Bangalore School on Statistical Physics XIV, Sept 2023

Statistical Mechanics of Complex Networks

Lecture 2: Metrics (or how to measure it?)

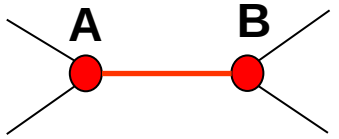
Sitabhra Sinha

The Institute of Mathematical Sciences, Chennai

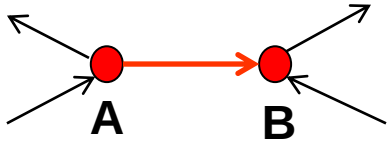
Networks: directed, weighted or signed

Adjacency matrices tell us about the presence or absence of links:
Is node A connected with node B ?

If the adjacency matrix is symmetric $\Rightarrow$ undirected network



Otherwise we have directed networks where a direction is associated with each link (A to B or B to A)



Many networks have links with heterogeneously distributed properties.

Connections in such systems can differ

- quantitatively by having a distribution of weights (that may for instance represent the strength of interaction) and/or
- qualitatively through the nature of their interactions, viz., positive (cooperative or activating) and negative (antagonistic or inhibitory)

How to characterize properties of a network ?

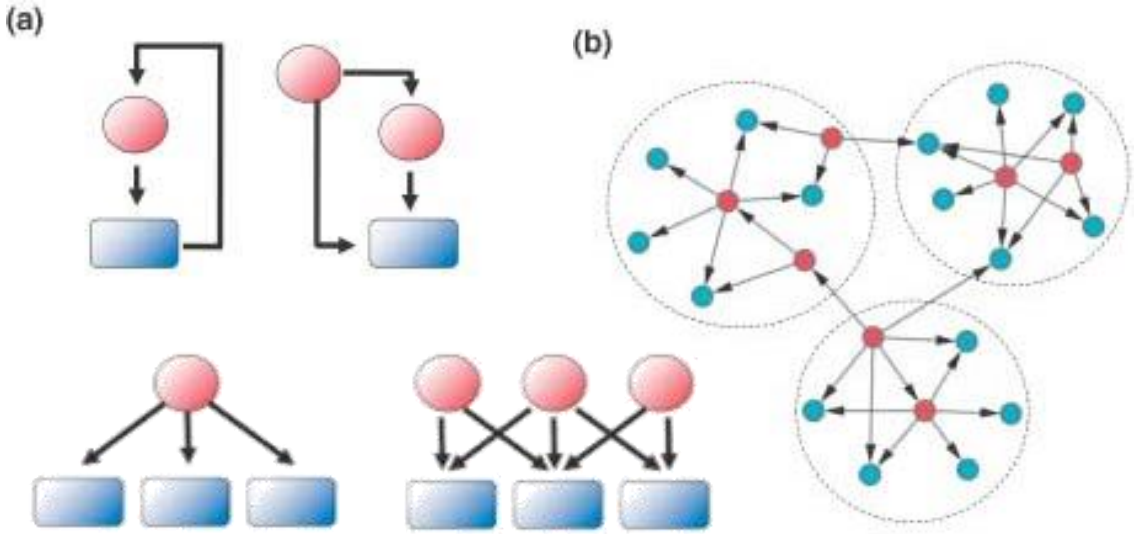
Local properties

Global properties

Micro

Meso

Macro



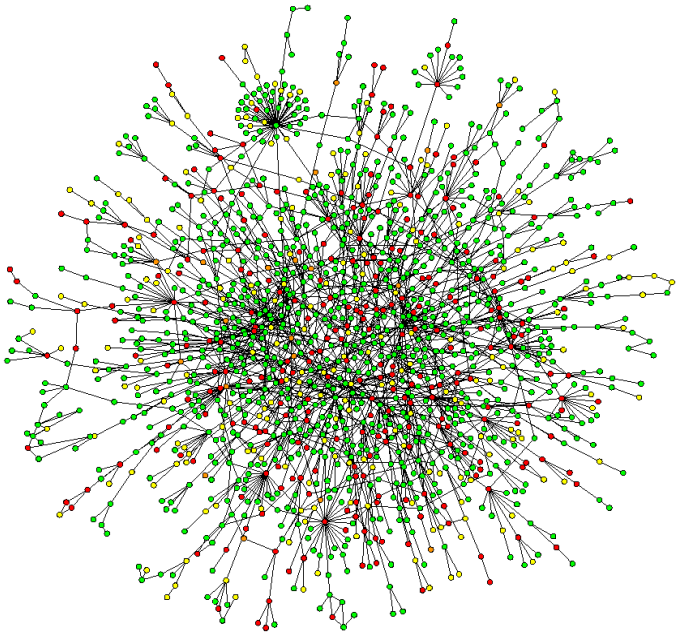


Image: Kim & Park, WIREs Syst Biol & Med, 2010

How cliquish is my neighbourhood ?

How fast can I travel to the farthest point of my network ?

Clustering

Path Length

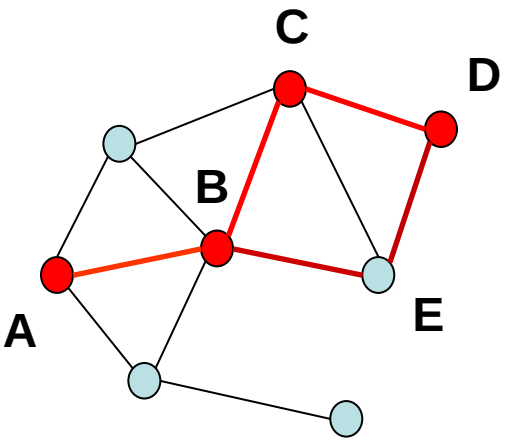
Network Metrics

"It becomes a science, only after you start measuring things quantitatively"

Paraphrasing our teacher during MSc in Calcutta University, Prof P N Ghosh

Trekking through a network

Undirected network

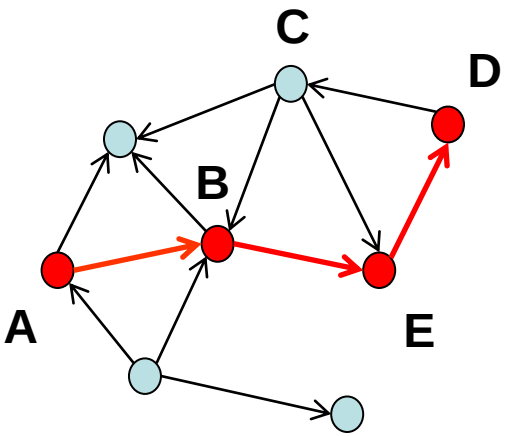


Network **path**: a sequence of nodes such that every consecutive pair is connected by a link in the network,

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

Directed network



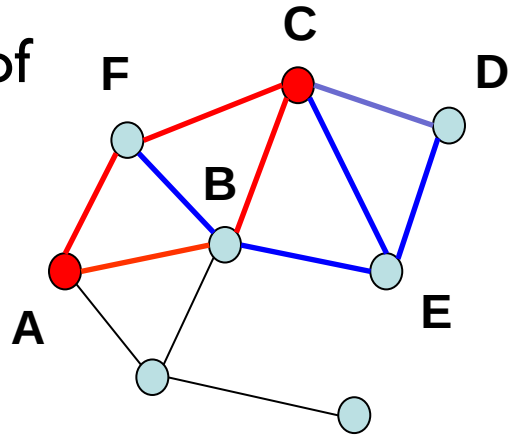
Example (Directed network): Unique path from A to D having length 3 is {A,B,E,D}

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)

Global property of networks: Geodesic

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 $[\,\mathbf{A}^r\,]_{ij} > 0$

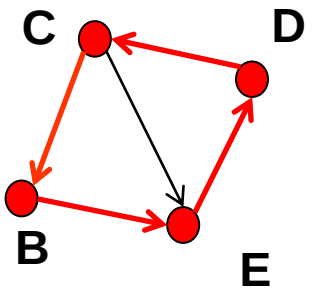
If a network has disconnected components, there may be no geodesic between members of one component and those of another $\Rightarrow$ infinite geodesic distance

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

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 2 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 $L_r = \sum_i A_{ik}\,A_{ky} \ldots A_{xi} = \sum_i [\,\mathbf{A}^r\,]_{ii} = \mathrm{Tr}\,\mathbf{A}^r$

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.

In undirected networks a 3-clique is the shortest non-trivial reciprocated relation and we can ask about their frequency in terms of clustering coefficient.

Local property of networks: Transitivity

A relation "R" is said to be transitive if *R(a,b)* and *R(b,c)* $\Rightarrow$ *R(a,c)*

E.g., "R" maybe equality, i.e., R(a,b): a = b

In networks, simplest relation between a pair of nodes is "connected by a link."
If this relation is transitive, it would mean that

- if nodes i and j are connected,
- and nodes j and k are connected,
- then nodes i and k are also connected

$\Rightarrow$ "the friend of my friend is also my friend."

Perfect transitivity occurs only in cliques

Partial transitivity: if i and j are connected, and j and k are connected, that makes it very likely that i and k will be connected, i.e., a closed triad will be formed between i,j and k

Measured by clustering coefficient C, i.e., the fraction of paths of length two which are closed (Extreme cases: Trees have C=0 , Cliques C=1)

Clustering coefficient

Clustering coefficient can also be defined as

C = 6 (no. of triangles)/(number of paths of length 2)

The factor of 6 arises because each triangle contains six distinct paths of length 2, all of them closed.

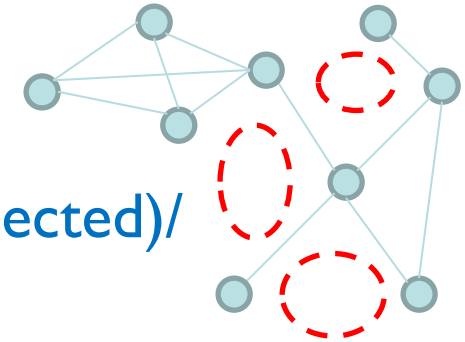
Or

C = 3 (no. of triangles)/(number of connected triples)

A connected triple means three nodes (i,j,k) with links (i,j) and (j,k) [The link (i,k) may be present or not present]

The factor of 3 arises because each triangle gets counted 3 times when we count the number of connected triples in it

Local clustering coefficient of a node



C_i = (no. of pairs of neighbors of i that are connected)/(total no. of pairs of neighbors of i = k_i (k_i-1)/2)

Local clustering can be used as a probe for the existence of "structural holes" (missing connections between neighbors) in a network – such "holes" reduce efficiency of information or traffic flow in network and increases the importance (power) of the central node around which the hole is created

Adapted from Newman, *Networks*

Triadic Closure

The clustering coefficient measures the average probability that two neighbors of a node are mutual neighbors.

In effect it measures the density of triangles in the networks and it is of interest because in many cases it is found to have values sharply different from what one would expect on the basis of chance.

If we consider a network with a given degree distribution in which connections between nodes are made at random, the clustering coefficient takes the value $C = (1/N) [\langle k^2 \rangle - \langle k \rangle]^2 / \langle k \rangle^3 = L/N$

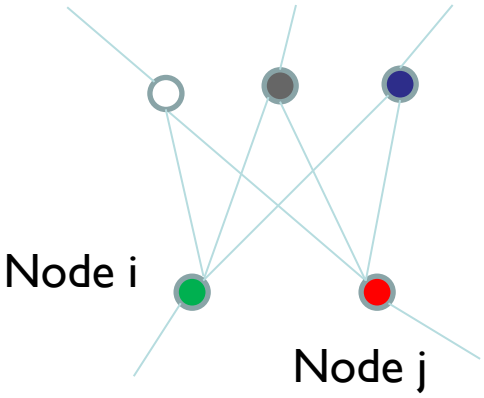
Thus, for large networks (N large) with finite first and second moments of the degree distribution, C is expected to be small

A relatively large value of C will imply connections happening through triadic closure: an "open" triad of vertices (i.e., a triad in which one node is linked to the other two, but the third possible link is absent) is "closed" by the addition of the last link, forming a triangle $\Rightarrow$ can result in modules or communities

Neighborhood Similarity

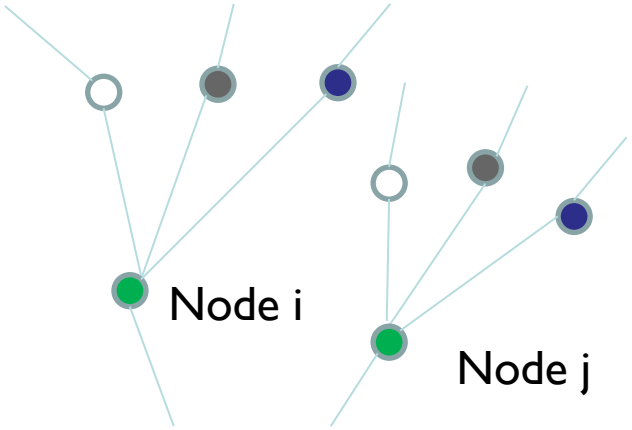
How does a web-site say
"If you like this (Q), you will probably like these (X, Y, Z) ?"
i.e., how is it possible to say which node (or nodes) is most similar to another given node in a specific network ?

- 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



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 ?

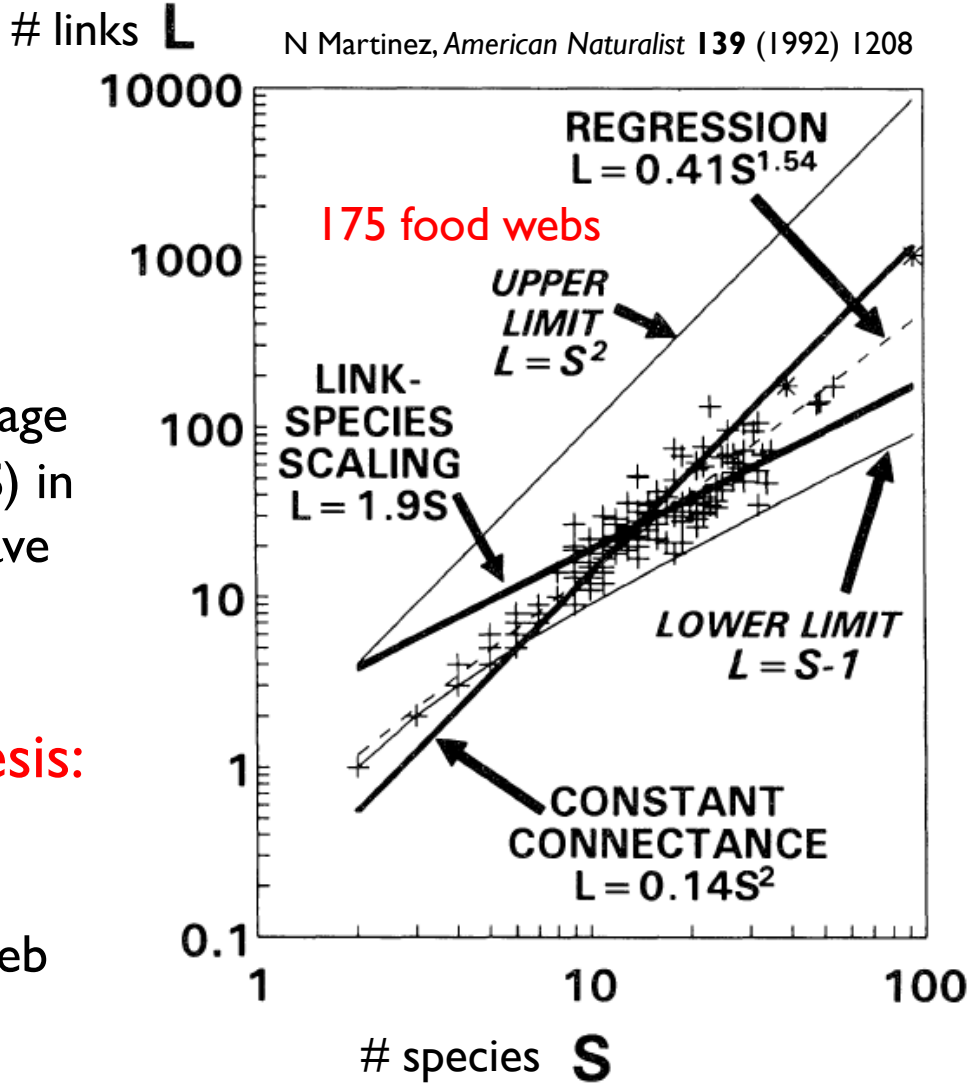
Question: How does the number of trophic links L in a food web vary with the number of trophic species S ?

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



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

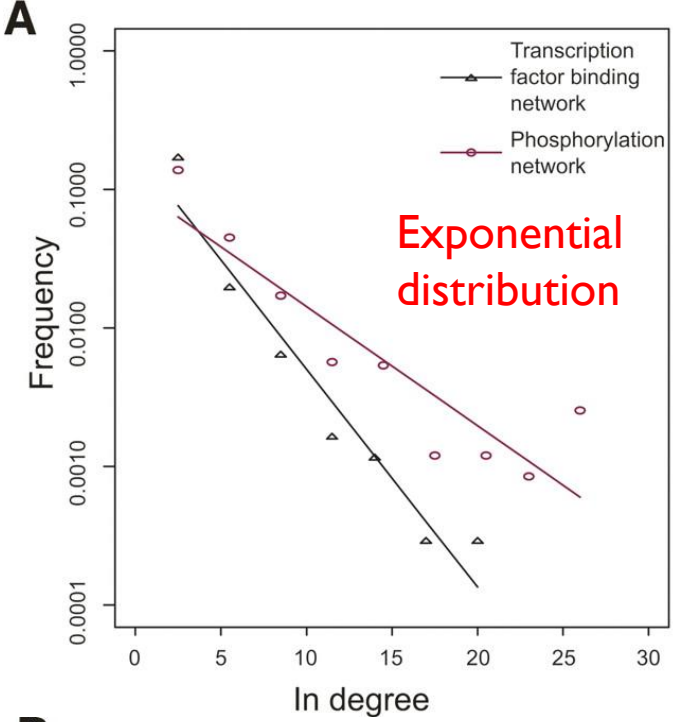
$L = \sum^{N}_{i=1} k_{i(in)} = \sum^{N}_{j=1} k_{j(out)} = \sum_{i,j} A_{ij}$

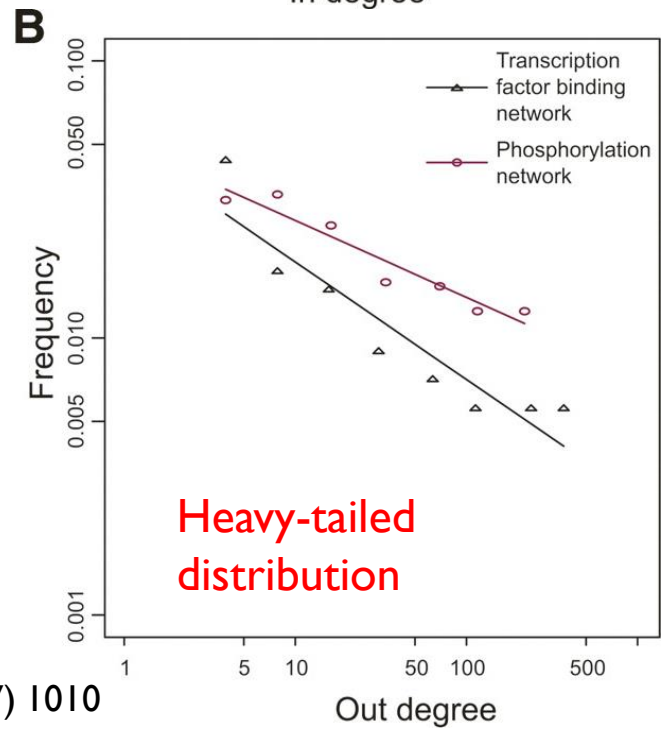
as each incoming end of a link is paired with an outgoing end of a link

$\Rightarrow$ 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

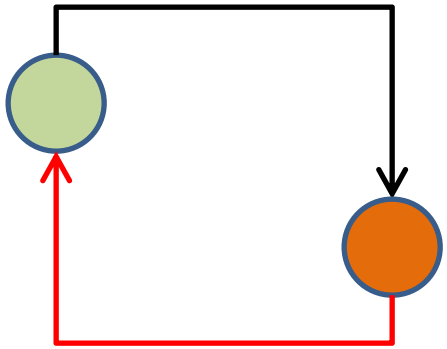




X Zhu et al. *Genes Dev.* **21** (2007) 1010

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: gene regulation or synaptic contacts

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)

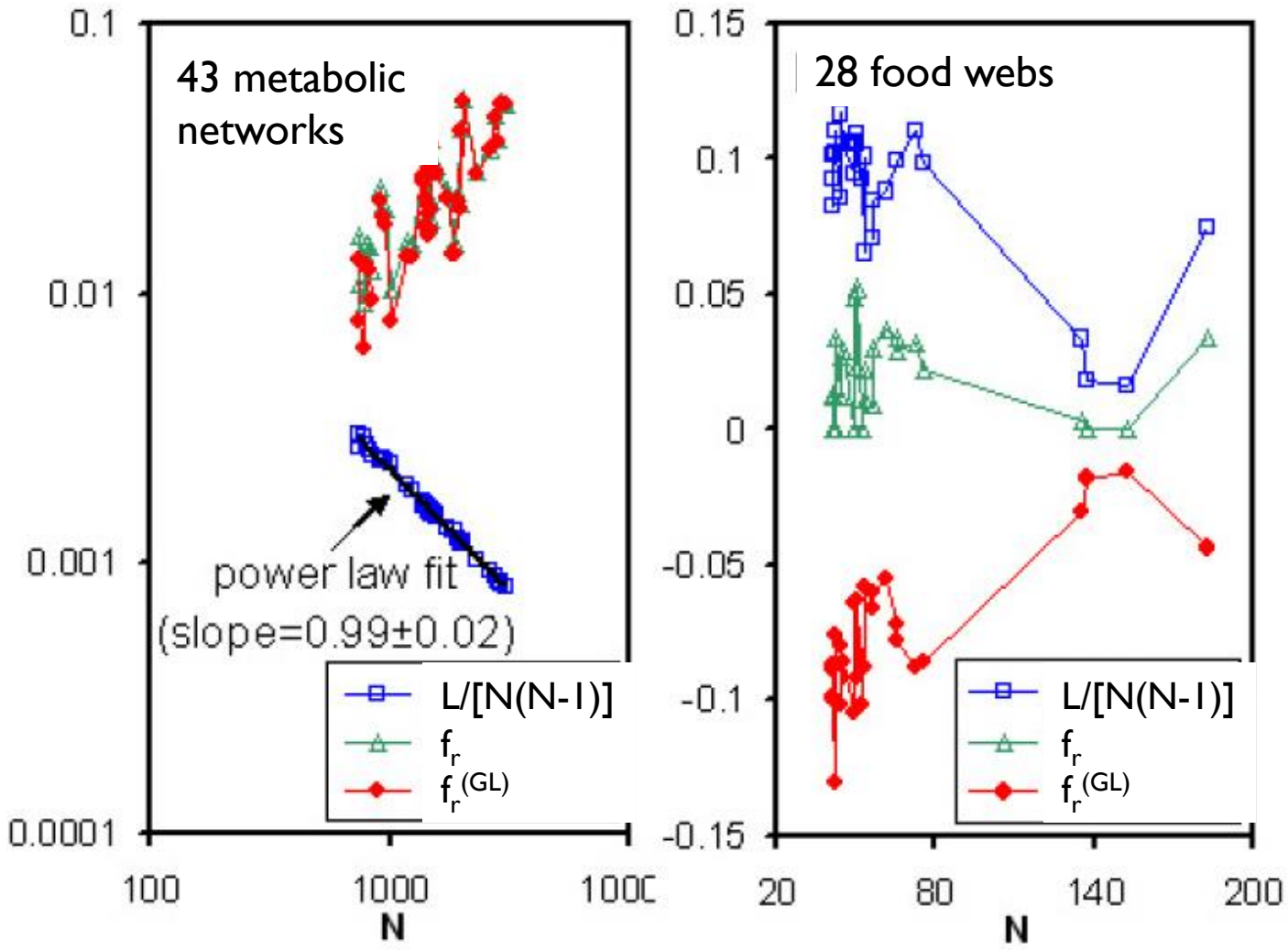
If there are no reciprocal edges, $[f_r^{(GL)}]_{min} = -\langle A\rangle/[1-\langle A\rangle]$

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.

Reciprocity in the biological world

- Neuronal network (*C. elegans* chemical synapses) shows reciprocity
- Metabolic networks are weakly reciprocal (reciprocity could be linked to potential reversibility of biochemical reactions
- Food webs are anti-reciprocal

Garlaschelli & Loffredo, PRL **93** (2004) 268701



Networks belonging to the same class (e.g., metabolic, neuronal or ecological) appear to have similar values of reciprocity

Next up: Network Models…

Assignment

1. Calculate the mean path length of (i) a chain of N nodes, and (ii) a 2-dimensional square lattice of N nodes.

2. If you consider each node in a ring of N nodes to be connected to their nearest and next-nearest neighbors (i.e., degree of each node is 4) what is its (i) mean path length and (ii) clustering coefficient.

3. Consider a (a) 2-dimensional square lattice with nearest neighbour connections only (degree of each node is 4), (b) 2-dimensional triangular lattice (degree of each node is 6), and (c) 2-dimensional square lattice with nearest neighbour as well as next nearest neighbour connections (degree of each node is 8) – calculate the clustering coefficient in each case.

*4. [Open-ended Numeric Exercise] Consider a 2-dimensional lattice of N nodes with nearest neighbour connections – try various values of N for your exercise – and sequentially replace 2 randomly chosen pairs of nearest nbr links (a,a') and (b,b') by (a,b) and (a',b') [unless the links already exist] with probability $r^{-\lambda}$ where r is the mean distance between link pairs (a,a') and (b,b') and λ (≥ 0) is a parameter. For $\lambda=0$, you should see something like the Watts-Strogatz "small-world" network while for $\lambda \to \infty$ the original lattice ("large-world" network) should not change. Around what value of λ would you see the transition from "small-world" to "large-world" ?