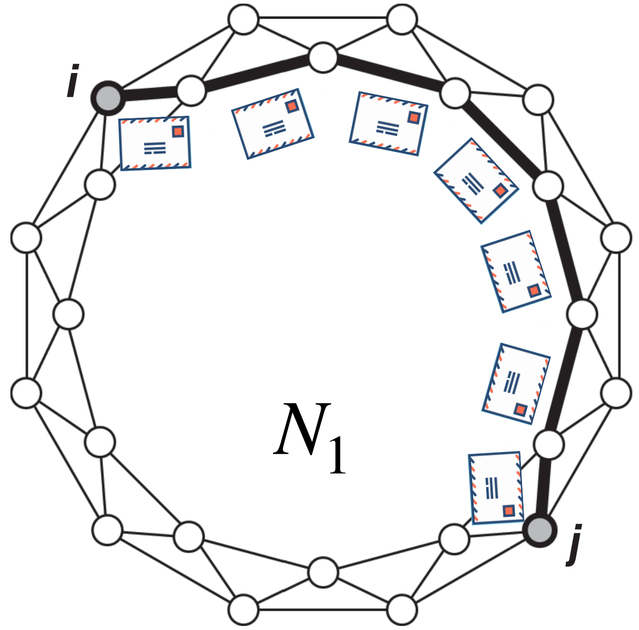
Measures and Metrics, Networks

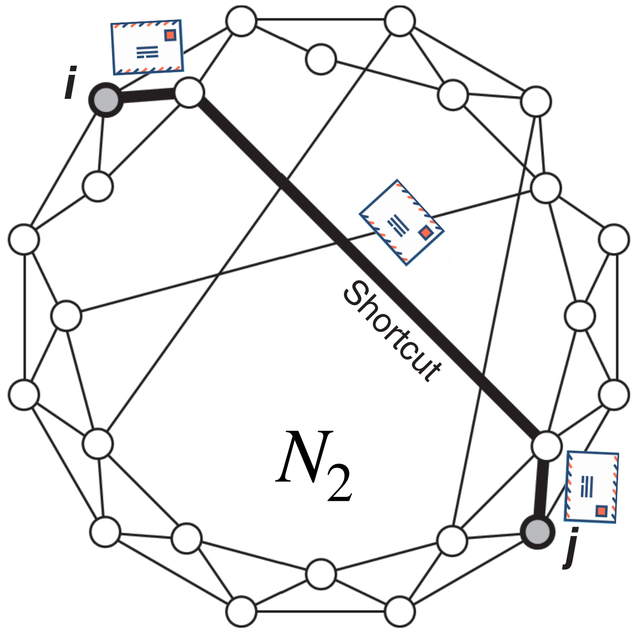
The Small-world Effect

A renowned (and measurable) network phenomenon is the small-world effect.

Informally, we have a small-world effect when we can find shorter-than-expected distances between pairs of nodes.

The typical example to illustrate a small-world effect is Milgram's experiment, where people were asked to get a letter from an initial holder to a distant target person by passing it from acquaintance to acquaintance through their social network. The letters that made it to the target did so in a remarkably small number of steps.

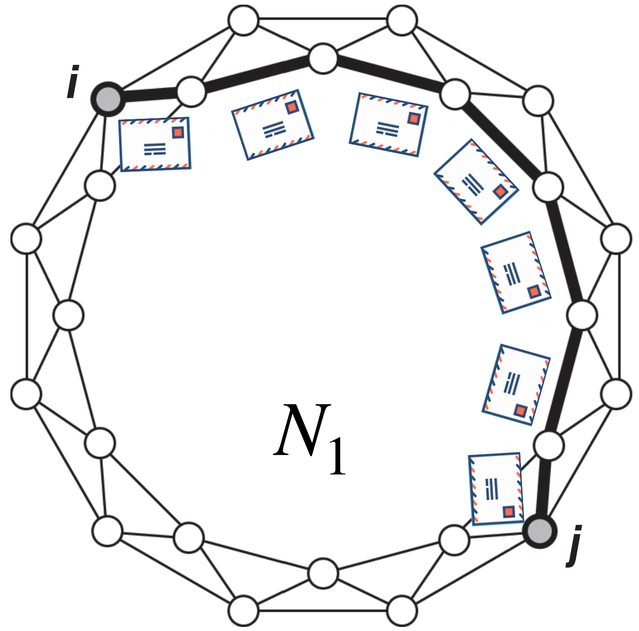


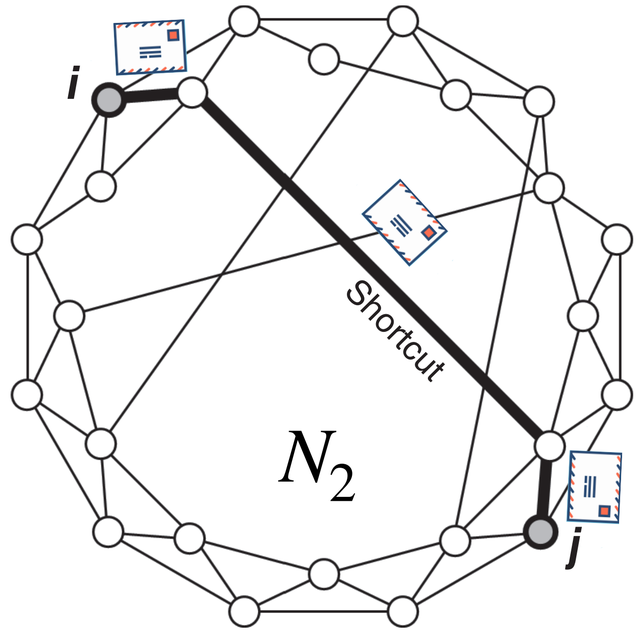


The Small-world Effect

Mathematically, let d_{ij} be the length of the shortest path through a network between nodes i and j; then, the mean distance ℓ_i for a node i corresponds to $\ell_i = \frac{\sum_j d_{ij}}{n}$ and the mean distance for the whole network corresponds to $\ell = \frac{\sum_i \ell_i}{n} = \frac{\sum_{ij} d_{ij}}{n^2}$ (for single-component networks).

Simplistically—as we will see more accurate measures using *random graphs*—a family of networks shows small-world effects when $\ell \propto \log n$ (i.e., when ℓ is directly proportional to $\log n$ by a constant k).

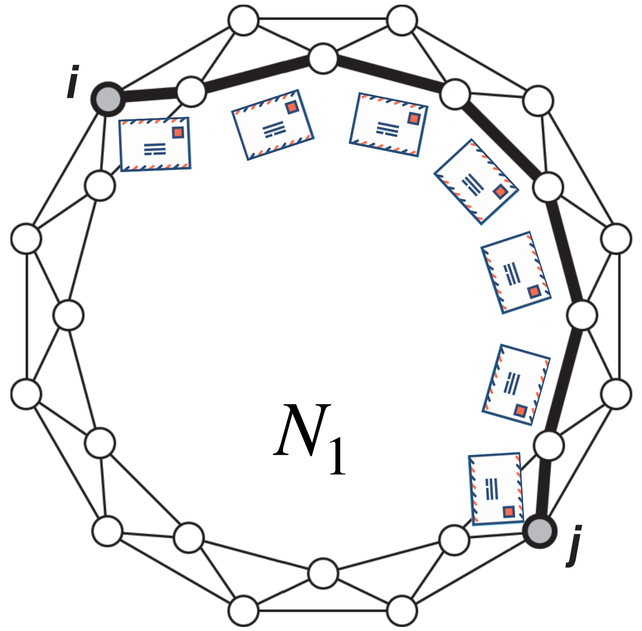


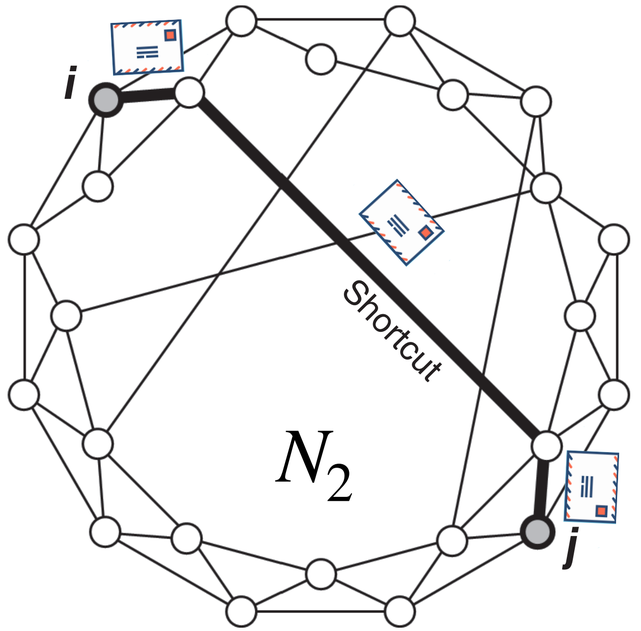


The Small-world Effect

Properties of small-world networks include:

- many highly-clustered groups (e.g., cliques) where all nodes are densely connected
- hubs that serve as “mediators” to shorten the lengths between other edges
- these networks are particularly robust to random perturbations (e.g., deletion of a random node rarely causes a sensible change of ℓ)—thanks to the low hub-to-leaf ratio. Vice versa, rare/selective deletions of hubs dramatically increase ℓ





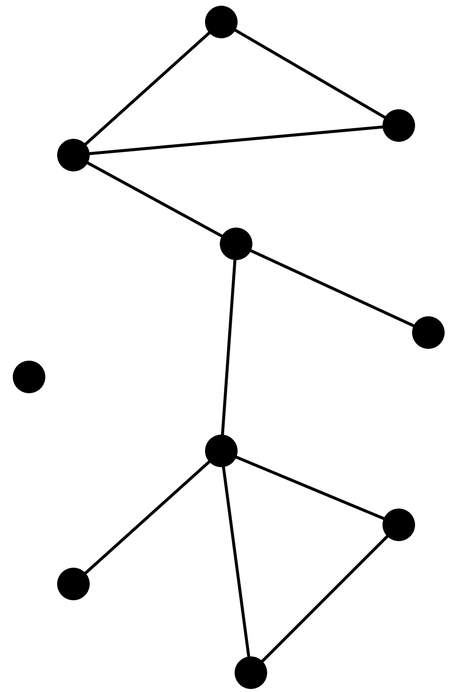
Degree Distribution

Reminder: the degree of a node corresponds to the number of edges attached to that node.

Consider an undirected network and let p_d be the fraction of nodes that have degree d. E.g., in the network on the right we have:

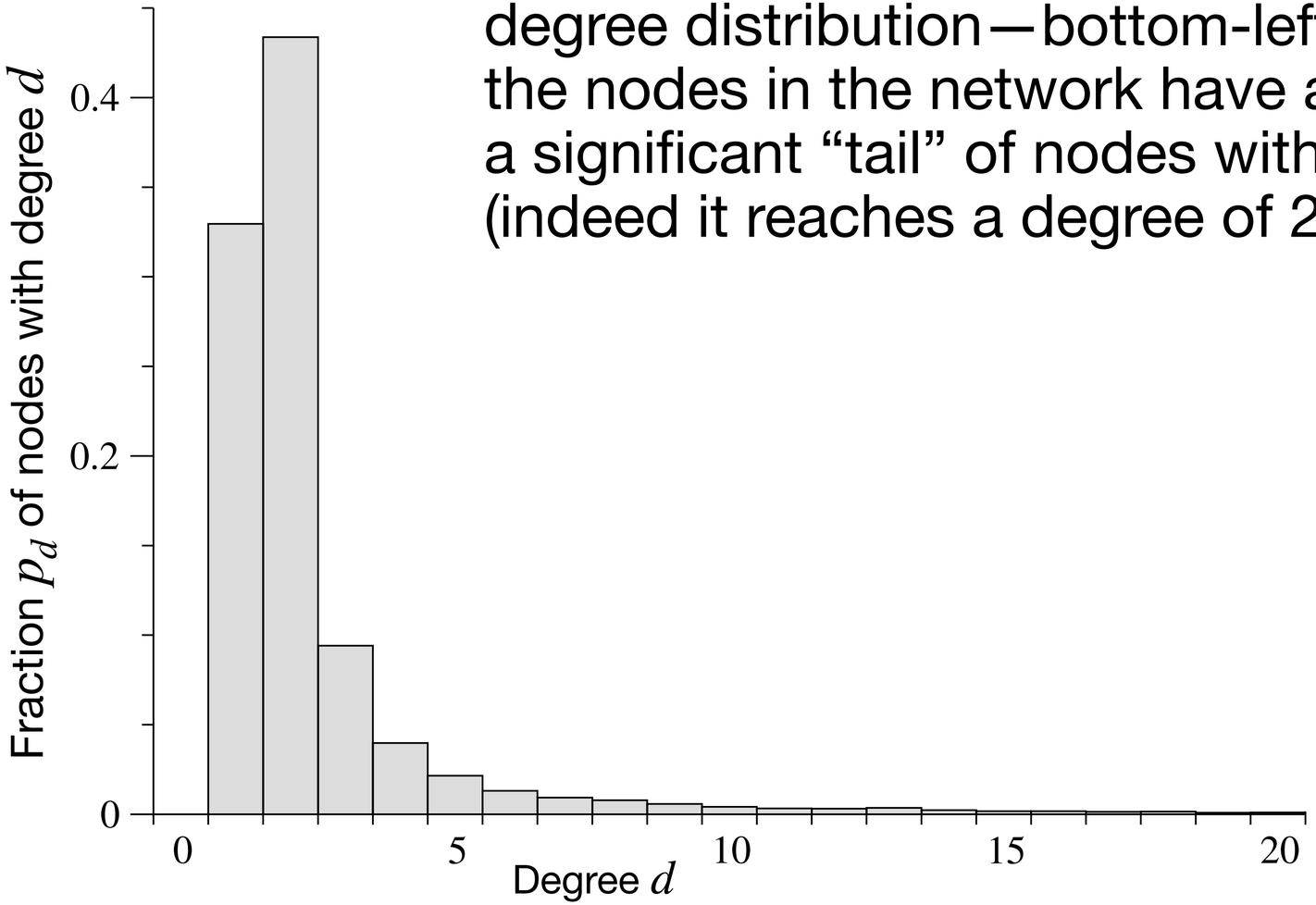
$$p_0 = 1/10 \quad p_1 = 2/10 \quad p_2 = 4/10 \quad p_3 = 2/10 \quad p_4 = 1/10 \quad p_{5+} = 0/10$$

That ratio is essentially the *probability* of a given node to have that degree.

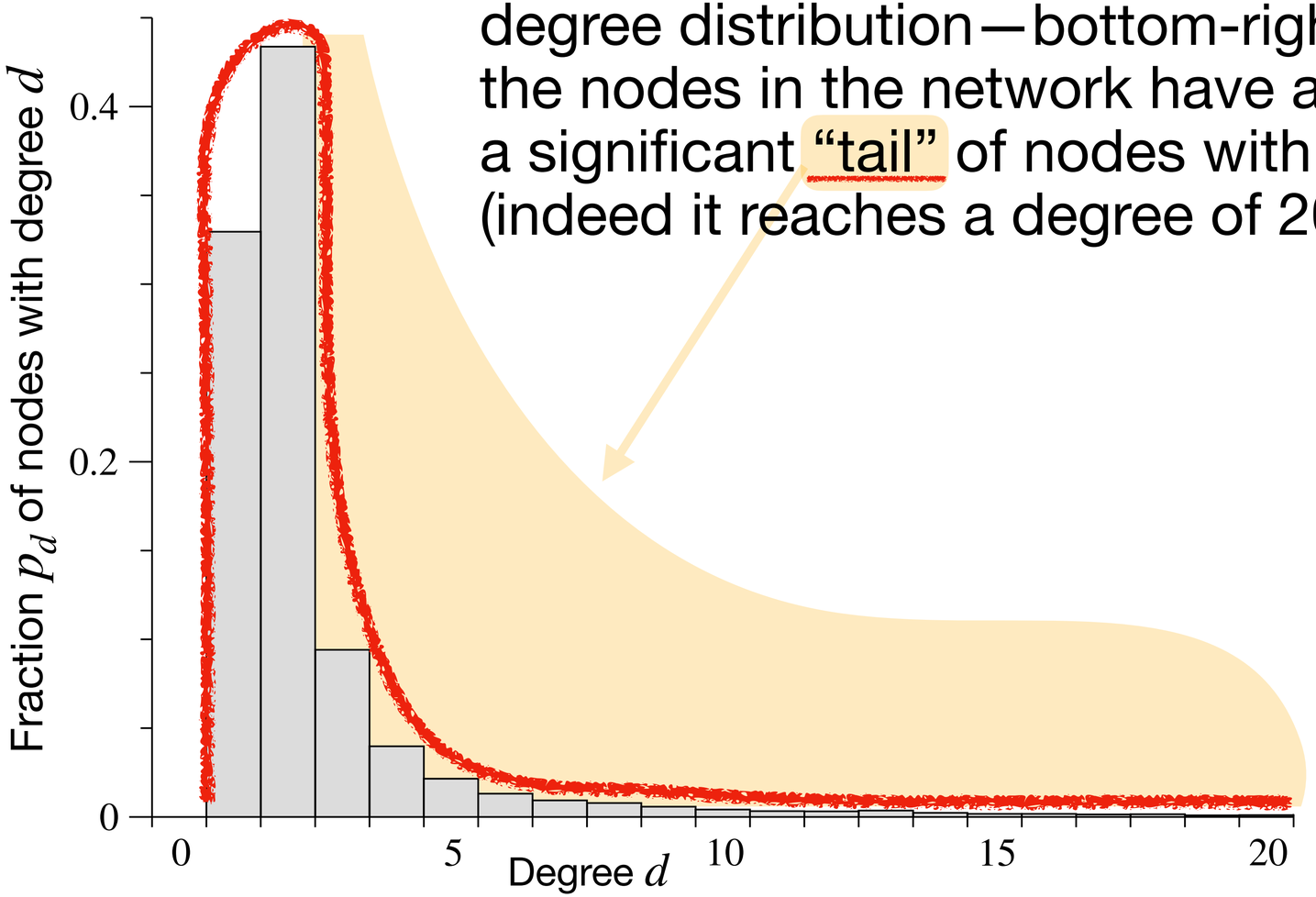


Power Laws and Scale-free Networks

Let us take the degrees of (a portion) of the Internet and plot the degree distribution—bottom-left. The figure shows that most of the nodes in the network have a low degree. However, there exists a significant “tail” of nodes with substantially higher degree (indeed it reaches a degree of 2000+).



Power Laws and Scale-free Networks

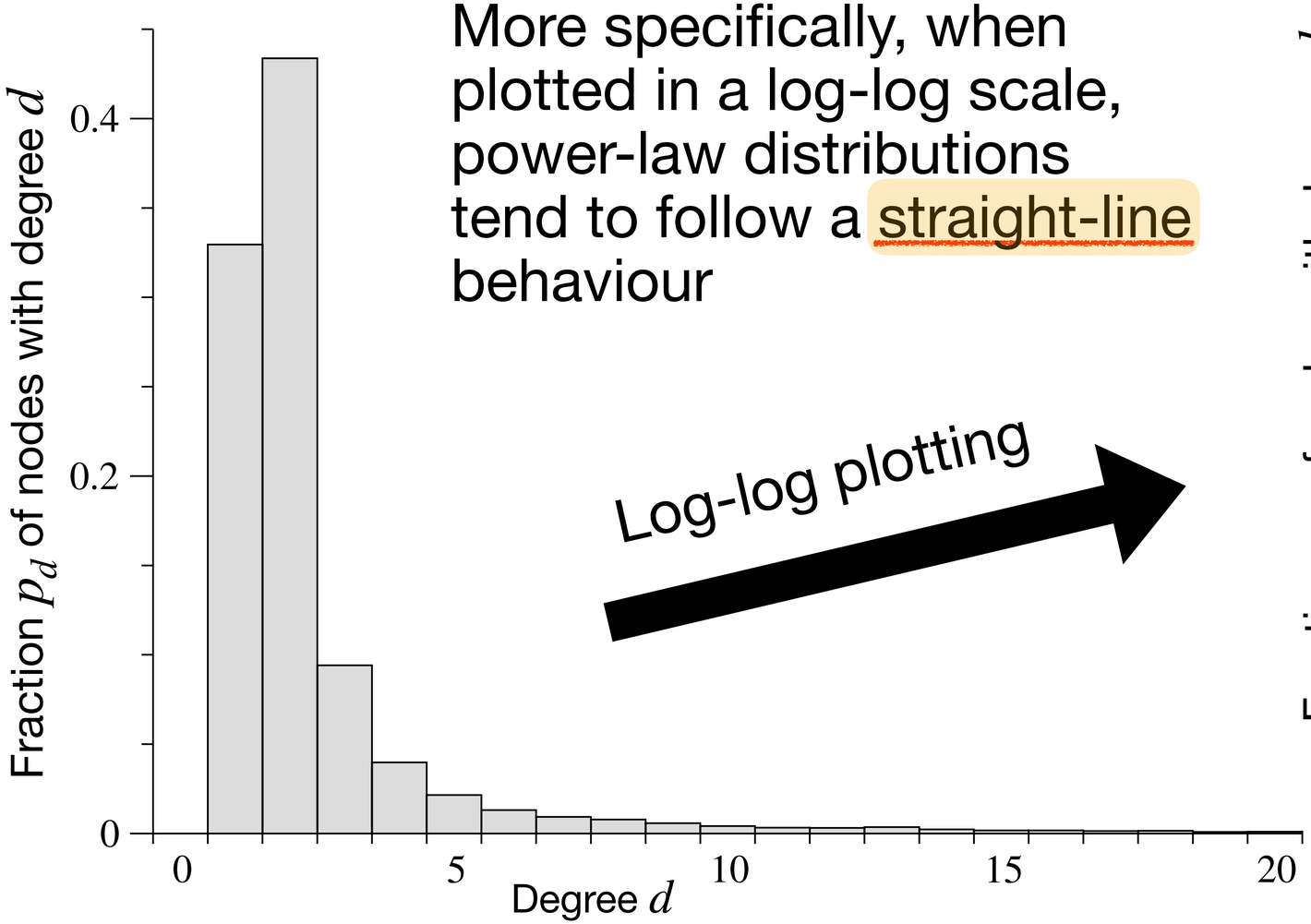


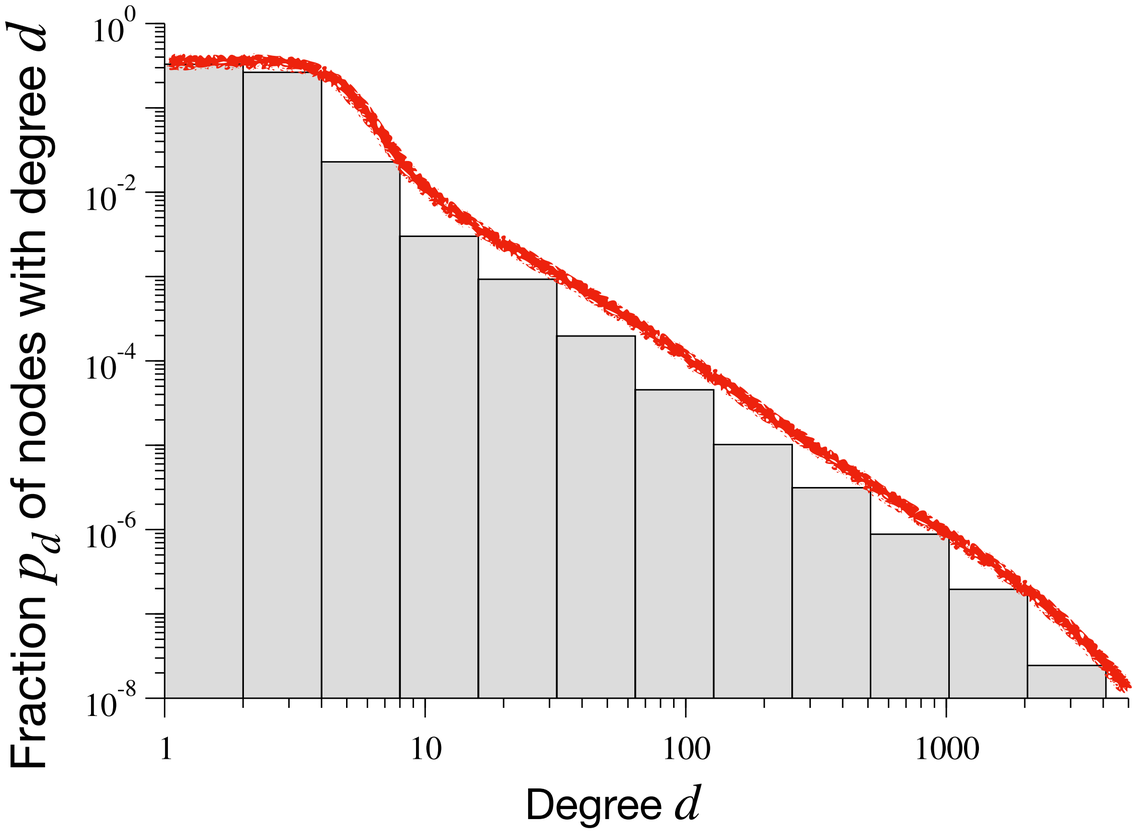
Let us take the degrees of (a portion) of the Internet and plot the degree distribution—bottom-right. The figure shows that most of the nodes in the network have a low degree. However, there exists a significant "tail" of nodes with substantially higher degree (indeed it reaches a degree of 2000+).

Power Laws and Scale-free Networks

More specifically, when plotted in a log-log scale, power-law distributions tend to follow a straight-line behaviour

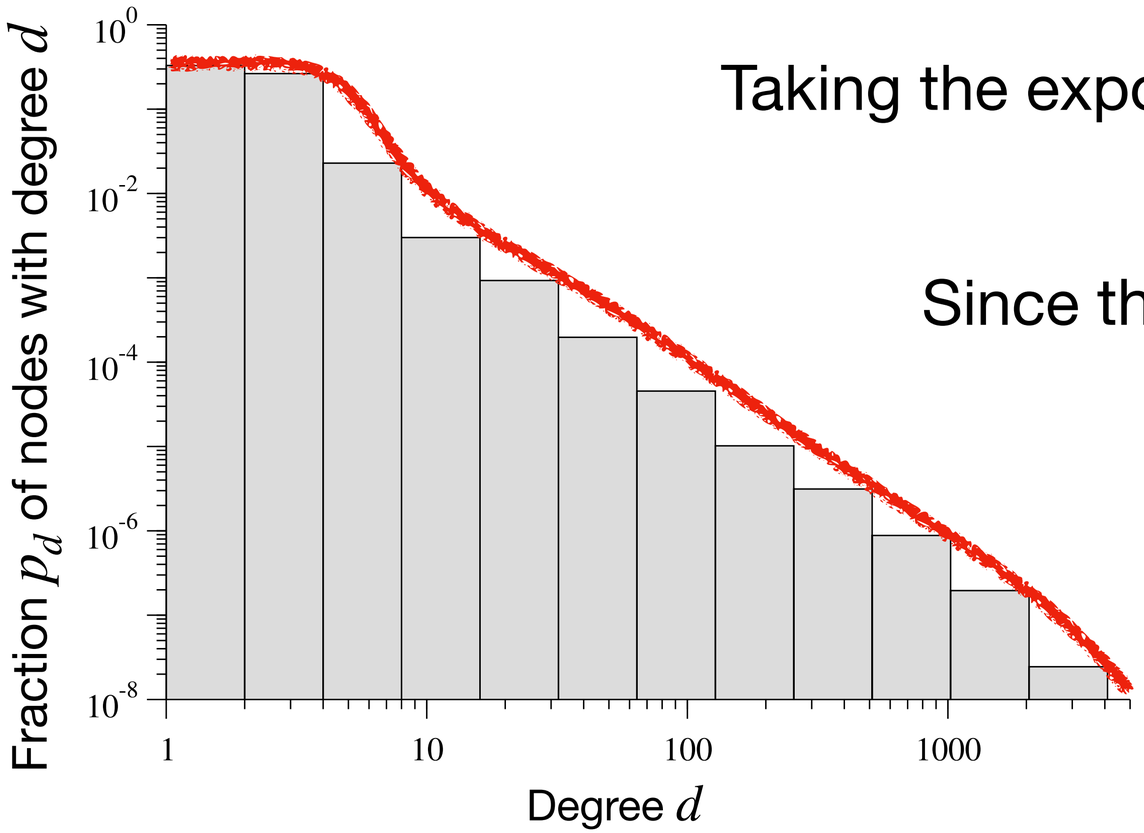
Log-log plotting





Power Laws and Scale-free Networks

Distributions of this kind are described by the formula $\ln p_d = -\alpha \ln d + c$ where α and c are constants that respectively modify the slope and normalise the curve of the distribution.

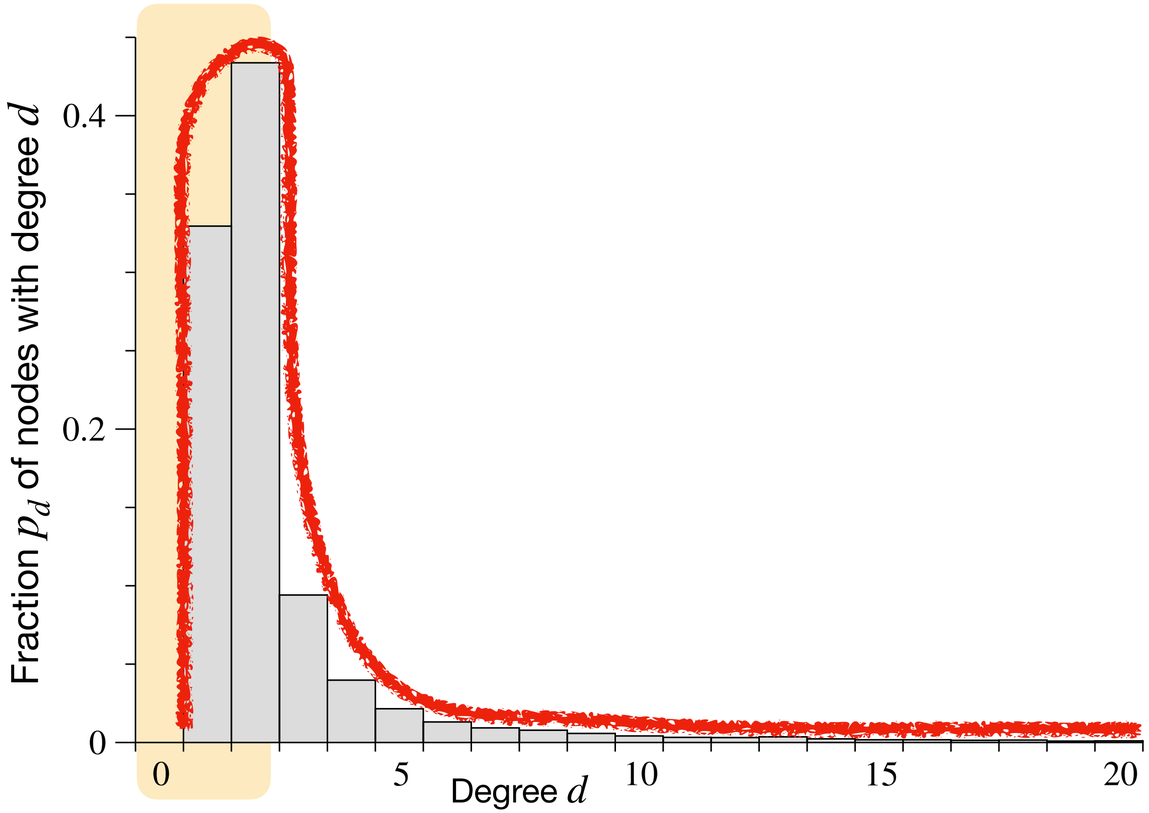


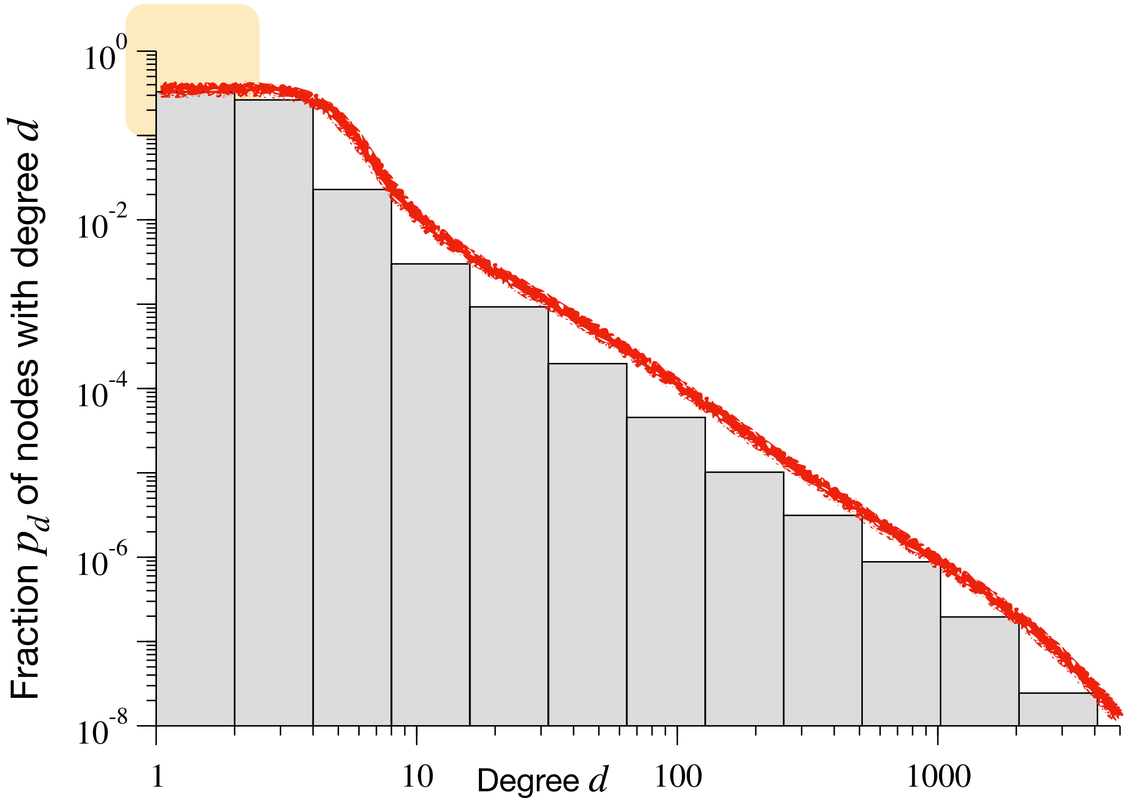
Taking the exponential of both sides of the formula, we have $p_d = Cd^{-\alpha}$ (with $C = e^c$).

Since the distribution is dependent on a power (with exponent α) of the degree d, it is called a "*power law*" distribution.

Power Laws and Scale-free Networks

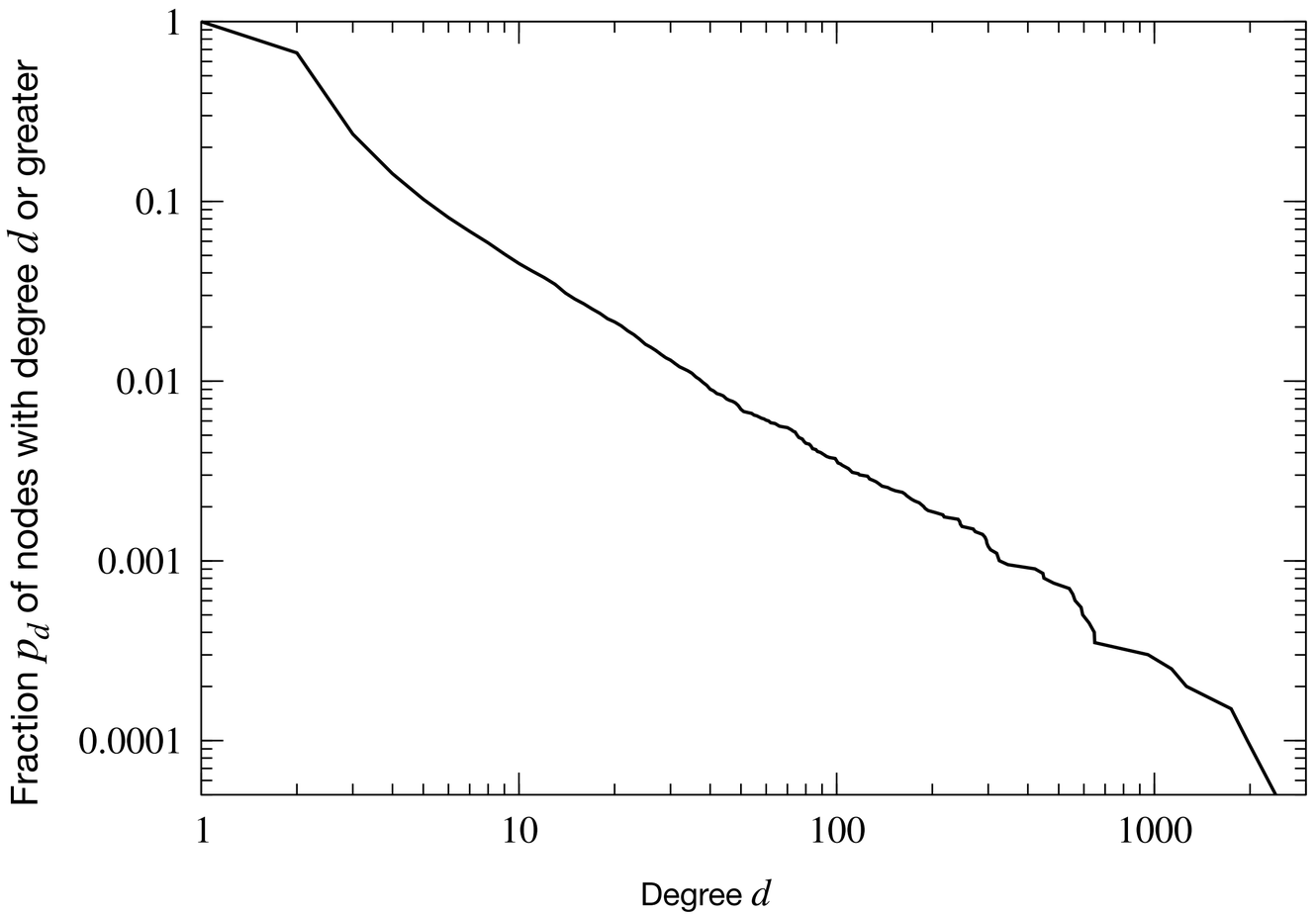
Detecting power-laws by just visualising the distribution (particularly in log-log form) cannot be trusted. Indeed, in our example we see a “deceiving” non-monotonically decreasing (direct scale) and non-straight (log-log scale) distribution curve.





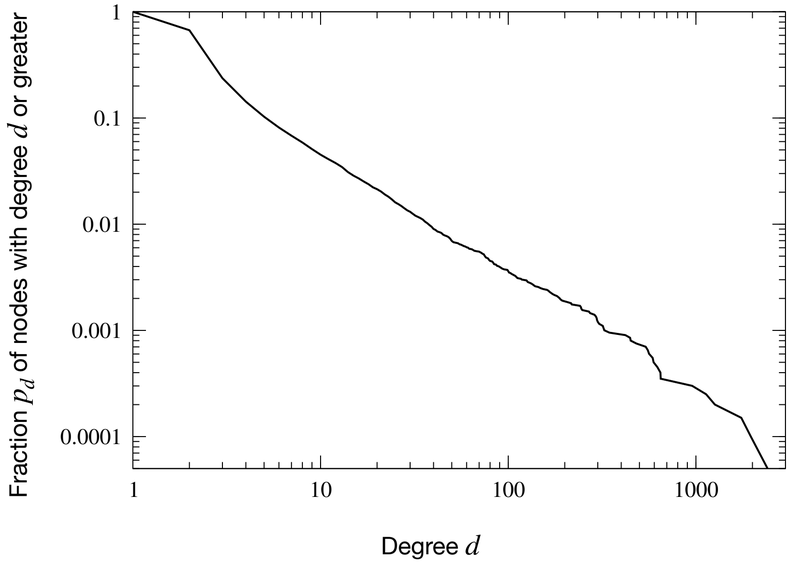
Power Laws and Scale-free Networks

To detect power-law behaviours, we can use the *cumulative distribution function*, which is defined by the formula $P_d = \sum_{d'=d}^{\infty} p_{d'}$, so that P_d is the fraction of nodes that have degree d or greater.



Power Laws and Scale-free Networks

Also here we are looking for a straight-line behaviour. However, while the curve lends itself to less-statistically-biased visual interpretations, we can get a precise measure of how close our distribution approximates a power-law by calculating the value of α.



Indeed, if $p_d = Cd^{-\alpha}$ then

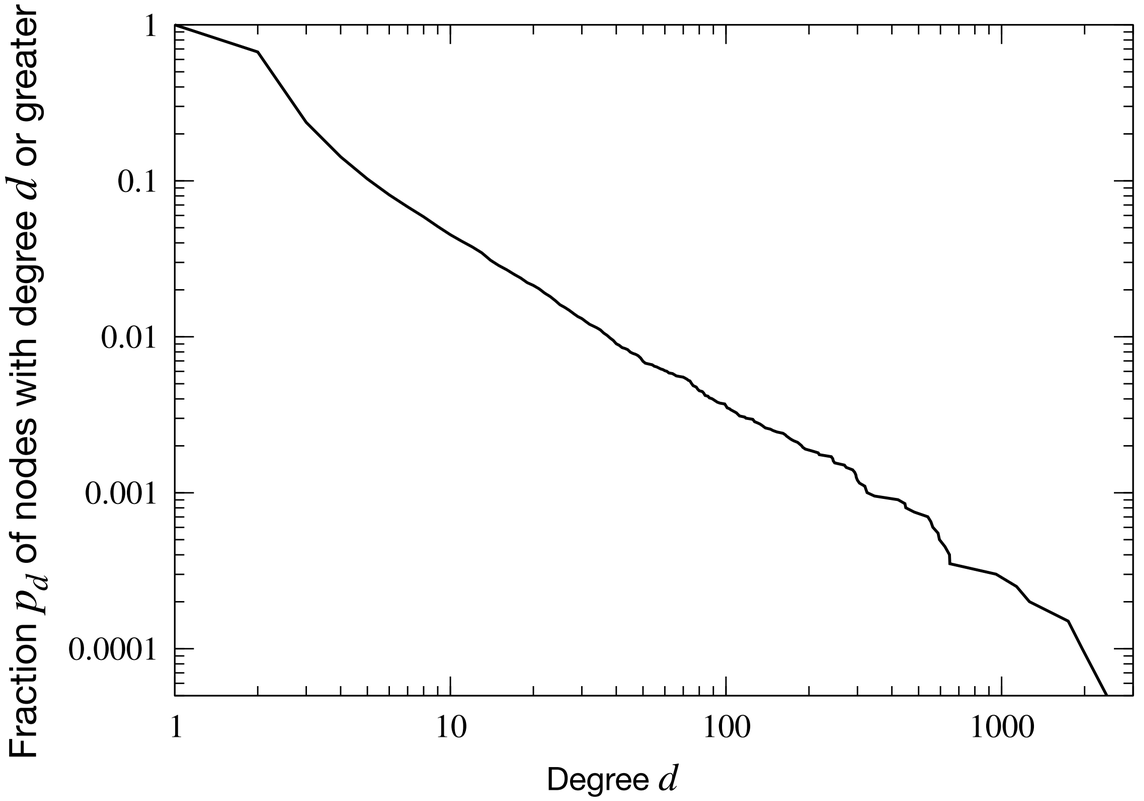
$$P_d = C\sum_{d'=d}^{\infty} d'^{-\alpha} \underbrace{\simeq}_{\text{Assuming } \alpha > 1} C\int_d^{\infty} d'^{-\alpha}\ \partial d' = \frac{C}{\alpha - 1} d^{-(\alpha-1)}$$

so that α becomes the exponent determining the distribution (on d) as

$$\alpha = 1 + n\left(\sum_i \ln \frac{d_i}{d_{min} - 1/2}\right)^{-1}$$

Power Laws and Scale-free Networks

Also here we are looking for a straight-line behaviour. However, while the curve lends itself to less-statistically-biased visual interpretations, we can get a precise measure of how close our distribution approximates a power-law by calculating the value of α.



Indeed, if $p_d = Cd^{-\alpha}$ then

$$P_d = C\sum_{d'=d}^{\infty} d'^{-\alpha} \underbrace{\simeq C\int_d^{\infty} d'^{-\alpha}\,\partial d'}_{\text{Assuming } \alpha > 1} = \frac{C}{\alpha - 1} d^{-(\alpha-1)}$$

so that α becomes the exponent determining the distribution (on d) as

$$\alpha = 1 + n\left(\sum_i \ln \frac{d_i}{d_{min} - 1/2}\right)^{-1}$$

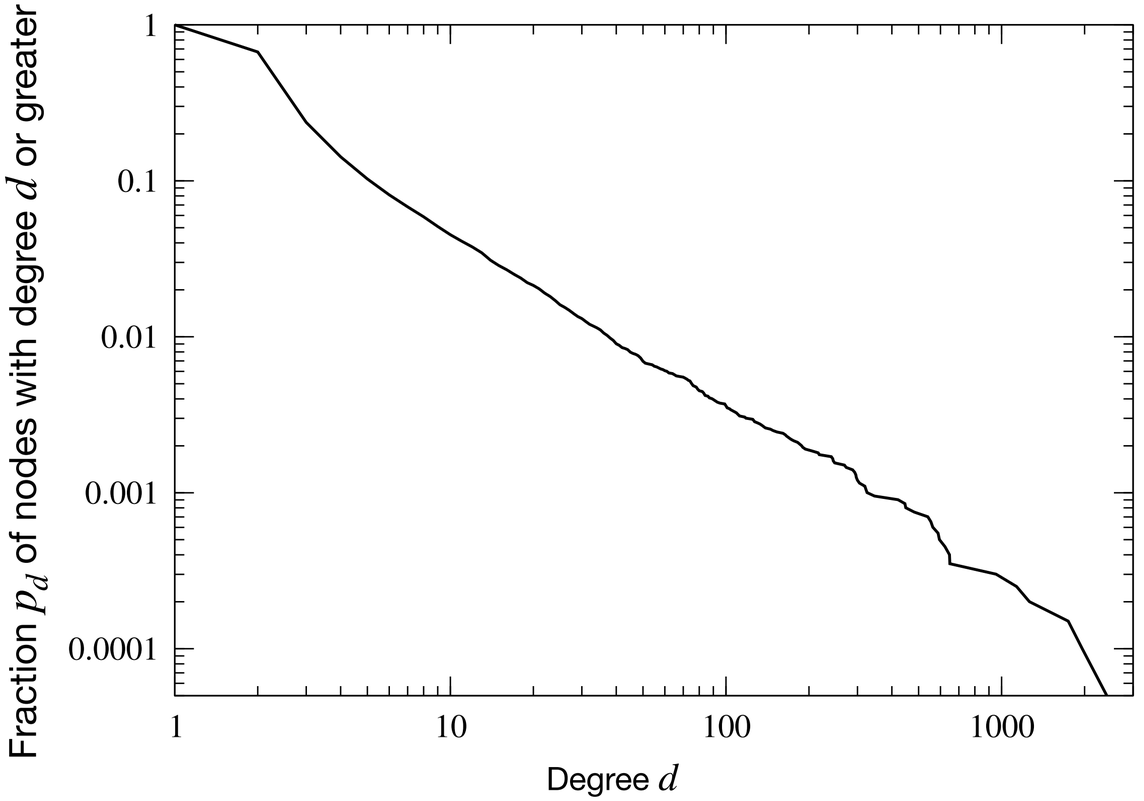
Empirically, in power-law distributions $2 \leq \alpha \leq 3$.

Power Laws and Scale-free Networks

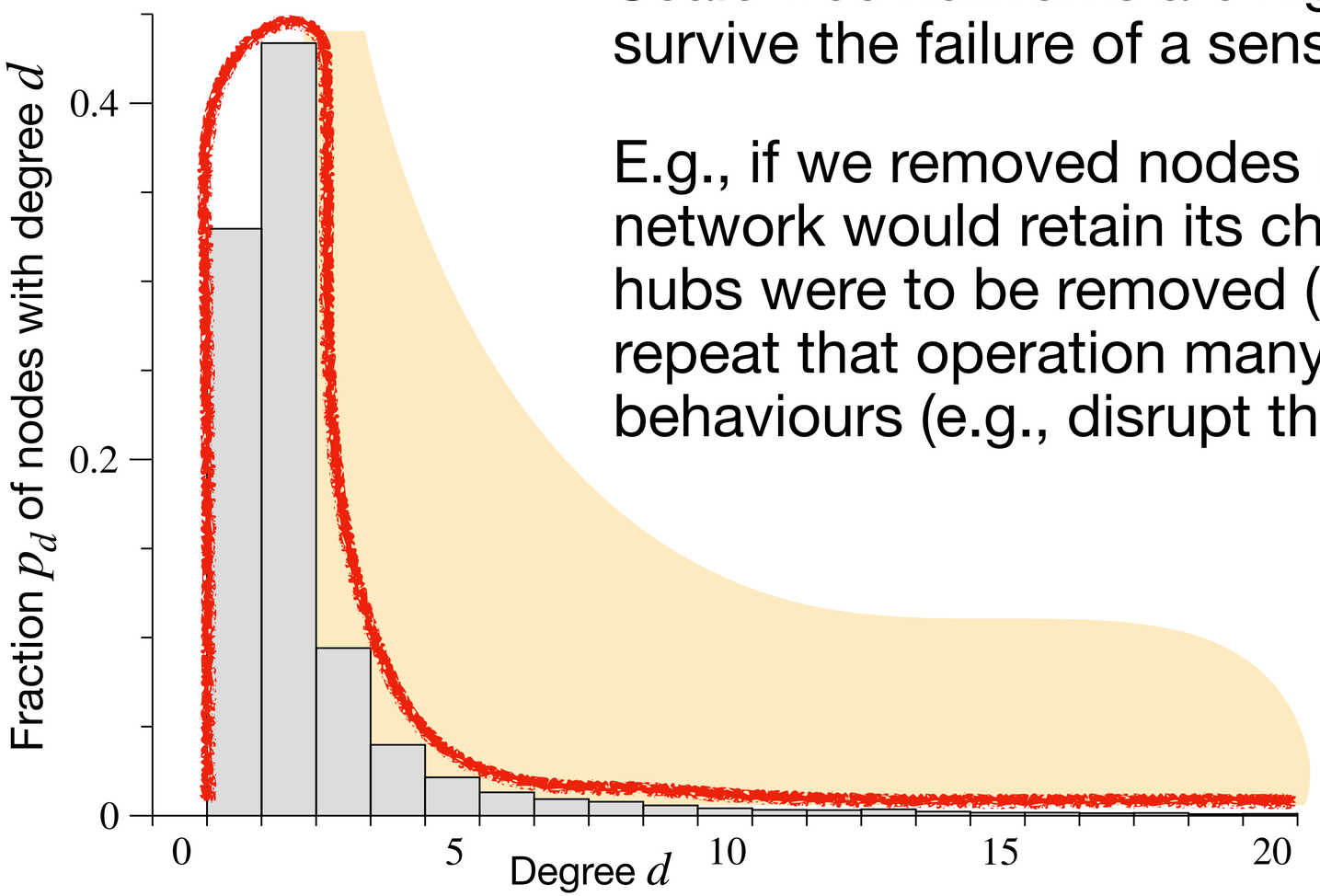
Networks whose degree distribution follows a power-law behaviour are usually called *scale-free networks*.

The reason for the name comes from the fact that power laws are scale-invariant, i.e., that scaling the argument, here d, by a constant factor just causes a multiplication of the original power-law relation by that constant.

This is also why we look for straight-line behaviours in log-log plots, which reduce the "noise" derived from constant multiplications.



Properties of Scale-free Networks



Scale-free networks are highly robust networks that can survive the failure of a sensible number of their nodes.

E.g., if we removed nodes randomly from the Internet, the network would retain its characterising behaviours. If central hubs were to be removed (by choice or luck), we should repeat that operation many times to significantly change the behaviours (e.g., disrupt the connectivity) of the network.

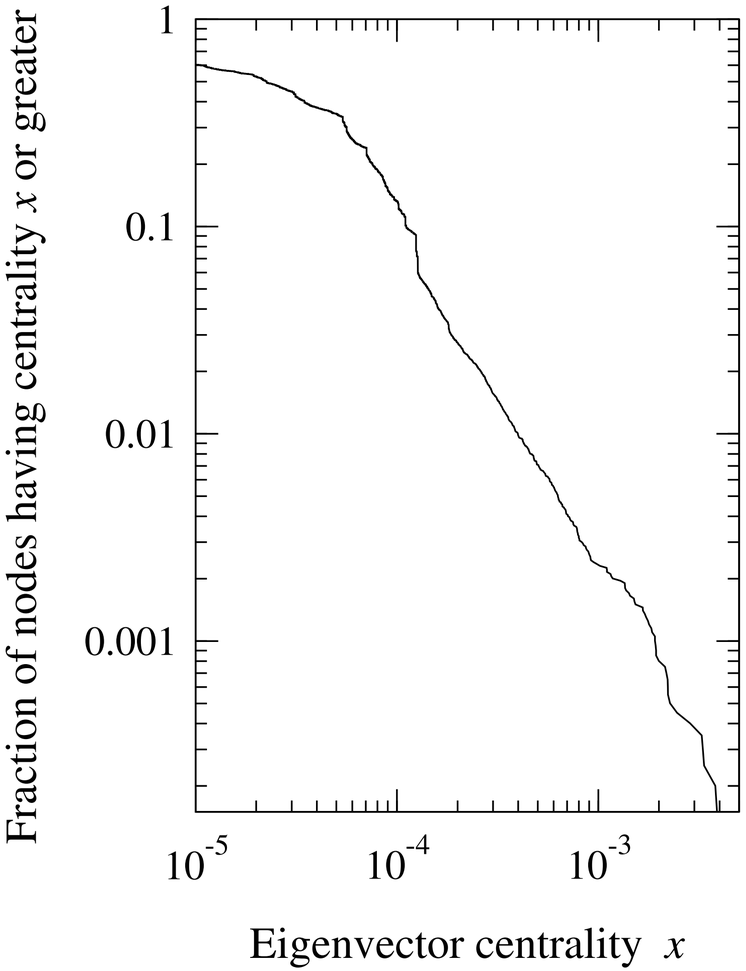
Distribution of Other Centrality Measures

The degree is not the only measure we can study the distribution of.

Other examples are eigenvector centrality (and its variants), betweenness centrality, and closeness centrality.

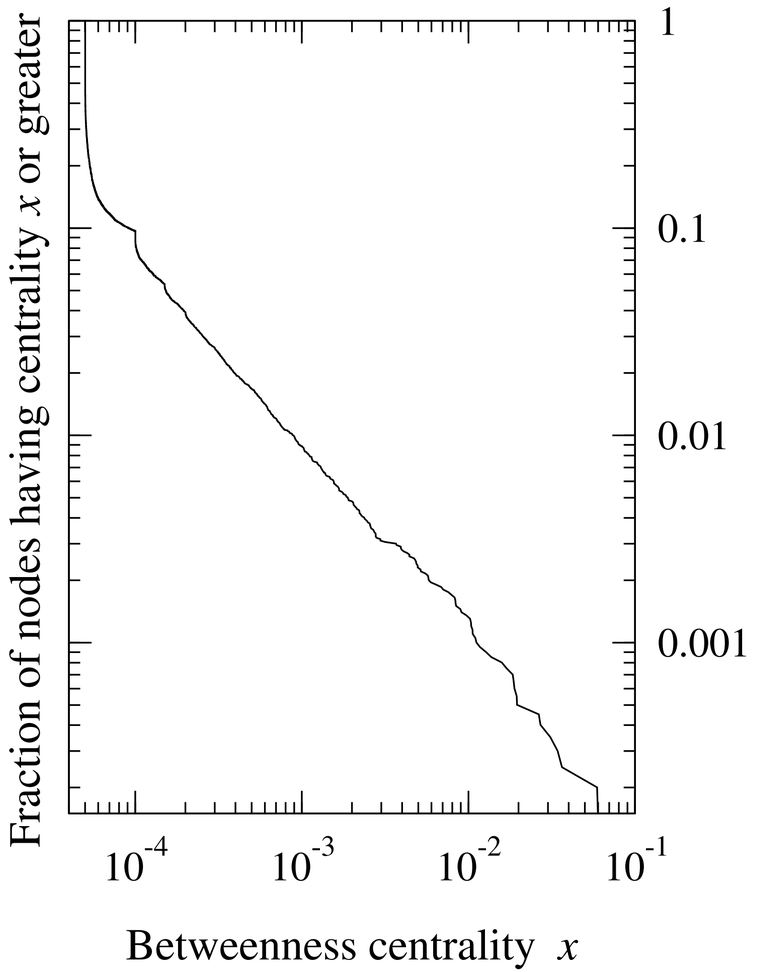
Eigenvector centrality is an extended form of degree (centrality), which takes into account not only how many neighbours a node has, but also how central those neighbours themselves are.

Eigenvector centrality often has a right-skewed distribution (similar to that of the degree). E.g., looking at the cumulative distribution of eigenvector centralities for the nodes of the Internet we see the typical straight line on the logarithmic scales.



Distribution of Other Centrality Measures

Betweenness centrality also tends to assume the same distribution — e.g., on the right, the cumulative distribution of betweenness for the nodes of the Internet.

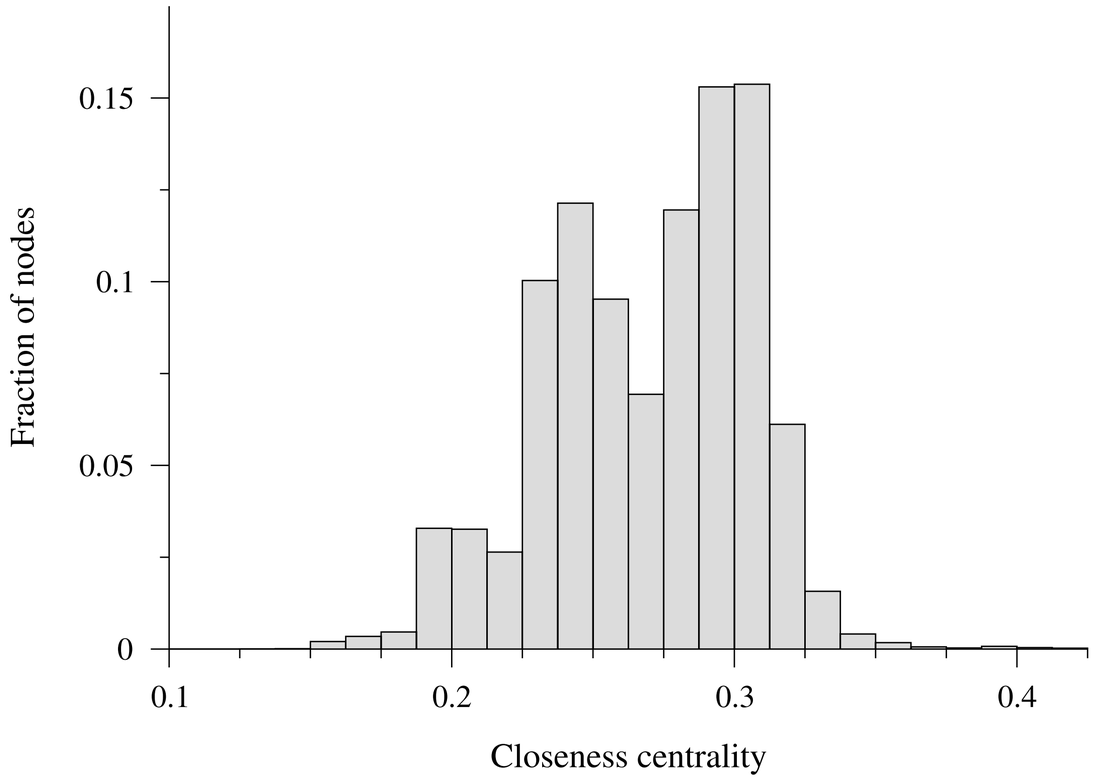


Distribution of Other Centrality Measures

Closeness centrality is an exception to that pattern. The measure is the reciprocal of the mean shortest-path distance from a node to all other reachable nodes.

The values of the mean distance typically have a small range, as they are limited by the diameter of the network, which is typically between 1 and $\log n$.

Hence, closeness centrality cannot have a broad distribution or a long tail.



Local Clustering Coefficient

The clustering coefficient quantifies the density of triads — i.e., strongly connected triangles of nodes — in a network.

Surprisingly, many large networks have a high clustering coefficient, i.e., there is typically a probability between about 10% and 60% that two neighbours of a node will be neighbours themselves.

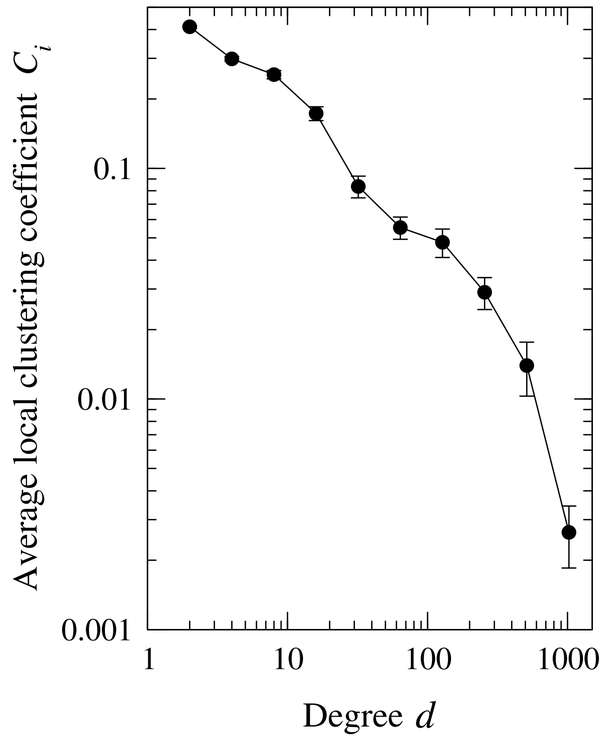
For example, a study on a large network of collaborations among physicists revealed a high clustering coefficient (0.45), which points to some underlying (non-random) pattern of selection of collaborators that gives rise to a high density of triangles.

Local Clustering Coefficient

Besides the network-level clustering coefficient, we can also study the distribution of (node-level) local clustering coefficient (the fraction of pairs of neighbours of node *i* that are themselves neighbours):

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

Interestingly, on average nodes with high degree tend to have low local clustering. E.g., looking at Internet nodes, their average local clustering coefficient and their degree d, we notice an inverse relation.



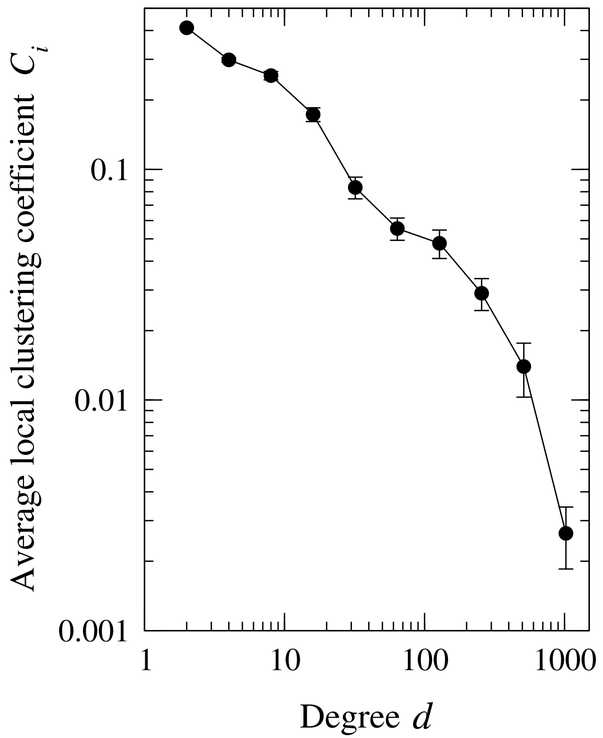
Local Clustering Coefficient

Besides the network-level clustering coefficient, we can also study the distribution of (node-level) local clustering coefficient (the fraction of pairs of neighbours of node *i* that are themselves neighbours):

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

An explanation of that phenomenon is that nodes tend to aggregate and connect internally within their “groups”.

Hence, in networks showing this behaviour, nodes that belong to small groups are constrained to have low degree but at the same time their local clustering coefficient tend to be larger because each group, being mostly detached from the rest of the network, boosts their internal clustering coefficient



Cohesion

The term “cohesion” indicates the likelihood of nodes being connected to each other. Notably, cohesion (the measure) does not indicate social aggregation — e.g., in a “hate” network a high network cohesion implies *less* social cohesion.

The simplest measure of cohesion is **density**, i.e., the ratio between the number of ties in the network with respect to the total number of possible ties $n(n-1)/2$.

While simple, density cohesion is not very useful as an absolute measure, e.g., in a 10-person network, a node is likely to have ties with all 9 others. On the contrary, in a 1000-person network it is much more unlikely that an actor has anything close to 999 ties with the rest of the members.

To avoid the issue of comparing sensibly different networks over density alone, we can resort to a cohesion measure on the average degree of the network. This is obtained by calculating the **average of the degrees** *(number of ties) of each node (i.e., the row sums of the adjacency matrix)*.

Cohesion and Connectedness

When measuring cohesiveness, it can be useful to consider network subgroups, specifically, to think about cohesion as the number and size of components in a network.

The simplest of these is the size of the main component: the bigger the main component (in terms of nodes), the greater the global cohesion of the network.

When more than one component exist, we can look at the number of components in the network. If c is the number of components and n that of nodes, we can obtain the **component ratio** as $(c-1)/(n-1)$, which has maximum value 1 when every node is isolate and minimum 0 when there is just one component.

Unfortunately, the component ratio is too-blunt of a measure as networks that vary in density and average degree may have the same component ratio.

Connectedness

Connectedness is a more sensitive measure of cohesion defined as the proportion of pairs of nodes that can reach each other by a path of any length — or, alternatively, the proportion of pairs of nodes that are located in the same component.

The formula for connectedness in directed non-reflexive networks is

$$\frac{\sum_{i \neq j} r_{ij}}{n(n-1)}$$

Where r_{ij} is 1 when i and j are in the same component, 0 otherwise.

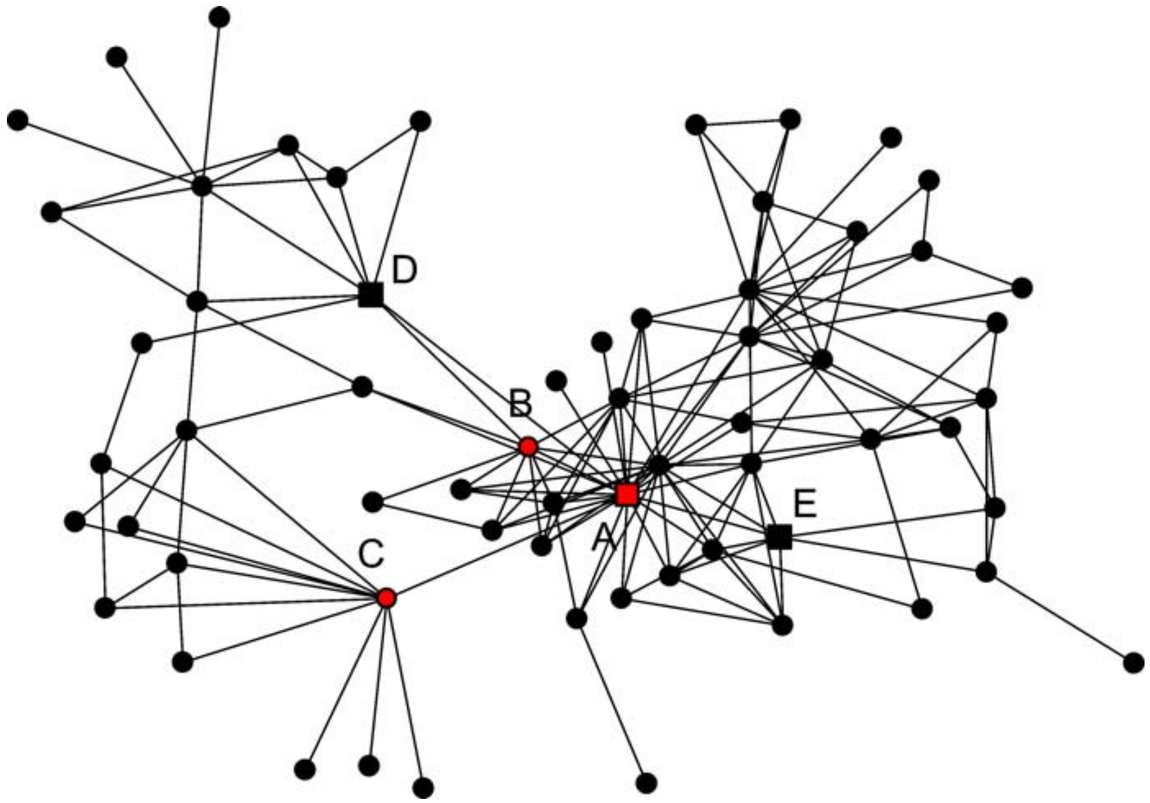
Inversi, we can define a cohesion measure, called **fragmentation**, as 1 minus connectedness, which gives the ratio of pairs of nodes that cannot reach each other by any means.

Connectedness

The typical usage of connectedness or fragmentation is in evaluating changes to a network either in reality or as part of a what-if simulation.

For example, if we are trying to prevent a terrorist organisation from coordinating attacks, we could figure out which key actors to arrest in order to maximally fragment the network.

A computer algorithm could search through the space of combinations of actors to determine a good set whose removal would maximally increase fragmentation.



Compactness

A variation on connectedness, called **compactness**, weights the paths connecting nodes inversely by their length:

$$\frac{\sum_{i \neq j} d_{ij}^{-1}}{n(n-1)}$$

Essentially, we replaced r_{ij} of connectedness with the reciprocal of the geodesic distance from i to j — with $d_{ij}^{-1} = 0$ when no path exists between i and j.

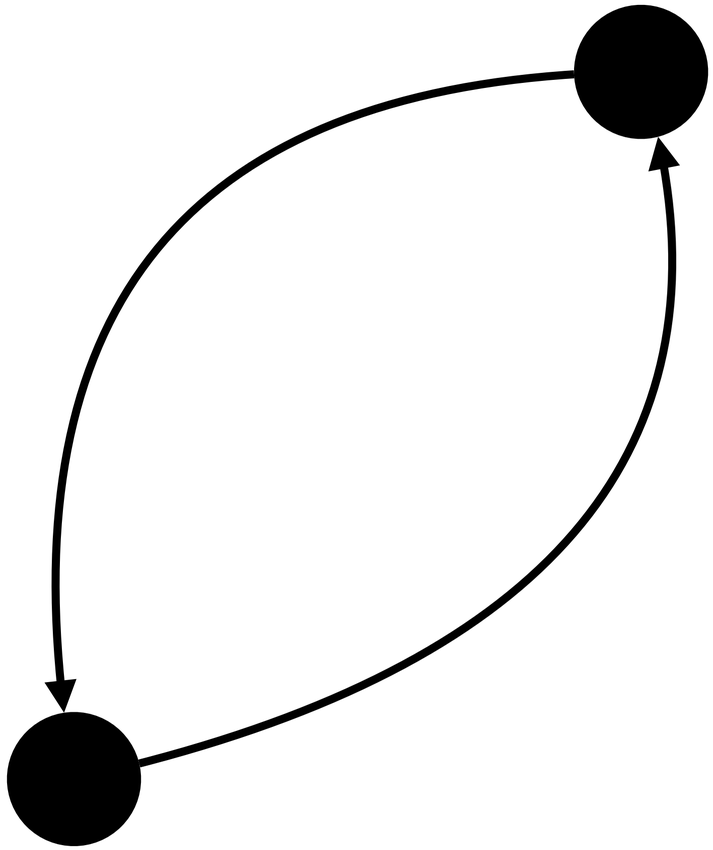
Intuitively, compactness considers network cohesion as a measure of how "easily" things can flow through it, accounting also for disconnected components — e.g., with compactness ≈ 1 nodes tend to be all connected and close (1-, 2-step paths).

Reciprocity

If ties are directed, we are often interested in the extent to which a tie from A to B is matched by one from B to A.

A simple measure of **reciprocity** is to count the number of reciprocated ties and divide these by the total number of ties.

A more refined measure is that of **symmetric pairs**, i.e., reciprocated ties together with the degenerate case where neither actors choose the other, that is, a reciprocated zero in the adjacency matrix.



Transitivity and Clustering Coefficients

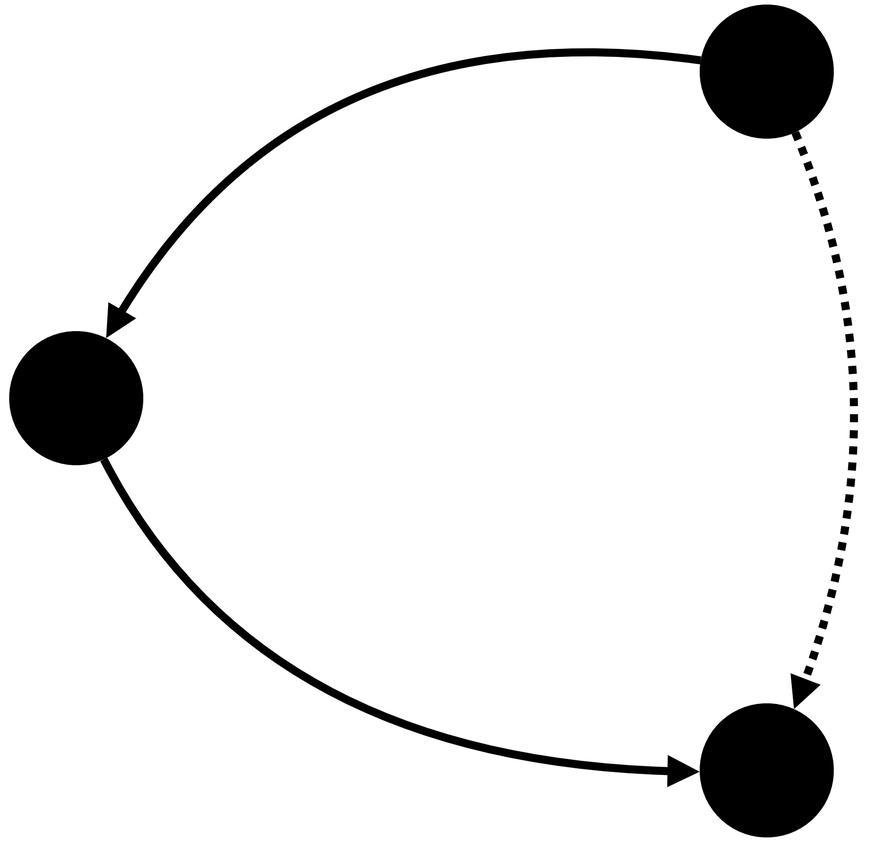
For many social relations we might expect that if A $\mathscr{R}$ B and B $\mathscr{R}$ C then A $\mathscr{R}$ C.

When this is the case we say that the triad is transitive. E.g., friends of friends are friends.

When networks have a high levels of transitivity, they assume a clustered structure.

To measure transitivity in *directed networks*, we count, across all possible triads i, j, and k, the proportion of triads for which $i \mathscr{R} j, j \mathscr{R} k$ and $i \mathscr{R} k$. Mathematically:

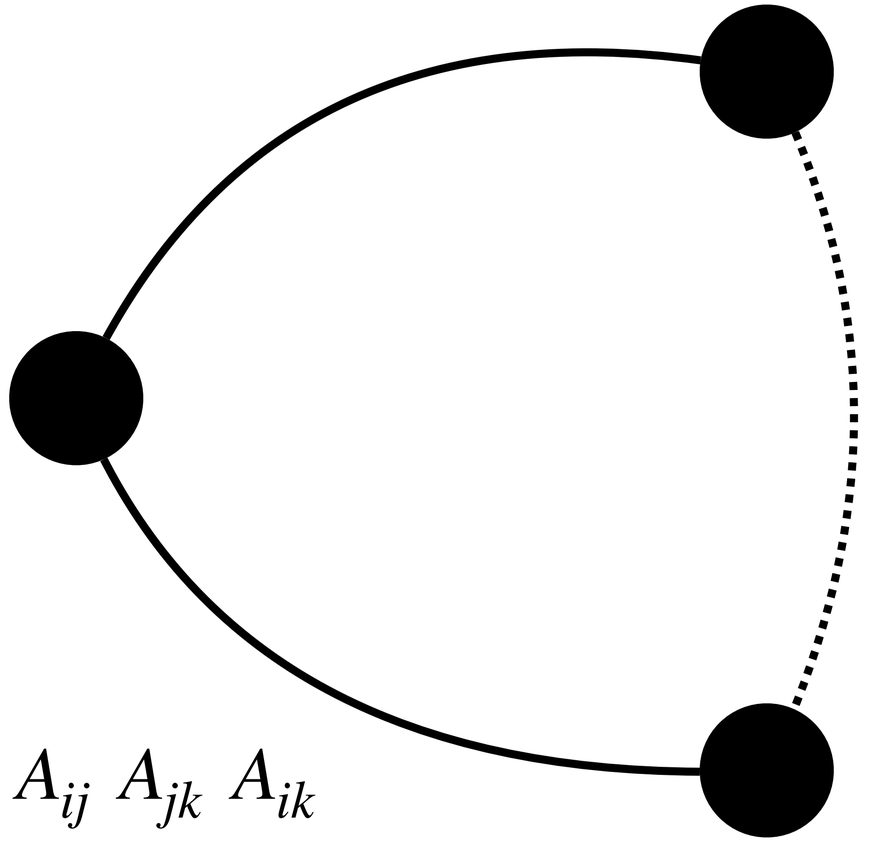
$$\frac{\sum_{i,j,k} A_{ij} A_{jk} A_{ik}}{\sum_{i,j,k} A_{ij} A_{jk}}$$



Transitivity and Clustering Coefficients

A declination of transitivity for undirected networks is the **clustering coefficient**, which captures the ratio between high- and low-density areas in a network.

Specifically, the most-used clustering coefficient is *the* **weighted overall clustering coefficient**, which, interestingly, mathematically corresponds to the formula for the transitivity coefficient of directed networks:



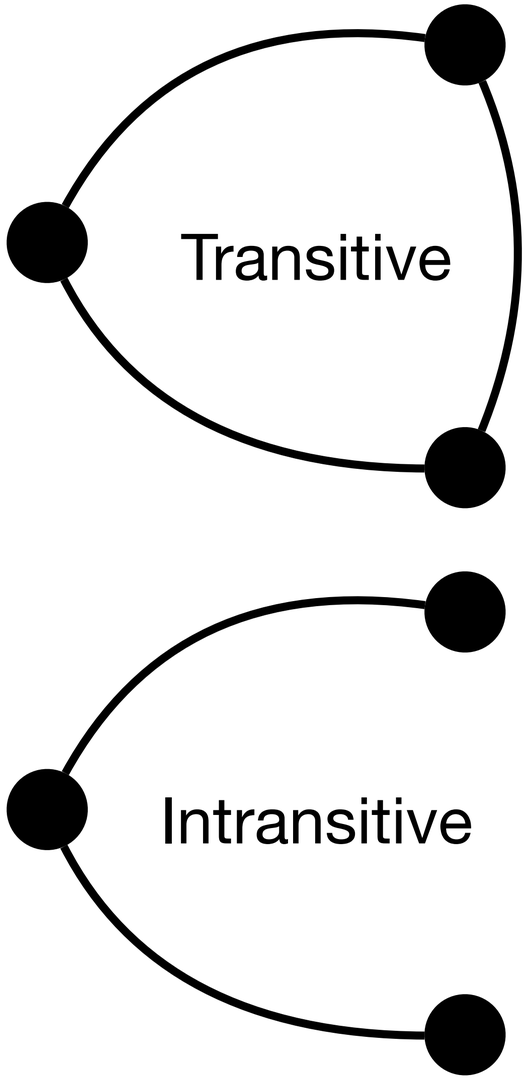
$$C_i = \frac{\text{number of pairs of neighbours of } i \text{ that are connected}}{\text{number of pairs of neighbours of } i} \Rightarrow \frac{\sum_{i,j,k} A_{ij}\ A_{jk}\ A_{ik}}{\sum_{i,j,k} A_{ij}\ A_{jk}}$$

Triad Census

Measuring transitivity involves counting the occurrences of at least two triadic configurations, which are labeled “transitive” and “intransitive”.

One measure of transitivity is the number of transitive triads divided by the number of transitive plus intransitive triads.

However, there are many other triadic configurations which could be used to characterise a network.



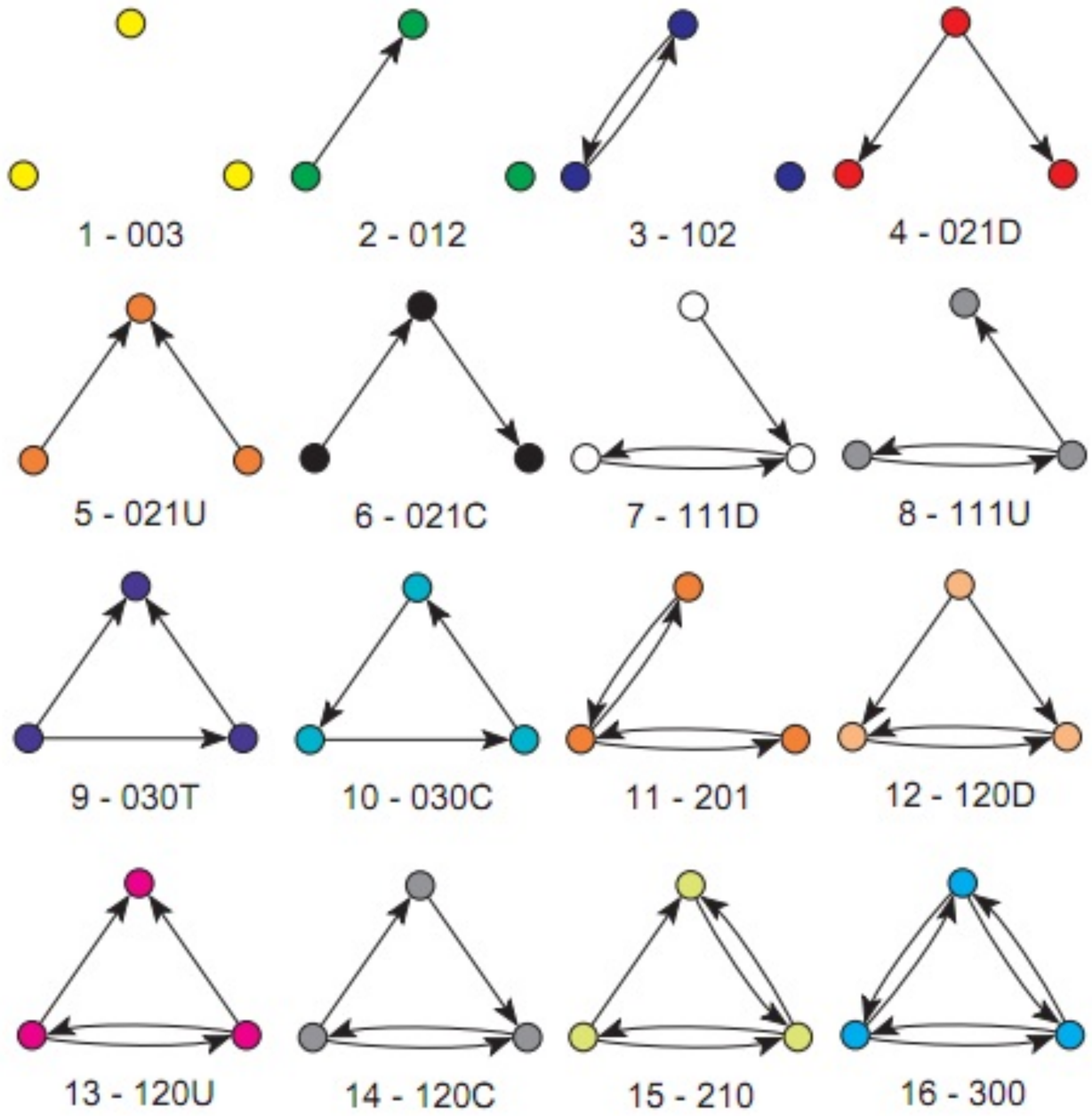
Triad Census

Specifically, for directed graphs we find 16 possible configurations, labelled following the MAN convention:

- Mutual, i.e., dyads with reciprocated ties;
- Asymmetric, i.e., dyads with unreciprocated ties;
- Null, i.e., dyads with no tie;

Where the label of a triad corresponds to the number of Ms, As, and Ns of the triad, e.g., 003 is a triad that has no mutual dyads (0), no asymmetric dyads (0), and has three unrelated (null) nodes (3).

Variants stand for **D**ownward, **U**pward, **C**yclic, and **T**ransitive



Triad Census

As an example, let us look at the triad census in a food web (who-eats-who) during the seasons, where nodes are species (or thereof aggregations).

In a food web, the transitive triad (030T), represents “omnivores”. That is, a species A eats species B, which eats C, but A also eats C, so A is eating at two separate levels of the food chain. A triad containing a mutual dyad, such as 120, reflects a pair of species that eat each other. This is not as rare as it sounds, but could also be due to the aggregation of different species into a single node.

Triad	Spring	Summer	Fall	Winter
003	4487	4359	4539	4906
012	1937	2001	1884	1663
102	75	71	88	118
021D	115	136	119	88
021U	259	300	273	180
021C	156	153	113	67
111D	25	27	44	37
111U	14	13	11	13
030T	46	54	46	39
030C	7	4	0	0
201	0	0	1	1
120D	8	6	7	8
120U	7	8	7	9
120C	1	5	5	5
210	3	3	3	6
300	0	0	0	0

Triad Census

A pattern we can observe is that in warmer seasons, we have many triads that begin with 0, meaning they have no mutual dyads.

Contrarily, in colder seasons, there are triads that have 1s or even 2s as the first number (Mutual). These are triads in which there are pairs that eat each other.

One explanation is that when the weather is warmer, there are more species available and there is no need to resort to reciprocal trophic interactions.

In winter, there is a kind of contraction of the ecosystem, with less variety available and more reciprocal interactions.

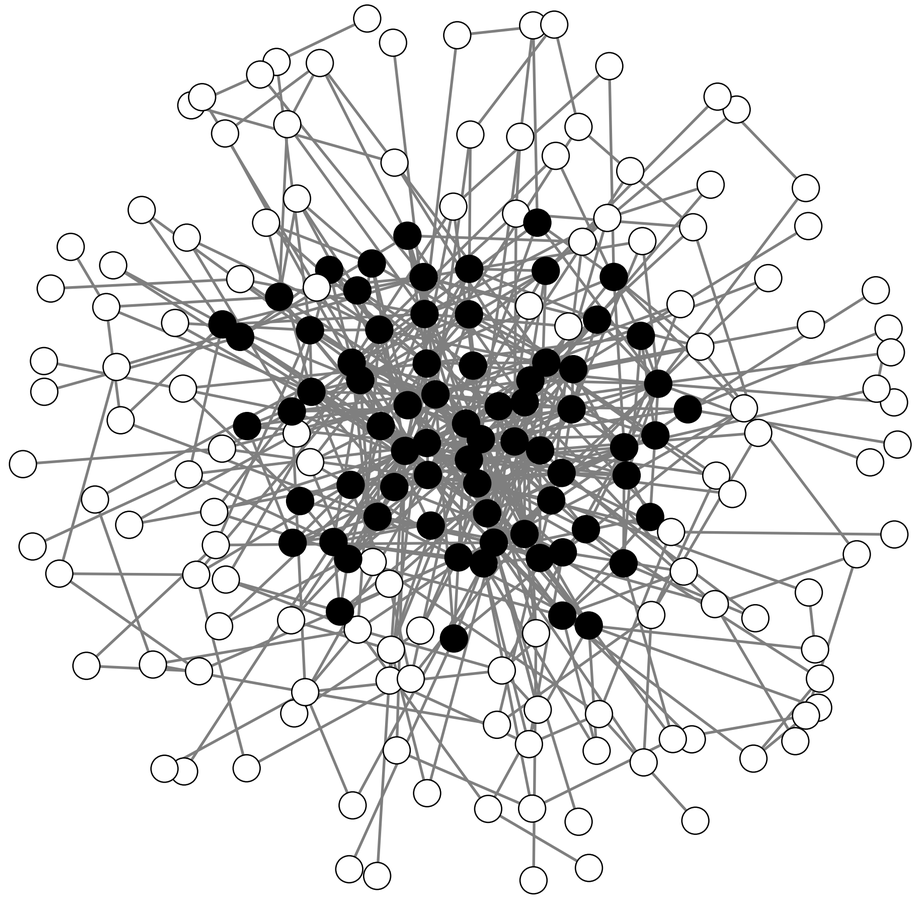
Triad	Spring	Summer	Fall	Winter
003	4487	**4359**	4539	**4906**
012	1937	**2001**	1884	**1663**
102	75	**71**	88	**118**
021D	115	**136**	119	**88**
021U	259	**300**	273	**180**
021C	**156**	153	113	**67**
111D	**25**	27	**44**	37
111U	**14**	13	**11**	13
030T	46	**54**	46	**39**
030C	**7**	4	**0**	**0**
201	**0**	**0**	**1**	**1**
120D	**8**	**6**	7	**8**
120U	**7**	8	**7**	**9**
120C	**1**	**5**	**5**	**5**
210	**3**	**3**	**3**	**6**
300	0	0	0	0

Centralisation and Core-periphery Indices

Centralisation refers to the extent a network is dominated by a single node. A maximally centralised network looks like a star: the node at the center of the network has ties to all other nodes, and no other ties exist.

More in general, we can measure the division of a network between a densely-connected core and a loosely-connected periphery.

One way to think of core-periphery structures is in terms of the average probabilities of edges within and between these two groups of nodes.

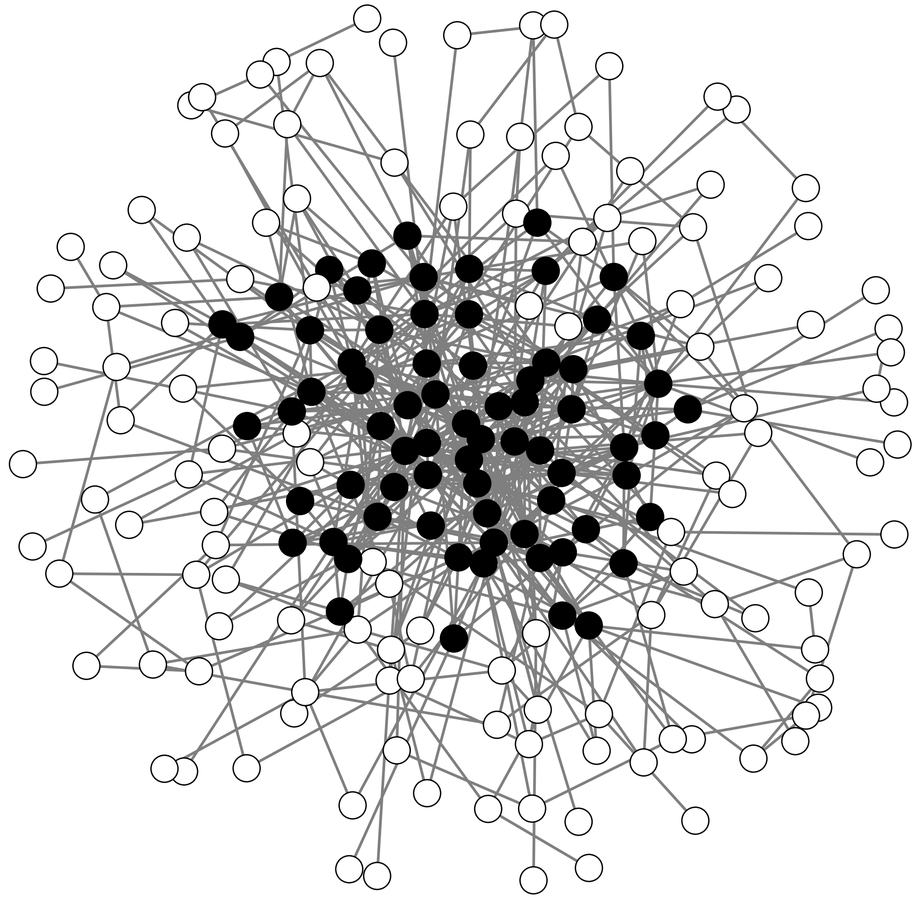


Centralisation and Core-periphery Indices

A simple method for finding the core-periphery structure assumes that the nodes in the core have higher degree than the nodes in the periphery and divide the nodes according to degree. While simple, the results returned by more sophisticated methods do not differ too much from this rudimentary degree-based division.

Another method is to find the *k*-cores of the network—a *k*-core is a group of nodes that each has connections to at least *k* other members of the group—"slicing" the network into different, nested layers.

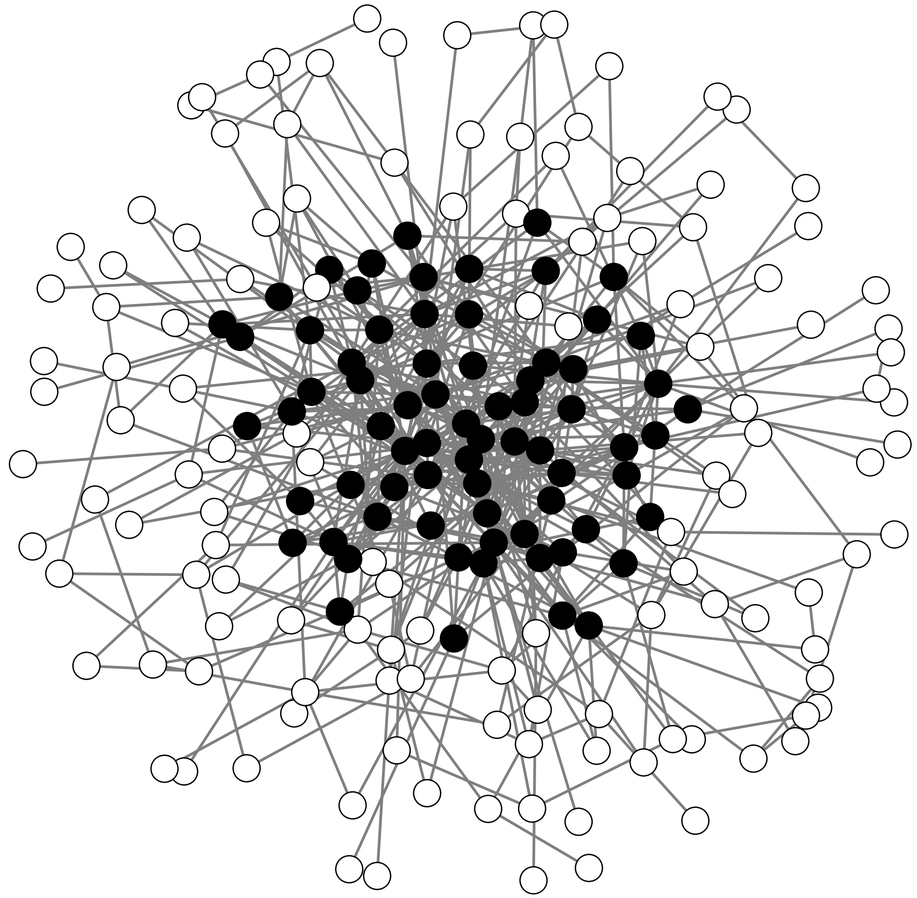
In both cases, core and peripheries can be multi-layered or dichotomised.



Centralisation and Core-periphery Indices

A more refined method for detecting dichotomic core–periphery structures relies on **finding the division** into core and periphery, defined by a value g_i, such **that** it **minimises** a measure ρ that calculates the difference between the number of edges in the periphery and the expected number of such edges if placed at random (simplified formula):

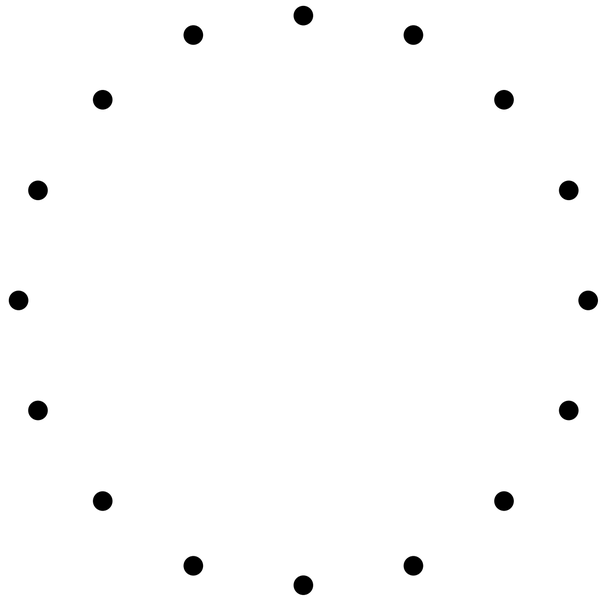
$$\rho = \frac{\sum_{ij} (A_{ij} - p_{ij})\ g_i\ g_j}{2} \quad \text{with} \quad g_k = \begin{cases} 0 & \text{if } k \in \text{core} \\ 1 & \text{otherwise} \end{cases}$$

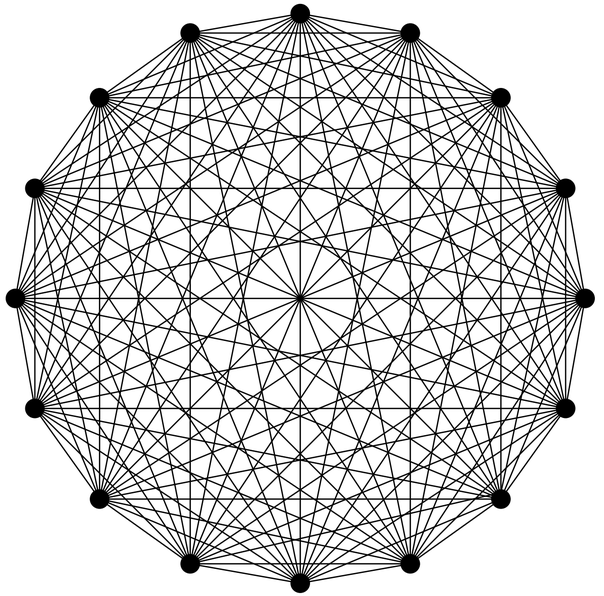


and p_{ij} equal to the average probability of the same of edges, if placed at random.

Random Graphs

A *random graph* is a model network in which the values of certain properties of the network are fixed, but the network is, in other respects, random.





Random Graphs

A *random graph* is a model network in which the values of certain properties of the network are fixed, but the network is, in other respects, random.

One of the simplest examples of a random graph is the one where we fix only the number of nodes *n* and the number of edges *m,* i.e., we choose *m* distinct pairs of nodes uniformly at random from all possible pairs and connect them with an edge. This model is often referred to by its mathematical name $G(n, m)$.

More specifically, we can define a random graph model as a family of networks defined by a probability distribution:

$$P(G) = \frac{1}{\binom{\binom{n}{2}}{m}}$$

pairs of nodes between which we could place an edge

ways of placing the *m* edges

Random Graphs

A special family of random graphs is that of $G(n, p)$, where we fix not the number of edges but the *probability* of edges between nodes, so that we have *n* nodes, but we place an edge between each distinct pair with independent probability p.

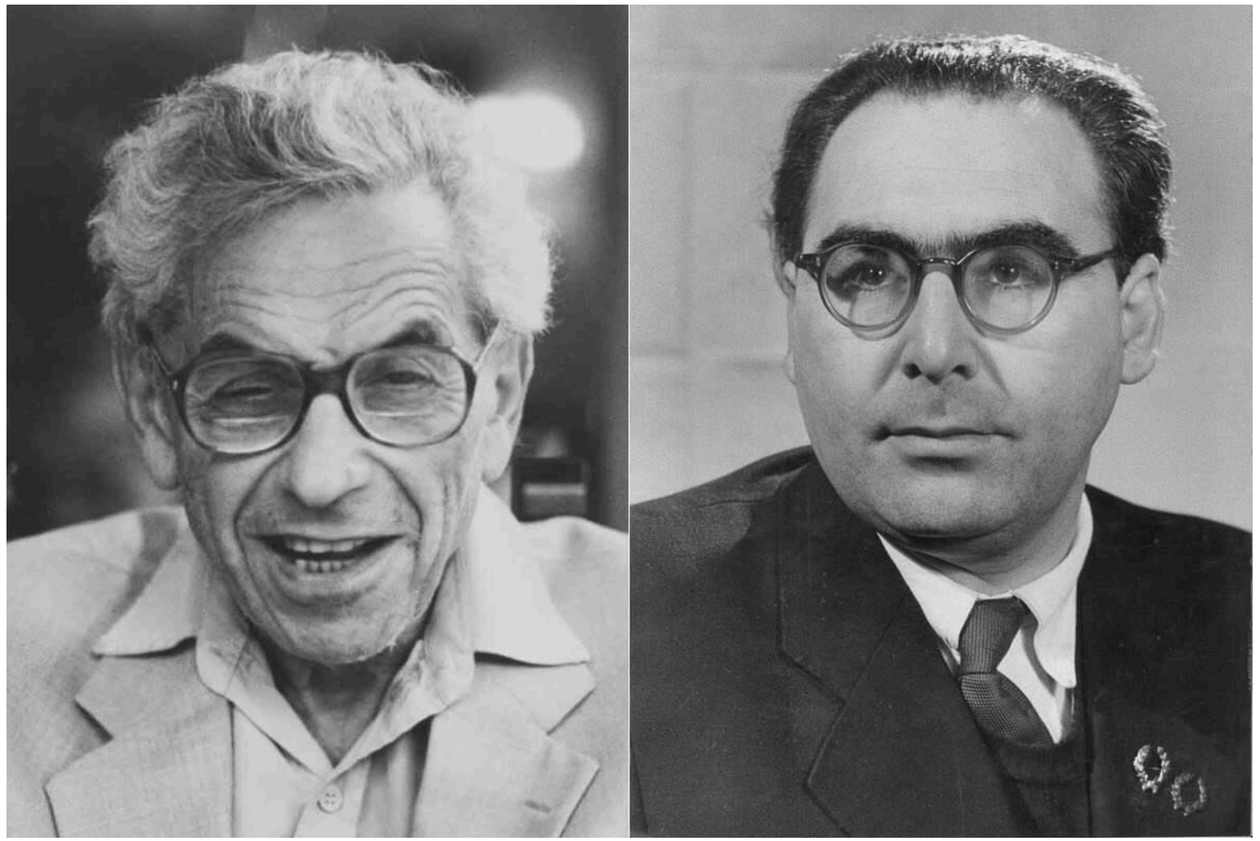
In this model the number of edges is not fixed.

$G(n, p)$ is the ensemble of simple networks with n nodes in which each network G appears with probability defined by the distribution $P(G) = p^m(1-p)^{\binom{n}{2}-m}$

$G(n, p)$ is most-closely associated with the names of Paul Erdős and Alfréd Rényi, who published a celebrated series of papers about the model in the late 1950s and early 1960s.

This is why it is frequent to find the model referred to as the “Erdős–Rényi model” or the “Erdős–Rényi random graph” or simply ER random graphs (as in UCINET).

Random Graphs



What makes the $G(n, p)$ family important is their capacity to simulate the tendency of real-world networks to become sparser and sparser the larger they grow—formally, their average degree grows slower than their size.

This is captured by the degree distribution of $G(n, p)$, which for smaller values of n follows a binomial Bernoullian distribution while for larger values approximates a Poissonian one.

The family is so fundamental to graph theory that ER graphs are frequently simply called “random graphs” and, although $G(n, p)$ fails to capture many other features of real networks, it is the main reference for random networks in network measures, besides being instrumental to explore graph theory in general.

Small-worldness and random graphs

One usage of random graphs is in the formal definition of small-worldness, i.e., the likelihood that a given network presents a small-world configuration calculated as $\sigma > 1$.

The calculation of σ is performed in three steps:

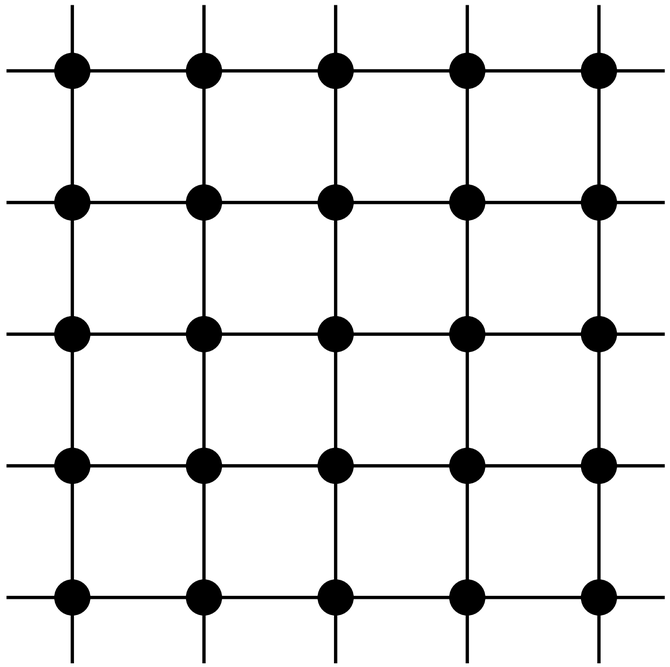
1. we calculate the ratio between the clustering coefficient in the network and the clustering coefficient of an ER random graph with the same size the network.
2. we calculate the ratio between the average path length in the network and the average path length of the random graph from 1.
3. we calculate the ratio between 1. and 2.

$$\sigma = \frac{\frac{C}{C_r}}{\frac{L}{L_r}}$$

Small-worldness, regular and random graphs

However, σ tends to be sensible to network size and density. This led to the definition of an additional measure, ω, which estimates the tendency of a network to resemble either a random or a regular one.

To define ω, we first need to introduce **regular graphs**, as another family of graphs where all nodes have the same degree. These regular graphs are also called **lattices**, since we can visualise these as grids or meshes of nodes where the same connection pattern repeats over the whole network.



Small-worldness, regular and random graphs

To calculate ω, we consider both a random graph and a regular one. We calculate ω in three steps:

where C and L are respectively the average clustering coefficient and average shortest path length of G.

1. we calculate the ratio between the average shortest path of the ER random graph with the same size the network and the average shortest path of the network;
2. we calculate the ratio between the average clustering coefficient of the network and the average clustering coefficient of a regular network of the same size;
3. we calculate the difference between 1. and 2.

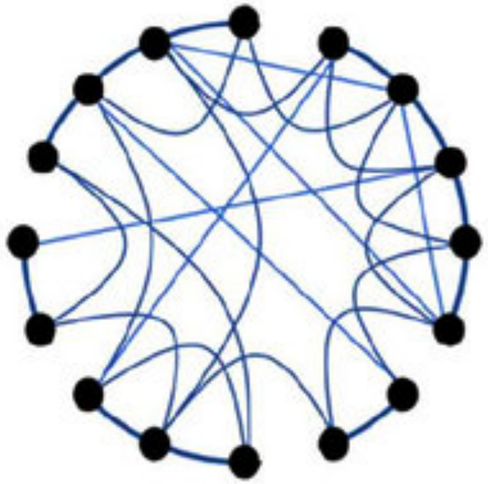
$$\omega = \frac{L_r}{L} - \frac{C}{C_l}$$

For $\omega < 0$ the network is closer to a lattice, for $\omega > 0$ it is closer to a random graph, for $\omega \approx 0$ the network approximates a small-work one.

Visual Recap

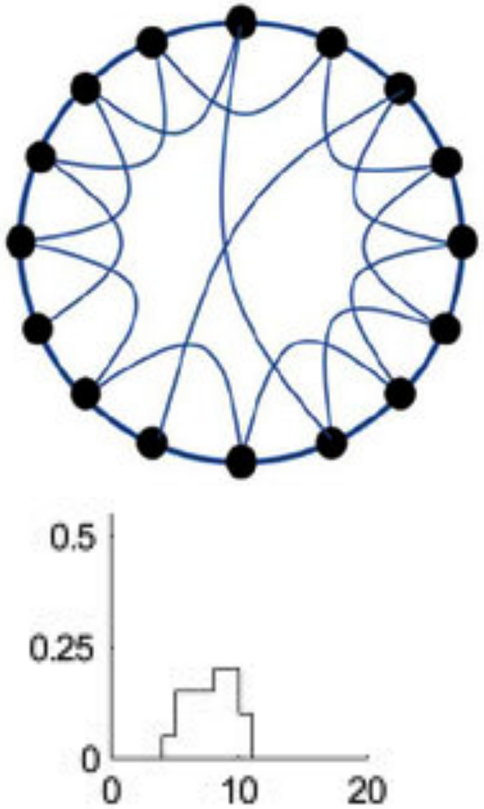
visualisation

degree distribution



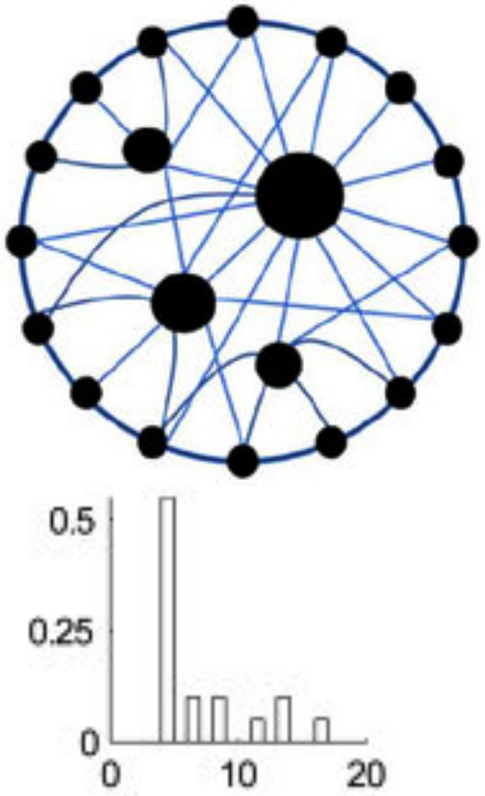
Bernoullian-to-Poissonian degree distribution, depending on n

ER random



Short average path length ($\ell \propto \log n$) and high local clustering ($\sigma > 1, \omega \approx 0$)

Small-world



Degree distribution follows a power law ($2 \leq \alpha \leq 3$)

Scale-free