Preliminaries

There will be 15 lectures, 3 2-hour practical sessions, and 4 1-hour examples classes.

The lectures will take place Wednesdays 4-5, week 1-8, and Fridays 2-3, week 1-7, in 1 South Parks Road. There will be examples classes on Thursdays weeks 2, 4, 6, 8, lasting one hour each; one class 3-4, the other class 4-5. They will take place in the Seminar Room in 1 South Parks Road. There will be practical sessions on Thursdays in weeks 3, 5, 7, lasting two hours each, 3-5pm. They take place in the Computer Room, 1 South Parks Road. The marker for your class is Vince Choo. The work is due Tuesdays at 12 noon, in 1 South Parks Road, weeks 2, 4, 6, and 8.

There will be homework assignments for the classes which will be marked, and the assignments may require some computer work.

For questions: reinert@stats.ox.ac.uk.

Reading


Outline

1 Introduction
2 Models for Networks
3 Stein’s Method for Normal Approximation
4 Stein’s Method for Poisson Approximation
5 Threshold Behaviour
6 Yule Processes
7 Shortest paths in small world networks
8 Other models and summaries
9 Sampling from networks
10 Fitting a model
11 Nonparametric methods
12 Statistical inference for vertex characteristics and edges
13 Motifs
14 Modules and communities
15 Further topics
1 Introduction

Networks are just graphs.

Often one would think of a network as a connected graph, but not always. In this lecture course we shall use network and graph interchangeably.

Here are some examples of networks (graphs).

Figure: Marriage relations between Florentine families; two different graphical representations

Figure: KEGG pathway: Parkinson’s disease; www.genome.jp/kegg/pathway/hsa/hsa05012.png, 11/4/2012
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

Unless the network is very small it appears like a hairball, and is difficult to analyse by just looking at it.

The study of networks has a long tradition in social science, where it is called *Social Network Analysis*. 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 did these networks evolve?
- How similar are these networks?
- How are these networks interlinked?
- What are the building principles of these networks? How is resilience achieved, and how is flexibility achieved?
From a statistical viewpoint, questions include

- How to best describe networks?
- How to infer characteristics of vertices in the network?
- How to infer missing links, and how to check whether existing links are not false positives?
- How to compare networks?
- 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.

1.9

1.1. What are networks?

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

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.

If both endpoints of an edge are the same, then the edge is a loop. Here we shall mainly exclude self-loops, as well as multiple edges between two vertices.

1.10

Example: Adjacency matrix for marital relation between Florentine families

(see Wasserman+Faust, p.744)

\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\]
A complete graph is a graph, without self-loops, such that every pair of vertices 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 vertex set $V$ is decomposed into two disjoint subsets, $U$ and $W$, say, such that there are no edges between any two vertices in $U$, and also there are no edges between any two vertices in $W$; all edges have one endpoint in $U$ and the other endpoint in $W$. The adjacency matrix $A$ can then be arranged such that it is of the form

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

1.2. Network summaries

To analyse and to compare networks, often low-dimensional summaries are used.

Some summaries concentrate on local features, such as local clustering, whereas other summaries concentrate on global features.

1.2.1. Degrees and the degree distribution

The degree $d(v)$ of a vertex $v$ is the number of edges which involve $v$ as an endpoint. The degree is easily calculated from the adjacency matrix $A$:

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

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

A degree distribution $(d_0, d_1, d_2, \ldots)$ of a graph on $n$ vertices is the vector of fraction of vertices with given degree:

$$d_i = \frac{1}{n} \times \text{number of vertices of degree } i.$$  

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

1.2.2. The local clustering coefficient

The local clustering coefficient of a vertex $v$ 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,

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

Here $0/0 := 0$.

The average clustering coefficient is defined as

$$\overline{C} = \frac{1}{|V|} \sum_{v \in V} C(v).$$
Note that
\[ \frac{1}{2} \sum_{u \neq v \in V, w \neq u, v \in V} a_{u,v}a_{w,v}a_{u,w} \]
is the number of triangles involving \( v \) in the graph. Similarly,
\[ \frac{1}{2} \sum_{u \neq w \in V} a_{u,v}a_{w,v} \]
is the number of 2-stars centred around \( v \) in the graph. The local clustering coefficient describes how "locally dense" a graph is.

1.2.3. The global clustering coefficient

The global clustering coefficient or transitivity is defined as
\[ C = \frac{3 \times \text{number of triangles}}{\text{number of connected triples}}. \]

Note that \( C \neq \overline{C} \) in general. Indeed \( C \) tends to be dominated by vertices with low degree, since they tend to have small denominators in the local clustering coefficient.

The global clustering coefficient in the Florentine family example is 0.1914894; the global clustering coefficient in the Yeast data is 0.1023149. The average clustering coefficient in the Florentine family example is 0.1395833.

1.2.4. The expected clustering coefficient

For models of random networks often we consider the expected clustering coefficient
\[ E(C) = \frac{3 \times E(\text{number of triangles})}{E(\text{number of connected triples})}. \]

Unfortunately all of the average clustering coefficient, the global clustering coefficient, and the expected clustering coefficient are often just called the clustering coefficient in the literature. Here we mean by clustering coefficient the global clustering coefficient, unless otherwise stated.

1.2.5. The average shortest path

In a graph a path from vertex \( v_0 \) to vertex \( v_n \) is an alternating sequence of vertices and edges, \( (v_0, e_1, v_1, e_2, \ldots, v_{n-1}, e_n, v_n) \) such that the endpoints of \( e_i \) are \( v_{i-1} \) and \( v_i \), for \( i = 1, \ldots, n \). The distance \( \ell(u, v) \) between two vertices \( u \) 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.
A graph is called connected if there is a walk between any pair of vertices in the graph, otherwise it is called disconnected.

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.

1.2.6. Small subgraphs and motifs

In addition to considering these general summary statistics, it has proven fruitful to describe networks in terms of small subgraphs, as these can be viewed as building-block patterns of networks.

Often a subgraph is called a motif when it is over-represented in the network. Over-representation is judged using a probabilistic model for the network.

Here we think of a motif as a subgraph with a fixed number of vertices and with a given topology, and we use the term interchangeably with small subgraph.

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.

1.2.7. Further summaries

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 vertices with special properties, such as hub vertices, which are vertices with high degree.

Looking at the "spectral decomposition", i.e. at eigenvectors and eigenvalues, of the adjacency matrix, provides another set of summaries, such as centrality.
Other popular summaries, to name but a few, are: the between-ness of an edge counts the proportion of shortest paths between any two vertices which pass through this edge. Similarly, the between-ness of a vertex is the proportion of shortest paths between any two vertices which pass through this vertex.

The connectivity of a connected graph is the smallest number of edges whose removal results in a disconnected graph.

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.

Milgram reported a median of six. 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.

For more details see for example http://www.columbia.edu/itc/sociology/watts/w3233/client_edit/big_world.html.
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. The randomness can be intrinsic to the network, and/or may stem from errors in the data.

To understand the randomness 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) vertex set $\mathcal{V}$ is given, say $|\mathcal{V}| = n$, and an edge between two vertices is present with probability $p$, independently of all other edges.

The expected number of edges is

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

Each vertex has $n - 1$ potential neighbours, and each of these $n - 1$ edges is present with probability $p$, and so the degree of a randomly picked vertex is $\text{Bin}(n - 1, p)$-distributed.

The expected number of triangles in the graph is

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

The expected clustering coefficient is $p$.

### Network

<table>
<thead>
<tr>
<th>Network</th>
<th>$n$</th>
<th>ave degree</th>
<th>$C$</th>
<th>$C_{\text{Bernoulli}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Los Alamos</td>
<td>52,909</td>
<td>9.7</td>
<td>0.43</td>
<td>0.00018</td>
</tr>
<tr>
<td>MEDLINE</td>
<td>1,520,251</td>
<td>18.1</td>
<td>0.066</td>
<td>0.000011</td>
</tr>
<tr>
<td>NCSTRL</td>
<td>11,994</td>
<td>3.59</td>
<td>0.496</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Note the discrepancies!

Also in real-world graphs often the shortest path length is much shorter than expected from a Bernoulli random graph with the same average vertex 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

In 1998, Watts and Strogatz published a ground-breaking paper with a new model for small worlds; the version currently most used is as follows. Arrange the $n$ vertices of $V$ on a lattice. Then hard-wire each vertex 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 vertices which are not hard-wired; the shortcuts are chosen independently, all with the same probability $p$.

The degree of a randomly chosen vertex has distribution $2k + Bin(n - 2k - 1, p)$.

If there are no shortcuts, then the average distance between two randomly chosen vertices is of the order $n$, the number of vertices. But as soon as there are just a few shortcuts, then the average distance between two randomly chosen vertices 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.

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 vertex degree distributions. It is often observed that vertices tend to attach to "popular" vertices; 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

$$d_k \sim Ck^{-\gamma}$$

for $k \to \infty$. Such behaviour is called *power-law behaviour*; the constant $\gamma$ 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 vertices linked by $m$ (parallel) edges. At every time $t \geq 2$ we add a new vertex with $m$ edges that link the new vertex to vertices already present in the network. We assume that the probability $\pi_i$ that the new vertex will be connected to a vertex $i$ depends on the degree $d(i)$ of $i$ so that

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

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

Other modifications vary the probability for choosing a given vertex.

If \( \alpha_1, \ldots, \alpha_L \) denote the proportion of the vertices of different types, \( \sum_\ell \alpha_\ell = 1 \), then for a vertex \( V \) picked uniformly at random,

\[
E(d(V)) = \sum_\ell \alpha_\ell \left( (L\alpha_\ell - 1)p_{\ell,\ell} + \sum_{k \neq \ell} L\alpha_k p_{k,\ell} \right).
\]

Daudin et al. (2008) 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.

### 2.4. Erdős-Renyi Mixture Graphs

An intermediate type of 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 vertices 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 vertices at the endpoints of the edge;

\[
p_{i,j} = \mathbf{P}(u, v) \in \mathcal{E} | u \text{ is of type } i, v \text{ is of type } j).
\]

### 2.5. The configuration model

The configuration model is a model which gives a graph which is uniformly chosen among all graphs with a given degree sequence. The degree sequence \((d(i), i = 1, \ldots, n)\) of a graph on \( n \) vertices is the vector of degrees for all the vertices.

Given a feasible degree sequence, \((d(i), i = 1, \ldots, n)\) and \( m = \frac{1}{2} \sum_i d(i) \), construct a graph as follows.

1. Give each vertex \( i \) a total of \( d(i) \) “stubs” of edges.
2. Choose two of the stubs uniformly at random and create an edge by connecting them to one another.
3. Choose another pair from the remaining \( 2m - 2 \) stubs and connect these.
4. Continue until all the stubs are used up.

In the resulting network every vertex has exactly the desired degree.

Here we assume that all degrees are at least 1.
We note that more than one set of matchings may result in the same network; if we labelled the stubs there would be typically many ways we can join up pairs of labelled stubs to create the same final configuration.

The model lends itself to analytical considerations through noting that the expected number of edges between vertices $i$ and $j \neq i$ is

$$p_{i,j} = \frac{d(i)d(j)}{2m-1}.$$ 

2.13

### Self-loops, multiple edges

The construction may create self-loops and multiple edges. But the density of self-loops and multiple edges tends to zero as the number of vertices tends to infinity.

The expected number of self-loops from $i$ to $i$ is

$$p_{i,i} = \frac{d(i)(d(i) - 1)}{2(2m-1)}.$$ 

Hence the expected number of self-loops is

$$\sum_{i=1}^{n} \frac{d(i)(d(i) - 1)}{2(2m-1)} = \frac{\text{E}(d(V)^2) - \text{Ed}(V)}{2(\text{Ed}(V) - 1/n)}$$

where $d(V)$ is the degree of a randomly chosen vertex; $\text{Ed}(V) = \frac{1}{n} \sum_{i=1}^{n} d(i) = \frac{2m}{n}$. As long as $\text{Ed}(V)^2$ remains bounded, the expected number of self-loops will remain bounded, whereas the number of edges will grow with the number of vertices in the network. Hence self-loops will be rare when the network is large.

2.14

### 2.6. Geometric random graphs

In geometric random graphs, introduced by Gilbert in 1961, $n$ points $X_1, \ldots, X_n$ are chosen independently at random according to a density $f(\cdot)$ on $\mathbb{R}^d$. A value $r > 0$ is chosen and we put an edge $(i,j)$ if $d(X_i, X_j) \leq r$. Here $d$ is a distance on $\mathbb{R}^d$; usually we choose the Euclidean distance.

The 2003 book by Penrose contains many more details.
2.7. Exponential random graph ($p^*$) models

Networks have been analysed for a long time in the social science literature, see for example the book by Wasserman and Faust. Here usually directed graphs are studied. Typical research questions could be

- 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 $X$ of a graph simultaneously, making it easy to incorporate dependence. The general form of the model is

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

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

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

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

2.8. Specific models for specific networks

Depending on the research question, it may make sense to build a specific network model. For protein-protein interaction networks gene duplication models have been proposed as models for network growth. When thinking of flows through networks, it may be a good idea to use weighted networks; the weights could themselves be random.
2.9. The need for the distribution of summary statistics under network models

For hypothesis testing we would like to know the distribution of the summary statistic under different network models. Such distributional results, if available at all, are of asymptotic nature. The appropriate asymptotic regime will depend on the parameter values - for example, in Bernoulli random graphs, the number of triangles will be approximately Poisson if \( p \) is small, and approximately normal if \( p \) is moderately large.

The next two chapters introduce a technique to obtain such results, and will show applications to graph statistics.

Facts of weak convergence

\[ P_n \xrightarrow{D} P \]

is equivalent to:

1. \( P_n(A) \to P(A) \) for each \( P \)-continuity set \( A \) (i.e. \( P(\partial A) = 0 \))
2. \( \int f \, dP_n \to \int f \, dP \) for all functions \( f \) that are bounded, continuous, real-valued
3. \( \int f \, dP_n \to \int f \, dP \) for all functions \( f \) that are bounded, infinitely often differentiable, continuous, real-valued.

For a random variable \( X \) with distribution \( \mathcal{L}(X) \), \( \mathcal{L}(X_n) \xrightarrow{D} \mathcal{L}(X) \) if and only if \( \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)] \) for all continuous, bounded real-valued functions \( f \) that are infinitely often differentiable.

3 Stein’s Method for Normal Approximation

The standard reference for this chapter is the book by Chen, Goldstein and Shao (2011).

Often distributional distances are phrased in terms of test functions. This approach is based on weak convergence.

Suppose that \( X_1, X_2, \ldots \) are (real-valued) random variables; for all \( n \) let \( P_n = \mathcal{L}(X_n) \) denote the distribution of \( X_n \), and \( F_n \) denote the cumulative distribution function (c.d.f.) for \( X_n \), so that \( F_n(x) = \mathbb{P}(X_n \leq x) \), with \( x \in \mathbb{R} \).

For c.d.f.s \( F_n, n \geq 0 \) and \( F \) on the line we say that \( F_n \) converges weakly (converges in distribution) to \( F \),

\[ F_n \xrightarrow{D} F \]

if for all continuity points \( x \) of \( F \) we have \( F_n(x) \to F(x) \) as \( n \to \infty \).

If \( F_n \xrightarrow{D} F \) then, for the associated probability distributions, we write

\[ P_n \xrightarrow{D} P. \]

Stein’s characterisation of the normal distribution Part 1

Let \( \mathcal{N}(\mu, \sigma^2) \) denote the normal distribution with mean \( \mu \) and variance \( \sigma^2 \), let \( Z \sim \mathcal{N}(0, 1) \) denote a standard normal variable, and \( \Phi \) the standard normal c.d.f. while \( \phi \) denotes its probability density function (p.d.f.).

**Theorem 3.1** (Stein 1972) If \( W \sim \mathcal{N}(0, 1) \), then

\[ \mathbb{E}[W f(W)] = \mathbb{E}[f'(W)] \]

(3.1)

for all absolutely continuous functions \( f : \mathbb{R} \to \mathbb{R} \) with \( \mathbb{E}|f'(W)| < \infty \).
Heuristic of Stein’s method:

For a random variable $W$ with $E[W] = 0$, $\text{Var}(W) = 1$, if

$$E[f'(W)] - E[Wf(W)]$$

is close to zero for many functions $f$, then $W$ should be close to $Z$ in distribution.

Proof of the theorem

Assume $W \sim \mathcal{N}(0, 1)$. Let $f : \mathbb{R} \to \mathbb{R}$ be an absolutely continuous function with $E|f'(W)| < \infty$. Hence $Ef'(W)$ exists, and

$$Ef'(W) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(w) e^{-w^2/2} dw$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} f'(w) \left( \int_{-\infty}^{w} -xe^{-x^2/2} dx \right) dw$$

$$+ \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} f'(w) \left( \int_{w}^{\infty} xe^{-x^2/2} dx \right) dw.$$

Proof continued

By Fubini’s Theorem it follows that

$$Ef'(W) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} \left( \int_{-\infty}^{0} f'(w) dw \right) (-x)e^{-x^2/2} dx$$

$$+ \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} \left( \int_{0}^{x} f'(w) dw \right) xe^{-x^2/2} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [f(x) - f(0)] xe^{-x^2/2} dx$$

$$= E[Wf(W)],$$

as required.

The Stein equation

Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded (measureable) function. We abbreviate

$$Nh = Eh(Z)$$

and call

$$f'(w) - wf(w) = h(w) - Nh$$

(3.2)

the Stein equation (for $h$).
Solving the Stein equation

Lemma 3.2
The unique bounded solution \( f(w) = f_h(w) \) of the Stein equation (3.2) is given by

\[
f_h(w) = e^{-\frac{w^2}{2}} \int_{-\infty}^{w} (h(x) - Nh)e^{-x^2/2} dx
given by (3.3)
\]

\[
= -e^{-\frac{w^2}{2}} \int_{w}^{\infty} (h(x) - Nh)e^{-x^2/2} dx.
given by (3.4)
\]

Boundedness of the solution

For \( g : \mathbb{R} \rightarrow \mathbb{R} \) let \( ||g|| = \sup_x |g(x)| \) denote the supremum norm.

Lemma 3.3
For \( h \) bounded and \( f_h \) its solution of the Stein equation (3.2) we have

\[
||f_h|| \leq \sqrt{\frac{\pi}{2}} ||h(\cdot) - Nh||.
given by (3.5)
\]

Proof
Multiplying both sides of (3.2) by \( e^{-\frac{w^2}{2}} \) gives

\[
(e^{-\frac{w^2}{2}} f(w))' = e^{-\frac{w^2}{2}} (h(w) - Nh).
given by (3.3)
\]

Integrating yields

\[
e^{-\frac{w^2}{2}} f_h(w) = \int_{-\infty}^{w} (h(x) - Nh)e^{-x^2/2} dx = -\int_{w}^{\infty} (h(x) - Nh)e^{-x^2/2} dx.\]

Re-arranging gives the required form.

The next lemma will show that the solution is bounded. The general solution to (3.2) is given by \( f_h(w) \) plus some constant multiple, say, \( ce^{\frac{w^2}{2}} \), of the solution to the homogeneous equation. Hence the solution is only bounded when \( c = 0 \).

Proof
From (3.3) and (3.4) we have that

\[
|f_h(z)| \leq ||h(\cdot) - Nh|| \int_{|w|}^{\infty} e^{-x^2/2} dx.
given by (3.5)
\]

For \( w \geq 0 \) the function

\[
g(w) = e^{\frac{w^2}{2}} \int_{|w|}^{\infty} e^{-x^2/2} dx\]

has derivative \( wg(w) - 1 \). Note that for \( w > 0 \)

\[
\int_{w}^{\infty} e^{-x^2/2} dx < \int_{w}^{\infty} \frac{x}{w} e^{-x^2/2} dx = \frac{e^{-w^2/2}}{w}.
given by (3.6)
\]

Hence \( wg(w) < 1 \) and \( g(w) \) has negative derivative for \( w \geq 0 \).

As \( g(w) = g(-w) \) the function achieves its maximum at \( w = 0 \).
Moreover
\[ g(0) = \sqrt{2\pi} P(Z \geq 0) = \sqrt{2\pi} \times \frac{1}{2} = \sqrt{\frac{\pi}{2}}, \]
proving the assertion.

Note that we have just shown that
\[ e^{\frac{x^2}{2}} \int_{|w|}^{\infty} e^{-x^2/2} dx \leq \sqrt{\frac{\pi}{2}}. \quad (3.7) \]

Proof

Assume that \( f \) is a bounded, continuous and piecewise continuously differentiable function with \( E|f'(Z)| < \infty \).

Let \( z \in \mathbb{R} \) and let \( g_z(w) = 1_{(-\infty,z]}(w) \) denote the indicator function of \( (-\infty,z] \), so that \( g_z(w) = 1 \) if \( w \leq z \), and 0 otherwise. The solution \( f_z \) of the Stein equation (3.2) for \( g_z \) is clearly continuous and piecewise continuously differentiable. Lemma 3.3 shows that \( f_z \) is also bounded. Now we replace \( w \) by \( W \) in (3.2), take expectations, and conclude that, for all \( z \in \mathbb{R} \),
\[ 0 = E f_z'(W) - E[Wf_z(W)] = P(W \leq z) - \Phi(z), \]
so \( W \sim \mathcal{N}(0,1) \), as required.

Stein’s characterisation of the normal distribution Part 2

Theorem 3.4
(Stein 1972) If (3.1) holds for all bounded, continuous and piecewise continuously differentiable functions \( f \) with \( E|f'(Z)| < \infty \), then \( W \sim \mathcal{N}(0,1) \).

Theorem 3.1 and Theorem 3.4 together provide Stein’s characterisation of the standard normal distribution.

Stein’s Method: summary

To assess how far the distribution of \( W \) is from the standard normal distribution: Let \( \mathcal{H} \) be a class of test functions which is sufficient for characterising weak convergence. Given a test function \( h \in \mathcal{H} \), solve for \( f = f_h \) in the Stein equation
\[ f'(w) - wf(w) = h(w) - Nh. \]

Then replace \( w \) by \( W \) and take expectations:
\[ E[f_h'(W)] - E[Wf_h(W)] = Eh(W) - Nh. \]

So
\[ \sup_{h \in \mathcal{H}} |Eh(W) - Nh| = \sup_{f_h} |Ef_h'(W) - E[Wf_h(W)]|. \]

Sometimes the right-hand side is easier to bound than the left-hand side! But often we need bounds on the second derivative of \( f_h \).
More bounds

**Lemma 3.5**

Let \( h : \mathbb{R} \to \mathbb{R} \) be absolutely continuous and let \( f = f_h \) be the solution of its Stein equation (3.2). Then

\[
\|f_h\| \leq 2\|h'\| \\
\|f_h'\| \leq \sqrt{\frac{2}{\pi}} \|h'\| \\
\|f_h''\| \leq 2\|h'\|.
\]

The proof of lemma is technical; see pp.39-41 in the book by Chen, Goldstein and Shao.

**3.16**

**Theorem: explicit bound on distance**

Let \( X, X_1, \ldots, X_n \) be i.i.d. with \( \mathbb{E}X = 0 \) and \( \text{Var}(X) = \frac{1}{n} \); put \( W = \sum_{i=1}^{n} X_i \). For any absolutely continuous function \( h : \mathbb{R} \to \mathbb{R} \) we have

\[
|\mathbb{E}h(W) - Nh| \leq \|h'\| \left( \frac{2}{\sqrt{n}} + \sum_{i=1}^{n} \mathbb{E}|X_i^3| \right).
\]

Note: this bound is true for any \( n \); nothing goes to infinity.

**3.17**

**Proof**

First note that \( \mathbb{E}W = 0 \) and \( \text{Var}(W) = 1 \). For \( i = 1, \ldots, n \) put \( W_i = W - X_i = \sum_{j \neq i} X_j \). Let \( h : \mathbb{R} \to \mathbb{R} \) be absolutely continuous. Then, for \( f = f_h \) solution of the Stein equation (3.2) for \( h \),

\[
\mathbb{E}Wf(W) = \sum_{i=1}^{n} \mathbb{E}X_if(W) \\
= \sum_{i=1}^{n} \mathbb{E}X_i f(W_i) + \sum_{i=1}^{n} \mathbb{E}X_i^2 f'(W_i) + R \\
= \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}f'(W_i) + R.
\]

Here, by Taylor expansion and Lemma 3.5

\[
|R| \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}|X_i^3|\|f''\| \leq \|h'\| \sum_{i=1}^{n} \mathbb{E}|X_i^3|.
\]

**3.18**

**Proof continued**

So

\[
\mathbb{E}\{f'(W)\} - \mathbb{E}\{f(W)\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{f'(W) - f'(W_i)\} + R
\]

and again by Taylor expansion

\[
\left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\{f'(W) - f'(W_i)\} \right| \leq \|f''\| \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}|X_i^3| \leq 2\|h'\| \frac{1}{\sqrt{n}},
\]

where we used Lemma 3.5 and the Cauchy-Schwarz inequality for the last step. This completes the proof.
Local dependence: decomposable random variables

The theorem extends to local dependence. Let $I$ be a finite index set and assume that $X_i, i \in I$ are mean zero random variables, finite variances. Put $W = \sum_{i \in I} X_i$; assume $\text{Var}(W) = 1$. Suppose that we can write $W = W_i + Z_i$ with $W_i$ independent of $X_i$ and such that there is a set $K_i \subset I$ with

$$Z_i = \sum_{k \in K_i} Z_{ik}, \quad i \in I.$$ 

Further assume that $W_i = W_{ik} + V_{ik}$ for $i \in I, k \in K_i$ so that $W_{ik}$ is independent of the pair $(X_i, Z_{ik})$. If we can find such a decomposition then we say that $W$ is decomposable.

This concept, as well as the next result, can be found in A.D. Barbour, M. Karonski, and A. Ruzinski (1989).

Proof

First note that

$$1 = \mathbf{E}\{W^2\} = \text{Var}(W) = \sum_{i \in I} \mathbf{E}(X_i W)$$ 

$$= \sum_{i \in I} (\mathbf{E}X_i W - \mathbf{E}X_i W_i) = \sum_{i \in I} \mathbf{E}(X_i Z_i)$$ 

$$= \sum_{i \in I} \sum_{k \in K_i} \mathbf{E}(X_i Z_{ik}).$$

Theorem for decomposable variables

Assume that $W$ is decomposable as above. Then, for any absolutely continuous $h$,

$$|\mathbf{E}h(W) - Nh|$$ 

$$\leq 2||h'|| \left\{ \frac{1}{2} \sum_{i \in I} \mathbf{E}(|X_i Z_i^2|)$$ 

$$+ \sum_{i \in I} \sum_{k \in K_i} (\mathbf{E}|X_i Z_{ik}V_{ik}| + \mathbf{E}|X_i Z_{ik}|\mathbf{E}|Z_i + V_{ik}|) \right\}.$$ 

Proof continued

Let $h : \mathbb{R} \to \mathbb{R}$ be absolutely continuous. Then, for $f = f_h$ solution of the Stein equation (3.2) for $h$,

$$\mathbf{E}\left(Wf(W) - f'(W)\right)$$ 

$$= \left\{ \mathbf{E}(Wf(W)) - \sum_{i \in I} \mathbf{E}(X_i Z_i f'(W_i)) \right\}$$ 

$$+ \left\{ \sum_{i \in I} \mathbf{E}(X_i Z_i f'(W_i)) - \sum_{i \in I} \sum_{k \in K_i} \mathbf{E}(X_i Z_{ik})\mathbf{E}(f'(W_{ik})) \right\}$$ 

$$+ \sum_{i \in I} \sum_{k \in K_i} \{\mathbf{E}(X_i Z_{ik}|f'(W_{ik}) - \mathbf{E}|f'(W)|)\}.$$
Proof continued

By Taylor expansion,

\[ W f(W) = \sum_{i \in I} X_i f(W) \]

\[ = \sum_{i \in I} X_i \left\{ f(W_i) + Z_i f'(W_i) + \frac{1}{2} Z_i^2 f''(W_i + \theta_i Z_i) \right\} \]

for some \( \theta_i \in [0, 1] \). Using the decomposability,

\[ \left| \mathbb{E}(W f(W)) - \sum_{i \in I} \mathbb{E}(X_i Z_i f'(W_i)) \right| \]

\[ \leq \left| \mathbb{E} \left( \sum_{i \in I} X_i f(W_i) \right) \right| + \frac{1}{2} ||f''|| \sum_{i \in I} \mathbb{E}(|X_i|Z_i^2). \]

Due to the independence and \( \mathbb{E}X_i = 0 \) the first term vanishes.

As \( W_{ik} = W_i - V_{ik} = W - Z_i - V_{ik} \), Taylor yields

\[ f'(W_{ik}) = f'(W) - (Z_i + V_{ik})f''(W - \rho_{ik}(Z_i + V_{ik})), \]

for some \( \rho_{ik} \in [0, 1] \), and hence

\[ |\mathbb{E}(f'(W_{ik})) - \mathbb{E}(f'(W))| \leq ||f''|| |\mathbb{E}Z_i + V_{ik}|, \]

and the theorem follows from using the bound \( ||f''|| \leq 2||h'|| \).

Using Taylor again,

\[ X_i Z_i f'(W_i) = \sum_{k \in K_i} X_i Z_{ik} f'(W_i) \]

\[ = \sum_{k \in K_i} X_i Z_{ik} \{ f'(W_{ik} + V_{ik} f''(W_{ik} + \theta_{ik} V_{ik})) \} \]

for some \( \theta_{ik} \in [0, 1] \). As \( W_{ik} \) and \( (X_i, Z_{ik}) \) are independent,

\[ \left| \sum_{i \in I} \mathbb{E}(X_i Z_{ik} f'(W_i)) - \sum_{i \in I} \sum_{k \in K_i} \mathbb{E}(X_i Z_{ik}) \mathbb{E}(f'(W_{ik})) \right| \]

\[ \leq ||f''|| \sum_{i \in I} \sum_{k \in K_i} \mathbb{E}|X_i Z_{ik} V_{ik}|. \]

Example: Triangles in Bernoulli random graphs

Let \( G = G(n, p) \) be a Bernoulli random graph on \( n \) vertices with edge probability \( p \) and adjacency matrix \( A \). Assume that \( n \geq 3 \).

Use as index set

\[ \Gamma_n = \{ \alpha = (u, v, w) : 1 \leq u < v < w \leq n \}. \]

Let \( T \) be the number of triangles in \( G \) and \( \alpha = (u, v, w) \). Setting

\[ Y_\alpha = Y_{u,v,w} = a_{u,v}a_{v,w}a_{u,w} \]

which equals 1 if \( u, v, w \) form a triangle, and 0 otherwise, we can write

\[ T = \sum_{\alpha \in \Gamma_n} Y_\alpha. \]

If \( |\{u, v, w\} \cap \{a, b, c\}| \leq 1 \) then \( Y_{u,v,w} \) and \( Y_{a,b,c} \) are independent. Let \( K_\alpha = K_{u,v,w} \) contain \( (u, v, w) \) plus all those indices which share exactly two vertices with \( \{u, v, w\} \).
We standardise

\[ X_\alpha = \frac{Y_\alpha - p^3}{\sqrt{\text{Var}(T)}}. \]

A straightforward calculation shows that

\[ \sigma^2 = \text{Var}(T) = \binom{n}{3} p^3 [1 - p^3 + 3(n - 3)p^2(1 - p)]. \]

We put \( W = \sum_{\alpha \in \Gamma} X_\alpha \),

\[ W_\alpha = \sum_{\beta \not\in K_\alpha} X_\beta \]

and

\[ Z_\alpha = \sum_{\beta \in K_\alpha} X_\beta. \]

Moreover we can write

\[ W_\alpha = \sum_{\gamma \not\in K_\alpha \cup K_\beta} X_\gamma + \sum_{\gamma \in K_\beta \setminus K_\alpha} X_\gamma. \]

Thus we have a decomposition which shows that \( W \) is decomposable.

**Proof.**

Note that \( |K_\alpha| = 3(n - 3) + 1 = 3n - 8 < 3n \). Then, for \( \beta \in K_\alpha \), we have \( |K_\beta \setminus K_\alpha| \leq 2n \). We can use the crude bound that \( |X_\alpha| \leq \frac{1}{s} \). Moreover

\[ \mathbb{E}|X_\alpha| = \frac{2}{s} p^3(1 - p^3), \quad \mathbb{E}(X_\alpha)^2 = \frac{1}{s^2} p^3(1 - p^3) \]

and

\[ \mathbb{E}|X^3_\alpha| = \frac{1}{s^3} (p^3(1 - p^3)^3 + (1 - p^3)p^9) \leq \frac{2p^3}{s^3}. \]

A normal approximation for the number of triangles

Let \( W \) denote the standardised number of triangles in a Bernoulli random graph with \( p \leq \frac{1}{2} \) and \( n \geq 3 \). Then, for any absolutely continuous \( h \),

\[ |\mathbb{E}h(W) - Nh| \leq \frac{27n^5p^3}{s^3}||h'||. \]

When \( p \) does not depend on \( n \) then this expression is of order \( O(n^{-1}) \).

This gives that

\[ \sum_{\alpha \in \Gamma} \mathbb{E}(|X_\alpha Z^2_\alpha|) \leq \sum_{\alpha \in \Gamma} \sum_{\beta \gamma \in K_\alpha} \mathbb{E}(|X_\alpha X_\beta X_\gamma|) \]

\[ \leq \sum_{\alpha \in \Gamma} \sum_{\beta \gamma \in K_\alpha} \mathbb{E}(X^2_\alpha) + \sum_{\alpha \in \Gamma} \sum_{\beta \gamma \in K_\alpha} \frac{2}{s^3} p^3(1 - p^3) \]

\[ \leq \binom{n}{3} \frac{22n^3p^3}{s^3} \leq \frac{11n^5p^3}{3s^3}. \]

Next, as \( \beta \in K_\alpha \iff \alpha \in K_\beta \),

\[ \sum_{\alpha \in \Gamma} \sum_{\beta \in K_\alpha} \mathbb{E}|X_\alpha Z_\beta V_\alpha| \leq \sum_{\alpha \in \Gamma} \sum_{\beta \gamma \in K_\alpha} \mathbb{E}(|X_\alpha X_\beta X_\gamma|) \]

\[ \leq \frac{11n^5p^3}{3s^3}. \]
Lastly

\[ \sum_{\alpha \in \Gamma_n} \sum_{\beta \in K_n} \mathbb{E}\{|X_{\alpha}Z_{\alpha\beta}|\} \mathbb{E}\{|X_{\alpha} + V_{\alpha\beta}|\} \]

\leq 2 \sum_{\alpha \in \Gamma_n} \sum_{\beta, \gamma \in K_n} \mathbb{E}\{|X_{\alpha}X_{\beta}|\} \mathbb{E}\{|X_{\gamma}|\}

\leq \left( \frac{n}{3} \right) \frac{30n^2}{\sigma^3} \left( p^3(1 - p^2) \right)^2 < \frac{5n^5}{\sigma^3} p^3.

When \( p \) does not depend on \( n \) then \( \sigma^2 \) is of order \( n^4 \) and thus the overall bound is of order \( n^{-1} \). Collecting the bounds finishes the proof.

### The dense regime and the sparse regime

For the normal approximation for the number of triangles in a Bernoulli \( G(n, p) \) graph to give a small bound we assumed that \( p \) does not vary with \( n \), or, more precisely, that \( \frac{n^2}{\sigma^3} \) is small. If \( p \) varies with \( n \), then the situation could be very different. For example if \( p = \frac{1}{n} \) then \( \sigma \) is of order 1, and \( \frac{n^5}{\sigma^3} p^3 \) can be very large.

In networks one distinguishes the dense regime and the sparse regime. In the dense regime the expected number of edges grows quadratically with the number of vertices, whereas in the sparse regime the expected number of edges grows linearly with the number of vertices.

### More formally

Recall that \( f = \Theta(g) \) if \( f \) is bounded both above and below by a constant times \( g \), asymptotically.

A Bernoulli \( G(n, p) \) graph is dense if \( \binom{n}{2} p = \Theta(n^2) \), that is, \( p = \Theta(1) \).

A Bernoulli \( G(n, p) \) graph is sparse if \( \binom{n}{2} p = \Theta(n) \), that is, \( p = \Theta \left( \frac{c}{n} \right) \) for some \( c = \Theta(1) \).

A Bernoulli \( G(n, p) \) graph is moderately dense if \( \binom{n}{2} p = \Theta(n \log n) \), that is, \( p = \Theta \left( \frac{c \log n}{n} \right) \) for some \( c = \Theta(1) \).

For counts of small subgraphs in the sparse regime, a Poisson approximation is more appropriate than a normal approximation.

### 4 Stein’s Method for Poisson approximation

The standard reference for this chapter is *Barbour, Holst and Janson (1992).*

Stein’s method for Poisson approximation is often called the Stein-Chen method for Poisson approximation.

Recall the Poisson(\( \lambda \)) distribution

\[ p_k = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, \ldots. \]

This is a discrete distribution, and for approximations on \( \mathbb{Z}^+ \) we would usually take total variation distance:

\[ d_{TV}(P, Q) = \sup_{A \subseteq \mathbb{Z}^+} |P(A) - Q(A)|. \]
A Stein characterisation Part 1

If \( Z \sim \text{Po}(\lambda) \) and if \( g : \mathbb{N}_0 \to \mathbb{R} \) is bounded, then

\[
\mathbb{E}\{\lambda g(Z + 1) - Zg(Z)\} = 0.
\]

To see this:

\[
\lambda \mathbb{E}g(Z + 1) = \lambda \sum_{j \geq 0} g(j + 1) e^{-\lambda \frac{j^2}{j!}} - \sum_{k \geq 1} g(k)e^{-\lambda \frac{k^2}{(k - 1)!}} = \sum_{k \geq 1} kg(k)e^{-\lambda \frac{k^2}{k!}} = \mathbb{E}\{Zg(Z)\}.
\]

Proof of the lemma

For \( g \) given by (4.2).

\[
\lambda g(j + 1) - jg(j) = \frac{j^2}{\lambda^2} e^{\lambda} \sum_{k=0}^{j} e^{-\lambda \frac{k^2}{k!}} [I(k \in A) - \text{Po}(\lambda)(A)]
\]

\[
-\frac{j^2}{\lambda^2} e^{\lambda} \sum_{k=0}^{j-1} e^{-\lambda \frac{k^2}{k!}} [I(k \in A) - \text{Po}(\lambda)(A)]
\]

\[
= \frac{j!}{\lambda^2} e^{\lambda} e^{-\lambda \frac{j}{j!}} [I(j \in A) - \text{Po}(\lambda)(A)]
\]

\[
= [I(j \in A) - \text{Po}(\lambda)(A)].
\]

Proof continued

The uniqueness of the solution follows by putting \( j = 0 \) in (4.1), from which it follows that any function \( h \) such that

\[
\lambda h(j + 1) -jh(j) = I(j \in A) - \text{Po}(\lambda)(A)
\]

must satisfy

\[
h(1) = g(1); \quad \text{then, by iteration in}
\]

\[
\lambda h(j + 1) -jh(j) = \lambda g(j + 1) - jg(j),
\]

it follows that

\[
h(j) = g(j) \quad \text{for all } j. \quad \text{This completes the proof}.
\]

The next lemma shows that this solution is bounded.
Lemma 4.2
For the solution of the Stein equation
\[ \lambda g(j + 1) - j g(j) = I(j \in A) - Po\{\lambda\}(A) \]
given in Lemma 4.1 we have that
\[ \sup_j |g(j)| \leq \min(1, \lambda^{-\frac{1}{2}}) \]
and
\[ \Delta g := \sup_j |g(j + 1) - g(j)| \leq \min(1, \lambda^{-1}) \]
Here \( \lambda^{-1} \) is often referred to as the magic factor.
The proof can be found in the book by Barbour, Holst and Janson, p.7.

The Stein-Chen method
Stein’s method for Poisson approximation is also called the Stein-Chen method. Summarised: Use sets \( A \subset \mathbb{Z}^+ \) and as Stein equation
\[ \lambda g(j + 1) - j g(j) = I(j \in A) - Po\{\lambda\}(A) \]
For any random variable \( W \) we hence have
\[ \lambda g(W + 1) - W g(W) = I(W \in A) - Po\{\lambda\}(A) \]
and taking expectations
\[ d_{TV}(\mathcal{L}(W), Po\{\lambda\}) \leq \sup_g \mathbb{E} \{ \lambda g(W + 1) - W g(W) \} \]
Here the sup is over all functions \( g \) which solve the Stein equation.

A Stein characterisation Part 2
Lemma 4.3
Suppose that for all bounded \( g : \mathbb{N}_0 \rightarrow \mathbb{R} \)
\[ \mathbb{E}\{\lambda g(W + 1) - W g(W)\} = 0. \]
Then \( W \sim Po(\lambda) \).
Proof: Take \( g \) as in Lemma 4.1, then
\[ 0 = \lambda \mathbb{E}\{g(W + 1) - W g(W)\} = \mathbb{P}(W \in A) - Po\{\lambda\}(A). \]
As the right-hand side vanishes for all sets \( A \subset \mathbb{Z}^+ \), it follows that \( W \) must have the Poisson distribution. The proof is complete.

Example: sum of independent Bernoulli variables
Let \( X_1, X_2, \ldots \) be independent Bernoulli(\( p_i \)) and \( W = \sum_{i=1}^{n} X_i \).
With \( \lambda = \sum_{i=1}^{n} p_i \)
\[ \mathbb{E}(W g(W)) = \sum_i \mathbb{E}(X_i g(W)) = \sum_i \mathbb{E}(X_i g(W - X_i + 1)) = \sum_i p_i \mathbb{E}(g(W - X_i + 1)) \]
and thus
\[ \mathbb{E}\{\lambda g(W + 1) - W g(W)\} = \sum_i p_i \mathbb{E}(g(W + 1) - g(W - X_i + 1)) = \sum_i p_i^2 \mathbb{E}(g(W + 1) - g(W - X_i + 1)|X_i = 1). \]
We conclude that
\[ |\mathbb{E}\{\lambda g(W + 1) - W g(W)\}| \leq \Delta g \sum_i p_i^2 \leq \min(1, \lambda^{-1}) \sum_i p_i^2. \]
The local approach

Again the argument generalises to local dependence. Let $I$ be an index set and let $X_\alpha, \alpha \in I$ be Bernoulli random variables, $X_\alpha \sim \text{Be}(p_\alpha)$, and put $W = \sum_{\alpha \in I} X_\alpha$. Let $\lambda = \sum_{\alpha \in I} p_\alpha$. Suppose that for each $\alpha \in I$ there exists a set $A_\alpha \subset I$ such that $X_\alpha$ is independent of $\sum_{\beta \notin A_\alpha} X_\beta$. Define

$$\eta_\alpha = \sum_{\beta \in A_\alpha} X_\beta.$$

**Theorem 4.4**

With $\lambda = E(W)$,

$$d_{TV}(L(W), Po(\lambda)) \leq \sum_{\alpha \in I} [(p_\alpha E(\eta_\alpha) + E(X_\alpha(\eta_\alpha - X_\alpha))] \min (1, \lambda^{-1}) .$$

**Proof**

Let $A \subset \mathbb{Z}^+$ and let $g$ be the corresponding solution of the Poisson Stein equation (4.1). Then

$$P(W \in A) - Po(\lambda)(A) = \lambda E g(W + 1) - EW g(W).$$

Note that

$$EW g(W) = \sum_{\alpha \in I} E X_\alpha g(W)$$

$$= \sum_{\alpha \in I} E X_\alpha g(W - X_\alpha + 1).$$

**Bounding the error terms**

Here

$$|R_1| = \left| \sum_{\alpha \in I} p_\alpha \{ E(g(W_\alpha + 1) - g(W - X_\alpha + 1)) \} \right|$$

$$\leq \sum_{\alpha \in I} p_\alpha \Delta g E(\eta_\alpha)$$

and

$$|R_2| = \left| \sum_{\alpha \in I} E\{ X_\alpha(g(W - X_\alpha + 1) - g(W_\alpha - X_\alpha + 1)) \} \right|$$

$$\leq \sum_{\alpha \in I} \Delta g E\{ X_\alpha(\eta_\alpha - X_\alpha) \}.$$

This finished the proof.
Example: Triangles in Bernoulli random graphs

Let $G = \mathcal{G}(n, p)$ be a Bernoulli random graph on $n$ vertices with edge probability $p$ and adjacency matrix $A$. Use as index set

$$\Gamma_n = \{ \alpha = (u, v, w) : 1 \leq u < v < w \leq n \}.$$

Let $X$ be the number of triangles in $G$ and $\alpha = (u, v, w)$. Setting

$$X_\alpha = X_{u,v,w} = a_{u,v}a_{v,w}a_{u,w}$$
which equals 1 if $u, v, w$ form a triangle, and 0 otherwise, we can write

$$W = \sum_{\alpha \in \Gamma_n} X_\alpha.$$

Note that $E(W) = \binom{n}{3} p^3$.

Proof

If $|\{u, v, w\} \cap \{a, b, c\}| \leq 1$ then $X_{u,v,w}$ and $X_{a,b,c}$ are independent. Let $A_\alpha = K_{u,v,w}$ contain $(u, v, w)$ plus all those indices which share exactly two vertices with $\{u, v, w\}$. Recall that $|A_\alpha| < 3n$.

Here $p_\alpha = p^3$ and

$$E(\eta_\alpha) \leq 3np^3.$$

Moreover

$$E\{X_\alpha(\eta_\alpha - X_\alpha)\} = \sum_{\beta \in A_\alpha, \beta \neq \alpha} p^5 < 3np^5.$$

This finishes the proof.

Poisson approximation for Triangles

Theorem 4.5
With $\lambda = \binom{n}{3} p^3$,

$$d_{TV}(\mathcal{L}(W), \text{Po}(\lambda)) \leq \binom{n}{3} p^3 (3np^3 + 3np^2) \min(1, \lambda^{-1}).$$

Thus, if $p = \frac{c}{n}$ then the bound is of order $cn^{-1}$.

If $p$ is much smaller than $\frac{1}{n}$ then the expected number of triangles is very small.

Similarly we can show Poisson and Normal approximations for counts of other small subgraphs. Note that the bounds tell us which of the approximations, if any, may be appropriate.

The approach generalises to non-equal edge probabilities, as long as the independence of the edges is maintained.
The clustering coefficient in Bernoulli random graphs

Using the Stein-Chen method, K. Lin (2007) showed that, for the average clustering coefficient in Bernoulli random graphs,

\[ C = \frac{1}{n} \sum_{v} C(v), \]

it is also true that, in distribution,

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

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

4.18

5 Threshold behaviour

In this chapter we shall see that Bernoulli random graphs display a threshold-type behaviour for the appearance of small subgraphs. For example for the number of triangles when \( np \) is much smaller than 1, then we do not expect to see triangles; when it is much larger than 1 we would expect to see many triangles. Here is a simulation.

5.1

A mathematical formulation

In a Bernoulli \( \mathcal{G}(n,p) \) random graph with \( p = \frac{c}{n} \), when \( c < 1 \) then the largest component has size \( O(\log n) \) and when \( c > 1 \) then the largest component has linear size with probability tending to 1 as \( n \) grows to infinity.

5.2

5.3
The depth-first search algorithm

To understand the threshold behaviour we use the approach by M. Krivelevich and B. Sudakov (2012). This is an example how an algorithm can be analysed to yield a probabilistic result.

The Depth-First Search (DFS) algorithm labels vertices of a vertex set \( V = \{1, \ldots, n\} \) of a graph \( G = (V, E) \) with three different labels - \( S \) for explored, \( T \) for unvisited, and \( U = V \setminus (S \cup T) \) for under investigation.

### Pseudo-code

Start: \( S = U = \emptyset, T = V \)

If \( U \neq \emptyset \): Let \( v \) be the last vertex added to \( U \).

1. If \( v \) has a neighbour \( u \) in \( T \) then delete \( u \) from \( T \) and insert it into \( U \).
2. If \( v \) does not have a neighbour in \( T \) then \( v \) is moved from \( U \) to \( S \).

If \( U = \emptyset \): choose the first vertex of \( T \) and move it from \( T \) to \( U \).

If \( U \cup T = \emptyset \): query all the remaining pairs of vertices in \( S = V \), then stop.

### Properties of DFS

The DFS algorithm can be used to reveal the connected components of \( G \). It starts revealing a connected component \( C \) at the moment the first vertex of \( C \) gets into (empty beforehand) \( U \) and completes discovering all of \( C \) when \( U \) becomes empty again.

A period of time between two consecutive empty-ings of \( U \) is called an epoch.

Each epoch corresponds to one connected component of \( G \).
DFS on a Bernoulli random graph

In a Bernoulli random graph at each stage of the algorithm an edge \( e \) is queried, and the answer is positive with probability \( p \).

Denote by \( X_i \in \{0, 1\} \) the outcome of the \( i \)th query in the DFS algorithm; then \( X_i, i = 1, 2, \ldots N \) are independent Bernoulli\((p)\)-variables. Here \( N = \binom{n}{2} \).

Denote by \( S(t), T(t), U(t) \) the sets \( S, T, U \) at time \( t \).

Some observations

Since the addition of a vertex (except the first vertex) in a connected component to \( U \) is caused by a positive answer to a query,

\[
|U(t)| \leq 1 + \sum_{i=1}^{t} X_i.
\]

At all times,

\[
|U(t)| = n - |S(t)| - |T(t)|.
\]

As long as \( T \neq \emptyset \), every positive answer to a query results in a vertex to be moved from \( T \) to \( U \). If after \( t \) queries still \( T(t) \neq \emptyset \), we have

\[
|S(t) \cup U(t)| \geq \sum_{i=1}^{t} X_i.
\]

Heuristic

If \( np < 1 \) then we see few edges, and epochs will be short.

If \( np > 1 \) then at any time the set of unexplored vertices \( T \) and the set of explored vertices \( S \) cannot be both large since there are no edges between them. Therefore, the set \( U \) of vertices under exploration (which form a path) must at some point be large. Hence the graph must contain a long path, which is only possible if it has a large connected component.

To make this argument precise we use Chebychev’s inequality and the Chernoff inequality:

Lemma 5.1
Let \( X \) have the \( \text{Binomial}(N, p) \)-distribution. Then, for all \( n \), and \( k > 0 \),

1. \( \mathbb{P}(|X - Np| \geq n^{\frac{2}{3}}) \leq n^{-\frac{4}{3}}Np(1 - p) \).

2. \( \mathbb{P}(X - Np > k) \leq e^{-\frac{k^2}{2(Np + k/N)}} \).

The proof is an exercise.
Corollary 5.2
Let $X_i, i = 1, \ldots, N$ be a sequence of i.i.d. Bernoulli trials with success probability $p$. Let $k, \ell \in \mathbb{N}$ such that $k > \ell p$. Then the probability that there is no (discrete) interval of length $\ell$ in \{1, \ldots, N\} in which at least $k$ of the random variables $X_i$ take the value 1 is at least 
\[(N - \ell + 1)e^{-\frac{(k-\ell p)^2}{2(\ell p + (k-\ell p)/3)}}.\]

Proof
Let $W$ denote the number of intervals of length $\ell$ in \{1, \ldots, N\} in which at least $k$ of the random variables $X_i$ take the value 1.

Using Markov's inequality,
\[P(W \geq 1) \leq E(W).\]

Now
\[E(W) = (N - \ell + 1)P\left(\sum_{j=1}^{\ell} X_i > k\right).\]

From Lemma 5.1 we have that
\[P\left(\sum_{j=1}^{\ell} X_i > k\right) = P\left(\sum_{j=1}^{\ell} X_i - \ell p > k - \ell p\right) \leq e^{-\frac{(k-\ell p)^2}{2(\ell p + (k-\ell p)/3)}}.\]

Combining these bounds yields the assertion.

Now we can prove the main result.

Theorem 5.3
Let $\varepsilon > 0$, and let $G \sim G(n, p)$.
1. Let $\varepsilon < 1$ and let $p = \frac{1 - \varepsilon}{n}$. Then the probability that all connected components of $G$ are of size at most $\frac{5}{2} \varepsilon^2 \log n$ tends to 1 as $n \to \infty$.
2. Let $\varepsilon < \sqrt{\frac{2}{5}}$ and let $p = \frac{1 + \varepsilon}{n}$. Then the probability that $G$ contains a path of length at least $\frac{1}{3} \varepsilon^2 n$ tends to 1 as $n \to \infty$.

Proof: Part 1
For Part 1, let $p = \frac{1 - \varepsilon}{n}$. Assume that $G$ contains a connected component $C$ with more than $\frac{5}{2} \log n$ vertices. Let $k + 1 = \lfloor \frac{5}{2} \log n \rfloor + 1$ be the smallest integer which is larger than $\frac{5}{2} \log n$.

Look at the epoch when $C$ was created. Consider the moment inside this epoch when the algorithm has found the $(k + 1)$th vertex of $C$ and is about to move it to $U$. Write $\Delta = S \cap C$ at this moment. Then $|\Delta \cup U| = k$, and thus the algorithm got exactly $k$ positive answers to its queries to random variables $X_i$ during the epoch.

At that moment during the epoch only pairs of edges touching $\Delta \cup U$ have been queried, and the number of such pairs is therefore at most
\[\binom{k}{2} + k(n - k) < kn.\]
From Corollary 5.2 with $\ell = kn$ and $N = \binom{n}{2}$ we get that the probability of such an event is at most
\[
\binom{n}{2} e^{-\frac{(k-nk)^2}{2(kn+k-nk)}} = \binom{n}{2} e^{-\frac{(kn)^2}{2kn}} = \binom{n}{2} e^{-\frac{k^2}{2(1-\frac{1}{3})}} \leq \binom{n}{2} e^{2(1-\frac{1}{3})} n^{-\frac{5}{2}} \leq e^{2(1-\frac{1}{3})} n^{-\frac{5}{2}} = e^{\frac{3}{2}n^{-\frac{1}{2}}}
\]
as $\varepsilon < 1$. This quantity tends to 0 as $n \to \infty$.

Indeed,
\[
\mathbb{P}
\left( |U(t)| > \frac{1}{3} n \right) \leq \mathbb{P} \left( 1 + \sum_{i=1}^{t} X_i > \frac{1}{3} n \right) \leq \exp \left\{ -\frac{\left( \frac{1}{3} n - (1+tp) \right)^2}{2 \left( tp + \frac{1}{3} \left( \frac{1}{3} n - (1+tp) \right) \right)} \right\}.
\]

Now
\[
\left( \frac{1}{3} n - (1+tp) \right)^2 \geq \frac{1}{9} (n^2 - 3n)
\]
and
\[
tp + \frac{1}{3} \left( \frac{1}{3} n - (1+tp) \right) \leq (\varepsilon(1+\varepsilon) + \frac{1}{3}) n \leq \left( \frac{10}{9} + \frac{1}{3} \right) n
\]
and so
\[
\frac{\left( \frac{1}{3} n - (1+tp) \right)^2}{2 \left( tp + \frac{1}{3} \left( \frac{1}{3} n - (1+tp) \right) \right)} \geq \frac{n}{13}.
\]

Note that $|S(t)||T(t)| \leq N_0$ as the algorithm must have queried all edges between $S(t)$ and $T(t)$. Moreover
\[
|T(t)| = n - |S(t)| - |U(t)| = \frac{2}{3} n - |U(t)|. \quad (5.1)
\]

If simultaneously $|U(t)| < \frac{1}{3} n$ and $|S(t)| = \frac{1}{3} n$ then $|T(t)| \geq \frac{1}{3} n$, which yields $|S(t)||T(t)| \geq \frac{1}{9} n > N_0$ for $n$ large enough, as $\varepsilon < \frac{\sqrt{3}}{3}$, and so this would give a contradiction. Hence
\[
\mathbb{P} \left( |S(N_0)| \geq \frac{1}{3} n; |U(t)| < \frac{1}{3} n \right) = 0.
\]

Proof of Part 2

Now let $p = \frac{1+\varepsilon}{n}$. Let $N_0 = \left\lfloor \frac{1}{2} \varepsilon n^2 \right\rfloor + 1$. We claim that after the first $N_0$ queries of the DFS algorithm we have $|U| \geq \frac{1}{5} \varepsilon^2 n$, with the contents of $U$ forming a path of desired length at that moment. Assume that the DFS algorithm has carried out the first $N_0$ queries.

First we would like to show that $|S(N_0)| < \frac{1}{3} n$ with probability at least $1 - e^{-\frac{1}{5} \varepsilon^2 n}$. Assume to the contrary that $|S(N_0)| \geq \frac{1}{3} n$. Then there is a time $t < N_0$ such that $|S(t)| = \frac{1}{3} n$; we must have that $t \geq \frac{1}{3} n$. For this time $t$, $|U(t)| \leq \frac{1}{3} n$ with high probability.
Now

\[ 1 \geq P\left( |S(N_0)| \geq \frac{1}{3}n \right) \cup \{ |U(t)| < \frac{1}{3}n \} \]

\[ = P\left( |S(N_0)| \geq \frac{1}{3}n \right) + P\left( |U(t)| < \frac{1}{3}n \right) + 0 \]

\[ \geq P\left( |S(N_0)| \geq \frac{1}{3}n \right) + 1 - e^{-\frac{n}{13}} \]

and so

\[ P(|S(N_0)| \geq \frac{1}{3}n) \leq e^{-\frac{n}{13}} \]

as claimed.

Now we consider only events on the set \( \{ |S(N_0)| < \frac{1}{3}n \} \). If \( |U(N_0)| \leq \frac{1}{5} \varepsilon^2 n \) then \( T \neq 0 \) by (5.1), and so the algorithm is still revealing the connected components of \( G \), and each positive answer it got resulted in moving a vertex from \( T \) to \( U \). From Lemma 5.1 we get that, with probability at least \( 1 - 2n^{-\frac{1}{4}} \), the number of positive answers at that point is at least \( N_0p - n^{\frac{2}{3}} \), and on that event,

\[ |S(N_0) \cup U(N_0)| \geq N_0p - n^{\frac{2}{3}}. \]

Now suppose that \( |U(N_0)| \leq \frac{1}{5} \varepsilon^2 n \); we would like to show that this is not possible when \( |S(N_0) \cup U(N_0)| \geq N_0p - n^{\frac{2}{3}} \). To see this, if \( |U(N_0)| \leq \frac{1}{5} \varepsilon^2 n \) and \( |S(N_0) \cup U(N_0)| \geq N_0p - n^{\frac{2}{3}} \), then

\[ |S(N_0)| \geq N_0p + \frac{1}{5} \varepsilon^2 n - n^{\frac{2}{3}}. \]

From (5.1), we would then get that

\[ N_0 \geq |S| \left( n - |S| - \frac{1}{5} \varepsilon^2 n \right) \]

\[ \geq \left( N_0p + \frac{1}{5} \varepsilon^2 n - n^{\frac{2}{3}} \right) \left( n - N_0p + n^{\frac{2}{3}} \right). \]

As \( N_0p \sim \frac{1}{2} \varepsilon (1 + \varepsilon) n \) we see that for \( n \) large enough, the right-hand side exceeds \( N_0pn = N_0(1 + \varepsilon) \), giving a contraction.

Hence, on the event \( |S(N_0) \cup U(N_0)| \geq N_0p - n^{\frac{2}{3}} \), we must have \( |U(N_0)| > \frac{1}{5} \varepsilon^2 n \), and hence we have a path of length at least \( \frac{1}{5} \varepsilon^2 n \) with probability tending to 1 as \( n \to \infty \). This finishes the proof.

For studying global properties such as the average shortest path length we use another set of tools: Yule processes.
6 Yule Processes

The main references for this chapter are the book by S.M. Ross (1995) and Steven Lalley’s lecture notes at http://galton.uchicago.edu/~lalley/Courses/312/Branching.pdf.

A Yule process is a process in which each member of a population acts independently and gives birth at an exponential rate $\lambda$. No one ever dies.

Let $X(t)$ represent the population size at time $t$. Then $\{X(t), t \geq 0\}$ is a (continuous-time) Markov chain with states $1, 2, \ldots$, and the only transition from state $i$ is to state $i + 1$.

Let $T_i$ denote the time it takes for the population size to go from $i$ to $i + 1$. Then the $T_i, i \geq 1$ are independent and $T_i$ is exponentially distributed with parameter $\lambda_i$:

$$P(T_i \leq t) = 1 - e^{-\lambda_i t}, \quad (6.1)$$

for $t \geq 0$.

### The birth times

Starting with a single individual, the time of the $i^{th}$ birth is $S_i = \sum_{j=1}^{i} T_j$. The following result holds.

**Lemma 6.1**

Consider a Yule process with $X(0) = 1$. Then, given that $X(t) = n + 1$, the birth times $S_1, \ldots, S_n$ are distributed as the ordered values from a sample of size $n$ from a population having density

$$f(x) = \begin{cases} \frac{\lambda e^{-\lambda(t-x)}}{1-e^{-\lambda t}}, & 0 \leq x \leq t; \\ 0, & \text{otherwise.} \end{cases}$$

**Proof (heuristical)**

Assume that we can treat densities as if they were probabilities. Then, for $0 = s_0 \leq s_1 \leq s_2 \leq \ldots \leq s_n \leq t$ we have

$$P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n | X(t) = n + 1) = \frac{1}{P(X(t) = n + 1)} P(S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n, X(t) = n + 1)$$

$$= \frac{1}{P(X(t) = n + 1)} P(T_i = s_i - s_{i-1}, i = 1, \ldots, n, T_{n+1} > t - s_n)$$

$$= \frac{1}{P(X(t) = n + 1)} \prod_{i=1}^{n} \left\{ i\lambda e^{-\lambda(s_i-s_{i-1})} \right\} e^{-(n+1)\lambda(t-s_n)}$$

$$= n!\lambda^n \frac{1}{P(X(t) = n + 1)} e^{-\lambda t} \prod_{i=1}^{n} e^{-\lambda(t-s_i)}.$$
AS $X(t)$ follows a geometric distribution with mean $e^{\lambda t}$ we have
\[ \mathbb{P}(X(t) = n + 1) = e^{-\lambda t} \left(1 - e^{-\lambda t}\right)^n. \]
Hence
\[ \mathbb{P} (S_1 = s_1, S_2 = s_2, \ldots, S_n = s_n | X(t) = n + 1) = n! \prod_{i=1}^{n} f(s_i), \]
as required.

How many are there?

**Lemma 6.2**

*Start the process with one individual. Then*
\[ \mathbb{E}X(t) = e^{\lambda t}. \]

The proof is an exercise.

Proof.
Condition on $X(t)$:
\[ \mathbb{E}[A(t)|X(t) = n + 1] = a_0 + t + \mathbb{E} \left\{ \sum_{i=1}^{n} (t - S_i)|X(t) = n + 1 \right\} = a_0 + t + n \int_{0}^{t} (t - x) \frac{\lambda e^{-\lambda(t-x)}}{1 - e^{-\lambda t}} \, dx, \]
where we used Lemma 6.1. Thus
\[ \mathbb{E}[A(t)|X(t)] = a_0 + t + (X(t) - 1) \frac{1 - e^{-\lambda t} - \lambda e^{-\lambda t}}{\lambda (1 - e^{-\lambda t})}. \]
Taking expectations and using Lemma 6.2 gives
\[ \mathbb{E}[A(t)] = a_0 + t + (e^{\lambda t} - 1) \frac{1 - e^{-\lambda t} - \lambda e^{-\lambda t}}{\lambda (1 - e^{-\lambda t})} = a_0 + \frac{e^{\lambda t} - 1}{\lambda}. \]

How old are they?

Start the process with one individual. Let $a_0$ be the age at time $t = 0$ of the initial individual. Then the sum $A(t)$ of all ages at time $t$ is
\[ A(t) = a_0 + t + \sum_{i=1}^{X(t)-1} (t - S_i). \]

**Lemma 6.3**

*We have that*
\[ \mathbb{E}A(t) = a_0 + \frac{e^{\lambda t} - 1}{\lambda}. \]

Note that if $a_0 = 0$ then the average age of an individual which is present at time $t$ is roughly
\[ \frac{\mathbb{E}A(t)}{\mathbb{E}X(t)} \approx \frac{1}{\lambda}. \]
Most individuals are very young.
Asymptotic growth

Theorem 6.4
Start the process with 1 individual. Then
\[ \frac{X(t)}{e^{\lambda t}} \to W \quad \text{almost surely}, \]
where \( W \) has the exponential distribution with parameter 1.

The proof of this fact is not straightforward. It is completed in a number of steps.

A general strong convergence result

Recall that \( S_m \) is the sum of \( m \) independent random variables.

Theorem 6.6
Let \( X_j \) be independent random variables with mean zero and finite variances \( \text{Var}(X_j) = \sigma_j^2 \). Assume that \( \sum_{j=1}^{\infty} \sigma_j^2 = \sigma^2 < \infty \). Then
\[ \lim_{n \to \infty} \sum_{j=1}^{n} X_j = S \]
exists and is finite with probability 1, and the limit variable \( S \) has mean 0 and variance \( \sigma^2 \).

The proof can be found in S. Lalley’s 2012 lecture notes.

Start of the proof

Lemma 6.5
With \( S_m \) the time of the \( m^{\text{th}} \) birth,
\[ \frac{X(t)}{e^{\lambda t}} \to W \quad (t \to \infty) \iff S_m - \log m \to -\log W \quad (m \to \infty). \]

Proof. First note that
\[ \frac{X(t)}{e^{\lambda t}} > \alpha \iff Z_i > e^{\lambda \alpha} \iff S_{e^{\lambda \alpha}} < t \]
and setting \( m = e^{\lambda \alpha} \) so that \( t = \log m - \log \alpha \), the above is equivalent to \( S_m < \log m - \log \alpha \), or
\[ S_m - \log m < -\log \alpha. \]

Hence the lemma follows.

Applying this result: a limit exists

Lemma 6.7
There is a random variable \( V \) such that
\[ S_m - \log m \to V \]
with probability 1.

Proof. We have that
\[ S_m = \sum_{i=1}^{m} T_i, \]
so that \( E S_m = \sum_{i=1}^{m} (\lambda_i)^{-1} \sim \lambda^{-1} \log m \) and and
\( \text{Var} T_i = (\lambda_i)^{-2} \), and
\[ \sum_{j=1}^{\infty} (\lambda_i)^{-2} \to \lambda^{-2} \zeta(2) < \infty \]
and with Theorem 6.6 we may conclude that the lemma holds.
The moment-generating function

To determine the distribution of \( V \), or equivalently of \( W = e^V \), we use the moment-generating function.

The moment generating function (m.g.f) of a random variable \( X \) is

\[
M_X(t) = \mathbb{E}e^{tX},
\]

where \( t \) is a real variable.

Distributions are determined by their moment-generating functions.

The moment-generating function of an exponential distribution with parameter \( \lambda \) is

\[
(1 - \frac{t}{\lambda})^{-1}; \quad t < \lambda.
\]

A branching argument

For \( X(t) \) starting with one individual, consider the time \( T_1 \) of the first birth. Then we can view the process as the initial individual disappearing and being replaced by two identical offspring particles, each of which creating their own Yule process; call these \( Z'_{t-T_1} \) and \( Z''_{t-T_1} \). Hence

\[
X(t) = 1(t < T_1) + 1(t \geq T_1)(Z'_{t-T} + Z''_{t-T}).
\]

Calculating the m.g.f.

So, for all \( \varepsilon > 0 \),

\[
M_W(t) = \int_0^1 \mathbb{E}e^{ut(W' + W'')} du
= \int_0^{1-\varepsilon} \mathbb{E}e^{ut(W' + W'')} du + \int_{1-\varepsilon}^1 \mathbb{E}e^{ut(W' + W'')} du
= (1 - \varepsilon) \int_0^1 \mathbb{E}e^{ut(1-\varepsilon)(W' + W'')} du
+ \int_{1-\varepsilon}^1 \mathbb{E}e^{ut(W')} \mathbb{E}e^{ut(W'')} du
= (1 - \varepsilon)M_W(t(1 - \varepsilon)) + \int_{1-\varepsilon}^1 M_W(ut)^2 du.
\]
Subtract $M_W(t(1-\varepsilon))$ from both sides and divide by $\varepsilon$ to get

$$\frac{1}{\varepsilon} (M_W(t) - M_W(t(1-\varepsilon))) = -M_W(t(1-\varepsilon)) + \frac{1}{\varepsilon} \int_{1-\varepsilon}^{1} M_W(ut)^2 \, du.$$ 

Now let $\varepsilon \to 0$ to conclude that

$$-tM_W'(t) = -M_W(t) + M_W(t)^2.$$ 

The only solution to this equation is a function of the form $\frac{a}{a-t}$, thus $W$ must have an exponential distribution. The mean of $W$ is

$$E \left\{ \frac{X(t)}{e^{\lambda t}} \right\} = 1$$

and hence the assertion of the theorem follows.

**Proof**

Since $T$ is a stopping time, for any integer $k \geq 1$ the event 

$\{T \geq k\} = \{T > k - 1\}$

depends only on the random variables $X_i$ for $i < k$, and hence is independent of $X_k$. In particular, if $j < k$ then

$$E(X_jX_k 1(T \geq k)) = EX_jE(X_k 1(T \geq k)) = 0.$$ 

Now suppose that $T$ is a bounded stopping time; then $T \leq m$ almost surely for some integer $m \geq 1$.

**Additional material: proof of Theorem 6.6**

First we use Wald’s Theorem:

**Theorem 6.8**

Let $X_1, X_2, \ldots$ be independent random variables with zero means and finite variances $\sigma_j^2$, and set $S_n = \sum_{i=1}^{n} X_j$. Let $T$ be a stopping time such that there is an integer $m$ such that $\mathbb{P}(T \leq m) = 1$. Then

$$E(S_T^2) = E \left\{ \sum_{j=1}^{m} \sigma_j^2 \right\}.$$ 

Thus,

$$E(S_T^2) = E \left( \sum_{k=1}^{m} X_k 1(T \geq k) \right)^2$$

$$= \sum_{k=1}^{m} E \left\{ X_k^2 1(T \geq k) \right\} + 2 \sum_{1 \leq j < k \leq m} E \{X_jX_k 1(T \geq k)\}$$

$$= \sum_{k=1}^{m} E \left\{ X_k^2 1(T \geq k) \right\}$$

$$= \sum_{k=1}^{m} E \left( X_k^2 \right) E \{1(T \geq k)\}$$

$$= \sum_{k=1}^{m} \sigma_k^2 \mathbb{P}(T \geq k)$$

$$= E \left( \sum_{k=1}^{m} T \sigma_k^2 \right),$$

as required.
Lemma 6.9

(A Maximal Inequality). Assume that in addition, \( \sum_{j} \sigma_{j}^{2} = \sigma^{2} < \infty \). Then

\[
\mathbb{P}\{\sup_{n \geq 1} |S_n| \geq \alpha\} \leq \frac{\sigma^{2}}{\alpha^{2}}. \tag{6.2}
\]

**Proof.** Let \( T = \inf\{n : |S_n| \geq \alpha\} \). Then

\[
\mathbb{P}\{\sup_{n \geq 1} |S_n| \geq \alpha\} = \mathbb{P}\{T < \infty\}.
\]

Pick \( m \); then for \( \min(T, m) \) Wald’s Theorem applies and we get

\[
\mathbb{E}\left(S_{\min(T,m)}^{2}\right) \leq \sigma^{2}.
\]

Hence

\[
\alpha^{2}\mathbb{P}(T \leq m) \leq \mathbb{E}\left(S_{\min(T,m)}^{2}\right) \mathbb{1}\{T \leq m\} \leq \mathbb{E}\left(S_{\min(T,m)}^{2}\right) \leq \sigma^{2}.
\]

Recall: the Borel-Cantelli Lemma

Let \( E_1, E_2, \ldots \) denote a sequence of events. If \( \sum_{i=1}^{\infty} \mathbb{P}(E_i) < \infty \) then

\[
\mathbb{P}(\text{an infinite number of the } E_i \text{ occur}) = 0.
\]

The proof follows from expressing the event that an infinite number of the \( E_i \) occur as \( \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} E_i \), and note that \( \cup_{i=n}^{\infty} E_i, n \geq 1 \), is a decreasing sequence of events. Hence

\[
\mathbb{P}\left(\cap_{n=1}^{\infty} \cup_{i=n}^{\infty} E_i\right) = \mathbb{P}\left(\lim_{n \to \infty} \cup_{i=n}^{\infty} E_i\right) \leq \lim_{n \to \infty} \mathbb{P}\left(\cup_{i=n}^{\infty} E_i\right) \leq \lim_{n \to \infty} \sum_{i=1}^{\infty} \mathbb{P}(E_i).
\]

Finishing the proof of Theorem 6.6

Finally, using (6.2), for every \( k \geq 1 \) we can find an \( n_k \) such that \( \sum_{j \geq n_k} \sigma_{j}^{2} \leq 8^{-k} \). By (6.2),

\[
\mathbb{P}\{\sup_{n \geq n_k} |S_n - S_{n_k}| \geq 2^{-k}\} \leq 8^{-k}4^{k} = 2^{-k}.
\]

As \( \sum_{k} 2^{-k} < \infty \) we may apply the Borel-Cantelli theorem, and hence it follows that \( S_n \) converges with probability 1, and therefore has a finite limit \( W \).

7 Shortest paths in small-world networks

Branching processes can be used to approximate the distribution of the shortest paths in small-world networks.

Let us consider the following continuous circle model: Let \( C \) be a circle of circumference \( L \). On this circle a Poisson \( (L \rho / 2) \) number of shortcuts added uniformly. Chords between points have length zero. Assume \( L \rho > 1 \).

Let \( D \) denote shortest distance between two randomly chosen points.
The growth process

Pick a point P at random from C.
The process walks from P at the same speed $2\rho$ in all possible directions, taking any shortcut that it can find. Denote by $R(t)$ the set of points that can be reached from P within time $t$.
Taking shortcut means initially creating a new intervals on the circle, but the process will in due time meet some areas that it has covered before, introducing dependence.

For the pure growth process $S(t)$ started at P let $M(t)$ number of intervals at time $t$, and let $s(t)$ total length of the circle covered at time $t$. For a Yule process with birth rate $2\rho$ we know from Lemma 6.2 and Lemma 6.3 that

$$E_M(t) = e^{2\rho t};$$
$$E_s(t) = \frac{1}{\rho} (e^{2\rho t} - 1).$$

A Yule process approximation

Compare this process to a pure growth (Yule) process $S(t)$: start at P, grow at rate $2\rho$.
Pick another point $P'$ at random from C. let an independent pure growth process run from that point, then the time at which the two independent pure growth processes will meet is approximately $\frac{1}{2}D$.

Heuristic for the shortest path

Let $N(t)$, $u(t)$ denote the corresponding quantities for the pure growth process started at the point $P'$. Run both pure growth processes from time 0. Then at time $t$ approximately $e^{4\rho t}$ pairs of intervals, and each has approximately length $\frac{1}{\rho}$. Let $V_t$ denote the number of intersecting pairs of intervals at time $t$, one from the process started at P, the other from the process started at $P'$. Then

$$V_t \approx \frac{2}{L\rho} e^{4\rho t}.$$

Thus the time scale of first encounter is

$$\tau_x = \frac{1}{2\rho} \left\{ \frac{1}{2} \log(L\rho) + x \right\}$$

and

$$V_{\tau_x} \approx 2e^{2x}.$$
Poisson approximation

We now derive a Poisson approximation for the number of intersections of intervals when started from \( P, P' \). Assume that there are \( m \) intervals \( I_1, \ldots, I_m \), with given lengths \( s_1, \ldots, s_m \), in the first process, and \( n \) intervals \( J_1, \ldots, J_n \), with given lengths \( u_1, \ldots, u_n \), in the second process, positioned uniformly on \( C \), independently. Let

\[
X_{ij} = 1(I_i \cap J_j \neq \emptyset)1(I_i \not\subseteq J_j)1(J_j \not\subseteq I_i)
\]

and put

\[
V = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}.
\]

Then

\[
EV = \lambda = \frac{2}{L} \sum_{i=1}^{m} \sum_{j=1}^{n} \min(s_i, u_j).
\]

Corollary 7.1

\[
d_{TV}(\mathcal{L}(V), Po(\lambda)) \leq \frac{4}{L}(m + n)l_{su}.
\]

Put

\[
l_{su} = \max\{\max_i s_i, \max_j u_j\}.
\]

Then

\[
EX_{ij} \leq \frac{2}{L} l_{su},
\]

and the \( X_{ij} \) are independent if indices do not intersect.

Thus we are in the setting of local dependence, and we can apply the Stein-Chen method for Poisson approximation with Theorem 4.4.

Proof

Let \( I = \{(i, j) : i \in \{1, \ldots, m\}, j \in \{1, \ldots, n\}\} \) and

\[
A_{(i,j)} = \{(k, \ell) \in I : \{k, \ell\} \cap \{i, j\} \neq \emptyset\}.
\]

Then, with the notation of Theorem 4.4,

\[
\sum_{\alpha \in I} p_\alpha \eta_\alpha \leq (m + n) \frac{2}{L} l_{su} \sum_{\alpha \in I} p_\alpha = \frac{2}{L} (m + n) l_{su} \lambda.
\]

Now a random intervals, of length \( b \) overlaps with a given interval of length \( a \) on the circle with probability \( \frac{2 \min(a,b)}{L} \), and so also

\[
\sum_{\alpha \in I} \mathbb{E}(X_\alpha (\eta_\alpha - X_\alpha)) \leq \frac{2}{L} (m + n) l_{su} \lambda.
\]

Dividing by \( \lambda \) gives the assertion.
Mixed Poisson approximation for $V_t$

Given $M(\tau_x) = m, N(\tau_x) = n$, and all the lengths $s_1, \ldots, s_m$ and $u_1, \ldots, u_n$ of the intervals,

$$V_{\tau_x} \approx \text{Poisson} \left( \frac{2}{L} \sum_{i=1}^{m} \sum_{j=1}^{n} \min(s_i, u_j) \right).$$

So

$$\mathbb{P}\{V_{\tau_x} = 0 | M(\tau_x) = m, N(\tau_x) = n, s_1, \ldots, s_m, u_1, \ldots, u_n\} \approx e^{-\frac{2}{L} \sum_{i=1}^{m} \sum_{j=1}^{n} \min(s_i, u_j)}.$$

Unconditioning

Unconditioning gives that

$$\mathbb{P}\{V_{\tau_x} = 0\} \approx \mathbb{E} e^{-\frac{4}{L} \int_0^{\tau_x} M(u)N(u)du}.$$  

Using Theorem 6.6, with $W', W$ being independent copies of the almost-sure limit of $e^{-2\rho t}M(t)$,

$$\int_0^{\tau_x} M(u)N(u)du = \int_0^{\tau_x} e^{4\rho u} (e^{-2\rho M(u)}e^{-2\rho N(u)}) du \approx \int_0^{\tau_x} e^{4\rho u} WW' du \approx WW' \frac{1}{4\rho} (e^{4\rho \tau_x} - 1) \approx \frac{1}{4\rho} WW' e^{2\rho} (\sqrt{L\rho})^2 = \frac{L}{4} WW' e^{2\rho},$$

and so

$$\mathbb{P}\{V_{\tau_x} = 0\} \approx \mathbb{E} e^{-e^{2\rho} WW'}.$$  

Branching process argument

We can determine the approximate distribution of the shortest path length more explicitly, see Barbour and Reinert (2001).

Theorem 7.2

$$\mathbb{E} \exp\{-e^{2\rho} WW'\} = \int_0^\infty \frac{e^{-y}}{1 + e^{2\rho y}} dy.$$
Proof (heuristical)

From Lemma 6.2 and Lemma 6.3,
\[ e^{-2\rho t} M(t) \to W \text{ a.s.; } \frac{s(t)}{M(t)} \to \frac{1}{\rho} \text{ a.s.} \]
and \( W \sim \text{Exp}(1) \), so with \( W, W' \) independent, \( \text{Exp}(1) \)-distributed
\[
\exp \left\{ -\frac{4}{L} \int_{0}^{\tau_{x}} M(v) N(v) dv \right\} \\
\sim \exp \left\{ -\frac{4}{L} WW' \int_{0}^{\tau_{x}} e^{4Lx} dv \right\} \\
\sim \exp \{-e^{2x} WW'\}.
\]
Hence
\[
\mathbb{E} \exp \{-e^{2x} WW'\} = \int_{0}^{\infty} \frac{e^{-y}}{1 + e^{2x} y} dy
\]
This finishes the proof.

7.14

Collecting the approximation:

Corollary 7.3
If \( T \) denotes a random variable with distribution given by
\[
\mathbb{P} [ T > x ] = \int_{0}^{\infty} \frac{e^{-y} dy}{1 + e^{2x} y}
\]
and \( D^* = \frac{1}{\rho} \left( \frac{1}{2} \log(\rho) + T \right) \), then
\[
\sup_{x} \left| \mathbb{P}[D \leq x] - \mathbb{P}[D^* \leq x] \right| = O \left( (\rho)^{-\frac{1}{2}} \log^2(\rho) \right).
\]
Note that \( \mathbb{E}(T) = \frac{1}{2} \gamma \approx 0.2886 \), where \( \gamma \) is Euler’s constant.

7.15

Remarks

The approach, with modifications, also applies to other networks, such as higher-dimensional small worlds and random intersection graphs. It is conjectured that the branching process approximation holds in a much wider setting, and it can serve as a heuristic when analytical results are not available.

When going from the discrete to the continuous model, the discretisation can make a lot of difference, in particular when the average shortest path length is small. Then, there may be a time at which there are hardly any shortcuts, but at the next discrete time step almost all vertices are connected. For more details see *A.D. Barbour and G. Reinert (2006)*.

8 Other models and summaries

Here we briefly review results for other summary statistics and for other models.

To define the occurrence of a subgraph \( m \) on \( k \) vertices, we follow the approach by Picard et al. (2007). Let \( I_{k} \) denote the set of all \( k \)-tuples of the vertex set \( V \). We consider \( \alpha = (i_1, \ldots, i_k) \in I_{k} \) a potential position of \( m \) in the graph \( G = (V, E) \). In order to match a position with an adjacency matrix, we order the elements of \( \alpha \) so that \( i_1 < i_2 < \ldots < i_k \).

Let \( Y_{\alpha}(m) \) be the indicator variable which equals 1 if \( m \) occurs at position \( \alpha \), and 0 otherwise.

In many models the distribution of \( Y_{\alpha}(m) \) does not depend on \( \alpha \) but is Bernoulli with probability of success \( \mu(m) \), the probability of occurrence of \( m \) at any position.
8.1 The distribution of vertex degrees in Bernoulli random graphs

In a Bernoulli $G(n, p)$ random graph, pick a vertex $V$ uniformly at random, and record its degree $d(V)$. We have seen that the distribution of $d(V)$ is Binomial$(n-1, p)$.

In the sparse regime, in distribution, $d(V) \approx \text{Po}((n-1)p)$.

Note that the vertex degrees in a graph are not independent. Indeed the dependence only vanishes asymptotically for very high degrees. So $d(V)$ does not stand for the average vertex degree $\overline{D}$.

8.2 The average degree

With $X = (X_{u,v})_{u,v=1,...,n}$ the adjacency matrix,

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

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

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

and so, in distribution,

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

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

In the dense regime, a normal approximation applies instead.

8.3 Expected subgraph counts

For a general subgraph $m$ on $k$ vertices, with edge indicators $m_{u,v}$, taking the value 1 if vertices $u$ and $v$ are connected in $m$, and 0 otherwise, then the occurrence of $m$ does not depend on the position $\alpha$. and $\mu(m)$ is a simple product,

$$\mu(m) = \prod_{u,v} P(X_{u,v} = 1)^{m_{u,v}} = \prod_{u,v} p^{m_{u,v}} = p^{m_{++}},$$

where $m_{++} = \sum_{u,v} m_{u,v}$ is twice the total number of edges in $m$, see Picard et al. (2007).

8.4 In the dense regime, the counts for small subgraphs are not asymptotically independent. Indeed in dense Bernoulli random graphs the number of edges already asymptotically determines the number of 2-stars and the number of triangles, see Reinert and Röllin (2010).
8.2 The distribution of summary statistics in Watts-Strogatz small worlds

Recall the model: $n$ vertices of $V$ are arranged on a lattice. Then hard-wire each vertex to its $k$ nearest neighbours on each side on the lattice. Introduce random shortcuts between vertices which are not hard-wired; the shortcuts are chosen independently, all with the same probability $\phi$.

Thus the shortcuts behave like a Bernoulli random graph, but the graph will necessarily be connected. The degree $d(V)$ of a randomly picked vertex $V$ is hence distributed as

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

taking the fixed lattice into account. Again a Poisson approximation holds when $p$ is small; see Lin (2007) for details.

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

**Figure:** An example for triangles in a small-world network: one shortcut here creates 2 triangles
8.3 The distribution of summary statistics in Barabasi-Albert models

The vertex 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:

8.10

The average pathlength

The average pathlength $\ell$ 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.

8.12

8.4 The distribution of summary statistics in the configuration model

Consider a configuration model graph on $n$ vertices, with given degrees $(d(i), i = 1, \ldots, n)$, and $\sum_{i=1}^n d(i) = 2m$. Recall that the expected number of edges between $i$ and $j \neq i$ is

$$p_{i,j} = \frac{d(i)d(j)}{2m - 1}.$$

8.13

The average clustering coefficient

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

$$E(C) \approx \frac{m - 1}{8} \frac{(\log n)^2}{n}.$$
Let $f_k$ denote the fraction of vertices in the network with degree $k$, so that the total number of vertices with degree $k$ is $nf_k$. Then the expected number of stubs of a randomly chosen vertex $V$ which attach to any vertex $\neq V$ which has degree $k$ is
\[
\frac{knf_k}{2m - 1} = \frac{f_k}{Ed(V) - \frac{1}{n}}.
\]
Here we used that $Ed(V) = \frac{2m}{n}$.
So the vertex we reach by following an edge is not a typical vertex in the network; it is more likely to have a high degree than a typical vertex.

### 8.5 The distribution of summary statistics in random geometric graphs

Recall that in geometric random graphs, $n$ points $X_1, \ldots, X_n$ are chosen independently at random according to a density $f(\cdot)$ on $\mathbb{R}^d$. A value $r = r_n > 0$ is chosen and we put an edge $(i, j)$ if $d(X_i, X_j) \leq r$. Here $d$ is a distance on $\mathbb{R}^d$; usually we choose the Euclidean distance.

Put $Y_i = \frac{1}{r}X_i$ so that we put an edge $(i, j)$ if $d(Y_i, Y_j) \leq 1$. If $Q = f(X_0)$ then $Y_i$ is approximately a Poisson process of rate $Qnr^d$. If $v_d$ denotes the volume of the unit ball in $d$ dimensions, then the degree of $X_0$, conditional on $Q$, is approximately Poisson distributed with parameter $Qnr^dv_d$.

For the number of induced subgraphs there is a formula for the mean. In the sparse regime when $nr^d = \Theta(1)$ we have that the mean scales as $n$.

If the mean converges to a finite value $\lambda$, then the number of induced subgraphs is approximately Poisson. This is called the super-sparse regime.

If the mean tends to infinity, then a normal approximation for the count holds, where the variance is given by a complicated formula.

### The size of the largest component

For asymptotics we need to think about an asymptotic regime. Assume that if there are $n$ vertices in the graph, then our degree distribution is such that there are $\lambda_i n + o(n)$ vertices of degree $i$. Here $\lambda_0, \lambda_1, \ldots$ are fixed. Put
\[
Q = \sum_i i(i - 2)\lambda_i.
\]
Then Molloy and Reed (1995) showed that
1. If $Q < 0$ then, almost surely, the size of the largest component is $O(\omega^2 \log n)$, where $\omega$ is the highest degree in the graph.
2. If $Q > 0$ then, almost surely, the size of the largest component is $\Theta(n)$, and the size of the second-largest component is $O(\log n)$. 
Connectivity

When working on a $d$-dimensional torus and $f(x)$ is the uniform density, the crucial quantity is

$$p_n = f_0 v d n^d,$$

then the degree of a randomly chosen vertex is approximately Poisson with parameter $np$. Then the probability that the graph is connected tends to 1 if $p_n = (1 + \varepsilon) \frac{\log n}{n}$, and it tends to 0 if $p_n = (1 - \varepsilon) \frac{\log n}{n}$. But when working on a square instead of a torus, the boundary effects cannot be ignored.

8.18

8.6 The expected subgraph counts in Erdős-Rényi mixture graphs

Assume that there are $Q$ different classes of vertices, with proportions $\alpha_1, \ldots, \alpha_Q$. Edges are independent conditionally on the classes of the vertices; with our usual edge indicator notation,

$$P(X_{i,j} = 1 | i \in q, j \in \ell) = p_{q\ell}.$$

The conditional occurrence probability of a subgraph $m$ given the class of each vertex is

$$P(Y_\alpha(m) = 1 | i_1 \in c_1, \ldots, i_k \in c_k) = \prod_{1 \leq u,v \leq k} p_{c_u,c_v}^{m_{uv}}.$$

Then

$$\mu(m) = \sum_{c_1=1}^Q \cdots \sum_{c_k=1}^Q \alpha_{c_1} \cdots \alpha_{c_k} \prod_{1 \leq u,v \leq k} p_{c_u,c_v}^{m_{uv}}.$$

8.19

8.7 The distribution of summary statistics in exponential random graph models

The distribution of the vertex 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 vertices, 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 $\theta$.

8.18

8.20

When only the average vertex 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. See Chatterjee and Diaconis (2011) for more details.

8.21
Weighted networks

So far we have only considered edges that form simple on/off connections between vertices. Sometimes it is useful to represent edges as having a strength, weight, or value to them, usually a real number. The weights may represent for example the capacity of an edge, or the volume of trade, or a reliability of an edge. If they represent the strength of a friendship then they could conceivably also take on negative values.

We can view weights as a special case of variables which are attached to edges. Indeed one could also attach variables to vertices, representing vertex properties.

Directed graphs

A special variable attached to an edge could be one which defines a direction on the edge; \( i \to j \) if there is a directed edge between \( i \) and \( j \). For example in food webs an edge could represent a predator-prey relation. Usually the adjacency matrix \( A \) of a directed networks has entries \( A_{ij} = 1 \) if there is an edge from \( j \) to \( i \) (note the direction of the edge, slightly counter-intuitive.).

A simple way of making a directed network undirected is to ignore the edge directions entirely. This procedure ignores a lot of information. Two alternatives are the co-citation matrix and the bibliographic coupling matrix.

9 Sampling from networks

Often the whole network is too large and we only observe a portion of the network. If we want to draw inference from the whole network based on the sample then we need to take the specifics of our sample into account.

- the co-citation matrix \( C = AA^T \); \( C_{i,j} \) gives the number of vertices that have outgoing edges pointing to both \( i \) and \( j \). Thinking of citation networks, the cocitation of two papers is the number of other papers that cite both. If we want to convert \( C \) into the adjacency matrix of an undirected network, just record whether or not \( C_{i,j} = 0 \).
- the bibliographic coupling matrix \( B = A^T A \); here, if \( i \) and \( j \) are papers, then \( B_{i,j} \) is the number of other papers that are cited by both \( i \) and \( j \).

Both of these matrices can be used as measures of vertex similarity.
Sampling schemes

The main sampling schemes for networks which we shall consider are

1. induced subgraph sampling - we take a random sample of vertices and observe all edges between the vertices in the sample.
2. snowball sampling - in a 1-hop snowball sample we start with one vertex, sample all of its edges and neighbours and, depending on the scheme, the edges between its neighbours too. In a 2-hop snowball sample we start with one vertex and sample all vertices within distance 2 of the vertex (i.e. at most two edges away), and depending on the scheme, also sample the edges between those neighbours.

9.1 Sampling designs

Here we assume that observations are made without measurement error, and that the only source of randomness is the sampling design. This will help to understand the issues which arise from sampling alone, not taking any probabilistic model for the network into account.

9.2

Suppose that we have a population \( U = \{1, \ldots, N_u\} \) of units and with each unit \( i \in U \) is associated a value \( y_i \) of interest. Let

\[
\tau = \sum_{i \in U} y_i
\]

be the population total, and let

\[
\mu = \frac{\tau}{N_u}
\]

be the population average. We also need the population variance

\[
\sigma^2 = \frac{1}{N_u} \sum_{i \in U} (y_i - \mu)^2.
\]

9.3

Let \( S = \{i_1, \ldots, i_n\} \) be a sample of \( n \) units from \( U \) and observe that we observe \( y_i \) for each element in the sample. The goal is to estimate \( \mu \) and \( \tau \).

If the sample \( S \) is chosen by drawing \( n \) units uniformly from \( U \), with replacement, then natural estimates of \( \mu \) and \( \tau \) are

\[
\hat{\mu} = \overline{y} = \frac{1}{n} \sum_{i \in S} y_i; \quad \hat{\tau} = N_u \overline{y}.
\]
These estimates are unbiased, and
\[ \text{Var}(\bar{y}) = \frac{1}{n} \sigma^2; \quad \text{Var}(\hat{\tau}) = \frac{1}{n} N_u \sigma^2. \]
These variances can be estimated in an unbiased fashion using
\[ \hat{\text{Var}}(\bar{y}) = \frac{1}{n} s^2; \quad \hat{\text{Var}}(\hat{\tau}) = \frac{1}{n N_u} s^2, \]
with
\[ s^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y})^2 \]
being the sample variance.

\section*{Mean}

Assume that \( \pi_i > 0 \) for all \( i \). Let \( Z_i = 1(i \in S) \). Then
\[ \mathbb{P}(Z_i = 1) = \pi_i \]
and
\[ \mathbb{E} \hat{\tau}_\pi = \mathbb{E} \sum_{i \in S} \frac{y_i}{\pi_i} = \mathbb{E} \sum_{i \in U} \frac{y_i}{\pi_i} Z_i = \sum_{i \in U} \frac{y_i}{\pi_i} \mathbb{E} Z_i = \sum_{i \in U} y_i = \tau, \]
so \( \hat{\mu}_\pi \) is an unbiased estimator for \( \tau \).

\section*{Variance}

Let
\[ \pi_{i,j} = \mathbb{P}(i \in S, j \in S) \]
with \( \pi_{i,i} = \pi_i \). Then
\[ \text{Var} \left( \hat{\mu}_\pi \right) = \sum_{i \in U} \frac{y_i^2}{\pi_i^2} \text{Var} Z_i + \sum_{i \in U} \sum_{j \in U, j \neq i} \frac{y_i y_j}{\pi_i \pi_j} \text{Cov}(Z_i, Z_j) \]
\[ = \sum_{i \in U} \frac{y_i^2}{\pi_i} (1 - \pi_i) + \sum_{i \in U} \sum_{j \in U, j \neq i} \frac{y_i y_j}{\pi_i \pi_j} (\pi_{i,j} - \pi_i \pi_j). \]
Hence
\[
\text{Var} \left( \hat{\mu}_\pi \right) = \sum_{i \in U} \sum_{j \in U} y_i y_j \left( \frac{\pi_{i,j}}{\pi_i \pi_j} - 1 \right).
\]

Assuming that \( \pi_{i,j} > 0 \) for all pairs, an unbiased estimator for \( \text{Var} \left( \hat{\mu}_\pi \right) \) is given by
\[
\hat{\text{Var}} \left( \hat{\mu}_\pi \right) = \sum_{i \in S} \sum_{j \in S} y_i y_j \left( \frac{1}{\pi_i \pi_j} - 1 \right).
\]

### 9.10 Example: Simple random sampling

Simple random sampling is random sampling without replacement. There are \( \binom{N_u}{n} \) possible simple random samples of size \( n \) from \( U \), and \( \binom{N_u-1}{n-1} \) contain unit \( i \). Hence
\[
\pi_i = \frac{\binom{N_u-1}{n-1}}{\binom{N_u}{n}} = \frac{n}{N_u}.
\]

Similarly, for \( i \neq j \),
\[
\pi_{i,j} = \frac{n(n-1)}{N_u(N_u-1)}.
\]

### 9.11 Induced subgraph sampling

In this design, a simple random sample of \( n \) vertices is selected from the vertex set \( \mathcal{V} \), without replacement, and edges are observed for all vertex pairs in the sample. Then, for any \( i \in \mathcal{V} \),
\[
\pi_i = \frac{n}{|\mathcal{V}|},
\]
and, for \( i \neq j \),
\[
\pi_{i,j} = \frac{n(n-1)}{|\mathcal{V}|(|\mathcal{V}| - 1)}.
\]

Note that these probabilities assume that we know \( |\mathcal{V}| \).
Star sampling

Star sampling is a special case of snowball sampling. A simple random sample $S$ of $n$ vertices is selected from the vertex set $\mathcal{V}$, without replacement. For each vertex in the sample all edges that are incident to it are observed. In this case, for any $i \in \mathcal{V}$,

$$\pi_i = \frac{n}{|\mathcal{V}|}$$

and, for $i \neq j$,

$$\pi_{i,j} = \mathbb{P}(i \neq S, j \neq S) = 1 - \frac{(|\mathcal{V}| - 2)n}{(\binom{|\mathcal{V}|}{2})}.$$  

If we include all adjacent vertices of all vertices in $S$ then $\pi_{i,j}$ stays the same, but the vertex inclusion probabilities change. For $i$,

$$\pi_{i} = \sum_{L \subseteq N(i)} (-1)^{|L| + 1} \mathbb{P}(L),$$

where $N(i)$ is the set of all vertices which are adjacent to $i$, and $i$ itself, and $\mathbb{P}(L)$ is the probability of selecting the set $L$ when obtaining the sample $S$.

Inclusion probabilities for snowball sampling become increasing intractable to calculate beyond star sampling.

Variants

Instead of attaching a variable $y_i$ to a vertex we could attach a variable $y_{i,j}$ to an edge. The Horvitz-Thompson estimator for the total is then

$$\hat{\tau}_{\pi} = \sum_{i,j \in \mathcal{V}} \frac{y_{i,j}}{\pi_{i,j}}.$$ 

For the variance we need the probability $\pi_{i,j,k,l}$ that vertices $i,j,k$ and $l$ are included in the sample. Using $y_{i,j}$ as the indicator that an edge is present, this approach can be used to estimate the total number of edges.

Similarly we could attach a variable $y_{i,j,k}$ to triples of vertices. Using $y_{i,j,k}$ as the indicator that a triangle is present, this approach can be used to estimate the total number of triangles in the graph.

9.2 Within-graph versus between graph sampling

Usually we assume that we have i.i.d. observations and can estimate the parameters from these observations. When analysing networks we often observe only 1 network. If the network has $n$ vertices then we observe $\binom{n}{2}$ edge indicator variables, but depending on the network model these indicator variables may not be independent.

Also sometimes our observations relate directly to summary statistics, and these may introduce dependence even when the edges in the model are independent. Here is an example: we calculate all shortest paths between vertices in a Watts-Strogatz small world.
The shortest path length in Watts-Strogatz small world: dependent sampling

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:

<table>
<thead>
<tr>
<th>$n$</th>
<th>$k$</th>
<th>$\phi$</th>
<th>$E.no$</th>
<th>mean p-value</th>
<th>max p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>300</td>
<td>1</td>
<td>0.01</td>
<td>3</td>
<td>1.74 E-09</td>
<td>8.97 E-08</td>
</tr>
<tr>
<td></td>
<td>0.167</td>
<td>50</td>
<td>0.1978</td>
<td></td>
<td>0.8913</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.01</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>1</td>
<td>0.003</td>
<td>3</td>
<td>1.65E-13</td>
<td>3.30 E-12</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>50</td>
<td>0.0101</td>
<td></td>
<td>0.1124</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.03</td>
<td>60</td>
<td>0.0146</td>
<td>0.2840</td>
</tr>
</tbody>
</table>

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

For statistical testing we would need to know the distribution of the average shortest path length within a network, rather than the distribution of the distance between two randomly chosen vertices. This is an open problem.

10 Fitting a model: parametric approaches

**Parametric** just means that we have a finite set of parameters which fully specify the model. Otherwise we call the setting non-parametric. For example:

**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.
10.1 Maximum likelihood estimation

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

The likelihood of the parameter given the data is just the probability of seeing the data we see, given the parameter. Maximum-likelihood estimators have attractive properties; under some regularity conditions they would not only converge to the true parameter as the sample size tends to infinity, but it would also be approximately normally distributed if suitably standardized, and we can approximate the asymptotic variance.

Example: Bernoulli random graphs

Our data is the network we see. We describe the data using the adjacency matrix, denote it by \( x \) here because it is the realisation of a random adjacency matrix \( X \). Recall that the adjacency matrix is the square \(|V| \times |V|\) matrix where each entry is either 0 or 1:

\[
x_{u,v} = 1 \text{ if and only if there is an edge between } u \text{ and } v.
\]

We denote by

\[
E = \{(u, v) \in V \times V : u \neq v\}
\]

the set of potential edges.

To maximise the likelihood, we often take logs, and then differentiate. Here this would give

\[
\ell(p; x) = \log \mathcal{L}(p; x) = |E| \log(1 - p) + t \log p - t \log(1 - p);
\]

and

\[
\frac{\partial \ell(p; x)}{\partial p} = -\frac{1}{1 - p} (|E| - t) + \frac{t}{p}.
\]

10.2

The likelihood of \( p \) being the true value of the edge probability if we see \( x \) is

\[
\mathcal{L}(p; x) = \prod_{(i,j) \in E} \{p^{x_{i,j}}(1 - p)^{1-x_{i,j}}\}
\]

\[
= (1 - p)^{|E|} \prod_{(i,j) \in E} \left(\frac{p}{1 - p}\right)^{x_{i,j}}
\]

\[
= (1 - p)^{|E|} \left(\frac{p}{1 - p}\right)^{\sum_{(i,j) \in E} x_{i,j}}.
\]

Note that \( t = \sum_{(i,j) \in E} x_{i,j} \) is the total number of edges in the random graph.

10.3
To find a maximum we equate this to zero and solve for $p$,

$$\frac{t}{p} = \frac{1}{1 - p}(|E| - t) \iff t(1 - p) = p(|E| - t)$$

$$\iff t = \frac{p|E|}{1 - p} \iff p = \frac{t}{|E|}.$$

We can check that the second derivative of $\ell$ is less than zero when $t \neq 0,1$, and a careful consideration of the cases $t = 0$ and $t = 1$ shows that the fraction of edges that are present in the network,

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

is our maximum-likelihood estimator.

Maximum-likelihood estimation also works well in Erdős-Rényi 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

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

the calculation of the normalising constant $\kappa$ becomes numerically impossible very quickly.

10.2 Markov Chain Monte Carlo estimation

When the likelihood is not available, Markov chain Monte Carlo methods can often be set up.

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; a random walk is an example.

A Markov chain is irreducible if any set of states can be reached from any other state in a finite number of moves.

The Markov chain 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.
There are mathematical definitions for these concepts, but we only need the main result here:

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 $X$ 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.

### 10.3 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 vertex degree, clustering coefficient, and shortest path length. Then we see whether our observed values are plausible under the estimated model. Note that the model should reflect the research question; typically no model will capture all aspects of the observed network.

Often, secretly we would like to find that the observed value of the statistics are not plausible under the null model! Because then we can reject, say, the simple random graph model, and conclude that something more complicated is going on.

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 vertex degree, clustering coefficient, and shortest path length. Then we see whether our observed values are plausible under the estimated model. Note that the model should reflect the research question; typically no model will capture all aspects of the observed network.

Often, secretly we would like to find that the observed value of the statistics are not plausible under the null model! Because then we can reject, say, the simple random graph model, and conclude that something more complicated is going on.

### Example: Florentine family data

In the Florentine family data, we observe 16 vertices, 20 edges, an average vertex degree of 2.5, and a clustering coefficient of 0.1914894. We want to assess the null hypothesis that the data come from a Bernoulli random graph.

Let us assume that the null hypothesis is true. Then we estimate

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

As in a Bernoulli random graph $\overline{D} \approx \frac{2}{5}Z$, where $Z \sim Po\left(\frac{n(n-1)p}{2}\right)$, under the null hypothesis the average node degree would be $\overline{D} \approx \frac{1}{8}Z$, where $Z \sim Po(20)$. The probability under the null hypothesis that $\overline{D} \geq 2.5$ would then be

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

so there is no reason to reject the null hypothesis.
Quantile-quantile plots

We could also assess the fit visually. A much used plot in Statistics is the *quantile-quantile plot*. The *quantiles* of a distribution are its "percent points"; for example the 0.5 quantile is the 50% point, i.e. the median. Mathematically, the *(sample) quantiles* $q_\alpha$, are defined for $0 \leq \alpha \leq 1$ so that a proportion of at least $\alpha$ of the data are less or equal to $q_\alpha$ and a proportion of at least $1 - \alpha$ is greater or equal to $q_\alpha$. There are many (at least 8) definitions of $q_\alpha$ if $\alpha n$ is not an integer.

We plot the quantiles of our observed (empirical) distribution against the quantiles of our hypothesised (null) distribution; if the two distributions agree, then the plot should result in a roughly diagonal line.

Example: normal distribution

Simulate 1,000 random variables from a normal distribution. Firstly: mean zero, variance 1; secondly: mean 1, variance 3. Both QQ-plots are satisfactory.

We can also use a quantile-quantile plot for two sets of simulated data, or for one set of simulated data and one set of observed data. The interpretation is always the same: if the data come from the same distribution, then we should see a diagonal line; otherwise not. Here we compare 1000 Normal (0,1) variables with 1000 Poisson (1) variables - clearly not a good fit.

11 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.
11.1 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 $T_{01}, T_{02}, \ldots$ 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 $\{t^*, t_{01}, \ldots, t_{0n}\}$
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,

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

The procedure is exact however small $n$ might be. However, 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). A reasonable rule is to choose $n$ such that $m \geq 5$. Note that we will need more simulations to test at smaller values of $\alpha$.

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.

$$P = P(T_0 > t^* | H_0).$$

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.

11.2

For random graphs, Monte Carlo tests often use shuffling edges with the number of edges fixed, or fixing the vertex 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 vertex 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 vertex degrees, which determine the number of edges) gives degenerate results. More generally, we have seen that vertex 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. Even in Bernoulli random graphs, when the expected number of edges is moderate, so that a normal approximation would hold for the number of edges, then, asymptotically, the number of 2-stars and the number of triangles is already completely determined by the number of edges!

11.2 Fitting scale-free networks

Barabasi and Albert introduced networks such that the distribution of the degree of a randomly picked vertex \( V \) is of the type

\[
P(d(V) = k) \sim Ck^{-\gamma}
\]

for \( k \to \infty \). Such behaviour is called power-law behaviour; the constant \( \gamma \) is called the power-law exponent. The networks are also called scale-free, because of the following property.
If $\alpha > 0$ is a constant, then

$$P(d(V) = \alpha k) \sim C(\alpha k)^{-\gamma} \sim C' k^{-\gamma},$$

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:

$$\log P(d(V) = k) \sim \log C - \gamma \log k$$

$$\log P(d(V) = \alpha k) \sim \log C - \gamma \log \alpha - \gamma \log k;$$

scaling the argument results in a linear shift of the log probabilities only. This equation also leads to the suggestion to plot the log 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 $\gamma$.

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

$$\log P(d(V) \geq k).$$

If the model is correct, then one can calculate that

$$\log P(d(V) \geq k) \sim C'' - (\gamma - 1) \log k.$$ 

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.

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

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

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 $10^9$ vertices. The criteria are not often matched.

11.14

11.15

12 Statistical inference for vertex characteristics and edges

In the statistical analysis of networks we are often interested in inferring "local" properties, such as the existence of an edge or the characteristics of a vertex, from the position of the vertex, or the potential edge, in the network. Such inference has a long tradition in social network analysis; there logistic regression and loglinear regression models are used. Here we look at two examples which are inspired by social network analysis.

12.1 Log-linear models

Suppose that we observe two variables $A$ and $B$, with possible outcomes $A_1, A_2$ and $B_1, B_2$. We summarise the data by a $2 \times 2$ table.

<table>
<thead>
<tr>
<th></th>
<th>$B_1$</th>
<th>$B_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_1$</td>
<td>$n_{11}$</td>
<td>$n_{12}$</td>
</tr>
<tr>
<td>$A_2$</td>
<td>$n_{21}$</td>
<td>$n_{22}$</td>
</tr>
<tr>
<td>$n_{+1}$</td>
<td>$n_{+2}$</td>
<td>$n$</td>
</tr>
</tbody>
</table>

Here the + sign in a subscript indicates a subscript that has been “summed out”. Let $p_{ij}$ denote the probability that the observation $(A, B)$ falls into the cell $(A_i, B_j)$. Assume that we have $n$ independent observations of the pair $(A, B)$. 

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 vertices 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, is is debated how useful the concept of “scale-free” behaviour is, as many biological networks contain relatively few vertices.
The likelihood

In this case we use a multinomial likelihood,

\[
\log L = \sum_i \sum_j n_{ij} \log p_{ij} = \sum_i \sum_j n_{ij} \log e_{ij} - n \log n,
\]

where \( e_{ij} = np_{ij} \) is the expected cell frequency of cell \((A_i, B_j)\).

The null hypothesis

Under the null hypothesis that \(A\) and \(B\) are independent we find that \(p_{ij} = p_{i+}p_{+j}\); i.e.

\[
H_0 : \log e_{ij} = \log e_{i+} + \log e_{+j} - \log n.
\]

Thus, under \(H_0\),

\[
\log e_{ij} = \alpha_i + \beta_j + \lambda,
\]

an additive combination of a term depending only on \(i\), and a term depending only on \(j\), and a constant. This is a linear model! In such sums we can define \(\lambda\) such that

\[
\sum \alpha_i = \sum \beta_j = 0.
\]

Then

\[
np_{ij} = e_{ij} = e^{\alpha_i + \beta_j + \lambda}.
\]

The simple \(2 \times 2\) table approach motivates the introduction of the log-linear model. Suppose that \(x \in \mathbb{R}^n\) are explanatory variables and \(Y \in \{0, 1, \ldots\}\) is the response variable. Then we may model

\[
\log(\mathbb{E}(Y|x)) = \theta^T x
\]

for some unknown parameter \(\theta \in \mathbb{R}^n\). Often we would assume that \(Y\) is Poisson distributed; then

\[
Y \sim P_0 \left( e^{\theta^T x} \right).
\]

The parameter \(\theta\) is usually estimated by maximum likelihood, and found numerically.

If \(Y\) is the unknown characteristic of a vertex and if \(x\) are the known characteristics of its neighbours, then a log-linear regression can be used to predict \(Y\) if we assume that edges and characteristics are independent.

In real networks this assumption may often be violated.

If we use network statistics as explanatory variables then the log-linear model can be viewed as an exponential random graph model. Thus maximum-likelihood estimation may not be numerically stable.
12.2 Inferring characteristics of proteins from their position in a protein interaction network

The most often used method is Majority Vote (M.V.): observe the functional characteristics which the nearest neighbours of the target protein possess, and to select the function which occurs most frequently.

We can improve on this method with ideas from loglinear models to develop a score which we can use to classify proteins, say. Here is an elaborate example from Chen et al. (2007).

12.7 As characteristics we consider structure (7 categories) and function (24 categories). From the protein-protein interaction network, we build an upcast set of category-category interactions. A category-category interaction is constructed by two characteristic categories from two interacting proteins.

Consider a protein $x$, within the set of all characteristic categories $S$, $S(x)$ includes the categories that protein $x$ is classified into. If two proteins $x$ and $y$ interact, the category-category interaction is the edge between two characteristic categories, $a$ and $b$ ($a \in S(x)$, $b \in S(y)$), from each of two proteins (denoted by $a \sim b$).

Our scoring method is based on the heuristic assumption that the likelihood for a specific category to be observed in the query protein is roughly proportional to the product of the relative frequencies of observing this category in all pairs around the neighbours of a query protein.

Let

$$C(a, x) = \prod_{\begin{subarray}{c} b \in S(n) \\ n \in B(x) \end{subarray}} f(a \sim b),$$

where $f(a \sim b)$ is the relative frequency of category-category interaction $\{a \sim b\}$ among all category-category interactions.

The score for the query protein $x$ with annotated protein-interaction network neighbours $B(x)$ to be in a specific category $a$ is proportional to the product $C(a, x)$.

We define our score $F(a, S(x))$ by

$$F(a, S(x)) := \frac{C(a, x)}{\sum_{k \in S} C(k, x)}.$$

The protein is then predicted to possess the characteristic category, or categories, with the highest score.

This score is derived as an analogy of the likelihood of observing category $a$ in $S(x)$ if all edges in the category interaction network occurred independently. Heuristically this score serves as a measure for the chance of protein $x$ having characteristic $a$. 

12.8 As characteristics we consider structure (7 categories) and function (24 categories). From the protein-protein interaction network, we build an upcast set of category-category interactions. A category-category interaction is constructed by two characteristic categories from two interacting proteins.

Consider a protein $x$, within the set of all characteristic categories $S$, $S(x)$ includes the categories that protein $x$ is classified into. If two proteins $x$ and $y$ interact, the category-category interaction is the edge between two characteristic categories, $a$ and $b$ ($a \in S(x)$, $b \in S(y)$), from each of two proteins (denoted by $a \sim b$).

Our scoring method is based on the heuristic assumption that the likelihood for a specific category to be observed in the query protein is roughly proportional to the product of the relative frequencies of observing this category in all pairs around the neighbours of a query protein.

Let

$$C(a, x) = \prod_{\begin{subarray}{c} b \in S(n) \\ n \in B(x) \end{subarray}} f(a \sim b),$$

where $f(a \sim b)$ is the relative frequency of category-category interaction $\{a \sim b\}$ among all category-category interactions.

The score for the query protein $x$ with annotated protein-interaction network neighbours $B(x)$ to be in a specific category $a$ is proportional to the product $C(a, x)$.

We define our score $F(a, S(x))$ by

$$F(a, S(x)) := \frac{C(a, x)}{\sum_{k \in S} C(k, x)}.$$

The protein is then predicted to possess the characteristic category, or categories, with the highest score.

This score is derived as an analogy of the likelihood of observing category $a$ in $S(x)$ if all edges in the category interaction network occurred independently. Heuristically this score serves as a measure for the chance of protein $x$ having characteristic $a$. 

12.9
The method, which we call the Frequency (F.) method, can be extended to include two or more protein characteristics in the prediction of a specific protein characteristic; we call this the E.F. method. Then the category in a category-category interaction is a vector containing all characteristics of the protein. Here we use function information as well to predict structure, and also includes structure information when predicting function.

The protein structure is predicted the class with the highest probability.

A function prediction is counted as correct if one of the best three predicted categories is correct.

The accuracy of a method is the ratio of the number of correctly predicted proteins, divided by the total number of predicted proteins.

We compare two prediction methods based on a paired $z$-test, using a normal approximation. While predictions for different proteins may not be independent, we claim that the dependence is weak enough to justify a normal approximation.

### The accuracies of structure prediction using different methods:

<table>
<thead>
<tr>
<th>Organism (DIP)</th>
<th>Predicted proteins</th>
<th>M.V.</th>
<th>F.</th>
<th>E. F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.Melanogaster</td>
<td>1262</td>
<td>0.35</td>
<td>0.17</td>
<td><strong>0.44</strong></td>
</tr>
<tr>
<td>C.Elegans</td>
<td>78</td>
<td>0.36</td>
<td>0.37</td>
<td>0.49</td>
</tr>
<tr>
<td>S.Cerevisiae</td>
<td>1608</td>
<td>0.39</td>
<td>0.31</td>
<td><strong>0.54</strong></td>
</tr>
<tr>
<td>E.Coli</td>
<td>150</td>
<td>0.57</td>
<td>0.70</td>
<td><strong>0.71</strong></td>
</tr>
<tr>
<td>M.Musculus</td>
<td>32</td>
<td>0.72</td>
<td>0.50</td>
<td>0.69</td>
</tr>
<tr>
<td>H.Sapiens</td>
<td>273</td>
<td>0.44</td>
<td>0.47</td>
<td>0.71</td>
</tr>
</tbody>
</table>

*Underline*: where the result outperforms M.V. with statistical significance

### The accuracies of function prediction using different methods:

<table>
<thead>
<tr>
<th>Organism (DIP)</th>
<th>Predicted proteins</th>
<th>M.V.</th>
<th>F.</th>
<th>E. F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>D.M</td>
<td>1275</td>
<td>0.53</td>
<td>0.67</td>
<td><strong>0.69</strong></td>
</tr>
<tr>
<td>C.E</td>
<td>85</td>
<td>0.38</td>
<td>0.55</td>
<td>0.71</td>
</tr>
<tr>
<td>S.C</td>
<td>1618</td>
<td>0.67</td>
<td>0.61</td>
<td>0.67</td>
</tr>
<tr>
<td>E.C</td>
<td>154</td>
<td>0.69</td>
<td>0.69</td>
<td>0.70</td>
</tr>
<tr>
<td>M.M</td>
<td>32</td>
<td>0.59</td>
<td>0.88</td>
<td>0.81</td>
</tr>
<tr>
<td>H.S</td>
<td>274</td>
<td>0.79</td>
<td>0.90</td>
<td>0.89</td>
</tr>
</tbody>
</table>

*Underline*: where the result outperforms M.V. with statistical significance
12.3 Logistic regression

In logistic regression the response variable is binary, \( Y \in \{0, 1\} \), and we are interested in \( p(x) = P(Y = 1|x) \). We model the odds ratio \( \frac{p}{1-p} \) using the logit transformation,

\[
\text{logit}(P(Y = 1|x)) = \log \left( \frac{p(x)}{1 - p(x)} \right) = \theta^T x,
\]

or, equivalently,

\[
p(x) = \frac{e^{\theta^T x}}{1 + e^{\theta^T x}}.
\]

Again, numerical maximum likelihood is used to estimate \( \theta \).

12.15

If the edges in the network are independent but may depend on some vertex characteristics \( x \), then the logistic regression model can be used to predict presence or absence of an edge. In real networks the independence assumption may not be justified, but we can still use ideas from logistic regression to derive a score which will be used for classification.

12.16

12.4 The ROC curve

In order to put our scores to work we choose a threshold; all pairs with scores above that threshold would be classified as interacting, while all pairs below that threshold would be classified as non-interacting.

The choice of threshold depends on the desired sensitivity and specificity. The sensitivity is the fraction of correct predictions among all predicted positive pairs and the specificity is the fraction of correct predictions among all predicted negative pairs.

To assess our scores we use a Receiver Operating Characteristic (ROC) curve, which is a useful technique for examining the performance of a classifier; in our case the score “interacting” or “non-interacting” for a pair of proteins. The curve plots sensitivity against (1 minus specificity). Each point on a ROC curve is generated by selecting a score threshold for a method. We move the cutoff along the range of the score and record different sensitivities and specificities of a method. The closer the curve is to the upper left hand corner (i.e., the larger the area under curve), indicating that sensitivity and specificity are both high, the better the predictive score.
12.5 Inferring protein interactions from protein characteristics using the protein interaction network

This time we use ideas from logistic regression to develop a score which we can use to predict and to validate protein interactions, based on the protein characteristics and the protein interaction network; see Chen et al. (2008) for more details. Here, network structure comes into play, based on the observation that the protein interaction network has a tendency to complete triangles as opposed to two-stars.

The protein interaction network is used to build an upcast set of triplets of characteristic vectors. Here, $A, B, C$ and $D$ denote protein characteristics, whereas different shapes indicate different proteins. A protein may possess more than one characteristic. Our triplets are triangles and lines of three characteristic vectors according to their interacting patterns. A characteristic line is a specific pattern constructed by three vectors with two vector interactions among them. A characteristic triangle is formed by three vectors interacting with each other.

Within the triplet interactions, we assess the odds to observe triangles versus lines around the query protein pair. More formally, let $t_{xy}$ be the total frequency of all characteristic triangles around the query protein pair $\{x, y\}$; denoting by $z \in B(x, y)$ the set of all common neighbours of $x$ and $y$ in the protein interaction network,

$$t_{xy} = \sum_{z \in B(x, y)} \left[ \sum_{v_a \in S(x), v_b \in S(y), v_c \in S(z)} f(v_a \sim v_c \sim v_b \sim v_a) \right],$$

where $f(v_a \sim v_c \sim v_b \sim v_a)$ is the frequency of triangle $\{v_a \sim v_c \sim v_b \sim v_a\}$ among all characteristic triangles in the prior data base.

Figure: Example for the upcast set. Here we consider a single characteristic, the characteristic vector for a protein is a 1-vector. There are 3 single-category proteins and 1 two-category protein in the protein interaction network, which result in an upcast set of six characteristic pairs $\{A - B, A - D, B - D, B - C, C - D\}$. 
Similarly, $t_{xy}$ is the total frequency of all characteristic lines around the query protein pair \( \{x, y\} \). We define the \textit{triangle rate score}, \( tri(x, y) \) for the protein pair \( \{x, y\} \) as the odds of observing triangles versus lines among triangles and lines in its neighbourhood,

\[
tri(x, y) = \frac{t_{xy}}{t_{xy} + l_{xy}}.
\]

(12.1)

Heuristically, the higher the triangle rate score is, the higher the chance one would observe an interaction between the query protein pair.

We have many more gold-standard negatives than positives. The unequal sizes of gold-standard sets may affect the ROC curve; when the cutoff is high, too many gold-standard negatives would cause a rapid increase in true negatives, which would result in artificially high specificity. To avoid this bias, we collect 300 samples of randomly selected pairs from the extensive GSN. Each sample is the same size as our GSP set. Predictions are verified against these 300 reference sets obtained by combining the GSP set and the sample from the GSN set. We test the difference between two ROC curves through a \( z \)-test for differences, at 5% significance level.

While we are never completely certain that a prediction is correct, we assume that a positive prediction is correct if it is contained in our gold-standard positive (GSP) set, and that a negative prediction is correct if it is contained in our gold-standard negative (GSN) set. The GSP set is based on 8,250 hand-curated interactions in MIPS complexes catalog (MIPS-GSP). These positive interactions are identified if two proteins are within the same complex and if the interactions are confirmed by various experimental techniques. The set of gold-standard negatives (GSN) are random protein pairs which neither share protein localisation, nor expression nor homologous interaction data.
Network motifs are small subgraphs with a fixed number of nodes and with a given topology. We are particularly interested in over-represented subgraphs. Such motifs seem to be conserved across species, suggesting a link between protein evolution and topological features of the protein interaction network.

Over-representation is judged in comparison to a statistical null model. Most commonly, significantly over-represented motifs are detected based on a conditional uniform graph test which preserve some characteristics of the network: for example keeping the degree distribution fixed, which results in a configuration model.

13.1 Asymptotics for motif counts

Picard et al. (2008) calculate not only the mean but also the variance for motifs on 3 and 4 nodes in undirected graphs under the models:
1. Bernoulli random graph (ER)
2. Random graphs with fixed degree sequence (FDD) (here they estimate the mean and standard deviation from simulations; and they also consider a version with fixed expected degrees)
3. Erdös-Renyi mixture models (ERMG).

Here are some examples of their results for the protein interaction networks of H. pylori, E. coli. For the mean:

<table>
<thead>
<tr>
<th>motif</th>
<th>obs</th>
<th>ER</th>
<th>FDD</th>
<th>ERMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. pylori</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-stars</td>
<td>14,113</td>
<td>5704.08</td>
<td>14,113</td>
<td>13,602.97</td>
</tr>
<tr>
<td>triangles</td>
<td>75</td>
<td>10.85</td>
<td>66.91</td>
<td>52.82</td>
</tr>
<tr>
<td>3-stars</td>
<td>112,490</td>
<td>7676.83</td>
<td>112,490</td>
<td>93,741.08</td>
</tr>
<tr>
<td>E. coli</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-stars</td>
<td>248,093</td>
<td>52,774.79</td>
<td>248,093</td>
<td>243,846.93</td>
</tr>
<tr>
<td>triangles</td>
<td>11,368</td>
<td>72.47</td>
<td>3579.49</td>
<td>10,221.17</td>
</tr>
<tr>
<td>3-stars</td>
<td>6,425.495</td>
<td>133,050.00</td>
<td>5,772,005.15</td>
<td>1,537,740.00</td>
</tr>
</tbody>
</table>

Note that in FDD the degree distribution is fixed to equal the degree sequence in the network.

For the standard deviation:

<table>
<thead>
<tr>
<th>motif</th>
<th>ER</th>
<th>FDD</th>
<th>ERMG</th>
</tr>
</thead>
<tbody>
<tr>
<td>H. pylori</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-stars</td>
<td>311.08</td>
<td>0</td>
<td>2659.18</td>
</tr>
<tr>
<td>triangles</td>
<td>3.40</td>
<td>7.80</td>
<td>20.41</td>
</tr>
<tr>
<td>3-stars</td>
<td>681.76</td>
<td>0</td>
<td>27,039.88</td>
</tr>
<tr>
<td>E. coli</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-stars</td>
<td>1281.87</td>
<td>0</td>
<td>51,676.68</td>
</tr>
<tr>
<td>triangles</td>
<td>8.90</td>
<td>68.58</td>
<td>3041.98</td>
</tr>
<tr>
<td>3-stars</td>
<td>5089.62</td>
<td>0</td>
<td>1,672,086.51</td>
</tr>
</tbody>
</table>

The expectation and variance strongly depend on the model we choose for comparison.
We see that the variance is very different from the mean, so a Poisson approximation is not adequate. Also the counts are relatively low compared to the potential number of structures on the networks, thus a normal approximation (z-score) is not appropriate. Instead we use a Polya-Aeppli distribution, which is a special case of a compound Poisson distribution. It is obtained when the clump size has a geometric distribution.

\[ P(k) = a^{k-1} (1 - a), \]

and if the number of clumps is Poisson distributed with parameter \( \lambda \), then we can write down the Polya-Aeppli distribution for the count \( W \),

\[ P(W = w) = e^{-\lambda} a^w \sum_{c=1}^{w} \frac{1}{c!} \left( \frac{\lambda(1-a)}{a} \right)^c \quad \text{if} \ w = 1, 2, \ldots, \]

and

\[ P(W = 0) = e^{-\lambda}. \]

Parameter estimation

The parameters of the Polya-Aeppli distribution can be calculated from the first two moments:

\[ a = \frac{\text{var} - \text{mean}}{\text{var} + \text{mean}} \]

and

\[ \lambda = (1 - a) \text{mean}. \]

These parameters can be estimated using the sample mean and the sample variance.

The Polya-Aeppli distribution is a better fit to the count distribution, compared to the normal distribution; this result is fairly consistent across motifs and across networks. In particular the normal approximation can lead to false positive results: some motifs may be thought as being exceptional while they are not.

13.2 Network to model comparison based on subgraph counts

Before reaching statistical conclusions we should assess whether the null model is suitable for subgraph counts. Here is a list of candidate small subgraphs again, see Przulj (2007), where they are called graphlets.
The Graphlet Degree Distribution Agreement (GDDA) is based on orbit degree distributions, which are based on the automorphism orbits of the 29 subgraphs on 2-5 vertices, as follows. Automorphisms are isomorphisms from a graph to itself, and together they form a permutation group. An automorphism orbit is a node that represents this group. Within the 29 subgraphs, 73 different orbits can be found and each one will have an associated orbit degree distribution. An orbit \(i\) from subgraph \(G_j\) has orbit degree \(k\) in the graph \(G\) if there are \(k\) copies of \(G_j\) in \(G\) which involve orbit \(i\).

Let \(d^j_G(k)\) be the sample distribution of the node counts for a given orbit degree \(k\) in a graph \(G\) and for a particular automorphism orbit \(j\). This sample distribution is then scaled by \(1/k\) in order that large degrees do not dominate the score, and normalized to give a total sum of 1,

\[
N^j_G(k) = \frac{d^j_G(k)/k}{\sum_{\ell=1}^{\infty} d^j_G(\ell)/\ell}.
\]

We compare two networks using

\[
D^j(G, H) = \frac{1}{\sqrt{2}} \left( \sum_{k=1}^{\infty} [N^j_G(k) - N^j_H(k)]^2 \right)^{\frac{1}{2}}.
\]

This is turned into an agreement by subtracting from 1, and the agreements are combined into a single value by taking the arithmetic mean over all \(j\),

\[
\text{GDDA} = \frac{1}{73} \sum_{j=0}^{72} (1 - D^j(G, H)).
\]
To interpret the output from a graph comparison based on GDDA, for the ER model and the GEO3D model, graphs of 500, 1000 and 2000 vertices with increasing graph density were generated using. The graphs were subsequently used as query networks in the software and compared with 50 networks of the same model, to ascertain typical GDDA scores if the model is correct.

We compare these scores with using protein interaction networks as input networks. In the figure each value represents the average agreement of 50 networks. The graph densities of the protein interaction networks considered are indicated on the top x axis. The graph density values where the expected number of occurrences of a specific subgraph is approximately equal to 1, for an ER graph with 500 and 2000 nodes are respectively indicated by the short and long arrows along the x axis.

While we see that similar models have overlapping GDDA scores, none of the models under investigation here fits the protein interaction network. We also tried preferential attachment and gene duplication networks. In Rito et al. (2012) we find that the Yeast protein interaction network is highly heterogeneous, which makes model fitting very difficult.

But remember - the model depends on the research question; here we were interested in finding motifs.

(Rito et al., 2010)
Why would protein interaction networks operate near the threshold of the appearance of small subgraphs? We conjecture:

A small number of potential interactions, so that interactions are specific, makes the network efficient.

Some redundancy in the network makes the network robust against small errors.

14 Modules and communities

Finding modules, or communities, which make up the network, has been of interest not only for social networks. Statistically speaking, we would like to apply a clustering method to find out more about the structure of the network. There is an abundance of clustering methods available.

14.1 The Newman-Girvan algorithm

A much used algorithmic approach is the algorithm by Newman and Girvan, see for example Newman and Girvan (2004). Recall that the betweenness of an edge is defined to be the number of shortest paths between node pairs that run along the edge in question, summed over all node pairs.

The algorithm of Girvan and Newman then involves simply calculating the betweenness of all edges in the network and removing the one with highest betweenness, and repeating this process until no edges remain. If two or more edges tie for highest betweenness then one can either choose one at random to remove, or simultaneously remove all of them.

Modularity

As a guidance to how many communities a network should be split into, they use the modularity. For a division with \( g \) groups, define a \( g \times g \) matrix \( e \) whose component \( e_{ij} \) is the fraction of edges in the original network that connect nodes in group \( i \) to those in group \( j \). Then the modularity is defined to be

\[
Q = \sum_i e_{i,i} - \sum_{i,j,k} e_{i,j}e_{k,i},
\]

the fraction of all edges that lie within communities minus the expected value of the same quantity in a graph where the nodes have the same degrees but edges are placed at random. A value of \( Q = 0 \) indicates that the community is no stronger than would be expected by random shuffling.
14.2 The latent position cluster model

Unfortunately the Newman-Girvan algorithm does not provide a measure for statistical significance. The approach by Handcock et al. (2007) in contrast can assess statistical significance of the clusters, for an Erdős-Renyi mixture model. They propose a new model, the latent position cluster model, under which the probability of an edge between two nodes depends on the distance between them in an unobserved Euclidean space, and the nodes’ locations in the latent space arise from a mixture of distributions, each corresponding to a cluster. They propose two estimation methods: a two-stage maximum likelihood method and a fully Bayesian method that uses Markov chain Monte Carlo sampling. The former is quicker and simpler, but the latter performs better.

14.3 Communities appear at different scales

Lewis et al. (2010) show that there is no one scale of interest in the community structure of the yeast protein interaction network, and that almost all proteins lie in a functionally homogeneous community at some scale. They are able to identify a range of resolution parameters that yield the most functionally coherent communities. Also they trace the community membership of a protein through multiple scales.

15 Further topics

Many networks, such as food webs, have a hierarchical structure.

Networks may also be dynamic; our data are snapshots in time.

Edges in networks may be of different types. For example in protein interaction networks there are different types of physical binding between proteins.

In mobile phone networks, short phone calls indicate a different type of interaction than long phone calls.

It is interesting to look at vertices which are “special” in networks, leading to roles in networks. These could be vertices with high degree, or these could be nodes which have low degree but high betweenness, indicating that they may link fairly separate part of the network, or modules. Depending on the scientific question one may like to identify other roles.

For example the yeast protein interaction network appears to exhibit organized modularity where a small proportion of proteins, the 'hubs', interact with many partners. These hubs fall into one of two categories: 'party' hubs, which interact with most of their partners simultaneously, and 'date' hubs, which bind different partners at different locations and times. The biological role of topological hubs may vary depending upon the timing and location of the interactions they mediate.
Further reading

Here is a list of papers and books which may not be found in the recommended reading list and sections within the books on that reading list.


