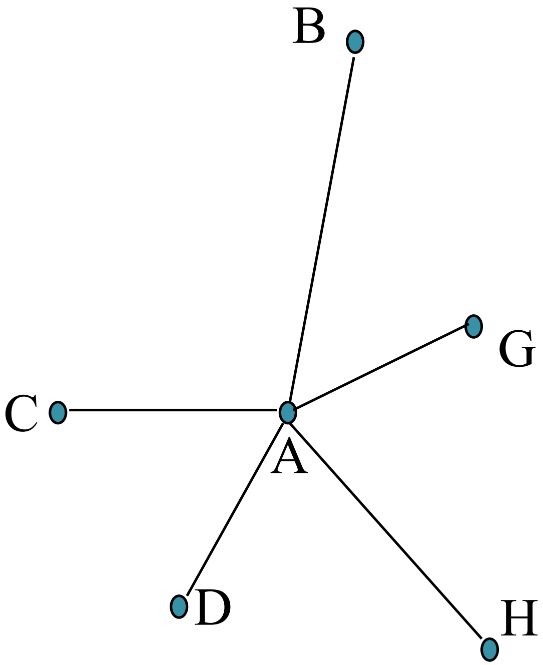
  - k is the number of v's neighbors. (k*(k-1)/2) is the *maximum* possible links among the neighbors
  - ***n* is the number of links connecting the *v*'s neighbors to each other**

D. J. Watts and Steven Strogatz (June 1998). Nature 393: 440–442.

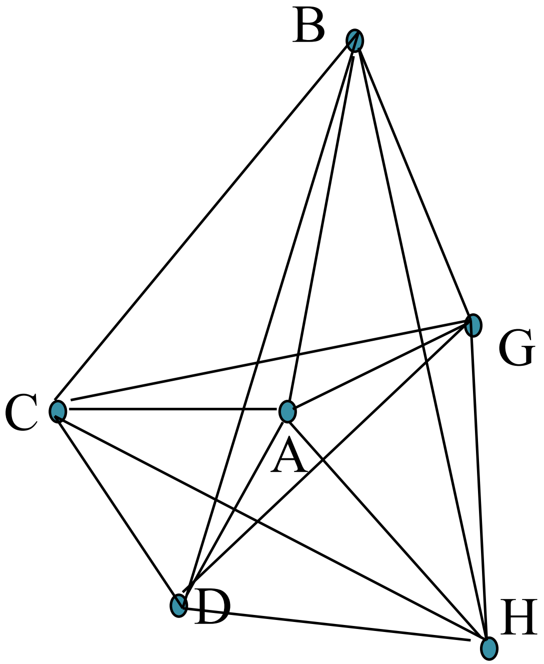
General Network Measure (6)

Example: clustering coefficient CC

- CC indicates how your friends know each other.

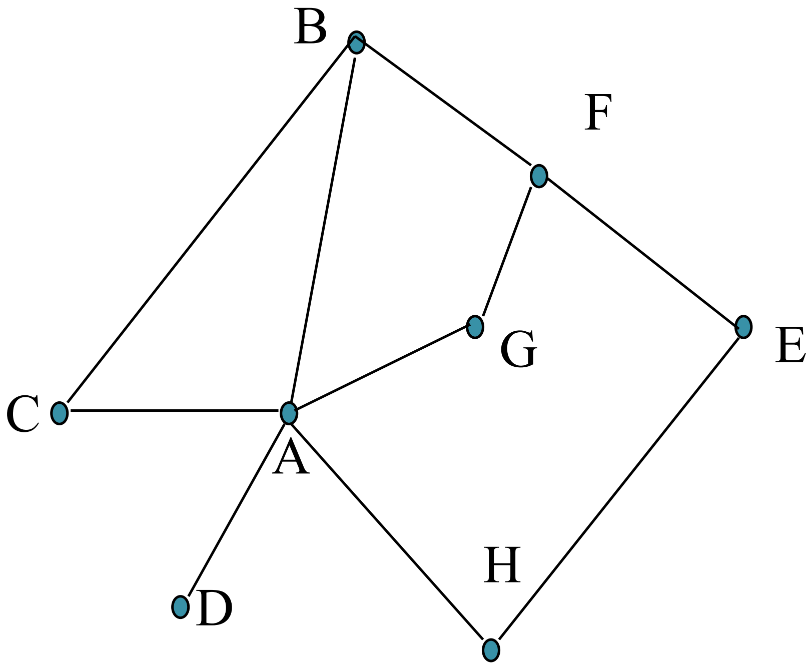


CC(A)=0



CC(A)=1

Example of clustering coefficient



$C(B)=2*n\ /\ (k*(k\text{-}1))$

$=2*1/(3*2)=1/3$

$k = 3$ (C, A, F), n=1 (CA)

$C(A)=2*n\ /\ (k*(k\text{-}1))$

$=2*1/(5*4)=1/10$

$k = 5$ (C, B, G, H, D),

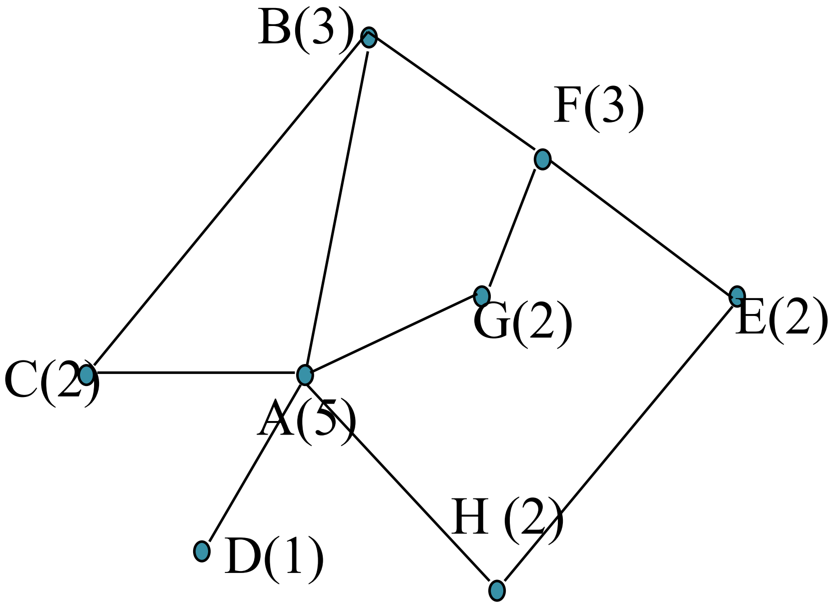
$n = 1$ (CB)

$C(A)$=1/10, $C(B)$=1/3, C(C)=1, C(D)=0,

C(E)=0, C(F)=0, C(G)=0, C(H)=0

Example: Clustering coefficient distribution $C(k)$

$C(k)$ is defined as the average clustering coefficient of all nodes with k links. For many real networks, $C(k) \sim k^{-1}$



$C(A)$=1/10, $C(B)$=1/3,
C(C)=1, C(D)=0,
C(E)=0, C(F)=0,
C(G)=0, C(H)=0

C(1)=C(D)=0
C(2)= [C(C)+ C(H)+ C(G)+C(E)]/4=1/4
C(3)= [C(B)+ C(F)] /2=1/6
C(4)= 0
C(5)= C(A)=1/10

What does k^{-1} mean?
Big k, C(k) is small
Small k, C(k) is big

Example: Clustering Coefficient Distribution

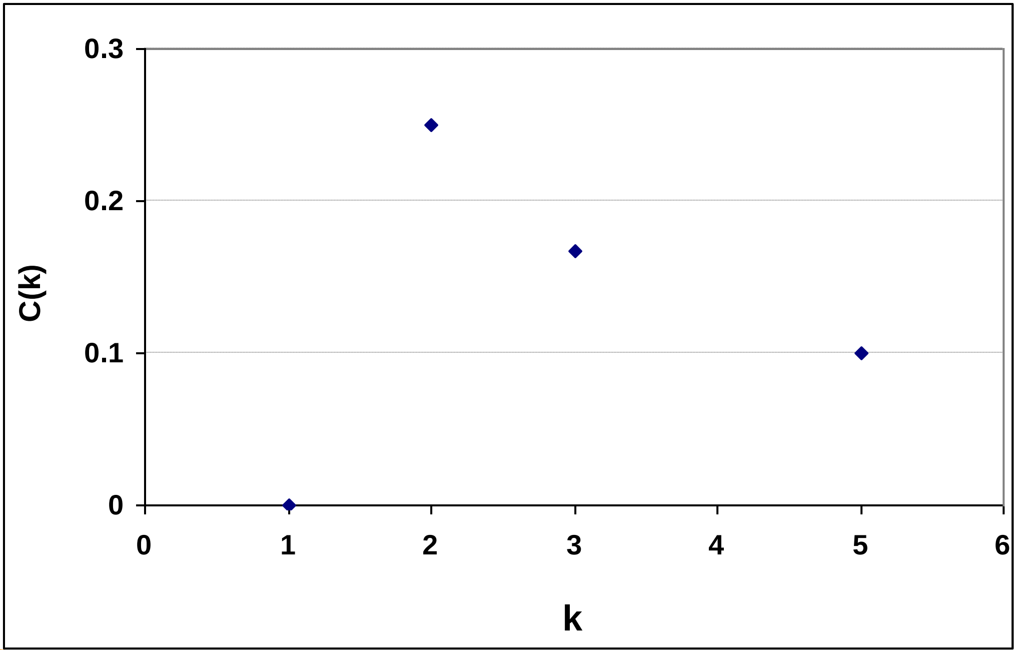
C(1)=C(D)=0

C(2)= [C(C)+ C(H)+ C(G)+C(E)]/4=1/4

C(3)= [C(B)+ C(F)] =1/6

C(4)= 0

C(5)= C(A)=1/10



Summary of Network Measures

1. degree k
2. degree distribution $P(k)$,
3. scale-free networks and degree exponent $P(k) \sim k^{-r}$,
4. shortest path length l and mean path length $<l>$
5. diameter (d): two definitions
6. density
7. clustering coefficient distribution $C(k)$

Graph & Network Mining: Concepts

- 1. General Network Measures
- 2. Typical Architecture of Networks
- 3. Social network analysis

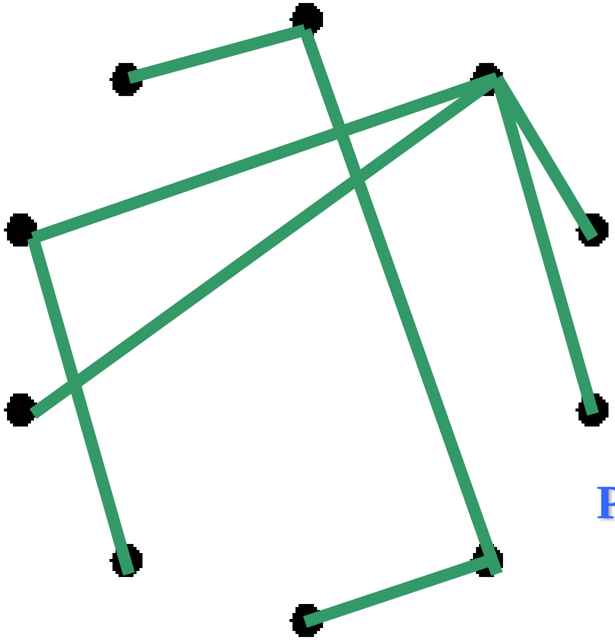
2. Typical Architecture of Networks

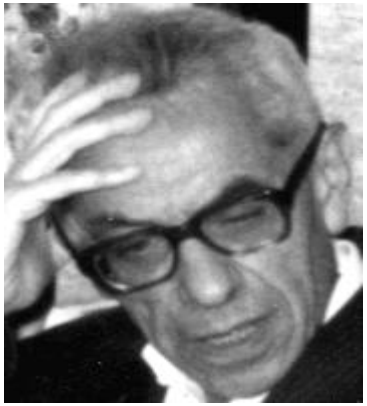
- 1. Random networks
- 2. Scale-free networks
- 3. Scale-free networks with modularity structure

Key difference: *degree distribution* and *clustering coefficient distribution*

2.1 Random Networks (Erdös-Rényi model 1960)

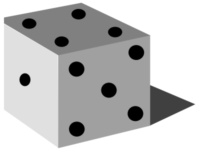
A random network starts with N nodes and connects each pair of nodes with probability *p*, which creates a graph with approximately *p*N(N–1)/2 randomly placed links





Pál Erdös (1913-1996)

Connect each pair of nodes with probability p



p=1/6

N=10

- **Democratic**
- **Random**

L≈(1/6)*(10*9/2)=7.5

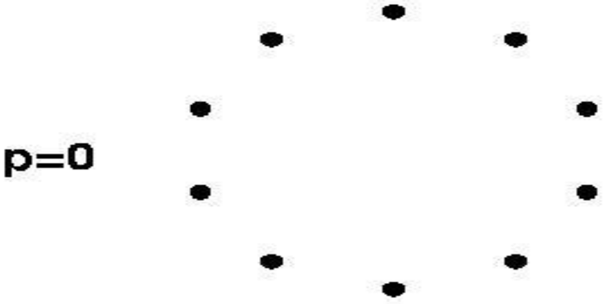
<k>≈2*L/N

=2*7.5/10=1.5

Examples of Random Networks

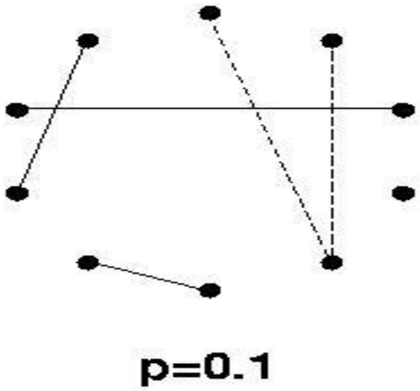
A graph with approximately *p*N(N–1)/2 randomly placed links (N=10, N*(N-1)/2=10*9/2=45)

0 edge

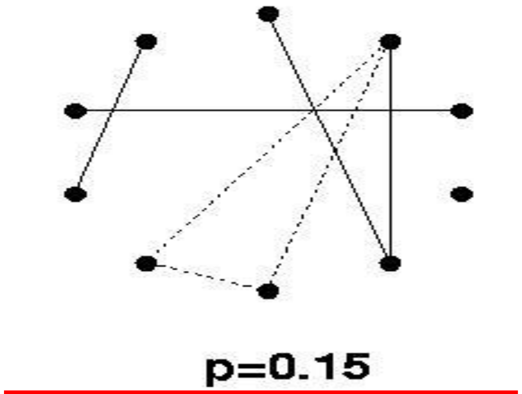


L=0*45=0

5 edges



L=0.1*45=4.5



7 edges

L=0.15*45=6.75

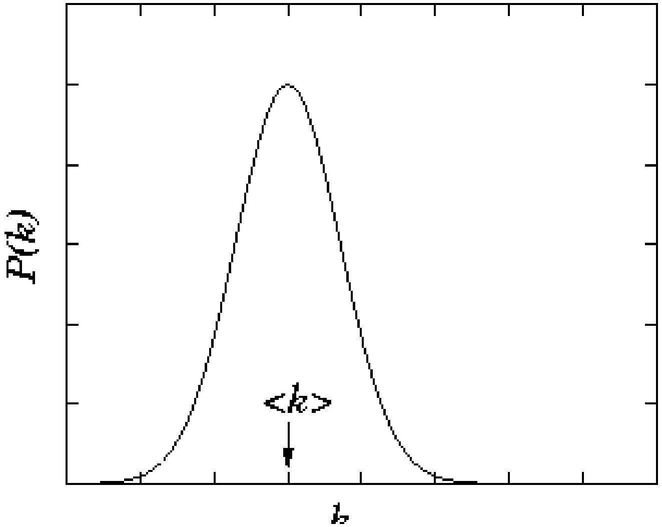
Democratic: Not very easy to form hubs since every node gets the same opportunity (p) to connect to other nodes

Random Networks (degree distribution)

St. Paul
Sioux Falls
Milwaukee
Detroit
Buffalo
Cedar Rapids
Chicago
Omaha
Cleveland
Peoria
Indianapolis
Philadelphia
Kansas City
St. Louis
Washington
Wichita
Lexington-fayette
Richmond
Tulsa
Nashville-davidson
Durham
Memphis
Chattanooga
Charlotte
Birmingham
Dallas
Shreveport
Jackson
Columbus
Savannah

Highway systems of the United States

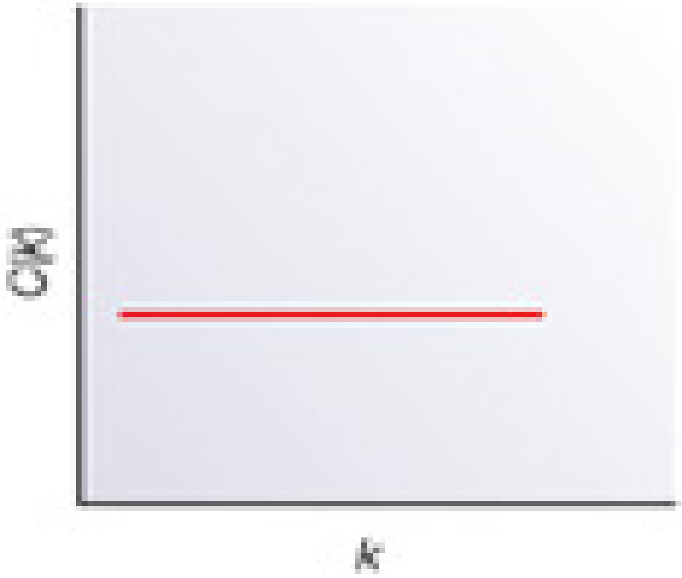
Degree Distribution *P*(*k*)



The node degrees follow a Poisson distribution, which indicates that most nodes have approximately the same number of links (close to the average degree <*k*>). The tail (high *k* region) of the degree distribution *P*(*k*) decreases exponentially, which indicates that nodes that significantly deviate from the average are extremely rare.

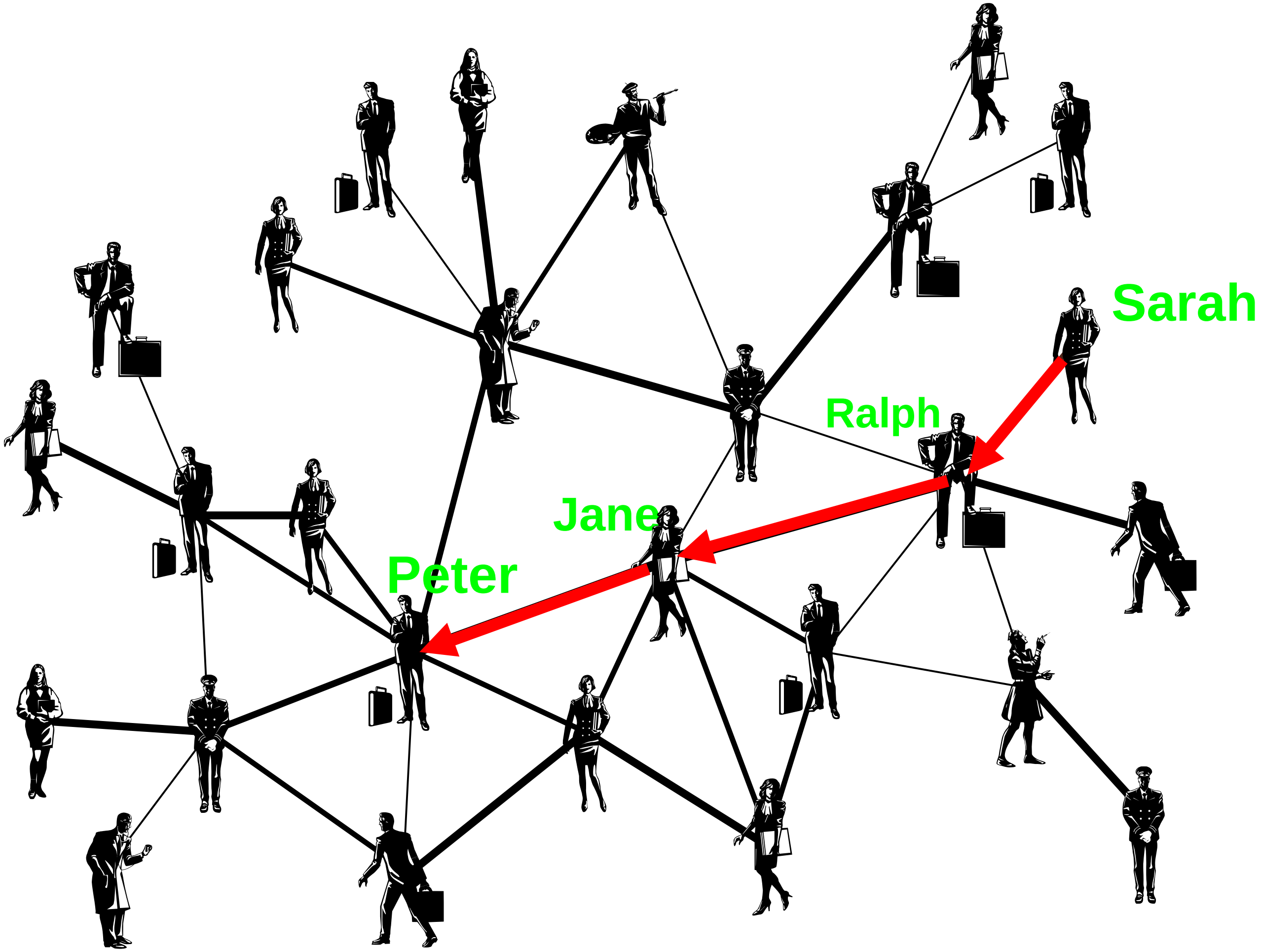
Random Networks (clustering coefficient and mean path length)

- The clustering coefficient is independent of a node’s degree, so $C(k)$ appears as a horizontal line if plotted as a function of k.



- The *mean path length* is proportional to the logarithm of the network size, $\langle l \rangle \sim \log N$, which indicates that it is characterized by the small-world property.

Small worlds



Society:

Six degrees

S. Milgram 1967

F. Karinthy 1929

WWW:

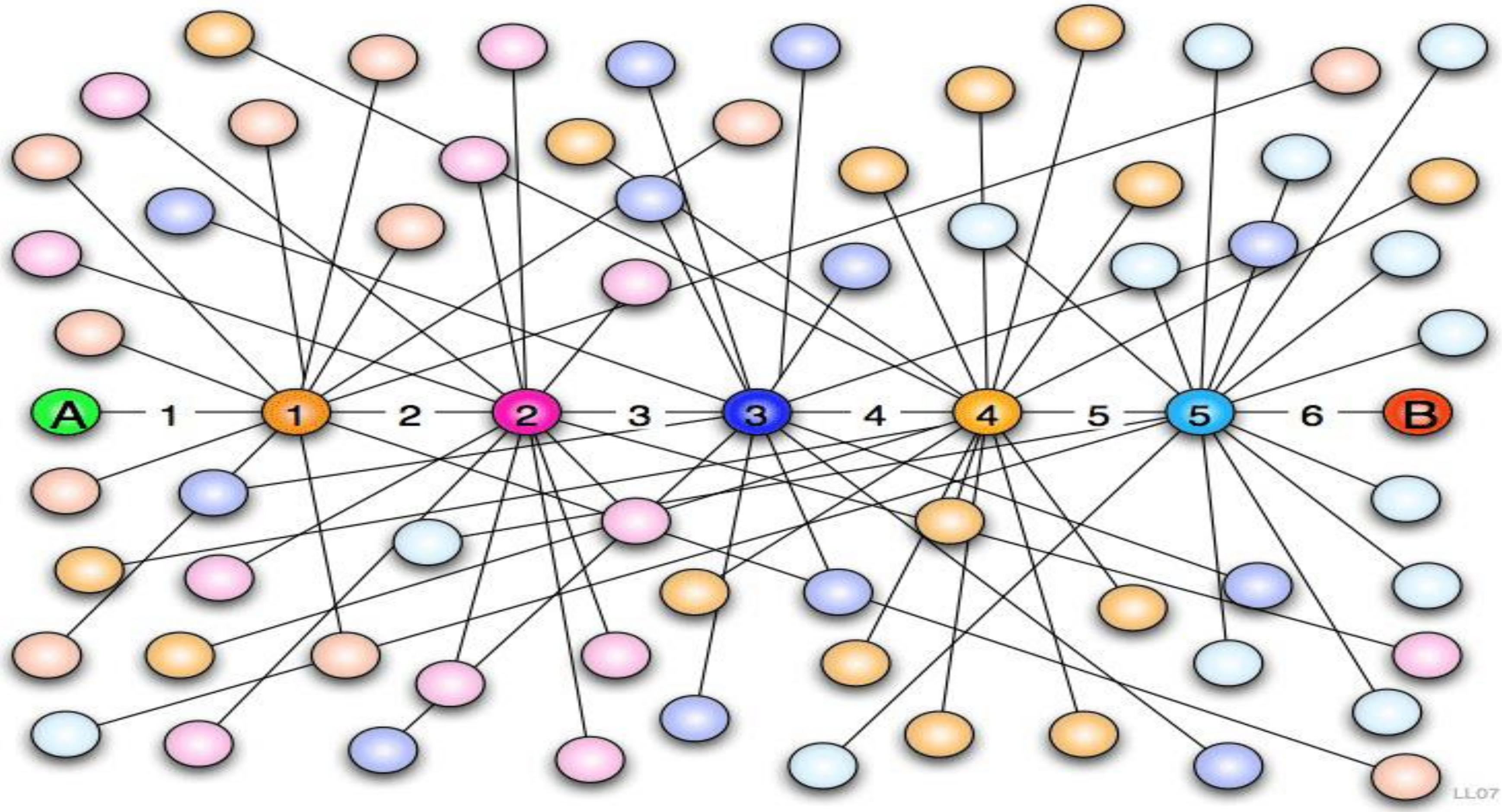
19 degrees

Albert *et al.* 1999

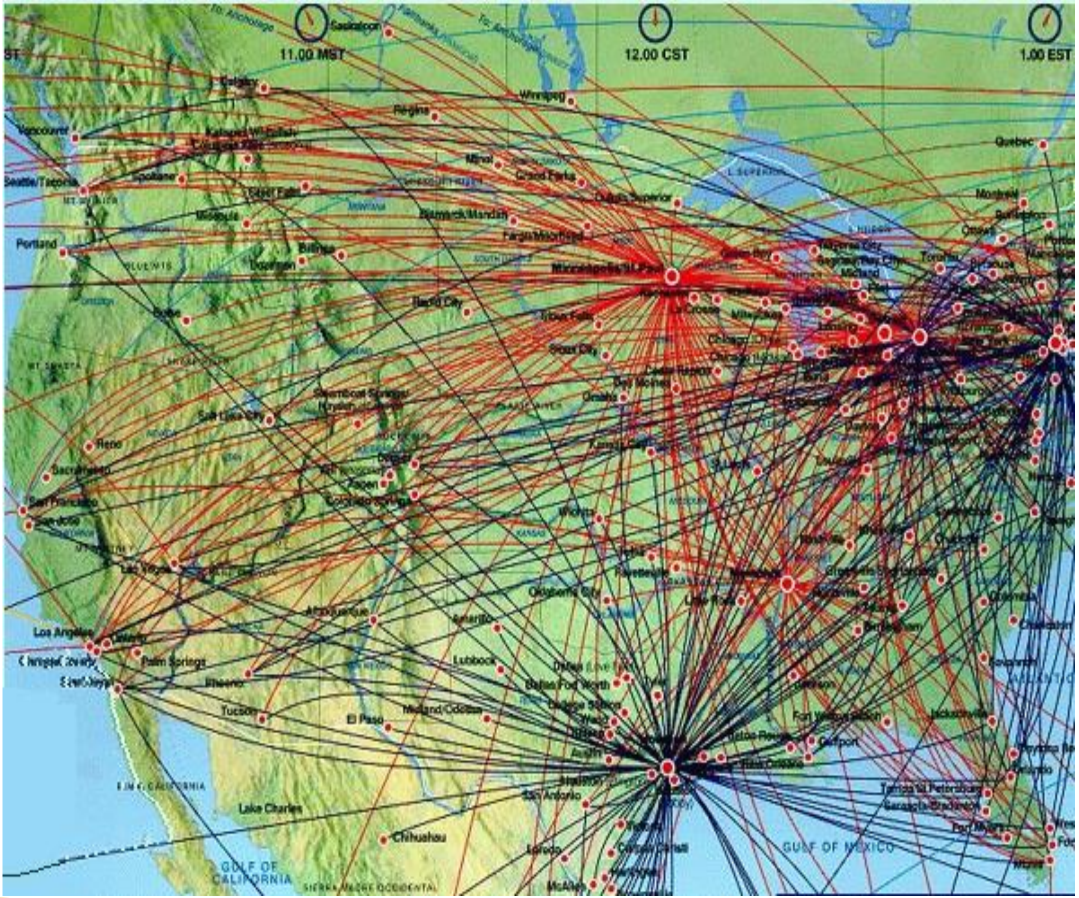
Milgram's experiment

- Measure the probability that two randomly selected people would know each other.
- Milgram chose individuals in the U.S. cities of Omaha (the starting points) and Boston (the end point) of a chain of correspondence.
- Information packets were sent to randomly selected individuals in Omaha, including basic information about a target contact person in Boston. A recipient was asked whether he knew the contact person? If yes, the person was to forward the letter directly to that person; else the person was to think of a friend they know that is more likely to know the target.
- Conclusion: the average path length fell around 5.5 or six-people in the United States are separated by about six people on average.

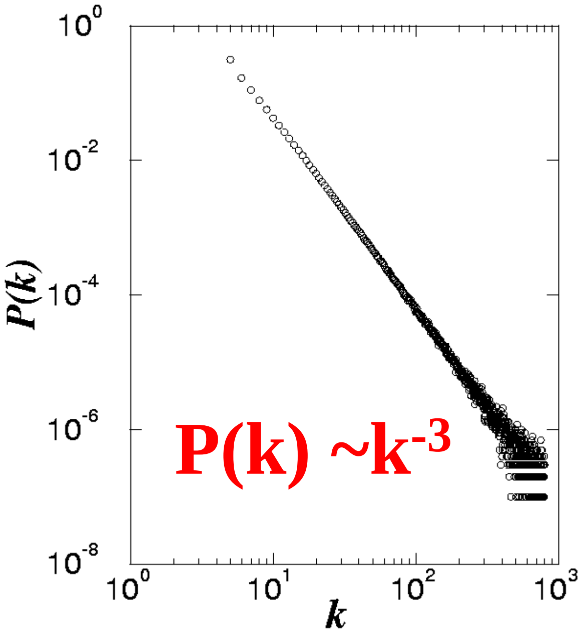
- **Six degrees of separation**: if a person is one step away from each person he or she knows and two steps away from each person who is known by one of the people he or she knows, then everyone is an average of six "steps" away from each person on Earth.



2.2 Scale-free networks



Degree Distribution follows power law



$P(k) \sim k^{-3}$

A.-L.Barabási, R. Albert, Science **286,** 509 (1999)

Many real world networks have the same architecture:

Scale-free networks

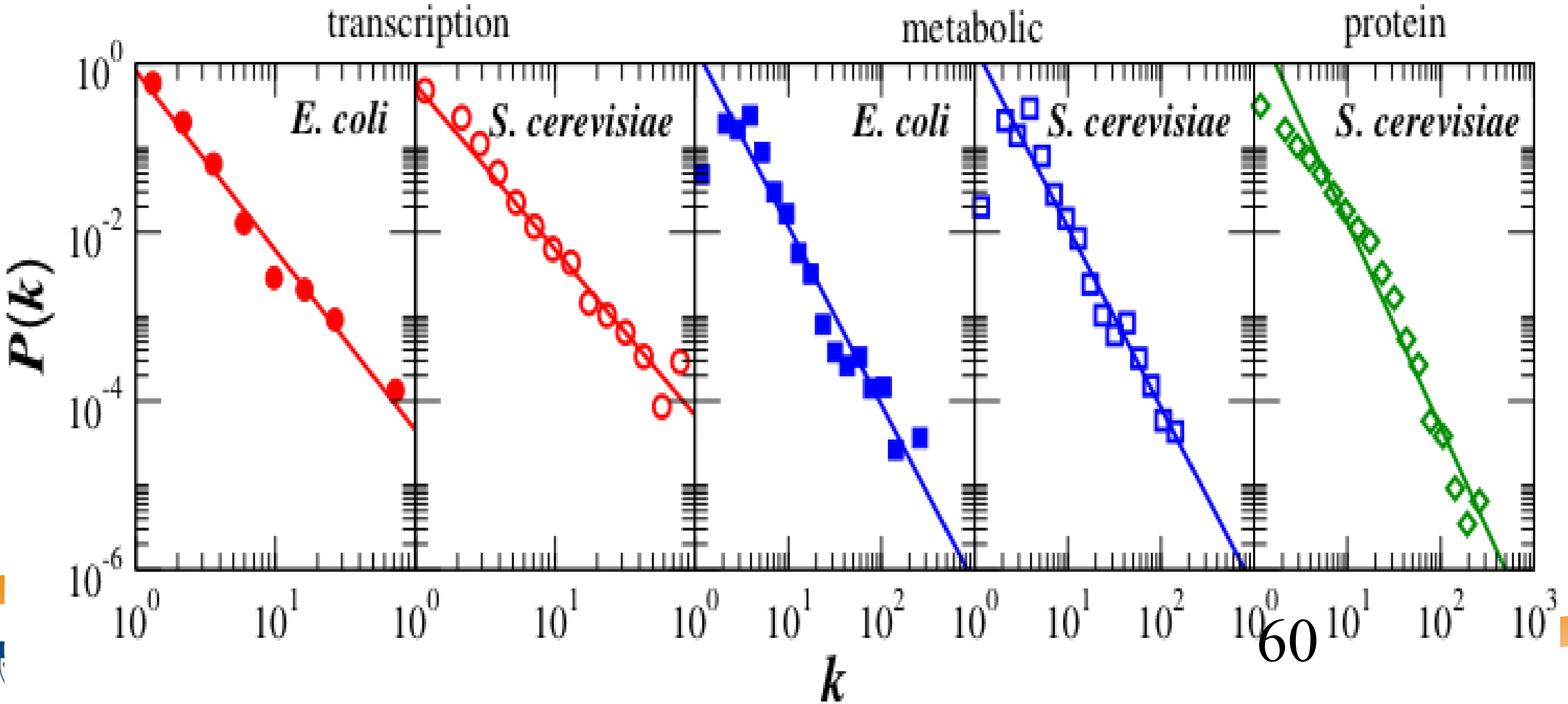
WWW, Internet (routers and domains), electronic circuits, movie actors, coauthorship networks, sexual web, instant messaging, email web, citations, phone calls, metabolic, protein interaction, protein domains, brain function web, linguistic networks, comic book characters, international trade, bank system, encryption trust net, energy landscapes, earthquakes, astrophysical network…

Cellular networks are scale-free

- **Degree distribution**

  P(k): probability that a node has *k* neighbors.

$$P(k) \sim k^{-r}$$

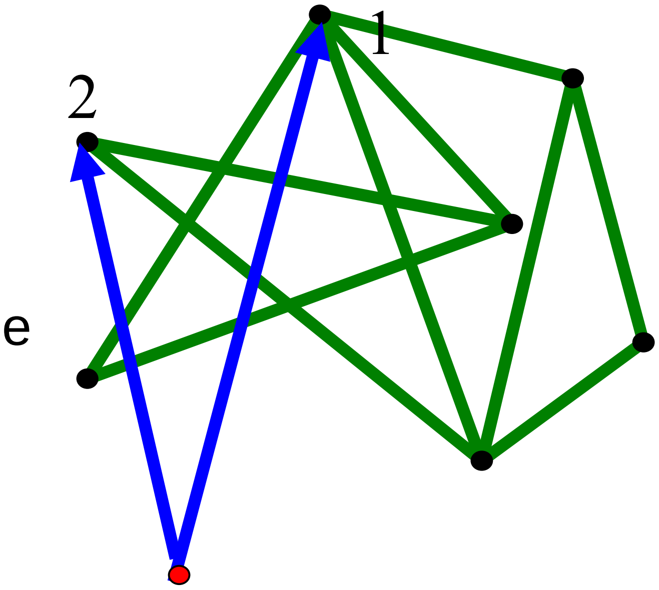


Why Scale-free Model

(1)GROWTH (different from random network model)

At every timestep, the network grows by adding a new node with *m* edges connected to the nodes already present in the system.

Network emerges through the subsequent addition of new nodes



A new node

(2) PREFERENTIAL ATTACHMENT

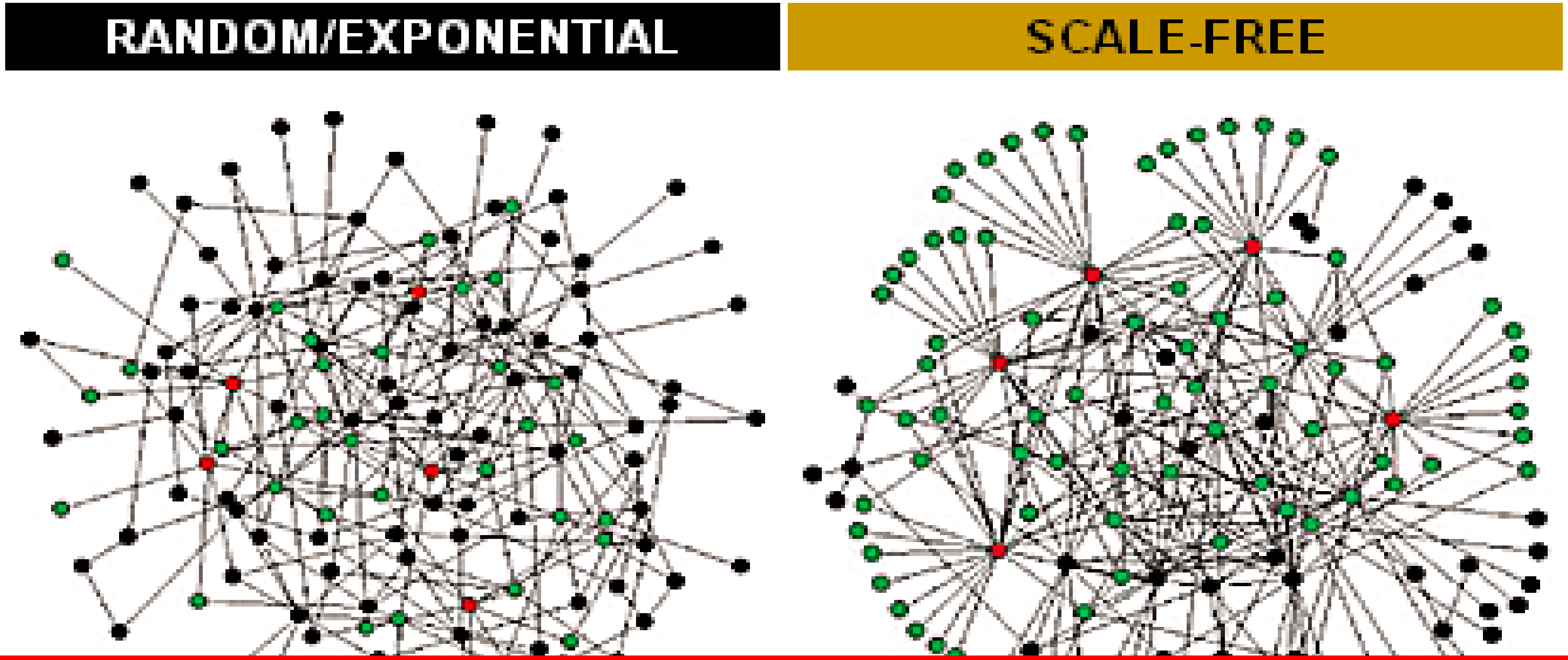
The probability **Π** that a new node will be connected to an existing node *i* depends on the connectivity k_i of that node

$$\Pi(k_i) = \frac{k_i}{\Sigma_j k_j}$$

The probability that the red node will connect to vertex 1 ($k_1 = 4$) is twice as large as connecting to node 2 ($k_2 = 2$)

Network Hubs in Scale-Free Networks

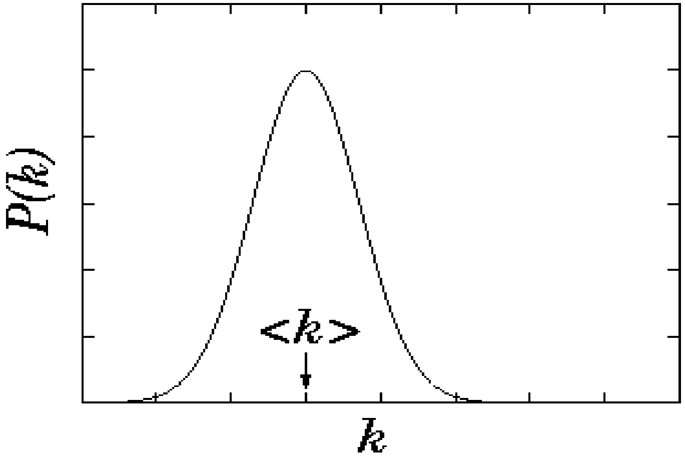
In the random network, the five nodes with the most links (in red) are connected to only 27% of all nodes (green). In the scale-free network, the five most connected nodes (red) are connected to 60% of all nodes (green).



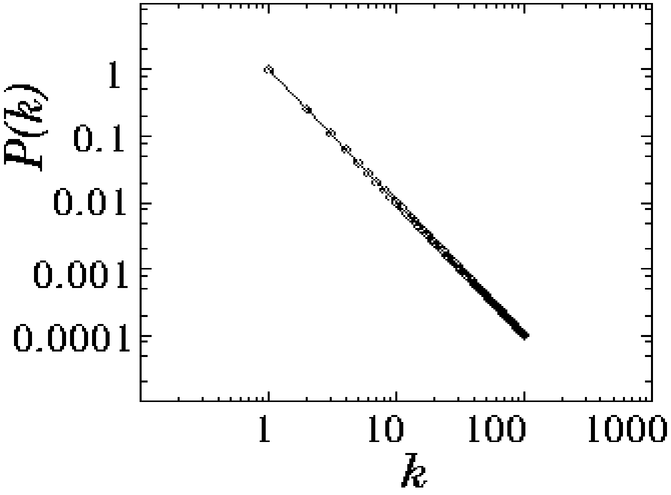
- The probability that a node is highly connected is statistically more significant than in a random graph
- The network’s properties are often being determined by a relatively small number of hubs

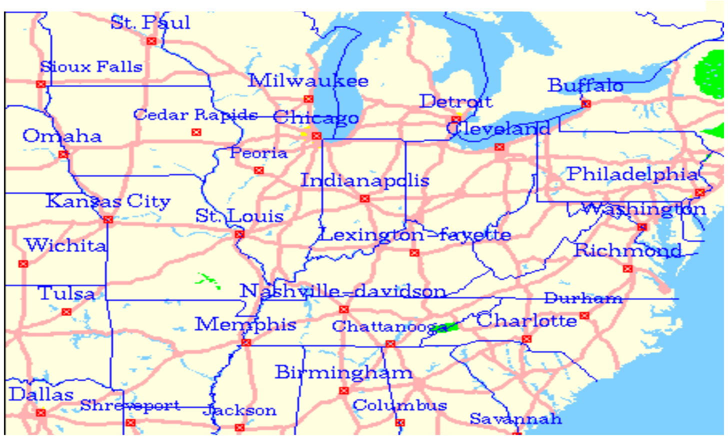
What Does This Mean?

Poisson distribution

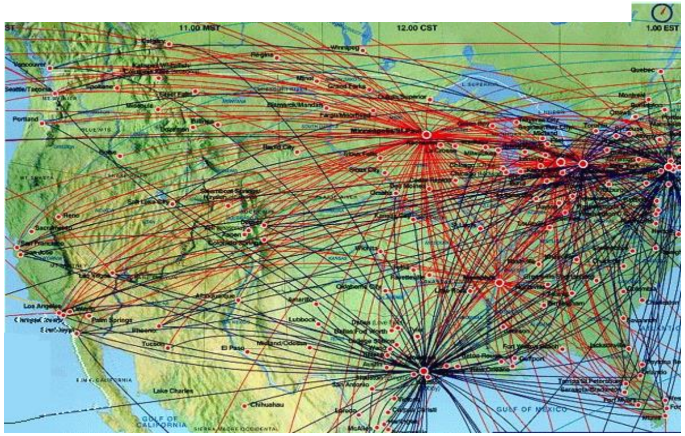


Power-law distribution (*long tail*) (or *line* in log-log plot)





Random Network



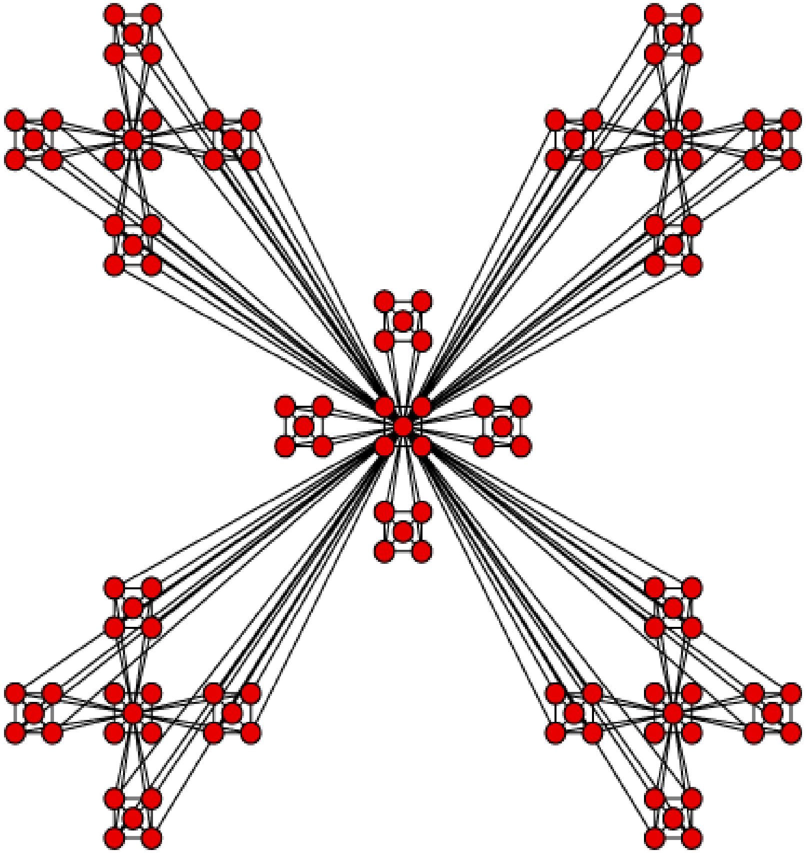
Scale-free Network

Nodes that significantly deviate from $<k>$ are extremely rare

The network properties are determined by a relatively small number of hubs

2.3 Scale-free networks with modularity structure

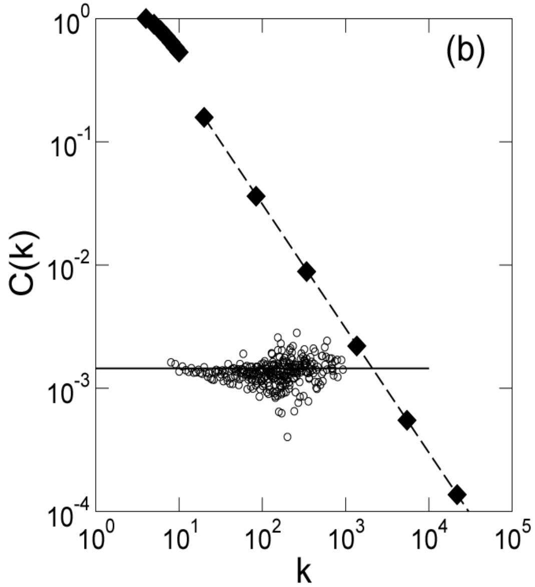
- To account for the coexistence of modularity, local clustering and scale-free topology in many real systems it assumed that clusters combine in an iterative manner, generating a hierarchical network

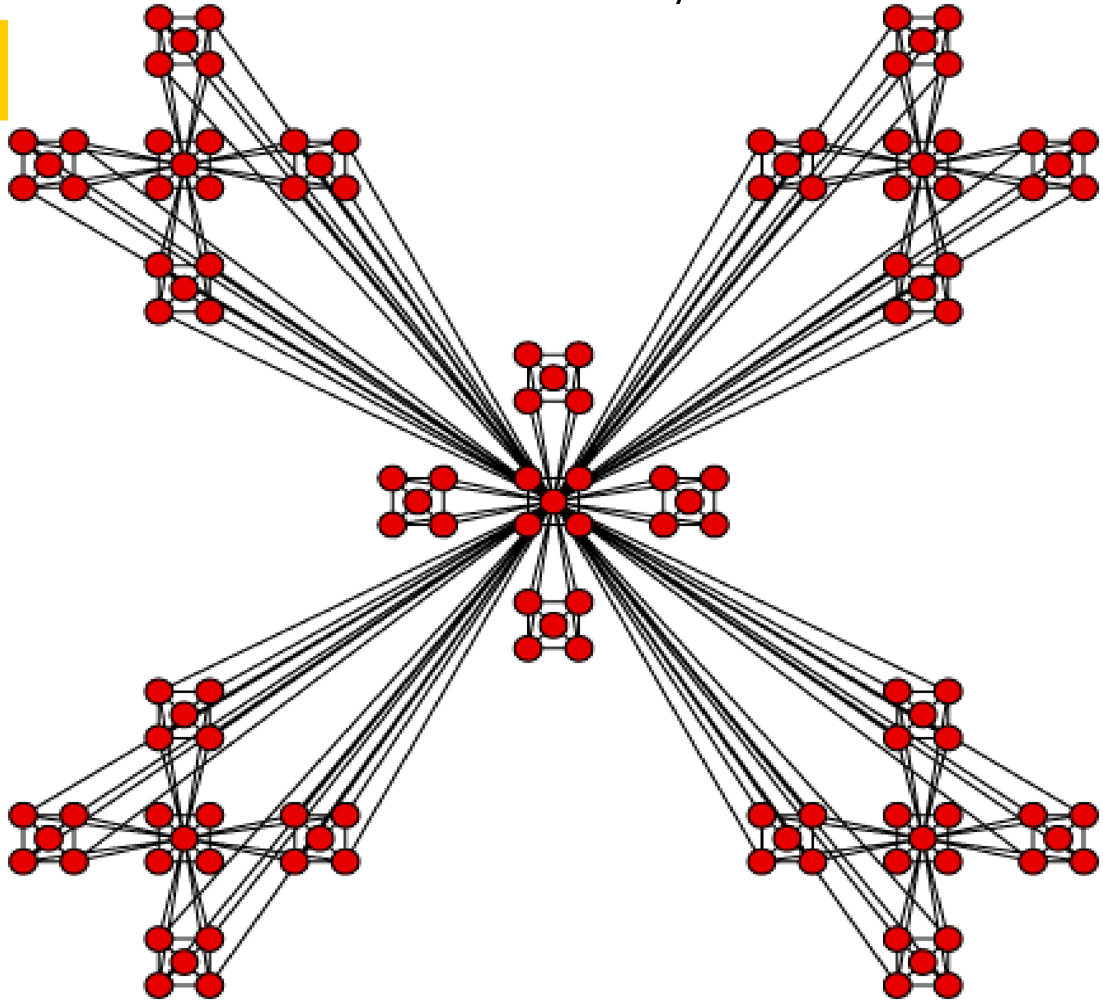


Scale-free networks with modularity structure

Major Characteristic

$$C(k) \sim k^{-1}$$

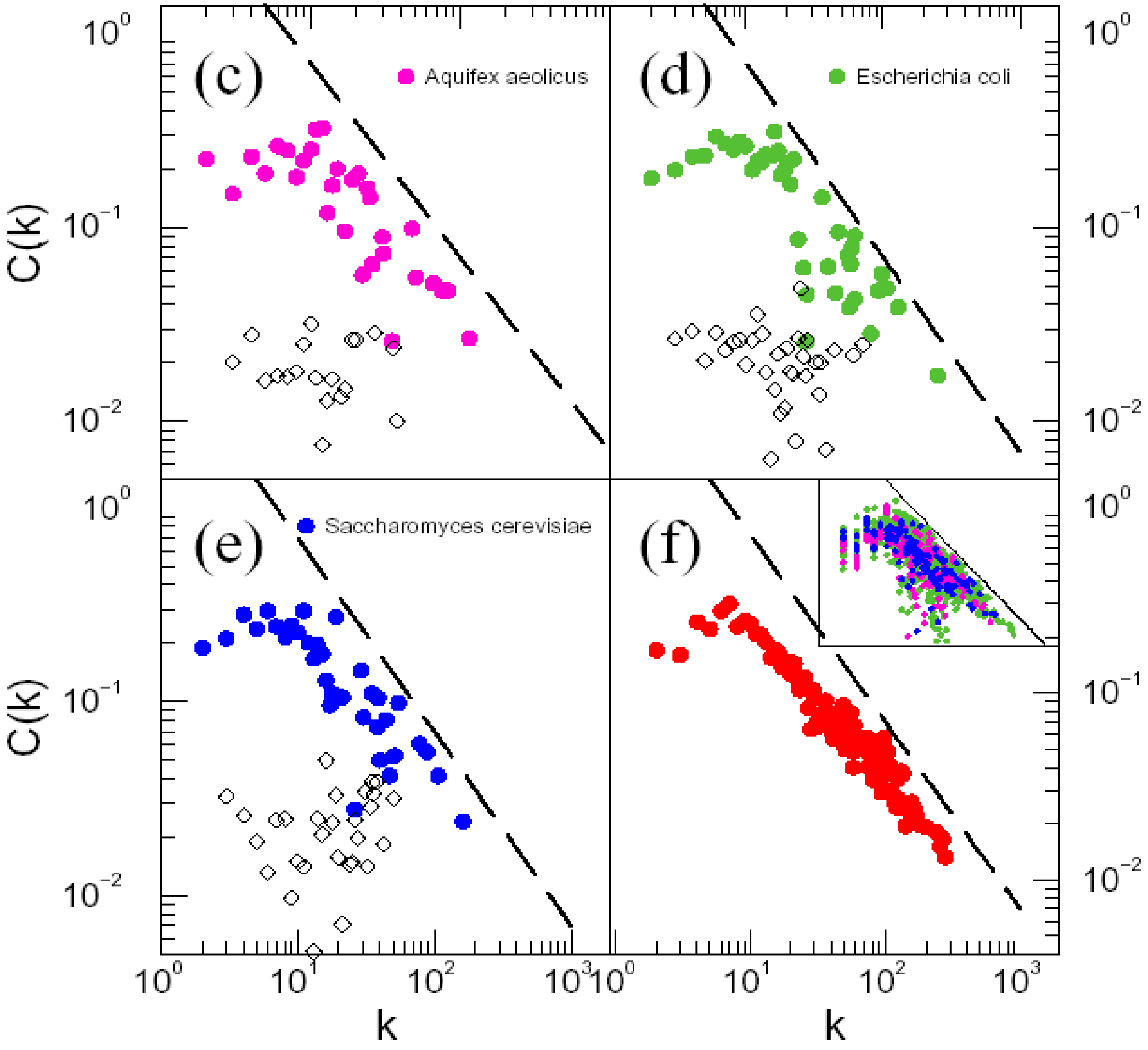




a small cluster of 5 densely linked nodes, four replicas connected to central node, producing a large 25-node module

Four replicas of this 25-node module are then generated and the peripheral nodes connected to the central node of the old module, producing a new module.

Scaling of the clustering coefficient C(k)

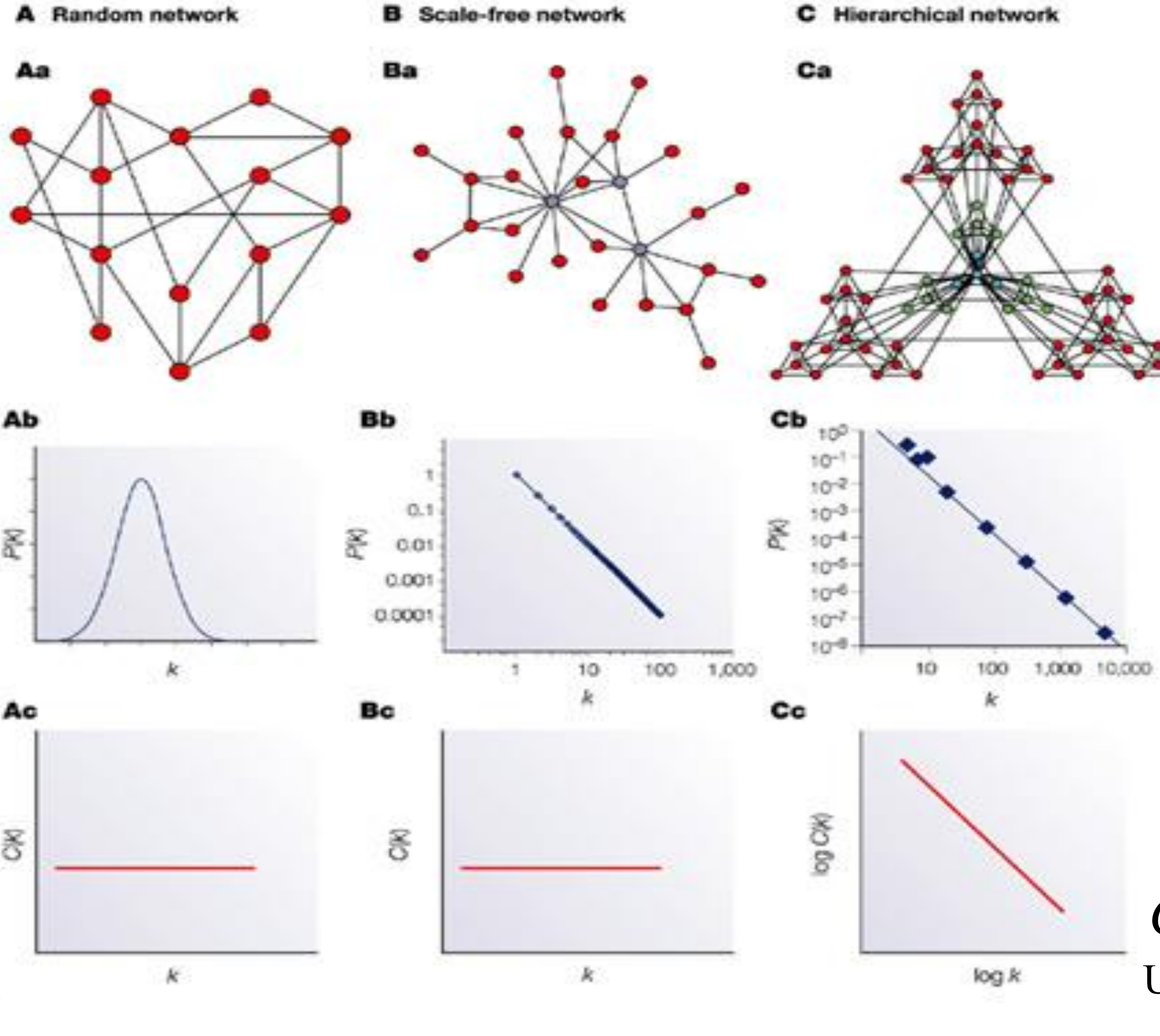


The examples of hierarchical networks.

Comparison of Global network properties

Degree distribution

Clustering coefficient distribution



Nature Reviews | Genetics

$C(k)=k^{-r}$

Usually r=1

Graph & Network Mining: Concepts

- 1. General Network Measures
- 2. Typical Architecture of Networks
- 3. Social network analysis

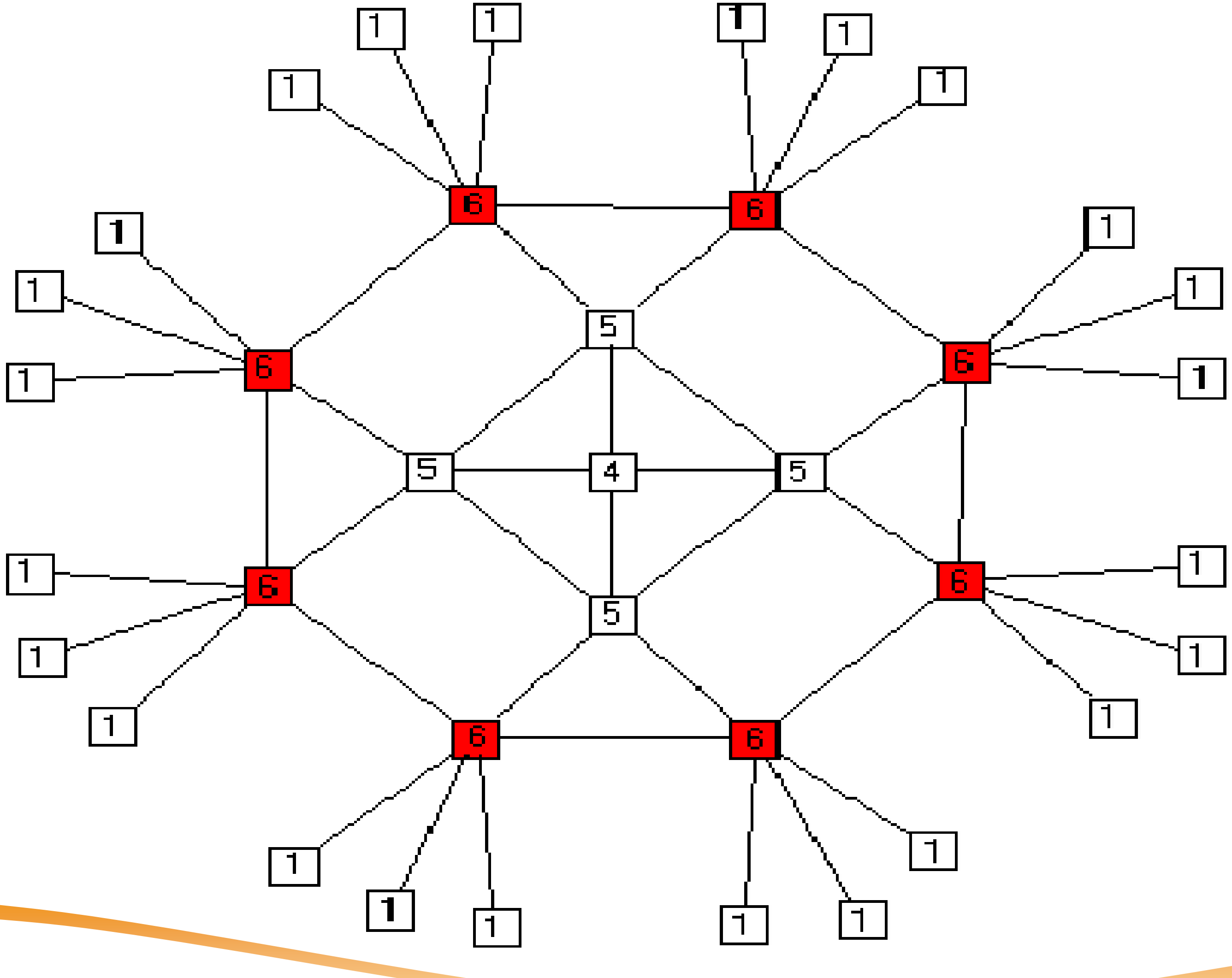
3. Social network analysis

- Social network is the study of social entities (people in an organization, called **actors**), and their interactions and relationships, which can be represented as a network or graph,
  - each vertex (or node) represents an actor and
  - each link represents a relationship.
- From the network, we can study the properties of its structure, and the role, position and prestige of each social actor.
- We can also find various kinds of sub-graphs, e.g., **communities** formed by groups of actors.
- SNA is relevant to advertising, national security, medicine, geography, politics, social psychology, etc.

Social network and the Web

- Social network analysis is useful for the Web because the Web is essentially a virtual society, and thus a virtual social network,
  - Each page: a social actor and
  - each hyperlink: a relationship.
- Many results from social network can be adapted and extended for use in the Web context.
- We study two types of social network analysis, **centrality** and **prestige**, which are closely related to hyperlink analysis and search on the Web.

Degree Centrality



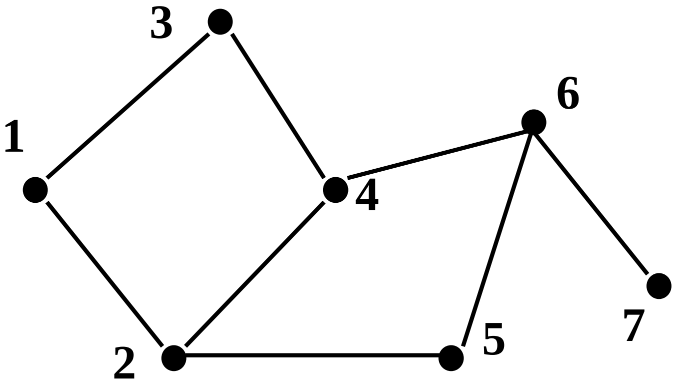
Degree Centrality –
We already know this

- Central actors are the most active actors that have most links or tiers with other actors.
- The degree of a node is the number of edges incident on it, which is the simplest indicator of how connected a node is within a graph
- In a directed graph, in-degree is the no. of incoming edges, and out-degree the no. of outgoing one
- For directed graphs, total degree = in-degree + out-degree (used in some software)

Closeness Centrality (Example)

- It is based on the closeness or distance. An actor x_i *is central if it can easily interact with all other actors.* That is, its distance to all other actors is short. Thus, the shortest distance is used to compute this measure. Let the shortest distance from actor *i* to actor *j be d*(*i, j*) *(measured as the number of links in a shortest path).*

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$



$C_c(1)=(7-1)/(1+1+2+2+3+4)=6/13$

1->2, 1->3, 1->4, 1->5, 1->6, 1->7

$C_c(4)=(7-1)/(2+1+1+2+1+2)=6/9$

4->1, 4->2, 4->3, 4->5, 4->6, 4->7

Closeness Centrality (Example)

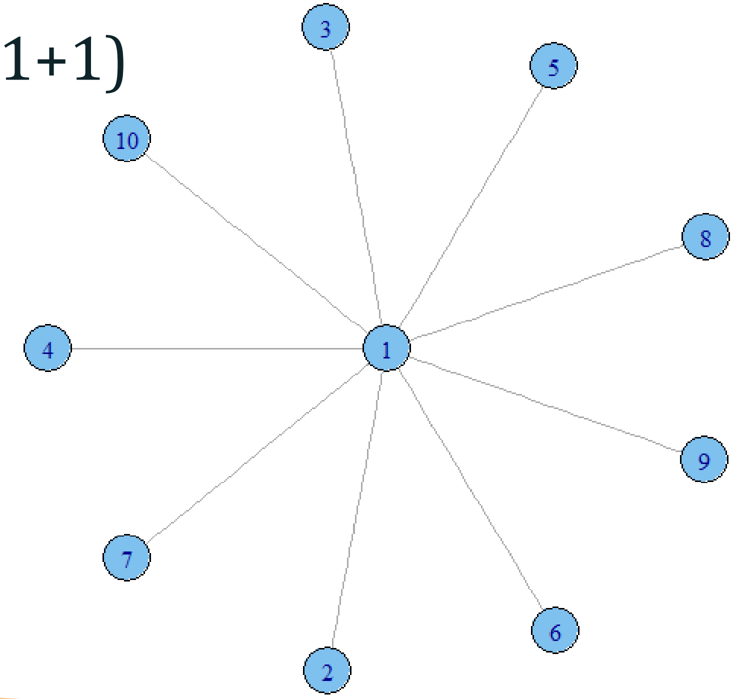
The value of this measure also ranges between 0 and 1 as *n -1 is the minimum* value of the denominator, which is the sum of the shortest distances from *i to all other actors.*

Example: Closeness Centrality gets the maximal value

$C_c(1) = (10\text{-}1)/(1+1+1+1+1+1+1+1+1)$

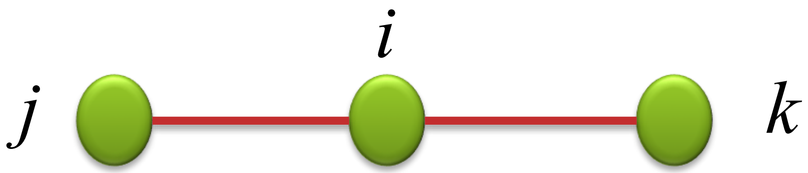
$=1$

$$C_C(i) = \frac{n-1}{\sum_{j=1}^{n} d(i,j)}$$



Betweenness Centrality

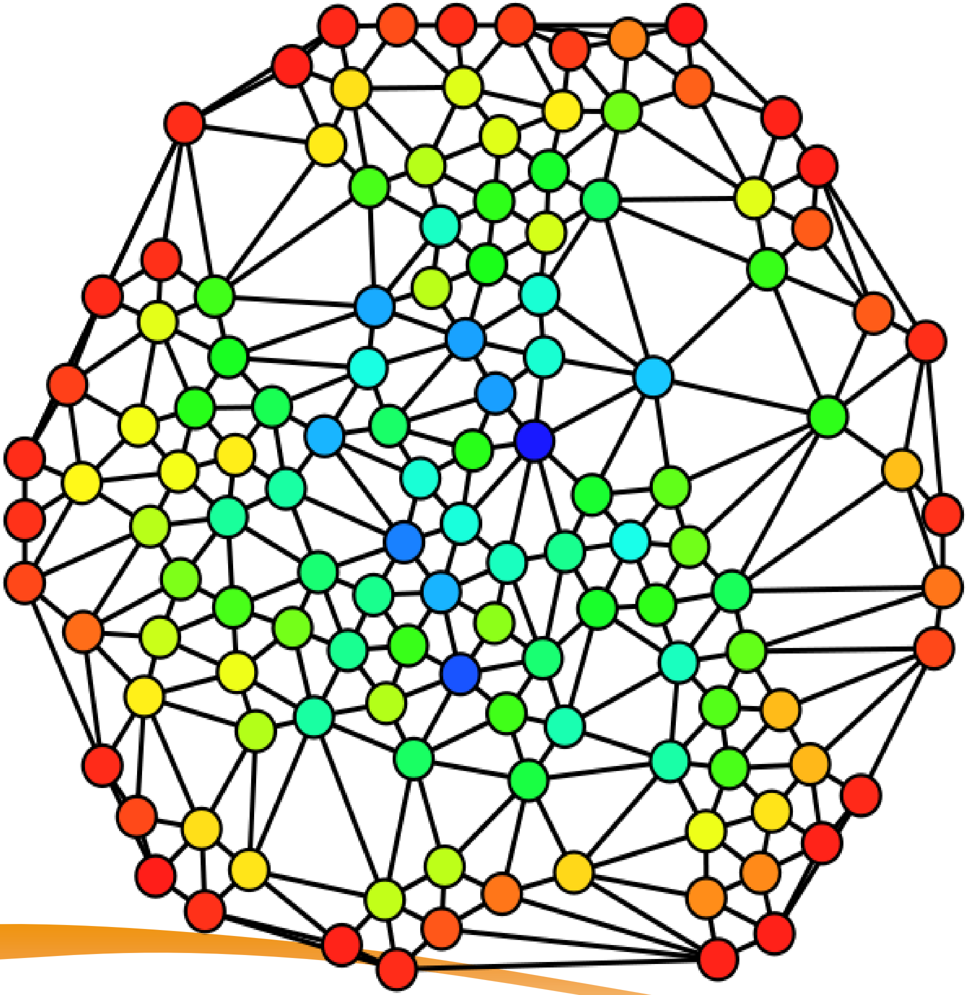
- If two non-adjacent actors *j* and *k* want to interact and actor *i* is on the path between *j* and *k*, then *i* may have some control over the interactions between *j* and *k*.

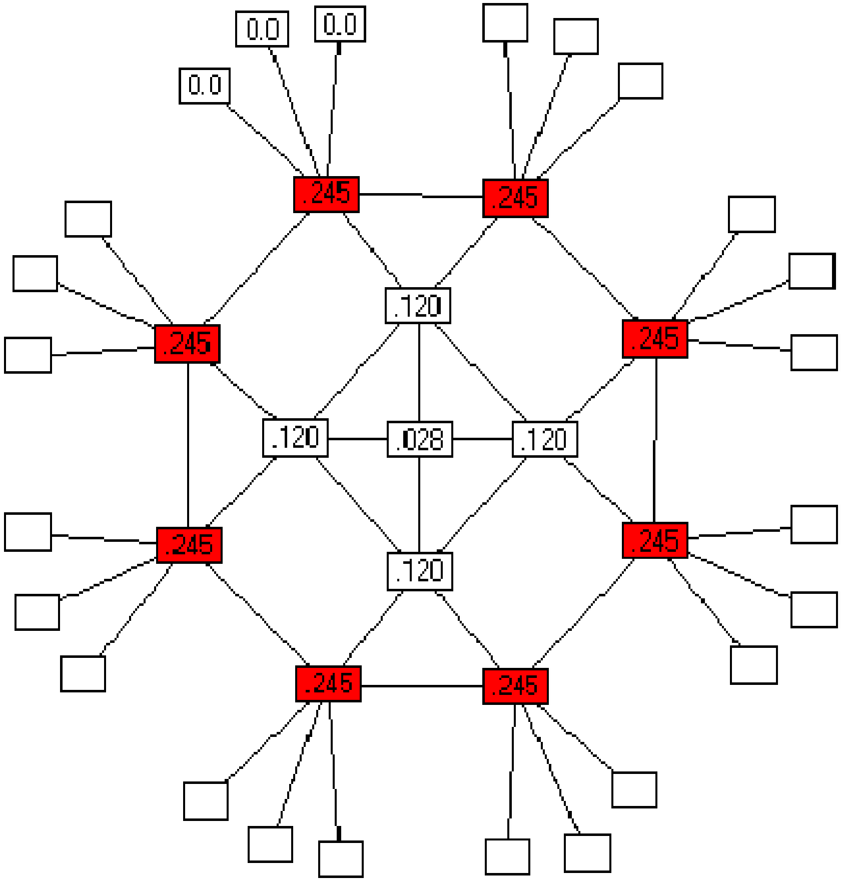


- Betweenness measures this control of *i* over other pairs of actors. Thus,
  - if *i* is on the paths of *many* such interactions, then *i* is an important actor.

Node Betweenness

from red=0 to blue=max

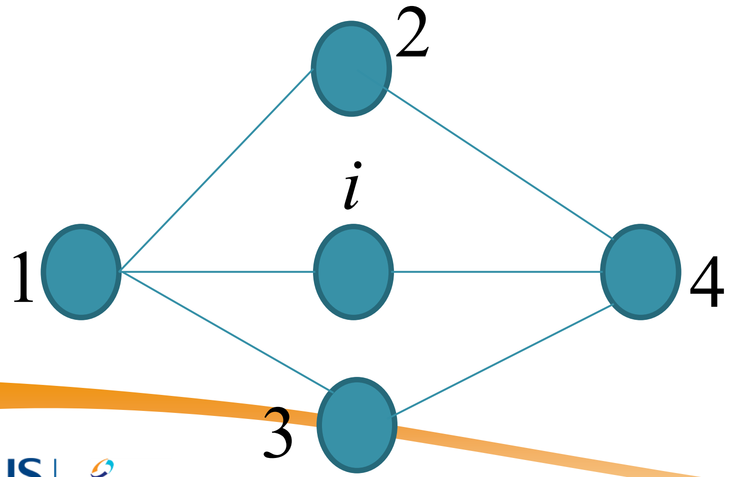




Example: Betweenness Centrality

- **Undirected graph**: Let p_{jk} be the number of shortest paths between actor j and actor k.
- The betweenness of an actor i is defined as the number of shortest paths that pass i $p_{jk}(i)$ normalized by the total number of shortest paths.

$$C_B(i) = \sum_{j<k} \frac{p_{jk}(i)}{p_{jk}}$$



$C_c(i)$=0/1+0/1+1/3+0/2+0/1+0/1=1/3

(1,2), (1,3), (1, 4), (2, 3), (2, 4), (3, 4)

Prestige

- Prestige is a more refined measure of prominence of an actor than centrality.
  - Distinguish: ties sent (out-links) and ties received (in-links).
- A prestigious actor has **extensive ties as a recipient**.
  - To compute the prestige: we use **only in-links.**
- We study two prestige measures. **Rank prestige** forms the basis of most Web page link analysis algorithms, including PageRank and HITS.

Prestige (cont ...)

- An actor is prestigious if it receives many in-links or nominations. Thus, the simplest measure of prestige of an actor *i* (*denoted by* $P_D(i)$) *is its in-degree.*

$$P_D(i) = \frac{d_I(i)}{n-1}$$

where $d_I(i)$ is the in-degree of *i* (the number of in-links of *i*). As in the degree centrality, dividing by *n* – 1 standardizes the prestige value to the range from 0 and 1. The maximum prestige value is 1 when every other actor links to or chooses actor *i*.

Proximity prestige

- The standard prestige of an actor *i* only considers the actors that are adjacent to *i*.
- The proximity prestige generalizes it by considering both the actors directly and *indirectly* linked to actor *i*.
- Let I_i be the set of actors that can reach actor *i*.
- The **proximity** is defined as distance of other actors to *i*.

Proximity prestige (cont ...)

- Let $d(j, i)$ denote the distance from actor j to actor i.
- We consider every actor j that can reach i, i.e., there is a directed path from j to i.

$$\frac{\sum_{j \in I_i} d(j, i)}{|I_i|}$$

- Let $|I_i|$ be the set of actors that can reach actor i. The proximity is defined as average distance of other actors to i.

Thank You

Contact: xlli@i2r.a-star.edu.sg if you have questions

References

1. http://barabasilab.com/

2. A.-L.Barabási & Z.N. Oltvai, **Nature**, 2004