Chapter 2

Real-World Networks

The real-world networks (aka empirical networks) could roughly be divided into four categories: biological networks, information networks, technological and infrastructure networks, and social networks. The study of a networked system comprises of a number of specific things that include the quantitative measurements of the system, development of proper quantification of the gathered data, and use of computer algorithms to analyze the data. Finally, one needs to do model the system using mathematics and statistics. These steps are often highly interdependent, and also present their own challenges. Especially measuring the networked systems is highly system dependent, and in most cases the tools developed for one system are useless for the measurement of another system. For example, most biological networks are constructed by using chemical and physical analysis in a laboratory. This method is useless if we want to gather a transportation network data. In this chapter, we will have a look at prominent networks in each category with respect to the systems they represent, and tools used to measure and study them.

2.1 Technological networks

Modern human society has developed a number of networks for various uses. The prominent ones of these include the Internet, power-grids, telephone networks, transportation networks like railways and highways, pipeline networks for water and fuel distribution and so on. In fact, major technological revolutions in human history are closely related to an addition of some technological network.

2.1.1 The Internet

The most recent such example is the Internet, which is the network of computers and other devices like cellphones that are connected to each other by wired or wireless means. Being a part of the network, these devices can communicate with each other. An important thing to realize here is that not all these machines are connected to each other. Instead, each machine is connected to only a few others, but nevertheless, each device is usually capable of communicating to any other in the network through the other devices. This 'network effect' is a key to the power of the Internet.

Although most of the devices in the Internet are used by the End users, these devices have only a single connection to the network, and hence they don't lie on the paths connecting different devices. This means that the devices like desktop computers and smartphones are incapable of facilitating the communication between other such devices, and so they lie on the periphery of the Internet. To route the data, a special type of computers, called *routers* are required.

If we imagine the Internet as made up of concentric circles with the end-user devices on the outermost circle, then the innermost circle or the core of the Internet is made up of special and powerful routers that are connected to each other by what are called as *trunk lines*. The routers in the core are owned by national governments and large communication companies. These are called *backbone service providers* or BSPs. The next circle consists of the routers owned by the Internet Service Providers or ISPs. The BSPs sell their service to ISPs which in turn provide the service to the end-users. Often one talks about the regional ISPs and local ISPs, dividing the middle circle into two, but many times the distinction is blurred.

Given this structure, let us briefly talk about how the Internet works. The primary function of the Internet is to transport a message (this could mean any type of data like a text message, video, a picture etc) from one device to another. The Internet is a packet switched network which means that the message to be transmitted is first broken into small data packets which are then reassembled at the destination. The way the packets are constructed from the original data is decided by a protocol called the *Internet Protocol* or IP for short. Each device in the Internet has its own IP address that identifies that particular device uniquely during the communication. Each data packet contains the IP address of its source (i.e. which device it originated at) and that of its destination (i.e. which device it is supposed to reach). Apart from this, a data packet also contains a special number called *time-to-live* or TTL. Initially its value is set to some positive integer (usually 64), and each time the packet hops to the next router, the value is decreased by 1. At any router if the TTL value reaches 0, the packet is discarded, and the discarding router sends a message to the source of the packet that the packet was discarded. The purpose of the TTL is to stop misdirected packets from wandering over the Internet forever (and hence stopping the congestion as well as saving the resources).

Measuring the structure of the Internet

It might sound at first sight that this is a relatively simple task. After all, the Internet is a human-made network, and so surely its structure must be completely known. This is far from reality! The main point of difficulty is that there is no central authority or the 'Internet Government' that looks after the operation of the Internet. If you want to add a router, you don't need to apply to anyone for a permission to do so. The structure of the Internet today is thus a result of many such 'local' actions by a large number of individuals over a long period of time. Thus, there is no map of the Internet that we can look at to determine its structure, and it must be done by some other means.

The most common way to construct the structure of the Internet is to find lots of paths in the network, and to construct the network from these paths. To get a better idea of what this means, see Fig. 2.1. In the left part, we have "somehow" found paths starting from vertex 2 to vertices 1 and 5. Notice that in general there could be several paths between a given pair of vertices. Here we are assuming that for each pair, we have found just one path. Similarly, in the right part of the figure, vertex 7 is the source, and we have found paths starting from it to the vertices 0 and 4. The edges that are not part of these paths are drawn thin, which just means that we have no idea of their existence. Now if we combining these two pictures (which actually means combining the data for the two sources 2 and 7), then we get at least a partial picture of the structure of the network (see Fig. 2.2).

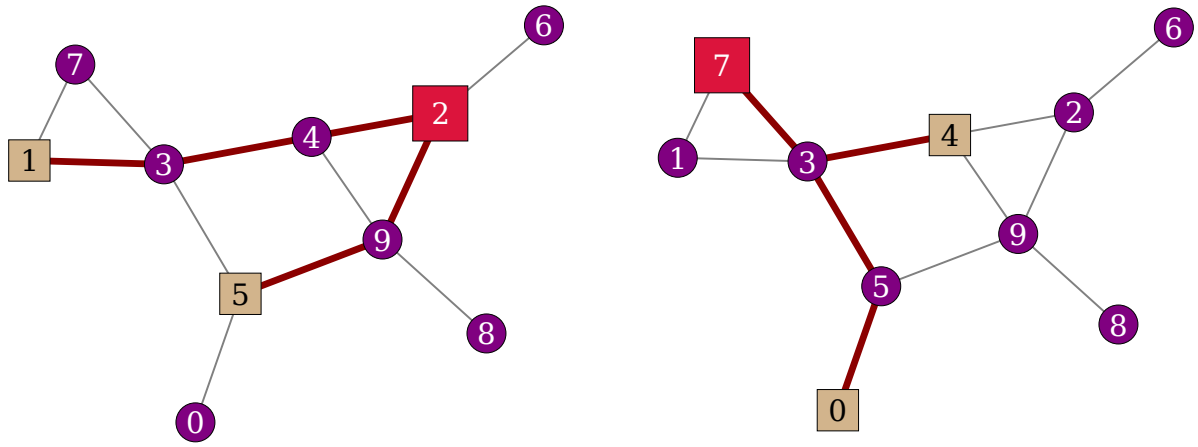

Figure 2.1: *Left*: Starting from vertex 2 as the source, two paths to the target vertices 1 and 7 are found (rendered thick and brown). *Right*: The source is changed to 7, and targets to 0 and 4. The edges that are not part of the paths so found are rendered in thin gray.

However, it should be noticed that because the our paths don't include all the edges in the network, some edges and some vertices will usually don't appear in our final measurement of the network obtained by

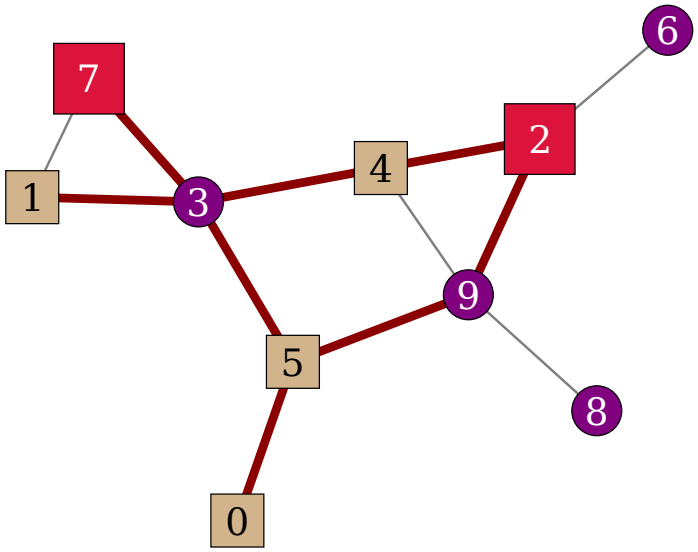

Figure 2.2: When the path data for the two sources in Fig. 2.1 is merged, we get an idea of the structure of the underlying network. Notice that some vertices (here 6 and 8) as well as some edges don't appear in the final measurement since our paths don't include all the edges in the network.

merging data from many sources. Such errors can be minimized by using many more sources and destinations, but some errors will inevitably occur as is the case in any type of measurement.

Two popular methods to gather paths data in the Internet are *Traceroute* and the use of *Routing Tables*. Here, we will not go into the details of the way these methods work, and the interested ones can have a look at the book by Newman.

2.1.2 The telephone network

One of the oldest electronic network still in use is the Telephone network. Although much has changed in the way we make telephone calls (including the advent of the *Wireless* technology), the structure of the telephone network has mostly remained almost the same. Even now, most wireless calls are first sent to the nearby transmission tower from which the signal is carried over the traditional telephone lines. The structure of the telephone network is quite simple, and is made up of three layers: the end-users, the local exchanges, and the long-distance offices. The end-users or the telephone subscribers are directly connected to the local exchanges by means of local lines. These exchanges in turn are connected to the long-distance offices by what are called the *trunk lines*. The long-distance offices are also connected to each other by trunk lines. The purpose of having such layers is to exploit the fact that most telephone calls are local! Most people make telephone calls to others who are geographically close to them. Thus, it makes no sense to use expensive trunk lines to make such calls, and only local exchanges can handle these calls. Sometimes, even the local exchanges are connected to each other by trunk lines so that some calls in geographically nearby local exchanges need not be handled by the long-distance offices.

In recent years, telephone companies have been increasingly sending the telephone data over the Internet instead of the traditional telephone network, and so in the near future the two networks may in fact get merged.

2.1.3 Power grids

This is a network with the power-generating stations and the power-switching stations as the vertices and the high voltage lines connecting them as the edges. The structure of the network is of interest at least for two reasons. First, the structure and the growth of the power-grid network contains clues about the geographic and economical situation of the region. Second and more important, often the failures in power-grids propagate through the network resulting in a cascade. Such cascades of failures result into blacking outs of entire regions, and so network scientists have a huge interest in understanding the dynamics of cascades in power-grids. Fortunately, power-grid networks are usually managed by single authority in most countries, and hence getting the data is relatively easy.

2.1.4 Transportation networks

Transportation networks consist of geographical regions like cities and villages connected to each other by means of means of transportation like railways, highways and airways. Fig. 2.3 shows the domestic air network of India. Similar to power-grids, the transportation networks reflect to a great extent the economic development and the inhabitance of the region.

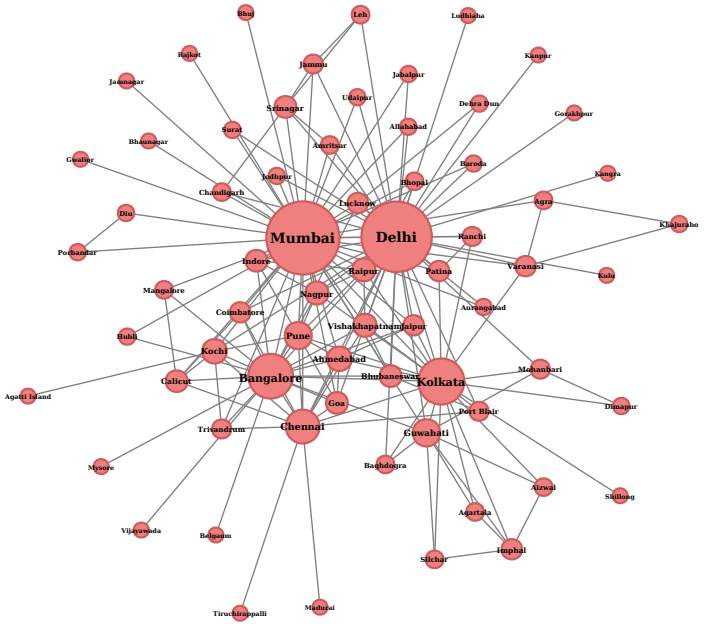

Figure 2.3: The Indian domestic air network. The node sizes are roughly proportional to the degree values.

2.1.5 Distribution networks

A large number of systems that "distribute" something take the form of a network. Some examples are pipeline networks used to distribute gas and water, routes used by courier companies, and natural systems like river networks. There has been relatively less amount of work on these networks although these are crucial for human society.

2.2 Information networks

Information networks consist of chunks of data connected to each other. These networks are unique in the sense that all information networks are human-made. Some of the examples are the World Wide Web,

citation networks, social networking sites such as Facebook and so on. Some of these networks such as Facebook can also be regarded as social networks, and as already mentioned earlier, such distinction is mostly a matter of convenience. Let us then have a brief look at some prominent information networks.

2.2.1 The World Wide Web

Perhaps the best known information network is the World Wide Web or WWW for short. The World Wide Web is a network in which vertices are the individual webpages, and edges are the hyperlinks that connect different webpages with each other. Although the term 'Internet' is casually used for the WWW, it should be kept in mind that the two network are completely different. As seen in the previous section, the Internet is the network of physical infrastructure like computers, routers and cables whereas the WWW is a virtual network of data chunks in the form of webpages connected by virtual hyperlinks. One could say that the Internet is a substrate for the WWW.

In the WWW, the hyperlinks are directed which means that if you can click a link on page A to go to another page B, usually the page B won't have a hyperlink pointing to page A. This directionality can be represented by placing a small arrow on the edges when the network is drawn so that the direction of the arrow shows that direction of the link itself. In the jargon of networks, we say that the World Wide Web is a directed network (See Fig. 2.4). This is not the only example of the directed network, and a large number of real-world networks are in fact directed.

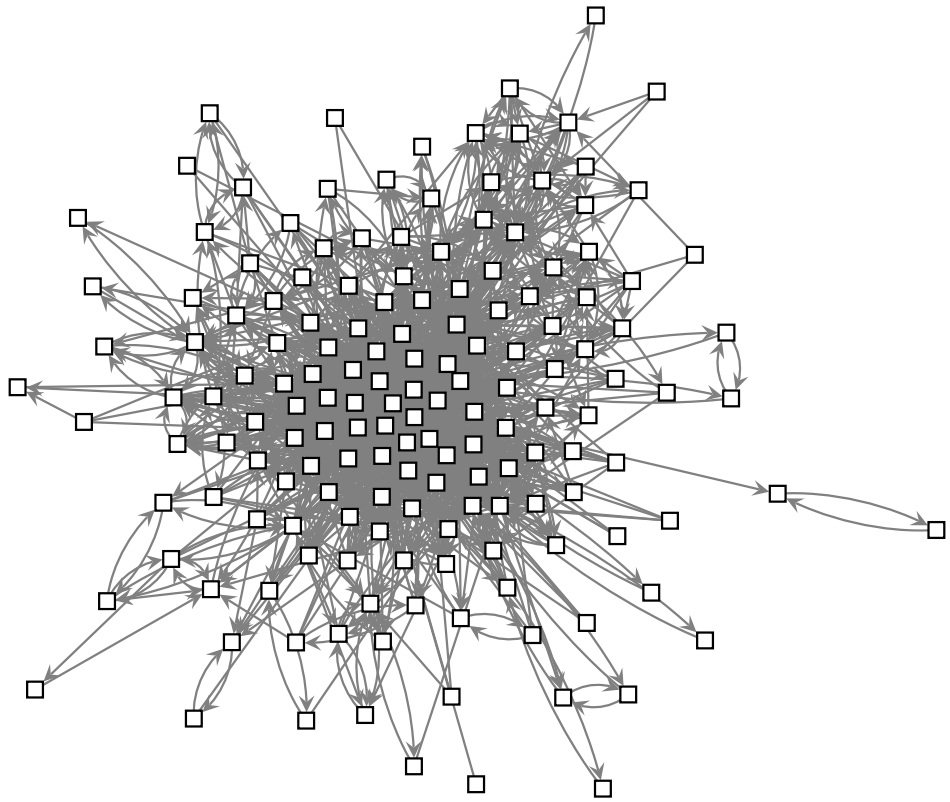

Figure 2.4: A small portion of the World Wide Web

Similar to the Internet, there is no central authority to look after the WWW, and hence its structure must be determined by indirect means. The most standard way to do this is to use a specialized computer program called a *crawler*. A crawler roughly works in the following fashion. We initialize it on a particular webpage. It then downloads the content of that page, and finds out hyperlinks embedded in that page and stores them. It then takes each of those hyperlinks, and visits the corresponding webpage, downloads its content and stores the hyperlinks in that page. This process is repeated to gather larger and larger picture of the World Wide Web. Although the process itself is straightforward, usually there are other things which restrict us from using a crawler to determine the structure of the whole of the WWW.

First, we must make sure that our crawler doesn't get trapped in a *loop*, a set of webpages that point to each other. This is a very common possibility, and most ready-made crawlers now have a capability to handle such things. In fact, whenever these crawlers see a webpage that they have already visited, they discard it immediately. However, apart from this minor obstacle, the shear size of the WWW makes it impossible for any one crawler to map all of it. Currently there are around 50 billion pages on the WWW, and most of us don't have infrastructure to map such a huge network. Also, several websites restrict crawlers from accessing their data. Also, many webpages are actually 'dynamic', in that they are created when requested (an example is the page of search results for Google search).

But perhaps the most important reason that crawlers can't see the whole of the WWW is that some pages are hidden from them simply because of the structure of the network. Imagine a webpage that has no incoming link. If we start our crawler on any other webpage, then it is impossible to locate it. Another possibility is to consider a webpage that has incoming links from many other webpages, but all of those webpages have no incoming links and so on. As we will see later, this is a general property of the directed networks.

Although we have been talking about employing crawlers to study the structure of the WWW network, it is not their primary purpose. Giants like Google who offer a search service need to maintain large databases of data on individual webpages. These data are processed when somebody searches for a particular term or topic. These companies use a large number of advanced crawlers to get such data. Since some webpages are preferred over others for various reasons like advertising, quality of content etc, inevitably the data gathered by these companies is highly biased. For this reason, one may prefer to use a separate crawler to gather unbiased data whenever possible.

2.2.2 Citation networks

Another important, although a less well-known, type of information networks is citation networks. Citing is an act of mentioning another event, information or a piece of knowledge. The most prominent example of citation networks is the network of scientific papers. When a scientific paper (also called a 'research paper') is written, the authors cite the papers that have already been written about the topic, and are somehow relevant to the paper being written. The relevances may include support for a claim being made, an argument against someone's claim in the past or just a simple indication of the work that has happened before. Because of this, the citation network provides information about the relatedness of various topics, and there has been a great interest in studying citation networks of scientific papers. Probably the first study of citation networks was carried out by Price in 1960s, and now specialized databases containing the citation data exist (for example, Google Scholar and Science Citation Index) that let researchers quickly build citation networks of interest.

Fig. 2.5 shows starting and ending of a typical scientific paper. The name 'PHYSICAL REVIEW E' at the top is the name of the journal in which is was published, and the numbers embedded in the text are the numbers corresponding to the papers being cited. The full references to these cited papers are written at the end of the paper.

PHYSICAL REVIEW E **91**, 062805 (2015)

Stochastic block model and exploratory analysis in signed networks

Jonathan Q. Jiang*
Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong, China
(Received 2 January 2015; published 15 June 2015)

We propose a generalized stochastic block model to explore the mesoscopic structures in signed networks by grouping vertices that exhibit similar positive and negative connection profiles into the same cluster. In this model, the group memberships are viewed as hidden or unobserved quantities, and the connection patterns between groups are explicitly characterized by two block matrices, one for positive links and the other for negative links. By fitting the model to the observed network, we can not only extract various structural patterns existing in the network without prior knowledge, but also recognize what specific structures we obtained. Furthermore, the model parameters provide vital clues about the probabilities that each vertex belongs to different groups and the centrality of each vertex in its corresponding group. This information sheds light on the discovery of the networks' overlapping structures and the identification of two types of important vertices, which serve as the cores of each group and the bridges between different groups, respectively. Experiments on a series of synthetic and real-life networks show the effectiveness as well as the superiority of our model.

DOI: 10.1103/PhysRevE.91.062805 PACS number(s): 89.75.Fb, 05.10.−a

I. INTRODUCTION

The study of networks has received considerable attention in recent literature [1–3]. This is mainly attributed to the fact that a network provides a concise mathematical representation for social [4,5], technological [6], biological [7–9], and other complex systems [1–3] in the real world, which paves the way for executing proper analysis of such systems' organizations, functions, and dynamics.

observed network structure, vertices with the same connection profiles are categorized into a predefined number of groups. The philosophy of these approaches is quite similar to that of the "role model" in sociology [30]—individuals having locally or globally analogous relationships with others play the same "role" or take up the same "position" [31]. It is clear to see that the possible topologies of the groups include community structure and multipartite structure, but they can be much, much wider

[1] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).
[2] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Not. Am. Math. Soc. **56**, 1082 (2009).
[3] S. Fortunato, Phys. Rep. **486**, 75 (2010).
[4] G. Palla, A.-L. Barabási, and T. Vicsek, Nature (London) **446**, 664 (2007).
[5] A. L. Traud, E. D. Kelsic, P. J. Mucha, and M. A. Porter, SIAM Rev. **53**, 526 (2011).
[6] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, IEEE Comp. **35**, 66 (2002).
[7] R. Guimerà and L. A. N. Amaral, Nature (London) **433**, 895
[23] T. Yang, Y. Chi, S. Zhu, Y. Gong, and R. Jin, Mach. Learn. **82**, 157 (2011).
[24] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, J. Mach. Learn. Res. **9**, 1981 (2008).
[25] H. W. Shen, X. Q. Cheng, and J. F. Guo, Phys. Rev. E **84**, 056111 (2011).
[26] T. P. Peixoto, Phys. Rev. X **4**, 011047 (2014).
[27] A. Decelle, F. Krzakala, and L. Zdeborova, Phys. Rev. Lett. **107**, 065701 (2011).
[28] A. Clauset, C. Moore, and M. E. J. Newman, Nature (London) **453**, 98 (2008).

Figure 2.5: A typical research paper showing how the past literature is cited and the actual citations written at the end of it.

Apart from this, other citation networks like patent citations and legal citations exist and have been studied by various researchers.

2.2.3 Peer-to-peer networks (P2P networks)

In these networks, the nodes are the computers that store some information usually in the form of discrete files (music, movies, e-books etc) and the links are the virtual links established for sharing these data. An important characteristic of P2P networks is that there doesn't exist a central server containing all the data. An absence of a central authority means that each computer in the network contains some data, but there is no computer that contains all the data. Many P2P networks are used often for illegal sharing of copyrighted materials although legal P2P networks are also quite common. An interesting network problem in the context of P2P networks is to search for a given piece of data on the network. The interested readers can have a look at the Newman's book.

2.2.4 Recommender networks

When you buy/view a product from an online seller like *Amazon*, typically you are also shown similar products under various headings (e.g. "You may also like"). Other services like *YouTube* also recommend 'similar videos' when you view a particular video. Such system that under the hood does this job of recommending products based on their similarity to the products customers showed interest in, is called a 'recommender system'. How does a typical recommender system work? Although a lot is involved in a modern day recommender system, its core is a network. In this 'recommender network', there are two types

of vertices: customers/users and products. The links in the network connect products to users but there are no links between the products or between the users. If a customer A has liked/bought products p and q then in the network there A is connected to p and q by links. Such networks with two types of vertices are quite common, and are called 'bipartite graphs'. We will study these later in the course.

The way a recommender system works then is this: suppose there are two users A and B both of which like many of the same products. Now suppose A has also recently used another product that B hasn't. It is reasonable to assume then that B would also be interested in that product, and so the product is recommended to B. A large number of computer algorithms have been developed around this idea, and these days companies like Amazon, YouTube and Netflix maintain large databases containing such bipartite graphs. In fact, in 2006, Netflix offered a huge prize for designing a better recommendation system to theirs (in fact 10% better!). Although this may look like a small improvement, for a big company like Netflix, such improvement usually means a remarkable increase in the profit. The prize was ultimately won by a large team of researchers (the BellKor's Pragmatic Chaos) who beat the Netflix system by 10.06%.

2.2.5 Keyword indexes

All of us have seen indexes at the end of technical books where several terms are listed along with the pages on which those terms have appeared in the book. Such indexes can easily (and usefully) be thought of as bipartite networks similar to recommender networks. In a keyword index network, one type of vertices is the technical terms or words, and the other type is the pages (usually in the form of page numbers). A word is connected by links to all the pages on which it appears. What is useful about constructing such network? First and foremost, this is just out of curiosity; we may just want to see if the structure of the network is affected by the topic of the book or whether there exist interesting structural patterns that tell us something about the subject of the book is organized. On a more pragmatic level, you may be interested in searching a large corpus of literature about a particular topic to find similar/related concepts/techniques. It is easy to see that we could do this in a manner similar to the one used in recommender networks.

In fact, search engines like Google inevitably need to maintain large databases for keyword indexes. The only difference is that instead of the pages of a book, the pages are the webpages. When a user queries a particular word/phrase, the first step that is taken is to search the available index to find the pages that contain that particular word before ranking them. When the data is already stored in the form of a bipartite graph, this becomes much easier (Probably this is the only practical way of storing such data!). However, the efficiency of such procedure potentially depends on the structure of the underlying keyword index network, and from this point of view, it is important that we study them. Unfortunately, not much has been explored about this side of the index networks, and there seems to be an ample space for work.

2.3 Social networks

As mentioned earlier, social science has a long and rich history of studying networks, and many tools that we use today to deal with networks were actually developed by social science researchers. A social network consists of vertices that are either humans, animals or their groups (e.g. a business firm), and the edges in a social network represent some kind of interactions like an acquintances (i.e. two people knowing each other), friendships, professional relations, sexual contact, family relations etc. Which interaction to use to form edges depends upon the type of questions that one wants to ask. For example, if the goal is to study the dynamics of a business firm, one would construct a network of the professional relationships rather than the network of friendships.

The first definitive study of social systems in the form of a network was done by a psychiatrist Jacob Moreno in 1930s. In one of his projects, he gathered data of friendships in schoolchildren (probably by asking them who their friends are), and constructed network pictures by hand depicting these friendships (See Fig. 4.1 from Newman's book). His 'sociograms' became so famous then, that a column was dedicated to his work in New York Times. He also termed this new discipline 'sociometry' although we now call it 'social network analysis'.

Another study done a few years later in a small town called Natchez in the United States has become famous by the name 'Southern Women Study'. In this study, the researchers compiled information about

attendance of social gatherings by 18 women from the high social class over a span of nine months. They used guest lists of the gatherings, descriptions of the event in the local newspapers etc to construct a bipartite network in which women are connected to the events they attended. In their network, they found that the women in the study are approximately divided into two groups that mostly attended different events without much overlap i.e. very few women attended events attended by both the groups. From modern point of view, this is an example of a community structure, and many sophisticated methods have been developed over the years to automatically extract such groups from large network datasets owing to their usefulness. (See Fig. 4.2 from Newman)

In spite of being very familiar to us, social networks remain the most difficult networks to understand given the complexity of the human beings and different types of relations that can exist among the individuals. Just to appreciate this argument, just try answering an innocent sounding question to yourself: how many friends do you have?

Now let us briefly look at various methods used to measure the structure of a social network.

2.3.1 Interviews and questionnaires

Arguably, the most obvious way to find out social relations is to ask people. Two methodologies exist which use this approach to gathering the social network data. In the first approach, one creates a *questionnaire*, a set of relevant questions that the participant in the study has agreed to answer. The questions may ask information other than that related solely to the structure of the network around them. For example, apart from asking them who their friends are, we may also ask what do they like to eat or where do you go for a trip etc. The same things can be asked by directly interviewing the participants instead of using a printed or online questionnaire. But an even better strategy is to combine the two approaches wherein the interviewer asks questions one by one from an already constructed questionnaire to a participant. The combined approach works better for several reasons. First, when a human is asking questions directly, participants tend to take the study more seriously, and answer the questions thoughtfully. This results in a significant reduction in the amount of vagueness and casual responses which in turn leads to better scientific conclusions. Second, when questionnaire is already constructed, there is no fear that the interviewer might forget some of the questions. Finally, if a certain question is unclear or confusing, the interviewer can assist the interviewee to understand it, thereby guaranteeing that an accurate response is received.

Certain things should be noted here. First, usually one conducts such studies in a restricted population or a community like a school or a company. In most cases, a participant from the group has ties outside the group, but during the study such ties are ignored. For example, a question might ask a child in a school to name his friends in the school. But the child probably also has friends outside the school, and because of the nature of the question, these ties are ignored. This means that not all the statistics drawn from the data are reliable. Also, it should be noted that questionnaires result in a directed network it is up to the participant to decide which people to name as their friends. Although we think that a friendship is a two-way relation, and hence if A mentions B as their friend, then B will necessarily name A as friend, that is not what is observed. Interestingly, it has been observed that around 50% of the edges in the networks so constructed are unidirectional. It turns out that this is related to an existence of social hierarchy or *ranking* in social groups, and one can infer the actual ranking from such data. There are other issues with the method. In many such questionnaires, the experimenter restricts the response of participants artificially: instead of asking who your friends are, the question may ask the interviewee to name their 10 best friends, and if someone has more than 10 friends, those won't be recorded. This cutting-off of the degree may sometimes result into erroneous conclusions. At the same time, such cutoff may reduce the vague answers (when restricted, people must think harder about the questions). It also means that there is less work for the experimenter.

2.3.2 Ego-centered networks

Usually a study that uses interviews and questionnaires are limited to small groups because of the efforts involved. An alternate method to gathering network data when the group of interest is large is to randomly sample individuals, and just interview them. These individuals are called **egos**, and in the study, they are asked to name their friends (called **alters**) and the friendships between them if known. In effect, we get a local snapshot of a network around an ego, and since relatively few egos are selected (that was the point of

sampling), we don't get to know the structure of the whole network. Nevertheless the method is useful for estimating the local quantities like the average degree, the average clustering coefficient, assortativity (we will study these later in the course) etc.

2.3.3 Direct observation

Another method to gather data on the social relations is to simply observe people interacting with each other. When observed for sufficiently long time, social ties in the group could be established. In fact, sometimes this is the only method to uncover social ties: if you are interested in the social network of animals, interviewing them makes no sense. A classic example of a network data gathered using direct observation is the karate club network that we saw in the introduction.

2.4 The small-world experiment

The psychologist Stanley Milgram became interested in the topic of *typical distance* in social networks in 1960s. We will study the notion of the *distance* in networks rigorously later, but for now think of it as the smallest number of hops you need to reach a given person by following acquintances. For example, consider any person that you know, say your friend or a family member or your teacher. Since you directly know him/her, we say that that person is at distance 1 from you. Let's call this person A. Now think of another person B whom A knows, but you don't know B. Then it is reasonable to say that the distance between you and B is 2, and so on. For any two nodes i and j in the network, we can think of the distance d_{ij} like this. For some pairs, it would be small, and would be large for others. But we can ask what is its average or typical value. If we think of all humans on earth as forming a giant social network, then there are around 8 billion nodes in the network, and hence it may feel that the average node-to-node distance in this network must be huge. One of the surprising findings in the field of networks is that this typical distance increases much more slowly with the size of the network than we think, and in fact Milgram had already heard of mathematical arguments suggesting so, and wanted to test those for the real-world.

The experiment he performed to achieve this has become a landmark in the history of networks, and is famously known as the small-world experiment. The experiment consisted of 96 volunteers chosen from the city of Omaha in USA, each one of which was given a booklet or a 'passport'. The participants were then asked to send these passports to a single target, who happened to be a friend of Milgram, in the city of Boston some 1500 miles away. Three pieces of information about the target were provided to the participants: his name, his location and his occupation (he was a stockbroker). But a participant was not allowed to send the passport directly to the target, and had instead been asked to send it to somebody they knew on the first name basis. Although each participant could in principle send the passports to a random acquaintance, they were encouraged to send them to that person who they thought would be somehow 'closer' to the target. Here closer doesn't necessarily mean geographically closer; even a person with a similar profession to that of the target can also be said to be closer to him. When that person receives the booklet (aka passport), he/she was asked to repeat the procedure until it reaches the intended target. All the people who received the booklets were also requested to note down their information in the booklet before they forwarded it. Out of 96 passports, 18 reached the targets, and other were somehow lost/discarded. This may sound a low number, but by modern standards, turns out to be exceptionally high.

From the passports that reached the target, Milgram could find the length of each of the chains that started with a participant in Omaha and ended in Boston, and hence the average length. He found the value to be 5.9 which is surprisingly low compared to the human population on earth or even in the USA (even in 1960s!). This is the origin of the phrase **Six degrees of separation** which (somewhat inaccurately) says that everybody in the world is just six handshakes away from each other.

Impressive as it may sound, the experiment and the conclusion cannot be completely accurate for several reasons (apart from the obvious reason that not all paths in the network are sampled). First, the sample of the participants in the experiment was not chosen uniformly randomly from all the available population; they were from the same city, and they responded to the newspaper advertisement. Similarly, the target was also not typical as he was probably chosen just because he happened to be Milgram's friend. There are other subtleties in the way the data is used to estimate the typical distance. Observe that for a given

passport, there is no guarantee that it followed the shortest path from the volunteer to the target. Hence, only the upper bound is known for the pair consisting of a participant and the target. If we had upper bounds for **all** the node pairs in the network, their average would have given us the upper bound on the actual average length. Since only some of those upper bounds are available, their average may be greater than the true average or may even be less than the true average. It is also instructive to think about the lost passports; it is reasonable to assume that they were discarded probably because they followed longer paths. Thus, the estimate that we have got is probably biased towards lower values. However, in spite of several such inaccuracies, the main conclusion of the experiment that the node-to-node distance in social networks is much smaller than the size of the network, is now well accepted. For example, in a 2011 study that used actual Facebook network data to find the average distance in the network found it to be around 4.7, a way too small number than the size of the Facebook graph (well over a billion at that time).

Funneling effect

Milgram observed another interesting thing in his experiment: the total 18 passports that reached the target, reached to him only through 3 of his acquintances although he surely had many more contacts in the network. It is as if most of his connectedness to the outside world was through those three contacts. This effect, called *funneling effect* has been observed in several other networks also, especially in citation networks. However, there are other studies that have claimed that no such effect exists in social networks.

Navigability

Another important aspect of the small-world experiment that Milgram himself didn't realize is the notion of *navigability*. This was highlighted in 2000 by Kleinberg. The essence of navigability is that the people who participated in the experiment surely didn't know the structure of the network they were part of, and so it should have been quite difficult for them to find short paths in the network. Nevertheless, they seem to have done impressive search just based on a few pieces of information about the target! Put differently, people seem to be excellent in navigating the social network around them in spite of lack of any knowledge of its structure.

An interesting variation of the small-world experiment to get more insight into the navigability was carried out by Killworth and Bernard in 1978. In their experiment, termed as the **reverse small-world experiment**, participants were asked what information they would like to know about the target if they were participants of the standard small-world experiment. Note that no actual messages or anything were passed from person to person. The experimenters were only interested in finding out how humans are able to find short paths based on incomplete information and what pieces of information are most relevant for the task. It was found that the most sought out characteristics about the targets were the name, location and the occupation of the target. Incidently, these are the exact three things that were provided to the participants in the Milgram's experiment!

Navigability is still an ill-understood area of network science, and on a broader level, connecting psychology of the humans to the network structure they form remains a formidable challenge.

Modern versions

In the digital age, it has become much easier to conduct experiments similar to the small-world experiments, and also on much larger scales. For example, Dodds et al repeated the small-world experiment in 2003 using emails instead of actual letters. In their experiment, 24000 chains were started with 18 targets in 13 different countries. Out of the 24000 chains, only 384 were completed which is 1.5% success rate (compare now this with 19% success rate in Milgram's experiment). Having at hand much larger data and better statistical methods than Milgram, Dodds et al reached the conclusion that the average distance is between 5 and 7, very similar to what Milgram had got.

2.4.1 Locating hidden populations with the help of networks

An interesting application of the existence of social networks is to locate or probe the hidden population. Certain populations like gangs of criminals, that of drug users, populations of illegal immigrants etc are hard

to probe because of the very fact that they don't want to reveal their nature or activities. As we will see in a moment, the structure of their social network can be cleverly exploited to gather their information. Different techniques exist for this purpose, but we will restrict ourselves to three prominent ones among them.

Snowball sampling

In this method, we first somehow locate just one or two people from the population we want to probe. We can then discuss and interview them. After we gain their trust, we can request to name some of their contacts, and after getting those contacts, request them to allow us to interview them. Since such requests are usually accompanied by someone who they are familiar with (first persons in our study), they are mostly successful. But then we can repeat this procedure, and ask these contacts to name some of their contacts and so on. This quickly leads to a data of a large number of people from the population, and with each round, we can see that the population increases like a snowball (hence the name). However, we should keep in mind that in the structure of the sampled network gathered in this fashion is highly biased for various reasons, and gathering the network structure is not actually the aim of the method. Rather, we simply leverage the existence of the network to reach more people from the hidden population.

Contact tracing

This is a method similar to the snowball sampling that is usually used to trace the people with a particular infection. As an example, suppose that we find a person who is infected with HIV. It is then important to find out other people in the population from who s/he might have got the infection or to whom s/he might have infected. To do this, we ask the person whether s/he had sexual contact with anyone or whether s/he used the same needle used by others (this happens with high probability if the person is a drug user). If we find such people, we go to them and repeat the procedure. Usually, we stop when we get a person who is not infected, and so in terms of sampling the network structure, this is an even more biased method. But similar to what is said above, the aim is to find out and heal the infected people rather than find out the structure of the network.

Random walk sampling

Sometimes a bias in the sampled data could be reduced to some extent if we modify the procedure used in the snowball sampling. There, once we get the names of all the contacts of a given person, we interview all of them. In the random walk sampling, we just choose one of these contacts randomly, and interview only him/her. To understand why this reduces the bias needs some understanding of the mathematics of networks, and we will come back to this issue once we have gathered sufficient amount of mathematical machinery.

2.5 Biological networks

2.5.1 Metabolic networks

Metabolism is a process which converts food and nutrients into useful biomolecules for biological cells. This happens in two stages: in the first stage (called *catabolic* metabolism), food is broken down into useful building blocks. Then in the second stage (called *anabolic* metabolism), these building blocks are assembled to form useful biomolecules. Each of these stages consist of a sequence of chemical reactions known as pathways, and the set of all chemical pathways forms a metabolic network. A rationale behind the study of the metabolic networks is that they would provide useful insights into the complex dynamics of a biological cell, which in turn would enable us to control it. Thus, the vertices in a metabolic network are various chemicals, and are known as *metabolites*. The metabolites by definition are small molecules (i.e. molecules consisting of not too many atoms) as opposed to macromolecules like DNA and proteins. The macromolecules are not products of any metabolic reactions, and other complex processes are used to make them.

One should not think of the process of metabolism as a simple sequence of reactions. The concentrations of the metabolites in the metabolic network vary widely and rapidly as per the requirements of the cell.

A biological cell is a highly dynamic entity that has a capability to respond to changes in its internal and external environment by switching on and off productions of various metabolites or even that of switching entire portions of the metabolic network. Usually, in a given chemical reaction inside a cell, a number of metabolites are involved some of which enter as *reactants*, and then several are produced as *products*. The mechanism by which cell controls these chemical reactions is the use of special macromolecules called *enzymes*. Most of these reactions are not thermodynamically favoured, and hence to increase the reaction rates, enzymes are used as catalysts. Enzymes are usually proteins but sometimes RNAs also act as enzymes.

A more accurate representation of a metabolic network is a bipartite network in which one type of vertex are metabolites and the other type of vertex is chemical reactions. A metabolite is connected to a reaction if it is involved in that reaction either as a reactant or a product. Even better, we could make the edges directed so that if a metabolite goes in a reaction, then directed edge is from the metabolite to the reaction. If a metabolite is a product of a reaction, we could make the direction of the edge from the reaction to the metabolite. Note that enzymes, in spite of being of paramount importance to the metabolic network, are not included in this representation. But this is easy to do; we simply include another type of vertex to represent enzymes! Then we could join the enzymes and the reactions they catalyze by undirected edges. Thus, the metabolic network is really a *tripartite* network that is partly directed and partly undirected. Although this is the most correct representation of a metabolic network, it is seldom used, and most often the network is represented simply as consisting of metabolites connected by directed edges from reactants to products. However, one should keep in mind that this discards a huge amount of information. For example, if two reactants are involved in a reaction that produces two products, there will be edges between both reactants and products. However, by looking at the network, it would be impossible to see that these four are really a single reaction.

2.5.2 Protein-protein interaction networks

As explained in the previous section, the cell controls the metabolic networks with the use of catalysts called enzymes. Most enzymes are proteins, and each enzyme is specific to only few reactions i.e. it can control only some of the reactions in the metabolic network. However, these proteins don't work in isolation, and need to interact with each other since each reaction is usually controlled by several proteins. Proteins are actually macromolecules, and the interaction between them is physical, not chemical (by this we mean that after interaction, the chemical structures of the proteins involved in the interaction don't change). The result of such interactions is the formation of what are known as *protein complexes*. As a naive model, you can imagine each protein to be a kind of a fork which can handle certain metabolites, and two proteins can bring two or more metabolites closer by interacting with each other. This increases the chance that the metabolites in question would interact chemically.

The set of all interactions between the proteins forms the protein interaction networks. It is easy to see that this network is undirected. It is important to note however, that many times more than two proteins interact with each other physically. However, in the network these interactions are represented by different edges, and it is impossible to see that more than two proteins are in fact part of the same protein complex. A remedy is to present this as a bipartite network of proteins and complexes; however, such representations are quite rare in practice.

2.5.3 Gene regulatory networks

So far we saw that the metabolic network inside the cell is controlled by proteins which do that by interacting with each other physically, and so form their own interaction network which controls the metabolic network. However, as mentioned earlier, proteins are macromolecules, and so are not the products of the metabolic reactions. But it is important for the cell to control their production to in turn control the metabolic network. This information about making of proteins is stored in the DNA of the cell. We can imagine proteins as a type of assembly toys that many of you would have played with in your childhood. In such toys, there are a few basic pieces which can be assembled in a variety of ways to make a variety of toys. The basic units for proteins are called **amino acids** and there are just 20 of them in all living organisms. Amino acids are the products of metabolic reactions and are small molecules. However, different combinations of these can give rise to thousands of types of proteins that in turn can control thousands of chemical reactions in the cell. In

fact, each protein is just a simple chain of large number of amino acids. When such a chain gets assembled inside the cell, depending upon the sequence of amino acids, it folds itself under thermodynamic forces in a specific structure or **conformation**. Because a given conformation is a result of a specific sequence of amino acids, and because conformation dictates which metabolites it can control, the cell needs different sequences (i.e. different proteins) to control different reactions.

A molecular machinery that makes proteins is called **ribosome** which is a complex made up of proteins and RNAs (I find proteins making other proteins somewhat cute!). But somebody needs to tell the ribosome which protein to make or equivalently which sequence of amino acids to use to form a chain. Exactly this information is stored in the DNA, and that is the primary function of the DNA in living organisms. DNA is also a macromolecules that is made up of special units called **nucleotides** arranged in the form of a long chain (in fact there are two such chains in each DNA molecule forming the famous *double helix structure*). There are exactly four types of nucleotides in living organisms: adenine, cytosine, guanine and thymine abbreviated as A, C, G and T. Each chain or strand of DNA is thus just a sequence of nucleotides which looks like 'GTTTCTCAGCCTTAGACCAGATAGCTGGTG...'. Each amino acid is encoded in this chain as a sequence of three nucleotides like GTT. Such sequence of nucleotides is called a **codon**. Thus, every protein, which is a finite sequence of amino acids, is encoded as a sequence of codons on the DNA. In fact, each DNA sequence codes for a large number of proteins, and so there needs to be something on the DNA to distinguish encodings for different proteins. For this, two special codons exist called **start codon** and **stop codon** that separate encodings for different proteins. Each such encoding between a start and stop codons is called a **gene**.

This might sound intimidating, but we are not done yet! Remember that even though a protein is encoded as a gene on the DNA, actual assembly of a protein is done by the ribosome. Thus, somebody must "note down" the sequence of codons on a gene and give it to a ribosome so that it can make that protein. Exactly this is done by a special enzyme molecule called **RNA polymerase**, and the "notepad" it uses to note down the sequence on a given gene is another molecule called RNA. This process of noting down or copying the gene is known as **Transcription**. Once the sequence is transcripted on the RNA, it is called "messenger RNA" (probably because it is a message from a DNA to the ribosome). Ribosome takes this sequence in the messenger RNA, and assembles the corresponding protein. We then say that the 'gene is expressed'. But hold on! Does that mean that every gene encoded on the DNA is continuously used to make the corresponding protein? No! As seen earlier, the main job of proteins is to act as catalysts in the metabolic reactions and hence to control the metabolic network. For this, the cell must be able to switch on and off the production of the proteins in the first place. To achieve this, the cell uses an ingenious mechanism. Recall that it is RNA polymerase that copies the sequence of codons in a gene. It does this by moving along the gene. But this is not a thermodynamically favored process, and to make it so, another type of molecules called **Transcription factors** is used. Transcription factors are are just proteins that attach to the starting region of the gene (called **promoter region**) and this kind of unlocks the gene which can then be "seen" by the RNA polymerase to copy it. A transcription factor need not always favor the process of transcription; it may also *inhibit* it. A given transcription factor is specific to one or more proteins. In other words, many proteins and RNAs are involved in the process of making other proteins (copying, building etc).

Now the interesting point: since transcription factors themselves are proteins, they are also encoded in other genes and are built by the same process mentioned above. Thus, to start the expression of a particular gene, another gene corresponding to the transcription factor of this gene should first be expressed. In other words, one gene can control (either promote or inhibit) the expression of another gene through its own expression, which in turn can control expression of other genes and so on. Thus, we can imagine a network of genes in which there is a directed link from gene A to gene B if the former controls the later. If we also want to include the information about the promotion or inhibition, we could include two types of links in the network. This network is the gene regulatory network of the cell. Although the gene regulatory network fundamentally controls the dynamics of living cells, its understanding is to a large extent incomplete, and the hope is that its better understanding will lead to answers to many pressing problems in genetics.

2.6 Brain networks

One of the primary functions of the human brain or the animal brain is to process information. Interestingly, this information processing can be viewed as a complex network in two different ways. Let us have a look at both.

2.6.1 Network of neurons

The brain is made up of, among other things, a specialized type of cells called **neurons**. An individual neuron can be thought of as the smallest unit that is used to process information. The brain of simple animals like a worm usually contains a few hundred neurons. On the other hand, the brain of a complex animal like human may contain more than 100 billion neurons! A simplistic model of the brain is a network with neurons as the nodes and connections between them as edges. An edge in this network is called a **synapse**, and so neurons communicate with each other via synapses. It is important to note that a synapse is directed, and so this network is a directed network. Fig. 2.6 shows the structure of the neural network for a simple worm called **C. elegans**. This network contains just around 300 neurons, and has been mapped completely and accurately. This is not the case for the more complex brains like human brains, and mapping the network structure remains a formidable challenge in this case.

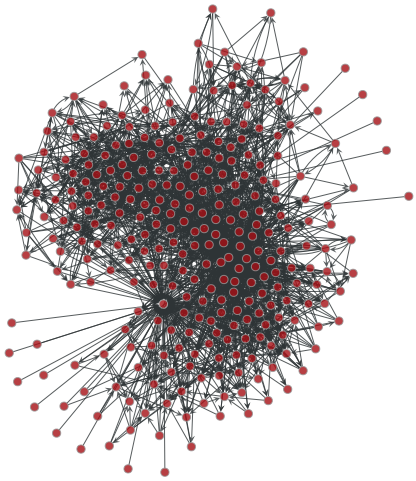

Figure 2.6: The network of neurons for the brain of the worm *C. elegans*. This network contains only around 300 neurons, and researchers have been able to map its structure completely.

We won't discuss here how an individual neuron functions or what is its structure; all that we need to know is that a given neuron has a number of incoming links and a number of outgoing links. The actual communication between the neurons happens with the help of an electrical signal called **action potential**. Each neuron can generate action potential that gets fed to the neurons it is connected to by its outgoing links. Thus, a neuron usually receives action potential from many different neurons via its incoming links, and when the sum of the action potentials exceeds a certain value, a *threshold* for that neuron, it generates its own action potential. This process is known as **firing**, and whenever this happens, we say that the neuron has fired. A synapse need not always increase the chance that the next neuron would fire. If a synapse increases the chances for the next neuron to fire, it is called an **excitatory synapse**, whereas if if it decreases the chances, it is called **inhibitory synapse**. The actual working of the network of neurons, in particular how it gives rise to things like decision making, emotions and consciousness is still not well understood, and is one of the most active areas in modern science.

2.6.2 Network of brain areas

Another way in which the brain can be thought of as the network is by considering individual areas of the brain as the nodes. Each such area would contain a large number of neurons, and this reduces the level of complexity of the neural networks. What is observed in the neuroscience studies is that the individual areas in the brain get excited during a particular activity. For example, a certain area might get excited in response to listening music, whereas another may get excited by hunger. More interestingly, one observes that for a given activity several areas of the brain may get activated. In this case, we could reasonably assume that the areas which get activated together are *functionally* connected. A popular way to detect such electrical activity in the brain and to find functional correlations in the brain areas is the use of fMRI (functional magnetic resonance imaging). Apart from these functional connections, it is also found that many brain areas are actually physically connected by bundles for neurons. A different technique called *diffusion MRI* is used to probe the existence of such connections.

Just like the study of the network of neurons, the study of network of brain regions is an active area of research, and in modern times has provided important insights into the way the brain works.

2.7 Ecological networks

An **ecosystem** is a group of biological species (organisms ond plants), that interact with each other in a restricted geographical region. Some examples of ecosystem are lakes, mountains, forests, and islands. A network with the species in an ecosystem as vertices and interactions as edges is called an *ecological network*. Usually ecosystems are to a good approximation isolated from the rest of the world in terms of interaction of species. For example, consider organisms in a particular lake. There could be different types of fishes, insects, and plants which can interact with each other. However, these rarely have any interaction with the animals outside of the lake. In a given ecosystem, many different types of interactions exist, and depending upon the type, we can think of different types of ecological networks.

2.7.1 Food webs

An obvious type of interaction between different species is the predator-prey interaction. If organisms of spcies A eat animals of species B, then A is called a predator of B and B is called a prey of A. Notice that this labelling is for a given pair of the species A and B: if species B eats species C, then in this interaction B is a predator and C is a prey. The network of species with predator-prey interaction is called a *food web*. You might have heard the term *food chain*, but this is a wrong view of the complex predation patterns in an ecosystem. It should also be quite obvious that a food web is a directed graph. If A is a predator of B, it looks natural to have a directed edge from A to B in the food-web. However, ecologists view these edges as carrying the flow of energy in the network, and so edges are directed from preys to predators in food-webs. Fig. 2.7 shows a food web for species from Antarctica (including humans!).

An important thing about food-webs deserves a mention. Sometimes several different species have the same or at least very similar pattern of predation and preyingi.e. all these species might have the same set of predators and the same set of preys. In such case, although the species are different, representating them as different nodes in the network doesn't convey any extra information. Therefore, often these species are merged into a single node, and we can think of that as a species in its own right. In ecology, they go by the name **trophic species**. In Fig. 2.7, two nodes birds and fish are examples of trophic species: although there are many different species of birds and fishes, they have the same predation pattern in the web.

An interesting property of the food-web is that it rarely contains cycles: if a species A preys on species B, and if the species B eats the species C, it is extremely rare that the species C would prey on A. Because of this reason, nodes in a food-web can be arranged vertically in such a way that all the arrows point in the up direction just like Fig. 2.7. The fact that this can be done may not be obvious at first sight, and we will look at this property of networks without cycles more closely later in the course. Different levels in this representation are sometimes called **trophic levels**, and sometimes they prove useful because a trophic level contains some information about the physical size and the population of a given species. Organisms with lower trophic level are usually smaller in physical size as well as more numerous than those with higher trophical level. This can sometimes be exploited in estimating populations of species for which actual

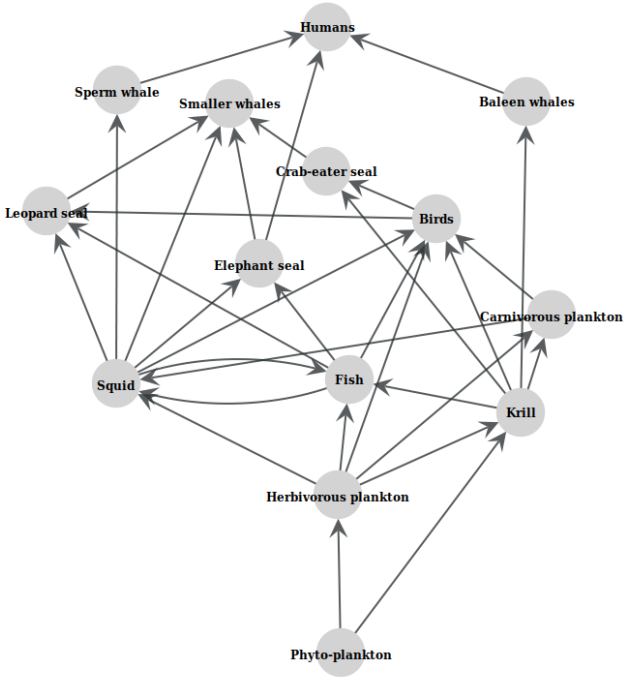

Figure 2.7: The food-web for the prominent species in Antarctica. The vertices are arranged so that almost all the arrows go upward.

counting is impossible (e.g. insects). Fig. 2.8 shows a bigger food-web from the Serengeti savana ecosystem from Tanzania.

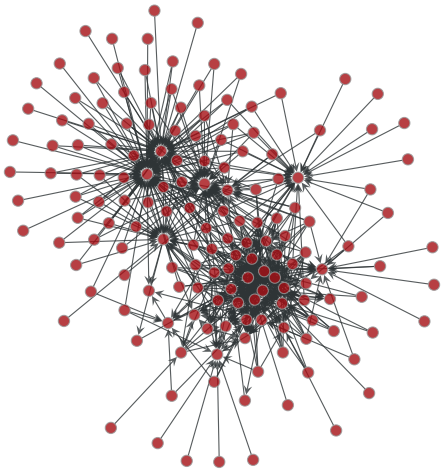

Figure 2.8: The food-web of the Serengeti savana ecosystem from Tanzania. There are total 161 species in this network.

2.7.2 Other ecological networks

Food-webs are not the only type of ecological networks although they are probably the most studied ones. Other types of networks that exist include **Host-parasite networks** and **mutualistic networks**. Host-parasite networks are similar to the food-webs except that the parasites usually don't kill their host. Mutu-

alistic networks represent different types of symbiotic relationships that are often present in organisms. Both these types are relatively unexplored, and it is reasonable to assume that many fascinating results about them are awaiting discoveries.