The Institute of Mathematical Sciences, Chennai

Fourth Summer Program on Dynamics of Complex Systems 2019 (DCS2019)

Networks

Tutorial 1:

Algorithms for network metrics

Sitabhra Sinha

The Institute of Mathematical Sciences, Chennai

How do you represent a network in a computer ?

How you store the information about the vertices and edges can affect speed of computation & memory usage !

- Vertices represented by unique labels, viz., 1, 2, 3, …, N
- To represent edges, one can use different possible representations, e.g.,
  - Adjacency matrix
    Simple (stored as 2-dimensional array of integers) and fast in finding/removing edges (O(1)) but for **sparse graphs** is inefficient in terms of use of memory and takes O(N) operations for neighbour enumeration
  - Adjacency list [Most popular data storage format]
    List containing labels of other vertices to which each vertex is connected: economical in terms of memory usage and takes O(L/N) operations for neighbour enumeration in **sparse graphs** but also for finding/removing edges
  - Adjacency tree
    Like adjacency list, but list of neighbors of each vertex is stored as a binary tree

Local Properties: Node degree

Degree k_i of a node i in a network is its number of connections

For an undirected network $k_i = \sum^{N}_{j=1} A_{ij}$

The total number of connections in the network $L = (1/2) \sum^{N}_{i=1} k_i$

as the two ends of every connection contribute to the degree of two nodes

The mean degree of a node in an undirected network $\langle k \rangle = 2L/N = (1/N) \sum^{N}_{i=1} k_i$

Regular networks: all nodes have the same degree

The maximum possible connections in a network with N nodes is

${}^{N}C_2 = (1/2)N(N-1)$

$\Rightarrow$The connection density (connectance) is

$\rho = L / ({}^{N}C_2)=2L/N(N-1)= \langle k \rangle/(N-1)$

The density of any network lies in the range [0,1] (e.g., $\rho = 1 \Rightarrow$ Clique)

Dense network: A network whose density ρ tends to a constant > 0 as $N \to \infty$

Sparse network: A network whose density $\rho \to 0$ as $N \to \infty$ (e.g., for networks whose average degree tends to a constant as no. of nodes increase)

Constant degree or constant connectance ?

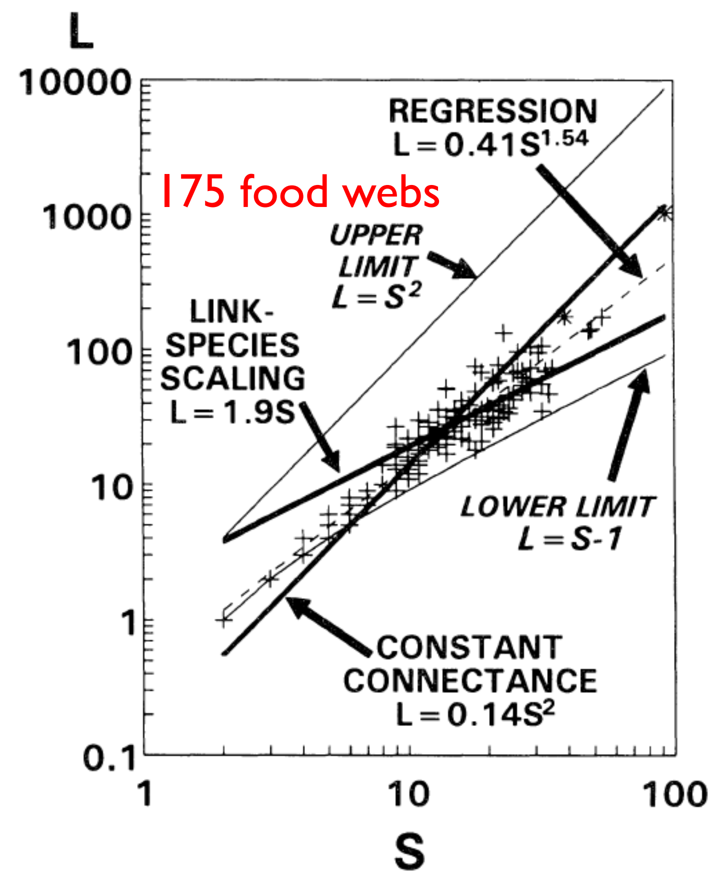
Question: How does the number of links L in a network increase with the number of nodes N in a class of networks is increased ?

Example:

- ❑ Trophic species: groups of organisms having identical sets of predators & prey
- ❑ Trophic links: feeding interactions directed from prey to predators

Link-species scaling law: On average the number of links (L) per species (S) in a food web is constant, i.e., species have constant avg degree independent of S

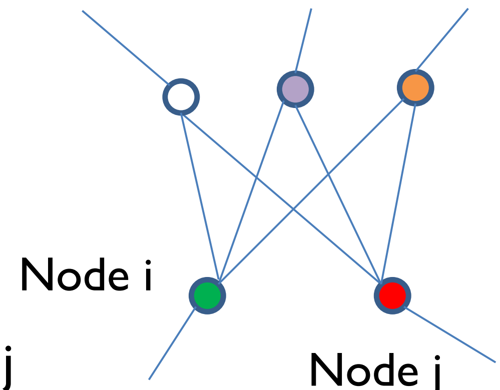
Constant-connectance hypothesis: The number of links (L) increases approximately as the square of functionally distinct species (S) in a web



N Martinez, *American Naturalist* **139** (1992) 1208

Similarity: Structural and Regular Equivalence

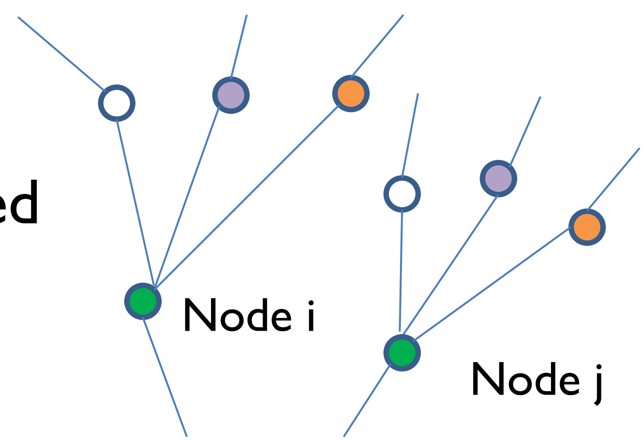
How does a web-site say
"If you like this (Q), you will probably like these (X,Y, Z) ?"
i.e., how is it possible to say which node (or nodes) is most similar to another given node in a specific network ?



- Two nodes in a network are **structurally equivalent** if they share many of the same network neighbors

Example: number of common neighbors n_{ij} of two nodes i,j
Cosine similarity = $\sum_k A_{ik} A_{kj} / [\sqrt{(\sum_k A_{ik}^2)}\sqrt{(\sum_k A_{kj}^2)}] = n_{ij} / \sqrt{(k_i\, k_j)}$

- Two **regularly equivalent** nodes need not necessarily share neighbors but when they have neighborhoods with similar properties



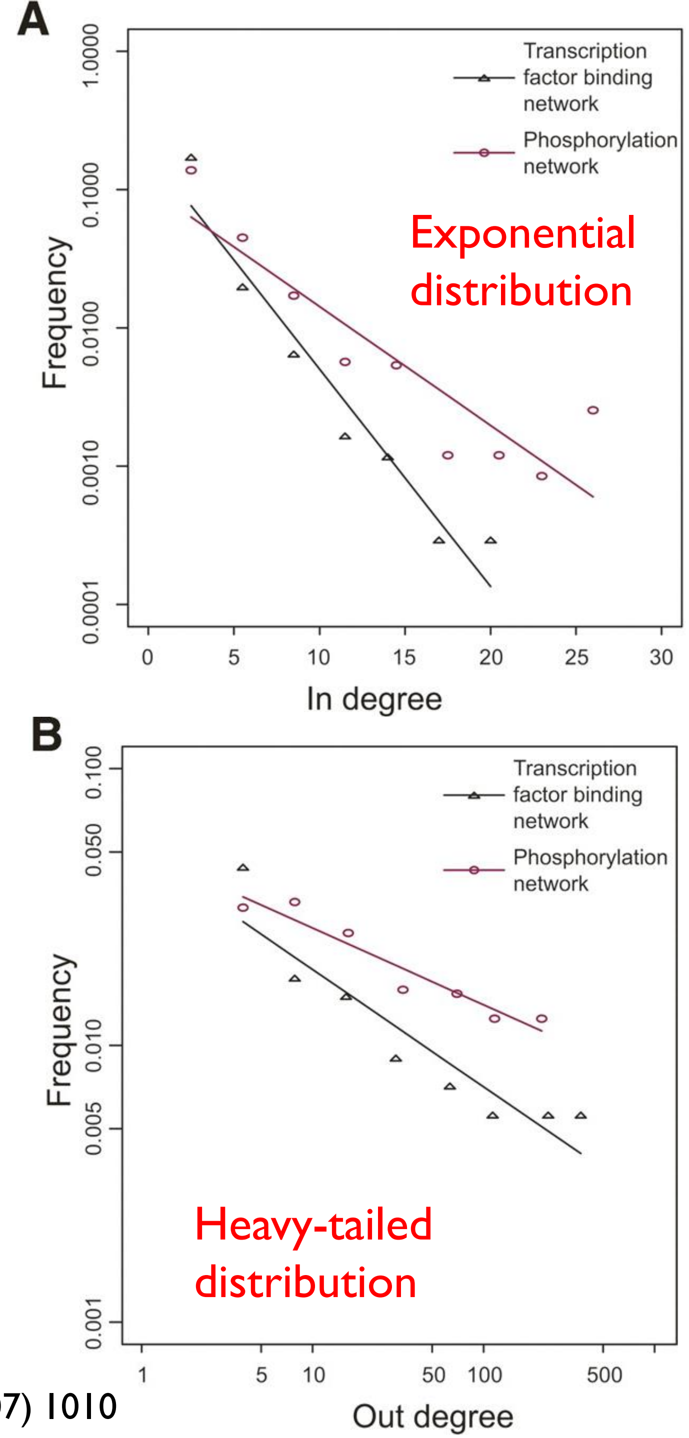
Degree in directed networks

In a directed network each node is associated with two types of degree
In-degree: number of incoming connections.
Out-degree: number of outgoing connections.

$A_{ij} = 1$ means there is connection from j to i
In-degree of node i : $k_{i\,(in)} = \sum^{N}_{j=1} A_{ij}$ and
Out-degree of node j: $k_{j\,(out)} = \sum^{N}_{i=1} A_{ij}$

Total number of connections in the network
$L = \sum^{N}_{i=1} k_{i(in)} = \sum^{N}_{j=1} k_{j(out)} = \sum_{i,j} A_{ij}$
as each incoming end of a link is paired with an outgoing end of a link
$\Rightarrow$ Mean in-degree $\langle k_{(in)} \rangle$ = Mean out-degree $\langle k_{(out)} \rangle = \langle k \rangle = L/N$

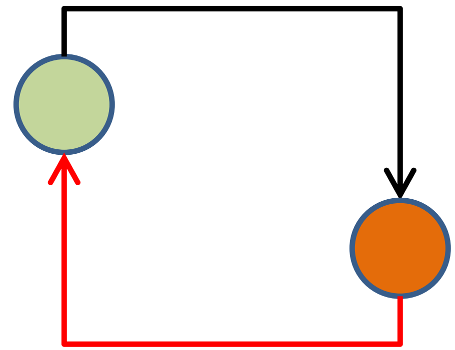
Question: In a network, do the high out-degree nodes also tend to have high in-degree ?
Distributions of in-degree and out-degree may have very different natures



X Zhu et al. *Genes Dev.* **21** (2007) 1010

Local properties of networks: Reciprocity

Just as we can ask if a node that sends out many links, also receive many connections from other nodes…
we can ask in directed networks that if node i sends a connection to j, whether node j also sends one to i



Example: think of "following" and "followers" of your social media account

Question: Are links between a pair of nodes reciprocated ?

The frequency of loops of length 2 is measured by reciprocity, i.e., the fraction of edges that are reciprocated $f_r = (1/L) \sum_{ij} A_{ij} A_{ji} = (1/L)\, \mathrm{Tr}\, A^2$

Alternatively, defined as correlation coefficient between corresponding entries of adjacency matrix

$f_r^{(GL)} = \sum_{i\neq j} (A_{ij} - \langle A \rangle)(A_{ji} - \langle A \rangle) \,/\, [\sum_{i\neq j} (A_{ij} - \langle A \rangle)^2]$

where $\langle A \rangle = \sum_{i\neq j} A_{ij} / N(N-1) = L/N(N-1)$ [Garlaschelli & Loffredo, PRL (2004)]

lies within -1 and +1 (>0 $\Rightarrow$ reciprocal, <0 $\Rightarrow$ anti-reciprocal)

If there are no reciprocal edges, $[f_r^{(GL)}]_{min} = -\langle A \rangle / [1 - \langle A \rangle]$

Dispersion of reciprocity among nodes measured by the standard deviation σ_f of $f_r^{(GL)}$ in terms of $f_r^{(GL)}(i,j)$ obtained when any link betn (i,j) is removed.

Calculating degree distribution

In adjacency list, information about neighbors for each vertex is maintained
To obtain degree for each node, we need simply count the number of entries in the neighbor set

For adjacency matrix, we need to sum together all the entries of i-th row or column to find the degree of the i-the node

Once the degree of all nodes $\{k_1, k_2, \ldots, k_N\}$ are known, create a histogram

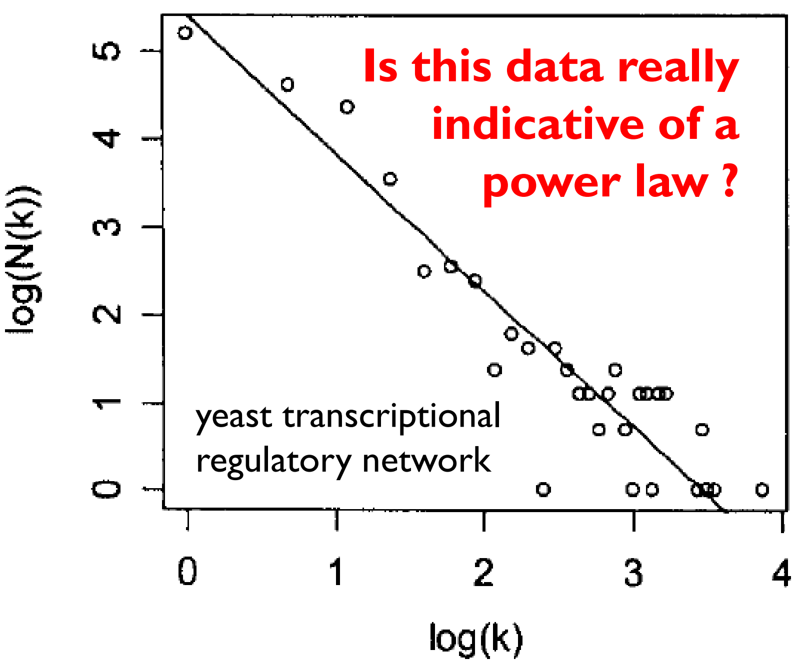
- Construct an array comprising k_{max} "bins" – each bin storing the number of vertices of a specific degree (up to the maximum degree).
- Set all array elements initially to zero.
- Run through each vertex in turn, find its degree q (say) and add 1 to the q-th bin.
- Once all N vertices have been gone through divide all array elements by N to obtain p_k.

Problem: For small bin widths, may look extremely non-uniform, but with larger bins we lose resolution
Solution: Construct complementary cumulative degree distribution $p_{k>K}$ by sorting the degrees in descending order, ranking them from 1 to N and plotting the rank divided by N as a function of the degree

Are real-world networks really scale-free ?

- Scale-free networks characterized by long-tailed degree distribution (power laws) proposed as unifying concept for complex systems.
- But many of these reports of scale-free networks are possibly just a result of bad statistics (a combination of extremely limited data and faulty analysis) !
- Almost any distribution seen over a small enough range in a double logarithmic scale would appear linear – and wrongly interpreted as power law

Guelzim *et al.*, 2002



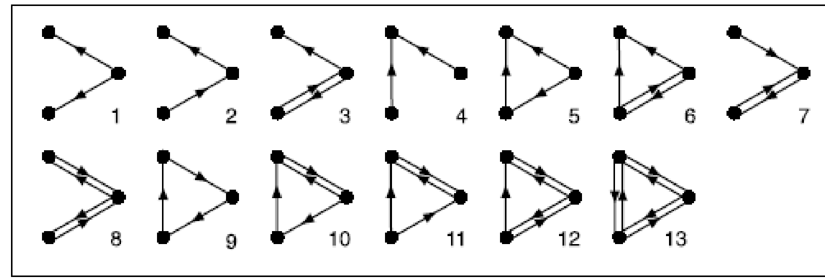
- To establish power laws from finite data one has to use unbiased techniques such as *maximum likelihood estimation.*
- Recent rigorous re-analysis of many of the data sets used by earlier studies that claimed power-law degree distributions for real-world networks have shown little evidence for scale-free nature!

  (E.g., R Khanin & E Wit, *J Comp Biol* **13** (2006) 810)

Local properties of networks: Motifs

The network may be built out of putting together recurring subnetworks of interactions

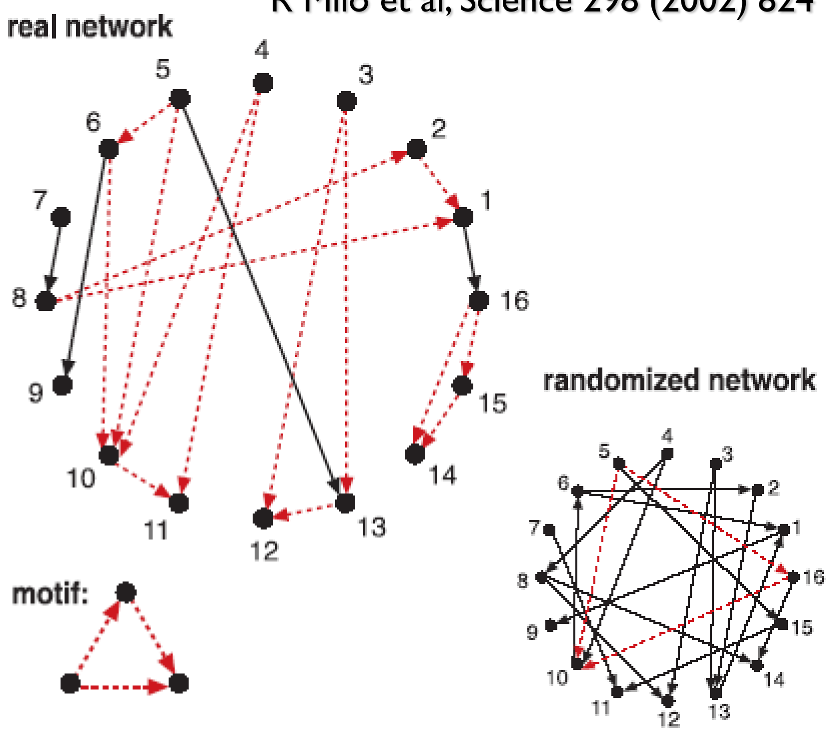
All connected three-node subgraphs

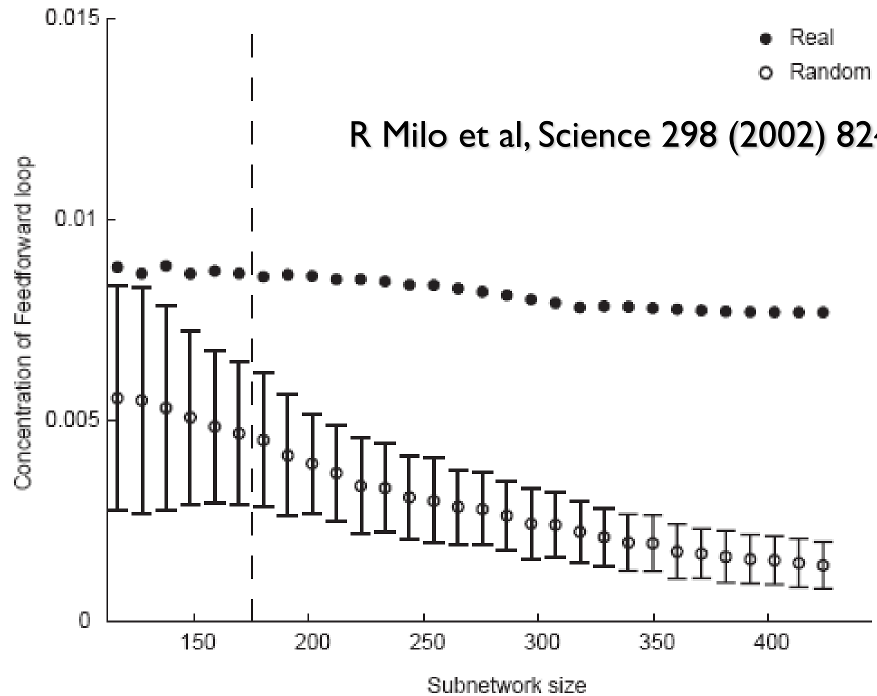


What are motifs ?

Subnetwork connection patterns that occur more frequently than expected in an equivalent random network

R Milo et al, Science 298 (2002) 824





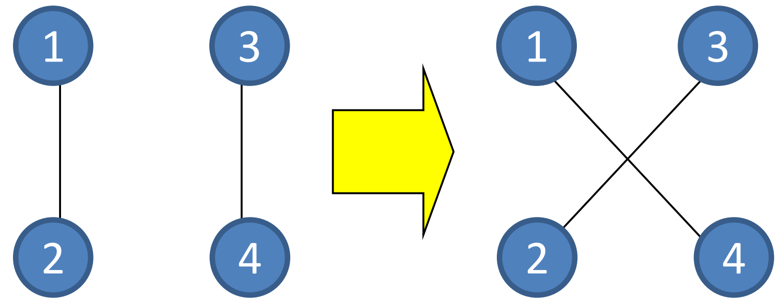
R Milo et al, Science 298 (2002) 824

Degree-preserved randomization of a network

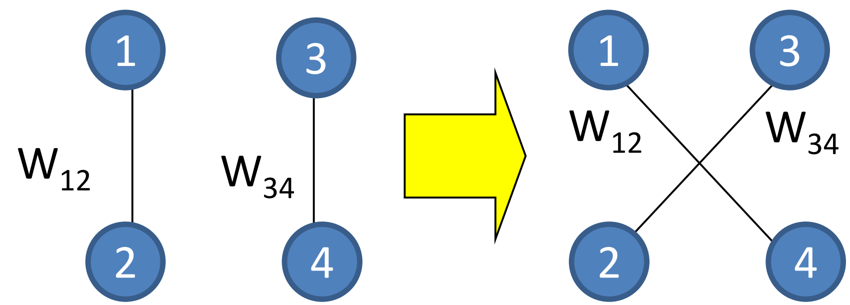
Randomization of unweighted network

maintaining degree of each node unchanged from its value in the empirical network by link swapping technique

Maslov, S. & Sneppen, K. Science 296,910–913 (2002).



Randomization of weighted network



Method I

•Degree preserved randomization of links

• link weights remaining with swapped connections

Advantage: preserves weight distribution

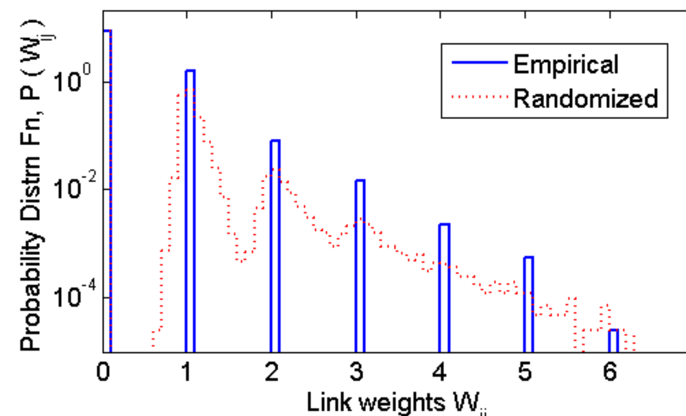
Drawback: does not preserve node strength

$s_i = \Sigma_j W_{ij}$

Method 2 *Bhattachraya et al, J Stat Mech (2008) P02002*

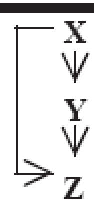
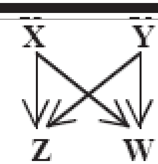
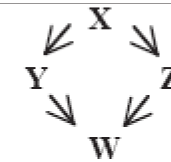
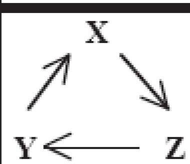
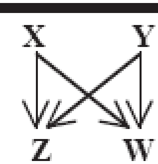
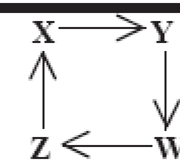
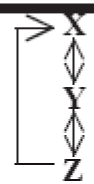
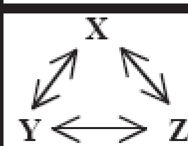
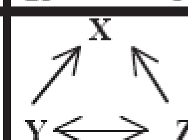
<u>Strength preserved</u> randomization of links by adjusting weights according to $w_{ij} \rightarrow w_{ij} + \delta_i (w_{ij}/\Sigma_j w_{ij})$ where $\delta i = s_i - \Sigma_j w_{ij}$ in order to balance si and $\Sigma_j w_{ij}$

Drawback: does not preserve weight distribution



Network motifs found in Technological Networks

R Milo et al, Science 298 (2002) 824

Network	Nodes	Edges	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score	N_{real}	$N_{rand} \pm SD$	Z score
Electronic circuits (forward logic chips)					**Feed-forward loop**			**Bi-fan**			**Bi-parallel**
s15850	10,383	14,240	424	2 ± 2	285	1040	1 ± 1	1200	480	2 ± 1	335
s38584	20,717	34,204	413	10 ± 3	120	1739	6 ± 2	800	711	9 ± 2	320
s38417	23,843	33,661	612	3 ± 2	400	2404	1 ± 1	2550	531	2 ± 2	340
s9234	5,844	8,197	211	2 ± 1	140	754	1 ± 1	1050	209	1 ± 1	200
s13207	8,651	11,831	403	2 ± 1	225	4445	1 ± 1	4950	264	2 ± 1	200
Electronic circuits (digital fractional multipliers)					**Three-node feedback loop**			**Bi-fan**			**Four-node feedback loop**
s208	122	189	10	1 ± 1	9	4	1 ± 1	3.8	5	1 ± 1	5
s420	252	399	20	1 ± 1	18	10	1 ± 1	10	11	1 ± 1	11
s838‡	512	819	40	1 ± 1	38	22	1 ± 1	20	23	1 ± 1	25
World Wide Web					**Feedback with two mutual dyads**			**Fully connected triad**			**Uplinked mutual dyad**
nd.edu§	325,729	1.46e6	1.1e5	2e3 ± 1e2	800	6.8e6	5e4±4e2	15,000	1.2e6	1e4 ± 2e2	5000

Global Properties: Calculating clustering

The local clustering coefficient of a node i is

$$C_i = \frac{\text{(number of pairs of neighbors of } i \text{ that are connected)}}{\text{(number of pairs of neighbors of } i\text{)}}$$

The denominator is just $\frac{1}{2} k_i (k_i - 1)$, trivial to obtain once degree of node i is known

To calculate the numerator

- ❑ go through every pair of distinct neighbors (p,q) of vertex i (with p<q)
- ❑ For each pair we determine whether an edge exists between them
- ❑ count up the number of such edges.

The overall clustering coefficient of a network is
$$C = \frac{3 \text{ (number of triangles)}}{\text{(number of connected triples)}}$$

The denominator is $\frac{1}{2} \sum_i k_i (k_i - 1)$, trivial to obtain once degree of node i is known

To calculate the numerator

- ❑ consider for every vertex i (=1, 2, …, N) each pair of neighbors (p, q) with p< q
- ❑ find whether they are connected by an edge
- ❑ add up the total number of such edges over all vertices

Structural Balance

A basic characterization of relationships between mutual acquaintances

Consider 3 individuals: Om, Pradeep and Xena. If

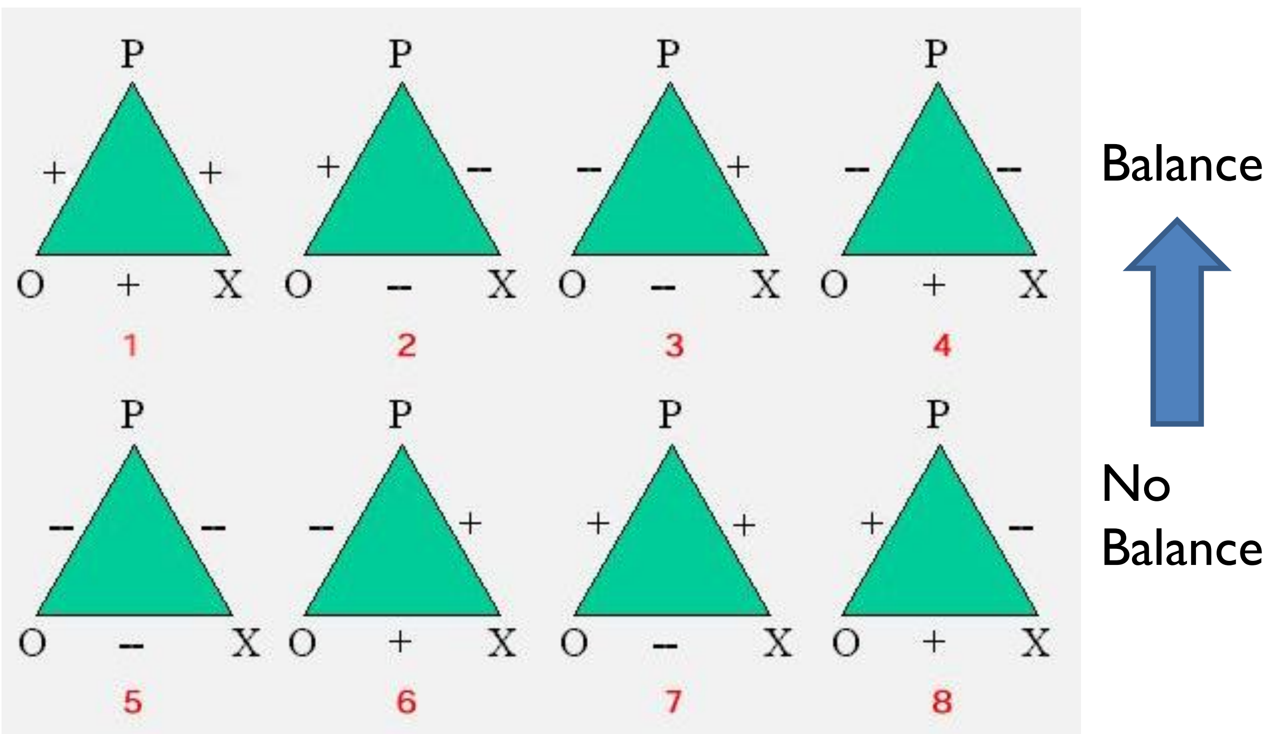
(1) Om and Xena are friends ⇒ OX: +ve

(2) Pradeep and Xena are friends ⇒ PX: +ve

(3) Om and Pradeep are enemies ⇒ OP: –ve

Tension

Fritz Heider (1896-1988)



Relationship triangles containing exactly 2 friendships are prone to transition to triangles with either 1 or 3 friendships ⇒ friend of my enemy is my enemy…

A single friendship may appear in a relationship triangle that initially had none ⇒ enemy of my enemy is my friend

F Heider (1946) Attitudes and cognitive organization. *J Psychol* **21**:107–112.

Structural Balance from triads to networks

Dorwin Cartwright (1915-2008)

Frank Harary (1921-2005)

Carwright & Harary (1956): Generalization of Heider's theory to network of N nodes

Psychol Rev **63**:277–293.

In a balanced network, *every* cycle (closed loop) is *balanced*, i.e., product of the signs of the links in the loop is +ve

A <u>complete</u> graph (a network where all pairs of nodes are connected) is balanced if each constituent triad is balanced

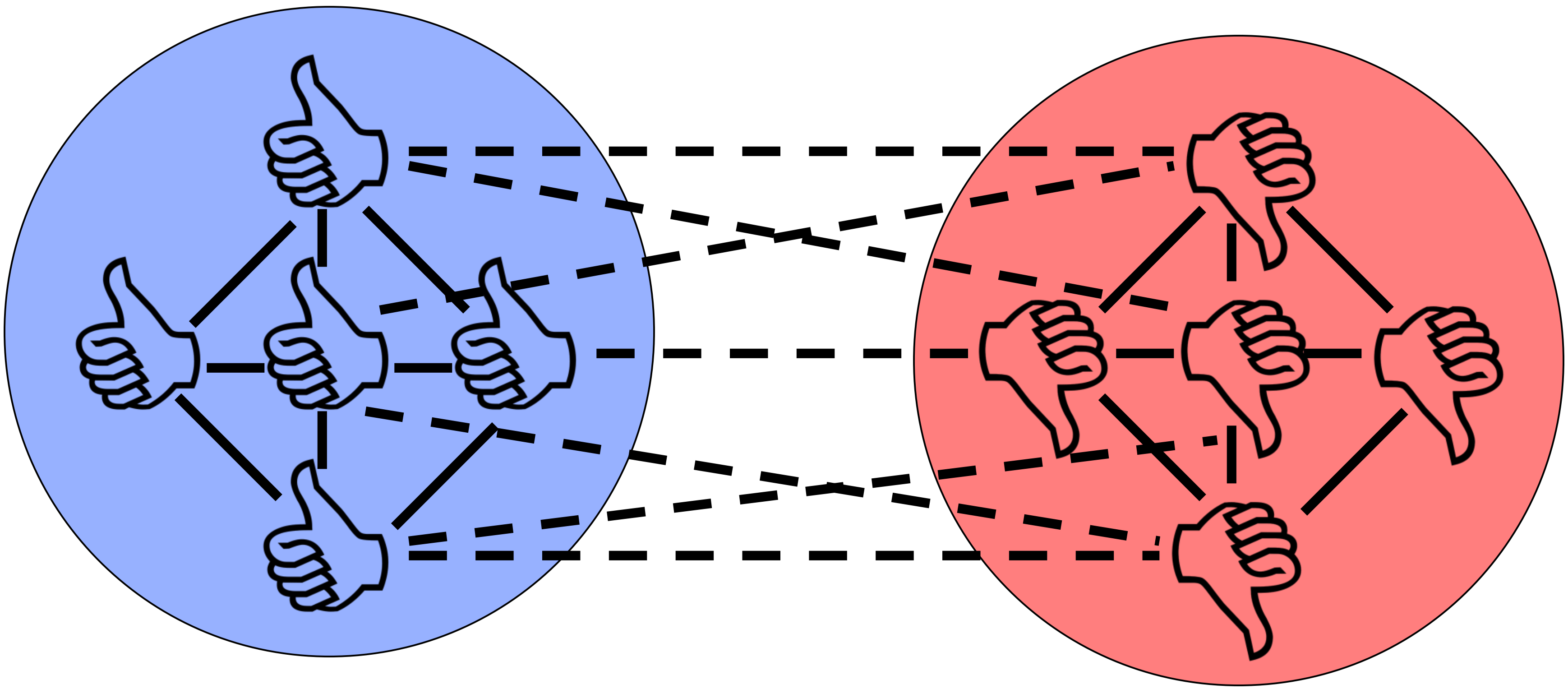
The local concept of balance results in non-trivial network structure

Any balanced network can be partitioned into two communities such that all edges inside each community are positive and all edges between nodes in opposite communities are negative (one of these communities may be empty)

Example

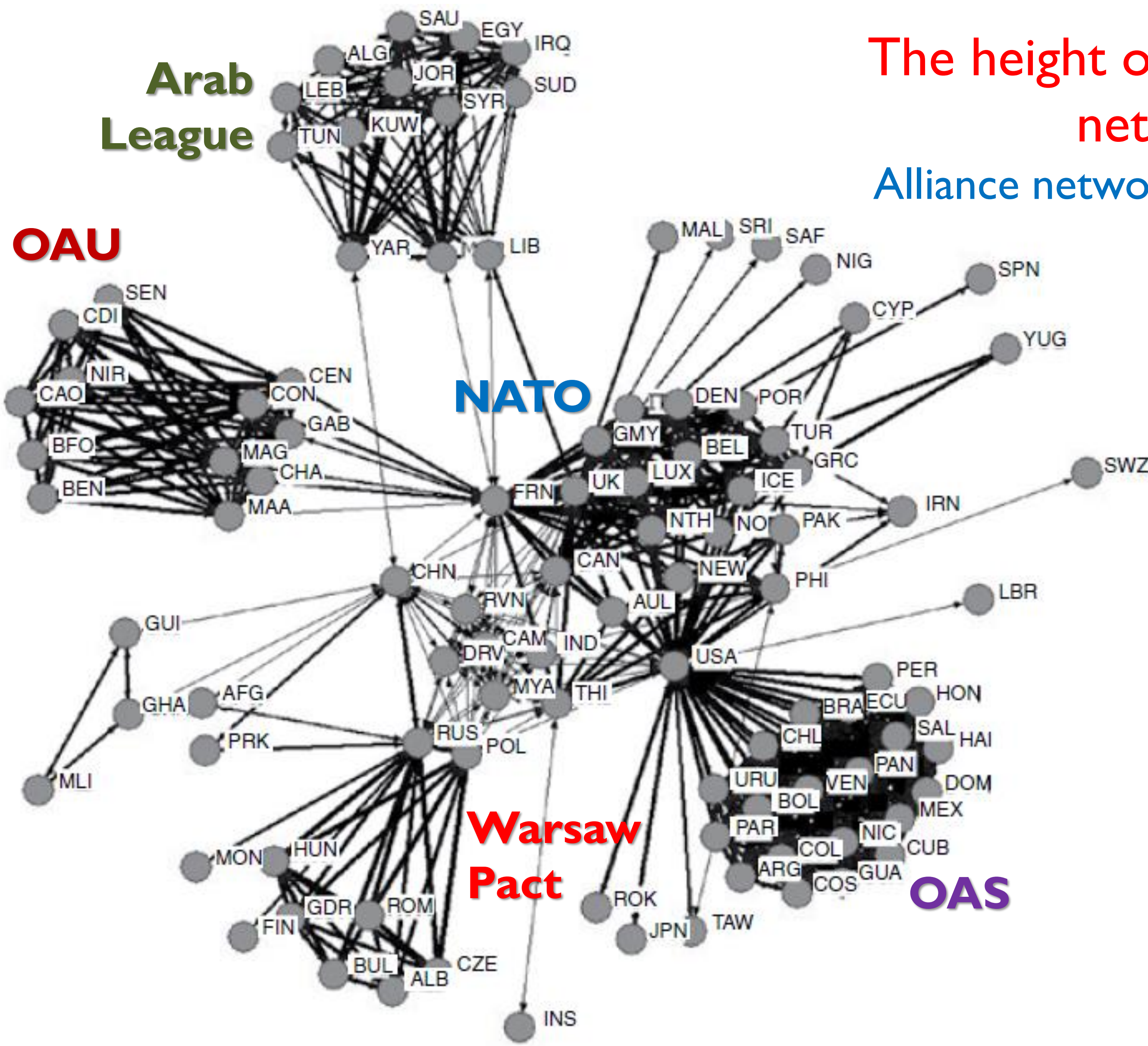
Consider two groups of individuals/organizations/nations

Within each group affiliative relations, between groups antagonistic relations



In absence of any external influence or noise, we expect the two communities to be unified and opposite in their response to any issue

Balance in Network of International Relations



The height of cold war from a network perspective

Alliance network of nations in 1962

As bipartite relations among countries that comprise major alliances change through events such as war, triads become unbalanced

$\Rightarrow$ creates tension

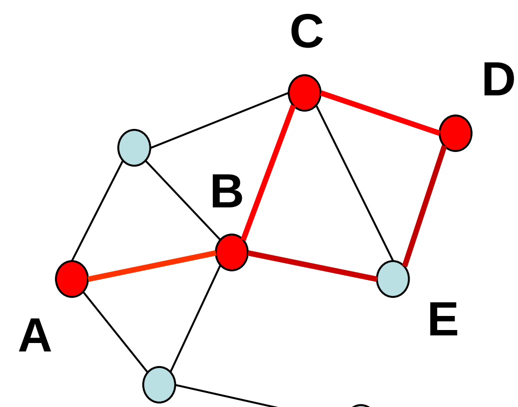
$\Rightarrow$ Reorganization into a balanced state involving new blocs and alliances

(Evolution to balance)

Z Maoz, *Networks of Nations* (Camb Univ Press, 2010)

Trekking through a network

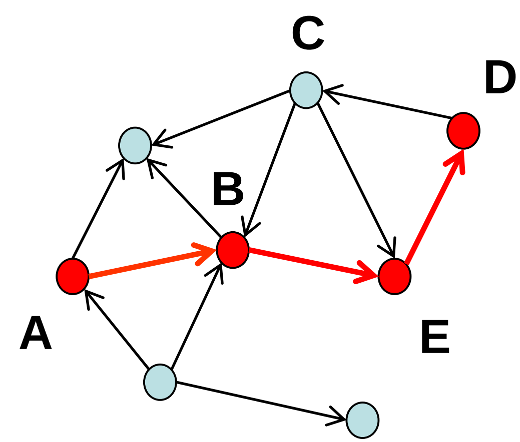
Undirected network



Network **path**: a sequence of nodes such that every consecutive pair is connected by a link in the network, i.e., a route across the nodes of a network traversing existing links

Network **path length**: number of links traversed ("hops") along a path to move from one node to another in the network

Directed network



Example (Undirected network): A path from A to D having length 3 is {A,B,C,D}
It is non-unique as another path of same length is {A,B,E,D}

Example (Directed network): Unique path from A to D having length 3 is {A,B,E,D}

Typically we focus on *self-avoiding paths* that do not intersect themselves, i.e., visit a node or link more than once (e.g., geodesics and Hamilton paths)

Number of paths of a given length

- For a network, element A_{ij} of the adjacency matrix $\mathbf{A}$ is 1 if there is node i and node j are connected by a link, and 0 otherwise.

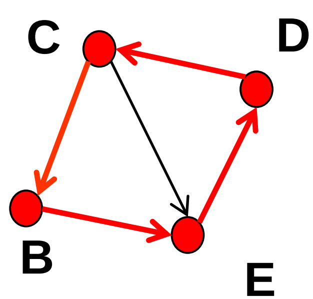
- The product $A_{ik} A_{kj}$ is 1 if there is a path of length 2 from j to i via k, and 0 otherwise.

- The total number of paths of length two from j to i, via any other vertex, is $N_{ij}^{(2)} = \sum_k A_{ik} A_{kj} = [\ \mathbf{A}^2\ ]_{ij}$

- In general, number of paths of length r is $N_{ij}^{(r)} = [\ \mathbf{A}^r\ ]_{ij}$

- If $i=j$ *(starting and ending points of a path are same)*, the path is a cycle or loop $\Rightarrow$ total number of cycles of length r in a network is

  $L_r = \sum_i A_{ik} A_{kj} = [\ \mathbf{A}^r\ ]_{ii} = \mathrm{Tr}\ \mathbf{A}^r$



it counts separately loops having same nodes but different starting points – i.e., {B,E,D,C,B} is considered different from {E,D,C,B,E}

- A *cycle* in a directed network has arrows on each of its links pointed in same way around the loop.

Acyclic networks

- Number of loops $L_r = \sum_i \lambda_i^r$ where λ are eigenvalues of **A**

- Thus, **acyclic networks** – i.e., networks having no cycles – will have a *nilpotent* adjacency matrix (all eigenvalues zero)

Trees

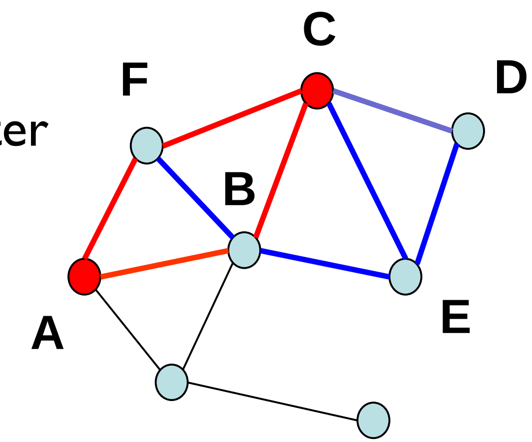
A *tree* is a connected, undirected network that contains no closed loops ("connected" $\Rightarrow$ every node is reachable from every other via some path through the network)

Represented with **root node** at base, with **branches** appearing from it and terminating in **leaf nodes**
Used for **hierarchical decomposition** of a network as in **dendrogram**

If a network contains two or more disconnected parts – all of whom are trees – it is called a *forest*

Exactly one path between any pair of nodes in a tree
Useful for defining lineage or phylogenetic distances

Geodesic or Shortest Path



A path between two vertices such that no path of a shorter length exists (necessarily self-avoiding)

Example (Undirected network): A geodesic of length 2 from A to C is {A,B,C}
It is non-unique as another path of same length is {A,F,C}

The length of a geodesic, called *geodesic distance* or *shortest path length*, is the shortest network distance between the nodes at the ends of the path

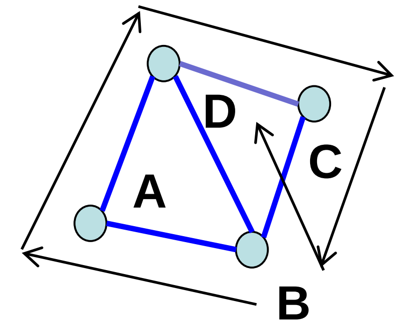
Defn. the smallest value of r such that $[\, \mathbf{A}^r \,]_{ij} > 0$

If a network has disconnected components, there may be no geodesic between members of one component and those of another $\Rightarrow$ infinite geodesic distance

Diameter of a network: length of the longest geodesic between any pair of nodes in the network for which a path actually exists.

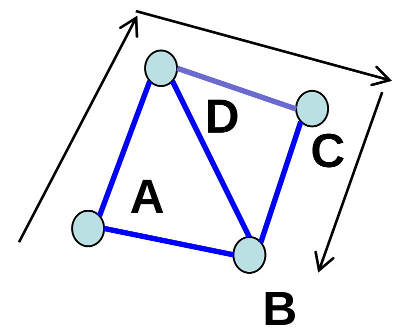
Eulerian path and Hamiltonian Path

An *Eulerian path:* path that traverses each link in a network exactly once.



A *Hamiltonian path*: a path that visits each node in a network exactly once. (by defn, self-avoiding)

A network can have one or many Eulerian or Hamiltonian paths, or none.



An Eulerian path need not be self-avoiding

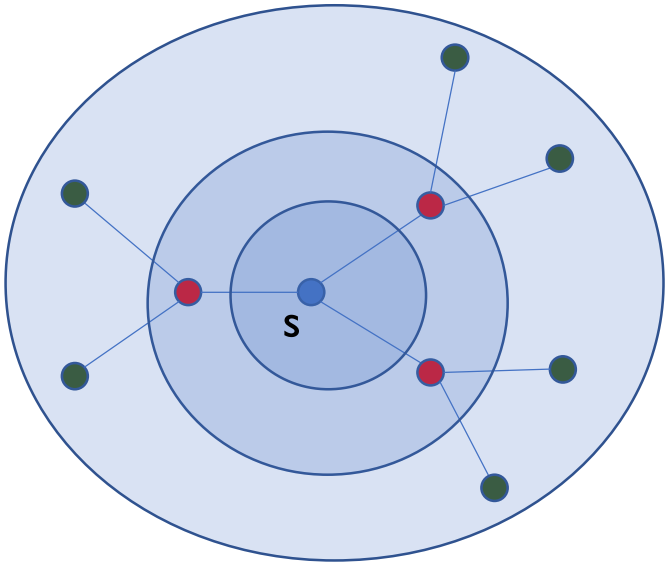
If there are any nodes of degree (number of links) > 2 in a network, an Eulerian path will have to visit those vertices more than once in order to traverse all their links.

Eulerian paths form the basis for solution of the Konigsberg Bridge problem by Euler

Global Properties: Calculating path length

Breadth-first search algorithm

finds the shortest (geodesic) distance from a single source vertex *s* to *every other* vertex in the network



- ❑ Start from vertex *s*
- ❑ Initially the distances to all other vertices are unknown
- ❑ Find all the neighbors of *s*
  By definition these have distance 1 from *s*.
- ❑ Then find all the neighbors of *those* vertices, excluding those already visited
  These vertices have distance 2 from s
- ❑ Then find their neighbors, excluding those already visited – these which have distance 3, and so on.
- ❑ On every iteration, the set of vertices visited grows by one step.
- ❑ Keep iterating until all nodes are visited

Hypergraphs

Typically, networks are defined by pairwise interactions between nodes. However, relations may be defined in terms of multilateral rather than just bilateral relations

Many biological processes/reactions involve several components participating together in an interaction, e.g.,

(i) substrate A is converted to product B on coming in contact with enzyme C

(ii) a protein complex that comprises more than 2 proteins

A generalized link connecting more than two nodes is a *hyperedge*

Hypergraph: A network with hyperedges

It is possible to represent a hypergraph by a **bipartite network** – a network consisting of two different types of nodes, with links occurring only between nodes of unlike type