

Graph Theory and Social Networks - part I

EE599: Social Network Systems

Keith M. Chugg

Fall 2014



USC University of
Southern California

Overview

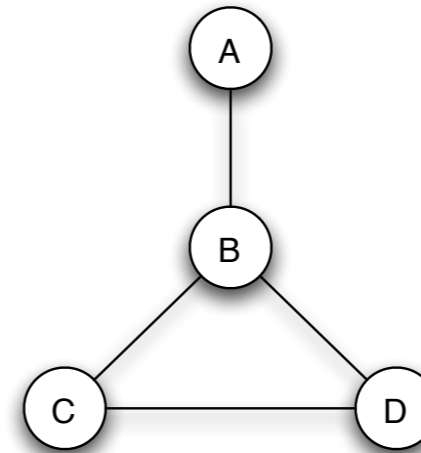
- Summary
- Graph definitions and properties
- Relationship and interpretation in social networks
- Examples

References

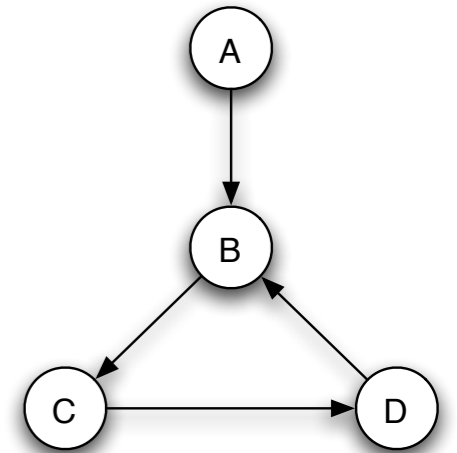
- Easley & Kleinberg, Ch 2
 - Focus on relationship to social nets with little math
- Barabasi, Ch 2
 - General networks with some math
- Jackson, Ch 2
 - Social network focus with more formal math

Graph Definition

- $G = (V, E)$
 - $V = \text{set of vertices}$
 - $E = \text{set of edges}$



(a) A graph on 4 nodes.



(b) A directed graph on 4 nodes.

Figure 2.1: Two graphs: (a) an undirected graph, and (b) a directed graph.

Easley & Kleinberg

- Modeling of networks
 - Vertex is a person (or entity)
 - Edge represents a relationship

Graph Example

Network Science	Graph Theory
network	graph
node	vertex
link	edge

Baraba'si, Ch 2

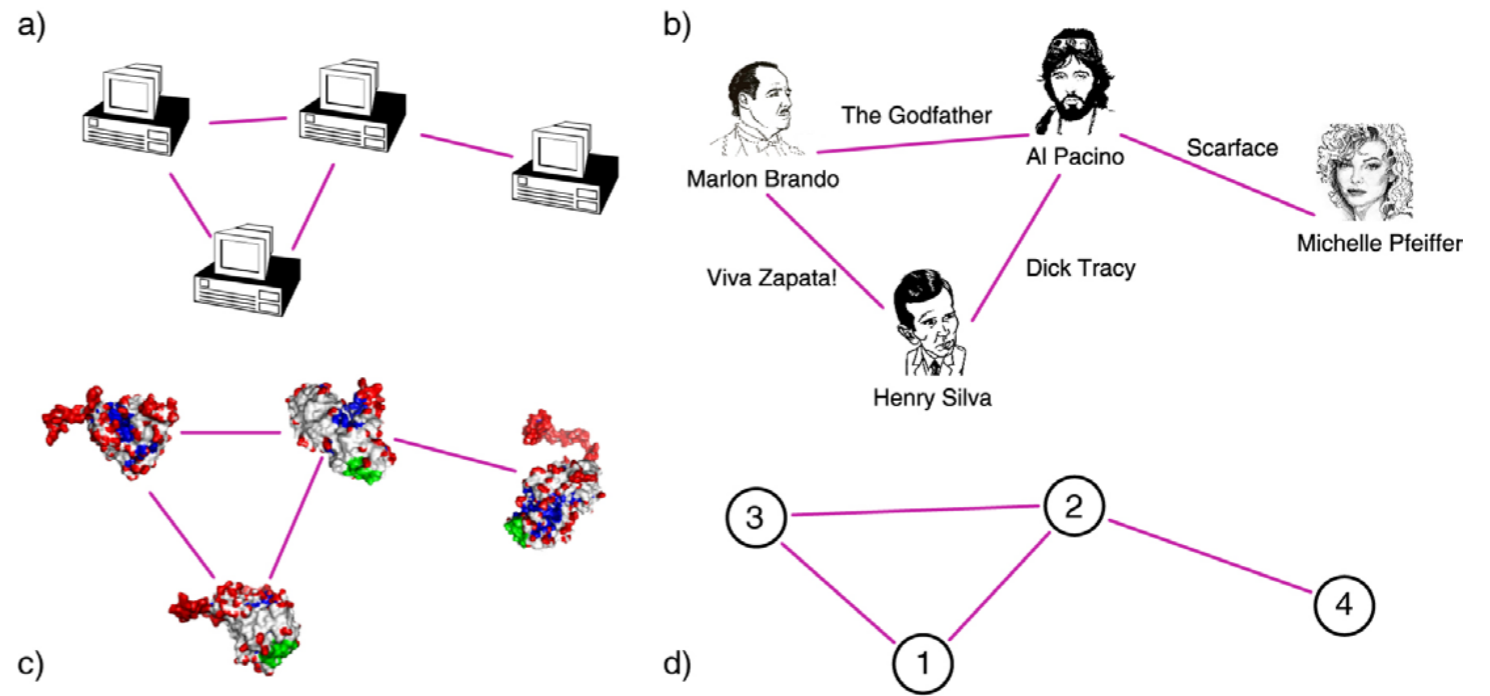


Image 2.3

Real systems of quite different nature can have the same network representation.

In the figure we show a small subset of (a) the *Internet*, where routers (specialized computers) are connected to each other; (b) the *Hollywood actor network*, where two actors are connected if they played in the same movie; (c) a *protein-protein interaction network*, where two proteins are connected if there is experimental evidence that they can bind to each other in the cell. While the nature of the nodes and the links differs widely, each network has the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links, shown in (d).

Baraba'si, Ch 2

Basic Graph Defs/Props

- Paths, walks, cycles
- Connectedness and components
 - Giant component
- Node degree, Node degree statistics
 - Sparseness
- Adjacency matrix
- Distance and diameter
 - Small World Phenomena

Complete Graph

- All nodes connect to all other nodes
- Maximum number of edges

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

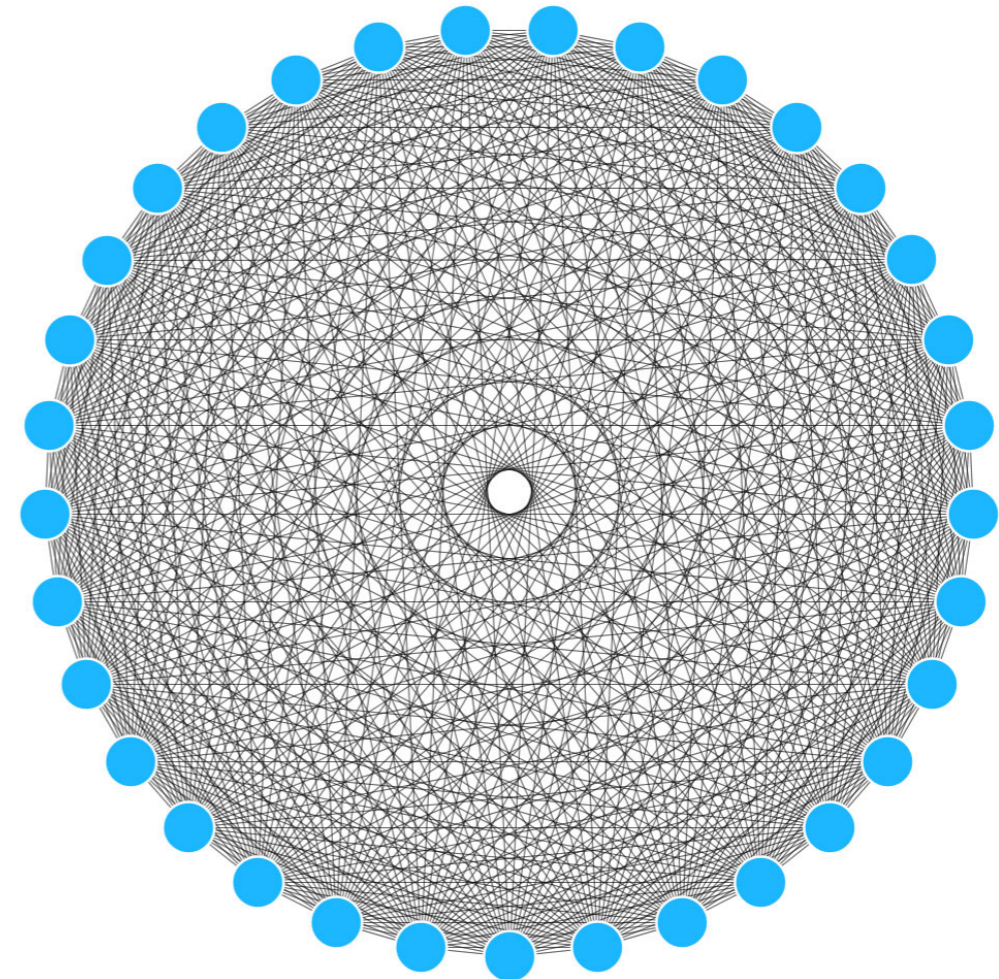


Image 2.5
Complete graph.

The figure shows a complete graph with $N = 16$ nodes and $L_{\max} = 120$ links, as predicted by Eq. (11). The adjacency matrix of a complete graph is $A_{ij} = 1$ for all $i, j = 1, \dots, N$ and $A_{ii} = 0$. The average degree of a complete graph is $\langle k \rangle = N - 1$.

Baraba'si, Ch 2

Example Networks

NETWORK NAME	NODES	LINKS	DIRECTED/ UNDIRECTED	N	L	⟨K⟩
Internet	routers	Internet Connections	Undirected	192,244	609,066	2.67
WWW	webpages	links	Directed	325,729	1,497,134	4.60
Power Grid	power plants, transformers	cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	subscribers	calls	Directed	36,595	91,826	2.51
Email	email addresses	emails	Directed	57,194	103,731	1.81
Science Collaboration	scientists	co-authorships	Undirected	23,133	186,936	16.16
Actor Network	actors	co-acting	Undirected	212,250	3,054,278	28.78
Citation Network	papers	citations	Directed	449,673	4,707,958	10.47
E. coli Metabolism	metabolites	chemical reactions	Directed	1,039	5,802	5.84
Yeast Protein Interactions	proteins	binding interactions	Undirected	2,018	2,930	2.90

Table 2.1

Network maps and their basic properties.

Baraba'si, Ch 2

- Real networks are sparse - far from fully-connected

Paths & Connectivity

- **Path**: sequence of links from node i to node $j \neq i$
- Repeat vertices
 - Walk: can repeat vertices
 - SocNet: often called path
 - Path: no repeat vertices (graph theory)
 - SocNet: “simple path”, “self-avoiding path”
- Two nodes are **connected** iff there is a path between them

Paths & Connectivity

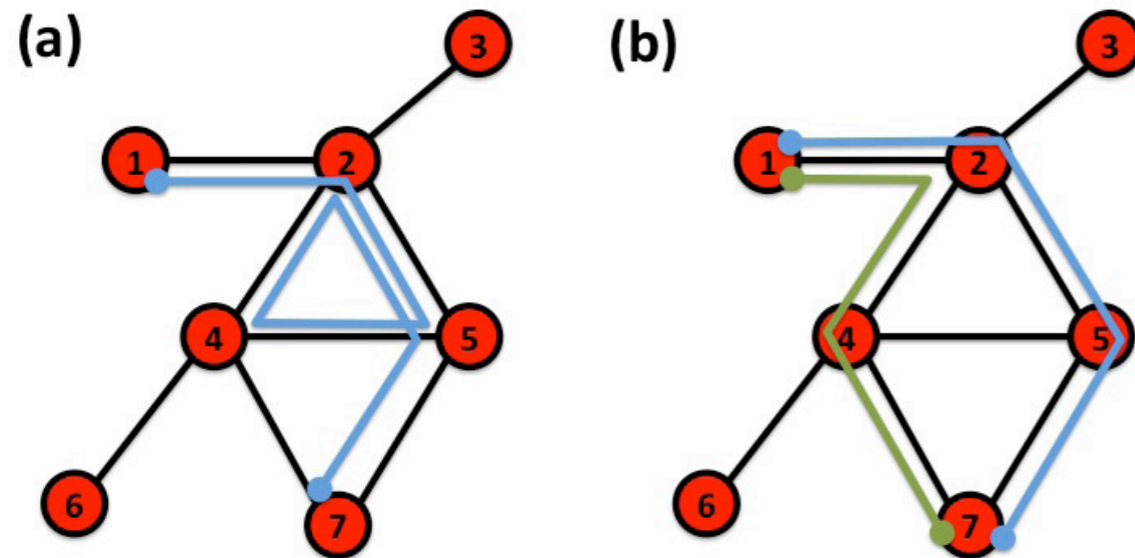


Image 2.11

The adjacency matrix is typically sparse.

(a) A path between nodes i_0 and i_n is an ordered list of n links $P_d = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \dots, (i_{n-1}, i_n)\}$. The length of this path is d . The path shown in (a) follows the route $1 \rightarrow 2 \rightarrow 5 \rightarrow 4 \rightarrow 2 \rightarrow 5 \rightarrow 7$, hence its length is $n = 6$.

(b) The shortest paths between nodes 1 and 7, representing the distance d_{17} , is the path with the fewest number of links that connect nodes 1 and 7. There can be multiple paths of the same length, as illustrated by the two paths shown in different colors. The network diameter is the largest distance in the network, being $d_{max} = 3$ here.

Baraba'si, Ch 2

- **Cycle**: a path where start and finish nodes are the same
- no repeat edges
- Path or Cycle length is number of edges

Connected Components

- **Connected graph:** every node is connected to every other node
- **Subgraph:** subset of vertices and edges from a given graph
- **(Connected) Components:** maximal (size) connected subgraphs

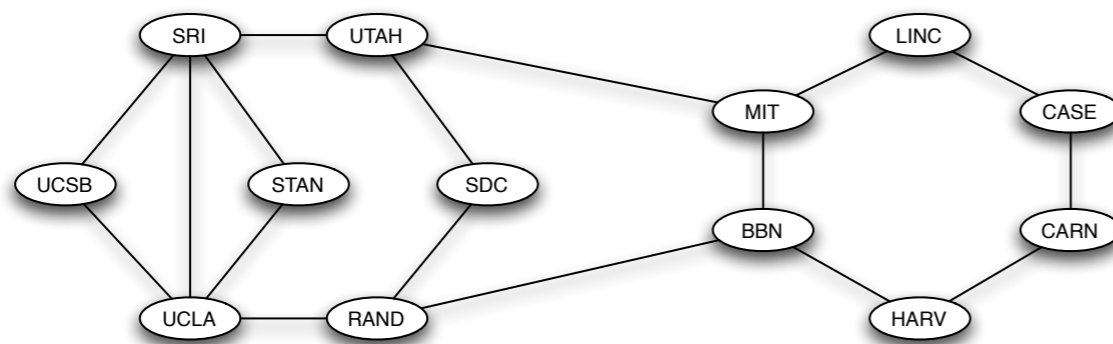


Figure 2.3: An alternate drawing of the 13-node Internet graph from December 1970.

Easley & Kleinberg

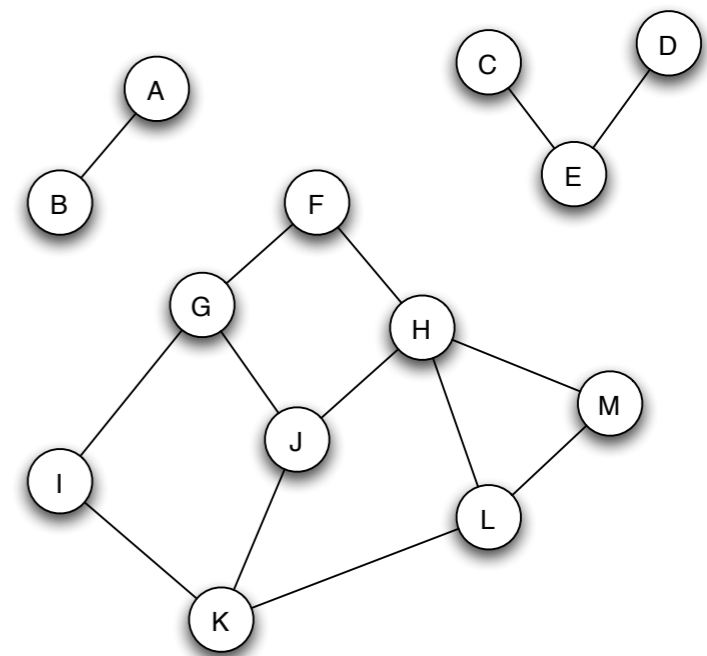


Figure 2.5: A graph with three connected components.

Easley & Kleinberg

Giant Component

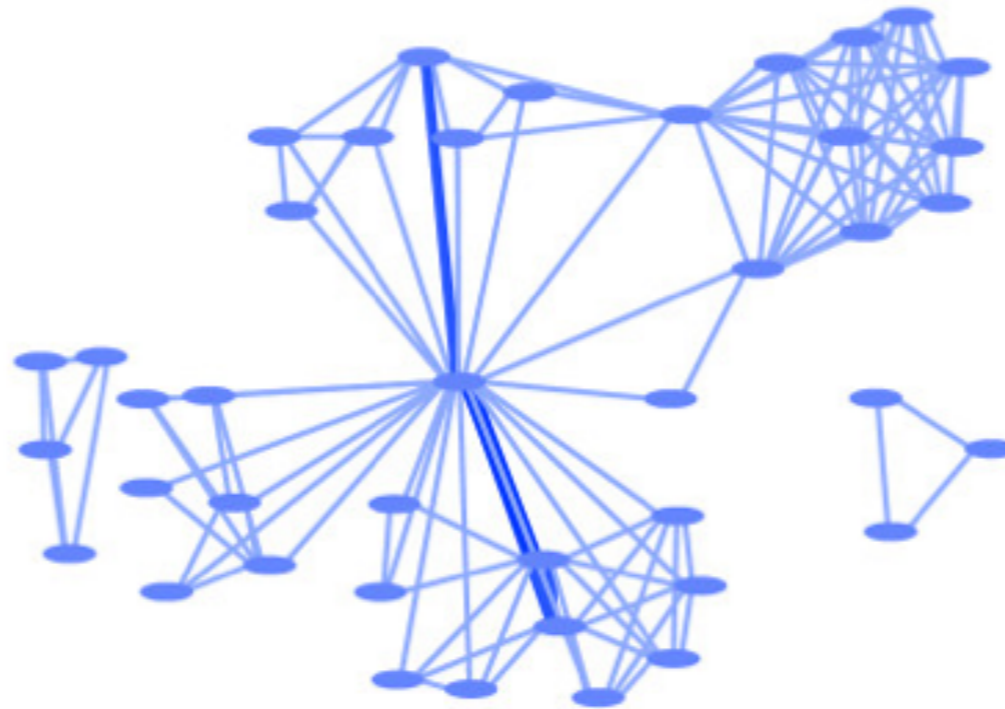
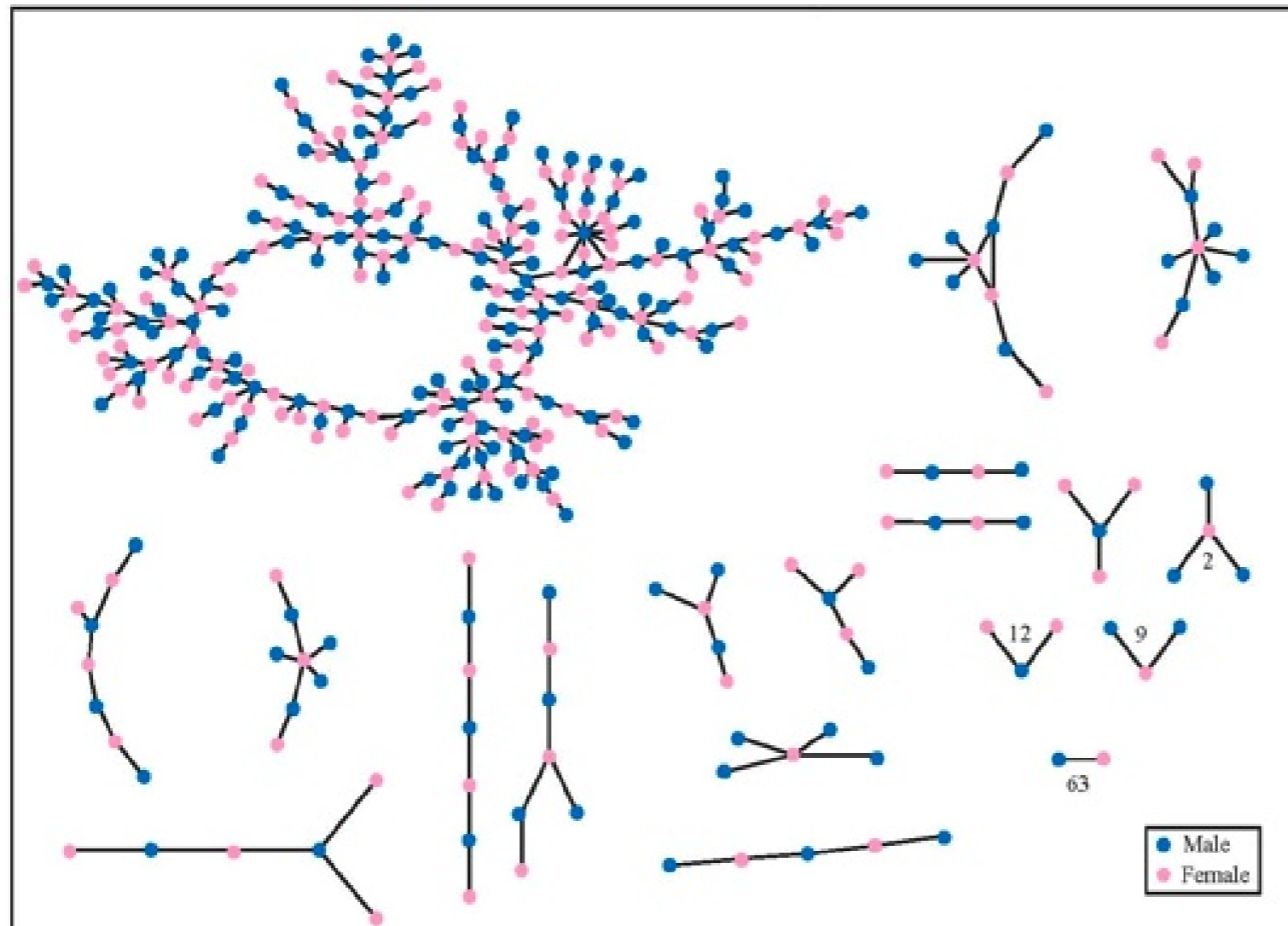


Figure 2.6: The collaboration graph of the biological research center *Structural Genomics of Pathogenic Protozoa (SGPP)* [134], which consists of three distinct connected components. This graph was part of a comparative study of the collaboration patterns graphs of nine research centers supported by NIH's Protein Structure Initiative; SGPP was an intermediate case between centers whose collaboration graph was connected and those for which it was fragmented into many small components.

Easley & Kleinberg

Giant Component



Giant component has over 100 people.

Next largest has 10

Figure 2.7: A network in which the nodes are students in a large American high school, and an edge joins two who had a romantic relationship at some point during the 18-month period in which the study was conducted [49].

Easley & Kleinberg

Giant Component

- Purposely vague term: means that there is a single, connected component with a large fraction of the vertices
 - Why does it exist?
 - Why is there only one?
 - Examples of social networks with 2 giant components?

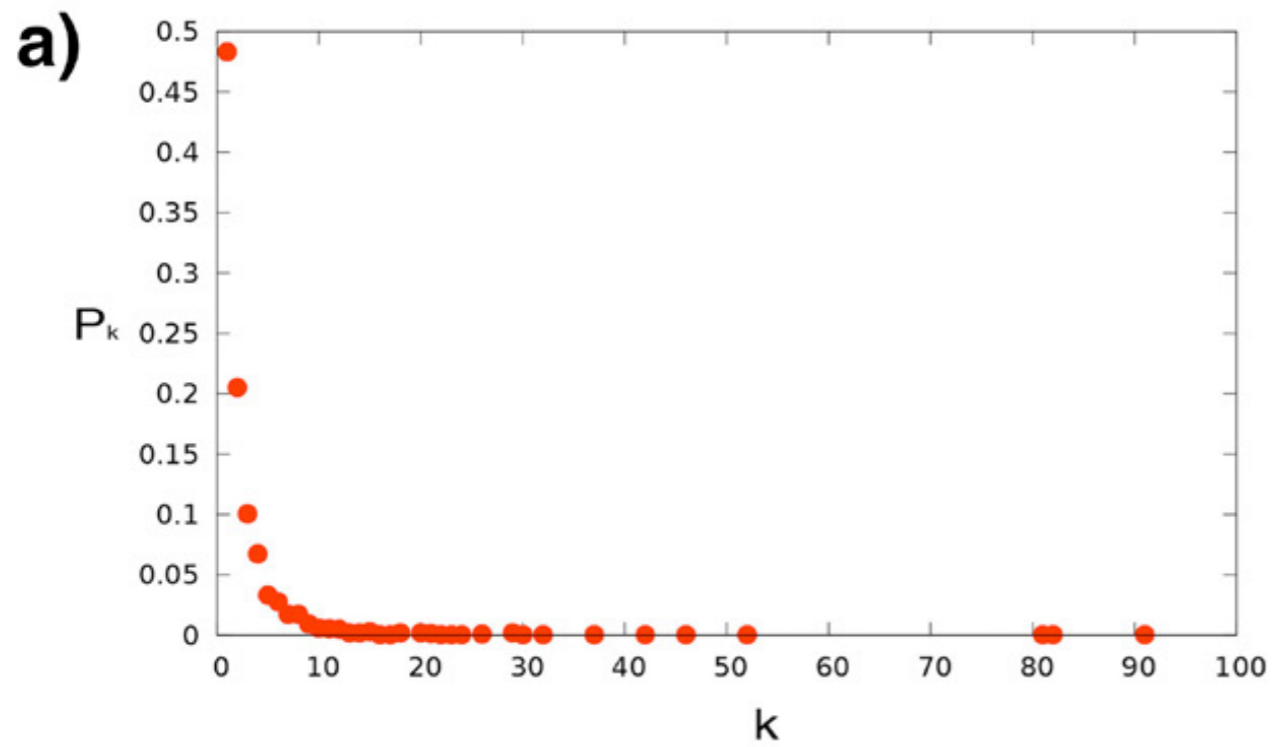
Node Degree

- **Node Degree:** number of edges connected to that node
- Statistical measure of node degree
 - “Complete” - degree distribution
 - “Incomplete” - mean, variance
 - These can be empirical or based on a random model

Node Degree

histogram (empirical)

probability mass function
(analytical)



Baraba'si, Ch 3

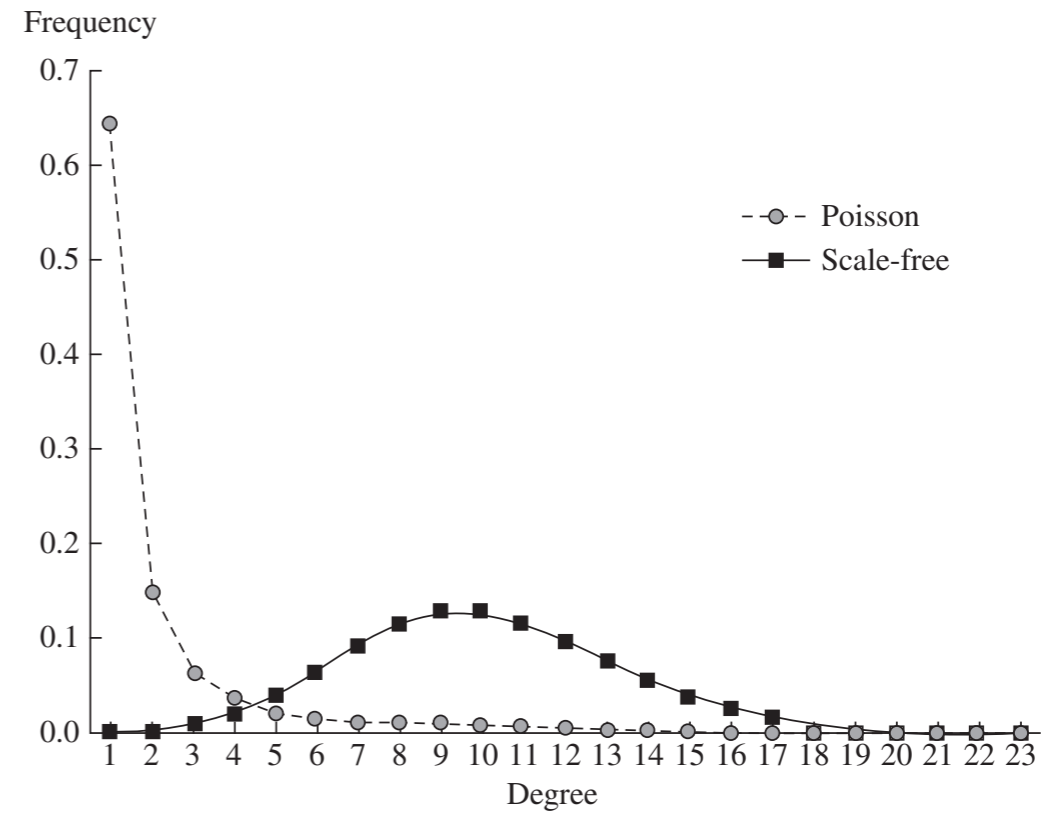


FIGURE 2.8 Comparing a scale-free distribution to a Poisson distribution.

Jackson

Example Networks

NETWORK NAME	NODES	LINKS	DIRECTED/ UNDIRECTED	N	L	⟨K⟩
Internet	routers	Internet Connections	Undirected	192,244	609,066	2.67
WWW	webpages	links	Directed	325,729	1,497,134	4.60
Power Grid	power plants, transformers	cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	subscribers	calls	Directed	36,595	91,826	2.51
Email	email addresses	emails	Directed	57,194	103,731	1.81
Science Collaboration	scientists	co-authorships	Undirected	23,133	186,936	16.16
Actor Network	actors	co-acting	Undirected	212,250	3,054,278	28.78
Citation Network	papers	citations	Directed	449,673	4,707,958	10.47
E. coli Metabolism	metabolites	chemical reactions	Directed	1,039	5,802	5.84
Yeast Protein Interactions	proteins	binding interactions	Undirected	2,018	2,930	2.90

Table 2.1

Network maps and their basic properties.

Baraba'si, Ch 2

- Real networks are sparse - far from fully-connected

Node Degree

degree distribution often plotted on log-log plot for real, sparse networks

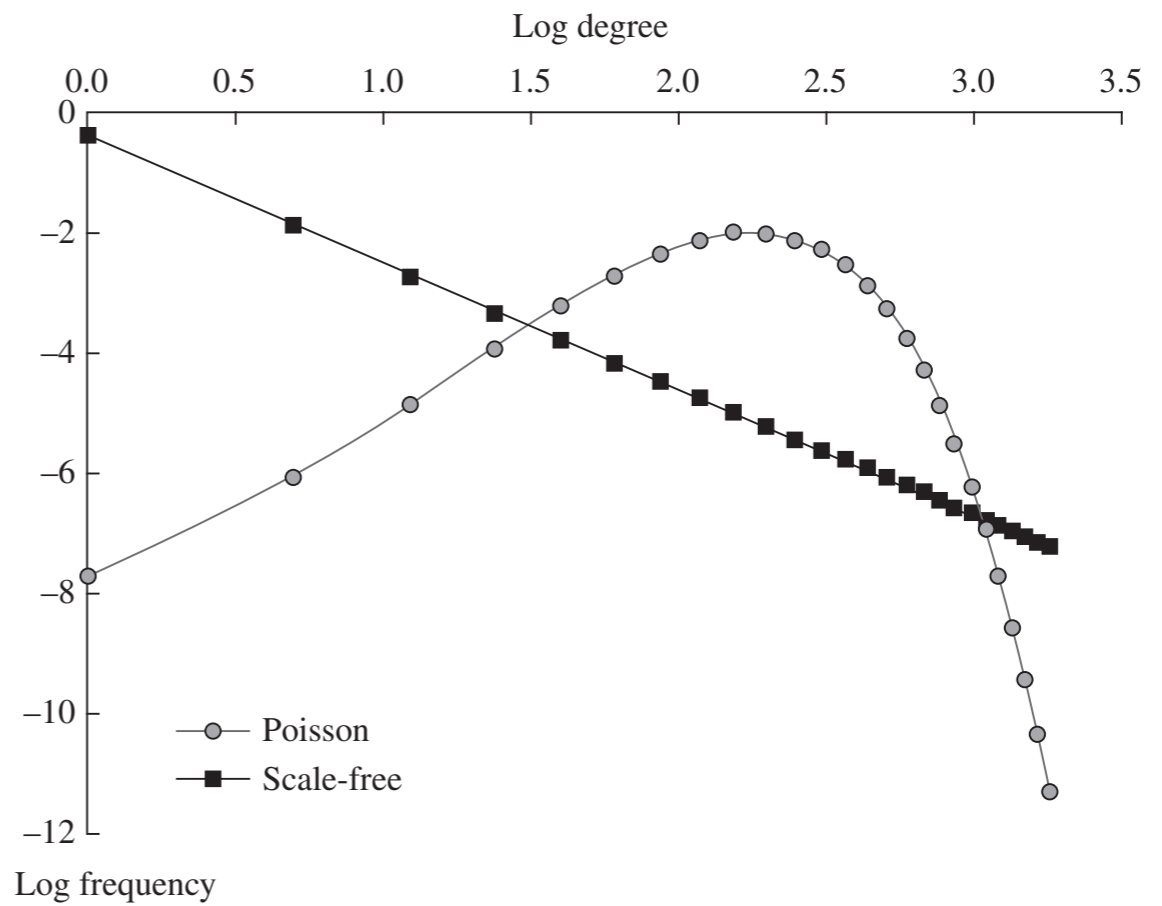


FIGURE 2.9 Comparing a scale-free distribution to a Poisson distribution: log-log plot.

Jackson

Node Degree

degree distribution often plotted on log-log plot for real, sparse networks

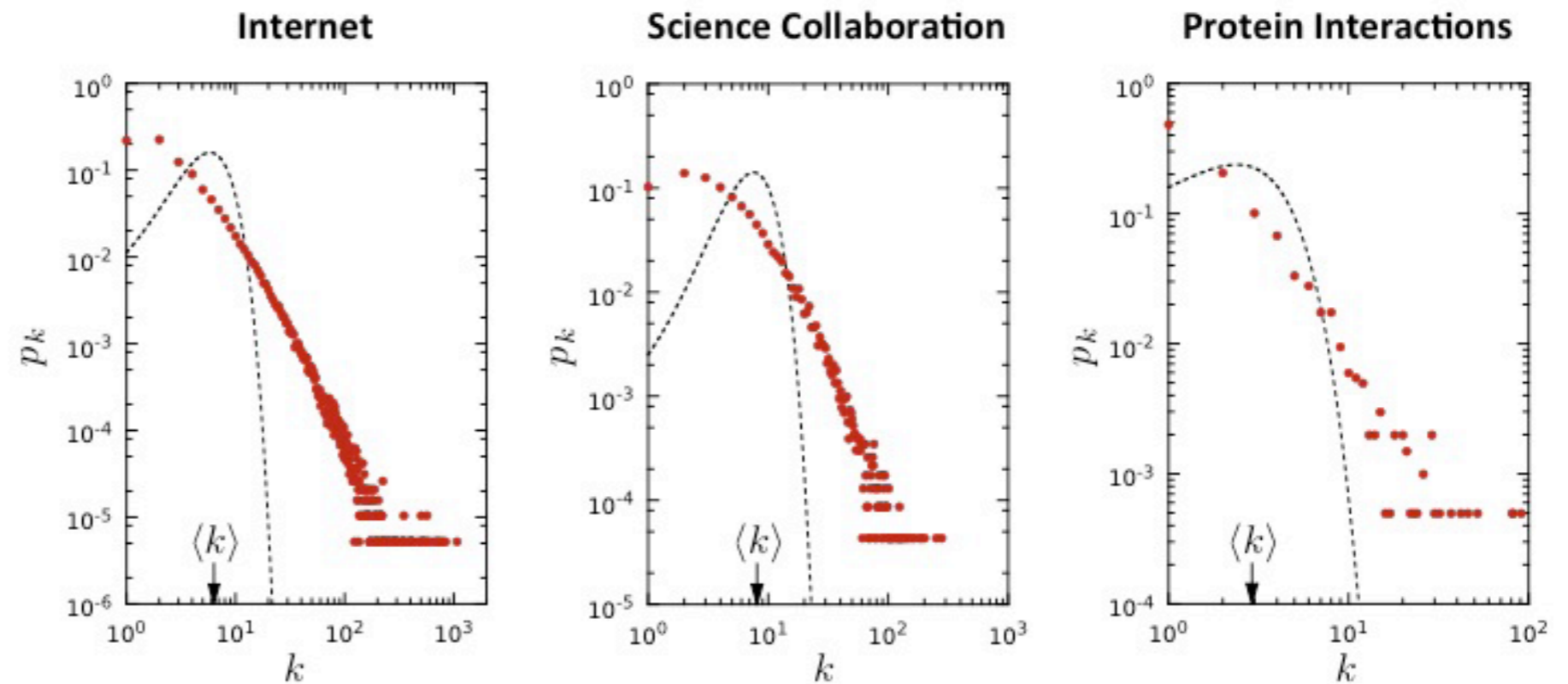


Image 3.5
Degree distribution of real networks.

The degree distribution of the Internet, science collaboration network, and the protein interaction network of yeast (Table 2.1). The dashed line corresponds to the Poisson prediction, obtained by measuring $\langle k \rangle$ for the real network and then plotting Eq. (8). The significant deviation between the data and the Poisson fit indicates that the random network model underestimates the size and the frequency of highly connected nodes, or hubs.

Baraba'si, Ch 3

Node Degree

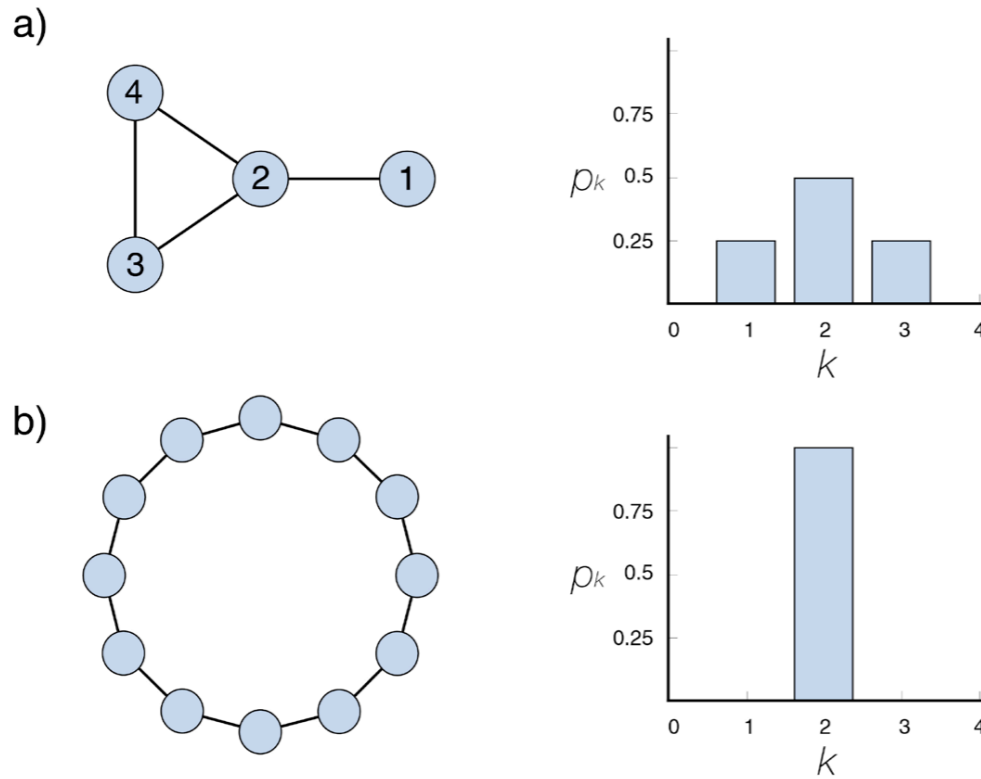


Image 2.4a
Degree distribution.

The degree distribution is defined as the $p_k = N_k/N$ ratio, where N_k denotes the number of k -degree nodes in a network. For the network in (a) we have $N = 4$ and $p_1 = 1/4$ (one of the four nodes has degree $k_1 = 1$), $p_2 = 1/2$ (two nodes have $k_3 = k_4 = 2$), and $p_3 = 1/4$ (as $k_2 = 3$). As we lack nodes with degree $k > 3$, $p_k = 0$ for any $k > 3$. Panel (b) shows the degree distribution of a one dimensional lattice. As each node has the same degree $k = 2$, the degree distribution is a Kronecker's delta function $p_k = \delta(k - 2)$.

Baraba'si

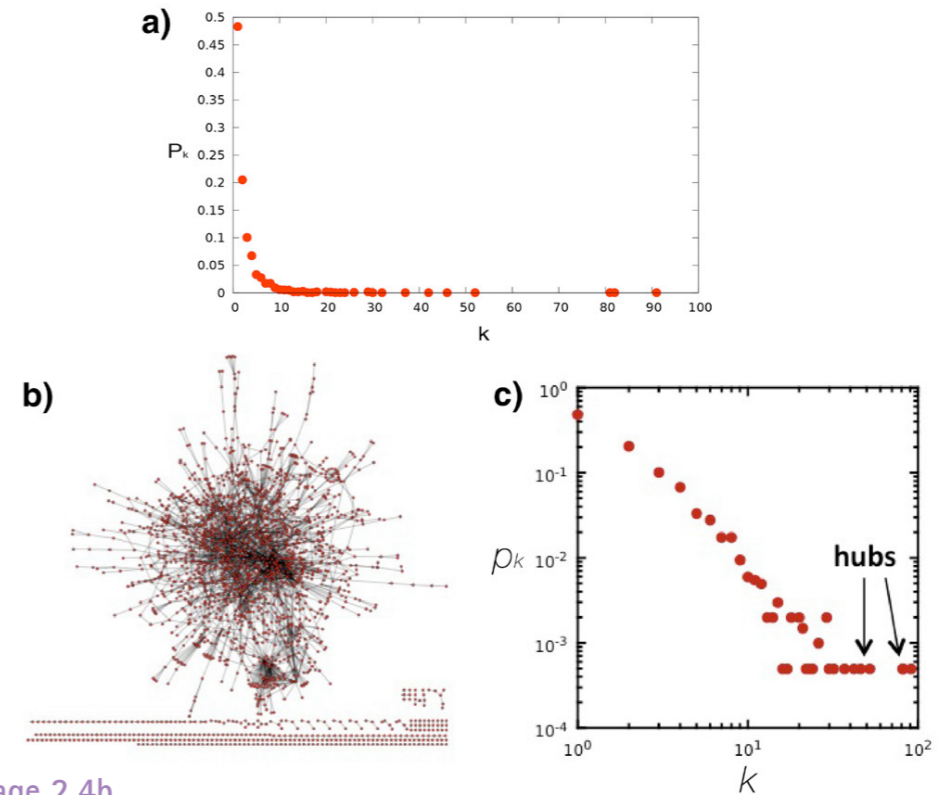


Image 2.4b

In many real networks, the node degree can vary considerably. For example, as the degree distribution (a) indicates, the degrees of the proteins in the protein interaction network shown in (b) vary between $k=0$ (isolated nodes) and $k=92$, which is the degree of the largest node, called a hub. There are also wide differences in the number of nodes with different degrees: as (a) shows, almost half of the nodes have degree one (i.e. $p_1=0.48$), while there is only one copy of the biggest node, hence $p_{92} = 1/N=0.0005$. (c) The degree distribution is often shown on a so-called log-log plot, in which we either plot $\log p_k$ in function of $\log k$, or, as we did in (c), we use logarithmic axes.

Distance & Diameter

- **Distance between two nodes:** length of the shortest path between i, j
- **Diameter:** largest distance between two distinct nodes
- **Girth:** length of the shortest cycle
- **Circumference:** length longest cycle

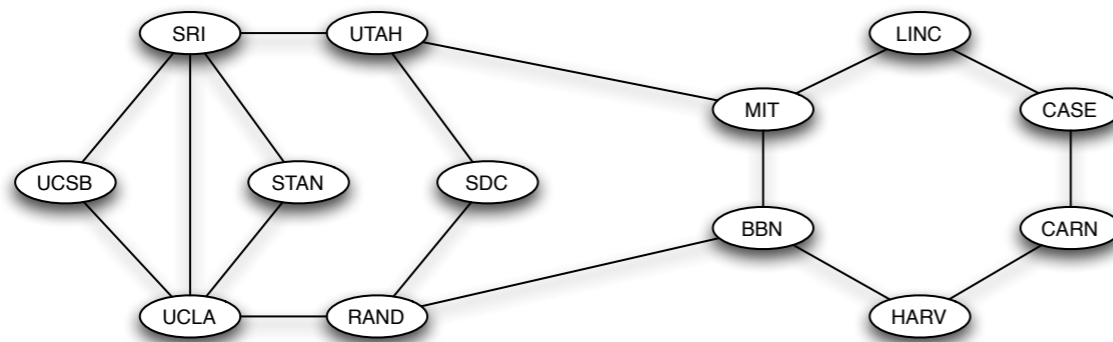


Figure 2.3: An alternate drawing of the 13-node Internet graph from December 1970.

Easley & Kleinberg

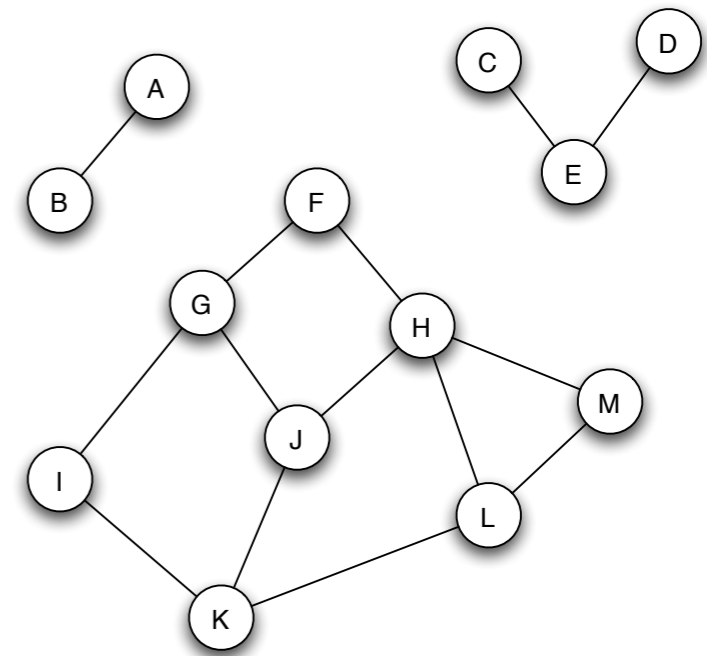


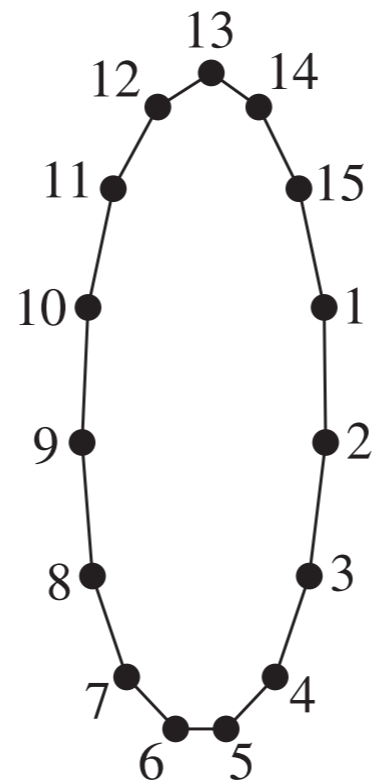
Figure 2.5: A graph with three connected components.

Easley & Kleinberg

Distance & Diameter

ave degree ~ 2

$$D \sim N/2$$



$$D \sim \log_2(N)$$

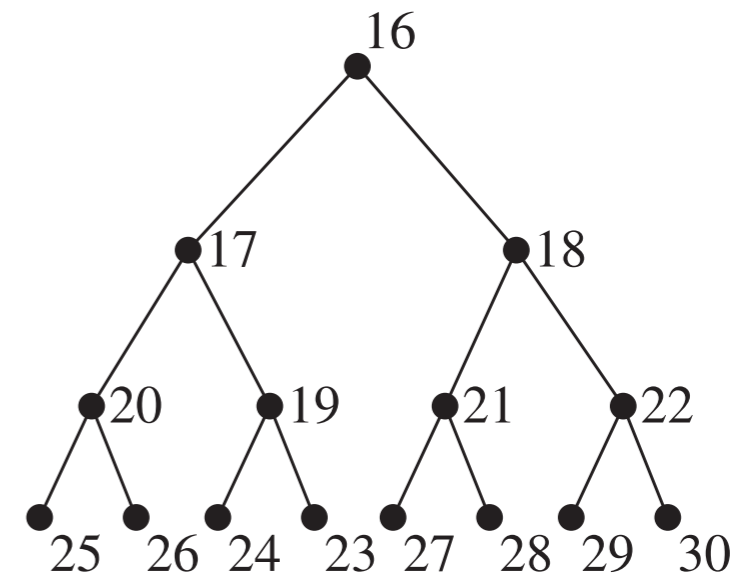


FIGURE 2.10 Circle and tree.

Jackson

Distance

As with degree, we can compute statistical measures of distance (empirical or analytical)

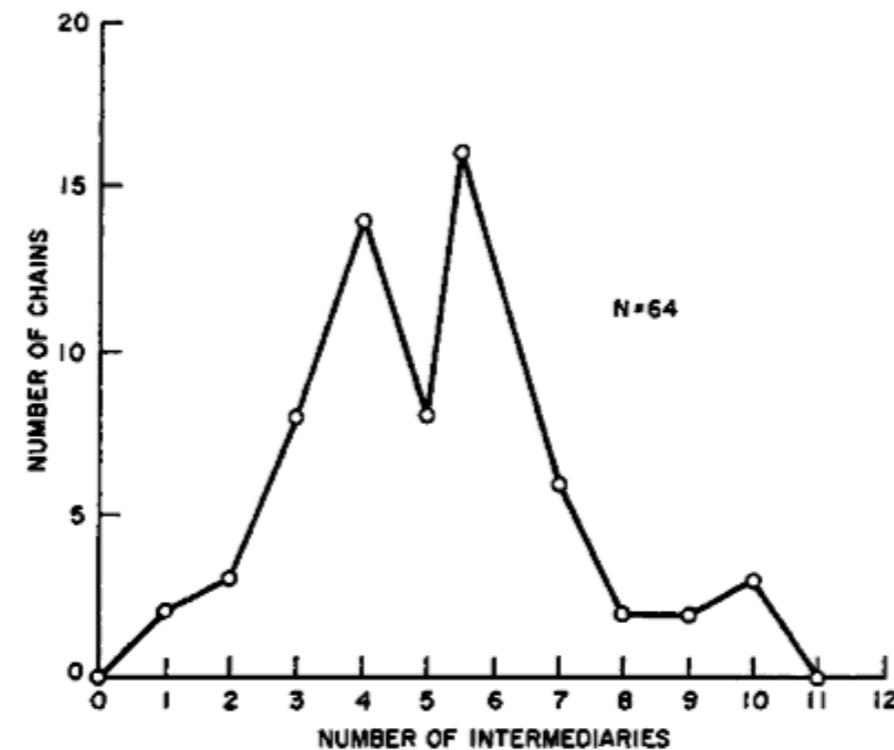


Figure 2.10: A histogram from Travers and Milgram's paper on their small-world experiment [391]. For each possible length (labeled "number of intermediaries" on the x -axis), the plot shows the number of successfully completed chains of that length. In total, 64 chains reached the target person, with a median length of six.

Easley & Kleinberg

Distance Distribution

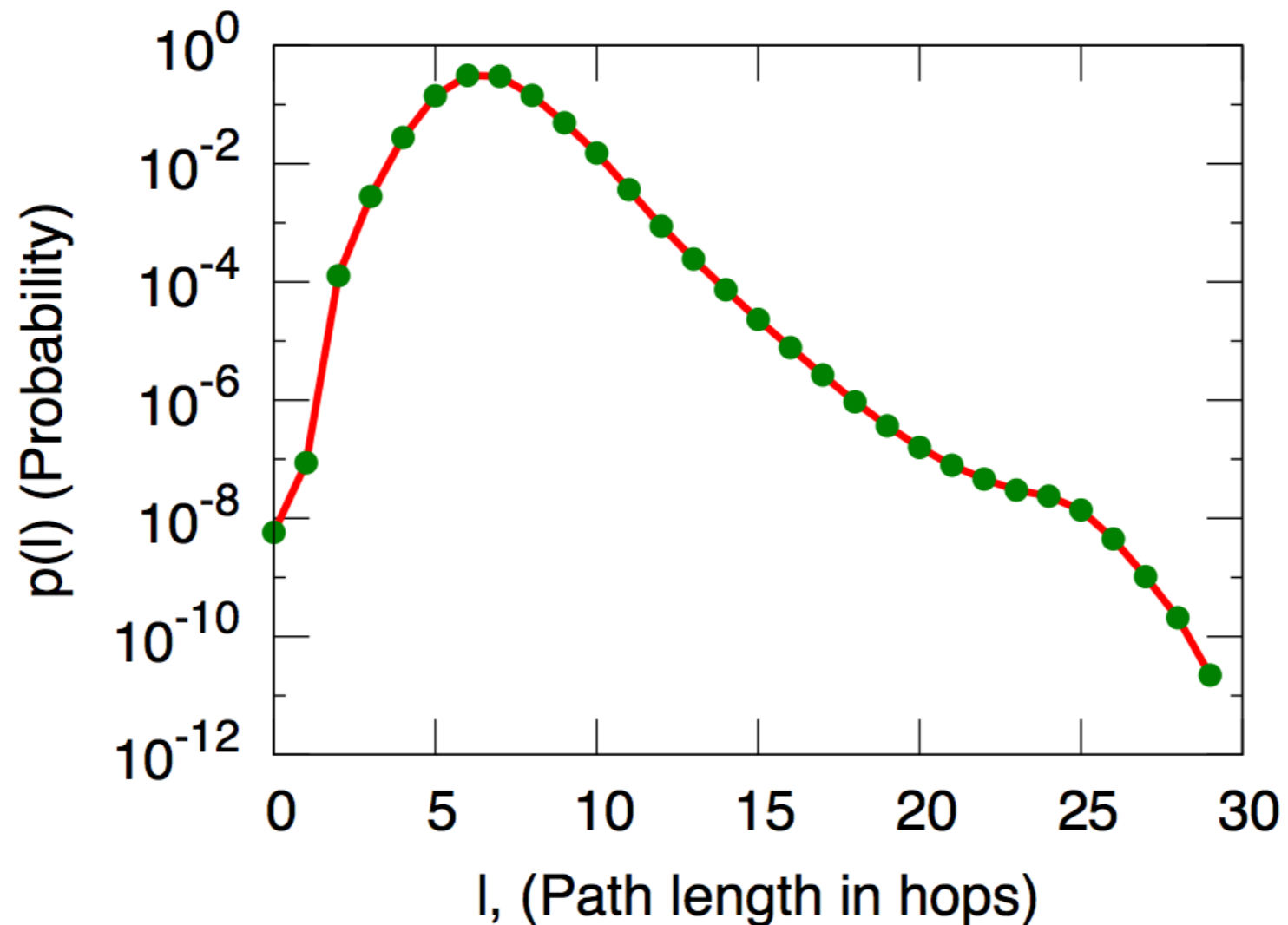


Figure 2.11: The distribution of distances in the graph of all active Microsoft Instant Messenger user accounts, with an edge joining two users if they communicated at least once during a month-long observation period [273].

Easley & Kleinberg

Small World Phenomena

AMERICAN MATHEMATICAL SOCIETY
MathSciNet 75
Mathematical Reviews 1940-2014
ISSN 2167-5163

Home | Preferences | **Free Tools** | About | Librarians | Terms of Use

MOBILE ACCESS

Search MSC | **Collaboration Distance** | Current Journals | Current Publications

MR Erdos Number = 4

Keith M. Chugg	coauthored with	P. Vijay Kumar	MR2028018 (2004j:94039)
P. Vijay Kumar	coauthored with	Krishnasamy Thiru Arasu	MR1872851 (2002m:05042)
Krishnasamy Thiru Arasu	coauthored with	Anthony B. Evans	MR1677785 (99k:05021)
Anthony B. Evans	coauthored with	Paul Erdős ¹	MR1016280 (90g:05049)


[Change First Author](#) | [Change Second Author](#) | [New Search](#)

Free Tool | [Help](#) | [Support Mail](#)

AMS
AMERICAN MATHEMATICAL SOCIETY

Mirror Sites [Providence, RI USA](#)

© Copyright 2014, American Mathematical Society
Privacy Statement



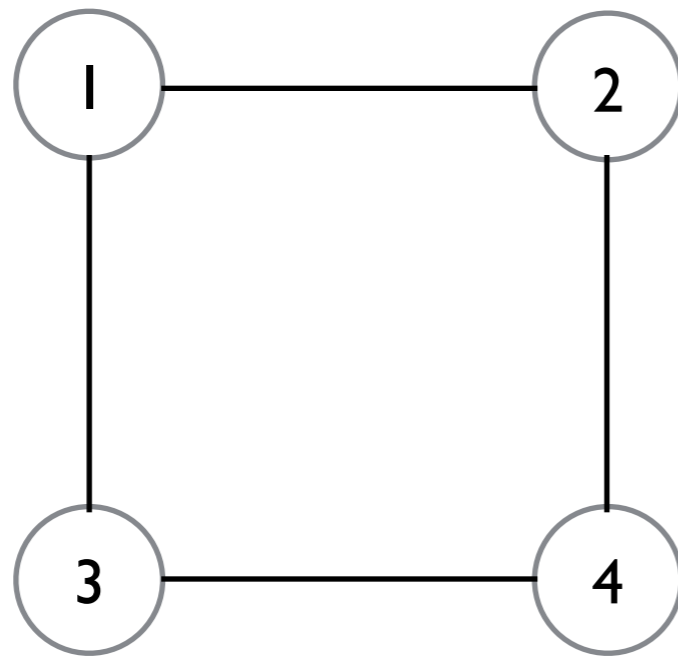
<http://www.ams.org/mathscinet/collaborationDistance.html>

Adjacency Matrix

$N \times N$ matrix A with

$$a_{ij} = \begin{cases} 1 & (i, j) \text{ is edge} \\ 0 & \text{else} \end{cases}$$

captures all info
about graph
discussed thus far

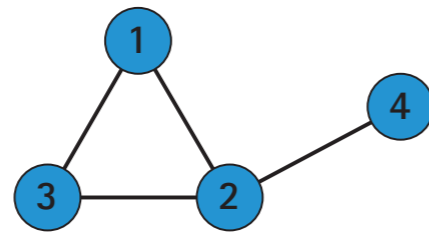


$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Adjacency Matrix

degree info

Undirected network



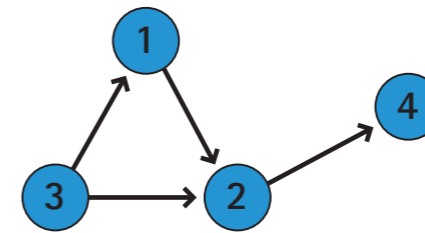
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$k_2 = \sum_{j=1}^4 A_{2j} = \sum_{i=1}^4 A_{i2} = 3$$

$$A_{ij} = A_{ji} \quad A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

Directed network



$$A_{ij} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$k_2^{in} = \sum_{j=1}^4 A_{2j} = 2$$

$$k_2^{out} = \sum_{i=1}^4 A_{i2} = 1$$

$$A_{ij} \neq A_{ji} \quad A_{ii} = 0$$

$$L = \sum_{i,j=1}^N A_{ij} \quad \langle k^{in} \rangle = \langle k^{out} \rangle = \frac{L}{N}$$

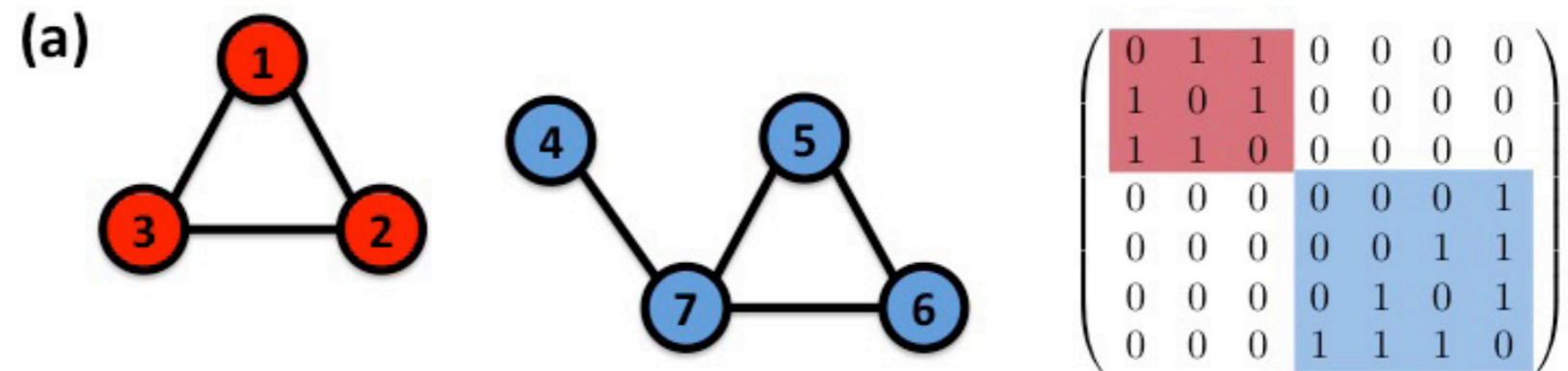
Image 2.7
The adjacency matrix.

Top: The elements of the adjacency matrix. The adjacency matrix of a directed (left column) and an undirected (right column) network. The figure highlights the fact that the degree of a node (in this case node 2) can be expressed as the sum over the appropriate column or row of the adjacency matrix. It also shows a few basic network characteristics, like the total number of links, (L), and average degree, ($\langle k \rangle$), expressed in terms of the elements of the adjacency matrix.

Baraba'si

Adjacency Matrix

connectedness



node 4 is a
“bridge”

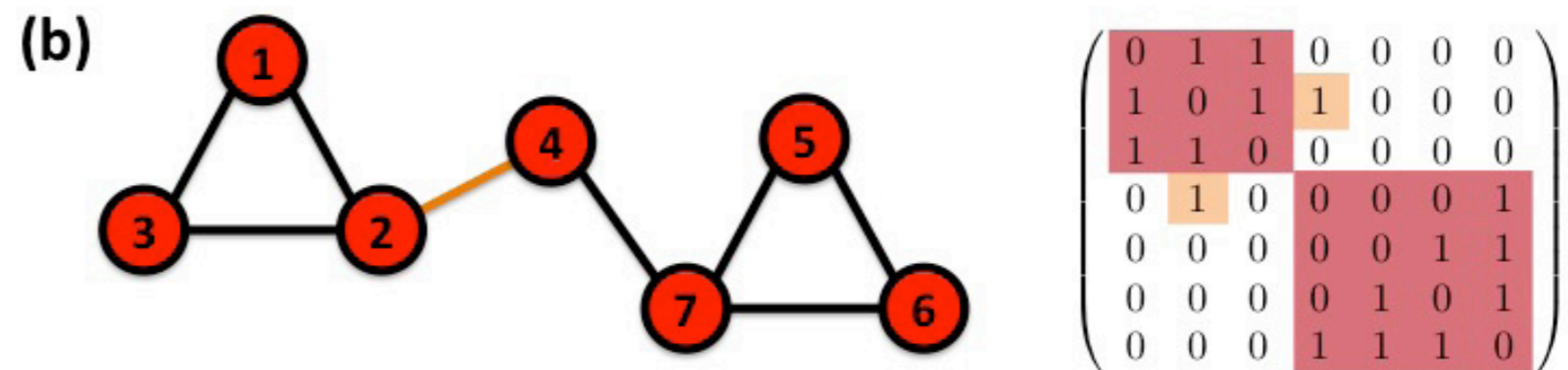
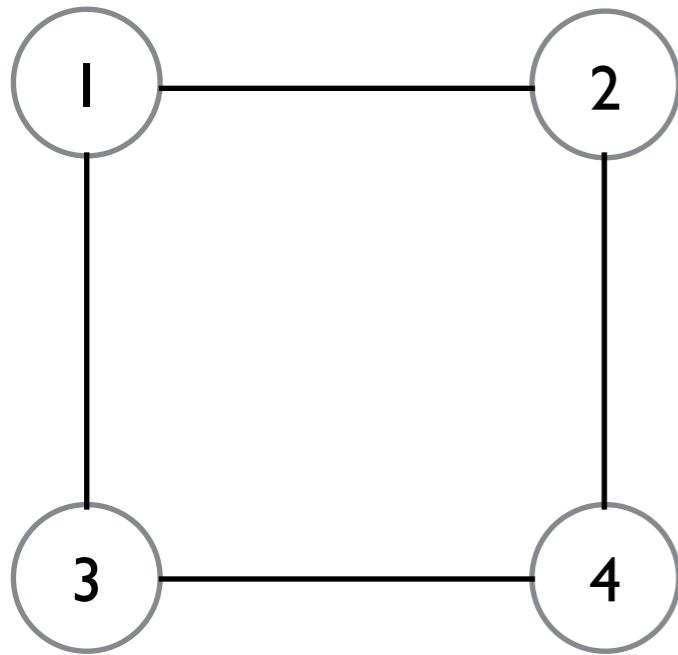


Image 2.14
Connected and disconnected networks.

Adjacency Matrix



$$A = S\Lambda S^{-1}$$



$$A^k = S\Lambda^k S^{-1}$$

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

1-hop connectivity

$$A^2 = \begin{bmatrix} 2 & 0 & 0 & 2 \\ 0 & 2 & 2 & 0 \\ 0 & 2 & 2 & 0 \\ 2 & 0 & 0 & 2 \end{bmatrix}$$

2-hop connectivity

$$A^3 = \begin{bmatrix} 0 & 4 & 4 & 0 \\ 4 & 0 & 0 & 4 \\ 4 & 0 & 0 & 4 \\ 0 & 4 & 4 & 0 \end{bmatrix}$$

3-hop connectivity

Distance by BFS

- Breadth first search:
 - Grow a tree from a given node, expanding outward
 - Increment hop-count at each expansion step
 - Disregard previously encountered nodes
- Advantage (potential)
 - Matrix multiplication complexity $\sim N*N$
 - BFS complexity $\sim N+L$ (L =number of edges)

Distance by BFS

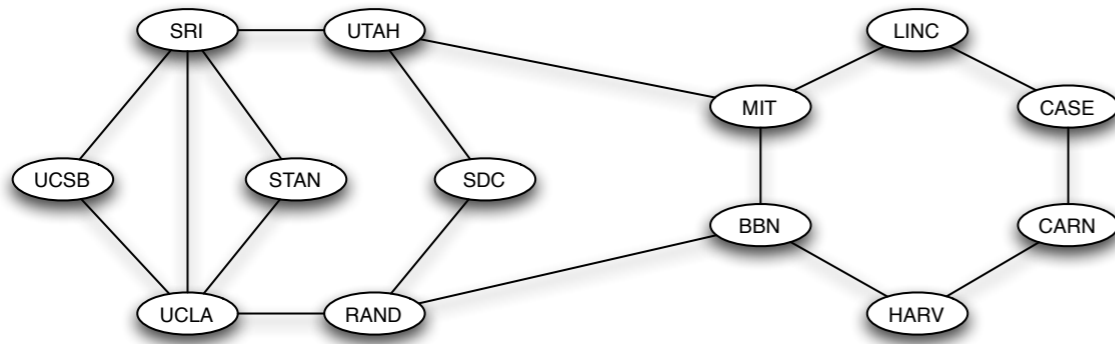


Figure 2.3: An alternate drawing of the 13-node Internet graph from December 1970.

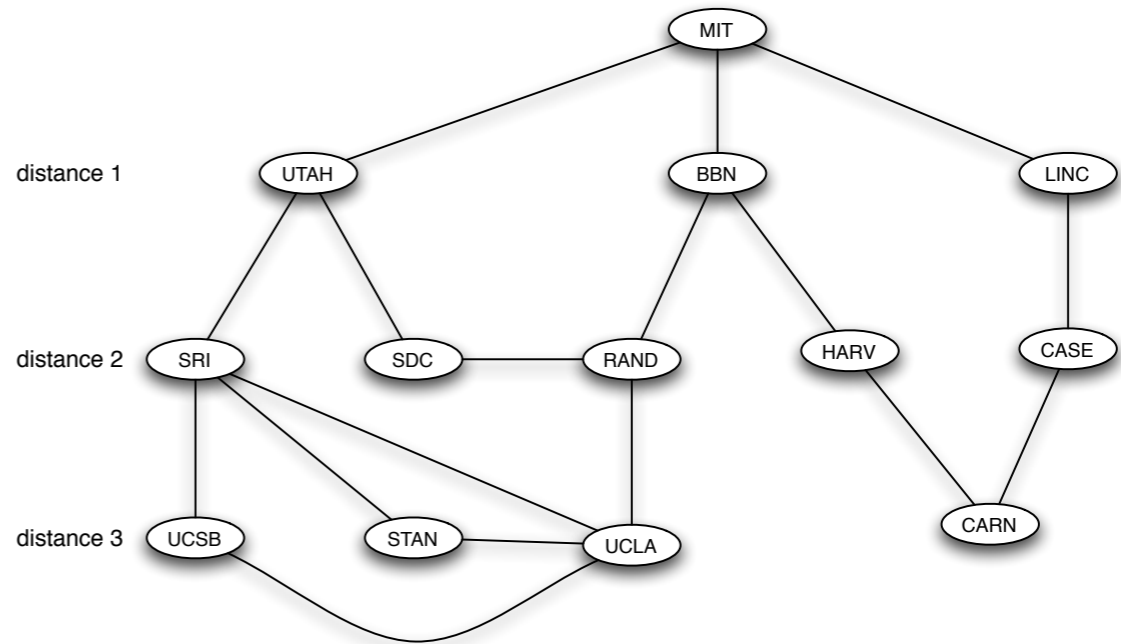


Figure 2.9: The layers arising from a breadth-first of the December 1970 Arpanet, starting at the node MIT.

Distance by BFS

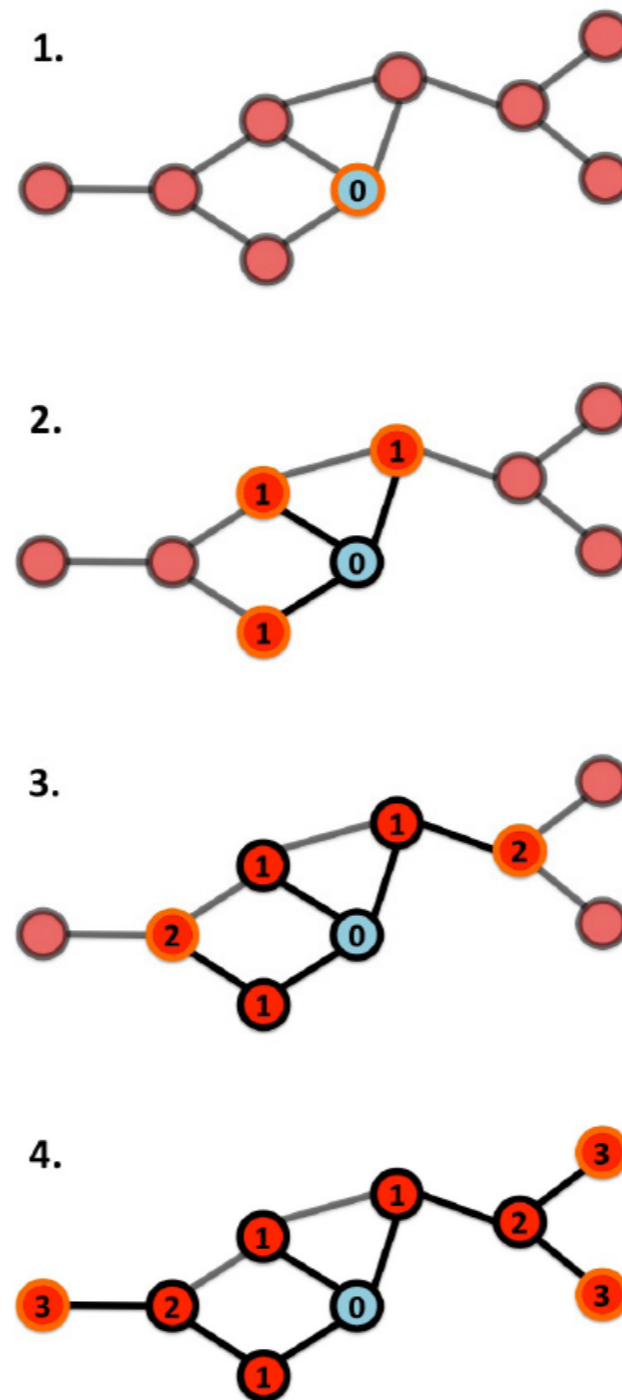
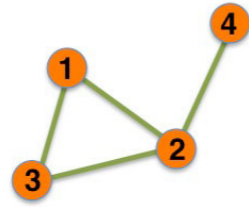


Image 2.13
The BFS algorithm applied to a small network.

Starting from the orange node, labeled "0", we identify all its neighbors, labeling them "1". Then we label "2" the unlabeled neighbors of all nodes labeled "1", and so on, in each iteration increasing the labels, until no node is left unlabeled. The length of the shortest path or the distance d_{0i} between node 0 and some other node i in the network is given by the label on node i . For example, the distance between node 0 and the leftmost node is $d_{03} = 3$.

Summary & Extensions

Undirected



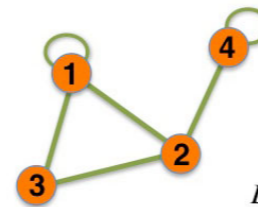
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{2L}{N}$$

UNDIRECTED NETWORK: a network whose links do not have a predefined direction. Examples: Internet, power grid, science collaboration networks, protein interactions.

Self-interactions



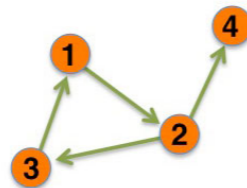
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii} \quad ?$$

SELF-INTERACTIONS: in many networks nodes do not interact with themselves, so the diagonal elements of adjacency matrix are zero, $A_{ii} = 0$, $i = 1, \dots, N$. In some systems self-interactions are allowed; in such networks, representing the fact that node i has a self-interaction. Examples: WWW, protein interactions.

Directed



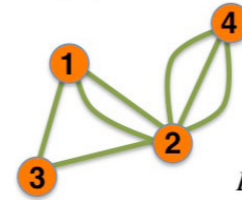
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} \neq A_{ji}$$

$$L = \sum_{i,j=1}^N A_{ij} \quad \langle k \rangle = \frac{L}{N}$$

DIRECTED NETWORK: a network whose links have selected directions. Examples: WWW, mobile phone calls, citation network.

Multigraph (undirected)



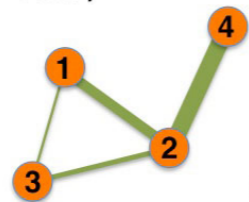
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \langle k \rangle = \frac{2L}{N}$$

MULTIGRAPH: in a multigraph nodes are permitted to have multiple links (or parallel links) between them. Hence A_{ij} can have any positive integer.

Weighted (undirected)



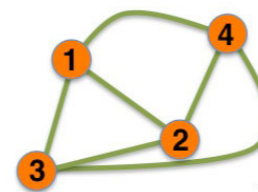
$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{ij} = A_{ji}$$

$$L = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \langle k \rangle = \frac{2L}{N}$$

WEIGHTED NETWORK: a network whose links have a predefined weight, strength or flow parameter. The elements of the adjacency matrix are $A_{ij} = 0$ if i and j are not connected, or $A_{ij} = w_{ij}$ if there is a link with weight w_{ij} between them. For unweighted (binary) networks, the adjacency matrix only indicates the presence ($A_{ij} = 1$) or the absence ($A_{ij} = 0$) of a link between two nodes. Examples: Mobile phone calls, email network.

Complete Graph (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A_{ii} = 0 \quad A_{i \neq j} = 1$$

$$L = L_{\max} = \frac{N(N-1)}{2} \quad \langle k \rangle = N-1$$

COMPLETE GRAPH: in a complete graph all nodes are connected to each other; no self-connections.