Chapter 10: Clustering Coefficients of Networks

In this chapter: We introduce the notion of the clustering coefficient. We define the Watts–Strogatz and the Newman clustering coefficients, give some of their mathematical properties and compare them.

10.1 Motivation

Many real-world networks are characterised by the presence of a relatively large number of triangles. This characteristic feature of a network is a general consequence of high transitivity. For instance, in a social network it is highly probable that if Bob and Phil are both friends of Joe then they will eventually be introduced to each other by Joe, closing a transitive relation, i.e. forming a triangle. Our relative measure is between the proportion of triangles existing in a network and the potential number of triangles it can support given the degrees of its nodes. In this Chapter we study two methods of quantifying this property of a network, known as its clustering coefficient.

10.2 The Watts–Strogatz clustering coefficient

The first proposal for a clustering coefficient was put forward by Watts and Strogatz in 1998. If we suppose that the clustering of a node is proportional to

$$C_i = \frac{\text{number of transitive relations of node } i}{\text{total number of possible transitive relations of node } i} \tag{10.1}$$

and t_i designates the number of triangles attached to node i of degree k_i then

$$C_i = \frac{t_i}{k_i(k_i-1)/2} = \frac{2t_i}{k_i(k_i-1)}\,. \tag{10.2}$$

Thus the average clustering coefficient of the network is

$$\overline{C} = \frac{1}{n}\sum_i C_i. \tag{10.3}$$

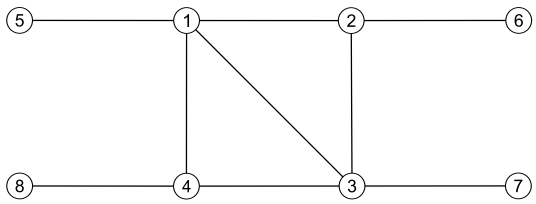


Figure 10.1: A network whose clustering coefficient can be calculated

Example 10.1

- Consider the network illustrated in Figure 10.1. Nodes 1 and 3 are equivalent. They both take part in two triangles and their degrees is 4. Thus,

$$C_1 = C_3 = \frac{2 \cdot (2)}{4 \cdot 3} = \frac{1}{3}. \tag{10.4}$$

In a similar way we obtain

$$C_2 = C_4 = \frac{1}{3}. \tag{10.5}$$

Notice that because nodes 5–8 are not involved in any triangle we have that $C_{i \geq 5} = 0$. Consequently,

$$\overline{C} = \frac{1}{8}\left(\frac{4}{3}\right) = \frac{1}{6}. \tag{10.6}$$

10.3 The Newman clustering coefficient

Another way of quantifying the global clustering of a network is by means of the Newman clustering coefficient, also known as the **transitivity index** of the network. Let $t = |C_3|$ be the total number of triangles, and let $|P_2|$ be the number of paths of length 2 in the network (representing all potential three-way relationships). Then,

$$C = \frac{3t}{|P_2|} = \frac{3|C_3|}{|P_2|}. \tag{10.7}$$

Example 10.2

- Consider again the network illustrated in Figure 10.1. We can obtain the number of triangles in that network by using the spectral properties of the adjacency matrix. That is,

$$t = \frac{1}{6}\,\mathrm{tr}(A^3) = 2. \tag{10.8}$$

The number of paths of length 2 in the network can be obtained using the following formula (which we will justify in Chapter 13).

$$|P_2| = \sum_{i=1}^{n} \binom{k_i}{2} = \sum_{i=1}^{n} \frac{k_i(k_i-1)}{2} = 18. \tag{10.9}$$

Thus,

$$C = \frac{3 \times 2}{18} = \frac{1}{3}. \tag{10.10}$$

Problem 10.1

- Consider the network illustrated in Figure 10.2. Obtain an expression for the average clustering, $\overline{C}$, and the network transitivity, C, in terms of the number of nodes n. Analyse your results as $n \to \infty$.

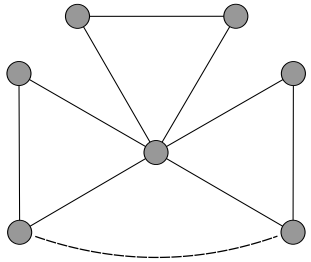


Figure 10.2: A network formed by triangles joined at a central node. The dashed line indicates the existence of other triangles.

To answer this problem, observe that there are two types of nodes in the network, which will be designated as i and j (see Figure 10.3). There is one node of type i and $n-1$ nodes of type j.

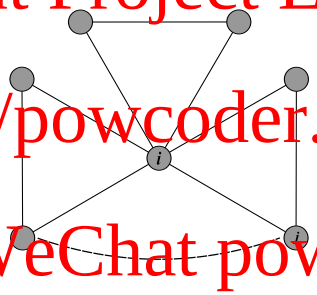


Figure 10.3: A labelling of Figure 10.2 to indicate nodes with different properties

The average clustering coefficient is then,

$$\overline{C} = \frac{C_i + (n-1)C_j}{n}. \tag{10.11}$$

Evidently, $C_j = 1$ and

$$C_i = \frac{2t}{k_i(k_i-1)}, \tag{10.12}$$

where t is the number of triangles in the network (note that node i is involved in all of them) and k_i is the degree of that node. It is easy to see that $k_i = 2t = n-1$. Then,

$$C_i = \frac{2t}{2t(2t-1)} = \frac{1}{2t-1} = \frac{1}{n-2} \tag{10.13}$$

and

$$\overline{C} = \frac{C_i + (n-1)C_j}{n} = \frac{\left(\frac{1}{n-2}\right) + (n-1)\cdot 1}{n} = \frac{1}{n(n-2)} + \frac{n-1}{n}. \tag{10.14}$$

Now, for the Newman transitivity coefficient we have

$$\begin{aligned} |P_2| &= \sum_u \binom{k_u}{2} = \frac{1}{2}\sum_u k_u(k_u-1) = \frac{(n-1)}{2}(2\times 1) + \frac{2t(2t-1)}{2} \\ &= 2t + t(2t-1) = t(2t+1). \end{aligned} \tag{10.15}$$

Thus,

$$C = \frac{3t}{|P_2|} = \frac{3t}{t(2t+1)} = \frac{3}{2t+1} = \frac{3}{n}. \tag{10.16}$$

As the number of nodes tends to infinity we have:

$$\lim_{n\to\infty} \overline{C} = \lim_{n\to\infty} \frac{1}{n(n-2)} + \lim_{n\to\infty} \frac{n-1}{n} = 0 + 1 = 1, \tag{10.17}$$

$$\lim_{n\to\infty} C = \lim_{n\to\infty} \frac{3}{n} = 0. \tag{10.18}$$

This indicates that the indices are accounting for different structural characteristics of a network.

In general, the Watts–Strogatz index quantifies how clustered a network is locally, while the Newman index indicates how clustered the network is as a whole. Often there is a good correlation between both indices for real-world networks as illustrated in Figure 10.4.

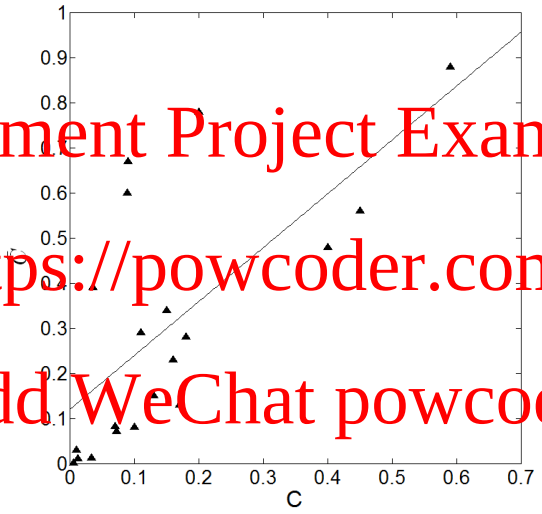


Figure 10.4: Correlation between Watts–Strogatz ($\overline{C}$) and Newman (C) clustering coefficients for 20 real-world networks.

Further Reading

Estrada, E., *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, 2011, Chapter 4.5.1.

Newman, M.E.J., *Networks: An Introduction*, Oxford University Press, 2010, Chapter 7.9.

Chapter 11: Random Models of Networks

In this chapter: We introduce simple and general models for generating random networks: the Erdös–Rényi model, the Barabási–Albert model and the Watts–Strogatz model. We study some of the general properties of the networks generated by using these models, such as their densities, average path length and clustering coefficient, as well as some of their spectral properties.

11.1 Motivation

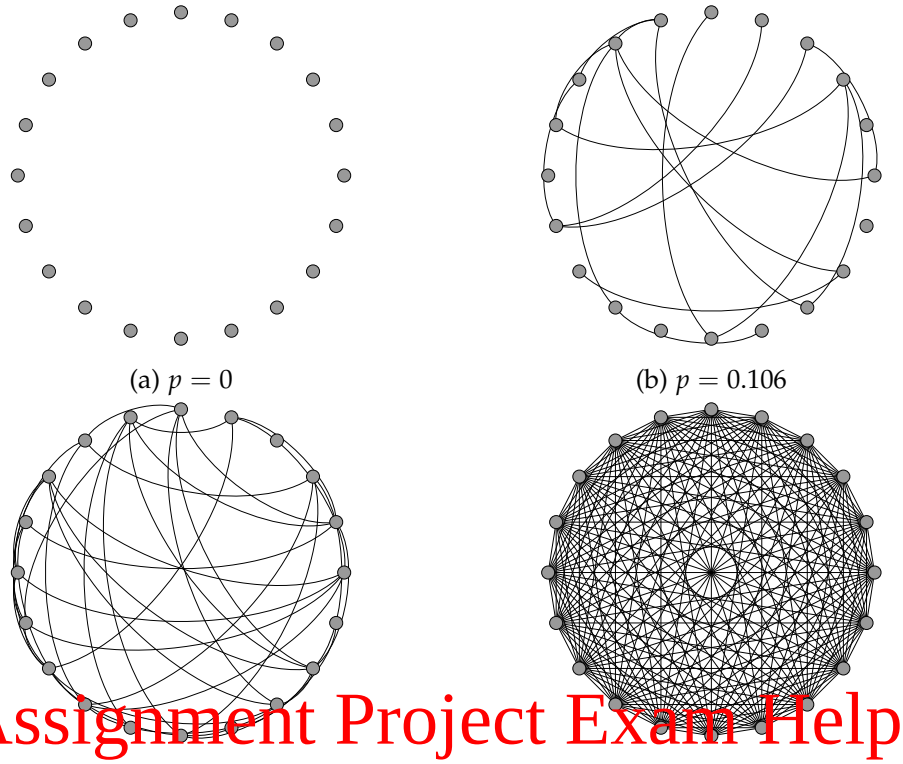
Every time that we look at a real-world network and analyse its most important topological properties it is worth considering how that network was created. In other words, we have to figure out what are the mechanisms behind the evolution of a group of nodes and links which give rise to the topological structure we observe.

Intuitively we can think about a model in which pairs of nodes are connected with some probability. That is, if we start with a collection of n nodes and for each of the $n(n-1)/2$ possible links, we connect a pair of nodes u, v with certain probability $p_{u,v}$. Then, if we consider a set of network parameters to be fixed and allow the links to be created by a random process, we can create models that permit us to understand the influence of these parameters on the structure of networks. Here we study some of the better known models that employ such mechanisms.

11.2 The Erdös–Rényi model of random networks

In this model, put forward by Erdös and Rényi in 1959, we start with n isolated nodes. We then pick a pair of nodes and with probability $p > 0$ we add a link between them. In practice we fix a parameter value p from which we generate the network. For each pair of nodes we generate a random number, r, uniformly from $[0, 1]$ and if $p > r$ we add a link between them. Consequently, if we select $p = 0$ the network will remain fully disconnected forever and if $p = 1$ we end up with a complete graph. In Figure 11.1 we illustrate some examples of Erdös–Rényi random networks with 20 nodes and different linking probabilities.

The Erdös–Rényi (ER) random network is written as either $G_{ER}(n, m)$ or $G_{ER}(n, p)$. A few properties of the random networks generated by this model are summarised below.



(a) $p = 0$ (b) $p = 0.106$

(c) $p = 0.265$ (d) $p = 1$

Figure 11.1: Erdös–Rényi random networks for different probabilities p

1. The expected number of edges is $\bar{m} = \frac{n(n-1)p}{2}$.

2. The expected node degree is $\bar{k} = (n-1)p$.

3. The degrees follow a Poisson distribution $p(k) = \frac{e^{-\bar{k}}\bar{k}^k}{k!}$ as illustrated in Figure 11.2 for ER random networks with 1000 nodes and 4000 links. The solid line is the expected distribution and the dots represent the values for the average of 100 realizations.

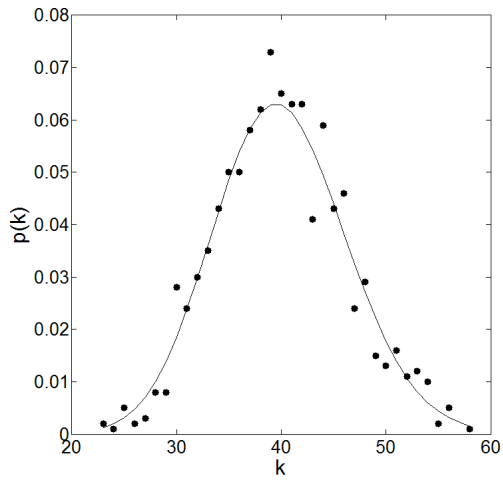


Figure 11.2: Erdös–Rényi degree distribution

4. The average path length for large n is

$$\bar{l}(G) = \frac{\ln n - \gamma}{\ln(pn)} + \frac{1}{2}, \tag{11.1}$$

where $\gamma \approx 0.577$ is the Euler–Mascheroni constant.

5. The average clustering coefficient is $\overline{C} = p$.

6. As p increases, most nodes tend to be clustered in one giant component, while the rest of nodes are isolated in very small components. In Figure 11.3 we illustrate the change of the size of the main connected component in an ER random network with 1000 nodes as a function of the linking probability.

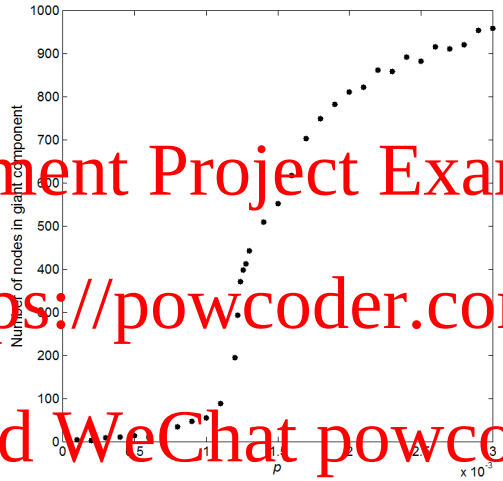


Figure 11.3: Connectivity of Erdös–Rényi random networks

7. The structure of $G_{ER}(n, p)$ changes as a function of $p = \bar{k}/(n-1)$ giving rise to the following three stages.

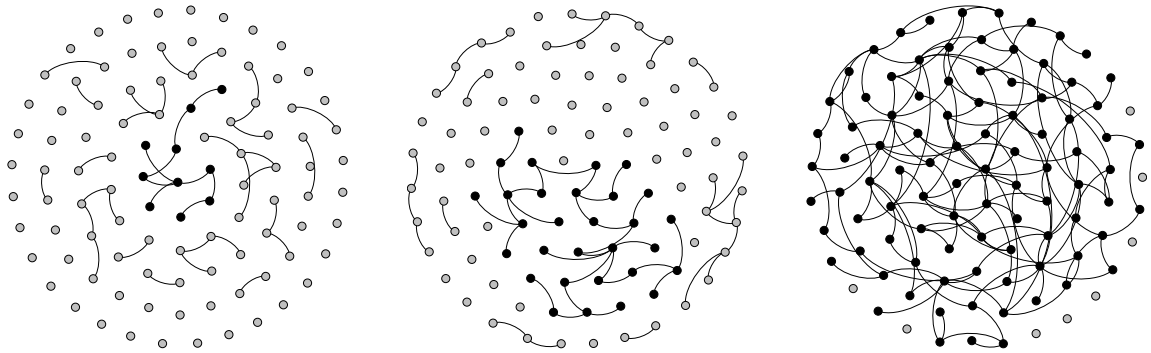
   (a) *Subcritical* $\bar{k} < 1$, where all components are simple and very small. The size of the largest component is $S = O(\ln n)$.

   (b) *Critical* $\bar{k} = 1$, where the size of the largest component is $S = O(n^{2/3})$.

   (c) *Supercritical* $\bar{k} > 1$, where the probability that $(f - \varepsilon)n < S < (f + \varepsilon)n$ is 1 when $n \to \infty$, $\varepsilon > 0$, and where $f = f(\bar{k})$ is the positive solution of the equation $e^{-\bar{k}f} = 1 - f$. The rest of the components are very small, with the second largest having size about $\ln n$.

   In Figure 11.4 we illustrate this behaviour for an ER random network with 100 nodes and different linking probabilities. The nodes in the largest connected component are drawn in a darker shade.

8. The largest eigenvalue of the adjacency matrix in an ER network grows proportionally to n so that $\lim_{n\to\infty} \frac{\lambda_1(A)}{n} = p$.



(a) $p = 0.0075$, $S = 10$ (b) $p = 0.01$, $S = 30$ (c) $p = 0.025$, $S = 92$

Figure 11.4: Emergence of a giant component in Erdös–Rényi random networks

9. The second largest eigenvalue grows more slowly than λ_1. In fact,

$$\lim_{n\to\infty} \frac{\lambda_2(A)}{n^\varepsilon} = 0$$

for every $\varepsilon > 0.5$.

10. The most negative eigenvalue grows in a similar way to $\lambda_2(A)$. Namely,

$$\lim_{n\to\infty} \frac{\lambda_n(A)}{n^\varepsilon} = 0$$

for every $\varepsilon > 0.5$.

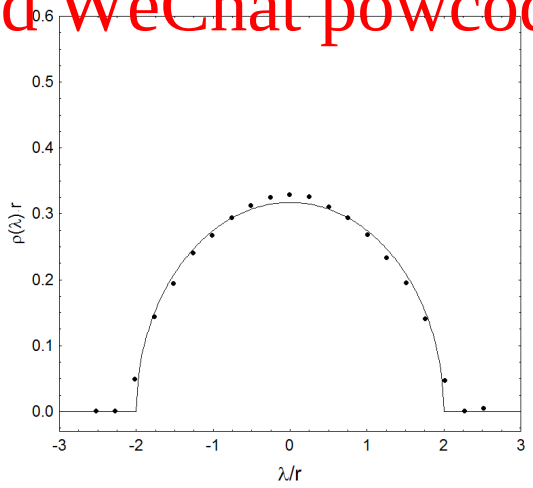


Figure 11.5: Spectral density for a network generated with the ER model

11. The spectral density of an ER random network follows Wigner's semicircle law. That is, almost all of the eigenvalues of an ER random network lie in the range $[-2r, 2r]$ where $r = \sqrt{np(1-p)}$ and within this range the density function is given by

$$\rho(\lambda) = \frac{\sqrt{4r^2 - \lambda^2}}{2\pi r^2}.$$

This is illustrated in Figure 11.5.

12. For an ER random graph we know that $\lim_{n\to\infty} \frac{\lambda_1}{n} = p$, and $\lim_{n\to\infty} \frac{\lambda_2}{n^\varepsilon} = 0$ for $\varepsilon > 0.5$. Thus,

$$\lim_{n\to\infty} \left(\frac{\lambda_1}{n} - \frac{\lambda_2}{n^\varepsilon} \right) = p,$$

which means that in the limit $\lambda_1 \geq n^\alpha \lambda_2$, for some $\alpha < 0.5$. The exponent α depends on the density of the network. For networks with very low density α is small, but as soon as the density of the network increases, this exponent approaches 0.5 asymptotically. This means that the spectral gap in an ER network is

$$\lambda_1 - \lambda_2 \geq (n^\alpha - 1)\, \lambda_2,$$

indicating that $\lambda_1 - \lambda_2$ grows with the density of the network. For instance, for ER networks with density 0.008 the spectral gap is about 3, while for density 0.08 it is about 60.

We will make the simplifying assumption $\lambda_2 \to 0$ as $n \to \infty$ in some of our examples.

11.3 The Barabási–Albert model

The ER model generates networks with Poisson degree distributions. However, it has been empirically observed that many networks in the real-world have a fat-tailed degree distribution of some kind, which varies greatly from the distribution observed for ER random networks. A simple model to generate networks in which the probability of finding a node of degree k decays as a power law of the degree was put forward by Barabási and Albert in 1999. We initialise with a small network with m_0 nodes. At each step we add a new node u to the network and connect it to $d \leq m_0$ of the existing nodes $v \in V$. The probability of attaching node u to node v is proportional to the degree of v. That is, we are more likely to attach new nodes to existing nodes with high degree. This process is known as preferential attachment.

We can assume that our initial random network is connected and of ER type with m_0 nodes, $G_{ER} = (V, E)$. In this case the Barabási–Albert (BA) algorithm can be understood as a process in which small inhomogeneities in the degree distribution of the ER network grow in time. A typical BA network is illustrated in Figure 11.6.

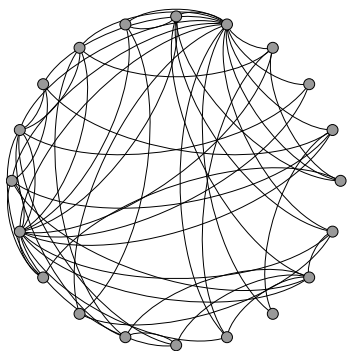
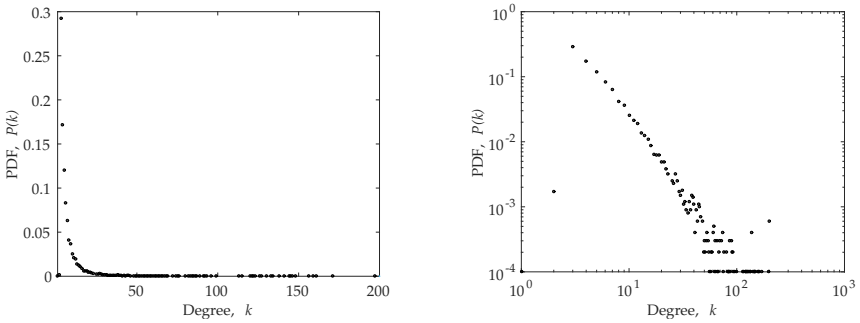


Figure 11.6: A Barabási–Albert network with $n = 20$ and $m = 4$.

Networks generated by this model have several global properties

1. The probability that a node has degree $k \geq d$ is given by

$$p(k) = \frac{2d(d-1)}{k(k+1)(k+2)} \approx k^{-3}. \tag{11.2}$$



(a) Linear scale (b) log–log scale

Figure 11.7: The characteristic power-law degree distribution of a BA network

That is, the distribution is close to a power-law as illustrated in Figure 11.7.

2. The cumulative degree distribution is $P(k) \approx k^{-2}$, illustrated on a log–log scale in Figure 11.8

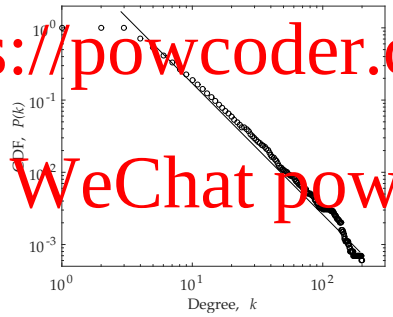


Figure 11.8: Cumulative degree distribution for a network generated with the BA model

3. The expected value for the clustering coefficient, $\overline{C}$, approximates $\frac{d-1}{8}\frac{\log^2 n}{n}$ as $n \to \infty$.

4. The average path length is given by

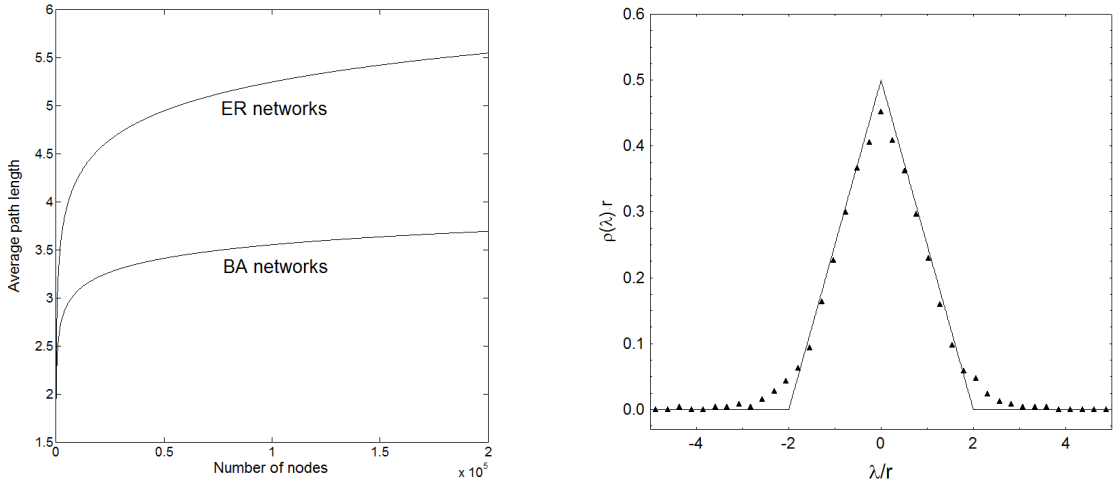
$$\bar{l} = \frac{\ln n - \ln(d/2) - 1 - \gamma}{\ln\ln n + \ln(d/2)} + \frac{3}{2}, \tag{11.3}$$

where again γ is the Euler–Mascheroni constant. Comparing (11.3) with (11.1) we find that for the same number of nodes and average degree, BA networks have smaller average path length than their ER analogues. We illustrate this in Figure 11.9a, which shows the change in the average path length of random networks created with the BA and ER models as the number of nodes increases.

5. The density of eigenvalues follows a triangle distribution

$$\rho(\lambda) = \begin{cases} (\lambda + 2r)/(4r^2), & -2 \leq \lambda/r \leq 0, \\ (2r - \lambda)/(4r^2), & 0 \leq \lambda/r \leq 2, \\ 0, & \text{otherwise.} \end{cases} \tag{11.4}$$

This is illustrated in Figure 11.9b.



(a) (b)

Figure 11.9: (a) Comparison of the small-worldness of BA and ER networks. (b) Spectral density of a model BA network

The BA model can been generalised to fit general power-law distributions where the probability of finding a node with degree k decays as a negative power of the degree: $p(k) \sim k^{-\gamma}$.

11.4 The Watts–Strogatz model

The phrase "six degrees of separation" is commonly used to express how surprisingly closely connected we are to each other in terms of shared acquaintances. The phrase originates from a famous experiment in network theory. Stanley Milgram carried out the experiment in 1967 he asked some randomly selected people in the U.S. cities of Omaha (Nebraska) and Wichita (Kansas) to send a letter to a target person who lived in Boston (Massachusetts) on the East Coast. The rules stipulated that the letter should be sent to somebody the sender knew personally. Although the senders and the target were separated by about 2,000 km and that there were 200 million inhabitants in the USA at the time, Milgram found two characteristic effects. First, the average number of steps needed for the letters to arrive to its target was around 6. And second, there was a large group inbreeding, which resulted in acquaintances of one individual feeding a letter back into his/her own circle, thus usually eliminating new contacts.

Although the ER model reproduces the first characteristic very well, i.e., that most nodes are separated by a very small average path length, it fails in reproducing the second. That is, the clustering coefficient in the ER network is very small in comparison with those observed in real-world systems. The model put forward by Watts and Strogatz in 1998 tries to sort out this situation.

First we form the circulant network with n nodes connected to k neighbours. We then rewire some of its links: each of the original links has a probability p (fixed beforehand) of having one of its end points moved to a new randomly chosen node. If p is too high, meaning almost all links are random, we approach the ER model.

The general process is illustrated in Figure 11.10. On the left is a circulant graph and on the right is a random ER network. Somewhere in the middle are the so-called "small-world" networks.

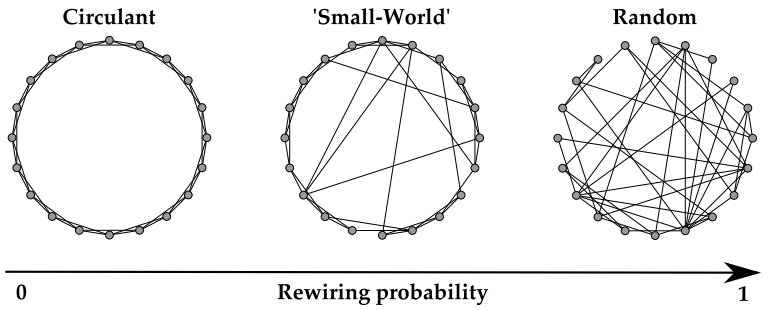
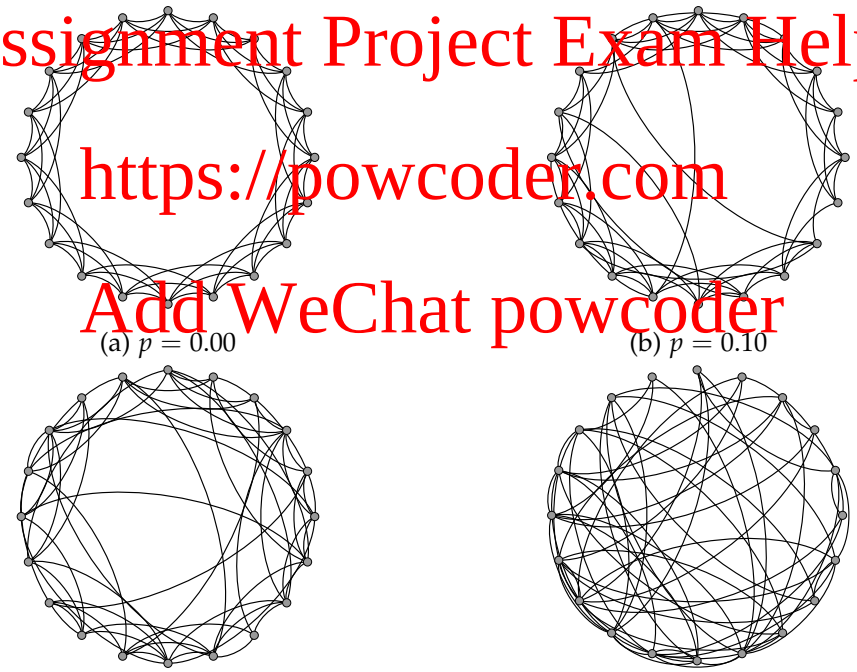


Figure 11.10: Schematic representation of the Watts–Strogatz rewiring process

In Figure 11.11 we illustrate the rewiring process, which is the basis of the Watts–Strogatz (WS) model for small-world networks. Starting from a regular circulant network with $n = 20, k = 6$ links are rewired with different choices of probability p.



(a) $p = 0.00$ (b) $p = 0.10$

(c) $p = 0.50$ (d) $p = 0.95$

Figure 11.11: Watts–Strogatz random networks for different rewiring probabilities p

Networks generated by the WS model have several general properties, listed below.

1. For $p = 0$ the average clustering coefficient is given by $\overline{C} = \dfrac{3(k-2)}{4(k-1)}$. For large values of k, $\overline{C}$ approaches 0.75. As p increases $\overline{C} \to k/n$. The decay is slow.

2. As p increases, the average path length decays very fast from that of a circulant graph,

$$\bar{l} = \frac{(n-1)(n+k-1)}{2kn}, \tag{11.5}$$

to approach that of a random network. In Figure 11.12 we illustrate the effect of changing the rewiring probability on both the average path length and clustering coefficient on random graphs generated using the WS model with 100 nodes and 5,250 links.

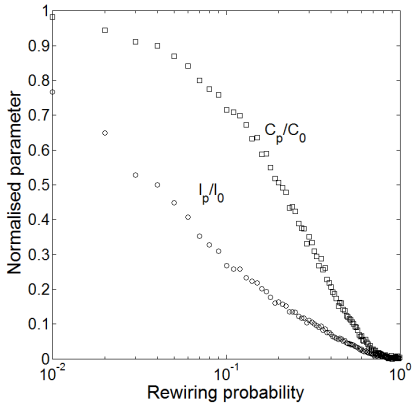


Figure 11.12: Changes in network statistics for Watts–Strogatz graphs

Problem 11.1

- Let $G_{ER}(n, p)$ be an Erdös–Rényi random network with n nodes and probability p. Use known facts about the spectra of ER random networks to show that if $\bar{k}$ is the average degree of this network then the expected number of triangles tends to $\bar{k}^3/6$ as $n \to \infty$.

  For any graph the number of triangles is given by

$$t = \frac{1}{6}\mathrm{tr}(A^3) = \frac{1}{6}\sum_{j=1}^{n} \lambda_j^3. \tag{11.6}$$

  In an ER graph we know that

$$\lim_{n\to\infty} \lambda_1 = np, \ \lim_{n\to\infty} \frac{\lambda_2}{n^\varepsilon} = 0, \ \lim_{n\to\infty} \frac{\lambda_n}{n^\varepsilon} = 0, \tag{11.7}$$

  and, since $|\lambda_i| \leq \max\{|\lambda_2|, |\lambda_n|\}$ for $i \geq 1$,

$$\lim_{n\to\infty} t = \frac{1}{6}\lambda_1^3 = \frac{1}{6}(np)^3. \tag{11.8}$$

  Since $\bar{k} = p(n-1)$, as $n \to \infty$,

$$t \to \frac{\bar{k}^3}{6}. \tag{11.9}$$

Problem 11.2

- The data shown in Table 11.1 belongs to a network having $n = 1000$ nodes and $m = 4000$ links. The network does not have any node with $k \leq 3$. Let $n(k)$ be the number of nodes with degree k. Determine whether this network was generated by the BA model.

  The probability that a node chosen at random has a given degree is shown in Table 11.2. A sketch of the plot of k against $p(k)$ in Figure 11.13 indicates that there is a fast decay of the probability

k	$n(k)$
4	343
5	196
10	23
20	3

Table 11.1: Degree frequencies in an example network

k	$p(k)$
4	0.343
5	0.196
10	0.023
20	0.003

Table 11.2: Probability distribution of degrees in an example network

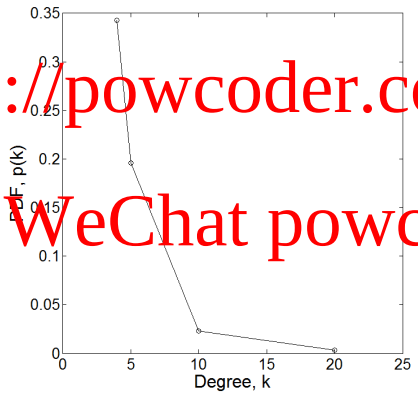


Figure 11.13: Degree distribution in an illustrative network

with the degree, which is indicative of fat-tailed degree distributions like the one produced by the BA model.

If the network was generated with the BA model it has to have a PDF of the form $p(k) \sim k^{-3}$ which means that $\ln p(k) \sim -3 \ln k + b$.

Given two degree values k_1 and k_2, the slope of a log-log plot is given by

$$m = \frac{\ln(p(k_2)) - \ln(p(k_1))}{\ln k_2 - \ln k_1}. \tag{11.10}$$

Using data from the table,

$$m = \frac{\ln(0.003) - \ln(0.196)}{\ln 20 - \ln 5} = -3.0149 \approx -3, \tag{11.11}$$

indicative of a network generated by the BA model.

Further Reading

Barabási, A.-L. and Albert, R., *Emergence of scaling in random networks*, Science **286**:509–512, 1999.

Bollobás, B., Mathematical results on scale-free random graphs, in Bernholdt, S. and Schuster, H.G. (eds.), *Handbook of Graph and Networks: From the Genome to the Internet*, Wiley-VCH, 1–32, 2003.

Bollobás, B., *Random Graphs*, Cambridge University Press, 2001.

Watts, D.J., Strogatz, S.H., Collective dynamics of 'small-world' networks, Nature **393**:440–442, 1998.

Chapter 13: Fragment-based Measures

In this chapter: We start with the definition of a fragment, or subgraph, in a network. We then introduce the concept of network motif and analyse how to quantify its significance. We illustrate the concept by studying motifs in some real-world networks. We then outline mathematical methods to quantify the number of small subgraphs in networks analytically. We develop some general techniques that can be adapted to the search for other fragments.

13.1 Motivation

In many real-life situations we are able to identify small structural pieces of a system which are responsible for certain functional properties of the whole system. Biologists, chemists and engineers usually isolate these small fragments of the system to understand how they work and gain understanding of their roles in the whole system. These kinds of structural fragments exist in complex networks. In Chapter 11 we saw that triangles can indicate transitive relations in social networks. They also play a role in interactions between other entities in complex systems. In this chapter we develop techniques to quantify some of the simplest but most important fragments or subgraphs in networks. We also show how to determine whether the presence of these fragments in a real-world network is just a manifestation of a random underlying process or that they signify something more significant.

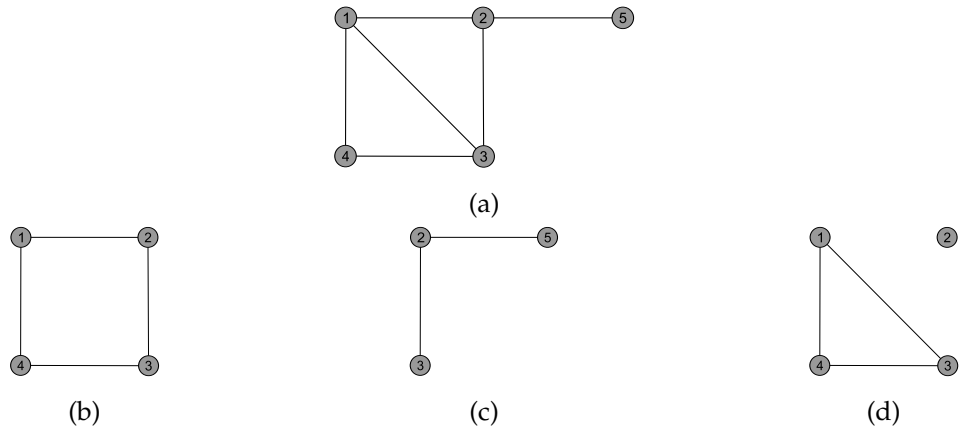


Figure 13.1: Three subgraphs (bottom line) in an undirected network (top)

In network theory fragments are synonymous with subgraphs. Typical subgraphs are illustrated in Figure 13.1. In general, a subgraph can be formed by one (connected subgraph) or more (disconnected subgraphs) connected components and they may be cyclic or acyclic.

13.2 Counting subgraphs in networks

In order to count subgraphs in a network we need to use a combination of algebraic and combinatorial techniques. First, we are going to develop some basic techniques which can be combined to count many different types of subgraph.

13.2.1 Counting stars

We know from previous chapters that the number of edges in a network can be obtained from the degrees of the nodes. An edge is a star subgraph of the type $S_{1,1}$. Thus

$$|S_{1,1}| = m = \frac{1}{2}\sum_{i=1}^{n} k_i. \tag{13.1}$$

The next star subgraph is $S_{1,2}$. Copies of $S_{1,2}$ in a network can be enumerated by noting that formed from any two edges incident to a common node. That is, $|S_{1,2}|$ is equal to the number of times that the nodes attached to a particular node can be combined in pairs. This is simply

$$|S_{1,2}| = \sum_{i=1}^{n} \binom{k_i}{2} = \frac{1}{2}\sum_{i=1}^{n} k_i(k_i - 1). \tag{13.2}$$

Since $S_{1,2}$ is the same as P_2, we have generated the formula we used in calculating transitivity in Chapter 10.

Similarly, the number of $S_{1,3}$ star subgraphs equals the number of times that the nodes attached to a given node can be combined in triples, namely,

$$|S_{1,3}| = \sum_{i=1}^{n} \binom{k_i}{3} = \frac{1}{6}\sum_{i=1}^{n} k_i(k_i - 1)(k_i - 2). \tag{13.3}$$

In general, the number of star subgraphs of the type $S_{1,s}$ is given by

$$|S_{1,s}| = \sum_{i=1}^{n} \binom{k_i}{s}. \tag{13.4}$$

13.2.2 Using closed walks

The idea of using closed walks (CWs) to count subgraphs is very intuitive and simple. Every time that we complete a CW in a network we have visited a sequence of nodes and edges which together form a subgraph. For instance, a CW that goes from a node to any of its neighbours and back again describes an edge, while every CW of length three necessarily visits all the nodes of a triangle. Note that CWs of length $l = 2d, d = 1, 2, \ldots$, that go back and forth between adjacent nodes also describe edges. Similarly, a CW of length $l = 2d + 1, d = 1, 2, \ldots$, visiting only 3 nodes of the network describes

a triangle. Keeping this in mind, we can design a technique to express the number of CWs as a sum of fragment contributions. We start by designating by μ_l the number of CWs of length l in a network. Then $\mu_0 = n$ and, in a simple network, $\mu_1 = 0$. In general, we know that

$$\mu_k = \text{tr}(A^k) = \sum_{i=1}^{n} \lambda_i^k$$

but we can rewrite the right-hand side of this expression in terms of particular subgraphs. Before continuing, visualise what happens with a CW of length 2. Each such walk represents an edge. But in an undirected network there are two closed walks along each edge (i, j), namely $i \to j \to i$ and $j \to i \to j$. Thus,

$$\mu_2 = 2|S_{1,1}| = 2|P_1|.$$

Similarly, there are 6 CWs of length 3 around every triangle $\triangle_{i,j,k}$ since we can start from any one of its three nodes and move either clockwise or counter-clockwise: $i \to j \to k \to i$; $i \to k \to j \to i$; $j \to k \to i \to j$; $j \to i \to k \to j$; $k \to i \to j \to k$; $k \to j \to i \to k$. So,

$$\mu_3 = 6|C_3|.$$

Things begin to get messy for longer CWs as there are several subgraphs associated with such walks. For example, a CW of length 4 can be generated by moving along the same edge four times. This can be done in two ways

$i \to j \to i \to j \to i$ and $j \to i \to j \to i \to j$.

We could also walk along two edges and then return to the origin in two ways,

$i \to j \to k \to j \to i$ and $k \to j \to i \to j \to k$.

There are two ways of visiting two nearest neighbours before returning to the origin,

$$j \to i \to j \to k \to j \text{ and } j \to k \to j \to i \to j.$$

And finally, there are eight ways of completing a cycle of length four in a square i, j, k, l since we can start from any node and go clockwise or anticlockwise. For example, starting from node i gives

$$i \to j \to k \to l \to i \text{ and } i \to l \to k \to j \to i.$$

Consequently,

$$\mu_4 = 2|P_1| + 4|P_2| + 8|C_4|.$$

Problem 13.1

- Let G be a regular network with $n = 2r$ nodes of degree k and spectrum

$$\sigma(G) = \{[k]^1, [1]^{r-1}, [-1]^{r-1}, [-k]^1\}. \tag{13.5}$$

  Find expressions for the number of triangles and squares in G.

  The number of triangles in a network is given by

$$t = \frac{1}{6}\text{tr}(A^3) = \frac{1}{6}\sum_{i=1}^{n} \lambda_i^3 = \frac{1}{6}\Big(k^3 + (-k)^3 + (r-1)\Big(1^3 + (-1)^3\Big)\Big) = 0.$$

The number of squares is given by $|C_4| = \mu_4/8 - |P_1|/4 - |P_2|/2$. Since each node has degree k,

$$|P_1| = \frac{kn}{2} = kr$$

and

$$|P_2| = \sum_{i=1}^{n} \binom{k}{2} = \frac{nk(k-1)}{2} = rk(k-1). \tag{13.6}$$

Given that

$$\mu_4 = \mathrm{tr}(A^4) = k^4 + (-k)^4 + (r-1) + (r-1)(-1)^4 = 2k^4 + 2(r-1), \tag{13.7}$$

we conclude that

$$|C_4| = \frac{k^4 + (r-1)}{4} - \frac{rk(k-1)}{2} - \frac{rk}{4} = \frac{k^4 + r - 1 - 2rk(k-1) - rk}{4} = \frac{k^4 + r - rk(2k-1) - 1}{4}. \tag{13.8}$$

Example 13.1

- A network of the type described in the last problem is the cube Q_3 which has the spectrum

$$\sigma(G) = \{[3]^1, [1]^3, [-1]^3, [-3]^1\}.$$

Applying the formula obtained for the number of squares gives

$$|C_4| = \frac{3^4 + 4 - 60 - 1}{4} = 6,$$

which is the number of faces on a cube.

13.2.3 Combined techniques

We start by considering a practical example. In this case we will be interested in computing the number of fragments of the type illustrated in Figure 13.2. This fragment is known as a tadpole $T_{3,1}$ subgraph.

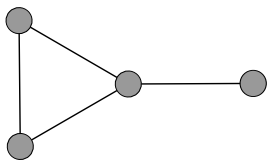


Figure 13.2: The tadpole graph $T_{3,1}$

The fragment $T_{3,1}$ is characterised by having a node which is simultaneously part of a triangle and of a path of length 1. We can combine the idea of calculating the number of triangles in which the node is involved with the number of nodes attached to it. Let t_i be the number of triangles attached to the node i and let $k_i > 2$ be the degree of this node. The number of nodes not in the triangle which are attached to i is just the remaining degree of the node, i.e., $k_i - 2$. Thus, the number of tadpole subgraphs in which the node i is involved is $k_i - 2$ times its number of triangles. Consequently,

$$|T_{3,1}| = \sum_{k_i > 2} t_i(k_i - 2). \tag{13.9}$$

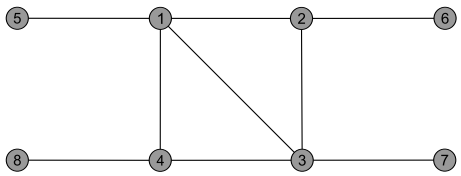


Figure 13.3: A network with many tadpole fragments

Example 13.2

- Find the number of fragments $T_{3,1}$ in Figure 13.3.

  Using (13.9) and concentrating only on those nodes with degree larger than 2 we have

  $$|T_{3,1}| = 2 \times (4-2) + 1 \times (3-2) + 2 \times (4-2) + 1 \times (3-2) = 10.$$

  Can you see all of them?

Let's add another edge and consider the fragment illustrated in Figure 13.4.

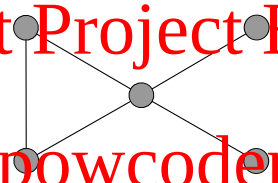


Figure 13.4: Illustration of the cricket graph

This subgraph is known as the cricket graph, which we designate by Cr. Here again, we can use a technique that combines the calculation of the two subgraphs forming this fragment. That is, this fragment is characterised by a node i that is simultaneously part of a triangle and a star $S_{1,2}$. Using t_i as before, we consider nodes for which $k_i > 3$.

If $t_i > 0$ then node i has $k_i - 2$ additional nodes which are attached to it. These $k_i - 2$ nodes can be combined in $\binom{k_i-2}{2}$ pairs to form all the $S_{1,2}$ subgraphs in which node i is involved.

The number of crickets involving node i is then

$$|Cr_i| = t_i \binom{k_i - 2}{2} \tag{13.10}$$

and hence

$$|Cr| = \sum_{k_i \geq 4} t_i \binom{k_i - 2}{2} = \frac{1}{2} \sum_{k_i \geq 4} t_i (k_i - 2) \cdot (k_i - 3). \tag{13.11}$$

Example 13.3

- Find the number of Cr fragments in the network illustrated in Figure 13.3.

  Only the nodes labelled as 1 and 3 have degree larger than or equal to 4. We have,

  $$|Cr| = \frac{1}{2}(2(4-2)(4-3) + 2(4-2)(4-3)) = 4. \tag{13.12}$$

  The four crickets are illustrated in Figure 13.5:

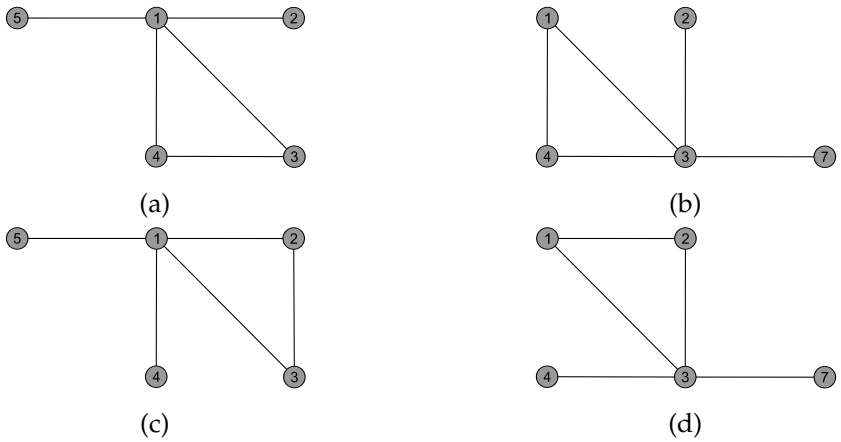


Figure 13.5: Illustration of the four cricket subgraphs within the network in Figure 13.3

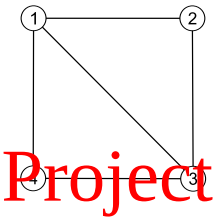


Figure 13.6: The diamond graph

13.2.4 Other techniques

The diamond graph (D) is characterised by the existence of two connected nodes (1 and 3) which are also connected by two paths of length 2 (1-2-3 and 1-4-3). It is illustrated in Figure 13.6. To calculate the number of diamonds in a network we note that the number of walks of length 2 between two **connected** nodes is given by $(A^2)_{ij}A_{ij}$ and hence that the number of pairs of paths of length two among two connected nodes i, j is given by

$$\binom{(A^2)_{ij}A_{ij}}{2}.$$

Consequently, the number of diamond subgraphs in a network is given by

$$|D| = \frac{1}{2}\sum_{i,j}\binom{(A^2)_{ij}A_{ij}}{2} = \frac{1}{4}\sum_{i,j}\Big((A^2)_{ij}A_{ij}\Big)\Big((A^2)_{ij}A_{ij}-1\Big). \tag{13.13}$$

Problem 13.2

- Find an expression for $|C_5|$, the number of pentagons in a network.

  A CW of length $l = 2d+1$ necessarily visits only nodes in subgraphs containing at least one odd cycle. So a CW of length 5 can visit only the nodes of a triangle, C_3; a tadpole, $T_{3,1}$; or a pentagon, C_5. Hence

  $$\mu_5 = a|C_3| + b|T_{3,1}| + c|C_5| \tag{13.14}$$

  and

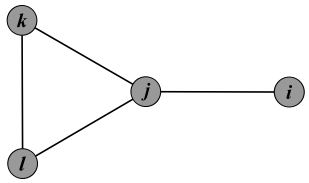
  $$|C_5| = \frac{1}{c}(\mu_5 - a|C_3| - b|T_{3,1}|). \tag{13.15}$$

We have seen how to calculate $|C_3|$ and $|T_{3,1}|$ hence our task is to determine the coefficients a, b and c.

To find a we must enumerate all the CWs of length 5 in a triangle. This can be done by calculating $\mathrm{tr}(A_C^5)$ where A_C is the adjacency matrix of C_3. From Chapter **??** we know that the eigenvalues of C_3 are 2, -1 and -1 hence

$$a = \sum_i \lambda_j^5 = 2^5 + 2(-1)^5 = 30. \tag{13.16}$$

To find b we can enumerate all the CWs of length 5 involving all the nodes of $T_{3,1}$. This is done in Figure 13.7 and we see that $b = 10$.



$i \to j \to k \to l \to j \to i$	$j \to i \to j \to l \to k \to j$
$i \to j \to l \to k \to j \to i$	$k \to l \to j \to i \to j \to k$
$j \to k \to l \to j \to i \to j$	$k \to j \to i \to j \to l \to k$
$j \to l \to k \to j \to i \to j$	$l \to j \to i \to j \to k \to l$
$j \to i \to j \to k \to l \to j$	$l \to k \to j \to i \to j \to l$

Figure 13.7: CWs of length 5 in $T_{3,1}$.

We could also proceed in a similar way as for the triangle, but in the tadpole $T_{3,1}$ not every CW of length 5 visits all the nodes of the fragment. That is, there are CWs of length 5 which only visit the nodes of the triangle in $T_{3,1}$. Thus,

$$b = \mathrm{tr}(A_T^5) - a \tag{13.17}$$

where A_T is the adjacency matrix of $T_{3,1}$. Computing A_T^5 explicitly we find that $\mathrm{tr}(A_T^5) = 40$ and, again,

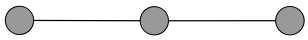
$$b = 40 - 30 = 10.$$

Finally, to find c note that there for every node in C_5 there is one CW of length 5 in a clockwise direction another counter-clockwise, e.g., $i \to j \to k \to l \to m \to i$ and $i \to m \to l \to k \to j \to i$. Thus, $c = 10$. Finally,

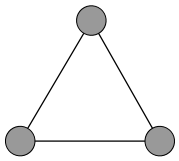
$$|C_5| = \frac{1}{10}\left(\mu_5 - 30|C_3| - 10|T_{3,1}|\right). \tag{13.18}$$

13.2.5 Formulae for counting small subgraphs

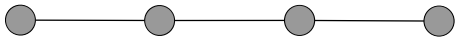
Formulae for a number of simple subgraphs can be derived using very similar techniques to the ones we've encountered so far. The results are summarised over the next few pages.

$F1$ 

$$|F_1| = \frac{1}{2}\sum_i k_i(k_i - 1)$$

$F2$ 

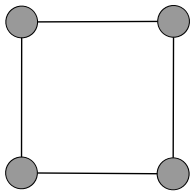
$$|F_2| = \frac{1}{6}\mathrm{tr}(A^3)$$

$F3$ 

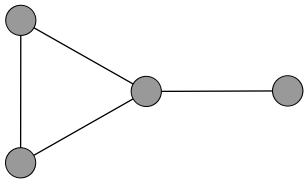
$$|F_3| = \sum_{(i,j)\in E} (k_i - 1)(k_j - 1) - 3|F_2|$$

$F4$ 

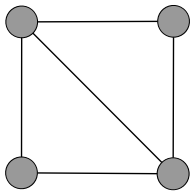
$$|F_4| = \frac{1}{6}\sum_i k_i(k_i - 1)(k_i - 2)$$

$F5$ 

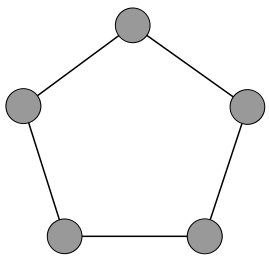
$$|F_5| = \frac{1}{8}(\mathrm{tr}(A^4) - 4|F_1| - 2m)$$

$F6$ 

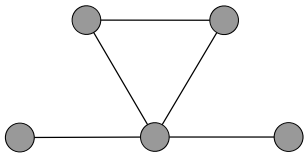
$$|F_6| = \sum_{k_i>2} t_i(k_i - 2)$$

$F7$ 

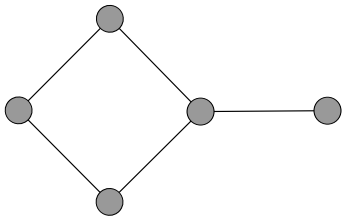
$$|F_7| = \frac{1}{4}\sum_{i,j}\Big((A^2)_{ij}A_{ij}\Big)\Big((A^2)_{ij}\cdot A_{ij} - 1\Big)$$

$F8$ 

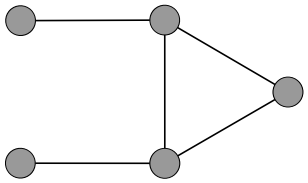
$$|F_8| = \frac{1}{10}(\mathrm{tr}(A^5) - 30|F_2| - 10|F_6|)$$

F9 

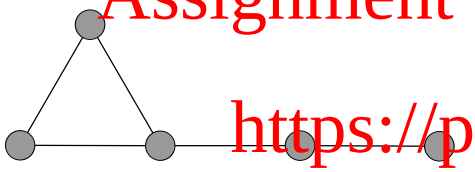
$$|F_9| = \frac{1}{2}\sum_{k_i \geq 4} t_i(k_i - 2)(k_i - 3)$$

F10 

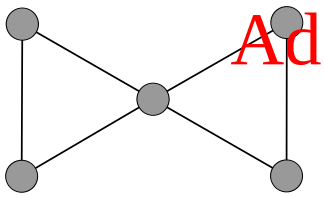
$$|F_{10}| = \frac{1}{2}\sum_i (k_i - 2) \times \sum_{i,j} \binom{(A^2)_{ij}}{2} - 2|F_7|$$

F11 

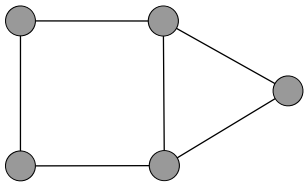
$$|F_{11}| = \sum_{(i,j)\in E} (A^2)_{ij}(k_i - 2)(k_j - 2) - 2|F_7|$$

F12 

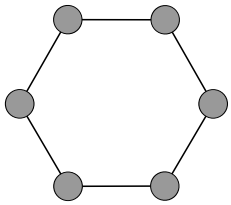
$$|F_{12}| = \sum_i t_i \left(\sum_{i \neq j} (A^2)_{ij} \right) - 6|F_2| - 2|F_6| - 4|F_7|$$

F13 

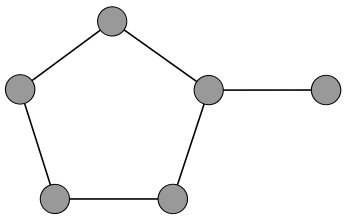
$$|F_{13}| = \frac{1}{2}\sum_i t_i(t_i - 1) - 2|F_7|$$

F14 

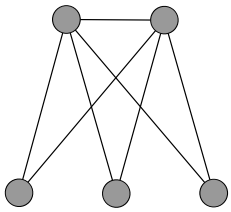
$$|F_{14}| = \sum_{(i,j)\in E} (A^3)_{ij}(A^2)_{ij} - 9|F_2| - 2|F_6| - 4|F_7|$$

F15 

$$|F_{15}| = \frac{1}{12}(\mathrm{tr}(A^6) - 2m - 12|F_1| - 24|F_2| - 6|F_3| \\ -12|F_4| - 48|F_5| - 36|F_7| - 12|F_{10}| - 24|F_{13}|)$$

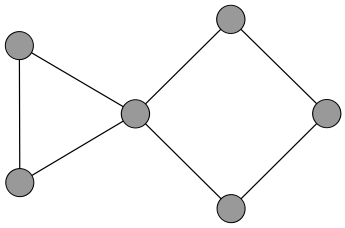
F16 

$$|F_{16}| = \frac{1}{2}\sum_i (k_i - 2)B_i - 2|F_{14}| \text{ where}$$
$$B_i = (A^5)_{ii} - 20t_i - 8t_i(k_i - 2) - 2\sum_{(i,j)\in E} [(A^2)_{ij}(k_j - 1) + t_j]$$



$F17$

$$|F_{17}| = \sum_{(i,j)\in E} \binom{(A^2)_{ij}}{3}$$



$F18$

$$|F_{18}| = \sum_i t_i \cdot \sum_{i\neq j} \binom{(A^2)_{ij}}{2} - 6|F_7| - 2|F_{14}| - 6|F_{17}|$$

13.3 Network motifs

We can use the techniques we have developed in this chapter to count the number of occurrences of a given fragment in a real-world network. Certain fragments arise inevitably through network connectivity. It is possible that the frequency with which they appear is similar to equivalent random networks. In this case, we cannot use the abundance of a fragment to explain any evolutionary mechanism giving rise to the structure of that network. However, if a given fragment appears more frequently than expected we can infer that there is some structural or functional reasons for the over expression. This is precisely the concept of a **network motif**. A subgraph is considered a network motif if the probability P of it appearing in a random network an equal or greater number of times than in the real-world network is lower than a certain cut-off value, which is generally taken to be $P_c = 0.01$.

In order to quantify the statistical significance of a given motif we use the Z-score which, for a given subgraph i, is defined as

$$Z_i = \frac{N_i^{real} - \langle N_i^{random} \rangle}{\sigma_i^{random}}, \tag{13.19}$$

where N_i^{real} is the number of times the subgraph i appears in the real network, $\langle N_i^{random} \rangle$ and σ_i^{random} are the average and standard deviation of the number of times that i appears in an ensemble of random networks, respectively. Similarly, the relative abundance of a given fragment can be estimated using the statistic

$$\alpha_i = \frac{N_i^{real} - \langle N_i^{random} \rangle}{N_i^{real} + \langle N_i^{random} \rangle}. \tag{13.20}$$

13.3.1 Motifs in directed networks

Motifs in directed networks are simply directed subgraphs which appear more frequently in the real-world network than in its random counterpart. The situation is more complex because the number of motifs with the same number of nodes is significantly larger than for the undirected networks (for instance, there are 7 directed triangles versus only one undirected); and in general there are no analytical tools for counting such directed subgraphs. However, there are several computational approaches that

allow the calculation of the number of small directed subgraphs and directed motifs in networks. In Figure 13.8 we illustrate some examples of directed triangles found in real-world networks as motifs.

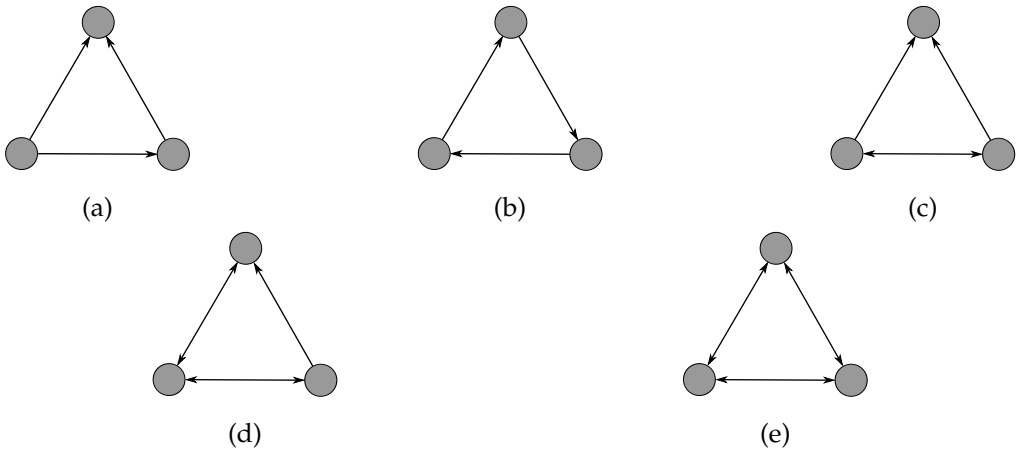


Figure 13.8: Motifs in directed networks (a) Feed-forward loop (neurons) (b) Three-node feedback loop (Electronic circuits) (c) Up-linked mutual dyad (d) Feedback with mutual dyads (e) Fully connected triad. Dyads are typical in the WWW.

A characteristic feature of network motifs is that they are network-specific. That is, what is a motif for one is not necessarily a motif for another. However, a family of networks can be identified if they share the same series of motifs. One can characterise it is by generating a vector whose ith entry gives the importance of the ith motif with respect to the other motifs in the network. The resulting component of the **significance profile** vector is given by

$$SP_i = \frac{Z_i}{\sum_j Z_j^2}. \tag{13.21}$$

13.3.2 Motifs in undirected networks

In Figure 13.9 we illustrate the relative abundance of 17 of the small subgraphs we've discussed for six complex networks representing different systems in the real-world. The average is taken over random networks whose nodes have the same degrees as the real ones. It can be seen that there are a few fragments which are over-represented in some networks while other fragments are under-represented. Fragments which appear less frequently in a real-world network than is expected in an analogous random one are called **anti-motifs**.

Problem 13.3

- The connected component of the protein–protein interaction network of yeast has 2224 nodes and 6609 links. It has been found computationally that the number of triangles in that network is 3530. Determine the relative abundance of this fragment in order to see whether it is a motif in this network.

  We use the formula

$$\alpha_i = \frac{t_i^{real} - \langle t_i^{random} \rangle}{t_i^{real} + \langle t_i^{random} \rangle},$$

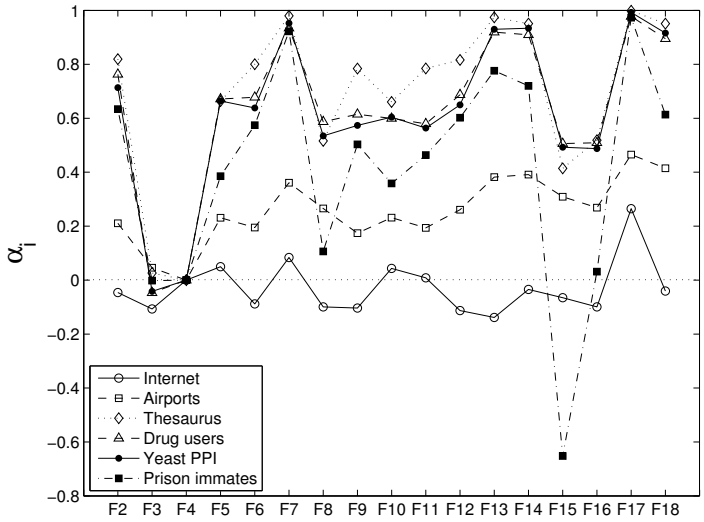


Figure 13.9: Motifs and anti-motifs in undirected networks

where we know that $t_i^{real} = 3530$. We have to estimate $\langle t_i^{random} \rangle$. Let us consider Erdös-Rényi random networks with 2224 nodes and 6609 links for which

$$p = \frac{2m}{n(n-1)} = 0.00267.$$

We also know that for large n, $\lambda_1 \to pn$ and all the other eigenvalues are negligible so we use the approximation

$$\langle t_i^{random} \rangle = \frac{1}{6} \sum_{j=1}^{n} \lambda_j^3 \approx \frac{\lambda_1^3}{6} = \frac{(np)^3}{6}.$$

Thus, $\langle t_i^{random} \rangle \approx 35$. This estimate is very good indeed. For instance, the average number of triangles in 100 realisations of an ER network is $\langle t_i^{random} \rangle = 35.4 \pm 6.1$. Using the value of $\langle t_i^{random} \rangle \approx 35$ we obtain $\alpha_i = 0.98$, which is very close to one. We conclude that the number of triangles in the yeast PPI is significantly larger than expected by chance and we can consider it as a network motif.

Further Reading

Alon, N., Yuster, R. and Zwick, U., Finding and counting given length cycles, Algorithmica **17**:209–223, 1997.

Milo, R. et al., Network motifs: Simple building blocks of complex networks, Science **298**:824–827, 2002.

Milo, R. et al., Superfamilies of evolved and designed networks, Science **303**:1538–1542, 2004.

Chapter 14: Classical Node Centrality

In this chapter: The concept of node centrality is motivated and introduced. Some properties of the degree of a node are analysed along with extensions to consider non-nearest neighbours. Two centralities based on shortest paths on the network are defined—the closeness and betweenness centrality—and differences between them are described. We finish this chapter with a few problems to illustrate how to find analytical expression for these centralities in certain classes of networks.

14.1 Motivation

The notion of centrality of a node first arose in the context of social sciences and is used in the determination of the most "important" nodes in a network. There are a number of characteristics, not necessarily correlated, which can be used in determining the importance of a node. These include its ability to communicate directly with other nodes; its closeness to many other nodes; and its indispensability to act as a communicator between different parts of a network.

Considering each of these characteristics in turn leads to different centrality measures. In this chapter we study such measures and illustrate the different qualities of a network that they can highlight.

14.2 Degree centrality

The degree centrality simply corresponds to degree and clearly measures the ability of a node to communicate directly with others. As we have seen, the degree of node i in a simple network G is defined using its adjacency matrix, A, as

$$k_i = \sum_{j=1}^{n} a_{ij} = (\mathbf{e}^T A)_i = (A\mathbf{e})_i. \tag{14.1}$$

So with degree centrality, i is more central than j if $k_i > k_j$. In a directed network, where in-degree and out-degree can be different, we can utilise degree to get two centrality measures, namely,

$$k_i^{in} = \sum_{i=1}^{n} a_{ji} = (\mathbf{e}^T A)_i, \quad k_i^{out} = \sum_{j=1}^{n} a_{ij} = (A\mathbf{e})_i. \tag{14.2}$$

The following are some elementary facts about the degree centrality. You are invited to prove these yourself.

1. $k_i = (A^2)_{ii}$.
2. $\sum_{i=1}^{n} k_i = 2m$, where m is the number of links.
3. $\sum_{i=1}^{n} k_i^{in} = \sum_{i=1}^{n} k_i^{out} = m$, where m is the number of links.

Examples 14.1

(i) Let us consider the network illustrated in Figure 14.1

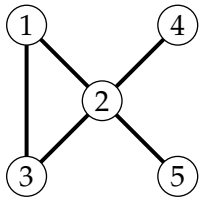


Figure 14.1: A simple labelled network

Since the adjacency matrix of the network is

$$A = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix}$$

the node degree vector is

$$\mathbf{k} = A\mathbf{e} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \\ 2 \\ 1 \\ 1 \end{bmatrix}.$$

That is, the degrees of the nodes are: $k(1) = k(3) = 2$, $k(2) = 4$, $k(4) = k(5) = 1$, indicating that the most central node is 2.

(ii) Let us consider the network displayed in Figure 14.2 together with its adjacency matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

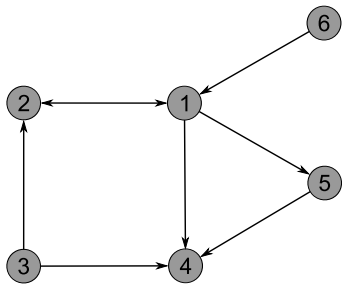


Figure 14.2: A labelled directed network

The in- and out-degree vectors are then obtained as follows:

$$\mathbf{k}^{in} = (\mathbf{e}^T A)^T = A^T \mathbf{e} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 3 \\ 1 \\ 0 \end{bmatrix},$$

$$\mathbf{k}^{out} = A\mathbf{e} = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix}.$$

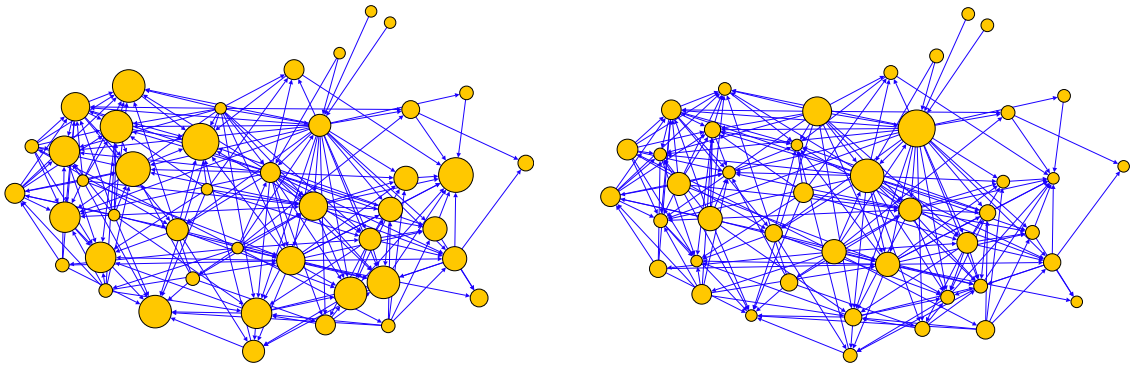
Nodes 3 and 6 are known as **sources** because their in-degrees are equal to zero but not their out-degrees. Node 4 is a **sink** because its out-degree is zero but not its in-degree. If both, the in- and out-degree are zero for a node, the node is isolated.

The most central node in sending information to its nearest neighbours is node 1 and in receiving information is node 4.

(iii) Let us now consider a real-world network. It corresponds to the food web of St Martin island in the Caribbean, in which nodes represent species and food sources and the directed links indicate what eats what in the ecosystem. Here we represent the networks in Figure 14.3 by drawing the nodes as circles with radius proportional to the corresponding in-degree in (a) and out-degree in (b).

The in and out-degree vectors are calculated in exactly the same way as in the last example. In this case, every node has a label which corresponds to the identity of the species in question. In analysing this network according to the in- and out-degree we can point out the following observations which are of relevance for the functioning of this ecosystem.

- Nodes with high out-degree are predators with a large variety of prey. Examples include the lizards *Anolis gingivinus* (the Anguilla Bank Anole) and *Anolis pogus*; and the birds the pearly-eyed thrasher and the yellow warbler.



(a) (b)

Figure 14.3: Food webs in St Martin with nodes drawns as circles of radii proportional to (a) in-degree (b) out-degree

- High in-degree nodes represent species and organic matter which are eaten by many others in this ecosystem, such as leaves, detritus and insects such as aphids.
- In general, top predators are not predated by other species, thus having significantly higher out-degree than in-degree.
- The sources, with zero in-degree are all birds: the pearly-eye thrasher, yellow warbler, kestrel and grey kingbird.
- Highly predated species are not usually prolific predators, thus they have high in-degree but low out-degree.
- The sinks are all associated with plants or detritus.

14.3 Closeness centrality

The closeness centrality of a node characterizes how close this node is from the rest of the nodes. This closeness is measured in terms of the shortest path distance. The closeness of the node i in an undirected network G is defined as

$$CC(i) = \frac{n-1}{s(i)}, \tag{14.3}$$

where the distance-sum $s(i)$ is calculated from the shortest path distances $d(i,j)$ as

$$s(i) = \sum_{j \in V(G)} d(i,j). \tag{14.4}$$

In a directed network a node has in- and out-closeness centrality. The first corresponds to how close this node is to nodes it is receiving information from. The out-closeness centrality indicates how close the node is from those it is sending information to. Recall that in directed networks the shortest path is a pseudo-distance due to a possible lack of symmetry.

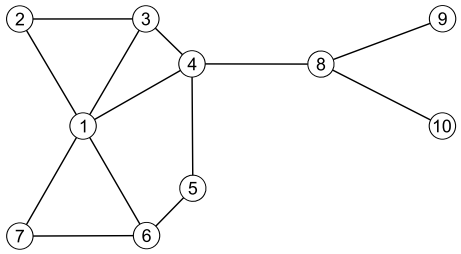


Figure 14.4: A network where closeness centrality does not match degree centrality

Examples 14.2

(i) Consider the network illustrated in Figure 14.4.

We start by constructing the distance matrix of this network, which is given by

$$D = \begin{bmatrix} 0 & 1 & 1 & 1 & 2 & 1 & 1 & 2 & 3 & 3 \\ 1 & 0 & 1 & 2 & 3 & 2 & 2 & 3 & 4 & 4 \\ 1 & 1 & 0 & 1 & 2 & 2 & 2 & 2 & 3 & 3 \\ 1 & 2 & 1 & 0 & 1 & 2 & 2 & 1 & 2 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1 & 2 & 2 & 3 & 3 \\ 1 & 2 & 2 & 2 & 1 & 0 & 1 & 3 & 4 & 4 \\ 1 & 2 & 2 & 2 & 1 & 1 & 0 & 3 & 4 & 4 \\ 2 & 3 & 2 & 1 & 2 & 3 & 3 & 0 & 1 & 1 \\ 3 & 4 & 3 & 2 & 3 & 4 & 4 & 1 & 0 & 2 \\ 3 & 4 & 3 & 2 & 3 & 4 & 4 & 1 & 2 & 0 \end{bmatrix}$$

The vector of distance-sum of each node is then

$$\mathbf{s} = D\mathbf{e} = (\mathbf{e}^T D)^T = [\ 15 \quad 22 \quad 17 \quad 14 \quad 19 \quad 20 \quad 21 \quad 18 \quad 26 \quad 26\ ]^T$$

And we use (14.3) to measure the closeness centrality of each node. For instance for node 1

$$CC(1) = \frac{9}{15} = 0.6.$$

The full vector of closeness centralities is

$$\mathbf{CC} = [\ 0.600 \quad 0.409 \quad 0.529 \quad \mathbf{0.643} \quad 0.474 \quad 0.450 \quad 0.428 \quad 0.500 \quad 0.346 \quad 0.346\ ]^T,$$

indicating that the most central node is the node 4. Notice that in this case the degree centrality identifies another node (namely 1) as the most important whereas in Figure 14.1 node 2 has both the highest degree and closeness centralities.

(ii) We consider here the air transportation network of the USA, where the nodes represent the airports in the USA and the links represent the existence of at least one flight connecting the two airports. In Figure 14.5 we illustrate this network in which the nodes are represented as circles with radii proportional to the closeness centrality.

The most central airports according to the closeness centrality are given below.

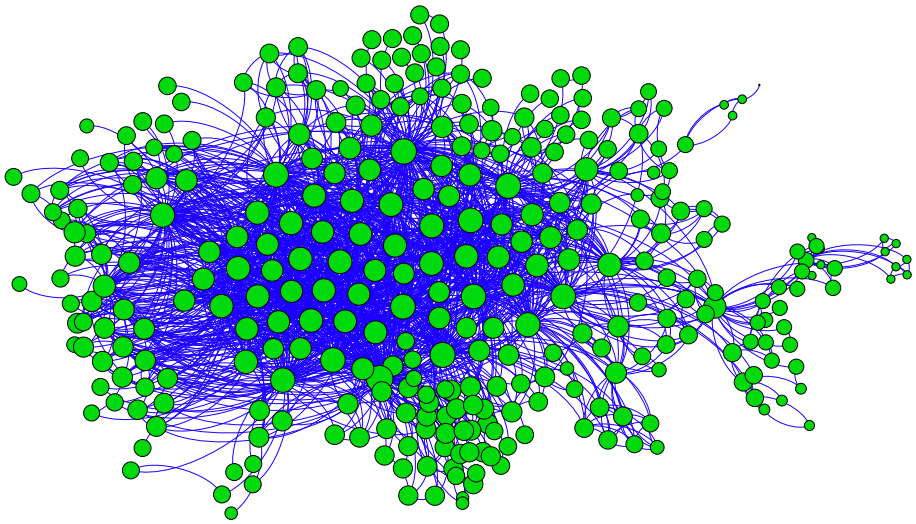


Figure 14.5: A representation of the USA air transportation network in 1997

Airport	Closeness centrality $\times 100$
Chicago O'Hare Intl	60.734
Dallas/Fort Worth Intl	55.444
Minneapolis-St Paul Intl	53.997
William B Hartsfield, Atlanta	53.560
San Francisco Intl	53.301
Lambert-St Louis Intl	52.875
Seattle-Tacoma Intl	52.623
Los Angeles Intl	52.456

The first four airports in this list (and the sixth) correspond to airports in the geographic centre area of continental USA. The other three are airports located on the west coast. The first group are important airports in connecting the East and West of the USA with an important traffic also between north and south of the continental USA. The second group represents airports with important connections between the main USA and Alaska, as well as overseas territories like Hawaii and other Pacific islands. The most highly ranked airports according to degree centrality are given below. Notice that the group of west coast airports is absent.

Airport	Degree centrality
Chicago O'Hare Intl	139
Dallas/Fort Worth Intl	118
William B Hartsfield, Atlanta	101
Pittsburgh Intl	94
Lambert-St Louis Intl	94
Charlotte/Douglas Intl	87
Stapleton Intl	85
Minneapolis-St Paul Intl	78

Problem 14.1

- Let $CC(i)$ be the closeness centrality of the ith node in the path network P_n labelled $1-2-3-4-\cdots-(n-1)-n$.

  a) Find a general expression for the closeness centrality of the ith node in P_n in terms of i and n only.

  b) Simplify the expressions found in a) for the node(s) at the centre of the path P_n (for both odd and even values of n).

  c) Show that the closeness centrality of these central nodes is the largest in a path P_n.

The solution can be arrived at as follows.

(a) We start by considering the sum of all the distance from one node to the rest of the nodes in the path by using the labelling $1-2-3-4-\cdots-(n-1)-n$,

Node	$\sum_{j\neq i} d_{ij}$
1	$1+2+\cdots+n-1$
2	$1+1+2+\cdots+n-2$
3	$2+1+1+2+\cdots+n-3$
$\vdots$	$\vdots$
i	$(i-1)+(i-2)+\cdots+2+1+1+2+\cdots+n-i$

It is important to provide here that for each node the sum of the distances corresponds to a right sum, i.e., the sum of the distances of all nodes to the right of the node i and a left sum, i.e., the sum of the distances of all nodes located to the left of i, $(i-1)+(i-2)+\cdots+2+1$. These two sums are given, respectively by

$$1+2+\cdots+n-i=\frac{(n-i)(n-i+1)}{2}, \tag{14.5}$$

$$(i-1)+(i-2)+\cdots+2+1=\frac{(i-1)i}{2}. \tag{14.6}$$

By substituting into the formula (14.3) we obtain

$$CC(i)=\frac{n-1}{\frac{(i-1)i}{2}+\frac{(n-i)(n-i+1)}{2}}, \tag{14.7}$$

which can be written as

$$CC(i)=\frac{2(n-1)}{(i-1)i+(n-i)(n-i+1)}. \tag{14.8}$$

(b) For a path with an odd number of nodes the central node is $i=\frac{n+1}{2}$. By substitution into (14.8) we obtain

$$CC\left(\frac{n+1}{2}\right)=\frac{2(n-1)}{\left(\frac{n+1}{2}-1\right)\frac{n+1}{2}+\left(n-\frac{n+1}{2}\right)\left(n-\frac{n+1}{2}+1\right)}, \tag{14.9}$$

which reduces to

$$CC\left(\frac{n+1}{2}\right) = \frac{4}{(n+1)}. \tag{14.10}$$

For a path with an even number of nodes the central nodes are $i = \frac{n}{2}$ and $i = \frac{n}{2}+1$. Now,

$$CC\left(\frac{n}{2}\right) = \frac{2(n-1)}{\left(\frac{n}{2}-1\right)\frac{n}{2} + \left(n-\frac{n}{2}\right)\left(n-\frac{n}{2}+1\right)}, \tag{14.11}$$

and in this case

$$CC\left(\frac{n}{2}\right) = \frac{4(n-1)}{n^2}. \tag{14.12}$$

We recover the same value when $i = \frac{n+1}{2}$.

(c) Simply consider

$$CC(i+1) - CC(i) = \frac{2(n-1)}{(i+1)i + (n-i-1)(n-i)} - \frac{2(n-1)}{i(i-1) + (n-i)(n-i+1)}. \tag{14.13}$$

Putting the right-hand side over a common denominator gives the numerator

$$4(n-1)(n-2i), \tag{14.14}$$

which is positive if $i < n/2$ and negative if $i > n/2$, so $CC(i)$ reaches its maximum value in the centre of the path.

14.4 Betweenness centrality

The betweenness centrality characterizes how important a node is in the communication between other pairs of nodes. That is, the betweenness of a node accounts for the proportion of information that passes through a given node in communications between other pairs of nodes in the network.

As for the closeness centrality, betweenness assumes that the information travels from one node to another through the shortest paths connecting those nodes. The betweenness of the node i in an undirected network G is defined as

$$BC(i) = \sum_j \sum_k \frac{\rho(j,i,k)}{\rho(j,k)}, \; i \neq j \neq k, \tag{14.15}$$

where $\rho(j,k)$ is the number of shortest paths connecting the node j to the node k, and $\rho(j,i,k)$ is the number of these shortest paths that pass through node i in the network.

If the network is directed, the term $\rho(j,i,k)$ refers to the number of directed paths from the node j to the node k that pass through the node i, and $\rho(j,k)$ to the total number of directed paths from the node j to the node k.

Examples 14.3

(i) We consider again the network used in Figure 14.4 and we explain how to obtain the betweenness centrality for the node labelled as 1. For this, we construct the following table in which we give the number of shortest paths from any pair of nodes that pass through the node 1, $\rho(j,1,k)$. We also report the total number of shortest paths from these pairs of nodes $\rho(j,k)$.

(j,k)	$\rho(j,1,k)$	$\rho(j,k)$	$\frac{\rho(j,i,k)}{\rho(j,k)}$
2,4	1	2	1/2
2,5	2	3	2/3
2,6	1	1	1
2,7	1	1	1
2,8	1	2	1/2
2,9	1	2	1/2
2,10	1	2	1/2
3,6	1	1	1
3,7	1	1	1
4,6	1	2	1/2
4,7	1	1	1
6,8	1	2	1/2
6,9	1	2	1/2
6,10	1	2	1/2
7,8	1	1	1
7,9	1	1	1
7,10	1	1	1
			12.667

The betweenness centrality of the node 1 is simply the total sum of the terms in the last column of the table,

$$BC(1) = \sum_j \sum_k \frac{\rho(j,1,k)}{\rho(j,k)} = 12.667.$$

Using a similar procedure, we obtain the betweenness centrality for each node:

$$\mathbf{BC} = [\ 12.667 \quad 0.000 \quad 2.333 \quad 20.167 \quad 2.000 \quad 1.833 \quad 0.000 \quad 15.000 \quad 0.000 \quad 0.000\ ]^T,$$

which indicates that the node 4 is the most central one, i.e., it is the most important in allowing communication between other pairs of nodes.

(ii) In Figure 14.6 we illustrate the urban street network of the central part of Cordoba, Spain. The most central nodes according to the betweenness correspond to those street intersections which surround the central part of the city and connect it with the periphery.

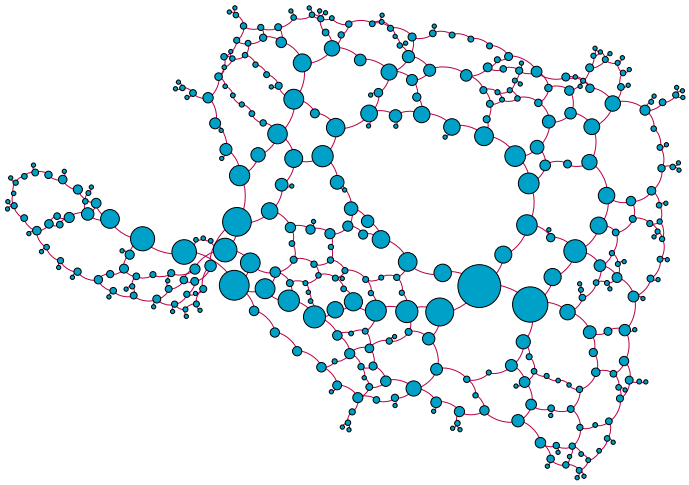


Figure 14.6: The street network of Cordoba with nodes of radii proportional to their betweenness centrality

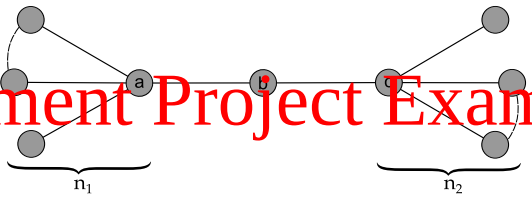


Figure 14.7: A network formed by joining two star networks together. Dashed lines indicate the existence of other equivalent nodes

Problem 14.2

- Let G be a tree with $n = n_1 + n_2 + 1$ and the structure displayed in Figure 14.7. State conditions for the nodes labelled a, b and c to have the largest value of betweenness centrality.

  We start by considering the betweenness centrality of node a. Let us designate by V_1 and V_2 the two branches of the graph, the first containing n_1 and the second n_2 nodes.

Fact 1 Because the network is a tree, the number of shortest paths from p to q that pass through node k, $\rho(p,k,q)$, is the same as the number of shortest paths from p to q, $\rho(p,q)$. That is, $\rho(p,k,q) = \rho(p,q)$.

Fact 2 There are n_1 nodes in the branch V_1. Let us denote by i any node in this branch which is not a and by j any node in V_2 which is not c. There are $n_1 - 1$ shortest paths from nodes i to node b. That is,

$$\rho(i,a,b) = n_1 - 1. \tag{14.16}$$

Fact 3 We can easily calculate the number of paths from a node i to any node in the branch V_2 which go through node a. Because there are $n_1 - 1$ nodes of type i and n_2 nodes in the branch V_2 we have

$$\rho(i,a,V_2) = (n_1 - 1)n_2.$$

Fact 4 Any path going from a node denoted by i to another such node passes through the node a. Because there are $n_1 - 1$ nodes of the type i we have that the number of these paths is given

by

$$\rho(i,a,i) = \binom{n_1 - 1}{2} = \frac{(n_1-1)(n_1-2)}{2}. \tag{14.17}$$

Therefore the total number of paths containing the node a, and consequently its betweenness centrality, is

$$\begin{aligned} BC(a) &= 2(n_1-1) + (n_1-1)(n_2-1) + \frac{(n_1-1)(n_1-2)}{2} \\ &= \frac{(n_1-1)(2n_2+n_1)}{2}. \end{aligned}$$

In a similar way we obtain

$$BC(c) = \frac{(n_2-1)(2n_1+n_2)}{2}. \tag{14.18}$$

To calculate the betweenness centrality for node b we observe that every path from the n_1 nodes in branch V_1 to the n_2 nodes in branch V_2 passes through node b. Consequently,

$$BC(b) = n_1 n_2. \tag{14.19}$$

Obviously, all nodes apart from a, b and c have zero betweenness centrality. We consider in turn the conditions for the three remaining nodes to be central.

In order for node a to have the maximum BC, the following conditions are necessary:

$$BC(a) > BC(c) \quad \text{and} \quad BC(a) > BC(b).$$

First,

$$BC(a) > BC(c) \Rightarrow \frac{(n_1-1)(2n_2+n_1)}{2} > \frac{(n_2-1)(2n_1+n_2)}{2} \Rightarrow (n_1^2+n_1) > (n_2^2+n_2) \Rightarrow n_1 > n_2,$$

and

$$BC(a) > BC(b) \Rightarrow \frac{(n_1-1)(2n_2+n_1)}{2} > n_1 n_2 \Rightarrow \frac{n_1(n_1-1)}{2} > n_2.$$

The second condition is fulfilled only if $n_1 \geq n_2$ and $n_1 > 3$. By combining both conditions we conclude that $BC(a)$ is the absolute maximum in the graph only in the cases when $n_1 > n_2$ and $n_1 > 3$.

By symmetry, $BC(c)$ is the absolute maximum if $n_2 > n_1$ and $n_2 > 3$.

Finally, for $BC(b)$ to be the absolute maximum we need

$$\begin{array}{c} BC(b) > BC(a) \\ n_1 n_2 > \dfrac{(n_1-1)(2n_2+n_1)}{2} \\ n_2 > \dfrac{n_1(n_1-1)}{2} \end{array} \quad \text{and} \quad \begin{array}{c} BC(b) > BC(c) \\ n_1 n_2 > \dfrac{(n_2-1)(2n_1+n_2)}{2} \\ n_1 > \dfrac{n_2(n_2-1)}{2} \end{array}$$

The two conditions are fulfilled simultaneously only if $n_1 = n_2 < 3$. That is, $BC(b)$ is the absolute maximum **only** when $n_1 = n_2 = 1$ or when $n_1 = n_2 = 2$, which correspond to P_2 and P_4, respectively (check this by yourself).

Further Reading

Borgatti, S.P., Centrality and network flow, Social Networks **27**:55–71, 2005.

Borgatti, S.P. and Everett, M.G., A graph-theoretic perspective on centrality, Social Networks **28**:466–484, 2006.

Brandes, U. and Erlebach, T. (Eds.), *Network Analysis: Methodological Foundations*, Springer, 2005, Chapters 3–5.

Estrada, E., *The Structure of Complex Networks: Theory and Applications*, Oxford University Press, 2011, Chapter 7.

Wasserman, S. and Faust, K., *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994, Chapter 5.

Chapter 15: Spectral Node Centrality

In this chapter: The necessity for considering the influence of a node beyond its nearest neighbours is motivated. We introduce centrality measures that account for long-range effects of a node, such as the Katz index, eigenvector centrality, the PageRank index and subgraph centrality. A common characteristic of these centrality measures is that they can be expressed in terms of spectral properties of the networks.

15.1 Motivation

Suppose we use a network to model a contagious disease amongst a population. Nodes represent individuals and edges represent potential routes of infection between these individuals. We illustrate a simple example in Figure 15.1. We focus on the nodes labelled 1 and 4 and ask which of them has the higher risk of contagion. Node 1 can be infected from nodes 2 and 3, while node 4 can be infected from 5 and 6. From this point of view it looks like both nodes are at the same level of risk. However, while 2 and 3 cannot be infected by any other node, nodes 5 and 6 can be infected from nodes 7 and 8, respectively. Thus, we can intuitively think that 4 is at a greater risk than 1 as a consequence of the chain of transmission of the disease. Local centrality measures like node degree do not account for a centrality that goes beyond the first nearest neighbours, so we need other kinds of measures to account for such effects. In this chapter we study these measures and illustrate the different qualities of a network that they can highlight.

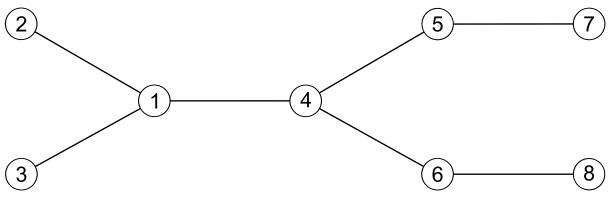


Figure 15.1: A simple network. Nodes 1 and 4 are rivals for title of most central

15.2 Katz centrality

The degree of the node i counts the number of walks of length one from i to every other node of the network. That is, $k_i = (A\mathbf{e})_i$. In 1953, Katz extended this idea to count not only the walks of length one, but those of any length starting at node i. Intuitively, we can reason that the closest neighbours have more influence over node i than more distant ones. Thus when combining walks of all lengths, one can introduce an attenuation factor so that more weight is given to shorter walks than to longer ones. This is precisely what Katz did and the Katz index is given by

$$K_i = \left[(\alpha^0 A^0 + \alpha A + \alpha^2 A^2 + \cdots + \alpha^k A^k + \cdots)\mathbf{e}\right]_i = \left[\sum_{k=0}^{\infty} (\alpha^k A^k)\mathbf{e}\right]_i . \tag{15.1}$$

The series in (15.1) is related to the resolvent function $(zI - A)^{-1}$. In particular, we saw in Example 12.4(iii) that the series converges so long as $\alpha < \rho(A)$ in which case

$$K_i = \left[(I - \alpha A)^{-1}\mathbf{e}\right]_i . \tag{15.2}$$

The Katz index can be expressed in terms of the eigenvalues and eigenvectors of the adjacency matrix. From the spectral decomposition $A = QDQ^T$ (see Chapter ??)

$$K_i = \sum_l \sum_j \mathbf{q}_j(i)\mathbf{q}_j(l)\,\frac{1}{1 - \alpha\lambda_j} . \tag{15.3}$$

When deriving his index, Katz ignored the contribution from $A^0 = I$ and instead used

$$\overline{K}_i = \left[\left((I - \alpha A)^{-1} - I\right)\mathbf{e}\right]_i . \tag{15.4}$$

While the values given by (15.2) and (15.4) are different, the rankings are exactly the same. We will generally use (15.2) because of the nice mathematical properties of the resolvent.

Example 15.1

- We consider the network illustrated in Figure 15.1. The principal eigenvalue of the adjacency matrix for this network is $\lambda_1 = 2.1010$. With $\alpha = 0.3$ we obtain the vector of Katz centralities

$$K = \begin{bmatrix} 3.242 & 1.972 & 1.972 & \mathbf{3.524} & 2.591 & 2.591 & 1.777 & 1.777 \end{bmatrix}^T .$$

  Node 4 has the highest Katz index, followed by node 1 which accords with our intuition on the level of risk of each of these nodes in the network.

15.2.1 Katz centrality in directed networks

In directed networks we should consider the Katz centrality of a node in terms of the number of links going in and out from a node. This can be done by considering the indices

$$K_i^{out} = \left[(I - \alpha A)^{-1}\mathbf{e}\right]_i , \quad K_i^{in} = \left[\mathbf{e}^T(I - \alpha A)^{-1}\right]_i .$$

The second index can be considered as a measure of the "prestige" of a node because it accounts for the importance that a node inherits from those that point to it.

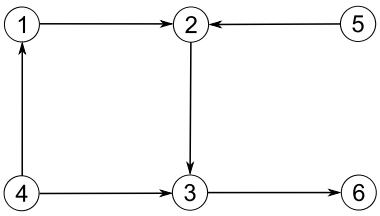


Figure 15.2: A directed network. In- and out- centrality measures vary

Example 15.2

- We measure the Katz indices of the nodes in the network illustrated in Figure 15.2.

  Using $\alpha = 0.5$ we obtain the Katz indices

  $$K^{in} = \begin{bmatrix} 1.50 & 2.25 & 2.62 & 1.00 & 1.00 & 2.31 \end{bmatrix}, \quad K^{out} = \begin{bmatrix} 1.88 & 1.75 & 1.50 & 2.69 & 1.88 & 1.00 \end{bmatrix}^T.$$

  Notice that nodes 2 and 3 are each pointed to by two nodes. However, node 3 is more central because it is pointed to by nodes with greater centrality than those pointing to 2. In fact, node 6 is more central than node 2 because the only node pointing to it is the most important one in the network. On the other hand, out-Katz identifies node 4 as the most central one. It is the only node having out-degree of two.

15.3 Eigenvector centrality

Let us consider the following modification of the Katz index:

$$\boldsymbol{\nu} = \left(\sum_{k=1}^{\infty} \alpha^{k-1} A^k \right) \mathbf{e} = \left(\sum_{k=1}^{\infty} \alpha^{k-1} \sum_{j=1}^{n} \mathbf{q}_j \mathbf{q}_j^T \lambda_j^k \right) \mathbf{e} = \left(\frac{1}{\alpha} \sum_{j=1}^{n} \sum_{k=1}^{\infty} (\alpha\lambda_j)^k \mathbf{q}_j \mathbf{q}_j^T \right) \mathbf{e} = \left(\sum_{j=1}^{n} \frac{\lambda_j}{1-\alpha\lambda_j} \mathbf{q}_j \mathbf{q}_j^T \right) \mathbf{e}.$$

Now, let the parameter α approach the inverse of the largest eigenvalue of the adjacency matrix from below, i.e., $\alpha \to 1/\lambda_1^-$. Then

$$\lim_{\alpha \to 1/\lambda_1^-} (1-\alpha\lambda_1)\boldsymbol{\nu} = \lim_{\alpha \to 1/\lambda_1^-} \left(\sum_{j=1}^{n} \frac{(1-\alpha\lambda_1)\lambda_j}{1-\alpha\lambda_j} \mathbf{q}_j \mathbf{q}_j^T \right) \mathbf{e} = \left(\lambda_1 \sum_{i=1}^{n} \mathbf{q}_1(i) \right) \mathbf{q}_1 = \gamma \mathbf{q}_1.$$

Thus the eigenvector associated with the largest eigenvalue of the adjacency matrix is a centrality measure conceptually similar to the Katz index. Accordingly, the eigenvector centrality of the node i is given by $\mathbf{q}(i)$, the ith component of the principal eigenvector $\mathbf{q}_1$ of A. Typically, we normalise $\mathbf{q}_1$ so that its Euclidean length is 1. By the Perron–Frobenius theorem we can choose $\mathbf{q}_1$ so that all of its components are nonnegative.

Examples 15.3

(i) The eigenvector centralities for the nodes of the network in Figure 15.1 are

$$\mathbf{q}_1 = \begin{bmatrix} 0.500 & 0.238 & 0.238 & \mathbf{0.574} & 0.354 & 0.354 & 0.168 & 0.168 \end{bmatrix}^T.$$

Here again node 4 is the one with the highest centrality, followed by node 1. Node 4 is connected to nodes which are higher in centrality than the nodes to which node 1 is connected to. High degree is not the only factor considered by this centrality measure. The most central nodes are generally connected to other highly central nodes.

(ii) Sometimes being connected to a few very important nodes make a node more central than being connected to many not so central ones. For instance, in Figure 15.3, node 4 is connected to only three other nodes, while 1 is connected to four. However, 4 is more central than 1 according to the eigenvector centrality because it is connected to two nodes with relatively high centrality while 1 is mainly connected to peripheral nodes. The transpose of the vector of centralities is

$$\begin{bmatrix} 0.408 & 0.167 & 0.167 & 0.5 & 0.408 & 0.167 & 0.167 & 0.167 & 0.408 & 0.167 & 0.167 & 0.167 & 0.167 \end{bmatrix}.$$

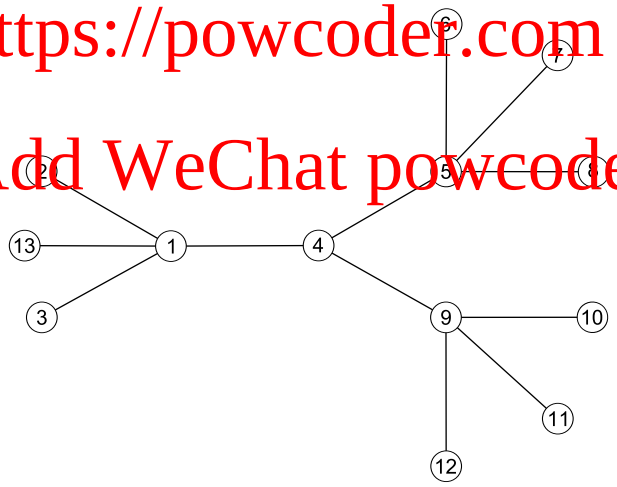


Figure 15.3: A network highlighting the difference between degree and eigenvector centrality

Problem 15.1

- Let G be a simple connected network with n nodes and adjacency matrix A with spectral decomposition QDQ^T. Let $N_k(i)$ be the number of walks of length k starting at node i. Let

$$\mathbf{s}_k(i) = \frac{N_k(i)}{\sum_{j=1}^{n} N_k(j)}$$

be the ith element of the vector $\mathbf{s}_k$. Show that if G is not bipartite then there is a scalar α such that as $k \to \infty$, $\mathbf{s}_k \to \alpha \mathbf{q}_1$ almost surely. That is, the vector $\mathbf{s}_k$ will tend to rank nodes identically to eigenvector centrality.

Since $A^k = QD^kQ^T$,

$$\mathbf{s}_k(i) = \frac{\mathbf{e}_i^T A^k \mathbf{e}}{\mathbf{e}^T A^k \mathbf{e}} = \frac{\mathbf{e}_i^T Q D^k Q^T \mathbf{e}}{\mathbf{e}^T Q D^k Q^T \mathbf{e}} = \frac{\mathbf{q}_i^T D^k \mathbf{r}}{\mathbf{r}^T D^k \mathbf{r}} = \frac{\mathbf{q}_i^T \overline{D}^k \mathbf{r}}{\mathbf{r}^T \overline{D}^k \mathbf{r}},$$

where $\mathbf{r} = Q^T\mathbf{e}$ and $\overline{D} = D/\lambda_1$.

Since G is connected and not bipartite, $|\lambda_1| > |\lambda_j|$ for all $j > 1$ so $\overline{D}^k \to \mathbf{e}_1\mathbf{e}_1^T$ as $k \to \infty$.[1] We have established that

$$\mathbf{s}_k(i) \to \frac{\mathbf{q}_i^T \mathbf{e}_1 \mathbf{e}_1^T \mathbf{r}}{\mathbf{r}^T \mathbf{e}_1 \mathbf{e}_1^T \mathbf{r}} = \alpha \mathbf{q}_i(1)$$

where $\alpha = 1/(\mathbf{e}_i^T\mathbf{r})$ and so

$$\lim_{k\to\infty} \mathbf{s}_k = \alpha \mathbf{q}_1,$$

as desired. Note that we require $\mathbf{e}_i^T\mathbf{r} \neq 0$ in this analysis, which is almost surely true for a network chosen at random.

15.3.1 Eigenvector centrality in directed networks

As with our other measures, we can define eigenvector centrality for directed networks. In this case we use the principal right and left eigenvectors of the adjacency matrix as the corresponding centrality vectors for the nodes in a directed network. If $A\mathbf{x} = \lambda_1\mathbf{x}$ and $A^T\mathbf{y} = \lambda_1\mathbf{y}$ then the elements of $\mathbf{x}$ and $\mathbf{y}$ give the right and left eigenvector centralities, respectively.

The right eigenvector centrality accounts for the importance of a node through the importance of nodes to which it points. It is an extension of the out-degree concept. On the other hand, the left eigenvector centrality accounts for the importance of a node by considering those nodes pointing towards a corresponding node and it is an extension of the in-degree centrality.

Example 15.4

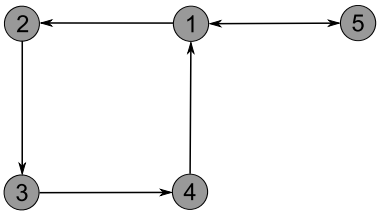


Figure 15.4: A directed network highlighting the difference between left and right eigenvector centrality

- The left and right eigenvector centralities of the network in Figure 15.4 are

$$\mathbf{x} = \begin{bmatrix} 0.592 & 0.288 & 0.366 & 0.465 & 0.465 \end{bmatrix}^T, \quad \mathbf{y} = \begin{bmatrix} 0.592 & 0.465 & 0.366 & 0.288 & 0.465 \end{bmatrix}^T.$$

  Notice the differences in the rankings of nodes 4 and 5. According to the right eigenvector both nodes ranked as the second most central. They both point to the most central node of the network

[1] Note that $\mathbf{e}_1\mathbf{e}_1^T$ is a matrix whose only nonzero element is a 1 in the top left hand corner.

according to this criterion, node 1. However, according to the left eigenvector, while node 5 is still the second most important, node 4 has been relegated to the least central one. Node 5 is pointed to by the most central node, but node 4 is pointed to only by a node with low centrality.

15.3.2 PageRank centrality

When we carry out a search for a particular term, a search engine is likely to return thousands or millions of related web pages. A good search engine needs to make sure that pages that are most likely to match the query are promoted to the front of this list and centrality measures are a key tool in this process.

By viewing the World Wide Web (WWW) as a giant directed network whose nodes are pages and whose edges are the hyperlinks between them, search engines can attempt to rank pages according to centrality. While this is only one of the factors involved nowadays, much of the initial success of Google has been credited to their use of their own centrality measure, which they dubbed PageRank.

PageRank is closely related to eigenvector centrality and it explicitly measures the importance of a web page via the importance of other web pages pointing to it. In simple terms, the PageRank of a page is the sum of the PageRank centralities of all pages pointing into it.

The first step in computing PageRank is to manipulate the adjacency matrix, A. The WWW is an extremely complex network and is known to be disconnected. In order to apply familiar analytic tools, such as the Perron–Frobenius theorem, we need to make adjustments to A. In practice, the simplest approach is to artificially rewire nodes which have no outbound links so that they are connected to all the other nodes in the network. That is, we replace A with a new matrix H defined so

$$H_{ij} = \begin{cases} a_{ij}, & k_i^{out} > 0, \\ 1 & k_i^{out} = 0. \end{cases} \tag{15.5}$$

PageRank can then be motivated by considering what would happen if an internet surfer moved around the WWW from page to page by picking out-links uniformly at random. The insight that the developers of Google had was that the surfer is more likely to visit pages that have been deemed important in that they have in-links from other important pages. Using the theory of **Markov chains** it can be shown that the relative frequency of page visits can be measured by the elements of the principal left eigenvector of the **stochastic matrix**

$$S = D^{-1}H,$$

where $D = \text{diag}(H\mathbf{e})$ is a diagonal matrix containing the out-degrees of the network with adjacency matrix H. Mathematically, PageRank is related to probability distributions so the vector of centralities is usually normalised to sum to 1.

Of course, computing this eigenvector with a network as big as the WWW (which has billions of nodes) is a challenge in itself. For reasons of expediency, an additional parameter α (not to be confused with the one previously used for the Katz index) is introduced and rather than working with S we work with

$$P = \alpha S + \frac{1-\alpha}{n}\mathbf{e}\mathbf{e}^T. \tag{15.6}$$

The parameter is motivated by the suggestion that every so often, instead of following an out-link, our surfer teleports randomly to another page somewhere on the internet, preventing the user from getting

stuck in a corner of the WWW. The value $\alpha = 0.85$ has been shown to work well in internet applications but there is no reason whatsoever to use this same parameter value when PageRank is used on general complex networks.

Example 15.5

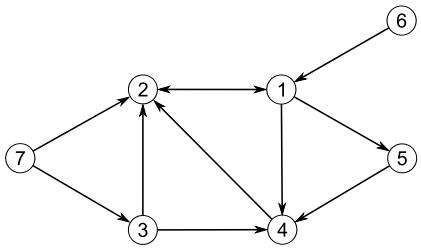


Figure 15.5: A directed network illustrating PageRank centrality

- The normalised PageRank of the network in Figure 15.5 is

$$\mathbf{PG} = \begin{bmatrix} 0.301 & 0.308 & 0.080 & 0.211 & 0.107 & 0.021 & 0.021 \end{bmatrix}^T$$

when $\alpha = 0.85$. Notice that node 1 has higher PageRank than node 4 due to its in-link from node 2. In this example, the rankings vary little as we change α.

15.4 Subgraph centrality

Katz centrality is computed from the entries of the matrix

$$K = \sum_{l=0}^{\infty} \alpha^l A^l.$$

We can easily generalise this idea and work with other weighted sums of the powers of the adjacency matrix, namely,

$$f(A) = \sum_{l=0}^{\infty} c_l A^l. \tag{15.7}$$

The coefficients c_l are expected to ensure that the series is convergent; they should give more weight to small powers of the adjacency matrix than to the larger ones; and they should produce positive numbers for all $i \in V$.

Notice that if the first of the three requirements hold then (15.7) defines a matrix function and we can use theory introduced in Chapter **??**. The diagonal entries, $f_i(A) = f(A)_{ii}$, are directly related to subgraphs in the network and the second requirement ensures that more weight is given to the smaller than to the bigger ones.

Example 15.6

- Let us examine (15.7) when we truncate the series at $l = 5$ and select $c_l = \frac{1}{l!}$ to find an expression for $f_i(A)$.

  Using information we collected in Chapter 13 on enumerating small subgraphs we know that

$$
\begin{aligned}
(A^2)_{ii} &= |F_1(i)|, && (15.8)\\
(A^3)_{ii} &= 2|F_2(i)|, && (15.9)\\
(A^4)_{ii} &= |F_1(i)| + |F_3(i)| + 2|F_4(i)| + 2|F_5(i)|, && (15.10)\\
(A^5)_{ii} &= 10|F_2(i)| + 2|F_6(i)| + 2|F_7(i)| + 4|F_8(i)| + 2|F_9(i)|. && (15.11)
\end{aligned}
$$

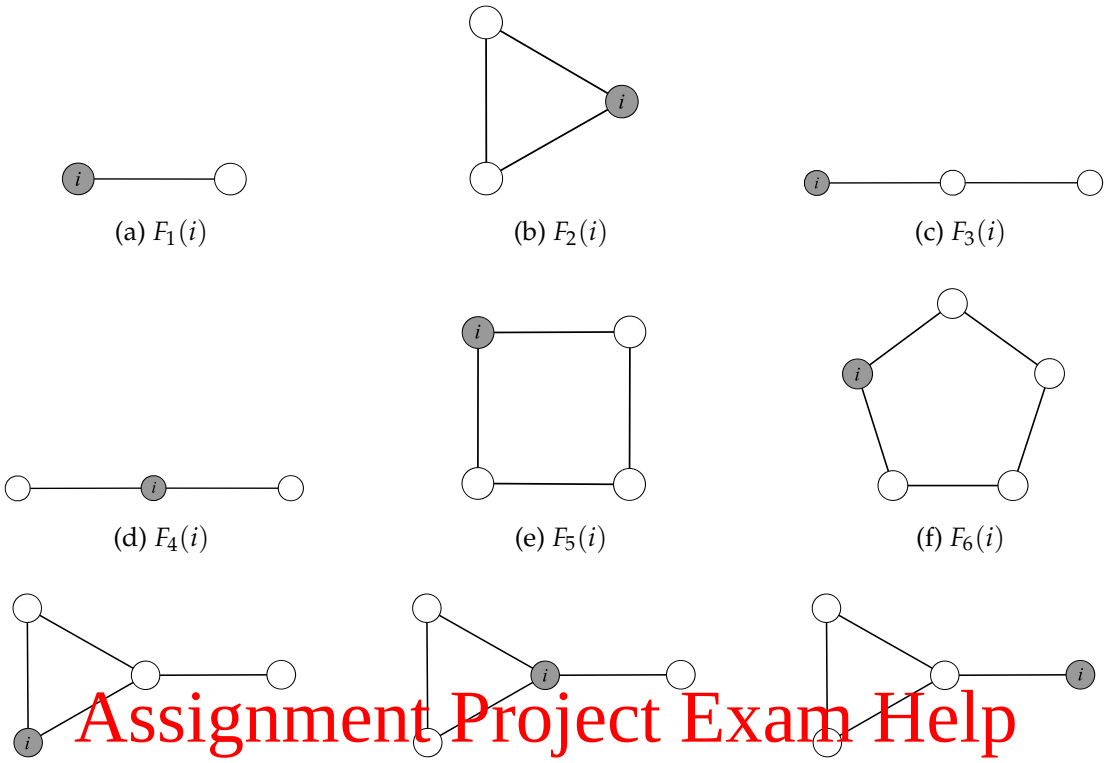
  where the rooted fragments are illustrated in Figure 15.6. So,

$$
\begin{aligned}
f_i(A) &= (c_2 + c_4)|F_1(i)| + (2c_3 + 10c_5)|F_2(i)| + (c_4)|F_3(i)| + (2c_4)|F_4(i)| \\
&\quad + (2c_4)|F_5(i)| + (2c_5)|F_6(i)| + (2c_5)|F_7(i)| + (4c_5)|F_8(i)| + (2c_5)|F_9(i)|.
\end{aligned} \tag{15.12}
$$

  By using $c_l = \frac{1}{l!}$ we get

$$
\begin{aligned}
f_i(A) &= \frac{13}{24}|F_1(i)| + \frac{5}{12}|F_2(i)| + \frac{1}{24}|F_3(i)| + \frac{1}{12}|F_4(i)| \\
&\quad + \frac{1}{12}|F_5(i)| + \frac{1}{60}|F_6(i)| + \frac{1}{60}|F_7(i)| + \frac{1}{30}|F_8(i)| + \frac{1}{60}|F_9(i)|.
\end{aligned} \tag{15.13}
$$

  Clearly, the edges (and hence node degrees) are making the largest contribution to the centrality, followed by paths of length two, triangles, and so on.



(a) $F_1(i)$ (b) $F_2(i)$ (c) $F_3(i)$

(d) $F_4(i)$ (e) $F_5(i)$ (f) $F_6(i)$

(g) $F_7(i)$ (h) $F_8(i)$ (i) $F_9(i)$

Figure 15.6: A collection of rooted fragments (roots designated by the letter i)

To define subgraph centrality we do not truncate (15.7) but work with the matrix functions which arise with particular choices of coefficients c_l. Some of the most well known are

$$EE_i = \left(\sum_{l=0}^{n} \frac{A^l}{l!}\right)_{ii} = \left(e^A\right)_{ii}, \tag{15.14}$$

$$EE_i^{odd} = \left(\sum_{l=0}^{n} \frac{A^{2l+1}}{(2l+1)!}\right)_{ii} = (\sinh(A))_{ii}, \tag{15.15}$$

$$EE_i^{even} = \left(\sum_{l=0}^{n} \frac{A^{2l}}{(2l)!}\right)_{ii} = (\cosh(A))_{ii}, \tag{15.16}$$

$$EE_i^{res} = \left(\sum_{l=0}^{n} \frac{A^l}{\alpha^l}\right)_{ii} = \left((I-\alpha A)^{-1}\right)_{ii}, \quad 0 < \alpha < 1/\lambda_1. \tag{15.17}$$

Notice that EE^{odd} and EE^{even} take into account only contributions from odd or even closed walks in the network, respectively. We will refer generically to EE as the subgraph centrality. Using the spectral decomposition of the adjacency matrix, these indices can be represented in terms of the eigenvalues and eigenvectors of the adjacency matrix as follows:

$$EE_i = \sum_{l=0}^{n} \mathbf{q}_l(i)^2 \exp(\lambda_l),$$

$$EE_i^{odd} = \sum_{l=0}^{n} \mathbf{q}_l(i)^2 \sinh(\lambda_l),$$

$$EE_i^{even} = \sum_{l=0}^{n} \mathbf{q}_l(i)^2 \cosh(\lambda_l),$$
$$EE_i^{res} = \sum_{l=0}^{n} \frac{\mathbf{q}_l(i)^2}{1-\alpha\lambda_l}, \quad 0 < \alpha < 1/\lambda_1.$$

Example 15.7

- We compare some centrality measure for the network illustrated in Figure 15.7.

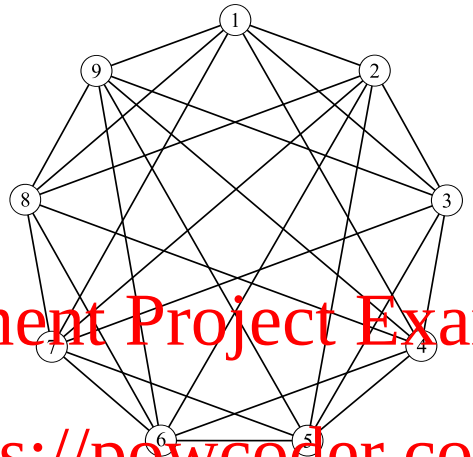


Figure 15.7: A regular graph with common degree 6

The regularity means that most centrality measures are unable to distinguish between nodes. The degree of each node is equal to 6. Also, because the network is regular, $\mathbf{q}_1 = \mathbf{e}/3$ and the closeness and betweenness centralities are uniform with $CC(i) = 0.8$ and $BC(i) = 2$ for all $i \in V$. Observe also that each node is involved in 10 triangles of the network.

The subgraph centrality, however, differentiates two groups of nodes $\{1, 3, 5, 6, 8\}$ with $EE_i = 45.65$ and $\{2, 4, 7, 9\}$ with $EE_i = 45.70$. This indicates that the nodes in the second set participate in a larger number of small subgraphs than those in the first group. For instance, each node in the second group takes part in 45 squares versus 44 for the nodes in the first group.

15.4.1 Subgraph centrality in directed networks

Subgraph centrality can be calculated for both directed and undirected networks using (15.14). Recall that a directed closed walk is a succession of directed links of the form $uv, vw, wx, \ldots, yz, zu$. This means that the subgraph centrality of a node in a directed network is $EE(i) > 1$ only if there is at least one closed walk that starts and returns to this node. Otherwise $EE(i) = 1$. The subgraph centrality of a node in a directed network indicates the "returnability" of information to this node.

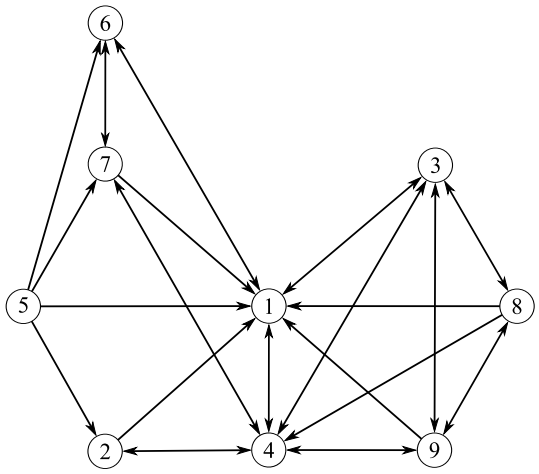


Figure 15.8: A network representing the observed flow of votes from 2000–2013 between groups of countries in the Eurovision song contest

Example 15.8

- In the Eurovision song contest, countries vote for their favourite songs from other countries. We can represent these countries as nodes and the votes as directed edges. The aggregate voting over the 2000–2013 contests has been measured with links weighted according to the sum of votes between countries over the 14 years.[2] The countries can be grouped together according to their pattern of votes. Groups which vote in a similar way are represented by the directed network illustrated in Figure 15.8. The labels correspond to countries as follows.

  1. Azerbaijan, Ukraine, Georgia, Russia, Armenia, Belarus, Poland, Bulgaria, Czech Republic.
  2. Netherlands, Belgium.
  3. Moldova, Romania, Italy, Israel.
  4. Macedonia, Albania, Serbia, Croatia, Slovenia, Boznia-Herzegovina, Montenegro, Turkey, Austria, France.
  5. Ireland, United Kingdom, Malta.
  6. Estonia, Lithuania, Latvia, Slovakia.
  7. Iceland, Denmark, Sweden, Norway, Finland, Hungary.
  8. Spain, Portugal, Germany, Andorra, Monaco, Switzerland.
  9. Greece, Cyprus, San Marino.

  The directed subgraph centrality for the groups of countries are

$$EE = \begin{bmatrix} 7.630 & 2.372 & 8.579 & \mathbf{12.044} & 1.000 & 2.950 & 3.553 & 5.431 & 5.431 \end{bmatrix}^T.$$

  The highest "returnabilities" of votes are obtained for groups 4, 3 and 1. The lowest returnability of votes is observed for group 5, which has no returnable votes at all, followed by group 2. Curiously, no countries from these two groups have won the contest in the last 14 years, while

[2] Details at `tinyurl.com/oq9kpj7`.

all the other groups (except group 3, which last won in 1998) have won the contest at least once in that time.

Further Reading

Langville, A.N. and Meyer, C.D., *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2006.

Estrada, E., *The Structure of Complex Networks. Theory and Applications*, Oxford University Press, 2011, Chapter 7.2.

Newmann, M.E.J., *Networks. An Introduction*, Oxford University Press, 2010, Chapter 7.