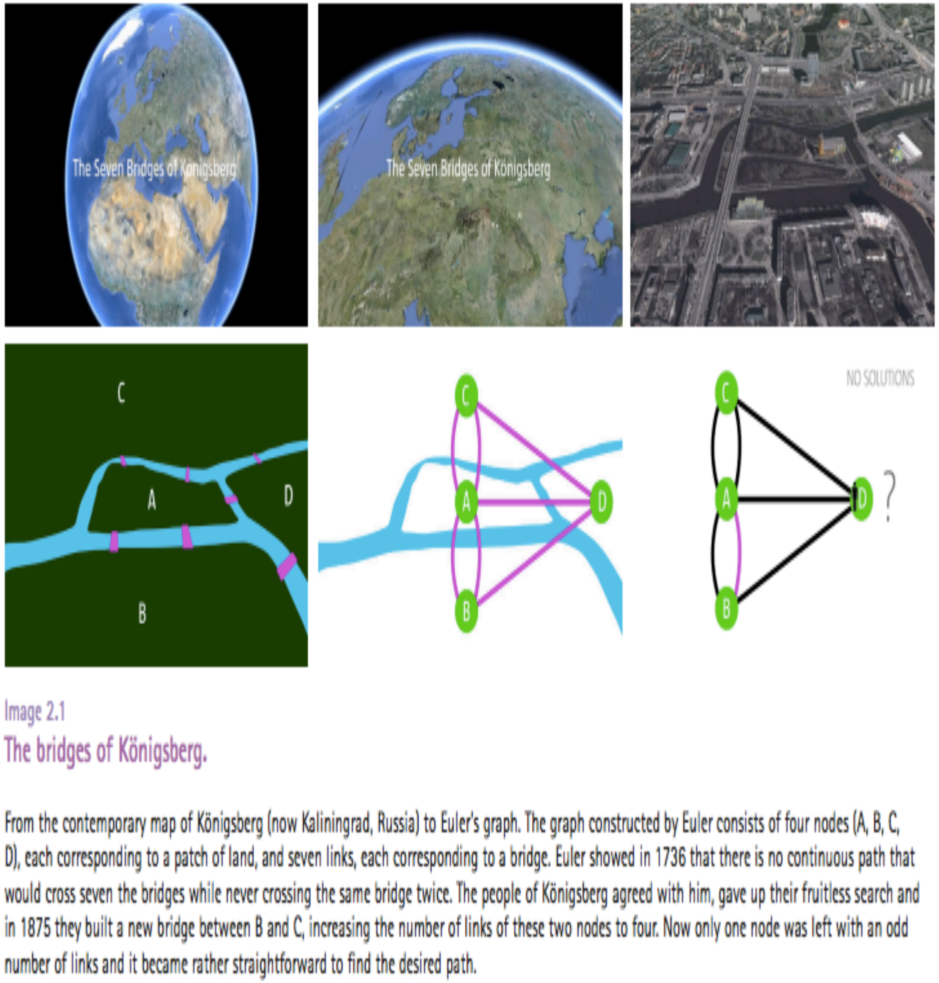

The Seven Bridges of Königsberg
The Seven Bridges of Königsberg
C
A
D
B
C
A
D
B
NO SOLUTIONS
C
A
D
?
B

Image 2.1

The bridges of Königsberg.

From the contemporary map of Königsberg (now Kaliningrad, Russia) to Euler's graph. The graph constructed by Euler consists of four nodes (A, B, C, D), each corresponding to a patch of land, and seven links, each corresponding to a bridge. Euler showed in 1736 that there is no continuous path that would cross seven the bridges while never crossing the same bridge twice. The people of Königsberg agreed with him, gave up their fruitless search and in 1875 they built a new bridge between B and C, increasing the number of links of these two nodes to four. Now only one node was left with an odd number of links and it became rather straightforward to find the desired path.

Complex Networks: Graph Theory

Ana Paula Couto

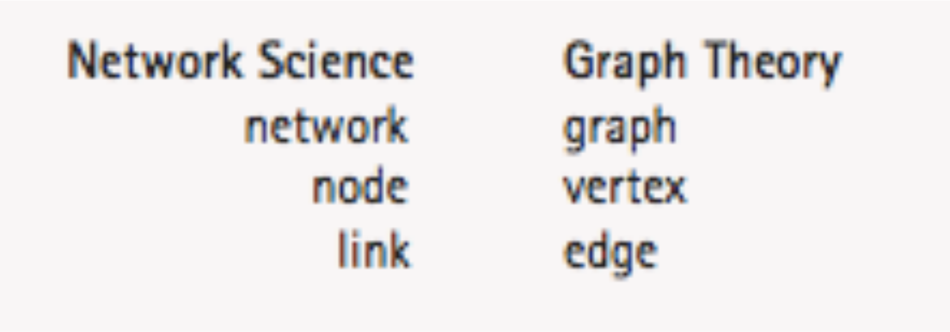
Computer Science Department

Universidade Federal de Minas Gerais

Network Science Book – Albert Barabasi

Complex networks: basic concepts

Proprieties on Complex networks are based on the basic concepts from graph theory



Network Science	Graph Theory
network	graph
node	vertex
link	edge

Network Science Book – Albert Barabasi

Network definition

Set of entities often called nodes or vertices and the direct interactions between them, called links or edges

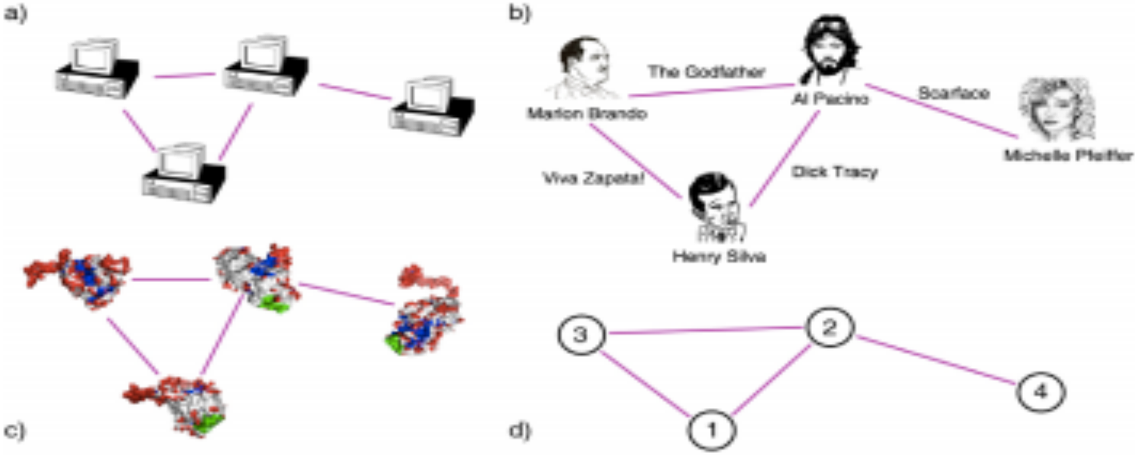


Image 2.3

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

Network Science Book – Albert Barabasi

Network definition

- Number of nodes (N): Total number of entities (components) in the system

- Number of links (L): Total number of interactions between the nodes

Network definition

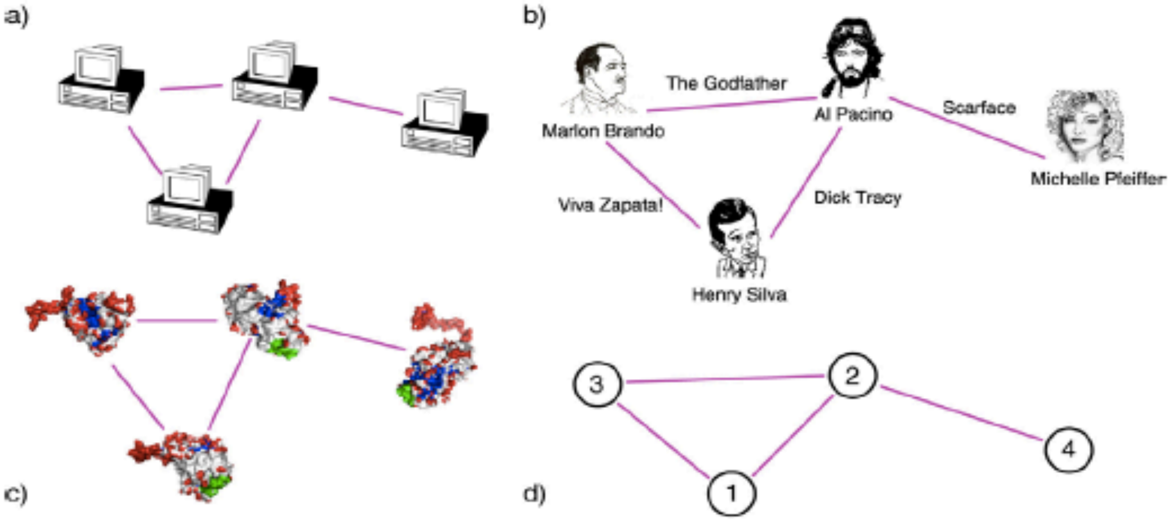


Image 2.3

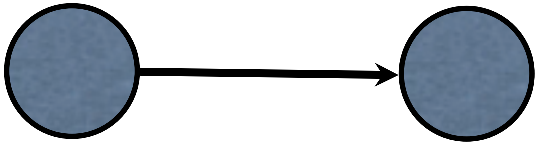
Real systems of quite different nature can have the same network representation.

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).

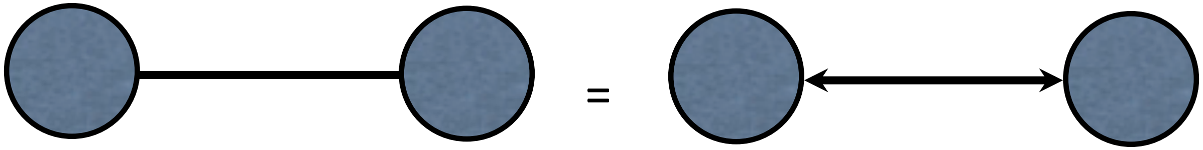
Network Science Book – Albert Barabasi

Network definition

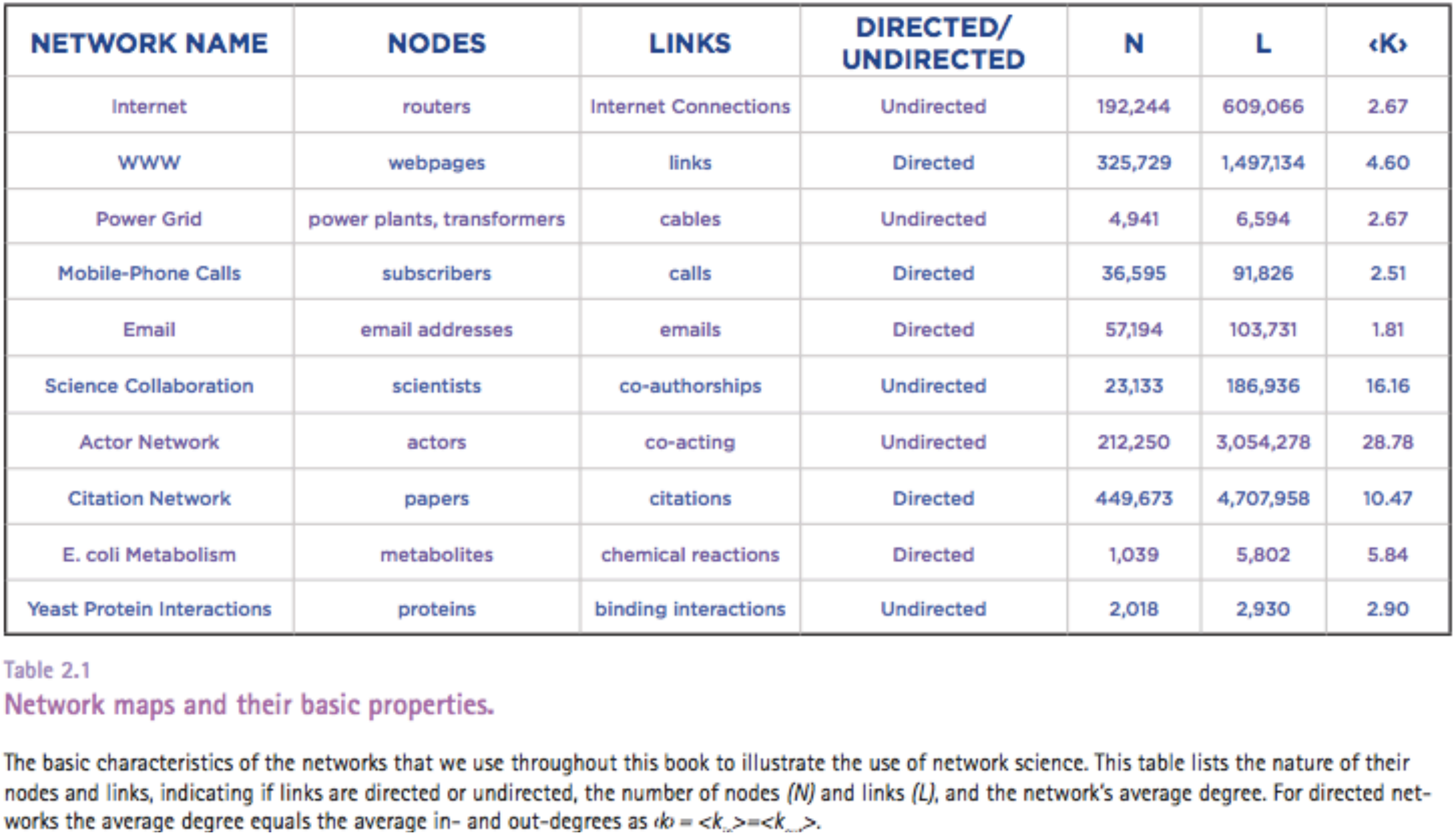
- Directed: All links are directed



- Undirected: All links are undirected



Examples



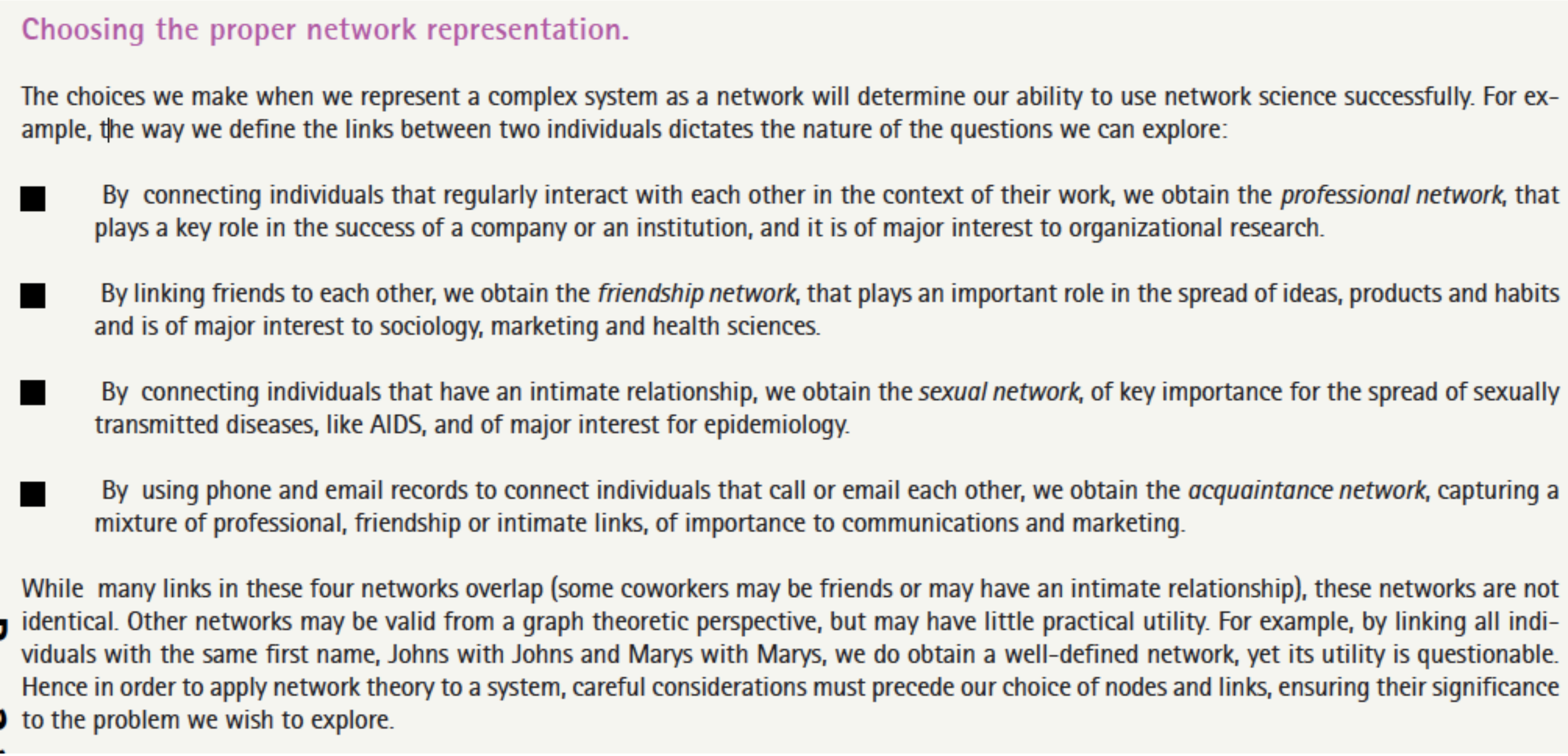
NETWORK NAME	NODES	LINKS	DIRECTED/ UNDIRECTED	N	L	‹K›
Internet	routers	Internet Connections	Undirected	192,244	609,066	2.67
WWW	webpages	links	Directed	325,729	1,497,134	4.60
Power Grid	power plants, transformers	cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	subscribers	calls	Directed	36,595	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.

The basic characteristics of the networks that we use throughout this book to illustrate the use of network science. This table lists the nature of their nodes and links, indicating if links are directed or undirected, the number of nodes *(N)* and links *(L)*, and the network's average degree. For directed networks the average degree equals the average in- and out-degrees as $\langle k \rangle = \langle k_{in} \rangle = \langle k_{out} \rangle$.

Network Science Book – Albert Barabasi



Choosing the proper network representation.

The choices we make when we represent a complex system as a network will determine our ability to use network science successfully. For example, the way we define the links between two individuals dictates the nature of the questions we can explore:

- By connecting individuals that regularly interact with each other in the context of their work, we obtain the *professional network*, that plays a key role in the success of a company or an institution, and it is of major interest to organizational research.

- By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

- By connecting individuals that have an intimate relationship, we obtain the *sexual network*, of key importance for the spread of sexually transmitted diseases, like AIDS, and of major interest for epidemiology.

- By using phone and email records to connect individuals that call or email each other, we obtain the *acquaintance network*, capturing a mixture of professional, friendship or intimate links, of importance to communications and marketing.

While many links in these four networks overlap (some coworkers may be friends or may have an intimate relationship), these networks are not identical. Other networks may be valid from a graph theoretic perspective, but may have little practical utility. For example, by linking all individuals with the same first name, Johns with Johns and Marys with Marys, we do obtain a well-defined network, yet its utility is questionable. Hence in order to apply network theory to a system, careful considerations must precede our choice of nodes and links, ensuring their significance to the problem we wish to explore.

Network Science Book – Albert Barabasi

NETWORK PROPERTIES

Node degree

- Number of links a node *i* has to other nodes
  - Number of mobile phone contacts an individual has in the call graph
  - Number of citations a research paper gets in the citation network
- Undirected Network: k_i
- Directed Network: k_i^{in} (incoming degree) and k_i^{out} (outgoing degree) , then:

$$k_i = k_i^{in} + k_i^{out}$$

From node degree

- Total number of links - Undirected Network

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i$$

- Mean degree - Undirected Network

$$\langle k \rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2L}{N}$$

From node degree

- Total number of links - Directed Network

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out}$$

- Mean degree - Directed Network

$$(k^{in}) = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = (k^{out}) = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} = \frac{L}{N}$$

Degree distribution

- Provides the probability p_k that a randomly selected node in network has degree k

$$\sum_{k=1}^{\infty} p_k = 1$$

- Central role in network theory: most network properties are based on p_k

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

Degree distribution

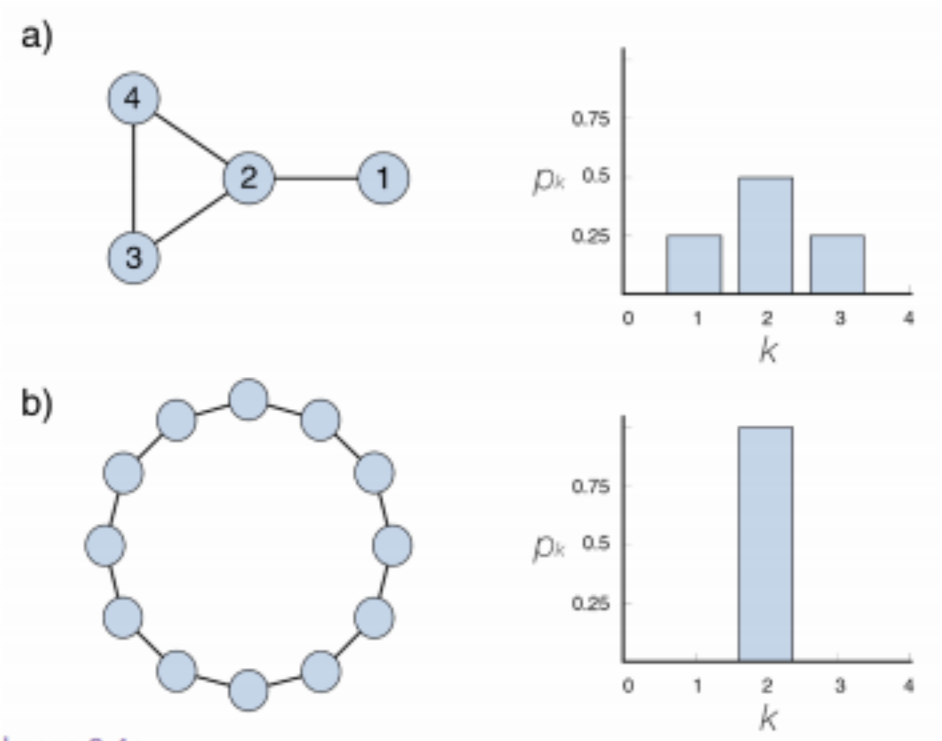


Image 2.4a

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 $p_k = \delta(k - 2)$.

Network Science Book – Albert Barabasi

Degree distribution

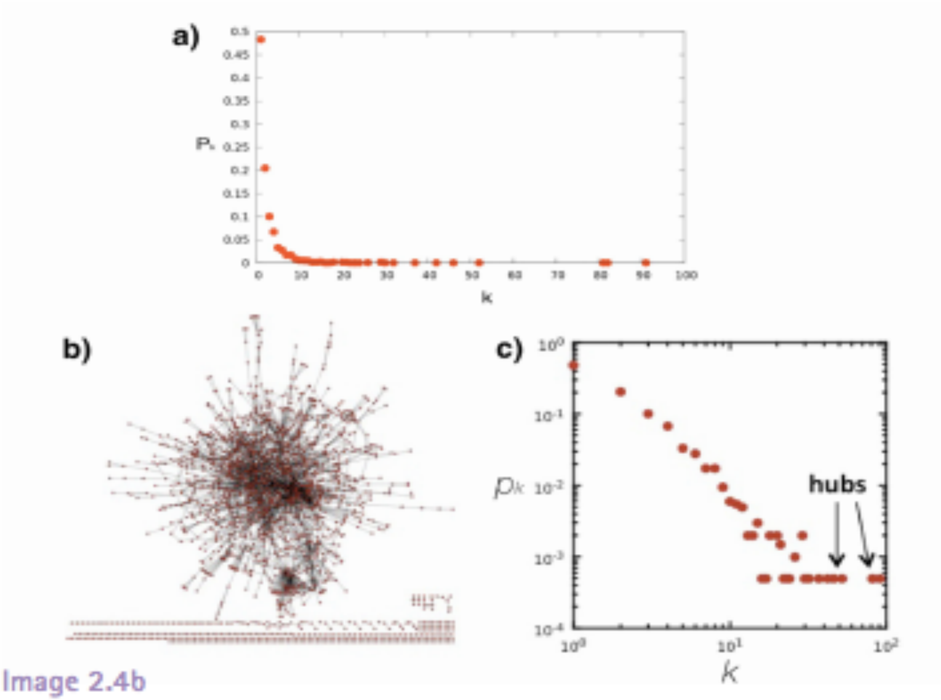


Image 2.4b

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.

Network Science Book – Albert Barabasi

Real networks are sparse

- L <<<< L_{max} (the total number of links present in a complete graph)

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

- WWW has about 1.5 million links. If WWW were to be a complete graph, $L_{max} \approx 10^{12}$ links

- The web graph has only a 10^{-6} fraction of the links it could have, making it a sparse network

Bipartite networks

- Network whose nodes can be divided into two disjoints sets U and V such that each link connects a U-node to a V-node

Bipartite networks

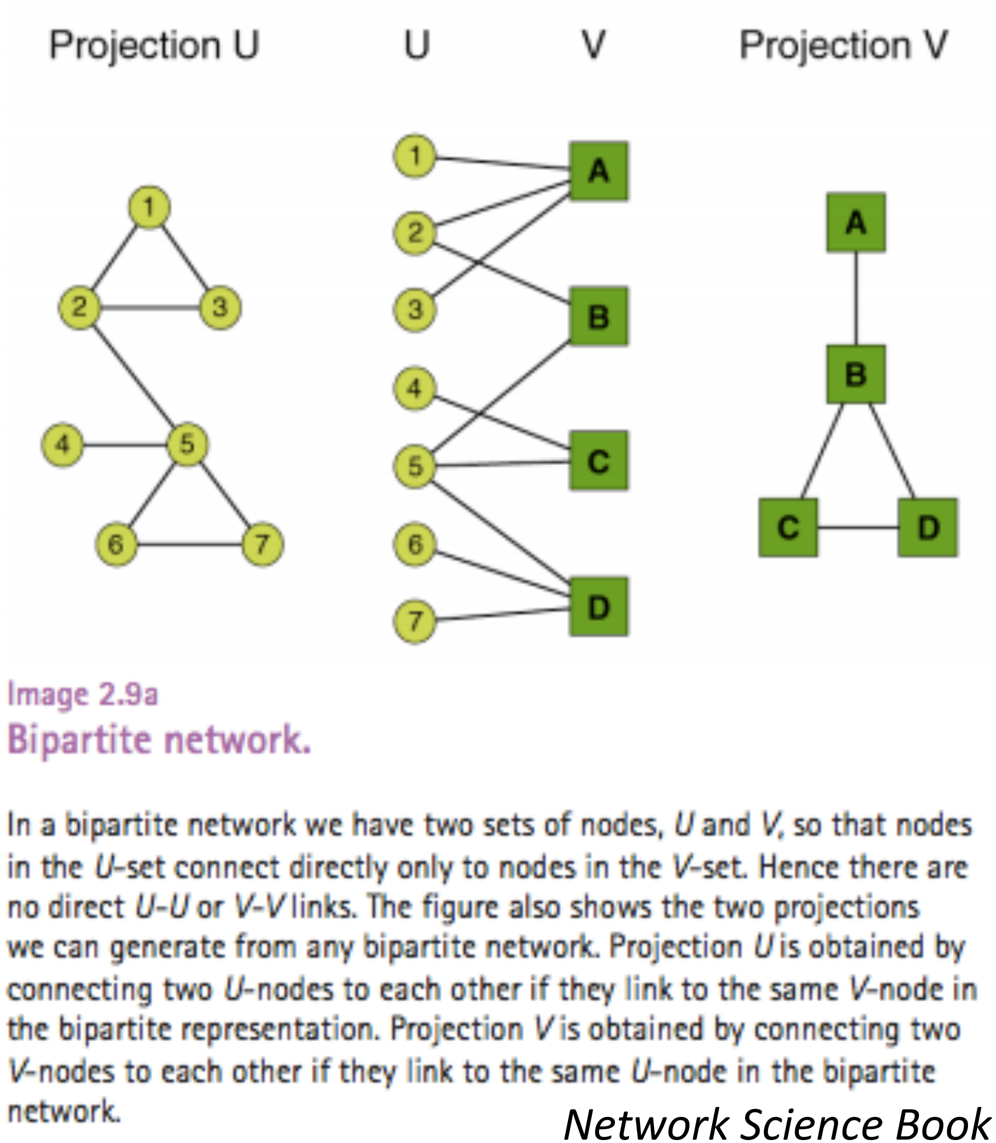
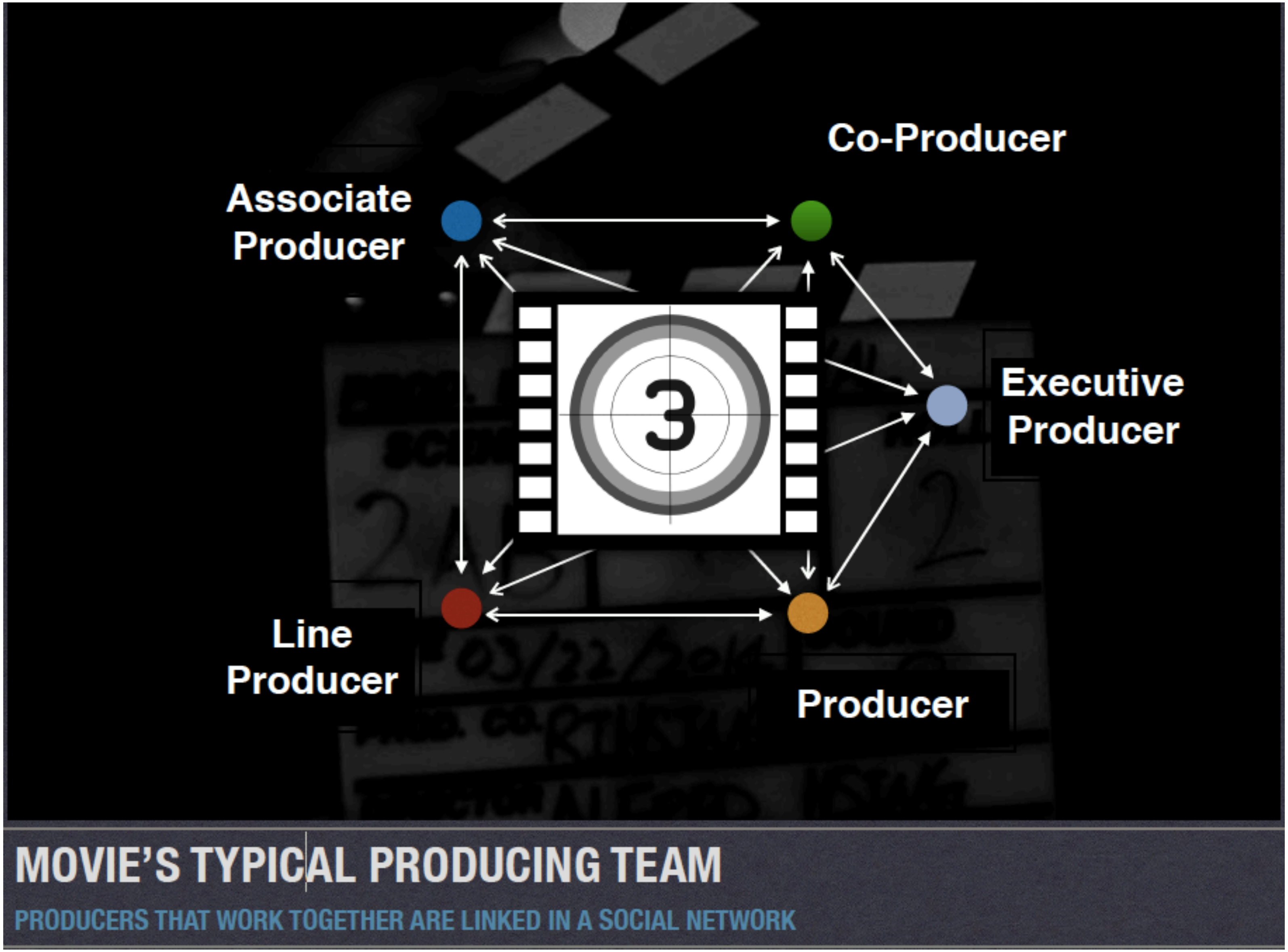


Image 2.9a
Bipartite network.

In a bipartite network we have two sets of nodes, *U* and *V*, so that nodes in the *U*-set connect directly only to nodes in the *V*-set. Hence there are no direct *U*-*U* or *V*-*V* links. The figure also shows the two projections we can generate from any bipartite network. Projection *U* is obtained by connecting two *U*-nodes to each other if they link to the same *V*-node in the bipartite representation. Projection *V* is obtained by connecting two *V*-nodes to each other if they link to the same *U*-node in the bipartite network.

Network Science Book – Albert Barabasi

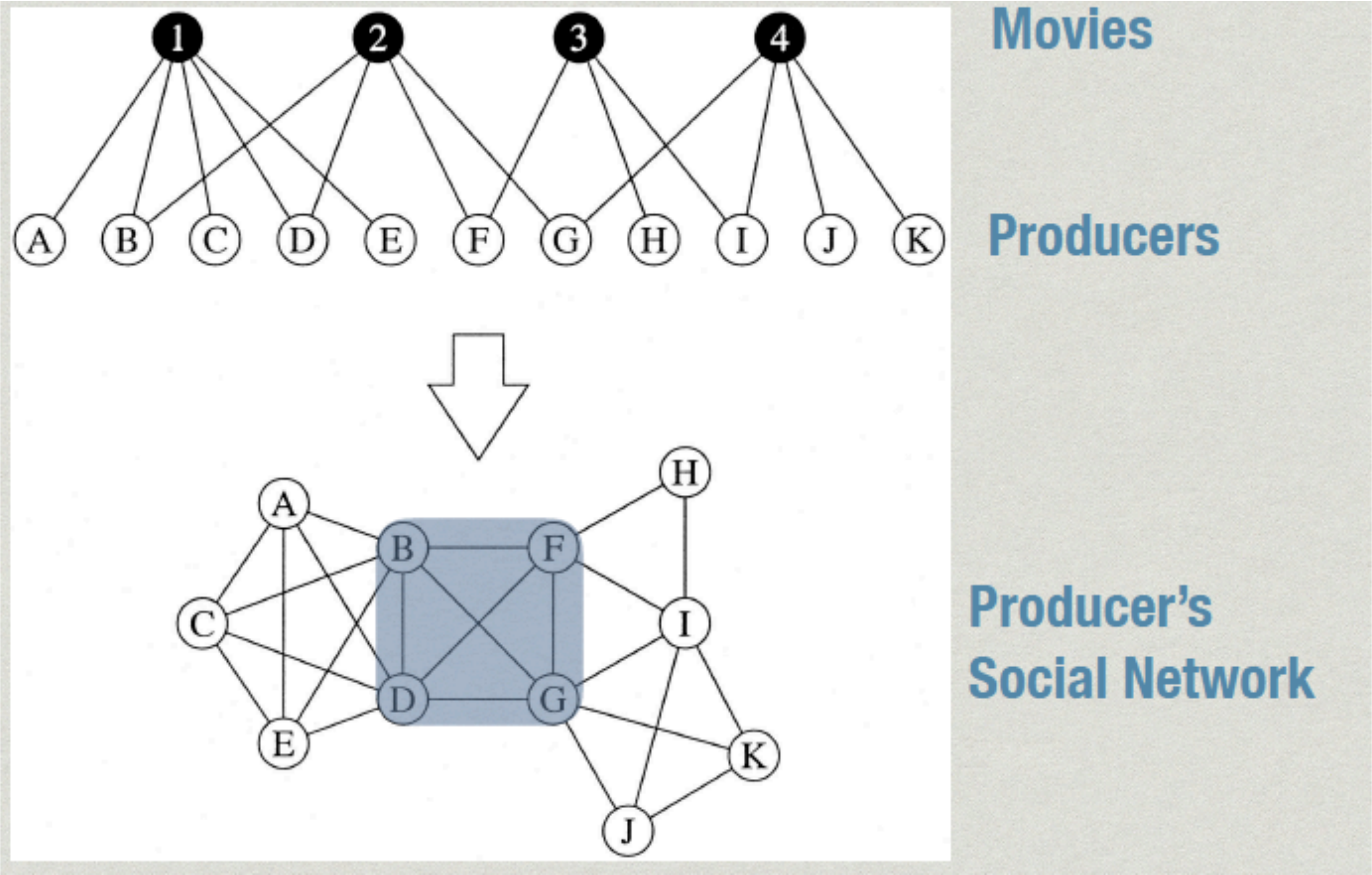
Bipartite networks - Examples



MOVIE'S TYPICAL PRODUCING TEAM

PRODUCERS THAT WORK TOGETHER ARE LINKED IN A SOCIAL NETWORK

Bipartite networks - Examples



Bipartite networks - Examples

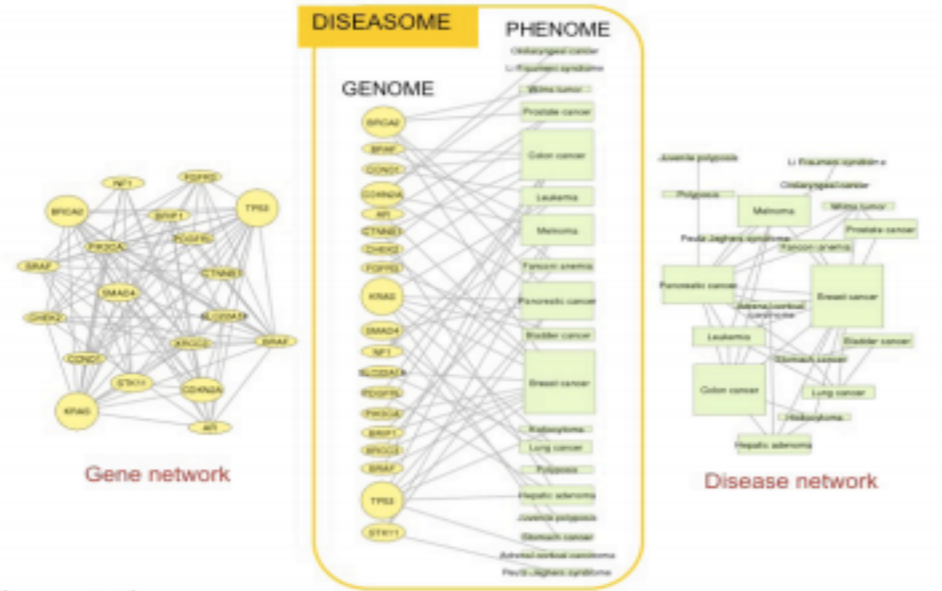


Image 2.9b
Bipartite network.

The *human diseaseome* is a bipartite network, whose nodes are diseases (*U*) and genes (*V*), in which a disease is connected to a gene if mutations in that gene are known to affect the particular disease [4]. One projection of the diseaseome is the *gene network*, whose nodes are genes, two genes being connected if they are associated with the same disease. The second projection is the disease network, whose nodes are diseases, two diseases being connected if the same genes are associated with them, indicating that the two diseases have common genetic origins. The figure shows a subset of the diseaseome, focusing on cancers. The full human diseaseome map, connecting 1,283 disorders via 1,777 shared disease genes. (After [4])

Network Science Book – Albert Barabasi

Tripartite networks - Examples

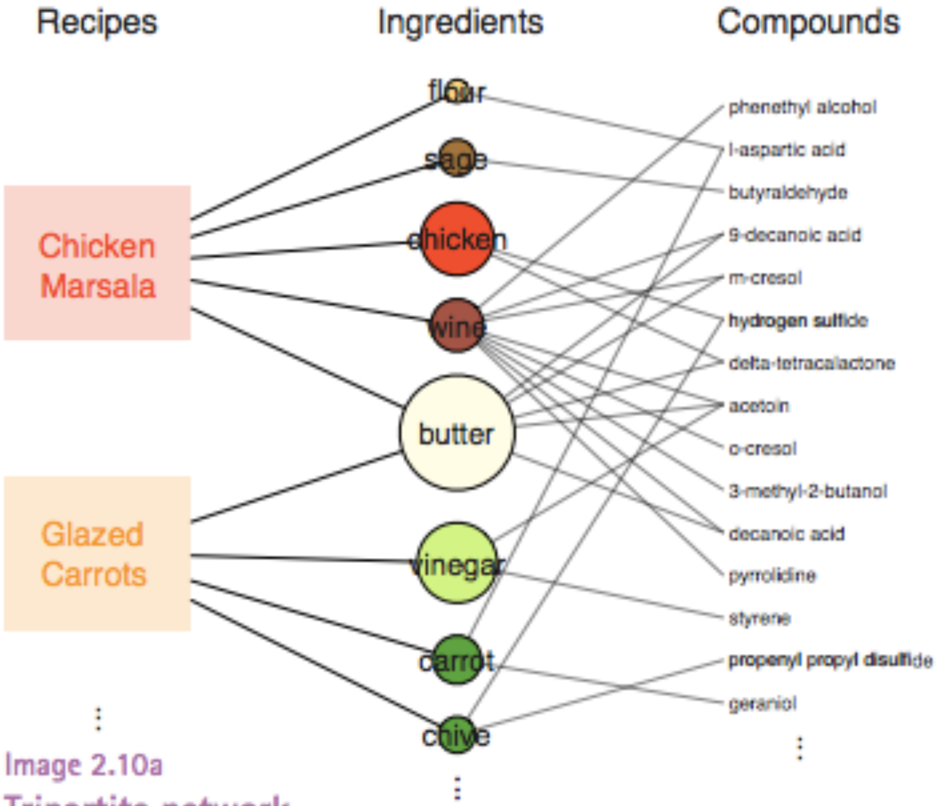


Image 2.10a

Tripartite network.

The tripartite recipe-ingredient-compound network, in which one set of nodes are recipes, like Chicken Marsala, the second set corresponds to the ingredients each recipe has (like flour, sage, chicken, wine, and butter for Chicken Marsala), and the third set captures the flavor compounds, or chemicals that contribute to the taste of a particular ingredient.

Network Science Book – Albert Barabasi

Tripartite networks - Examples

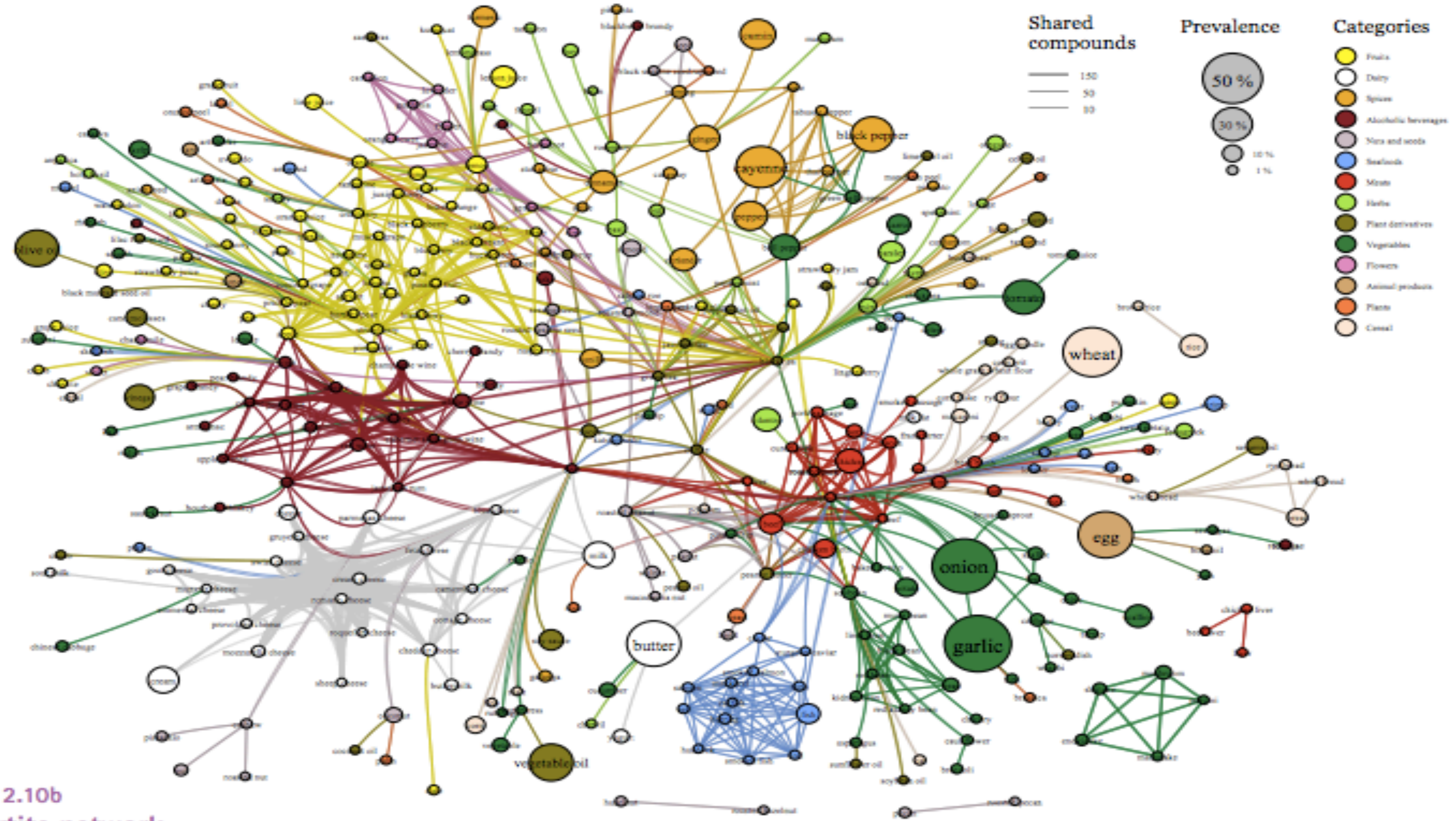


Image 2.10b

Tripartite network.

A projection of the tripartite network, resulting in the ingredient network, often called the flavor network. Each node denotes an ingredient; the node color indicating the food category and node size reflects the ingredient prevalence in recipes. Two ingredients are connected if they share a significant number of flavor compounds, link thickness representing the number of shared compounds between the two ingredients (After [12]).

Network Science Book – Albert Barabasi

Paths and distances

- What is the distance between two webpages on the WWW network?
- What is the distance between two individuals who may or may not know each other?

A path is a route that runs along the links of the network, its length representing the number of links the path contains.

Shortest path

- The shortest path between nodes *i* and *j* is the path with fewest number of links

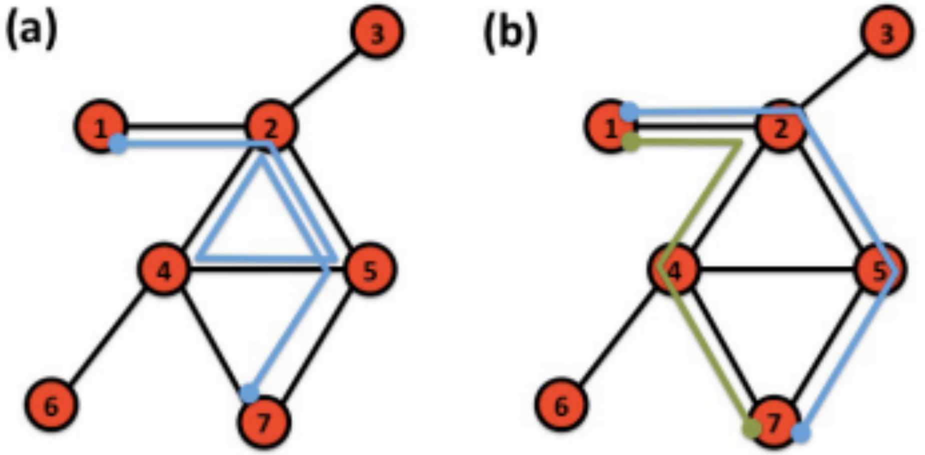


Image 2.11

The adjacency matrix is typically sparse.

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

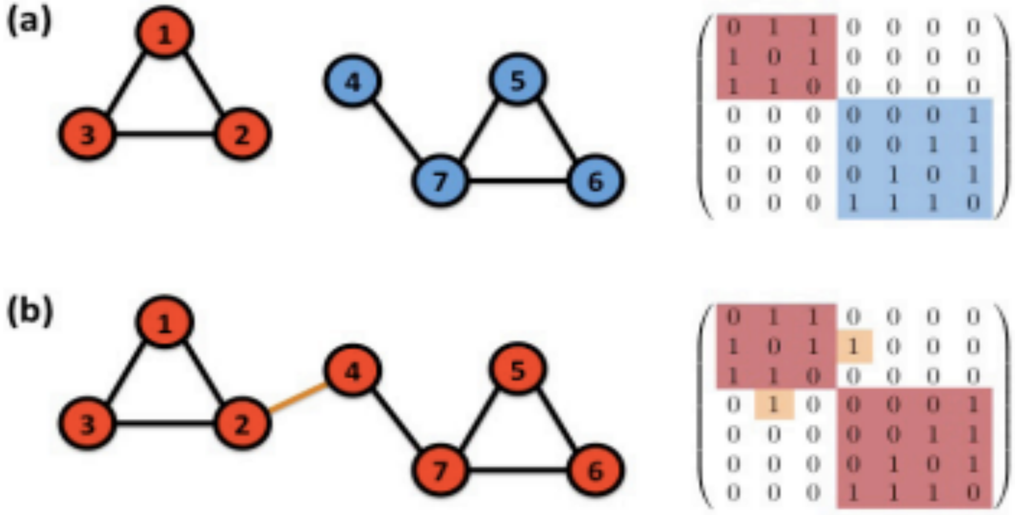
Network diameter

It is the maximal shortest path in the network

Connectedness and components

- Networks are built to ensure connectedness
  - Phone network without the possibility of calling any valid phone number
  - Email network with access only to a small number of addresses

Connectedness and components



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

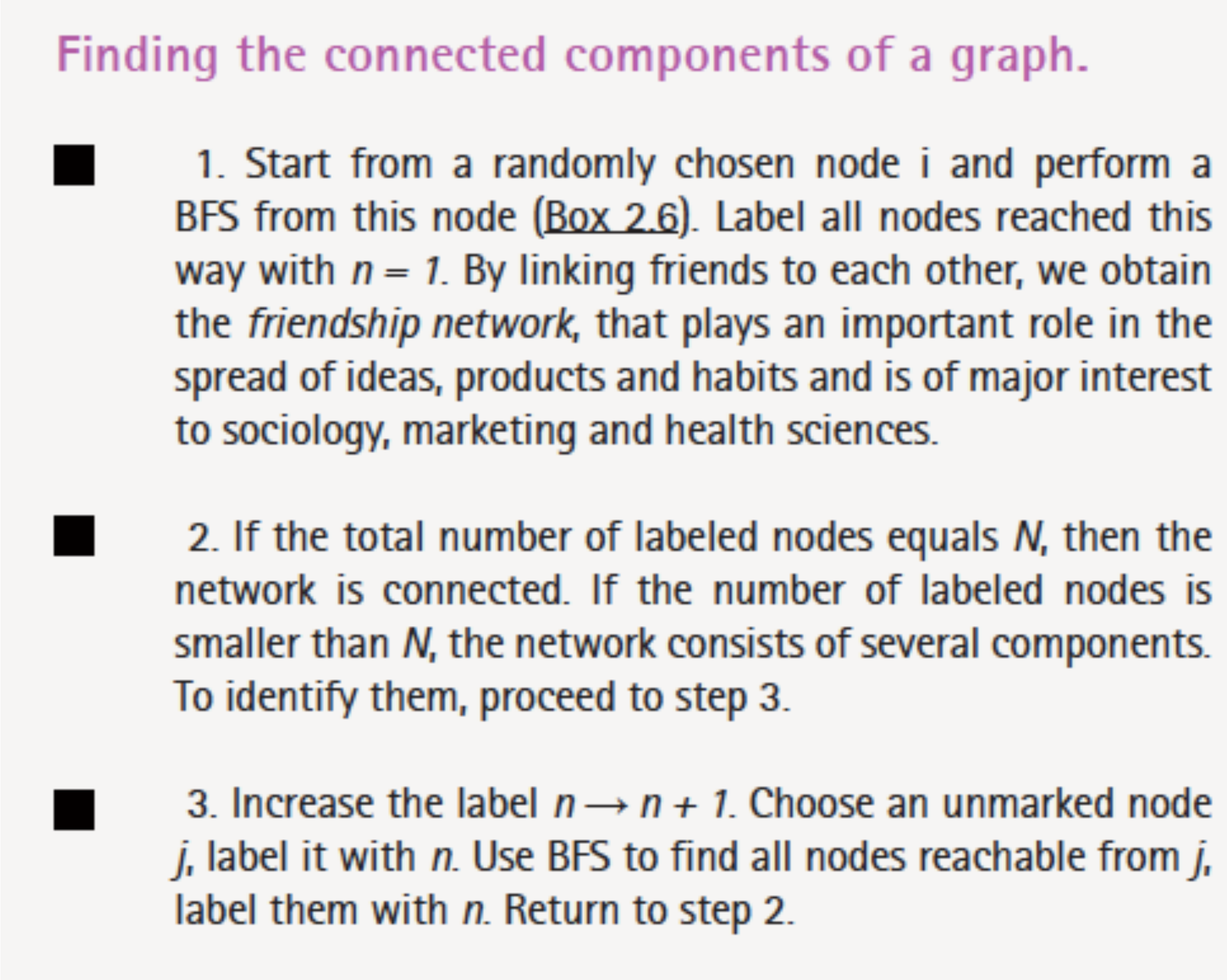
Image 2.14

Connected and disconnected networks.

(a) The network consists of two disconnected components, i.e. there is a path between any pair of nodes in the (1,2,3) component, as well in the (4,5,6,7) component. However, there are no paths between nodes that belong to different connected components. The right panel shows the adjacently matrix of the network. If the network consists of disconnected components, the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements of the matrix are contained in square blocks along the diagonal of the matrix and all other elements are zero.

(b) The addition of one link, called a *bridge*, can turn a disconnected network into a single connected component. Now there is a path between every pair of nodes in the network. Consequently the adjacency matrix cannot be written in a block diagonal form.

Connectedness and components



Finding the connected components of a graph.

- 1. Start from a randomly chosen node i and perform a BFS from this node (Box 2.6). Label all nodes reached this way with $n = 1$. By linking friends to each other, we obtain the *friendship network*, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.

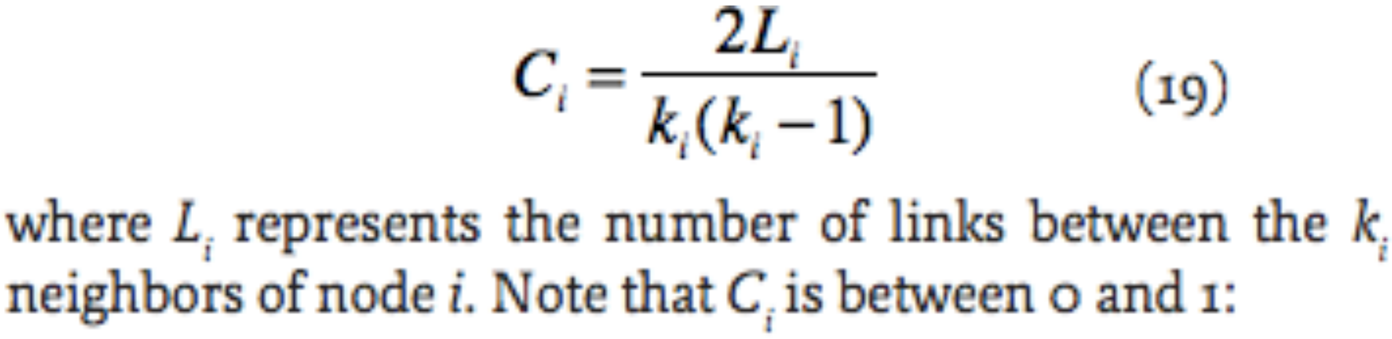
- 2. If the total number of labeled nodes equals N, then the network is connected. If the number of labeled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.

- 3. Increase the label $n \rightarrow n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them with n. Return to step 2.

Clustering coefficient

- Local clustering coefficient captures the degree to which the neighbors of a given node link to each other
- In social networks: *friend of my friend is also my friend*

Clustering coefficient



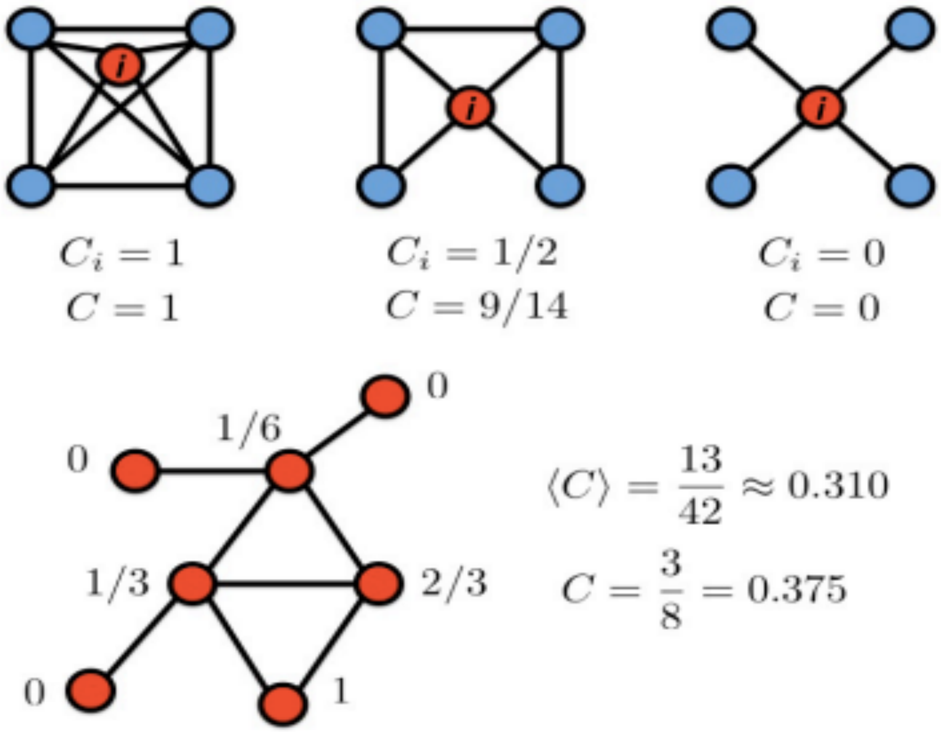
$$C_i = \frac{2L_i}{k_i(k_i - 1)} \qquad (19)$$

where L_i represents the number of links between the k_i neighbors of node i. Note that C_i is between 0 and 1:

Clustering coefficient

- C_i = 0 if none of the neighbors of node i link to each other;
- C_i = 1 if the neighbors of node i form a complete graph, i.e. they all link to each other (Image 2.7).
- In general C_i is the probability that two neighbors of a node link to each other: C = 0.5 implies that there is a 50% chance that two neighbors of a node are linked.
- In summary C_i measures the network's local density: the more densely interconnected the neighborhood of node i, the higher is C_i.

Clustering coefficient



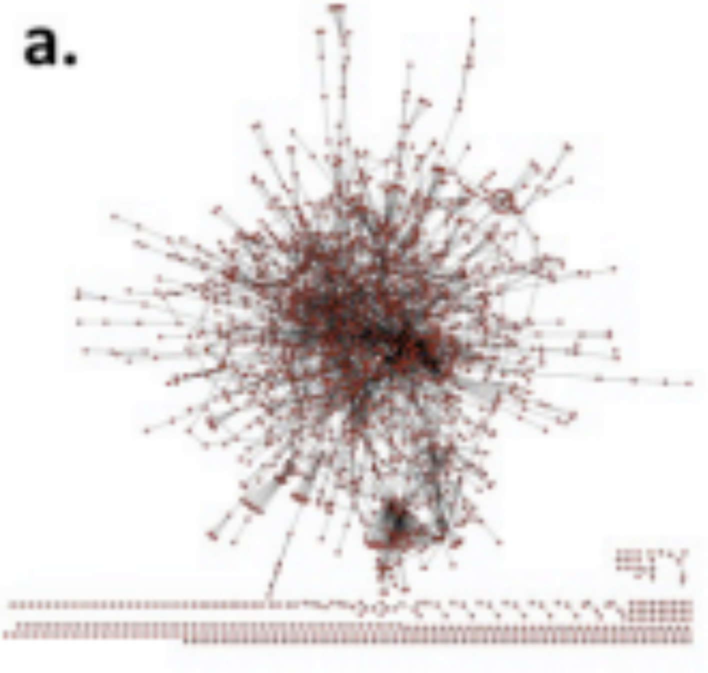
i
$C_i = 1$
$C = 1$
$C_i = 1/2$
$C = 9/14$
$C_i = 0$
$C = 0$
0
1/6
0
1/3
2/3
0
1
$\langle C \rangle = \frac{13}{42} \approx 0.310$
$C = \frac{3}{8} = 0.375$

Image 2.15

Clustering Coefficient.

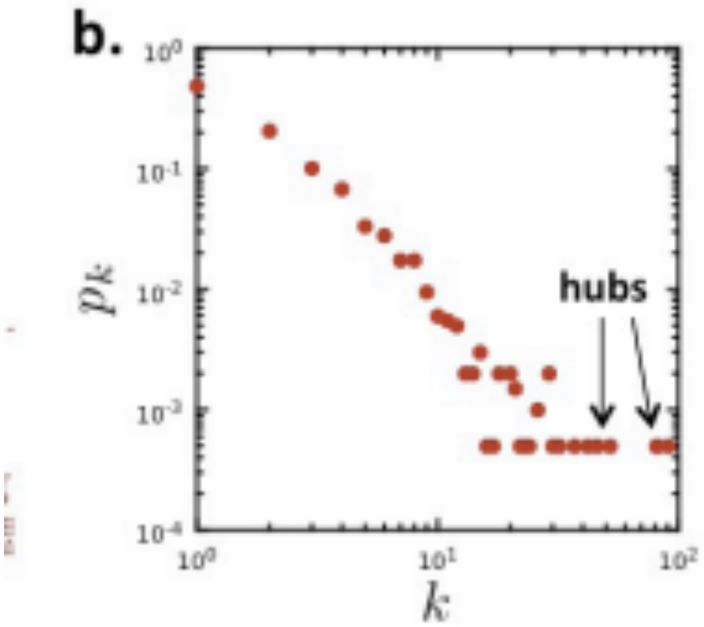
The local clustering coefficient, C_i , of the central node with degree k_i=4 for three different configurations of its neighborhood. The clustering coefficient measures the local density of links in a node's vicinity. The bottom figure shows a small network, with the local clustering coefficient of a node shown next to each node. Next to the figure we also list the network's average clustering coefficient <C>, according to Eq. (20), and its global clustering coefficient C, declined in Appendix A, Eq. (21). Note that for nodes with degrees k_i=0,1, the clustering coefficient is taken to be zero.

Case study: Protein-Protein Interaction Network



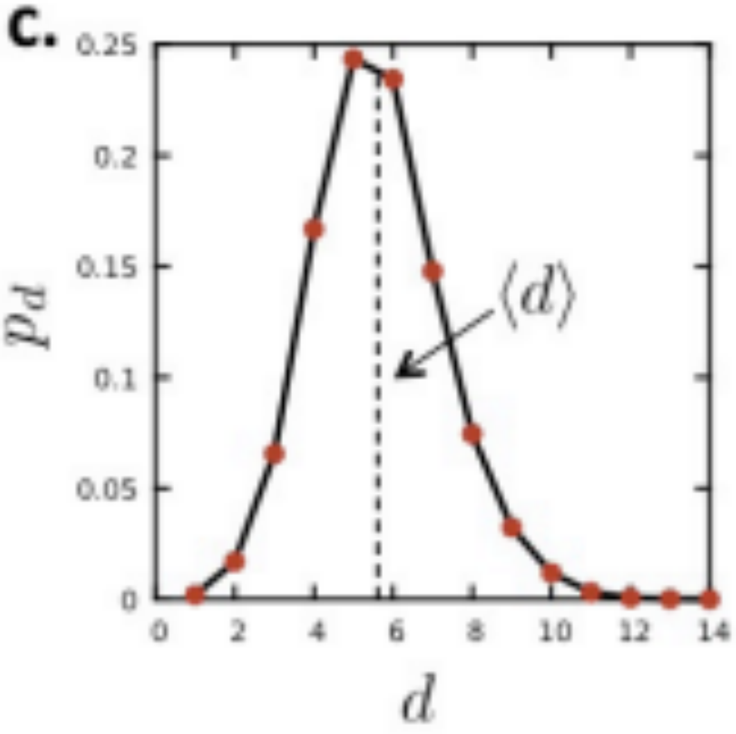
(a) The protein-protein interaction (PPI) network of yeast, a network frequently studied not only by biologists, but also by network scientists. The nodes of the network are proteins and links correspond to experimentally documented protein-protein binding interactions. The figure indicates that the network, consisting of N=2,018 nodes and L=2,930 links, has a giant component that connects 81% of the proteins, several smaller components, and numerous isolated proteins that do not interact with any other node.

Case study: Protein-Protein Interaction Network



(b) The degree distribution, p_k, of the PPI network, providing the probability that a randomly chosen node has degree k. As $N_k = Np_k$, the degree distribution provides the number of nodes with degree k. The degree distribution indicates that proteins of widely different degrees coexist in the PPI: most nodes have only a few links, a few, however, have dozens of links, representing the hubs of the network.

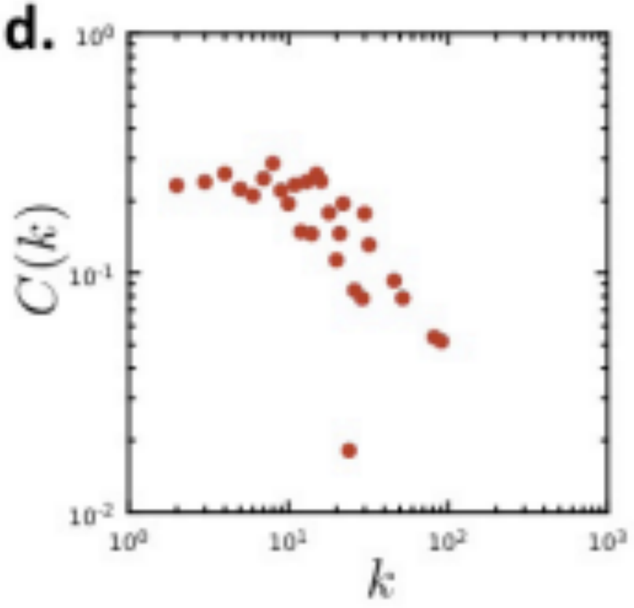
Case study: Protein-Protein Interaction Network



(c) The distance distribution, pd for the PPI network, providing the probability that two randomly chosen nodes have a distance d between them (shortest path). The dotted line shows the average path length, which is ‹d› =5.61.

Case study: Protein-Protein Interaction Network

d.

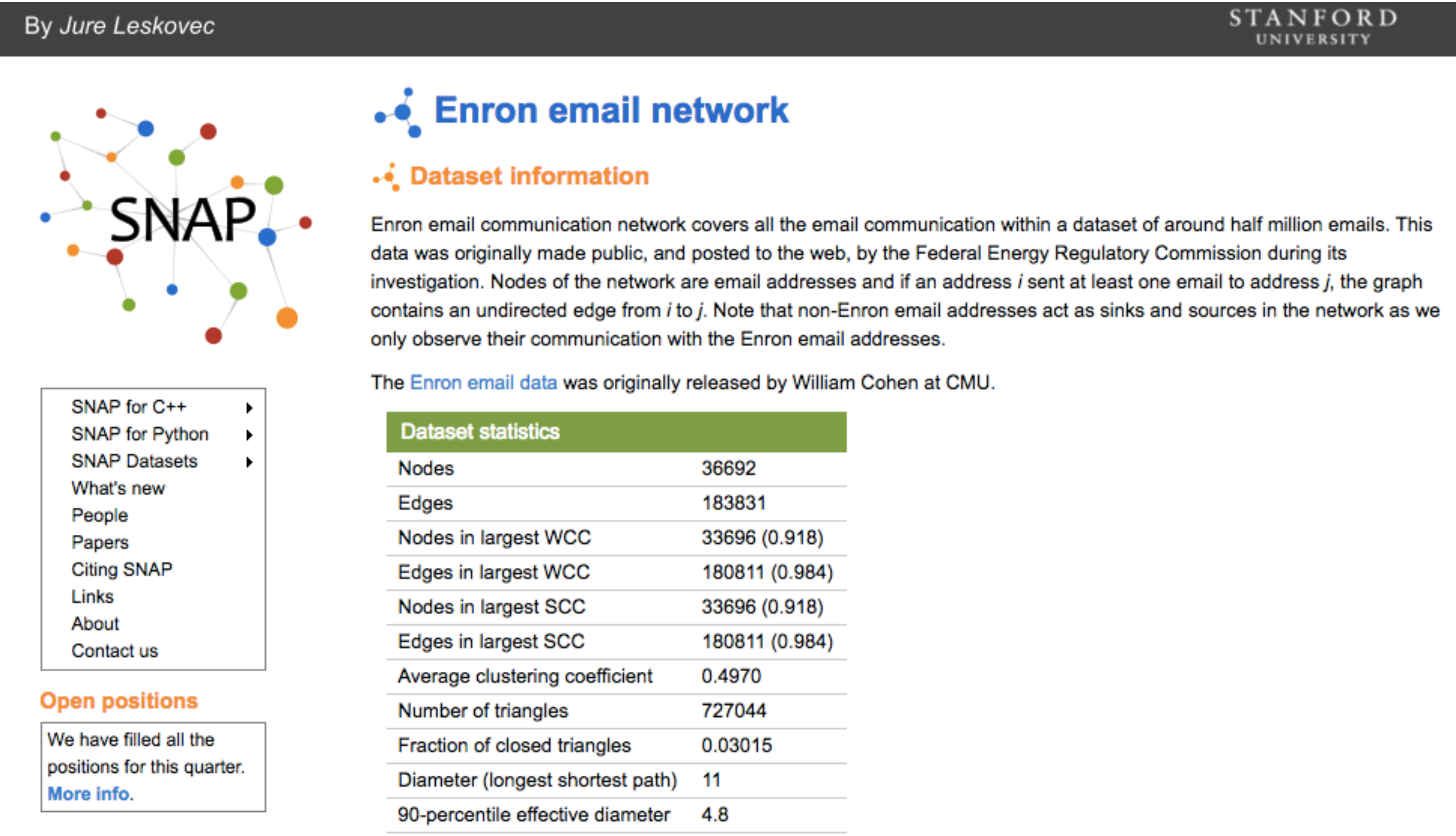


(d) The dependence of the average clustering coefficient on the node's degree, *k*. The *C(k)* function is measured by averaging over the local clustering coefficient of all nodes with the same degree *k*.

NetwokX

NetworkX is a Python language software package for creation, manipulation, and study of the structure, dynamics, and function of complex networks.

https://networkx.github.io/documentation/networkx-1.10/overview.html



By *Jure Leskovec*

STANFORD UNIVERSITY

SNAP

- SNAP for C++
- SNAP for Python
- SNAP Datasets
- What's new
- People
- Papers
- Citing SNAP
- Links
- About
- Contact us

Open positions

We have filled all the positions for this quarter. **More info**.

Enron email network

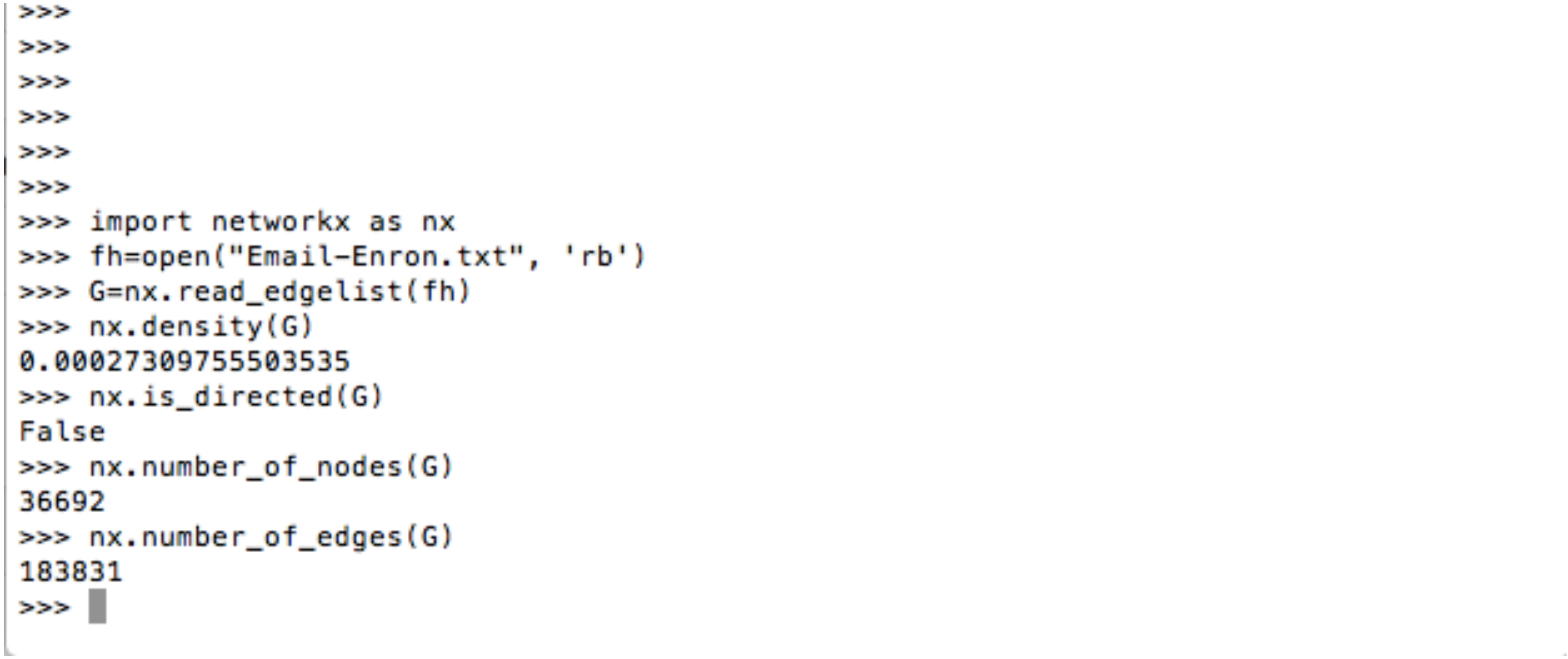
Dataset information

Enron email communication network covers all the email communication within a dataset of around half million emails. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. Nodes of the network are email addresses and if an address *i* sent at least one email to address *j*, the graph contains an undirected edge from *i* to *j*. Note that non-Enron email addresses act as sinks and sources in the network as we only observe their communication with the Enron email addresses.

The Enron email data was originally released by William Cohen at CMU.

Dataset statistics	
Nodes	36692
Edges	183831
Nodes in largest WCC	33696 (0.918)
Edges in largest WCC	180811 (0.984)
Nodes in largest SCC	33696 (0.918)
Edges in largest SCC	180811 (0.984)
Average clustering coefficient	0.4970
Number of triangles	727044
Fraction of closed triangles	0.03015
Diameter (longest shortest path)	11
90-percentile effective diameter	4.8

NetwokX



```
>>>
>>>
>>>
>>>
>>>
>>>
>>> import networkx as nx
>>> fh=open("Email-Enron.txt", 'rb')
>>> G=nx.read_edgelist(fh)
>>> nx.density(G)
0.00027309755503535
>>> nx.is_directed(G)
False
>>> nx.number_of_nodes(G)
36692
>>> nx.number_of_edges(G)
183831
>>> 
```