

Statistical Inference for Networks

Graduate Lectures

HILARY TERM 2009

PROF. GESINE REINERT

Overview

1: *Network summaries.* What are networks? Some examples from social science and from biology. The need to summarise networks. Clustering coefficient, degree distribution, shortest path length, motifs, between-ness, second-order summaries. Roles in networks, derived from these summary statistics, and modules in networks. Directed and weighted networks. The choice of summary should depend on the research question.

2: *Models of random networks.* Models would provide further insight into the network structure. Classical Erdős-Renyi (Bernoulli) random graphs and their random mixtures, Watts-Strogatz small worlds and the modification by Newman, Barabasi-Albert scale-free networks, exponential random graph models.

3: *Fitting a model: parametric methods.* Deriving the distribution of summary statistics. Parametric tests based on the theoretical distribution of the summary statistics (only available for some of the models).

4: *Statistical tests for model fit: nonparametric methods.*

Quantile-quantile plots and other visual methods. Monte-Carlo tests based on shuffling edges with the number of edges fixed, or fixing the node degree distribution, or fixing some other summary. The particular issue of testing for power-law dependence. Subsampling issues. Tests carried out on the same network are not independent.

Suggested reading

1. U. Alon: *An Introduction to Systems Biology Design Principles of Biological Circuits*. Chapman and Hall 2007.
2. S.N. Dorogovtsev and J.F.F. Mendes: *Evolution of Networks*. Oxford University Press 2003.
3. R. Durrett: *Random Graph Dynamics*. Cambridge University Press 2007.
4. W. de Nooy, A. Mrvar and V. Bagatelj. *Exploratory Social Network Analysis with Pajek*. Cambridge University Press 2005.
5. S. Wasserman and K. Faust: *Social Network Analysis*. Cambridge University Press 1994.
6. D. Watts: *Small Worlds*. Princeton University Press 1999.

Further reading:

1. L. da F. Costa, F.A. Rodrigues, P.R. Villas Boas, G. Travieso (2007). Characterization of complex networks: a survey of measurements. *Advances in Physics* 56, Issue 1 January 2007, 167 - 242.
2. J.-J. Daudin, F. Picard, S. Robin (2006). A mixture model for random graphs. Preprint.
3. O. Frank and D. Strauss (1986). Markov graphs. *Journal of the American Statistical Association* **81**, 832-842.
4. K. Nowicky and T. Snijders (2001). Estimation and Prediction for Stochastic Blockstructures. *Journal of the American Statistical Association* **455**, Vol. 96, pp. 1077-1087.
5. S. Milgram (1967). The small world problem. *Psychology Today* **2**, 60–67.

6. A.D. Barbour and G. Reinert (2001). Small Worlds. *Random Structures and algorithms* 19, 54 - 74.
7. A.D. Barbour and G. Reinert (2006). Discrete small world networks. *Electronic Journal of Probability* 11, 12341283.
8. G. Barnard (1963). Contribution to the discussion of Bartlett's paper. *J. Roy. Statist. Soc. B*, 294.
9. F. Chung and L. Lu (2001). The diameter of sparse random graphs. *Advances in Applied Math.* 26, 257–279.
10. M. Dwass (1957). Modified randomization tests for nonparametric hypotheses. *Ann. Math. Stat.* **28**, 181–187. A. Fronczak, P. Fronczak and Janusz A. Holyst (2003). Mean-field theory for clustering coefficients in Barabasi-Albert networks *Phys. Rev. E* 68, 046126.
11. A. Fronczak, P. Fronczak, Janusz A. Holyst (2004). Average path length in random networks *Phys. Rev. E* 70, 056110.

12. P. Hall and D.M. Titterton (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *J. Roy. Statist. Soc. B*, 459.
13. K. Lin (2007). Motif counts, clustering coefficients, and vertex degrees in models of random networks. D.Phil. dissertation, Oxford.
14. F. Marriott (1979). Barnard's Monte Carlo tests: how many simulations? *Appl. Statist.* 28, 75–77.
15. M.E.J. Newman (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 323351.
16. M. P. H. Stumpf, C.Wiuf and R. M. May (2005). Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc Natl Acad Sci U S A*. 2005 March 22; 102(12): 4221 - 4224.

17. C. J. Anderson, S. Wasserman, B. Crouch (1999). A p^* primer: logit models for social networks. *Social Networks* 21, 37–66.
18. P. Chen, C. Deane, G. Reinert (2007) A statistical approach using network structure in the prediction of protein characteristics. *Bioinformatics*, Vol 23, 17, 2314-2321.
19. P. Chen, C. Deane, G. Reinert (2008) Predicting and validating protein interactions using network structure. Submitted.
20. J.-D. Han, N. Bertin, T. Hao, D. S. Goldberg, G.F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430, 8893.
21. Handcock, M.S., Raftery, A.E., and Tantrum, J.M. 2007. Model-based clustering for social networks. *Journal of the Royal Statistical Society*. **170**:122.

22. P. W. Holland, S. Leinhardt (1981). An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76, 33-65.
23. F. Kepes ed. (2007). *Biological Networks. Complex Systems and Interdisciplinary Science Vol. 3*. World Scientific, Singapore.
24. Newman, M.E.J. and Girvan, M. 2004. Finding and evaluating community structure in networks. *Physical Review E*. 69 026113.
25. G. Palla, A.-L. Barabasi, T. Vicsek (2007). Quantifying social group evolution. *Nature* 446, 664–667.
26. F. Picard, J.-J. Daudin, M. Koskas, S. Schbath, S. Robin (2008). Assessing the exceptionality of network motifs. *Journal of Computational Biology* 15, 1–20.
27. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L.

Barabasi (2002). Hierarchical Organization of Modularity in Metabolic Networks. *Science* 297, 1551.

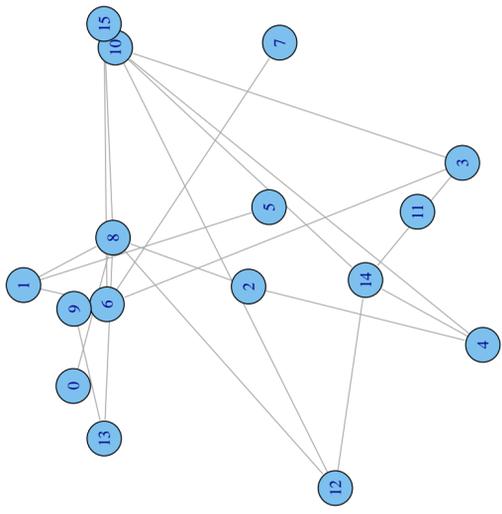
28. G.L. Robins, T.A.B. Snijders, P. Wang, M. Handcock, P. Pattison (2007). Recent developments in exponential random graph (p^*) models for social networks. *Social Networks* 29, 192-215 (2007). <http://www.stats.ox.ac.uk/~snijders/siena/RobinsSnijdersWangHandcockPattison2007.pdf>
29. G.L. Robins, P. Pattison, Y. Kalisha, D. Lushera (2007). An introduction to exponential random graph (p^*) models for social networks. *Social Networks* 29, 173-191.

1 Network summaries

1.1 What are networks?

Networks are just graphs. Often one would think of a network as a connected graph, but not always. In these lectures we shall use *network* and *graph* interchangeably.

Here are some examples of networks (graphs).



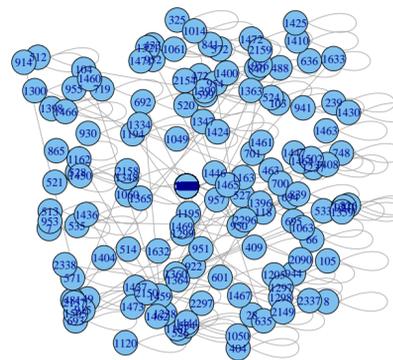
12

Marriage relations between Florentine families.

8: Medici

14: Strozzi

Yeast: A plot of a connected subset of Yeast protein interactions.



Networks arise in a multitude of contexts, such as

- metabolic networks
- protein-protein interaction networks
- spread of epidemics
- neural network of *C. elegans*
- social networks
- collaboration networks (Erdős numbers ...)
- Membership of management boards
- World Wide Web
- power grid of the Western US

The study of networks has a long tradition in social science, where it is called *Social Network Analysis*; see also Krista Gile's talks. The networks under consideration are typically fairly small. In contrast, starting at around 1997, statistical physicists have turned their attention to large-scale properties of networks. Our lectures will try to get a glimpse on both approaches.

Research questions include

- How do these networks work? Where could we best manipulate a network in order to prevent, say, tumor growth?
- How do biological networks evolve? Could mutation affect whole parts of the network at once?
- How similar are networks? If we study some organisms very well, how much does that tell us about other organisms?
- How are networks interlinked?
- What are the building principles of these networks? How is resilience achieved, and how is flexibility achieved? Could we learn from real-life networks to build man-made efficient networks?

From a statistical viewpoint, questions include

- How to best describe networks?
- How to infer characteristics of nodes in the network?
- How to infer missing links, and how to check whether existing links are not false positives
- How to compare networks from related organisms?
- How to predict functions from networks?
- How to find relevant sub-structures of a network?

Statistical inference relies on the assumption that there is some randomness in the data. Before we turn our attention to modelling such randomness, let's look at how to describe networks, or graphs, in general.

1.2 What are graphs?

A *graph* consists of *nodes* (sometimes also called *vertices*) and *edges* (sometimes also called *links*). We typically think of the nodes as actors, or proteins, or genes, or metabolites, and we think of an edge as an interaction between the two nodes at either end of the edge. Sometimes nodes may possess characteristics which are of interest (such as structure of a protein, or function of a protein). Edges may possess different weights, depending on the strength of the interaction. For now we just assume that all edges have the same weight, which we set as 1.

Mathematically, we abbreviate a graph G as $G = (V, E)$, where V is the set of nodes and E is the set of edges. We use the notation $|S|$ to denote the number of elements in the set S . Then $|V|$ is the number of nodes, and $|E|$ is the number of edges in the graph G . If u and v are two nodes and there is an edge from u to v , then we write that $(u, v) \in E$, and we say that v is a *neighbour* of u .

If both endpoints of an edge are the same, then the edge is a *loop*. For now we exclude self-loops, as well as multiple edges between two nodes.

Edges may be *directed* or *undirected*. A *directed graph*, or *digraph*, is a graph where all edges are directed. The *underlying* graph of a digraph is the graph that results from turning all directed edges into undirected edges. Here we shall mainly deal with undirected graphs.

Two nodes are called *adjacent* if they are joined by an edge. A graph can be described by its *adjacency matrix* $A = (a_{u,v})$. This is a square $|V| \times |V|$ matrix. Each entry is either 0 or 1;

$$a_{u,v} = 1 \text{ if and only if } (u, v) \in E.$$

As we assume that there are no self-loops, all elements on the diagonal of the adjacency matrix are 0. If the edges of the graph are undirected, then the adjacency matrix will be symmetric.

The adjacency matrix entries tell us for every node v which nodes are within distance 1 of v . If we take the matrix product $A^2 = A \times A$, the entry for (u, v) with $u \neq v$ would be

$$a^{(2)}(u, v) = \sum_{w \in V} a_{u,w} a_{w,v}.$$

If $a^{(2)}(u, v) \neq 0$ then u can be reached from v within two steps; u is within distance 2 of v . Higher powers can be interpreted similarly.

A *complete* graph is a graph such that every pair of nodes is joined by an edge. The adjacency matrix has entry 0 on the diagonal, and 1 everywhere else.

A *bipartite* graph is a graph where the node set V is decomposed into two disjoint subsets, U and W , say, such that there are no edges between any two nodes in U , and also there are no edges between any two nodes in W ; all edges have one endpoint in U and the other endpoint in W . An example is a network of co-authorship and articles; U could be the set of authors, W the set of articles, and an author is connected to an article by an edge if the author is a co-author of that article. The adjacency matrix A can then be arranged such that it is of the form

$$\begin{bmatrix} 0 & A_1 \\ A_2 & 0 \end{bmatrix}.$$

1.3 Network summaries

The *degree* $deg(v)$ of a node v is the number of edges which involve v as an endpoint. The degree is easily calculated from the adjacency matrix A ;

$$deg(v) = \sum_u a_{u,v}.$$

The *average degree* of a graph is then the average of its node degrees.

(For directed graphs we would define the *in-degree* as the number of edges directed at the node, and the *out-degree* as the number of edges that go out from that node.)

The *clustering coefficient* of a node v is, intuitively, the proportion of its "friends" who are friends themselves. Mathematically, it is the proportion of neighbours of v which are neighbours themselves. In adjacency matrix notation,

$$C(v) = \frac{\sum_{u,w \in V} a_{u,v} a_{w,v} a_{u,w}}{\sum_{u,w \in V} a_{u,v} a_{w,v}}.$$

The (*average*) *clustering coefficient* is defined as

$$C = \frac{1}{|V|} \sum_{v \in V} C(v).$$

Note that

$$\sum_{u,w \in V} a_{u,v} a_{w,v} a_{u,w}$$

is the number of triangles involving v in the graph. Similarly,

$$\sum_{u,w \in V} a_{u,v} a_{w,v}$$

is the number of $\mathcal{2}$ -stars centred around v in the graph. The clustering coefficient is thus the ratio between the number of triangles and the number of 2-stars. The clustering coefficient describes how "locally dense" a graph is. Sometimes the clustering coefficient is also called the *transitivity*.

The clustering coefficient in the Florentine family example is 0.1914894; the average clustering coefficient in the Yeast data is 0.1023149.

In a graph a *path* from node v_0 to node v_n is an alternating sequence of nodes and edges, $(v_0, e_1, v_1, e_2, \dots, v_{n-1}, e_n, v_n)$ such that the endpoints of e_i are v_{i-1} and v_i , for $i = 1, \dots, n$. A graph is called *connected* if there is a walk between any pair of nodes in the graph, otherwise it is called *disconnected*. The *distance* $\ell(u, v)$ between two nodes u and v is the length of the shortest path joining them. This path does not have to be unique.

We can calculate the distance $\ell(u, v)$ from the adjacency matrix A as the smallest power p of A such that the (u, v) -element of A^p is not zero.

In a connected graph, the *average shortest path length* is defined as

$$\ell = \frac{1}{|V|(|V| - 1)} \sum_{u \neq v \in V} \ell(u, v).$$

The average shortest path length describes how "globally connected" a graph is.

Example: *H. Pylori* and Yeast protein interaction network comparison:

	n	ℓ	C
H.Pylori	686	4.137637	0.016
Yeast	2361	4.376182	0.1023149

Node degree, clustering coefficient, and shortest path length are the most common summaries of networks. Other popular summaries, to name but a few, are: the *between-ness of an edge* counts the proportion of shortest paths between any two nodes which pass through this edge. Similarly, the *between-ness of a node* is the proportion of shortest paths between any two nodes which pass through this node. The *connectivity* of a connected graph is the smallest number of edges whose removal results in a disconnected graph.

In addition to considering these general summary statistics, it has proven fruitful to describe networks in terms of *motifs*; these are building- block patterns of networks such as a feed-forward loop, see the book by *Alon*. Here we think of a motif as a subgraph with a fixed number of nodes and with a given topology. In biological networks, it turns out that motifs seem to be conserved across species. They seem to reflect functional units which combine to regulate the cellular behaviour as a whole.

The decomposition of *communities* in networks, small subgraphs which are highly connected but not so highly connected to the remaining graph, can reveal some structure of the network. Identifying *roles* in networks singles out specific nodes with special properties, such as hub nodes, which are nodes with high degree.

Summaries based on spectral properties of the adjacency matrix.

If λ_i are the eigenvalues of the adjacency matrix A , then the spectral density of the graph is defined as

$$\rho(\lambda) = \frac{1}{n} \sum_i \delta(\lambda - \lambda_i),$$

where $\delta(x)$ is the delta function. For Bernoulli random graphs, if p is constant as $n \rightarrow \infty$, then $\rho(\lambda)$ converges to a semicircle.

The eigenvalues can be used to compute the k th moments,

$$M_k = \frac{1}{n} \sum_i (\lambda_i)^k = \frac{1}{n} \sum_{i_1, i_2, \dots, i_k} a_{i_1, i_2} a_{i_2, i_3} \cdots a_{i_{k-1}, i_k}.$$

The quantity nM_k is the number of paths returning to the same node in the graph, passing through k edges, where these paths may contain nodes that were already visited. Because in a tree-like graph a return path is only possible going back through already visited nodes, the presence of odd moments is an indicator for the presence of cycles in the graph.

The *subgraph centrality*

$$Sc_i = \sum_{k=0}^{\infty} \frac{(A^k)_{i,i}}{k!}$$

measures the "centrality" of a node based on the number of subgraphs in which the node takes part. It can be computed as

$$Sc_i = \sum_{j=1}^n v_j(i)^2 e^{\lambda_j},$$

where $v_j(i)$ is the i th element of the j th eigenvector.

The structure of a network is related to its reliability and speed of information propagation. If a random walk starts on node i going to node j , the probability that it goes through a given shortest path $\pi(i, j)$ between these vertices is

$$\mathcal{P}(\pi(i, j)) = \frac{1}{d(i)} \sum_{b \in \mathcal{N}(\pi(i, j))} \frac{1}{d(b) - 1},$$

where $d(i)$ is the degree of node i , and $\mathcal{N}(\pi(i, j))$ is the set of nodes in the path $\pi(i, j)$ excluding i and j . The *search information* is the total information needed to identify one of all the shortest paths between i and j ,

$$S(i, j) = -\log_2 \sum_{\pi(i, j)} \mathcal{P}(\pi(i, j)).$$

The above network summaries provide an initial go at networks. Specific networks may require specific concepts. In protein interaction networks, for example, there is a difference whether a protein can interact with two other proteins simultaneously (party hub) or sequentially (date hub). In addition, the research question may suggest other summaries. For example, in fungal networks, there are hardly any triangles, so the clustering coefficient does not make much sense for these networks.

Excursion: Milgram and the small world effect.

In 1967 the American sociologist Milgram reported a series of experiments of the following type. A number of people from a remote US state (Nebraska, say) are asked to have a letter (or package) delivered to a certain person in Boston, Massachusetts (such as the wife of a divinity student). The catch is that the letter can only be sent to someone whom the current holder knew on a first-name basis. Milgram kept track of how many intermediaries were required until the letters arrived; he reported a median of six; see for example [http : //www.ua.f.edu/northern/big_world.html](http://www.ua.f.edu/northern/big_world.html). This made him coin the notion of *six degrees of separation*, often interpreted as everyone being six handshakes away from the President. While the experiments were somewhat flawed (in the first experiment only 3 letters arrived), the concept of *six degrees of separation* has stuck.

2 Models of random networks

In order to judge whether a network summary is "unusual" or whether a motif is "frequent", there is an underlying assumption of randomness in the network.

Network data are subject to various errors, which can create randomness, such as

- There may be missing edges in the network. Perhaps a node was absent (social network) or has not been studied yet (protein interaction network).
- Some edges may be reported to be present, but that recording is a mistake. Depending on the method of determining protein interactions, the number of such *false positive* interactions can be substantial, of around 1/3 of all interactions.
- There may be transcription errors in the data.
- There may be bias in the data, some part of the network may have received higher attention than another part of the network.

Often network data are snapshots in time, while the network might undergo dynamical changes.

In order to understand mechanisms which could explain the formation of networks, mathematical models have been suggested.

2.1 Bernoulli (Erdős-Renyi) random graphs

The most standard random graph model is that of Erdős and Renyi (1959). The (finite) node set V is given, say $|V| = n$, and an edge between two nodes is present with probability p , independently of all other edges. As there are

$$\binom{n}{2} = \frac{n(n-1)}{2}$$

potential edges, the expected number of edges is then

$$\binom{n}{2}p.$$

Each node has $n - 1$ potential neighbours, and each of these $n - 1$ edges is present with probability p , and so the expected degree of a node is $(n - 1)p$. As the expected degree of a node is the same for all nodes, the average degree is $(n - 1)p$.

Similarly, the average number of triangles in the graph is

$$\binom{n}{3}p^3 = \frac{n(n-1)(n-2)}{6}p^3,$$

and the average number of 2-stars is

$$\binom{n}{3}p^2.$$

Thus, with a bit of handwaving, we would expect an average clustering coefficient of about

$$\frac{\binom{n}{3}p^3}{\binom{n}{3}p^2} = p.$$

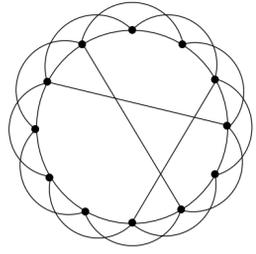
In a Bernoulli random graphs, your friends are no more likely to be friends themselves than would be a two complete strangers. This model is clearly not a good one for social networks. Below is an example from scientific collaboration networks (*N. Boccaro, Modeling Complex Systems, Springer 2004, p.283*). We can estimate p as the fraction of average node degree and $n - 1$; this estimate would also be an estimate of the clustering coefficient in a Bernoulli random graph.

Network	n	ave degree	C	$C_{Bernoulli}$
Los Alamos archive	52,909	9.7	0.43	0.00018
MEDLINE	1,520,251	18.1	0.066	0.000011
NCSTRL	11,994	3.59	0.496	0.0003

Also in real-world graphs often the shortest path length is much shorter than expected from a Bernoulli random graph with the same average node degree. The phenomenon of short paths, often coupled with high clustering coefficient, is called the *small world phenomenon*. Remember the Milgram experiments!

2.2 The Watts-Strogatz model

Watts and Strogatz (1998) published a ground-breaking paper with a new model for small worlds; the version currently most used is as follows. Arrange the n nodes of V on a lattice. Then hard-wire each node to its k nearest neighbours on each side on the lattice, where k is small. Thus there are nk edges in this hard-wired lattice. Now introduce random shortcuts between nodes which are not hard-wired; the shortcuts are chosen independently, all with the same probability.



If there are no shortcuts, then the average distance between two randomly chosen nodes is of the order n , the number of nodes. But as soon as there are just a few shortcuts, then the average distance between two randomly chosen nodes has an expectation of order $\log n$. Thinking of an epidemic on a graph - just a few shortcuts dramatically increase the speed at which the disease is spread.

It is possible to approximate the node degree distribution, the clustering coefficient, and the shortest path length reasonably well mathematically; we may come back to these approximations later.

While the Watts-Strogatz model is able to replicate a wide range of clustering coefficient and shortest path length simultaneously, it falls short of producing the observed types of node degree distributions. It is often observed that nodes tend to attach to "popular" nodes; popularity is attractive.

2.3 "The" Barabasi-Albert model

In 1999, Barabasi and Albert noticed that the actor collaboration graph and the World Wide Web had degree distributions that were of the type

$$\text{Prob}(\text{degree} = k) \sim Ck^{-\gamma}$$

for $k \rightarrow \infty$. Such behaviour is called *power-law behaviour*; the constant γ is called the *power-law exponent*. Subsequently a number of networks have been identified which show this type of behaviour. They are also called *scale-free random graphs*. To explain this behaviour, Barabasi and Albert introduced the *preferential attachment* model for network growth.

Suppose that the process starts at time 1 with 2 nodes linked by m (parallel) edges. At every time $t \geq 2$ we add a new node with m edges that link the new node to nodes already present in the network. We assume that the probability π_i that the new node will be connected to a node i depends on the degree $deg(i)$ of i so that

$$\pi_i = \frac{deg(i)}{\sum_j deg(j)}.$$

To be precise, when we add a new node we will add edges one at a time, with the second and subsequent edges doing preferential attachment using the updated degrees.

This model has indeed the property that the degree distribution is approximately power law with exponent $\gamma = 3$. Other exponents can be achieved by varying the probability for choosing a given node.

Unfortunately the above construction will not result in any triangles at all. It is possible to modify the construction, adding more than one edge at a time, so that *any* distribution of triangles can be achieved.

2.4 Erdős-Renyi Mixture Graphs

An intermediate model with not quite so many degrees of freedom is given by the Erdős-Renyi mixture model, also known as *latent block models* in social science (*Nowicky and Snijders (2001)*). Here we assume that nodes are of different types, say, there are L different types. Then edges are constructed independently, such that the probability for an edge varies only depending on the type of the nodes at the endpoints of the edge. *Robin et al* have shown that this model is very flexible and is able to fit many real-world networks reasonably well. It does not produce a power-law degree distribution however.

2.5 Exponential random graph (p^*) models

Networks have been analysed for "ages" in the social science literature, see for example the book by *Wasserman and Faust*. Here usually digraphs are studied, and typical research questions are

- Is there a tendency in friendship towards transitivity; are friends of friends my friends?
- What is the role of explanatory variables such as income on the position in the network?
- What is the role of friendship in creating behaviour (such as smoking)?
- Is there a hierarchy in the network?
- Is the network influenced by other networks for which the membership overlaps?

Exponential random graph (p^) models* model the whole adjacency matrix of a graph simultaneously, making it easy to incorporate dependence. Suppose that \mathbf{X} is our random adjacency matrix. The general form of the model is

$$Prob(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\left\{\sum_B \lambda_B z_B(\mathbf{x})\right\},$$

where the summation is over all subsets B of the set of potential edges,

$$z_B(\mathbf{x}) = \prod_{(i,j) \in B} x_{i,j}$$

is the network statistic corresponding to the subset B , κ is a normalising quantity so that the probabilities sum to 1, and the parameter $\lambda_B = 0$ for all \mathbf{x} unless all the variables in B are mutually dependent.

The simplest such model is that the probability of any edge is constant across all possible edges, i.e. the Bernoulli graph, for which

$$Prob(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\{\lambda L(\mathbf{x})\},$$

where $L(\mathbf{x})$ is the number of edges in the network \mathbf{x} and λ is a parameter.

For social networks, Frank and Strauss (1986) introduced *Markov dependence*, whereby two possible edges are assumed to be conditionally dependent if they share a node. For non-directed networks, the resulting model has parameters relating only to the configurations *stars of various types, and triangles*. If the number $L(\mathbf{x})$ of edges, the number $S_2(\mathbf{x})$ of two-stars, the number $S_3(\mathbf{x})$ of three-stars, and the number $T(\mathbf{x})$ of triangles are included, then the model reads

$$Prob(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\{\lambda_1 L(\mathbf{x}) + \lambda_2 S_2(\mathbf{x}) + \lambda_3 S_3(\mathbf{x}) + \lambda_4 T(\mathbf{x})\}.$$

By setting the parameters to particular values and then simulating the distribution, we can examine global properties of the network.

2.6 Specific models for specific networks

Depending on the research question, it may make sense to build a specific network model. For example, a gene duplication model has been suggested which would result in a power-law like node degree distribution. For metabolic pathways, a number of Markov models have been introduced. When thinking of flows through networks, it may be a good idea to use weighted networks; the weights could themselves be random.

3 Fitting a model: parametric methods

Bernoulli (Erdős-Renyi) random graphs

In the random graph model of Erdős and Renyi (1959), the (finite) node set V is given, say $|V| = n$. We denote the set of all potential edges by E ; thus $|E| = \binom{n}{2}$. An edge between two nodes is present with probability p , independently of all other edges. Here p is an unknown parameter.

In classical (frequentist) statistics we often estimate unknown parameters via the method of maximum likelihood.

Example: Bernoulli random graphs.

Our data is the network we see. We describe the data using the adjacency matrix, denote it by \mathbf{x} here because it is the realisation of a random adjacency matrix \mathbf{X} . Recall that

$x_{u,v} = 1$ if and only if there is an edge between u and v .

The likelihood of p being the true value of the edge probability if we see \mathbf{x} is

$$\mathcal{L}(p; \mathbf{x}) = (1-p)^{|E|} \left(\frac{p}{1-p} \right)^{\sum_{(i,j) \in E} x_{i,j}}.$$

If $t = \sum_{(i,j) \in E} x_{i,j}$ is the total number of edges in the random graph, then

$$\hat{p} = \frac{t}{|E|}$$

is our maximum-likelihood estimator.

Maximum-likelihood estimation also works well in Erdős-Renyi Mixture graphs when the number of types is known, and it works well in Watts-Strogatz small world networks when the number k of nearest neighbours we connect to is known. When the number of types, or the number of nearest neighbours, is unknown, then things become messy.

In Barabasi-Albert models, the parameter would be the power exponent for the node degree, as occurring in the probability for an incoming node to connect to some node i already in the network.

In exponential random graphs, unless the network is very small, maximum-likelihood estimation quickly becomes numerically unfeasible. Even in a simple model like

$$Prob(\mathbf{X} = \mathbf{x}) = \frac{1}{\kappa} \exp\{\lambda_1 L(\mathbf{x}) + \lambda_2 S_2(\mathbf{x}) + \lambda_3 S_3(\mathbf{x}) + \lambda_4 T(\mathbf{x})\}$$

the calculation of the normalising constant κ becomes numerically impossible very quickly.

3.1 Markov Chain Monte Carlo estimation

A Markov chain is a stochastic process where the state at time n only depends on the state at time $n - 1$, plus some independent randomness. It is *irreducible* if any set of states can be reached from any other state in a finite number of moves, and it is *reversible* if you cannot tell whether it is running forwards in time or backwards in time. A distribution is *stationary* for the Markov chain if, when you start in the stationary distribution, one step after you cannot tell whether you made any step or not; the distribution of the chain looks just the same.

If a Markov chain is irreducible and reversible, then it will have a unique stationary distribution, and no matter in which state you start the chain, it will eventually converge to this stationary distribution.

We make use of this fact by looking at our target distribution, such as the distribution for \mathbf{X} in an exponential random graph model, as the stationary distribution of a Markov chain.

This Markov chain lives on graphs, and moves are adding or deleting edges, as well as adding types or reducing types. Finding suitable Markov chains is an active area of research.

The `ergm` package has MCMC implemented for parameter estimation. We need to be aware that there is no guarantee that the Markov chain has reached its stationary distribution. Also, if the stationary distribution is not unique, then the results can be misleading. Unfortunately in exponential random graph models it is known that in some small parameter regions the stationary distribution is not unique.

3.2 Assessing the model fit

Suppose that we have estimated our parameters in our model of interest. We can now use this model to see whether it does actually fit the data.

To that purpose we study the (asymptotic) distributions of our summary statistics *node degree*, *clustering coefficient*, and *shortest path length*. Then we see whether our observed values are plausible under the estimated model.

3.3 The distribution of summary statistics in Bernoulli random graphs

In a Bernoulli random graph on n nodes, with edge probability p , the network summaries are pretty well understood.

3.3.1 The degree of a random node

Pick a node v , and denote its degree by $D(v)$, say. The degree is calculated as the number of neighbours of this node. Each of the other $(n - 1)$ nodes is connected to our node v with probability p , independently of all other nodes. Thus the distribution of $D(v)$ is Binomial with parameters n and p , for each node v .

Typically we look at relatively sparse graphs, and so a Poisson approximation applies. If \mathbf{X} denotes the random adjacency matrix, then, in distribution,

$$D(v) = \sum_{u:u \neq v} X_{u,v} \approx \text{Poisson}((n-1)p).$$

Note that the node degrees in a graph are not independent. So $D(v)$ does **not** stand for the average node degree.

How about the average degree of a node? Denote it by \bar{D} . Note that the average does not only take integer values, so would certainly not be Poisson distributed. But

$$\bar{D} = \frac{1}{n} \sum_{v=1}^n D(v) = \frac{2}{n} \sum_{v=1}^n \sum_{u < v} X_{u,v},$$

noting that each edge gets counted twice. As the $X_{u,v}$ are independent, we can use a Poisson approximation again, giving that

$$\sum_{v=1}^n \sum_{u < v} X_{u,v} \approx \text{Poisson} \left(\frac{n(n-1)}{2} p \right)$$

and so, in distribution,

$$\bar{D} \approx \frac{2}{n} Z,$$

where $Z \sim \text{Poisson} \left(\frac{n(n-1)}{2} p \right)$.

3.3.2 The clustering coefficient of a random node

Here it gets a little tricky already. Recall that the *clustering coefficient* of a node v is,

$$C(v) = \frac{\sum_{u,w \in V} X_{u,v} X_{w,v} X_{u,w}}{\sum_{u,w \in V} X_{u,v} X_{w,v}}.$$

The ratio of two random sums is not easy to evaluate. If we just look at

$$\sum_{u,w \in V} X_{u,v} X_{w,v} X_{u,w}$$

then we see that we have a sum of dependent random variables.

Most 3-tuples (u, w, v) and (r, s, t) , though, will not share an index, and hence $X_{u,v}X_{w,v}X_{u,w}$ and $X_{r,s}X_{s,t}X_{r,t}$ will be independent. The dependence among the random variables overall is hence weak, so that a Poisson approximation applies. As

$$E \sum_{u,w,v \in V} X_{u,v}X_{w,v}X_{u,w} = \binom{n}{3}p^3,$$

we obtain that, in distribution,

$$\sum_{u,w,v \in V} X_{u,v}X_{w,v}X_{u,w} \approx \text{Poisson} \left(\binom{n}{3}p^3 \right).$$

Similarly,

$$E \sum_{u,w \in V} X_{u,v}X_{w,v} = \binom{n}{2}p^2.$$

K. Lin (2007) showed that, for the average clustering coefficient

$$C = \frac{1}{n} \sum_v C(v)$$

it is also true that, in distribution,

$$C \approx \frac{1}{n \binom{n}{2} p^2} Z,$$

where $Z \sim \text{Poisson} \left(\binom{n}{3} p^3 \right)$.

Example. In the Florentine family data, we observe a total number of 20 edges, an average node degree of 2.5, and an average clustering coefficient of 0.1914894, with 16 nodes in total. Under the null hypothesis that the data come from a Bernoulli random graph we estimate

$$\hat{p} = \frac{20}{\binom{16}{2}} = \frac{20 \times 2}{16 \times 15} = \frac{1}{6},$$

and the average node degree would be

$$\bar{D} \approx \frac{1}{8}Z,$$

where $Z \sim \text{Poisson}(20)$. The probability under the null hypothesis that $\bar{D} \geq 2.5$ would then be

$$P(Z \geq 2.5 \times 8) = P(Z \geq 20) \approx 0.55,$$

so no reason to reject the null hypothesis.

3.3.3 Shortest paths: Connectivity in Bernoulli random graph

Erdős and Renyi (1960) showed the following "phase transition" for the connectedness of a Bernoulli random graph.

If $p = p(n) = \frac{\log n}{n} + \frac{c}{n} + o\left(\frac{1}{n}\right)$ then the probability that a Bernoulli graph, denoted by $\mathcal{G}(n, p)$ on n nodes with edge probability p is connected converges to $e^{-e^{-c}}$.

The *diameter* of a graph is the maximum diameter of its connected components; the diameter of a connected component is the longest shortest path length in that component.

Chung and Lu (2001) showed that, if $np \geq 1$ then, asymptotically, the ratio between the diameter and $\frac{\log n}{\log(np)}$ is at least 1, and remains bounded above as $n \rightarrow \infty$.

If $np \rightarrow \infty$ then the diameter of the graph is $(1 + o(1))\frac{\log n}{\log(np)}$. If $\frac{np}{\log n} \rightarrow \infty$, then the diameter is concentrated on at most two values.

In the Physics literature, the value $\frac{\log n}{\log(np)}$ is used for the average shortest path length in a Bernoulli random graph. This has hence to be taken with a lot of grains of salt.

While we have some idea about how the diameter (and, relatedly, the shortest path length) behaves, it is an inconvenient statistics for Bernoulli random graphs, because the graph need not be connected.

3.4 The distribution of summary statistics in Watts-Strogatz small worlds

Recall that in this model we arrange the n nodes of V on a lattice. Then hard-wire each node to its k nearest neighbours on each side on the lattice, where k is small. Thus there are nk edges in this hard-wired lattice. Now introduce random shortcuts between nodes which are not hard-wired; the shortcuts are chosen independently, all with the same probability ϕ .

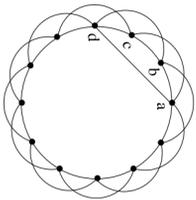
Thus the shortcuts behave like a Bernoulli random graph, but the graph will necessarily be connected. The degree $D(v)$ of a node v in the Watts-Strogatz small world is hence distributed as

$$D(v) = 2k + \text{Binomial}(n - 2k - 1, \phi),$$

taking the fixed lattice into account. Again we can derive a Poisson approximation when p is small; see K.Lin (2007) for the details.

For the clustering coefficient there is a problem - triangles in the graph may now appear in clusters. Each shortcut between nodes u and v which are a distance of $k + a \leq 2k$ apart on the circle creates $k - a - 1$ triangles automatically.

Thus a Poisson approximation will not be suitable; instead we use a *compound Poisson distribution*. A compound Poisson distribution arises as the distribution of a Poisson number of clusters, where the cluster sizes are independent and have some distribution themselves. In general there is no closed form for a compound Poisson distribution.



The compound Poisson distribution also has to be used when approximating the number of 4-cycles in the graph, or the number of other small subgraphs which have the clumping property.

It is also worth noting that when counting the joint distribution of the number of triangles and the number of 4-cycles, these counts are not independent, not even in the limit; a bivariate compound Poisson approximation with dependent components is required. See Lin (2007) for details.

3.4.1 The shortest path length

Let \mathcal{D} denote shortest distance between two randomly chosen points, and abbreviate $\rho = 2k\phi$. Then (Barbour + Reinert) show that uniformly in $|x| \leq \frac{1}{4} \log(n\rho)$,

$$\begin{aligned} & \mathbf{P} \left(\mathcal{D} > \frac{1}{\rho} \left(\frac{1}{2} \log(n\rho) + x \right) \right) \\ &= \int_0^\infty \frac{e^{-y}}{1 + e^{2xy}} dy + O \left((n\rho)^{-\frac{1}{5}} \log^2(n\rho) \right) \end{aligned}$$

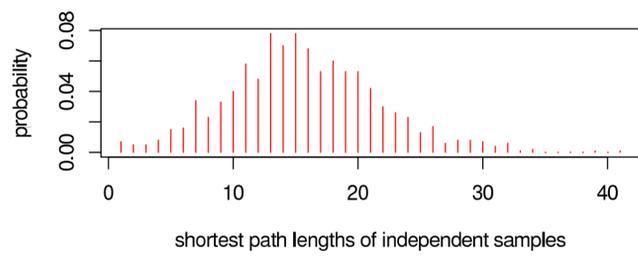
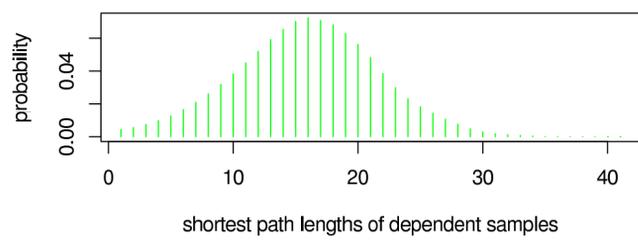
if the probability of shortcuts is small. If the probability of shortcuts is relatively large, then \mathcal{D} will be concentrated on one or two points.

Note that \mathcal{D} is the shortest distance between two randomly chosen points, **not** the average shortest path. Again the difference can be considerable (Computer exercise).

Dependent sampling

Our data are usually just one graph, and we calculate all shortest paths. But there is much overlap between shortest paths possible, creating dependence

Simulation: $n = 500, k = 1, \phi = 0.01$



Simulation: dependent vs independent

We simulate 100 replicas, and calculate the average shortest path length in each network. We compare this distribution to the theoretical approximate distribution; we carry out 100 chi-square tests:

n	k	ϕ	$E.no$	mean p-value	max p-value
300	1	0.01	3	1.74 E-09	8.97 E-08
		0.167	50	0.1978	0.8913
	2	0.01	6	0	0
1000	1	0.003	3	1.65E-13	3.30 E-12
		0.05	50	0.0101	0.1124
	2	0.03	60	0.0146	0.2840

Thus the two statistics are close if the expected number $E.no$ of shortcuts is large (or very small); otherwise they are significantly different.

3.5 The distribution of summary statistics in Barabasi-Albert models

The node degree distribution is given by the model directly, as that is how it is designed.

The clustering coefficient depends highly on the chosen model. In the original Barabasi-Albert model, when only one new edge is created at any single time, there will be no triangles (beyond those from the initial graph). The model can be extended to match any clustering coefficient, but even if only two edges are attached at the same time, the distribution of the number of the clustering coefficient is unknown to date.

The expected value, however, can be approximated. Fronczak *et al.* (2003) studied the models where the network starts to grow from an initial cluster of m fully connected nodes. Each new node that is added to the network created m edges which connect it to previously added nodes. The probability of a new edge to be connected to a node v is proportional to the degree $d(v)$ of this node. If both the number of nodes, n , and m are large, then the expected average clustering coefficient is

$$EC = \frac{m-1}{8} \frac{(\log n)^2}{n}.$$

The average pathlength ℓ increases approximately logarithmically with network size. If $\gamma = 0.5772$ denotes the Euler constant, then Fronczak *et al.* (2004) show for the mean average shortest path length that

$$E\ell \sim \frac{\log n - \log(m/2) - 1 - \gamma}{\log \log n + \log(m/2)} + \frac{3}{2}.$$

The asymptotic distribution is not understood.

3.6 The distribution of summary statistics in exponential random graph models

The distribution of the node degree, clustering coefficient, and the shortest path length is poorly studied in these models. One reason is that these models are designed to predict missing edges, and to infer characteristics of nodes, but their topology itself has not often been of interest.

The summary statistics appearing in the model try to push the random networks towards certain behaviour with respect to these statistics, depending on the sign and the size of their factors θ .

When only the average node degree and the clustering coefficient are included in the model, then a strange phenomenon happens. For many combinations of parameter values the model produces networks that are either full (every edge exists) or empty (no edge exists) with probability close to 1. Even for parameters which do not produce this phenomenon, the distribution of networks produced by the model is often bimodal: one mode is sparsely connected and has a high number of triangles, while the other mode is densely connected but with a low number of triangles. Again: active research.

4 Statistical tests for model fit: nonparametric methods

What if we do not have a suitable test statistic for which we know the distribution? We need some handle on the distribution, so here we assume that we can simulate random samples from our null distribution. There are a number of methods available. It is often a good idea to use plots to visually assess the fit, first via a quantile-quantile plot. Then what?

4.1 Monte-Carlo tests

The Monte Carlo test, attributed to Dwass (1957) and Barnard (1963), is an exact procedure of virtually universal application and correspondingly widely used.

Suppose that we would like to base our test on the statistic T_0 . We only need to be able to simulate a random sample T_{01}, T_{02}, \dots from the distribution, call it F_0 , determined by the null hypothesis. We assume that F_0 is continuous, and, without loss of generality, that we reject the null hypothesis H_0 for large values of T_0 . Then, provided that $\alpha = \frac{m}{n+1}$ is rational, we can proceed as follows.

1. Observe the actual value t^* for T_0 , calculated from the data
2. Simulate a random sample of size n from F_0
3. Order the set $\{t^*, t_{01}, \dots, t_{0n}\}$
4. Reject H_0 if the *rank* of t^* in this set (in decreasing order) is $\geq m$.

The basis of this test is that, under H_0 , the random variable T^* has the same distribution as the remainder of the set and so, by symmetry,

$$\mathbf{P}(t^* \text{ is among the largest } m \text{ values}) = \frac{m}{n+1}.$$

The procedure is exact however small n might be; increasing n increases the power of the test. The question of how large n should be is discussed by Marriott (1979), see also Hall and Titterington (1989).

An alternative view of the procedure is to count the number M of simulated values $> t^*$. Then $\hat{P} = \frac{M}{n}$ estimates the true significance level P achieved by the data, i.e.

$$P = \mathbf{P}(T_0 > t^* | H_0).$$

In discrete data, we will typically observe ties. We can break ties randomly, then the above procedure will still be valid.

Unfortunately this test does not lead directly to confidence intervals.

For random graphs, Monte Carlo tests often use shuffling edges with the number of edges fixed, or fixing the node degree distribution, or fixing some other summary.

Suppose we want to see whether our observed clustering coefficient is "unusual" for the type of network we would like to consider. Then we may draw many networks uniformly at random from all networks having the same node degree sequence, say. We count how often a clustering coefficient at least as extreme as ours occurs, and we use that to test the hypothesis.

In practice these types of test are the most used tests in network analysis. They are called *conditional uniform graph tests*.

Some caveats:

In Bernoulli random graphs, the number of edges asymptotically determines the number of triangles when the number of edges is moderately large. Thus conditioning on the number of edges (or the node degrees, which determine the number of edges) gives degenerate results. More generally, we have seen that node degrees and clustering coefficient (and other subgraph counts) are not independent, nor are they independent of the shortest path length. By fixing one summary we may not know exactly what we are testing against.

”Drawing uniformly at random” from complex networks is not as easy as it sounds. Algorithms may not explore the whole data set.

”Drawing uniformly at random”, conditional on some summaries being fixed, is related to sampling from exponential random graphs. We have seen already that in exponential random graphs there may be more than one stationary distribution for the Markov chain Monte Carlo algorithm; this algorithm is similar to the one used for drawing at random, and so we may have to expect similar phenomena.

4.2 Scale-free networks

Barabasi and Albert introduced networks such that the distribution of node degrees is of the type

$$\text{Prob}(\text{degree} = k) \sim Ck^{-\gamma}$$

for $k \rightarrow \infty$. Such behaviour is called *power-law behaviour*; the constant γ is called the *power-law exponent*. The networks are also called *scale-free*:

If $\alpha > 0$ is a constant, then

$$\text{Prob}(\text{degree} = \alpha k) \sim C(\alpha k)^{-\gamma} \sim C'k^{-\gamma},$$

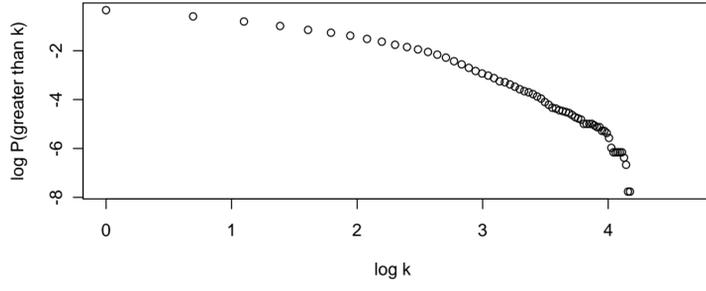
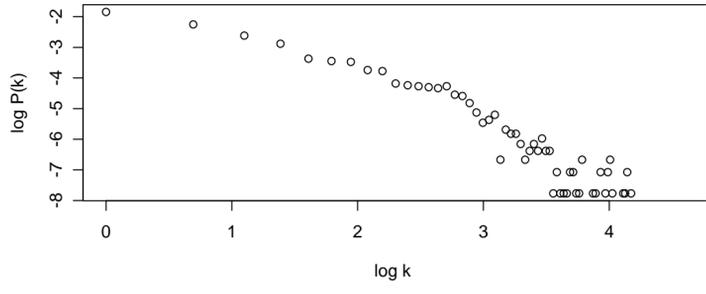
where C' is just a new constant. That is, scaling the argument in the distribution changes the constant of proportionality as a function of the scale change, but preserves the shape of the distribution itself.

If we take logarithms on both sides:

$$\begin{aligned}\log \text{Prob}(\text{degree} = k) &\sim \log C - \gamma \log k \\ \log \text{Prob}(\text{degree} = \alpha k) &\sim \log C - \gamma \log \alpha - \gamma \log k;\end{aligned}$$

scaling the argument results in a linear shift of the log probabilities only. This equation also leads to the suggestion to plot the $\log \text{relfreq}(\text{degree} = \alpha k)$ of the empirical relative degree frequencies against $\log k$. Such a plot is called a *log-log plot*. If the model is correct, then we should see a straight line; the slope would be our estimate of γ .

Example: Yeast data.



These plots have a lot of noise in the tails. As an alternative, *Newman (2005)* suggests to plot the log of the empirical cumulative distribution function instead, or, equivalently, our estimate for

$$\log \text{Prob}(\text{degree} \geq k).$$

If the model is correct, then one can calculate that

$$\log \text{Prob}(\text{degree} \geq k) \sim C'' - (\gamma - 1) \log k.$$

Thus a log-log plot should again give a straight line, but with a shallower slope. The tails are somewhat less noisy in this plot.

In both cases, the slope is estimated by least-squares regression: for our observations, $y(k)$ (which could be log probabilities or log cumulative probabilities, for example) we find the line $a + bk$ such that

$$\sum (y(k) - a - bk)^2$$

is as small as possible.

As a measure of fit, the sample correlation R^2 is computed. For general observations $y(k)$ and $x(k)$, for $k = 0, 1, \dots, n$, with averages \bar{y} and \bar{x} , it is defined as

$$R = \frac{\sum_k (x(k) - \bar{x})(y(k) - \bar{y})}{\sqrt{(\sum_k (x(k) - \bar{x})^2)(\sum_k (y(k) - \bar{y})^2)}}.$$

It measures the strength of the linear relationship.

In linear regression, $R^2 > 0.9$ would be rather impressive. However, the rule of thumb for log-log plots is that

1. $R^2 > 0.99$
2. The observed data (degrees) should cover at least 3 orders of magnitude.

Examples include the World Wide Web at some stage, when it had around 10^9 nodes. The criteria are not often matched.

A final issue for scale-free networks: It has been shown (*Stumpf et al. (2005)*) that when the underlying real network is scale-free, then a subsample on fewer nodes from the network will not be scale-free. Thus if our subsample looks scale-free, the underlying real network will not be scale-free.

In biological network analysis, it is debated how useful the concept of "scale-free" behaviour is, as many biological networks contain relatively few nodes.

For Bernoulli random graphs we have a reasonable grasp on the distribution of our network summaries, we can use maximum-likelihood estimation and we can use the distribution of the summaries for testing hypotheses. For Watts-Strogatz small worlds some results are available. A main observation is that, in contrast to Bernoulli random graphs, even for counting triangles a compound Poisson approximation is needed, rather than a Poisson approximation. The underlying issue is that triangles (and also other motifs) occur in clumps.

For random graphs, Monte Carlo tests often use shuffling edges with the number of edges fixed, or fixing the node degree distribution, or fixing some other summary. Then we use a Monte Carlo test to see whether our test statistic is unusual, compared to graphs drawn at random with the same number of edges, or the same node degree distribution, say. The results have to be interpreted carefully.