Network and Spatial Analyses

26. February 2020

Lecture:

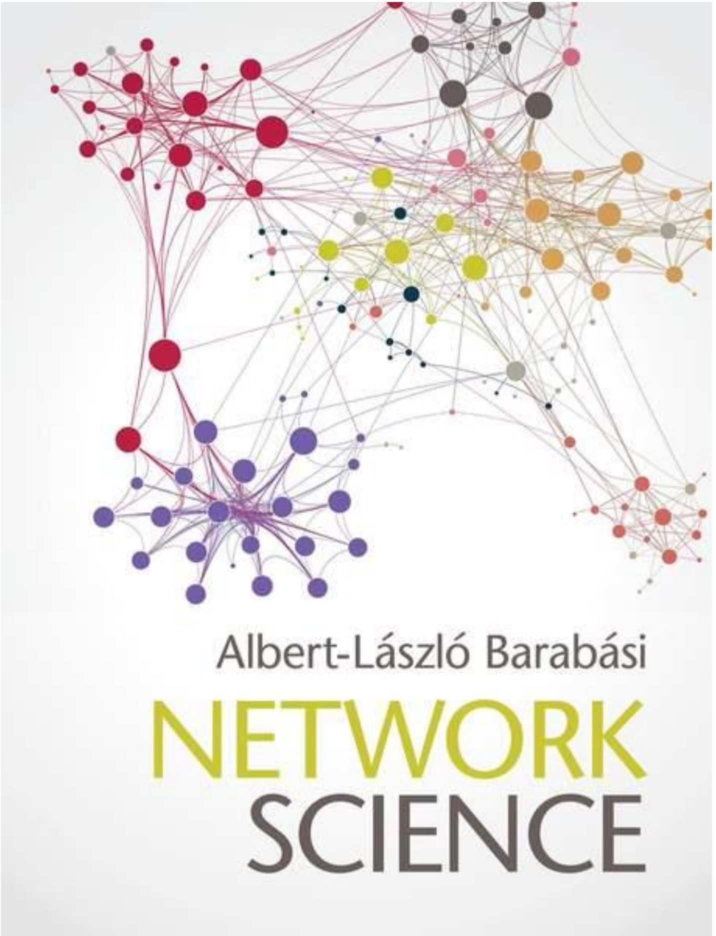
Introduction to Graph Theory I.

- Degree distribution
- Adjacency Matrix
- Bipartite Networks
- Paths and Distances

Seminar:

Intro network exercise in R

Course Github page: https://github.com/bokae/anet_course

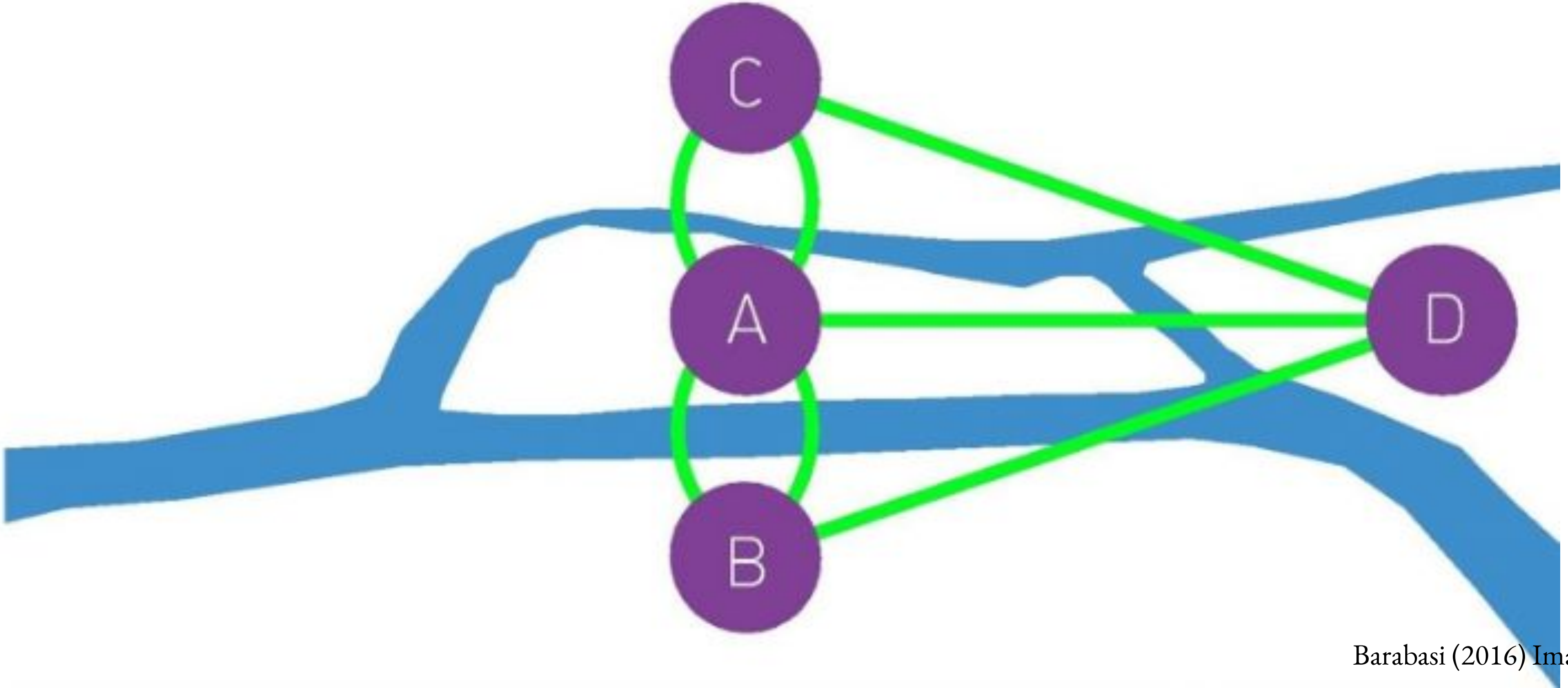


- Networks and Graphs
- Average Degree and Degree Distribution
- Adjacency Matrix
- Weighted Networks
- Bipartite Networks
- Paths and Distances
- Clustering
- Centrality Measures
- Application

Graphs, tables and pictures from Barabási, A. L. (2016). Network science. Cambridge university press. if we don’t indicate differently!

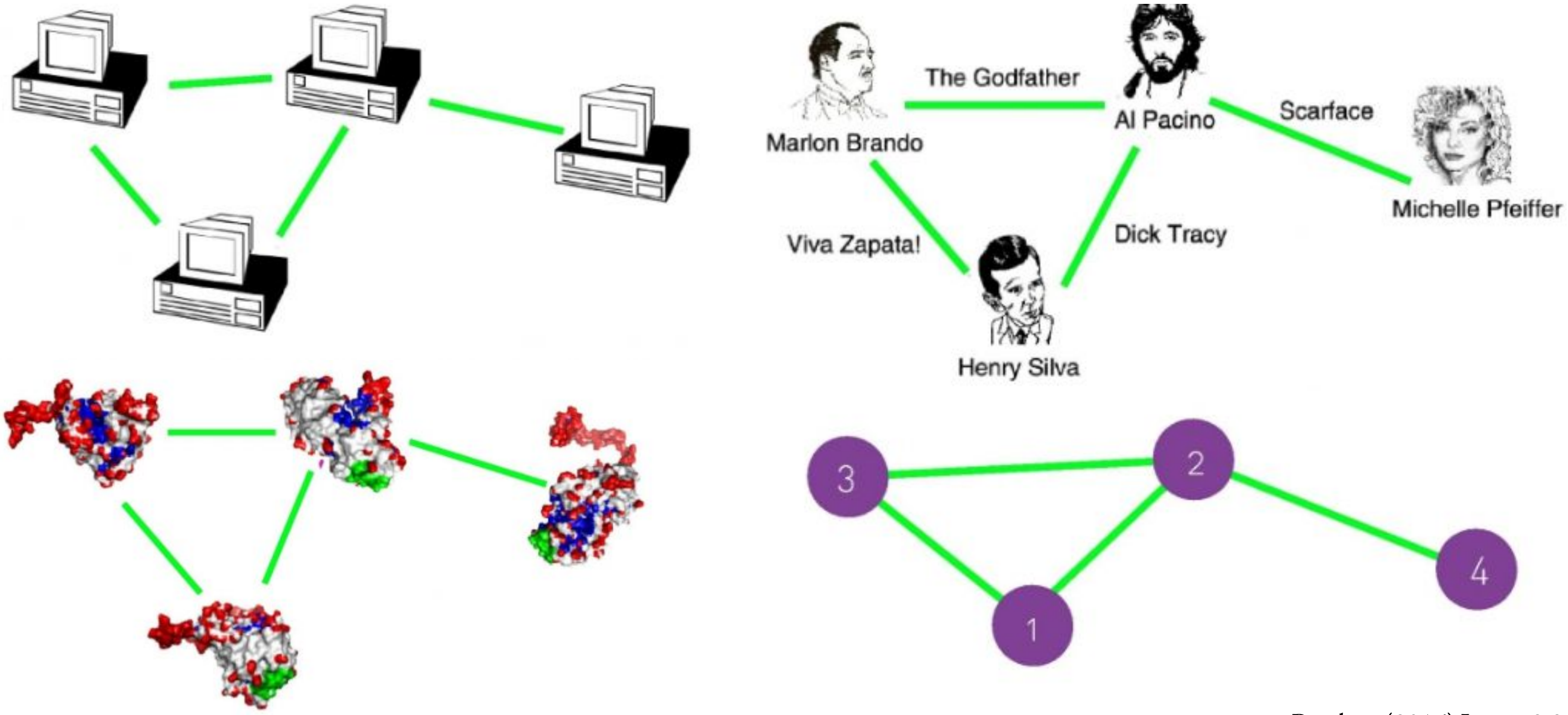
The Bridges of Königsberg

Can we find a continuous path that would cross the seven bridges while never crossing the same bridge twice?



Barabasi (2016) Image 2.1

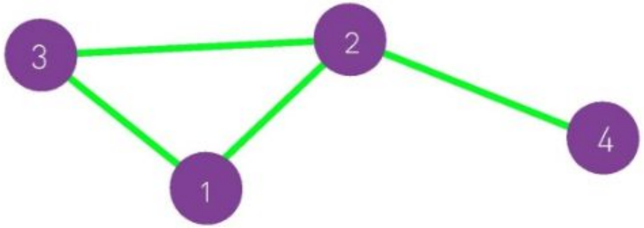
Networks and Graphs



Barabasi (2016) Image 2.2

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
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	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

Networks and Graphs



Average Degree

We can express the total number of links, L, as the sum of the node of degrees, $k_1=2$, $k_2=3$, $k_3=2$, $k_4=1$

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

In undirected networks average degree:

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

In case of directed network the total number of links is a sum of outgoing and ingoing degrees:

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

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

Networks and Graphs

Degree Distribution

Degree Distribution p_k gives the probability that a given node has a degree of k, normalized to 1:

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

,where there are N number of nodes, the normalized distribution can be expressed by:

$$p_k = \frac{N_k}{N}$$

,thus the number of degree-k nodes given by the degree distribution $N_k = Np_k$.

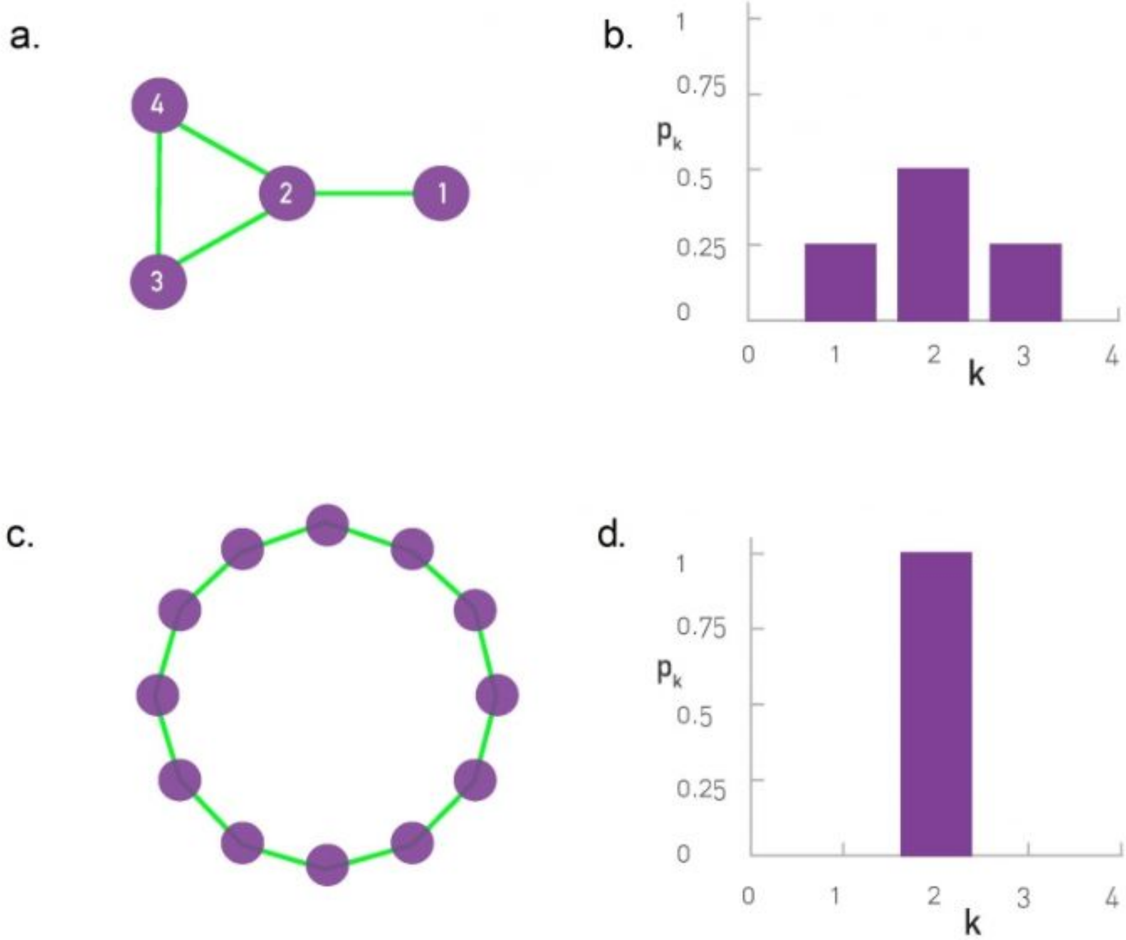
Networks and Graphs

Degree Distribution

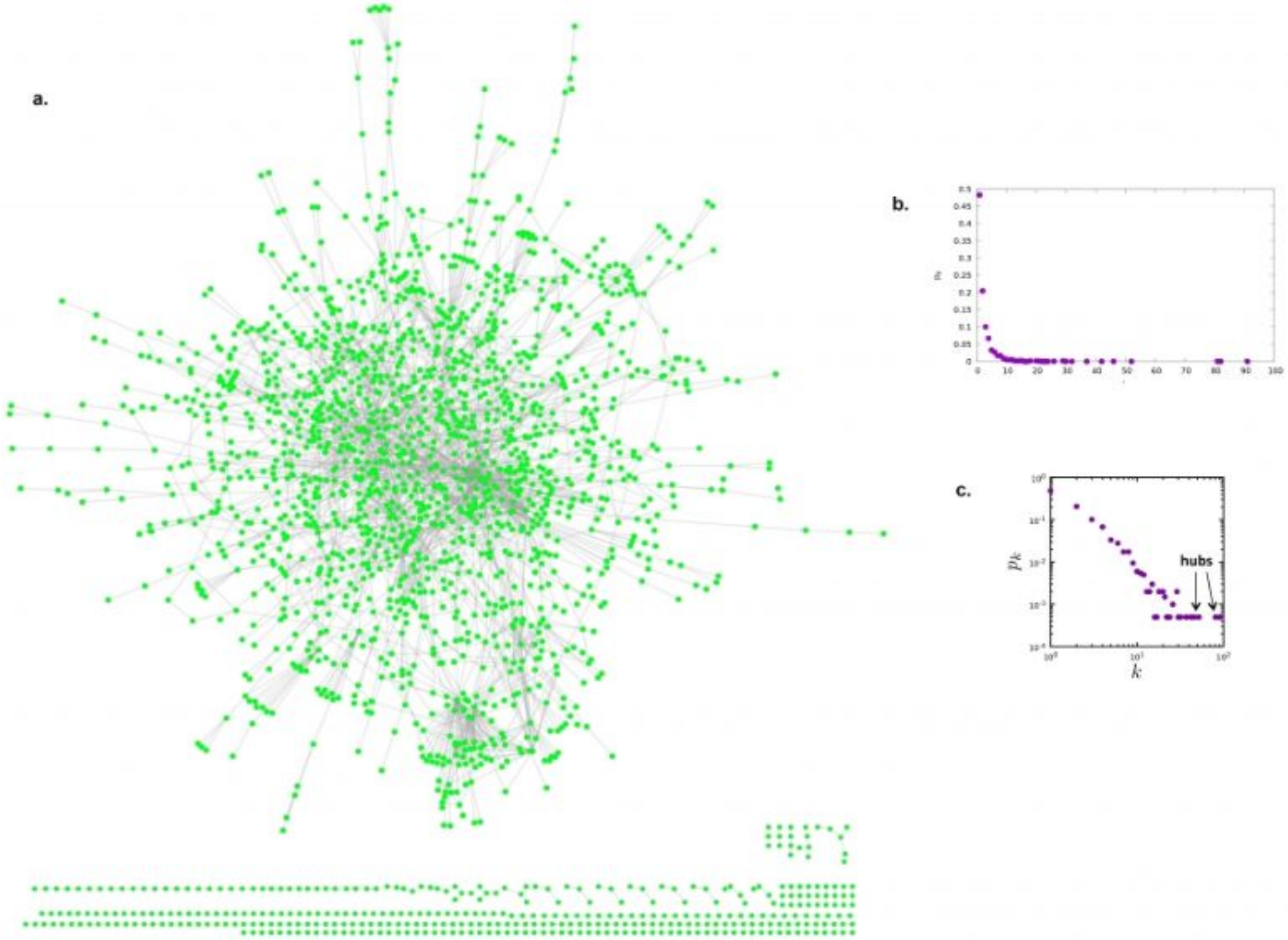
Degree Distribution plays a key role in Graph and Network Theory. Most of the structural network properties require the calculation of p_k.

e.g Average Degree can be expressed as a function of Degree Distribution:

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

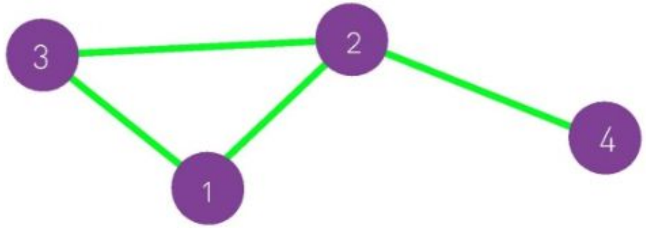


a.
4
2
1
3
b.
1
0.75
pk
0.5
0.25
0
0
1
2
3
4
k
c.
d.
1
0.75
pk
0.5
0.25
0
0
1
2
3
4
k



Barabasi (2016) Image 2.4

Adjacency Matrix



To keep track of the links we can provide a complete edge list:
{(1,2);(1,3);(2,3);(2,4)}

We can translate it into an Adjacency Matrix, where network of N nodes has N rows and N columns:

- $A_{ij} = 1$ if there is a connection between *i* and *j*,
- $A_{ij} = 0$ no connection between *i* and *j*,
- in an undirected network $A_{ij} = A_{ji}$.

The degree ki can be expressed by the sum of either the rows or the columns of the matrix:

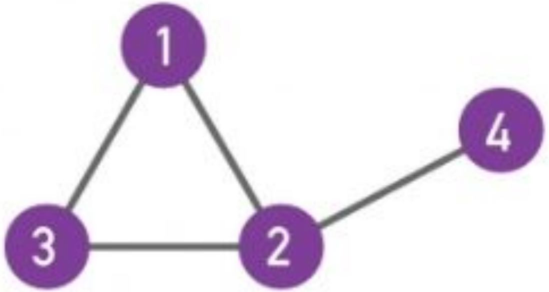
$$k_i = \sum_{i=1}^{N} A_{ij} = \sum_{j=1}^{N} A_{ij}$$

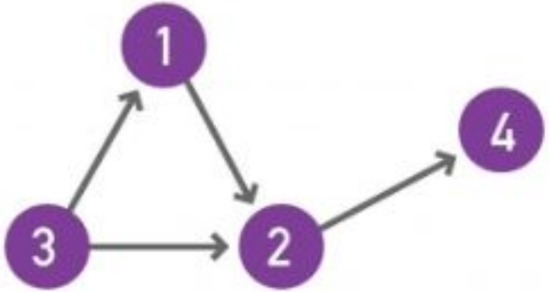
$$2L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} = \sum_{ij}^{N} A_{ij}$$

Adjacency Matrix

$$A_{ij} = \begin{matrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{matrix}$$

Adjacency Matrix





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

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

Adjacency Matrix

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

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

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

$$k_2^{\text{in}} = \sum_{j=1}^{4} A_{2j} = 2,\ k_2^{\text{out}} = \sum_{i=1}^{4} A_{i2} = 1$$

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

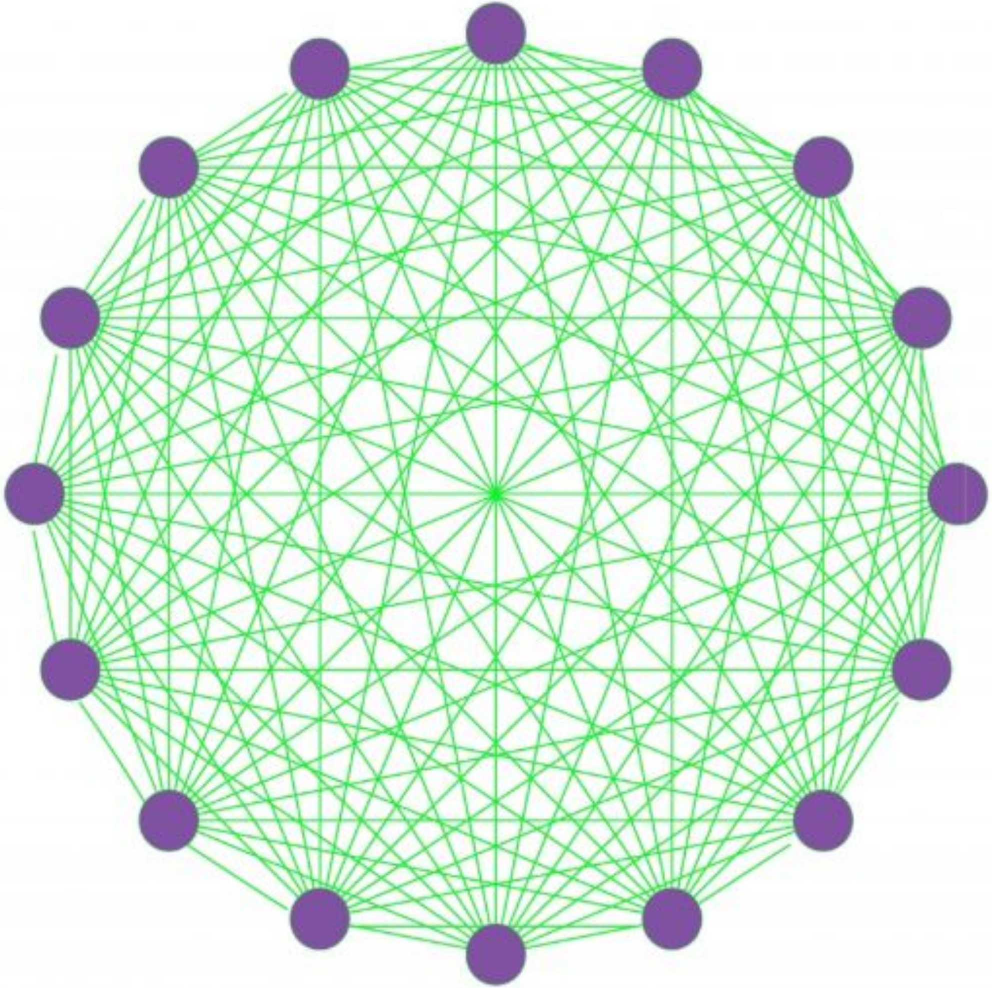
$$A_{ij} \neq A_{ji} \qquad A_{ii} = 0$$

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

$$L = \sum_{i,j=1}^{N} A_{ij}$$

$$\langle k \rangle = \frac{2L}{N}$$

$$\langle k^{\text{in}} \rangle = \langle k^{\text{out}} \rangle = \frac{L}{N}$$



The total number of links varies widely.

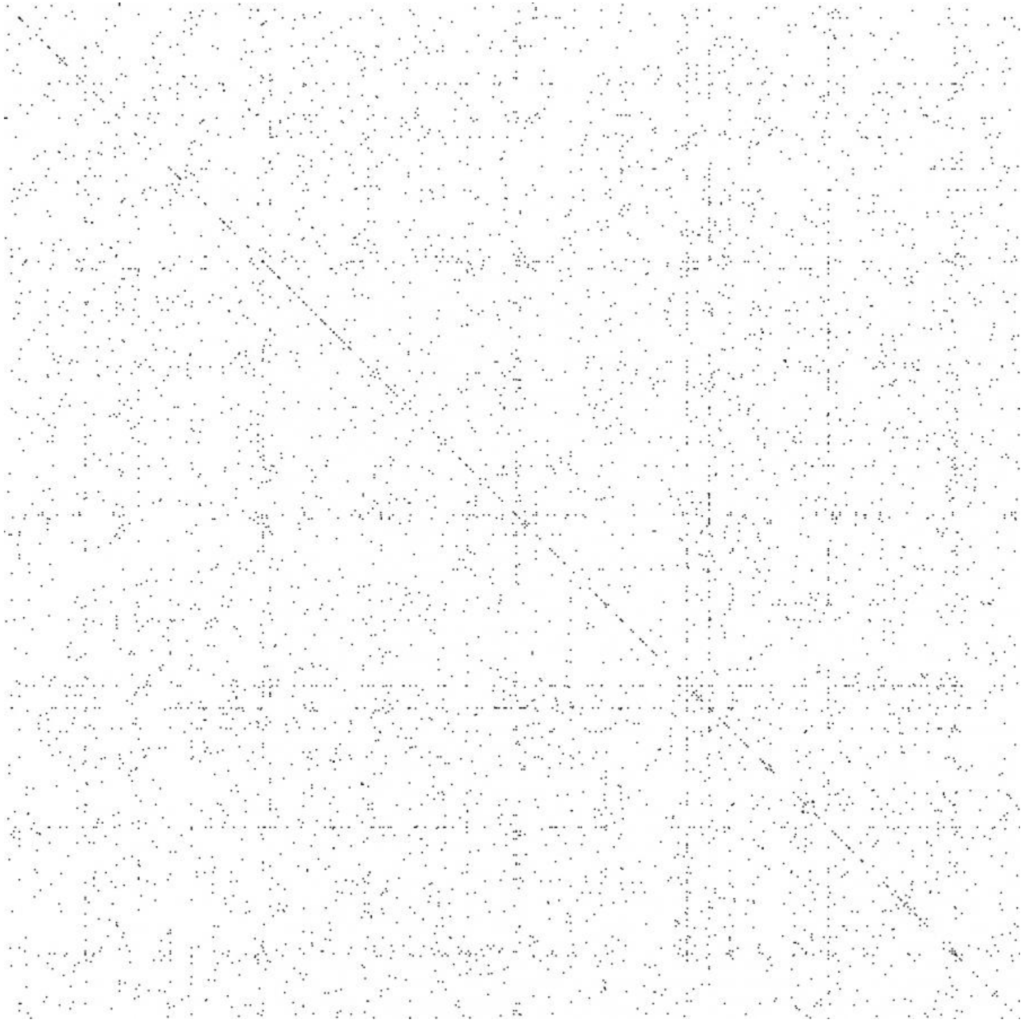
$$L_{min} = 0$$

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

Real networks are sparse!

$L << L_{\text{max}}$.

Barabasi (2016) Image 2.6



The total number of links varies widely.

$$L_{min} = 0$$

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

Real networks are sparse!

$L << L_{\text{max}}$.

Overwhelming fraction of elements are zero!

Barabasi (2016) Image 2.7

Weighted Networks

Unweighted Networks

$$A_{ij} = 1$$

A_{ij} = 1 if there connection between i and j

- is there a phone call?
- any export/import between i and j?

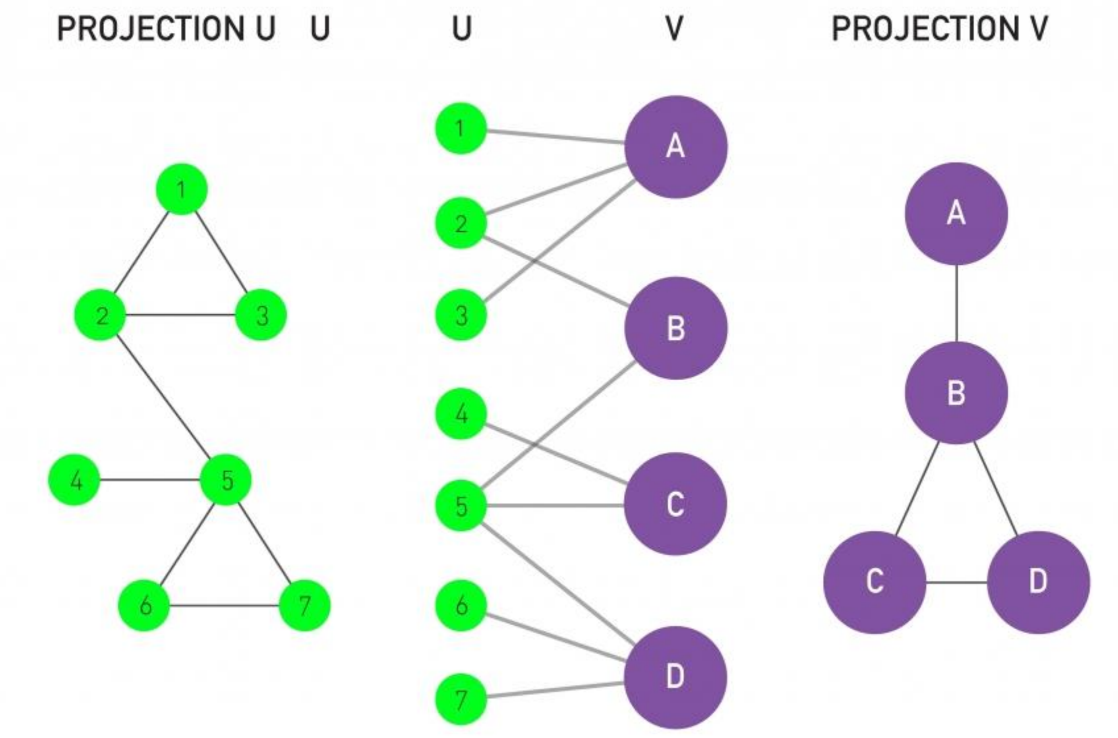
Weighted Networks

$$A_{ij} = w_{ij}$$

$A_{ij} = w_{ij}$ the extent of connection between i and j

- how many phone calls have happened?
- the extent of export/import between countries?

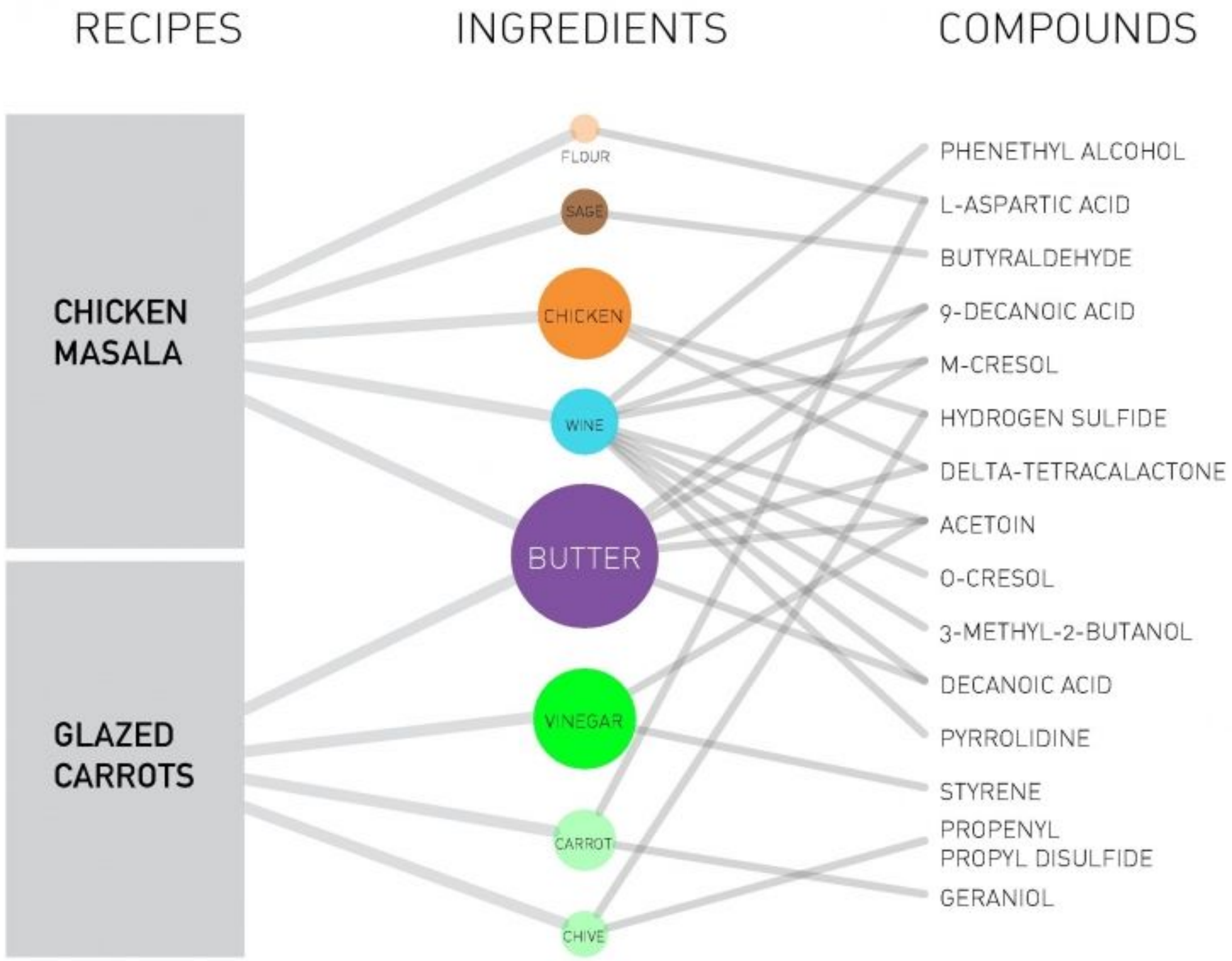
Bipartite Networks



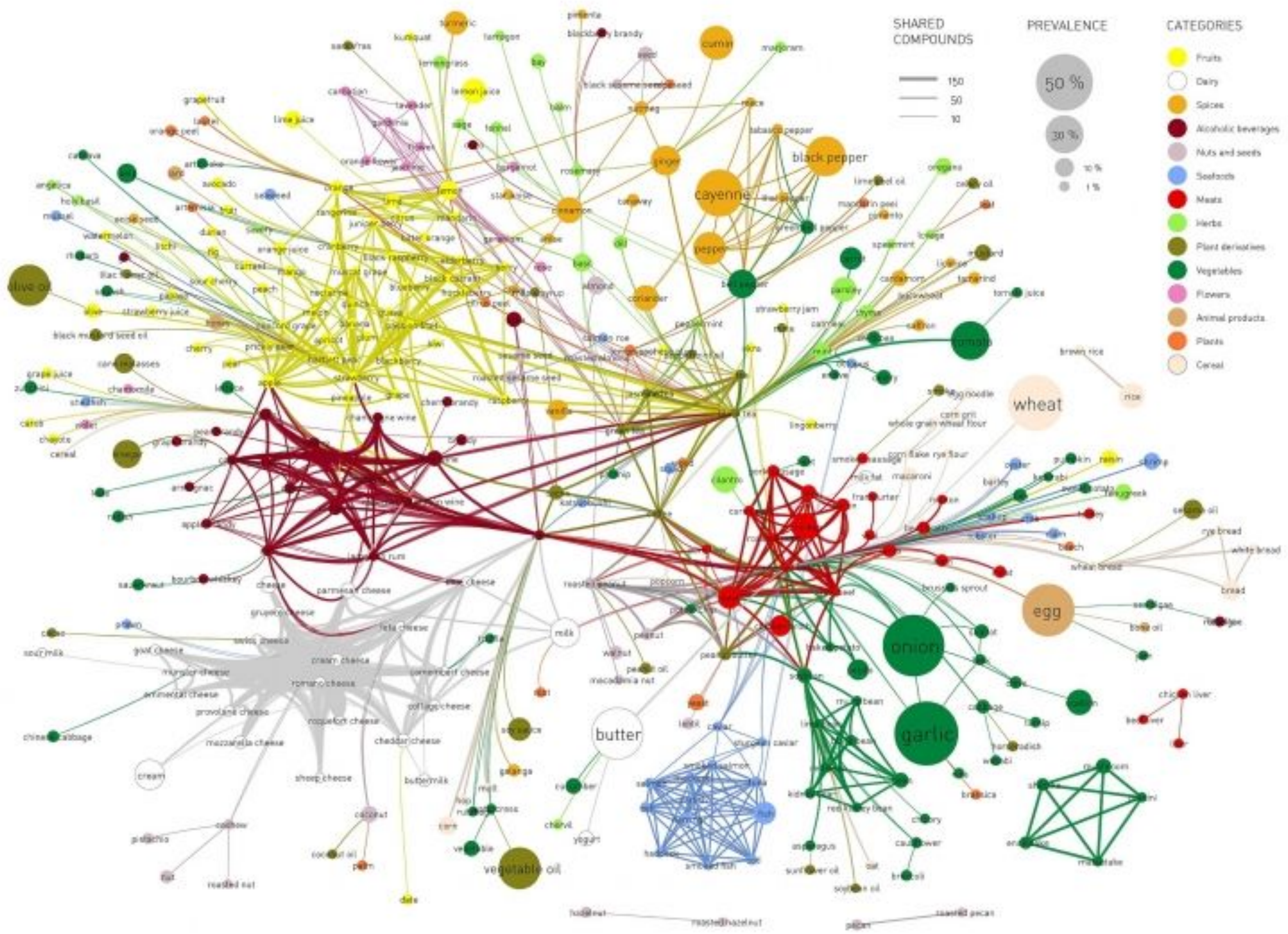
Nodes can be divided into two disjoint sets U and V

movie network:
U(actors)
V(movies)

co-publication network:
U(inventors/researchers)
V(patents/publications)



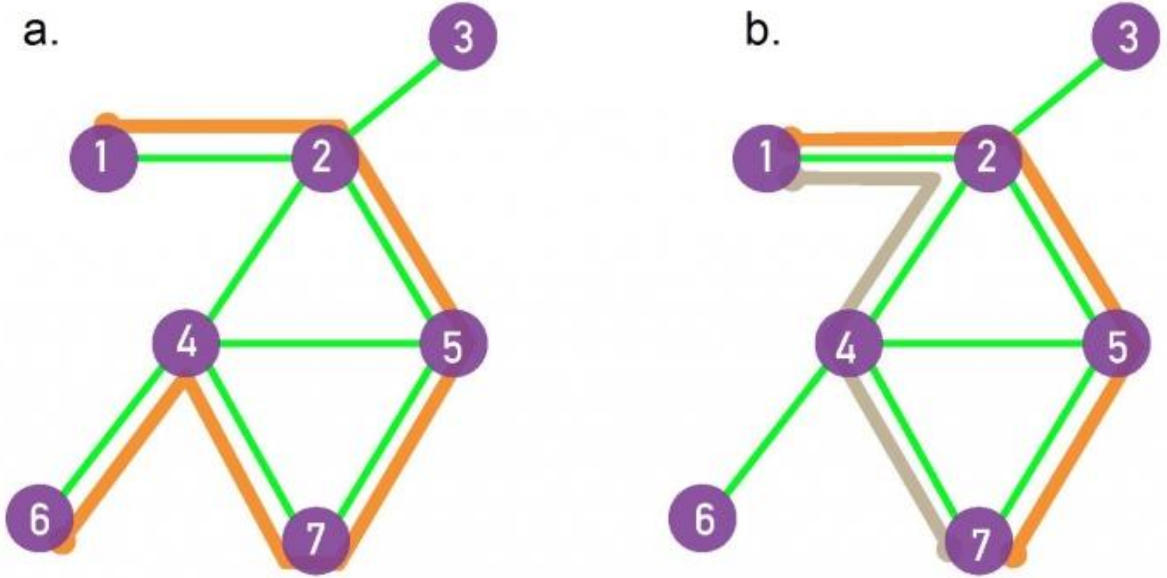
Barabasi (2016) Image 2.11



Barabasi (2016) Image 2.11

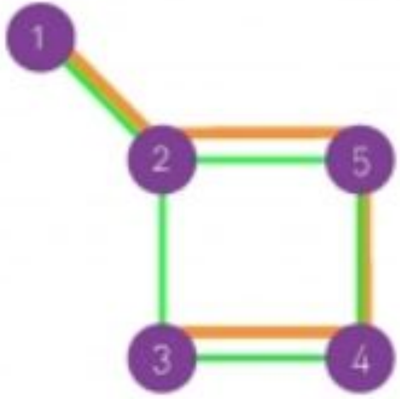
Paths and Distances

To determine the probability of interaction between two components of a system we usually use the physical distance between the agents, e.g. distance between two atoms or distance between two planets.



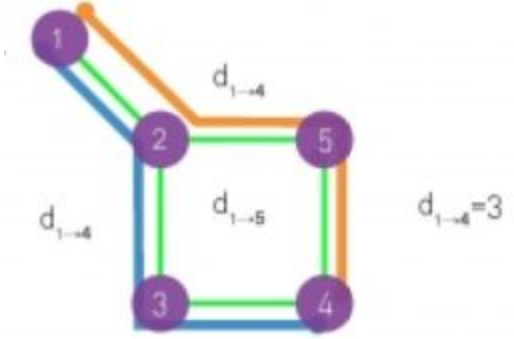
Paths and Distances

- **Path** is a sequence of connected nodes

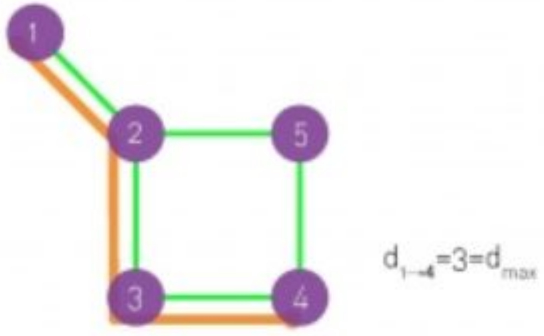


Paths and Distances

- **Path** is a sequence of connected nodes
- **Shortest path** (*d*) is the shortest distance between two nodes.



Paths and Distances



- **Path** is a sequence of connected nodes
- **Shortest path** (d) is the shortest distance between two nodes.
- **Diameter** (d_{max}) is the shortest path between the two furthest points of the graph.
  $d_{max}(1,4)=3$

Paths and Distances

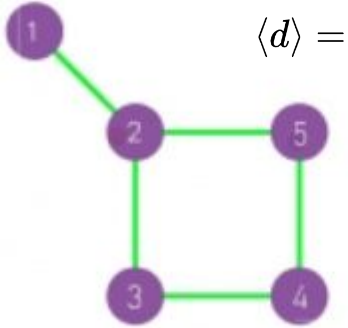
- **Path** is a sequence of connected nodes

- **Shortest path** (*d*) is the shortest distance between two nodes.

- **Diameter** (d_{max}) is the shortest path between the two furthest points of the graph.
  $d_{max}(1,4)=3$

- **Average Path Length** (*<d>*) is the average of the shortest path between all pair of nodes.
  <d>=1.6

$$\langle d \rangle = \frac{1}{N(N-1)} \sum_{i,j=1;i\neq j} d_{ij}$$



$$\langle d \rangle = [d_{1\to2}+d_{1\to3}+d_{1\to4}+d_{1\to5}+ \\ +d_{2\to3}+d_{2\to4}+d_{2\to5}+ \\ +d_{3\to4}+d_{3\to5}+ \\ +d_{4\to5}]/10=1.6$$

Paths and Distances

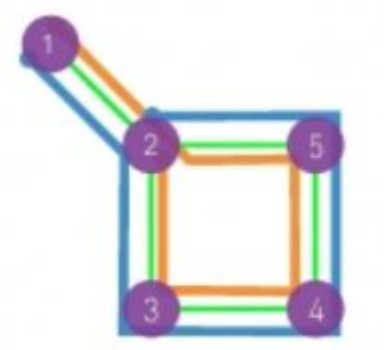
- **Path** is a sequence of connected nodes

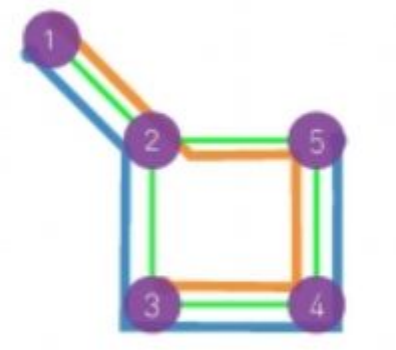
- **Shortest path** (d) is the shortest distance between two nodes.

- **Diameter** (d_{max}) is the shortest path between the two furthest points of the graph.
  $d_{max}(1,4)=3$

- **Average Path Length** ($<d>$) is the average of the shortest path between all pair of nodes.
  $<d>=1.6$

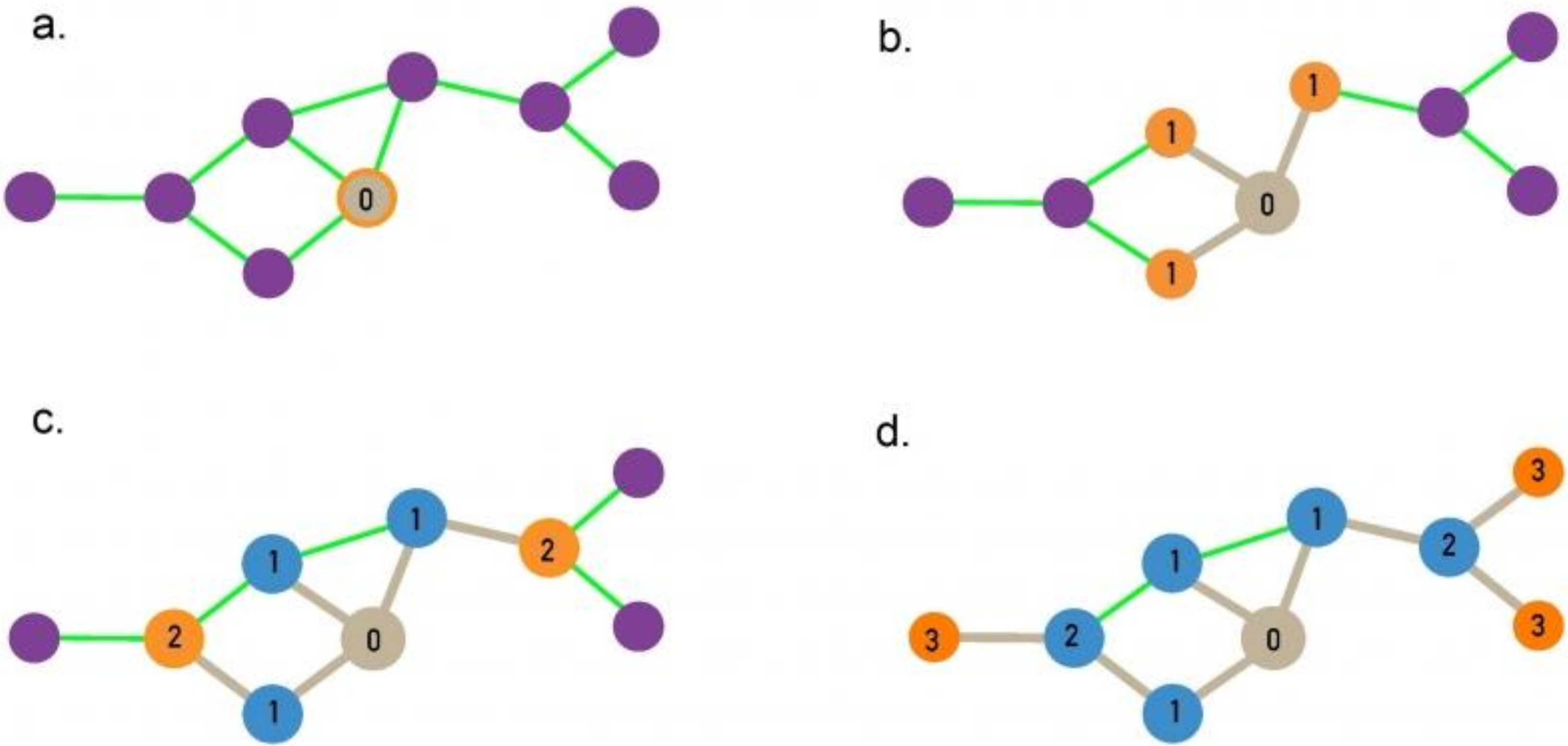
- **Eulerian Path** and **Hamiltonian Path**

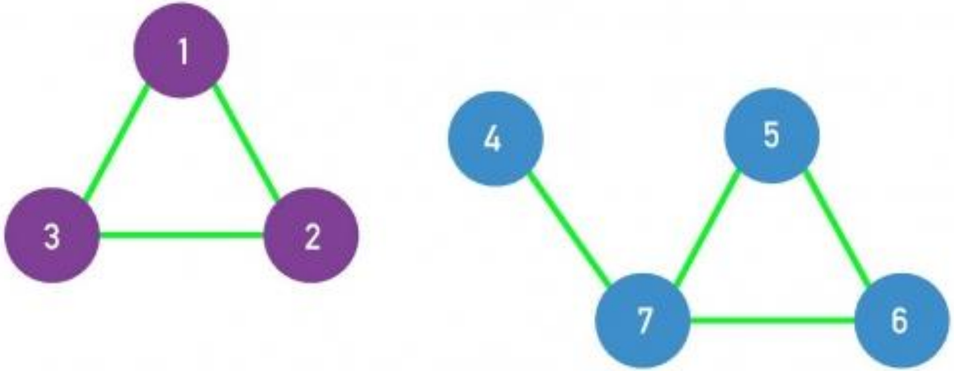




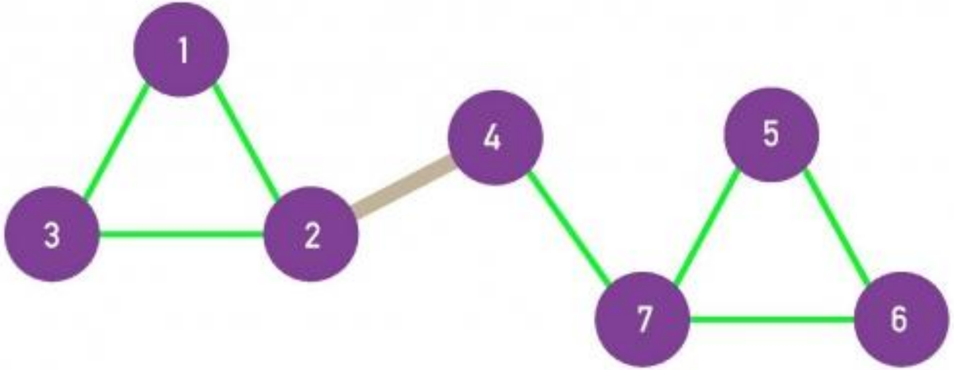
Paths and Distances

Breadth-First Search Algorithm





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



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

Clustering

Clustering coefficient shows the degree to which the neighbors of a given node connected to each other.

$$C_i = \frac{2L}{k_i(k_i-1)}$$

,where L is the number of links between k_i neighbors of node i.

$C_i = 0$ if there is no connection between the neighbors of i.
$C_i = 1$ if the neighbors of i form a complete graph.
$C_i = 0.5$ implies that there is a 50% chance that two neighbors of i are connected to each other.

Average Clustering Coefficient captures the degree of clustering of the whole network, or in other words it shows the probability that two neighbors of a randomly selected node connected to each other:

$$\langle C \rangle = \frac{1}{N}\sum_{i=1}^{N} C_i$$

Clustering

Clustering Coefficient

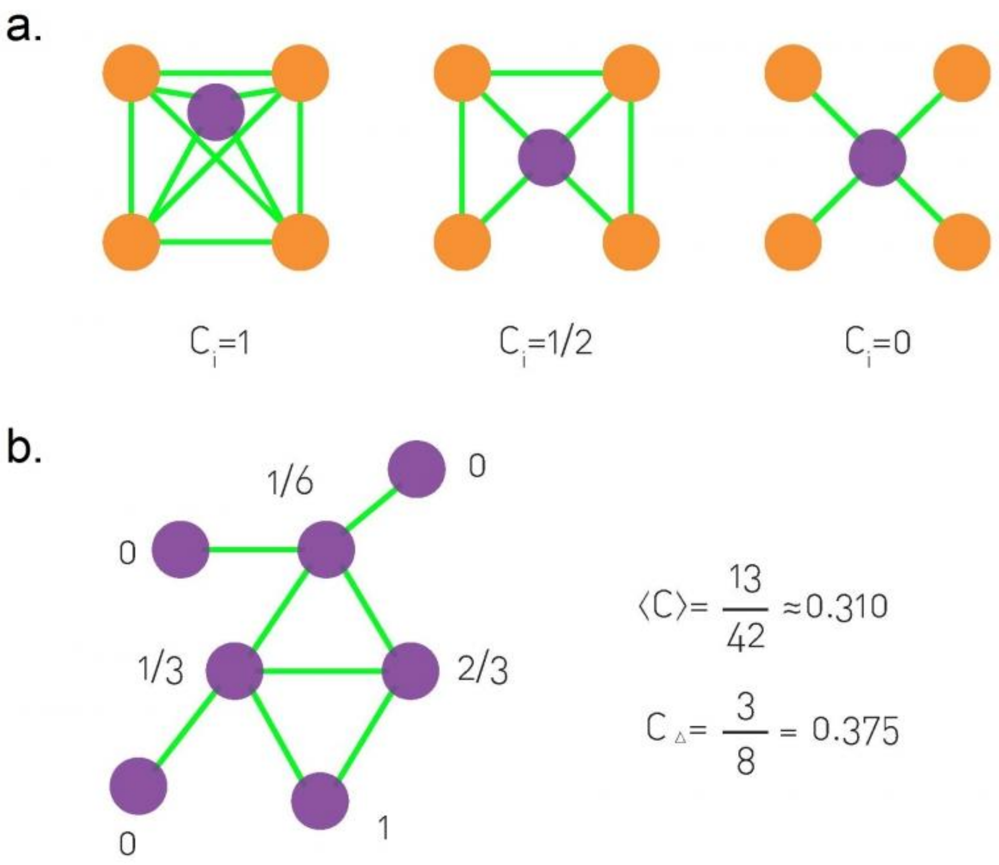
$$C_i = \frac{2L}{k_i(k_i - 1)}$$

Average Clustering Coefficient

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$

Global Clustering Coefficient

$$C_\Delta = \frac{3 \times \text{ Number of } \Delta}{\text{Number of Connected Triples}}$$



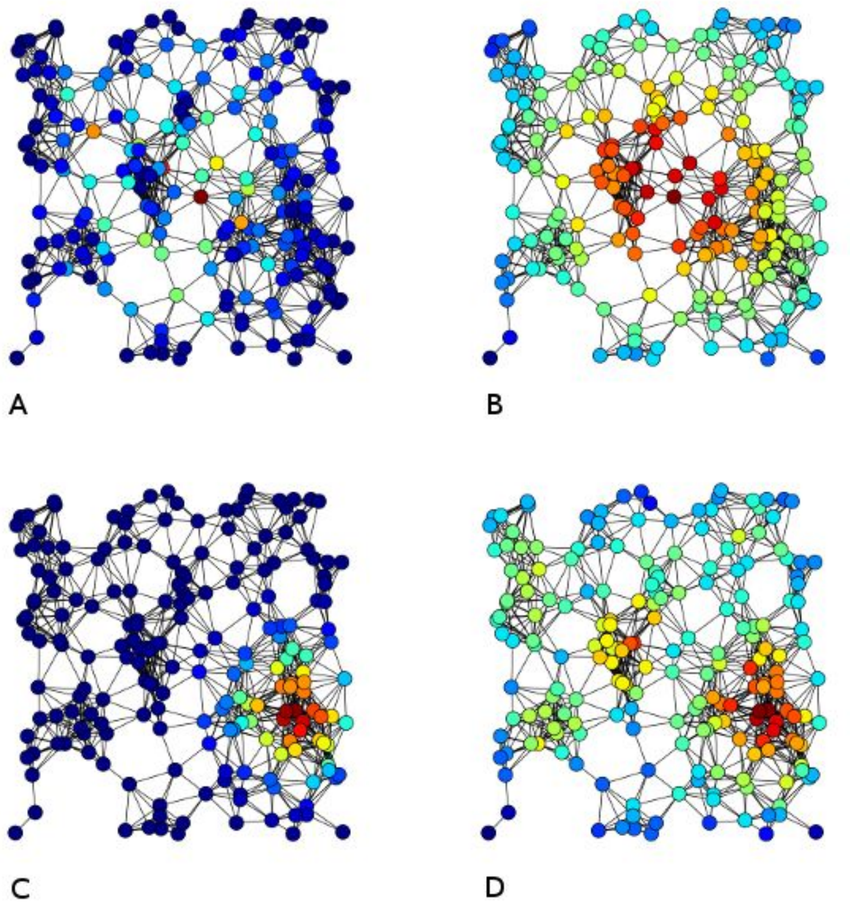
Node Centrality Measures

A) Betweenes captures the node's role as a bridge between groups of nodes. Betweenes in about how critical a bode is to the networks functioning as a bridge point between other parts of the network. Linton Freeman (1977,1991)

$$\text{Betweenness}_i = \sum_{i \neq j \neq = k} \frac{p_{jk}(i)}{p_{jk}}$$

,where p_{jk} is the total number of shortest path between j and k.

To normalize: $\frac{n-1(n-2)}{2}$



Node Centrality Measures

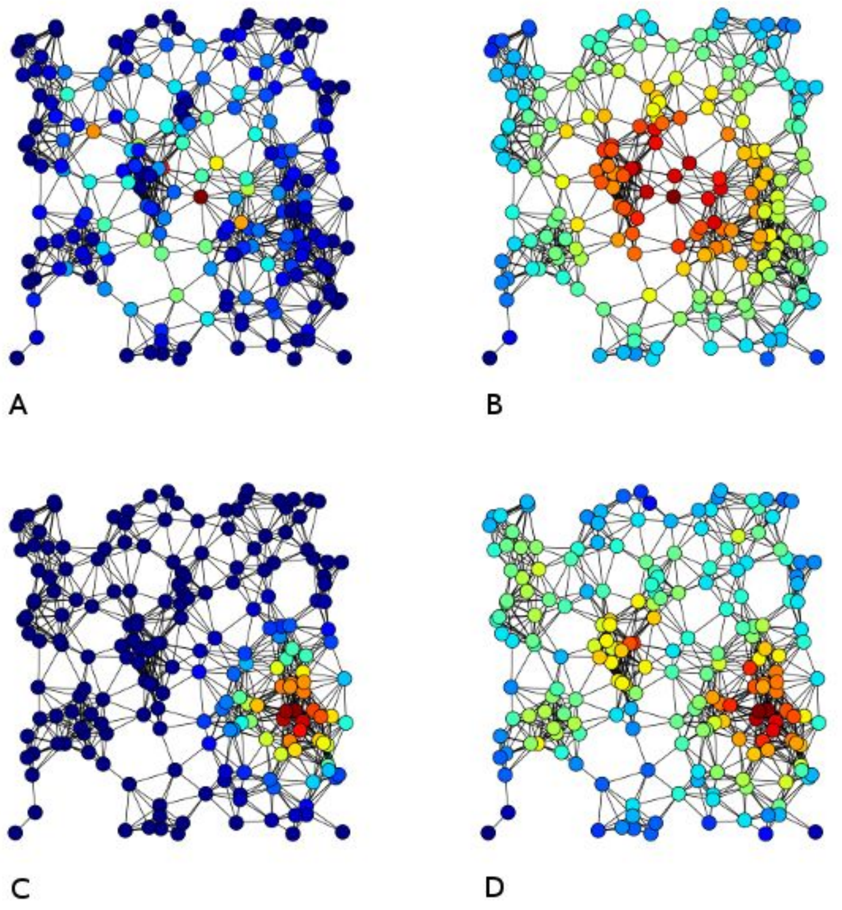
B) Closeness captures how close a node is to any other randomly selected node in the network. That is how quickly can the mode reach other nodes in the network.

It can be shown as the reciprocal of farness:

$$\text{Closeness}_i = \frac{1}{\sum_{i \neq j} d_{ij}}$$

The most central node has the lowest of total distance. If there is no path between *i* and *j*, then the path length replaced by the total number of nodes.

To normalize: $\frac{1}{(n-1)}$



Node Centrality Measures

C) Eigenvector centrality how important a node in a network based on the significance of the nodes that given node is connected to. So instead of looking on the number of connections it represents the value of the connections.

$$E_i = \sum_{j \in N(i)} \sum_j A_{ij} E_j$$

E_i depends not just the number of connections $|N(i)|$ but also on the quality of the connections E_j.
This measure used by web-search engines trying to rank the relative importance of websites.