1

Introduction

The database UniProt (The-UniProt-Consortium, 2021; www.uniprot.org) contains lists of proteins, with each protein annotated with a wealth of information about its structure and biological importance. For *Homo sapiens*, for example, in January 2021, there were 77 027 proteins listed. The interactions between proteins, for instance within individual cells, are the basic elements out of which the reactions that allow an organism to function are built. Those known interactions are collected in the database BioGRID (Stark *et al.*, 2006; https://thebiogrid.org); in January 2021, again for *Homo sapiens*, there were 553 923 (non-redundant, physical) interactions listed, and in the following 5 months, more than 20 000 further interactions were added. On the basis of this information, it is natural to build a network of vertices and edges, in which each protein is represented by a vertex, and there are edges between pairs of proteins that have been observed to have at least one interaction; a 'protein–protein interaction', or PPI, network. One might then expect the structure of the network to yield interesting biological information, and indeed it may. As an example, it has been shown to be possible to build the phylogenetic tree, relating a collection of species, that is currently accepted as being correct, using only the topological similarities between their protein interaction networks (Ali *et al.*, 2014); see Example 25.6.1. Indeed, such network information may be preserved over much longer evolutionary times than sequence similarity, which forms the basis of traditional tree reconstruction methods.

We noted earlier one example of how networks, as representations of complex data sets, have become useful tools in many areas of application. Sociologists and ethologists have long been interested in networks of relationships (edges) between individuals (vertices), and in their implications, for instance, in the workplace or in family groups. Here, it is natural for edges also to have a direction, in that a relationship between a manager and a subordinate or between an adult and a juvenile is not reflexive, and each direction of the interaction may have its own significance. Some networks, such as telephone and communication networks, are physical, with vertices representing equipment such as switches and computers, and the edges representing cables or microwave links. For these, the capacity of both cables and switches are important quantities, so that both edges and vertices naturally have 'weights' attached to them, which are relevant when optimizing the flow of traffic within the network. The World Wide Web forms a network, with web pages as vertices and hyperlinks as (directed) edges between them, and platforms such as Facebook and Twitter (now called X) also have associated networks. Some networks are small and can easily be pictorially represented: the Florentine marriage data of Padgett & Ansell (1993), described in §2.1, have only 16 vertices and 20 edges. Others, like the World Wide Web, are huge, and the sets of vertices and edges are forever changing. Most information about its graphical structure is obtained by exploration, using web crawlers, and cannot be expected to be complete. Protein–protein interaction networks themselves are derived from protein databases in which

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press

many proteins have not been extensively researched, so that there may be many interactions that have not yet been observed; as noted earlier, the number of interactions in BioGRID is still growing. Thus, there are likely to be edges that should be present in the network, but are not. There may also be edges present that should not be, owing to experimental error; the actual number of edges is unknown.

Faced with data in the form of a network, a first question to ask is 'what does the network look like?' For a small network, and even for some larger networks, it is possible to draw pictures to represent them, so that patterns and structures can be recognized by eye. However, most visualizations are two or three dimensional, enhanced by colour. The actual topology of a network need not correspond to that of a low-dimensional space, so that such a representation need not make sense and could even be deceptive. Similar problems have long been recognized in the more classical field of cluster analysis, where a pictorial representation can at times be very useful but may also distort the truth; see Myatt & Johnson (2009, Chapter 2). There are many methods for depicting networks, and they are the subjects of much research; see, for example, the compilation in Bekos & Chimani (2024). Using the R package `igraph`, the standard layout in which a network is drawn is the so-called Fruchterman–Reingold layout; sometimes we use other layouts, such as a circle layout. The choice of best layout is subjective; see Li, Liu & Wang (2016) for an empirical study with 30 participants.

For larger networks, it becomes progressively more difficult to find simple representations, though a number of methods have been developed (Battista *et al.*, 1999; De Oliveira & Levkowitz, 2003). Since even a network with 100 vertices already has almost 5 000 potential edges, and the number of potential edges grows as the *square* of the number of vertices, the sheer size of the structure rapidly makes depiction very challenging. Indeed, the objects that a network might be expected to resemble are other (better known) networks, and these are still much too complicated to describe in simple terms. Hence, tools for network analysis are required that go beyond visualization.

For extremely large networks, such as the World Wide Web, the complete network is usually not known, even at a given instant. Local exploration then becomes the most feasible procedure for investigating its structure. Starting from a single web page, its neighbours can be examined, and the neighbours of its neighbours (and the connections between them), and so on; one could think of rings of neighbours that are successively explored. In such a procedure, the numbers of vertices to be examined at each step is typically a multiple of the number at the previous step, much as in contact tracing for an epidemic disease, growing roughly geometrically with distance from the original vertex. This rapidly makes a full examination of each ring of neighbours impractical. Nonetheless, some questions of interest about the network can be answered in this way. For instance, by examining the neighbourhood rings up to distance r initiated from two different web pages, it can be checked whether there is a path of length at most $2r$ linking them, by seeing whether the two sets of vertices have non-empty intersection. In such a way, a brute force assessment of the distribution of the distance between pairs of web pages could be made: though, in a network whose scope is unknown, choosing 'random' pairs of web pages is itself a difficult problem. Another procedure, designed to avoid the geometric growth of neighbourhood rings, is to choose a *single* vertex at each step as the next to be examined, the vertex being chosen at random from those neighbours of the current vertex that have not already been examined. This yields a collection of vertices that have been selected by a random walk on the vertex set of the network. Then, to form a sample sub-network, all the edges that link them are included.

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press

Such sampling methods are also used for eliciting information from PPI networks. For instance, in research on a particular disease, a certain set of proteins may be known to be implicated in its development. It is then advantageous to examine the sub-network consisting of a neighbourhood of these proteins, to find more proteins that may also be associated with the disease. This is important because, if treating the disease involves suppressing the generation of some of the proteins responsible for the disease, their roles may be taken over by other proteins, which would also need to be suppressed. These can be expected to be proteins that belong to the sub-network. Illustrations of such applications are to be found in Examples 3.3.15 and 19.2.3, and in §21.5.

The subsets of vertices that are obtained by these methods are necessarily linked by edges to the initial set of vertices. As a result, the number of edges between them may exceed what would be expected of vertex pairs chosen at random from the entire vertex set. This effect may be enhanced if vertices close to each other on a path are more likely also to be joined directly by an edge, and vertices with large numbers of neighbours are more likely to be represented in such samples than those with few (isolated vertices never being visited, unless initially chosen). Thus, the topology within the sample sub-network may differ from that of the entire network. This has to be taken into account, when using the sub-network to deduce properties of the full network.

Even if a network is available in its entirety, it may be too complicated to comprehend in full detail. It may nonetheless still be possible to pick out some salient features, much as, in classical statistics, a random sample may be summarized by mean and standard deviation. In a network, there are many possible summary statistics that might be considered; for example, the *(edge) density* of the network (the proportion of possible edges that have actually been observed), the proportion of vertices that are isolated, the number of connected components in the graph, the numbers of copies of specific small subgraphs (such as triangles) in the network and so on. A variety of possibilities are discussed in Chapter 3. These, in themselves, only give a restricted picture of the network. However, a more informative image can then be obtained by matching the values of the statistics to those typical of some known, well-defined, family of networks, with the hope that a good match can be found. If so, it may be that further properties of the network can be derived, by analogy, from those of the best matching known networks. Furthermore, if the numbers of copies of certain small subgraphs (patterns) in the network are much larger (or much smaller) than what would be expected in a known family of networks, this may give information about what the network represents. For such judgements to be made, it is important to be able to describe the distributions of the values of the statistics to be expected from the known families of networks, and this in turn necessitates the probabilistic analysis of the best-known network models. If only a sample of the entire network can be obtained, any comparison with a well-known family has to be made by comparing it to the statistics of samples obtained according to the same procedure from the known family, as in §21.5.

What constitutes a 'known' family of networks? The most straightforward answer is perhaps a network that has been generated according to some relatively simple set of rules. The easiest of these is the rule that each possible edge is actually present in the network with some fixed probability p, and absent with the complementary probability $(1 - p)$, independent of all other edges. This rule gives rise to the family of Bernoulli random graphs – a whole family, because, for each specified number n of vertices, the probability p can take any value between 0 and 1 – and many properties of Bernoulli random graphs are by now well understood. Note that, using such a rule, the resulting graph is a particular outcome of a *random* process, and, in principle, *any* graph on n vertices could arise as a possible outcome. However, *typical* outcomes have many features

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press

that are observed in more or less every outcome, so that, in this sense, Bernoulli random graphs have a structure that can be observed and understood. For instance, the number of edges present, in a Bernoulli random graph on n vertices that has edge probability p, has the binomial distribution $\mathrm{Bi}(N, p)$, where $N := \binom{n}{2}$. Hence the density of the graph, obtained by dividing the number of edges by N, has mean p and variance $N^{-1}p(1-p)$, which is small if n is large. In this sense, a Bernoulli random graph 'typically' has density close to its value of p, and, if an observed network were to be compared with Bernoulli random graphs, the obvious comparison to make would be with Bernoulli random graphs that have p chosen equal to the observed density. Other typical properties of Bernoulli random graphs can deduced from the basic definition of the Bernoulli family, but with more difficulty, since the arrangement of the edges into a network complicates matters. For instance, the distribution of the number of triangles in a Bernoulli random graph already involves dependencies, arising because the occurrences of triangles that share an edge are dependent events. For finding such distributions, suitable probabilistic methods need to be devised.

The rule that gives rise to the Bernoulli random graphs is usually too simple to describe the networks encountered in practice. Because all edges are present or absent independently of all other edges, the resulting graphs are those exhibiting a complete absence of systematic structure, and their main use is perhaps as a null model, to indicate what is to be expected if a network indeed has no structure. Real-life networks have often arisen as a result of mechanisms that are unknown, but which it may be useful to attempt to describe; the way in which the networks differ from Bernoulli random graphs may give some information as to what the mechanisms may be. Real-life networks can also be compared to the graphs that are typical of other sets of rules, to see if they give a better fit; to judge this, probabilistic analysis of the different network models is needed. Protein–protein interaction networks turn out to be rather different from the graphs typical of the standard models described in this book. However, a set of rules for generating graphs that is specifically based on a crude description of the way in which proteins may evolve in nature, known as the duplication–divergence model, turns out to yield graphs that have subgraph counts resembling those found in PPI networks better than the others; see §4.6 and Example 25.6.1. Thus, the network structure appears to contain information about how it evolved, and this, in turn, can influence our understanding of the evolutionary mechanisms that brought it about; see Gibson & Goldberg (2009) for some of the biological implications.

There are many features of networks that convey information about the context from which the network arose; a number of them are discussed in Chapter 3. One such is the set of degrees (numbers of neighbours) of its vertices. As an elementary example, vertices of very high degree, in a sociological context, may correspond to very influential individuals (Friedkin, 1998, Chapter 6). On a more sophisticated level, the lack of uniformity in the collection of vertex degrees in a network may be one indication of the way in which the network has been generated, and how it may differ from realizations of Bernoulli random graphs; see §4.4 for a family of examples. Another aspect of a network that may be particularly worth investigating is whether or not there are groups of vertices that have many more edges between them than would be expected, if vertices were joined, in some sense, at random. Here, the number of edges to be expected within a group of vertices should be computed relative to the degrees of the vertices, so that the information is (as far as possible) separate from that inherent in the vertex degrees. Finding such groups, a process known as community detection (see Chapter 24), can be useful, for instance, for discovering groups of proteins that cooperate in a common set of reactions, perhaps reflecting a particular function

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press

within a cell. For instance, Luecken (2016) finds such groups in a PPI network of breast cancer hypoxia data and uses them to shed new light on the genes that are involved in vesicle transport; this, in turn, is related to the invasiveness of a tumour.

The examples described previously illustrate that the structure of the network itself may be an object of considerable scientific interest, and that finding an appropriate model fit may help to answer questions of practical importance. In particular, a network offers an explicit description of the relationships between entities, coded as vertices. However, in many applications, it is the entities themselves that are the main objects of interest, and the aim is to investigate and compare various of their attributes, whose measurements constitute the primary source of data, while making proper allowance for their inter-relationships. To take a classical problem, suppose that several quantities are measured at each vertex, and that the value of one of them is to be modelled as a response to the other values; for instance, the relationship between secondary school grade and academic attainment at university, as considered in Example 25.2.2. For independent observations, this would correspond to a (possibly non-linear) regression problem, for which least squares regression offers a straightforward means of inference. However, the measurements at pairs of vertices that are linked by an edge can be expected to be related, just because the vertices themselves are close to one another; university students in the same working or social groups may well have an influence on each other's progress. Such problems have long been considered for geographical data, and a key step there is to model the correlation structure in a way that matches the geography. The underlying geographical space is typically a Euclidean space, frequently in two dimensions. Unfortunately, not all networks are well adapted to correlation structures of this sort, since there may be no natural underlying Euclidean space, and, even if there is one, finding it already poses a complicated problem. If estimation is carried out using least squares, assuming the wrong correlation structure may not be disastrous for estimation, but it is damaging to any measure of precision of the estimates, such as confidence intervals and significance tests. So solving such problems for general networks is of considerable importance.

Other statistical problems arise from the network structure itself. For instance, if a network has been split by some algorithm into several communities, a question that it is very natural to pose is whether two of the communities should be amalgamated or whether a particular community should be split. Indeed, such judgements are often made within the algorithm itself. To decide such problems, in a statistical framework, a model should be chosen for the structure of the network that is rich enough to allow for different assignments to communities, but simple enough for tests, maybe based on simulation, to be carried out. Once again, this may make good sense for networks that can be reasonably matched to known model families, such as the Erdős–Rényi mixture models and stochastic blockmodels introduced in §§4.3.2 and 4.4.6, but is more questionable for those that do not.

Another problem concerning the structure of the network is to detect errors in its specification. Many protein interactions are not present in BioGRID because the relevant experiments have not been carried out, as is evidenced by the rapid growth of the number of such interactions over time. However, there may be a high false positive rate as well (Deane *et al.*, 2002). In Braun *et al.* (2009), it is found that none of the standard experimental techniques for detecting PPIs detected more than 36% of a known set of interactions. Hence, methods for validating the presence and absence of edges, using network data, can be important (Chen, Deane & Reinert, 2008); some of these are discussed in §25.4.

Many proteins, in particular those in lesser studied organisms, do not have informative func-

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press

tional annotations. Since proteins often cooperate to achieve a certain biological function, proximity within a network, or comparison with analogous networks in related organisms, can be used to predict which functions a protein can take on (Chen, Deane & Reinert, 2007); see also §25.5. Using the network to predict the values of variables associated with a vertex from those measured at other vertices is not restricted to PPI networks; an example related to political activism is given in Idan & Feigenbaum (2019).

Understanding the structure of a network can be important for many reasons. A few of these have already been mentioned in the context of PPI networks. In applying measures to control the disease HIV, the structure of typical networks of sexual contacts is an important input. For instance, can 'high risk' groups of individuals be identified, whose members each have many contacts within the group? And, if so, to what extent should control measures be concentrated on such groups, if the best outcomes are to be attained? The rate at which an airborne disease spreads has also been shown to depend on the structure of the network describing contacts between individuals (Volz, 2008) through more than just the average number of contacts per individual; COVID-19 provides such an example. The structure of an infrastructure network is important, in relation to its stability, should some of the edges or vertices be removed (fail); this leads to questions as to what structural elements in a network are important to achieve stability, and how best to incorporate them without increasing the cost of building and maintaining the network; see, for example, the survey in Pérez-Rosés (2018). Analysis of terrorist networks, such as that in Example 26.2.1, can be used to motivate and evaluate strategies designed to disrupt them.

This book provides a collection of ideas and techniques that could be put to use to address such questions. For reasons of presentation, attention is mainly restricted to undirected and unweighted networks. Some of these ideas and techniques are interesting in their own right; all of them are useful in applications. Most of our statistical analyses have been carried out using R (R Core Team, 2019), and in particular the `igraph` and `statnet` packages. R has also been used to create the plots, except for those taken from other sources. In applications, as with any statistical problem, subject specific information is likely to be available, and this needs to be incorporated into the analysis; for example, in the spread of foot-and-mouth disease, details of the relationship between farms and abattoirs, and of the mix of livestock on the farms (Keeling *et al.*, 2001). In the same way, particular applications lead to particular questions that need answering. Those expert in the field of application should be consulted at every stage of the analysis, to make sure that the right information is being used and that the right questions are being investigated. Given this input, the book offers a toolkit for the following:

- devising appropriate summary statistics;
- constructing a plausible model;
- describing the typical behaviour of the statistics under the model;
- fitting the model to the data, and, if necessary, modifying the model; and
- drawing statistical conclusions from the analysis.

We hope that the reader will enjoy putting these ideas into practice.

https://doi.org/10.1017/9781009651752.002 Published online by Cambridge University Press