Chapter 9

Graphs and Networks

9.1 Examples of Networks in the Real World

Over the last few decades the term "network" has gained increasing importance. Many phenomena in the real world can in fact beneficially be viewed as networks. One might even want to call the 21st century the century of networks.

Newman (2003) makes a distinction between four types of networks, social networks, information networks, technological networks, and biological networks (note that this is only one way to classify networks, many other alternatives would be equally valid). Social networks include friendships, business partners, sexual relations, scientific communities; information networks, also called "knowledge networks", comprise citation networks, the World Wide Web, peer-to-peer networks (which can also be viewed as social networks), relations between word classes in a thesaurus, and preference networks (connecting individuals and objects of their preference such as books or films). Technological networks include electrical power grids, airline routes, networks of roads, railways, and pedestrian traffic, telephone networks, delivery networks, the Internet, other kinds of computer networks, etc. Finally, biological networks encompass metabolic pathways, food webs (e.g. how whales and dophins feed), water cycles (or in general ecological networks), brain networks, spread of diseases (e.g. bird flu, SARS, AIDS).

A particularly interesting type of biological network that received a lot of attention in the context of the human genome project, are the genetic regulatory network: Rather than assuming that genes in the genotype are mapped onto traits in the phenotype, the idea is that there are complex networks of genes interacting, which are responsible for the developmental processes. This enumeration of

networks could be extended almost indefinitely.

9.2 The discovery of small world networks

In the 1960's, Stanley Milgram did an experiment using real people and the US Postal service. He distributed letters all addressed to a stock broker in Boston, MA among 160 people from Kansas and Nebraska and asked them to forward this letter to the stock broker in Boston. The task was to only send or give the letter to someone they knew personally and who they thought should be more likely to know the stock broker (or someone else who might know him). About 3/4 of the letters were lost but those letters that made it to the stock broker travelled from person to person in a small number of steps (the "six degrees of separation"). This phenomenon holds of all kinds of social networks, e.g. for connections between black people in L.A. and whites in New York, or, as empirically determined by a German newspaper, between a kebab shop owner in Frankfurt and Marlon Brando.

Another experiment which started out as a kind of game was the Oracle of Kevin Bacon. As a database served the Internet Movie Data Base (www.imdb.com) which contains over 800,000 actors and all the movies that they played in. If you consider actors as nodes and movies as links or edges, almost all actors are connected to each other via movies that they or others played in together. Kevin Bacon was used as a center to which the average number of steps was computed and it turns out to be quite small (≈ 3). A similar point can be made about the famous mathematician Paul Erdös, who co-authored over 1500 publications with other mathematicians. In this case, the nodes designate mathematicians, the links co-authorship. The Erdös-number represents the number of steps a particular mathematician is away from Erdös himself.

In 1998, Watts and Strogatz published a seminal paper in Nature, which investigated how a few far-reaching connections can dramatically reduce the average path length between any two nodes even in a highly clustered network.

9.3 Some basic concepts for graphs and networks

While graph theory is mostly interested in characteristics of individual nodes or specific paths, network theory focuses on global properties as characterized by various statistical measure. Here is a list of basic concepts:

- vertex (pl. vertices): The fundamental unit of a network, also called a node in computer science, or an actor in sociology.

- edge: The line connecting two nodes. This is also called a link (or an arc) in computer science, or a tie in sociology.

- graph: a set of n nodes (or vertices) and k edges (or links, arcs). Graphs can be directed or undirected. Formall, we can write a graph as an ordered pair: $G = (V, E)$ with vertices (nodes) V and edges (links) E. This can be represented in a diagram such as Figure 9.1 where:

$$G = (\{A, B, C\}, \{(A, B), (B, C), (A, C)\})$$

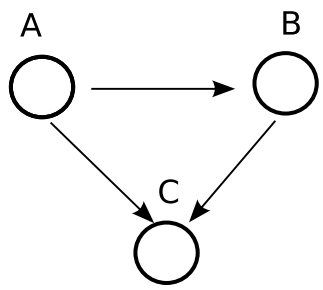

Figure 9.1: Example of a directed graph.

Tiny networks as the three-node example can be examined "by eye", but for very large networks statistical measures are usually used to study the global characteristics of a network.

- directed/undirected: An edge is directed if it runs in only one direction (typically indicated by unidirectional arrows). A graph is directed if all of its edges are directed (sometimes called a "digraph"). A graph is undirected if all its links are symmetric.

- degree: The number of edges connected to a node. A directed graph has both an in-degree and an out-degree for each node, which are the number of incoming and outgoing edges respectively.

- component: The component to which a node belongs is the set of nodes that can be reached from it. In a directed graph, a vertex has both an in-component and an outcomponents (corresponding to the nodes from which it can be reached, and the nodes that can be reached from the current node).

- distance: The distance between a source node j and a target node i is equal to the shortest path.

- adjacency matrix: The adjacency matric or connection matrix of a graph is a $n \times n$ with entries $a(i, j) = 1$ if node j connects to node i, and $a(i, j) = 0$, if there is no connection from node j to node i.

- distance matrix: The entries of the distance matrix $d(i, j)$ correspond to the distance between node j and node i. If no path exists, $d(i, j) = \infty$.

- path: A path is an ordered sequence of distinct nodes and links, linking a source node j to a target node i. No connection or node is visted twice in a path. The length of a path is equal to the number of distinct connections.

- cycle: A cycle is a path that links a node to itself.

And here are some examples of properties or measures that characterize entire networks and are not defined for individual nodes:

- average path length: The average path length, also called average degree of separation (or charcteristic path length) is the mean over all the path lengths in the network, i.e. for all pairs of nodes (i,j).

- average degree of a node: The average degree is the mean over all degrees in the network.

- distribution of degrees: Often we are not only interested in the average, but how the degrees are distributed (e.g. how many nodes with low degress, how many with high degrees, etc.).

- clustering coefficient: The number of connections between the nodes that are connected to a particular node. Intuitively, in a friendship network, this coefficient measure to what extent my friends are also friends of each other. Formally it is the maximum number of connections between the number divided by the actual number of connections.

- average clustering coefficient: Calculate the clustering coefficient for every node and take the mean.

- betweenness: Betweenness indicates the number of shortest paths going through a particular node. This number is used, for example, to characterize airports in a flight network.

- random network: A network with uniform connection probabilities and a binomial degree distribution. All nodes have roughly the same degree ("single-scale").

- scale-free networks: Gaph with a power-law degree distribution. "Scale-free" means that degrees are not grouped around one characteristic average degree (scale), but can spread over a very wide range of values, often spanning several orders of magnitude. Scale-free networks play a particularly important role in network theory.

- aristocratic networks: Because scale-free networks have a highly uneven degree distribution, they are sometime called aristocratic networks.

- egalitarian networks: In these types of networks, the nodes have roughly equal degree.

9.4 Analytical Approach to an Undirected, Regular Lattice

Consider a regular graph (Figure 9.2) with $n = 10$ nodes and a neighbourhood of $k = 2$. Neighbourhood of k means that each node is connected to those neighbours that are k or fewer nodes away. The number of edges in such a network is $n \cdot k$. The average degree is $2 \cdot k$.

9.4.1 Computing the Average Path Length

For a graph with 13 nodes and a neighbourhood of $k = 2$, the average path length (avp) is:

$$avp = \frac{4 \cdot (1 + 2 + 3)}{13} \approx 1.84$$

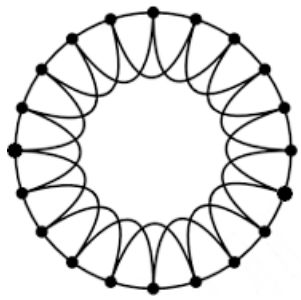

Figure 9.2: Undirected, regular graph with 10 nodes and a neighbourhood of 2.

In general, for this kind of regular graph:

$$avp \approx \frac{2k \cdot (1 + 2 + 3 + \ldots + \frac{n}{2k})}{n}$$

Using

$$1 + 2 + 3 + \ldots + x = \sum_{i=1}^{x} i = \frac{x(x+1)}{2} \approx \frac{x^2}{2},$$

we obtain:

$$avp \approx \frac{2k(\frac{n}{2k})^2}{2n} = \frac{n}{4k} \tag{9.1}$$

The larger the number of nodes n, the more steps are required. The larger the neighborhood (the degree), the smaller the number of steps.

9.4.2 Random Rewiring Procedure (after Watts and Strogatz, 1998)

We start with a regular lattice (as described above), a ring of n nodes, each node being connected to its k nearest neighbors (see Figure 9.3) by undirected edges. We choose a node and an edge that connects it to its nearest neighbor. With probability p, we connect this edge to a node chosen at random from the entire network (duplicate edges are forbidden). This process is repeated moving around the ring. Next, edges that connect the nodes to their second nearest neighbors are chosen and rewired in the same way as before. (As there are $\frac{nk}{2}$ edges in the entire graph, the procedure stops after $\frac{k}{2}$ laps). For $p = 0$, we have a fully structured

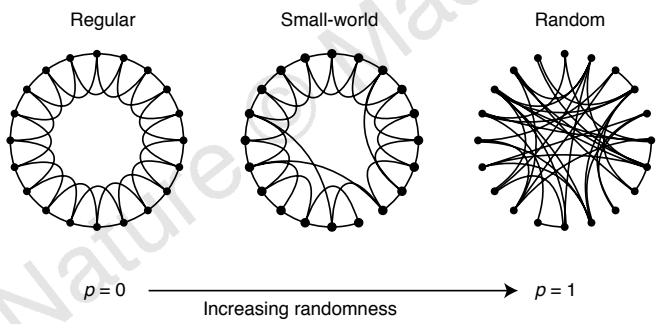

Figure 9.3: Random rewiring procedure (from Watts and Strogatz, 1998, Fig. 1, p. 441).

lattice. As p increases, the disorder increases, until at $p = 1$ we have a fully random network.

Figure 9.4 shows the characteristic path length and the clustering coefficient as a function of the probability p. The scale for p is log. It is interesting to observe that on the one hand the characteristic path length drops quite dramatically, even for low p (corresponding to relatively few rewired connections), whereas the clustering coefficient (corresponding to our local environment, so to speak, if we take our social network) remains quite high even for relatively high values of p.

9.4.3 Random Networks

Formally, assume that we have a random network with an adjacency matrix such as:

$$\begin{array}{c|ccc} & V_1 & \dots & V_n \\ \hline V_1 & 0 & & \\ \vdots & & \ddots & \\ V_n & & & 0 \end{array}$$

The maximum number of edges m_{max} in such a graph (with zeros in the diagonal) is:

a) for a directed graph $m_{\text{max}} = n(n \times 1)$, and

b) for an undirected graph $m_{\text{max}} = \frac{1}{2}n(n \times 1)$.

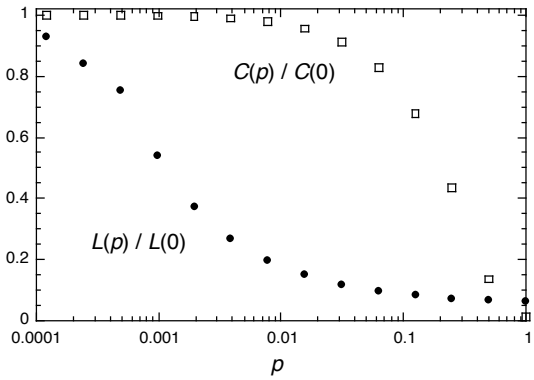

Figure 9.4: Characteristic path length $L(p)$ and clustering coefcient $C(p)$ for the family of randomly rewired graphs described in Fig. 9.3 (from Watts and Strogatz, 1998, Fig. 2, p. 441).

Assuming statistical independence for the connections, we can compute the probability that an undirected graph has m edges given that nodes are connected with probability p:

$$p^m \, (1-p)^{m_{\max}-m}$$

The mean degree d in such a random graph is approximately:

$$d \approx p(n-1)$$

Problems arise when random graphs are used to represent the real world:

- The assumption of statistical independence is often not valid
- The degree of adjacent vertices often differs strongly
- Random graphs have no community structure
- Navigation using local rules is not used

9.4.4 Clustering Coefficient (Degree of Clustering)

Informally, this measures whether “my friends are also friends amongst each other”.

More formally, the clustering coefficient C is defined as:

$$C = \frac{\text{number of acutal edges between neighbours}}{\text{number of possible egdes between neighbours}} \tag{9.2}$$

For a regular graph of the kind shown in figure 9.2, we have for example $C = \frac{3}{6} = 0.5$ for $k = 2$. More generally, it can be shown that for this kind of graph:

$$C = \frac{3k - 3}{4k - 2} \tag{9.3}$$

For large values of k, $C \approx \frac{3}{4}$.

9.4.5 Understanding small world networks

How is it possible that everybody is connected to everybody via just 6 people, if there are over 6×10^9 people on earth? One approach is: assume everyone knows 50 people. Then, I am connected to 50 people directly, via 2 steps already with $50 \cdot 50 = 2500$ and at 6 steps, I am already connected to 15,625,000,000 people, which is more than the current population of the planet.

Of course, this is a flawed calculation because acquaintances are highly clustered, in other words, the 50 people I know do not know 50 different people, but there is likely to be a significant overlap, i.e. my friends will often also be each other's friends.

The Strength of Weak Ties

Consider the whole population of human beings (6 billion) connected as a regular network with a neighbourhood of 50. Using Equation (9.1), the average path length for such a case can be computed as:

$$avp \approx \frac{6 \times 10^9}{4 \cdot 50} \approx 3 \times 10^7$$

If you replace 2 out of 10,000 connections with a random connection, the avp drops to about 8. If you replace 3 out of 10,000 connections with a random connection, the avp drops to about 5! (instead of 3×10^7).

This means that while your close friends (strong ties) are mostly connected among each other, they do not contribute very much to the overall connectedness of the network. On the other hand, loose connections (weak ties), your distant

acquaintances, are the links to whole new clusters in the network and therefore much more important in terms of average path lengths.

9.5 Growing Networks

Networks that grow in the real world typically develop a different structure than the model analyzed by Watts and Strogatz. One example is the Internet. In the Internet, one finds many computers with only one connection, but only few computers with a large number of connections. One can say, the network has a "hub structure".

If the distribution of the number of nodes is plotted versus the number of links from each node on a log-log scale, the result is a straight line. Such a distribution is called a"power law" (as we have seen) or"scale-free". In real-world networks, such power law distributions are ubiquitious. The same distribution can also be found in many other real world phenomena such as the size of earthquakes vs. the frequency with which they occur, or the size of avalanches vs. their frequency.

Although the Internet and the WWW are distinct entities – the Internet is physical while the WWW is logical – they both show the same structure and exhibit small world properties (meaning short avp). In the WWW for example, any web page can be accessed with about 19 clicks. The question that is often asked: Given that the WWW continues to grow at the present rate, will it still work in a few years. Estimates for an increase of factor of 1000 in the number of webpages forecast an avp of about 21 clicks - fortunately. In recent years the term "small world" has been used in a more precise way: in a "small world network" the value of avp scales logarithmically or slower with network size than for fixed mean degree.

Table 9.1: Examples of the different types of small world networks.

Egalitarian	Aristocratic
neurons in C.elegans	Internet
	WWW
web of roads and railways	sexual contacts
power networks	scientific papers linked by citations

The "hub structure" is enhanced by the nature of the growth process: new nodes tend to associate with existing hubs in order to get connected to as many

other nodes as possible. This method, also called "preferential attachment", makes existing hubs even more connected. Such networks are also called *aristocratic* networks, whereas networks with approximately the same degree of each node (e.g. Watts and Strogatz) are called *egalitarian* networks.

Case Study: Airline Transportation Networks

As an example for one of the world's busiest airport take Atlanta's Hartsfield-Jackson Intl. with over 88 Mio passengers in 2005. Analyses have shown that an airport operating close to half its maximum capacity will have many delays. When the airport system developed, it first became an aristocratic network. But when the maximum capacity of the big hubs was approached, the structure changed to a more egalitarian one as it became more attractive to also use smaller airports. In contrast to the Internet and the WWW, it is very costly to increase the capacity of an airport. Also, the capacity of airports cannot easily be expanded indefinitely. This is why it seemed cheaper to move to nearby smaller airports that still have spare capacity.

Guimerà et al. (2005) did a study on 3,883 locales (villages, towns, and cities with at least one airport) and established links between them if they are connected by nostop passenger flights. They found that the air transportation network is a small-world network for which the number of nonstop connections from a given city and the number of shortest paths going through a given city have distributions that are scale-free. However, in contrast to what one would expect, the most-connected cities are not ncessarily the most "central", that is, the cities through which most shortest paths go. This is because of the existence of several distinct "communities". Figure 9.5 shows the cumulative degree and betweenness distributions (i.e. $P(> k/z)$, where k is the degree, and z the average degree in the network; and $P(> b)$ where b is the betweenness). (Cumulative means the fraction of nodes that have degree greater than or equal to k). Surprisingly, there are only few cities with large degree and high betweenness. Anchorage has low degree (compared to other cities) but high betweenness, which is due to the fact that, because of politicial reasons most flights from Alaska go through Anchorage, rather than directly to other destinations which give the city a kind of key position. In other words, Anchorage lies at the interface between two (or several) communities (corresponding roughly to sub-networks, or components).

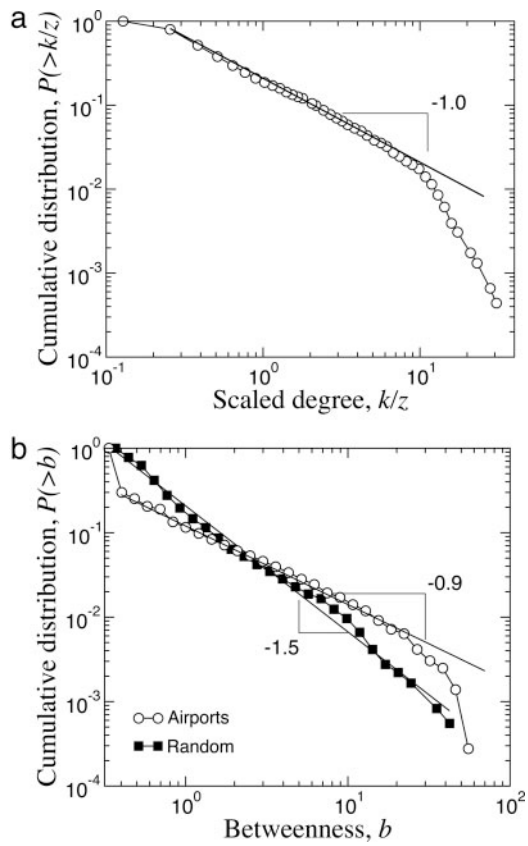

Figure 9.5: Degree and betweenness distributions of the worldwide air transportation network. (a) Cumulative degree distribution plotted in double-logarithmic scale. The degree k is scaled by the average degree z of the network. The distribution displays a truncated power-law behavior with exponent $1.0?? \pm 0.1$. (b) Cumulative distribution of normalized betweennesses plotted in double-logarithmic scale. The distribution displays a truncated power-law behavior with exponent $0.9?? \pm 0.1$. For a randomized network with exactly the same degree distribution as the original air transportation network, the betweenness distribution decays with an exponent $1.5?? \pm 0.1$. A comparison of the two cases clearly shows the existence of an excessive number of large betweenness values in the air transportation network (from Guimerà et al., 2005, Fig. 1, p. 7796).

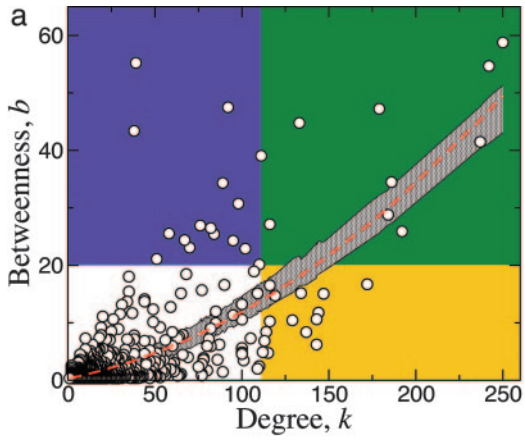

Figure 9.6: Betweenness as a function of the degree for the cities in the worldwide air transportation network (circles). For the randomized network, the betweenness is well described as a quadratic function of the degree (dashed line) with 95% of all data falling inside the gray region. In contrast to the strong correlation between degree and betweenness found for randomized networks, the air transportation network comprises many cities that are highly connected but have small betweenness and, conversely, many cities with small degree and large betweenness. We dene a blue region containing the 25 most central cities in the world and a yellow region containing the 25 most connected cities. Surprisingly, we nd there are only a few cities with large betweenness and degree (green region, which is the intersection of the blue and yellow regions). (from Guimerà et al., 2005, Fig. 2a), p. 7796).

The next Figure 9.6 shows a plot of degree vs. betweennes. Interestingly, there are only few cities that have high degree and at the same time high betweenness (the green area).

The average degree in Asia and Middle East is 3.5, in the global network only 1 step larger, 4.4. The growth rate is roughly logarithmic. The longest path is between Mount Pleasant in the Falkland Islands, and Wasu in Papua New Guinea which requires 15 different flights.

Applications and Relevance of these Analyses

What can we learn from these analysis? Questions we would like to be able to ask are

- How robust, how stable is the network agains attacks? (e.g. against terrorist attacks?) What about security
- identify bottlenecks, potential for congestion (e.g. betweenness)
- predict the spread of viruses and diseases

9.6 Analysis of threats

The analysis of networks is of practical importance, especially when you want to find out how brittle your network (e.g. a power network for the SBB) is.

9.6.1 Network resilience to deletion of nodes

To analyse this, randomly chosen nodes are deleted from the network (preferably in simulation), and the effect on the average path length (avp) is calculated. Previously, we have identified two kinds of small-world networks, aristocratic and egalitarian networks. Simulations have shown that if 5% of the nodes are destroyed, the avp of aristocratic networks remains unchanged, while it increases by $\approx 12\%$ in egalitarian networks. If 28% of the nodes are destroyed, egalitarian networks disintegrate completely, while aristocratic are still mostly connected (this is called graceful degradation). Note however that if nodes are not destroyed randomly but instead the hubs are attacked, an aristocratic network is more vulnerable.

9.6.2 Percolation models

Other threats are e.g. the spread of disease or of viruses. For such an analysis it is crucial to determine whether a node is "occupied/not occupied" and which nodes have the highest degree as these should be the first protected against "occupation" (infection). In order to find these nodes, one follows edges at random. The probability of finding a particular node is proportional to the degree of a node.

Similar ideas are used by Google for WWW search: the information contained both in the vertex and in the edge (the hyperlinks) is used (as well as e.g. the "authority" of a node).

9.7 Biological networks

As mentioned earlier, examples can be given indefinitely. We briefly look at three of them. The discipline of bio-informatics capitalizes, among other things, on the analysis of networks. Here are a few case studies:

- genetic regulatory networks (see Human Genome Project); models of GRNs for modeling ontogenetic development of artificial creatures, such as Josh Bongard's block pushers (e.g. Pfeifer and Bongard, 2007).
- motifs in brain networks (e.g. Milo et al., 2002; Sporns and Kötter, 2004, papers available from the lecture webpage).
- ant networks for foraging, pheromone trails; application to load-balancing in telecommunication networks.

9.8 "Tipping points"

A good way to understanding the emergence of fashion trends, the ebb and flow of crime waves, the transformation of unknown books into bestseller, the rise of teenage smoking, is to view them in terms of epidemics. Epidemics have three essential characteristics:

(i) they are contagious,

(ii) little causes can have big effects, and

(iii) change happens not gradually but at one dramatic moment.

And all epidemics have tipping points. A popular science book that nicely describes the idea behind "tipping points" is Malcolm Caldwells book "The Tipping Point" (2001). Here are a few examples. Because of the highly non-linear nature of the dynamics of large networks, they are hard to analyze analytically, and simulations are often the preferred means of study.

9.8.1 The Schelling Model of Segregation

As early as 1969, Thomas Schelling was interested in how micro-level preferences for like-colored neighbors might manifest themselves at the macro-level Schelling

(1969). Neighborhood relations can be depicted as networks where the nodes represent individuals (or families) and the links are only short-range indicating neighborhood (i.e. is neighbor of). In the simulation models that people have applied, there are typically constraints on the maximum number of neighbors, e.g. four. He defined a "movement rule":

(1) the agent computes the fraction of neighbors who are its own color;

(2) if this number is greater than or equal to its preference, the agent is considered satisfied;

(3) if this number is smaller than its preference, it will move to a different neighborhood.

Here is a summary of the results. If agents want at least 25% of their neighbors to be of the same color there is already a certain degree of segregation. However if you increase this threshold to 50%, but still allowing half of the neighbors to be of a different color, there is complete segregation. One of the points this study illustrates is that what looks like a dramatic issue at the macroscopic level, may not be reflected in individual preferences.

9.8.2 The Reappearance of "Hush Puppies"

"Hush Puppies" are legendary shoes, soft suede slip-ons, that made their debut in 1958. After a number of years they had almost completely gone from the market, selling a mere 30,000 pairs a year.

Then, to everyones surprise, the company, in a span of two years after its rebirth in 1994 managed to increase its sales to over 1,000,000 pairs, setting the fashion trends among young Americans for the second time. The reason might

have been, as Caldwell suggests, that just a few kids in New York wore them to a trendy club, in other words, a very minor event that eventually lead to the spread of the "fad" over the entire continent.

9.9 Network Motifs: Simple Building Blocks of Complex Networks

As we have seen, many networks in the real world are extremely large and their dynamics is very complex. Prominent examples of highly complex networks are genetic regulatory networks, biological brains (in particular the human brain), food webs, virus-spread networks (e.g. bird flu, SARS, AIDS), and WWW and the Internet. As a consequence, their structure and behavior is hard to understand. Or put differently, what would it mean to understand these networks? Often, it is difficult to try and explain them on the basis of first principles. What has been tried instead, or in addition, is to look for certain building blocks, the "network motifs". Motifs are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in randomized networks.

When we conceptualize a phenomenon in the real world as a network we have to define the nodes and links. For example, genetic regulatory networks are directed graphs, in which the nodes represent the genes. Edges are directed from a gene that produces a particular chemical (called the transcription factor) to a gene that is being regulated by this chemical (as described in Bongards simple model of genetic regulatory networks that control the growth of the "block pushers", e.g. Bongard). In food webs, the nodes represent groups of species. Edges are directed from a node representing a predator to the node representing its prey. In neural networks, nodes represent neurons (or neuron classes), and edges represent synaptic connections between the neurons.

Figure 9.7 shows all 13 types of three-node connected subgraphs.

Figure 9.9 shows a real-world network and several randomized networks. The networks are balanced in that each node in the randomized network has the same number of incoming and outgoing edges as does the corresponding node in the real network. The so-called feed-forward loop motif (number 5 in Figure 9.7) is detected. Note that its frequency is much higher in the real-world network that in the randomized ones.

The big question then is to figure out what these motifs might mean in terms of information processing in the network; identifying motif patterns is nice, but it

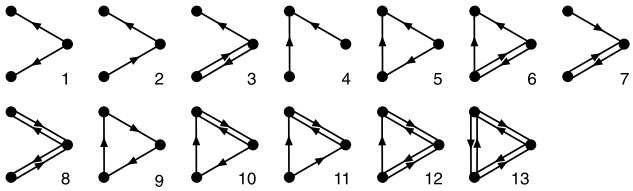

Figure 9.7: All 13 types of three-node connected subgraphs. From Milo et al. (2002).

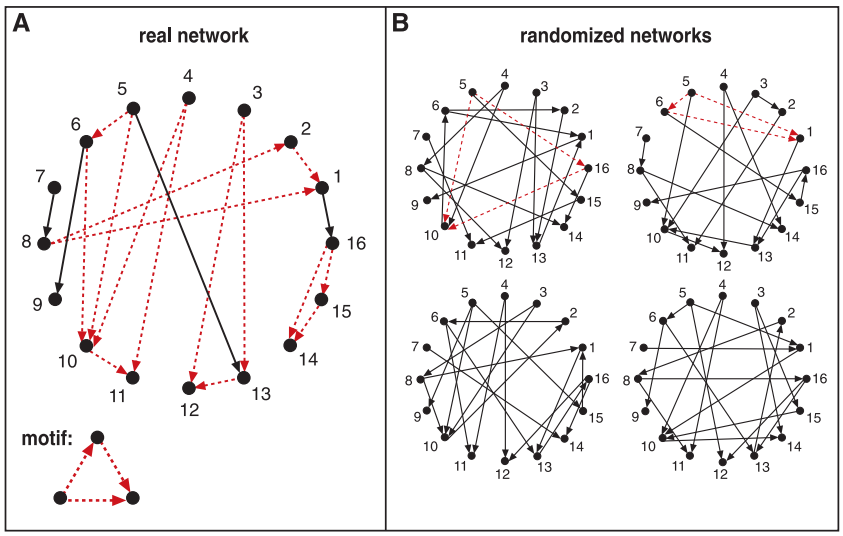

Figure 9.8: Real-world network and several randomized networks.

would be even nicer if they had a plausible interpretation of their functionality.

Milo et al. (2002) found that genetic regulatory networks, neural networks, and electronic circuits (forward logic chips) all contained the three-node feedforward loop motif (which seems to be in agreement with anatomical observations of triangular connectivity in some neural systems), and the four-node bi-fan motif. This similarity may point to a fundamental similarity in the design constraints of both types of networks. Both networks carry information from sensory components (sensory neurons/transcription factors regulated by chemical signals) to effectors (motor neurons/structural genes being produced under certain conditions).

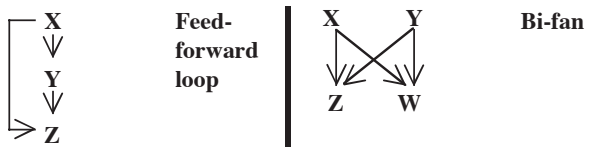

Figure 9.9: Feedforward loop (3 nodes) motif, and bi-fan (4 nodes) motif.

One of the important reasons why networks in the real world have a more pronounced motif structure than random network is that they are constrained by particular types of growth rules which in turn depend on the specific nature of the network.

9.9.1 Motifs in Brain Networks

When analyzing brain networks, often a distinction between structural and functional motifs is made (e.g. Sporns and Kötter, 2004). Structural motifs represent anatomical building blocks, whereas functional motifs represent elementary processing modes of a network. Functional motifs refer to specific combinations of nodes and connections (contained within structural motifs) that may be selectively activated in the course of neural information processing.

It has been found that the repertoire of functional interactions (i.e. functional motifs) is large and highly diverse, while their physical architecture is constructed from structural motifs that are less numerous and less diverse. This makes sense because a large functional repertoire facilitates flexible and dynamic processing, while a small structural repertoire promotes efficient encoding and assembly.

As can be seen from these considerations, it will be extremely important to have efficient network-processing algorithms.